

Ten ideas that might help improve the utility of the SSASD for studying education policies

Judith D. Singer
Harvard Graduate School of Education

Discussant comments for the NRC Symposium on the Use of
School-Level Data for Evaluating Federal Education Programs
December 2005*

Over 40 years ago, John Tukey (1962) published a wonderful paper in the *Annals of Mathematical Statistics* entitled “The Future of Data Analysis.” In addition to presaging his work in exploratory data analysis, Tukey mused more broadly about what he called the “tools and attitudes” needed by professional statisticians and data analysts who tackle real-world problems. Two quotes from that paper—one by Tukey; another attributed to his colleague, Martin Wilk—seem particularly appropriate for today’s discussion:

“Far better an approximate answer to the right question ... than an exact answer to the wrong question.”

“The hallmark of good science is that it uses models and ‘theory’ but never believes them.”

I first read Tukey’s paper in graduate school and I invoke his wisdom whenever asked to provide guidance to researchers attempting to address important questions with flawed data collected for other purposes. As I was reading the materials to prepare these remarks, my original first sentence was “Is there any way I can be optimistic? Is the SSASD so profoundly flawed that nothing is possible?” Seeking guidance—or solace—I reread Tukey’s paper and asked myself: “What would John Tukey have said (or more accurately given my training, what would Fred Mosteller say)?” I have little doubt that both of them would want all of us in this auditorium to be as constructive and forward thinking as possible.

So with the stark realities of the limitations of the SSASD in mind, I’d like to offer ten ideas for either improving the use of the current dataset or improving its construction for future use. Because Bob Linn and Laura Hamilton have commented extensively on issues of outcome measurement—which are so challenging that they may well trump every idea I’m going to offer—I’m going to focus on issues of statistical modeling and data analysis. I

* 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.

present these ideas recognizing that the dataset has fundamental—dare I say insurmountable?—limitations for addressing many of the questions of interest to the Department. Yet rather than squander my time belaboring problems, I'm going to muse aloud about two broad classes of ideas: Half focus on improving the statistical models that could be used; half focus on additional analyses or additional sources of data that might be added to the database itself. The lines between these categories are not sharp, but together, I hope that these ideas might stimulate a broader conversation about how we—as an educational research community—can improve our understanding of education policies.

Five issues of statistical modeling

1. Let's be sure that our statistically models are carefully, and fully, specified. Elizabeth Stuart wisely makes the case that researchers seeking to draw causal inferences from this database need to be much more explicit about their underlying causal model. I agree fully and would extend this point further, applying it to the entire class of statistical models described in the background readings. It is not enough to state an analytic approach (e.g., regression or multilevel modeling) and a comparison group (however constructed); researchers must specify an explicit model for modeling the outcome in relation to predictors and must then clearly state all the attendant assumptions. These include assumptions about both the structural portion of the model (e.g., is it linear, non-linear, or discontinuous?) and the stochastic portion (e.g., what error structure makes sense?).

At this point in the history of educational research, I think it's safe to say that the appropriate statistical model usually starts with the student, recognizes the role of the teacher and classmates, and only then builds to the school (and school district level). Every time I think that the old "unit of analysis question" has been put to rest, I read a paper or attend a meeting where a lack of data or a misunderstanding about multilevel modeling throws me back 25 years. Just because the SSASD may not have individual and class level data doesn't eliminate the need to fully specify the multilevel model you would fit were the data available. Whether you have the data to estimate its parameters is the real question of course, but specifying what you would do if you did is a necessary first step (Singer, 1998). Starting from this ideal should clarify which parameters you would like to estimate and should help you understand how the constraints that result from not having that data limit your ability to draw inferences (causal or otherwise).

2. As a corollary, let's fully understand the consequences of omitting one or more levels from our analyses. In the past few years, there has been much interesting work investigating the effects of omitting a level from what should be a more fully specified multilevel model (e.g., Moerbeek, 2004). Most of this work focuses on the consequences of ignoring a higher level of nesting (i.e., ignoring the group composition). But the same idea should be applied to the present situation in which lower levels of nesting are being ignored. Researchers analyzing the SSASD need to evaluate the implications of these omissions explicitly and determine whether the inferences that they believe they're making can be supported under closer scrutiny. Unfortunately, in many cases, I think the answer will be no, but perhaps there are conditions under which the answer might be "maybe."

3. Let's clarify what we think a "school" represents when it's used as a unit of analysis.

Because the current SSASD provides only school level data, I found myself asking the philosophical question: Are there any analyses in which the school itself is appropriately considered an individual level? In other words, are there research questions and measures where the appropriate lowest level unit to measure and analyze is the school? Certainly there are characteristics of aggregates—e.g., principal characteristics, school size, or other measures of student composition—that are measured appropriately at the aggregate level. If there are substantively interesting outcomes that can be measured appropriately at this higher level, fruitful analyses might result. As I raise this possibility, I think it bears repeating that one cannot infer individual level relationships—which is what is often sought in these analyses—from aggregate data. The ecologic fallacy is alive and well some 60 years after its birth (King, 1997; Morgenstern, 1995).

4. Let's critically evaluate our assumptions and conduct thorough sensitivity analyses.

What assumptions are we willing to make and what assumptions are patently untenable? The Stable Unit Treatment Value Assumption required for causal inference is undoubtedly violated in most randomized trials in education let alone the observational or quasi-experimental studies we're discussing here. It makes little sense to put our heads in the sand, with a wink and a nod, saying all will be well. Researchers need to do a better job of carefully identifying the assumptions attendant to their underlying models and evaluate the sensitivity of their results to varying—or violating—those assumptions. This is particularly true when the focus on school level analyses—and the lack of individual level data—forces us to be ignorant about the magnitude of the underlying intraclass correlation. Yet it is this intraclass correlation's magnitude that determines the appropriateness of that aggregate level analysis (Singer, 1987). Thus I would urge everyone working with this data base to take the time to conduct thorough and thoughtful sensitivity analyses; my hunch is that you'll be surprised at what you find.

5. Beware the perils of standardization. In several of the background papers, researchers standardized their outcome, their predictors, or both. I understand the desire for standardization—especially when measures differ across grades and states—but many of the arguments used to justify the approach are fundamentally flawed. One line of reasoning is that standardization helps identify the "relative importance" of predictors. Unfortunately, identifying the most important predictor in a statistical model isn't that easy and standardization does little to help (Greenland, et al 1986; Healy, 1990). The other line of reasoning suggests that standardization facilitates comparison of findings across samples, yet in reality, standardization can do just the opposite (Willett, Singer and Martin, 1998). Differences across samples in the standard deviations of either the predictors or the outcome can lead to mistaken conclusions about similarities or differences in effects. The problem is even worse in longitudinal studies, where standardization within waves places unnecessary and unusual constraints on trajectories (Singer & Willett, 2003). The bottom line is that standardization will not solve the problem of non-equatable measures. Only better measurement will.

Fives additional types of analysis we might conduct or data we might collect

6. Let's appreciate the role of descriptive analyses. Tukey concluded his 1962 paper by noting that “We need to face up to the need for *both indication and conclusion in the same analysis.*” It may be heresy to say these days, but not every statistical analysis must generate a causal inference. Don't get me wrong. I fully support the current emphasis on causal inference as a much needed corrective to the decades when educators dismissed the possibility of conducting randomized trials and the need for studies that permit causal attribution. Yet just as epidemiologists have learned much from descriptive analyses—especially of longitudinal data—researchers can use the SSASD to generate interesting descriptive information that would be invaluable for both hypothesis generation and the design of future intervention studies. Let's celebrate the efforts that have gone into creating a fully representative data base. Here is a complete accounting of every school in this country. Surely a descriptive or even a relational analysis based on these data—even one that cannot support a causal inference—can have value.

7. Can we use the database to identify and evaluate the effects of natural experiments? Economists have made great strides by using extant data to identify natural experiments that can be used to test hypotheses about treatment effects (Angrist & Krueger, 2001; Meyer, 1995; Rosenzweig & Wolpin, 2000) Perhaps the most compelling of these are studies in which either forces of nature (e.g., Hurricane Katrina) or differences in government policies (e.g., different eligibility levels for scholarships across place and time) create a situation in which exogenous variation—or at least variation beyond an individual's control—provides a treatment assignment mechanism that is not entirely self-selected. This approach has been used to study the effects of treatments such as the number of years of schooling (with quarter of birth serving as the naturally varying assignment mechanism) to class size (with discontinuities guaranteed by laws creating widely different class sizes in schools of roughly equal overall size). These methods are not without problems, but I think it's worth asking whether the SSASD could be modified to provide data necessary to support the identification of natural experiments. If so, we might have another tool for evaluating treatment effectiveness in situations in which a randomized trial is not possible.

8. Can we use the database to document the changing compositions of our schools? Most analyses of the SSASD examine changes in outcome data. Yet the SSASD—as currently constructed and even more so with future additions—can also tell us about how schools themselves are changing. We could use the database to address questions about changing demographic compositions, changing teaching forces, changing class sizes, and other attributes of schools (and school districts). Here, we would not be seeking to make causal inferences about program effectiveness but rather describing how covariates change over time. Information like this could be invaluable to those designing future studies involving primary data collection.

9. Could the SSASD provide a comprehensive inventory of interventions being conducted in schools? As currently constructed, the SSASD doesn't contain much information about the various "treatments" being applied to students, teachers, classes and schools. Researchers with access to information on participation instead use school IDs to identify which schools are in specific programs (e.g, Reading First, a charter school). Could the SSASD serve as a clearinghouse describing all the many interventions currently being implemented in schools? Imagine the evaluation of a medical intervention—a new drug or a new drug education program for physicians—where the researchers didn't collect comprehensive data on the full list of medications each patient was taking. This is what's happening when we evaluate the effectiveness of a specific intervention, ignoring all the other treatments being applied. Many schools are participating in more than one study. This is especially true in the small number of very large school districts that serve a relatively large fraction of our population. And if this recommendation would come to pass, it's worth remembering that many organizations other than the US Department of Education intervene in schools. Foundations, community agencies, private philanthropies, social service agencies, not-for-profits and even for-profit organizations like textbook publishers are all active and increasingly so.

Like all top ten lists, I've saved the most important idea for last:

10. Could the SSASD include data on individual students and teachers? As Y. M. Thum points out "longitudinal student data is the key evidence base." Yet the current database fails to include the very evidence that could provide even the most basic picture of individual achievement (let alone student gains over time). I'm not just speaking of moving from percentage passing to means and standard deviations; I'm asking whether it would be possible to provide the full roster of individual student level data. After all, it is being collected; that's where these summary statistics come from. Why not include the constituent data that took all this time and taxpayer money to collect? I'm not sure if there is sufficient political will to make this change. But I believe it's worthy of discussion because it's only with individual student data that we'll begin to really answer questions about educational effectiveness. Look at the wealth of new knowledge that has been generated in other fields because social security numbers are used to track earnings, Medicaid claims, and student loan repayments. It's appropriate at a symposium such as this one to revisit the assumption that individual student data are unobtainable, because without them, we will never be able to tackle what are undoubtedly the most fundamental questions in education: questions about how students change over time (Singer & Willett, 2003).

Summing up

In offering these suggestions, I have tried to frame them in ways that Department of Education officials and researchers analyzing the SSASD will find productive. I admit that I found many flaws in some of the analyses I read and the constraints of the current data base are real and severe.

But in the spirit of seeking “an approximate answer to the right question” I think it’s worth asking whether the data base has the potential—if adequately analyzed and potentially restructured—to provide us with some indication—if not a well-crafted causal inference—about the effects of educational policies.

References

- Angrist, J. D. & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4) 69-85.
- Greenland, S., Schlesselman, J. J., & Criqui, M. H. (1986). The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology*, 123(2), 203-208.
- Healey, M. J. R. (1990). Measuring importance. *Statistics in Medicine*, 9, 633-627.
- King, G. (1997). *A Solution to the Ecological Inference Problem*. Princeton: Princeton University Press.
- Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics*, 13(2), 151-161.
- Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, 39(1) 129-149.
- Morgenstern, H. (1995). Ecologic studies in epidemiology: Concepts, principles and methods. *Annual Review of Public Health*, 16, 61-81.
- Rosenzweig, M. R. & Wolpin, K. I. (2000). Natural “natural experiments” in economics. *Journal of Economic Literature*. 38, 827-874.
- Singer, J. D. (1987). An intraclass correlation model for the effects of group characteristics on individual outcomes in studies of multilevel data, *The Journal of Experimental Education*, 55(4), 219-228.
- Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 25, 323-355.
- Singer, J. D. & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.
- Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, 33(1), 1-67.

Willett, J. B., Singer, J. D., and Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology: Statistical models and methodological recommendations. *Development and Psychopathology*, *10*, 395-426.