

**Comments on Papers on Theoretical and Methodological Issues in the
Use of School-Level Data for Evaluating Federal Education Programs**

Robert L. Linn

University of Colorado at Boulder

National Center for Research on Evaluation, Standards, and Student Testing

Comments prepared for a Symposium on the Use of School-Level Data for Evaluating Federal Education Programs. Washington, DC: The Board on Testing and Assessment. The National Academies, December 8, 2005.

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.

Comments on Papers on Theoretical and Methodological Issues in the Use of School-Level Data for Evaluating Federal Education Programs

The four papers by Elizabeth Stuart, Yeow Meng Thum, Michael Scriven, and Gary Miron presented as part of the symposium on the use of school-level data provide an excellent foundation for considering the theoretical and methodological issues in using the School-Level State Assessment Score Database to support conclusions about causal effects. The papers by Elizabeth Stuart and Michael Scriven provide thoughtful theoretical overviews. Yeow Meng Thum's paper provides a more detailed discussion of a particular analytical approach – value-added modeling – that has attracted increasing interest in the last few years. Although applicability to the school-level database may be a question because of the lack of individual student data, a better understanding of the strengths and limitations of value-added models is certainly desirable. In his paper, Gary Miron focuses on a specific and quite controversial intervention – charter schools - and delves into the relative strengths and weaknesses of various research approaches that differ in their rigor and the degree to which they can support causal conclusions about charter school contributions to student achievement in comparison to the contributions of non-charter schools.

Stuart has clearly stated the goal of analyses seeking to support cause and effect inferences. As she says, the goal is to identify interventions that teach students more effectively. Charter schools are an example of a currently interesting intervention that some people expect to lead to more effective teaching and student learning. Certainly it would be desirable to be able to draw defensible cause and effect conclusions interventions such as increasing

funding for school libraries, the expansion of school choice, or the introduction of charter schools.

Stuart concludes that the school-level database can be used to reach cause and effect conclusions about interventions such as the examples just given, but wisely cautions that considerable care must be taken in using the database for this purpose. She also notes that additional data on school-level characteristics needs to be included in the analyses.

There is widespread agreement that a randomized experiment is the preferred study approach for making defensible causal inferences, but there are many situations where that ideal is not possible. Stuart suggests that even when randomized experiments cannot be used they still provide a template for thinking about the requirements for supporting causal inferences with the data that are available. While she argues that causal inferences can be justified without randomized experiments she adds several caveats. Like Stuart, Miron, places randomized experiments at the top of his list of alternative study designs that have been used in efforts to reach causal conclusions.

Scriven reminds us that in education randomized experiments often lack the essential features that make them the “gold standard” for pharmaceutical investigations. It is often impossible to devise and maintain a double blind experiment in education. Rather, as Scriven suggests the randomized experiment deteriorates into a “zero blind” experiment where comparison and treatment group participants not only know which condition they are in but they learn about many of the features of the other condition and may, in fact, implement some of those features. His point that a zero blind randomized experiment is actually a weak design is well taken.

Scriven makes a convincing case that there are alternatives to randomized experiments that can support causal inferences. But, he suggests that serious consideration needs to be given to the question of “whether the *present* kind of data is good enough to answer the *presently* important questions about the *current* crop of interventions.” He makes a strong case for the use of what he calls a General Elimination Model that attempts to eliminate various possible causes identified through theoretical analyses.

Stuart provides a convincing argument that we need to think in terms of replicating a randomized experiment to the extent possible through matching on observed covariates or through the use of other approaches such as an interrupted time series. She reminds us of Don Rubin’s concept of *strongly ignorable treatment assignment* as an alternative to random assignment. As she notes, two conditions must be satisfied to justify the conclusion of strongly ignorable treatment assignment. First, treatment assignment is independent of potential outcomes given the observed covariates, and second, there is a positive probability of receiving each treatment for all values of the observed covariates.

The second of these conditions can be particularly challenging in attempting to answer “the *presently* important questions about the *current* crop of interventions” using the school-level database. The fact that interventions of interest often are assigned in ways that make it very difficult to identify an appropriate comparison group according to the second requirement. This is particularly true given the information that is and is not included in the database, suggesting that, at the very least, the database needs to be supplemented with

additional data that makes it more feasible to match treatment and comparison schools.

The stable unit treatment value assumption discussed by Stuart is a major challenge for those wishing to use the school-level database to make causal inferences about interventions. The assumption has two components. First, potential outcomes are not affected by treatment assignment of other units and second component is that the treatment is administered to all units in the treatment and control group and there is only one version of each treatment.

The first assumption is often violated in studies in education because there is treatment spill over. For the school-level database it is hard to know from the available information how much of a threat this is to reaching valid conclusions.

The second assumption is also problematic in education studies. It is not always plausible that there is only one version of each treatment in educational settings, and hard to evaluate from information available in the school-level database. In fact, there may be many versions of treatment and many versions of control conditions. The Follow-through implementation studies conducted in the 1970s, for example, indicated that even in with randomization there were often many variations on the treatment actually delivered in different classrooms.

Charter school studies provide another example. There are many versions of charter schools and there are many versions of non-charter schools. What is to prevent effective charter school activities from being copied by their non-charter schools comparisons? The wide variation in both the strength and direction of effects reported by Miron for studies of charter schools may be

explainable, at least in part, by the between-state differences in what charter schools are and how they function.

Identifying a comparison group against which the performance of the treatment group can be judged is a major challenge in using the school-level database to evaluate federal education programs. There are at least two major challenges in selecting a comparison group. First, there is the challenge of finding comparison schools that look like treatment schools. By design, interventions are often target a particular category of schools, such as schools with a large concentration of students from low-income families. Targeting the intervention to reach students with the greatest need makes the treatment schools systematically different than the potential comparison schools.

Second, there is the challenge of obtaining enough information to be able to determine if the comparison and treatment schools are similar. This second challenge is behind Stuart's suggestion that database developers either include more extensive data on schools or suggest ways to get it.

Yeow Meng Thum's paper provides a fairly detailed description of an analytical approach to growth modeling, an instance of what has come to be known as value-added modeling. There certainly is a growing interest in growth modeling, in general, and value-added models, in particular. Tennessee is well known for the value-added analyses that the state has used for a number of years. Several other states, e.g., Colorado and Ohio, are moving in that direction. On November 21, 2005 Secretary Spellings announced pilot program that invites states to submit proposals for growth models as approach to assessing adequate yearly progress. The pilot program is likely to further increase the use of value-added analyses.

In her letter to Chief State School Officers, Secretary Spellings listed 7 “core principles” that a growth model would have to meet to be approved. First, the model “must ensure that all students are proficient by 2013-14 and set annual goals to ensure that the achievement gap is closing for all groups of students” (Spellings, 2005). Thus, the fixed achievement target of 100% proficient or above in 2014 is maintained. Other core principles include the requirement of setting “high expectations for low achieving students, while not setting [them on] student demographic or school characteristics,” and the requirement to make separate accountability decisions for reading/language arts and mathematics (Spellings, 2005).

The “value-added” terminology implies a causal interpretation. When teacher or school value-added results are reported it is assumed that it is the teacher or the school that is having an effect rather than some other factor such as students’ families, student background, or student peers in the school.

I agree, however, with Thum’s caution about the use of value-added models to make causal inferences about the quality of schools or teachers. Attributing value-added results to differences in school quality is not justifiable unless a variety of alternative explanations for the results can be eliminated, and there are many other alternate explanations. For example differences in educational support from home for students in schools A and B or differences in composition effects created by the peer groups in the two schools, may provide alternative explanations for the results and are not easily eliminated. School characteristics are confounded with many factors (e.g., socio-economic status, composition of the peer group, and levels of parental support).

Most applications of value-added models estimate gains based on a span of grades, the earliest of which is likely to be grade 2 or 3. In this way the value-added models control for differences in student achievement at the time of the earliest grade included in the analysis, but they do not rule out the possibility that achievement differences in kindergarten and grade 1 confound the value-added estimates. As Raudenbush (2004) has noted, “measured cognitive status prior to school entry is the most important confounder in studying school effects” (p.13).

Value-added models that use prior student achievement but do not include other student background are also subject to the criticism that the excluded variable might bias the estimates. In this regard, Raudenbush (2004) has argued that,

“... the estimation of gains does not necessarily eliminate all confounding. A critic might argue that unmeasured student characteristics predict gains students can expect and the schools they attend. This criticism is impossible to refute, though Ballou, Sanders, and Wright (2004) provide evidence that use of longitudinal data in multiple subject areas virtually eliminates the need to control for the usual confounders (ethnicity, gender, and poverty)” (p. 13).

Basically, value-added models yield results that give potentially valuable descriptive information about schools and/or classroom teachers.

Miron noted that many of the early charter school studies relied on cross-sectional comparisons of student achievement. Such a weak design could hardly be expected to shed much light on the effectiveness of charter schools in promoting student achievement. The most common approach used in more

recent charter school studies has been the comparison of achievement of successive cohorts of students. This is clearly an improvement over studies relying solely on cross-sectional comparisons. As noted by Miron, however, the interpretation of results for successive cohorts rests on two assumptions: first, that students in successive years will have more exposure to the treatment, and second, that the successive cohorts of students are comparable. Neither of these assumptions is very tenable. Thus, it is no wonder that the results of charter school studies have been so highly contested.

Use of residual gains and statistical filtering to try to isolate charter school effects are helpful, albeit not foolproof, ways making the most of the available data for successive cohorts. The odds-ratio analyses described by Miron are also potentially useful in dealing with state test score data that are reported only by performance categories. I wonder, however, how much difference well-known state-to-state variation in the stringency of their performance standards might make in the odds-ratio analyses. I agree with Stuart's suggestion that the school-level database would be more useful if average standard scores rather than only percents above a cut score could be included in the database for the schools in each state. I take it that Miron would also agree with this suggestion since he refers to percent above cut scores as the "crudest measures of student performance."

Together the four papers provide many helpful suggestions for thinking about causal inferences and the use of the school-level database for evaluating federal education programs. They should contribute to improvements in the evaluations that are undertaken using the database.

References

- Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29 (1), 37-65.
- Miron, G. (2005). *The constructive use of existing data and research for evaluating charter schools*. Paper presented at a Symposium on the Use of School-Level Data for Evaluating Federal Education Programs,. Washington, DC: The Board on Testing and Assessment, The National Academies, December 8.
- Raudenbush, S. W. (2004). Schooling, statistics, and poverty: Can we measure school improvement? The ninth annual William H Angoff Memorial Lecture. Princeton, NJ: Educational Testing Service.
- Scriven, M. (2005). *Can we infer causation from cross-sectional data?* Paper presented at a Symposium on the Use of School-Level Data for Evaluating Federal Education Programs,. Washington, DC: The Board on Testing and Assessment, The National Academies, December 8.
- Spellings, M. (November 21, 2005). Letter to Chief State School Officers, announcing growth model pilot program, with enclosures. Available at: <http://www.ed.gov/nclb/landing.jhtml>.
- Stuart, E. (2005). *Estimating school-level causal effects using SSASD*. Paper presented at a Symposium on the Use of School-Level Data for Evaluating Federal Education Programs,. Washington, DC: The Board on Testing and Assessment, The National Academies, December 8.
- Thum, Y. M. (2005). *Designing gross productivity indicators: A proposal for connecting accountability goals, data and analysis*. Paper presented at a

Symposium on the Use of School-Level Data for Evaluating Federal Education Programs,. Washington, DC: The Board on Testing and Assessment, The National Academies, December 8.