

## Comments on School-Level Database Design and Use\*

William D. Schafer  
University of Maryland

It is certainly a challenge to try to provide fresh insight after all that has gone on in this symposium. In these remarks, I'll emphasize a few things that have been discussed, perhaps twisting some of it a little, and a few of the things I say will be new.

### Considering Independent Variables

One focus has been about the design of the database. A useful exercise in trying to improve the database would be to determine who might have a need for it, what sorts of questions they might want to address, and what their data needs would be for those questions.

One anticipated use of the database is to evaluate educational interventions. As it is currently designed, the database seems most helpful to persons who know that specific schools are implementing particular programs, such as a district, which is evaluating some of its own schools, or an externally funded program that has identified its implementation sites.

But suppose a school's leadership wants to survey results that stem from innovations in schools that are like theirs demographically. They can see if and when other schools improved, but they have no way to investigate the programs that may have

---

\* 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.

led to more or less success. Or, there may be interest in the success or failure of programs with various combinations of elements that exist together in popular interventions.

This suggests that it would be helpful to find a way to describe educational programs in schools, and especially to describe what was done differently, year by year. Perhaps some sort of description could be contained in an external file that can easily be linked to the present database? Other ways might be easier to accomplish. But even if this could be done for a sample of schools, it should make the database more helpful to at least one class of user.

A significant barrier to describing how the programs in schools vary is that we do not know what variables to measure and include in a database. This is a problem typically faced by qualitative researchers and they have developed methodology to deal with it. Perhaps their approaches to designing a database could be used to inform the developers of the present effort about ways to provide elaborate, yet accessible descriptions of the “independent” variables in school change research.

#### Student-Level Inferences

Some powerful growth models have been described, but to what extent can they be applied to the data in our database? We have heard discussions about using data at the student level versus the school level. Here are three lines of thought we might profit from:

- The first line is to consider whether we are including all currently available information in the database that can afford inferences about student-level results without student-level data. For example, at least one paper and some

of the comments presented during these two days described analysis problems that stem from lack of data about variability in the database. Consistent with the guidelines of the APA Publication Manual, we should insist on sufficient statistics for everything we report in the database. For example, if we are reporting means, then standard deviations and sample sizes reported should be reported also.

- The second line poses a question: Are there any feasible additions to the database that would enable using student-level data? Is it possible for the database to contain some actual student-level data? Ways have been explored to introduce some student sampling and/or use alterations of data in order to protect privacy. Failing that, are there ways to coordinate and otherwise facilitate access by researchers to public, state-level databases?
- The third line adds another “me too” to David Thissen’s in support of one of Judith Singer’s points from yesterday, but with a twist. Judith and David noted that we are missing what is normally the basic level in hierarchical linear models: data from the student. But if that is a given, are there assumptions that we could make about student level data that would enable us to learn at least some of what we could from student-level growth models, when we have only school-level data? If we can answer this question, then we can evaluate these assumptions with other data in a coordinated research effort, which is an idea that stems from Donald Rubin’s remarks.

### Growth and NCLB

Predicting future growth from past growth is, of course, hazardous. The conventional wisdom is that assessment changes result in brief growth on the new measures with lesser increases thereafter. We are now seeing a large number of new assessments as a result of NCLB and can anticipate seeing growth for a few years as a result. However, linear projections of growth may over-predict and make schools appear on-track when in fact they are not. So, perhaps we should be thinking about the sustainability of growth and whether growth targets are realistic. This will become a more important question in a few years, especially in the several states that, following Ohio's lead, have back-loaded their adequate yearly progress (AYP) target increases. This point naturally leads to consideration of where achievement-level cut points are best placed.

#### Proficiency Standard Cut Points

Under NCLB, adequate yearly progress (AYP) is set based on the proportion of students who achieve at least the state's achievement level they call "proficient." The various state targets, which increase according to approved schedules, are required to converge at 100% in 2014. As we move toward 2014, then, we should expect all states to feel increased pressure to make changes in the criteria for AYP from what we can anticipate to be more and more failing schools. This may affect our beliefs about which states are setting appropriate cut points for the proficient achievement level. In the end, this is a policy issue rather than a technical one because its legitimacy depends on anticipated effects (expected impacts) rather than scientific understandings or insights.

Once we have been able to research the actual effects of the proficiency-level cut-points that have been set, we may eventually conclude that the states with the most

lenient achievement targets have the most positive influences on their educational programs. When the data become available, it would be interesting to see the correlation across states between the NAEP scale cuts for the proficient achievement level and growth on NAEP (as well as, after adjustments, state) tests.

#### State-Determined Content and Achievement Standards

As has been noted in earlier discussions, we are not like other countries in that we do not have a national curriculum. National content standards exist, such as NAEP's frameworks and the statements of the national councils of teachers in various disciplines, but states use these as advisory and are free to (and do) set their own content goals. They also set their own achievement level (proficiency) goals. What is the role of testing in this environment?

I have been thinking recently about what I have been calling the Fundamental Accountability Mission, which is to test every student on what he or she is supposed to be learning. For a statewide testing program, then, the test's blueprint (i.e., the rules used to sample the content-cognition domain for creation of new forms) can be used to define what the curricular goals are from the state perspective, as Popham has suggested it should.

If these content goals are indeed different from state to state, then why ever would we want to link the state tests to NAEP or to each other? Especially when we consider the effects of a testing program on what goes on in schools; even NAEP might become a negative influence on what a state education agency is trying to do since it could draw instruction away from the state's content specifications. Fortunately, some studies have shown that the content differences between states and NAEP are not all that severe.

Nevertheless, when there are differences, it is to the advantage of states to emphasize their content standards as opposed to NSEP frameworks in both instruction and testing.

### Ceiling and Floor Effects

A problem researchers should be aware of is the influence of arbitrary maxima and minima in student data. Because item response vectors with all zeros or all maximum scores lead to indeterminate theta estimates, states typically set arbitrary values for LOSS – the lowest obtainable scale score – and HOSS – the highest obtainable scale score that are assigned to these problematic vectors. It is also possible for other response vectors to be assigned HOSS or LOSS. HOSS doesn't occur much if at all (state tests usually have high ceilings), but LOSS does occur, and sometimes frequently, in part because state tests are based on grade-level content and students frequently fall below grade-level standards. Since states set these values arbitrarily, they can vary across states and even across grades and content areas within states. Additionally, state policies differ on whether students who are absent for tests, for various reasons, are declared missing or are assigned LOSS.

A problem with LOSS is that, as an extreme score, it can arbitrarily affect means. States have currently come to use “percents above cut” instead of means to report data, thus avoiding the need to address the influence of extreme scores. Researcher, though, often want to compare score data rather than counts of students, and the natural choice is the mean. But when tests with different LOSSes are involved, a better choice might be a

median or some other percentile. They can be indexed and have standard errors and may be used in meta-analyses.

### Inferring Causality

We have heard about the need for randomization to conditions in experiments and the point has been (rightly) made that randomization is the process of choice for controlling what Campbell and Stanley have called selection threats to internal validity. Failing randomization, we have heard about matching, covariates, and propensity scores among other ways to introduce controls for selected variables. We have also heard about using change scores where students are their own controls.

The concept of propensity scores as a means of inferring causality from non-experimental studies has been proposed. Some professionals have argued that propensity scores allow a researcher to infer causality, even lacking randomization. I have two areas of concern about inferring causality when propensity scores are used in place of randomization:

- Propensity scores do not necessarily control all relevant confounds. Here is a counter-example: A program works because the principals who chose it were excited about it and conveyed their enthusiasm to their staff. There is no way a researcher will include “principal enthusiasm” as a covariate. This is an implementation issue. Another sort of deficiency is that the researcher may not have available a measurement of what is actually the crucial characteristic that causes participants to become participants in treatments. Also, characteristics of groups of participants may determine treatments. This is especially an issue in using a large-scale database such as this one, in which it

is unclear how much of the assignments to programs is due to student-level and/or school-level and/or district-level characteristics.

- There is a better way to infer causality, especially the development and testing of theory through replication. Actually, even absent theory, a collection of observations – multiple, independent replications, even with pre-post data in individual schools as study sites, with school characteristics as study variables – seems to me to be at least as strong in ability to infer causation as an analysis of one study using propensity scores.

The second point above suggests we may need to go beyond the one-study paradigm and consider where that leads us. If we expect to base important decisions on observations or understandings, it follows that we should develop criteria for the nature and degree of support the understandings need to have.

How are we to use non-experimental data meaningfully? One thought is to use it for hypothesis generation and then to test the hypotheses with independent studies. As we all know, we act every day on assailable understandings about what are really hypotheses. Indeed, all our understandings are in principle assailable. So our focus should not be on establishing when evidence is unassailable, but on establishing when evidence is sufficient.

Should we require that studies supporting our crucial observations or understandings be replicated? Should we require that they have survived a peer evaluation process? Should we require that there have been serious attempts to falsify them? How do we determine that an observations or understanding is crucial in the first place? Some discussion of questions like these would be helpful for our field.

### Meta-Analysis

If we move toward analyses of multiple, independent studies, we may want to look more closely at meta-analysis. In using a school-level database, meta-analyses seem naturally likely to be based on effect sizes defined within schools. Most often, effect size is thought of as between treatment and control conditions. However, facilitating study of effect sizes based on change over time (within or across cohorts; although power should be ample either way, within-cohort change would yield smaller standard errors because the samples would be related) might be a way to think about making the database more useful.

This would allow interventions to be compared across states since each school becomes a study in the meta-analysis and by using meta-analysis, generalizability of findings across states (or school-level characteristics) can be hypothesized and studied. Effect size for a one-year gain is obvious, but as we have heard in the symposium, effect size could be defined in terms of trend components, allowing for meta-analysis of growth patterns.

Effect sizes have been criticized in the symposium because they are indexed to structurally different variabilities. Within a state, schools (and programs within schools) vary in heterogeneity, which affects the denominators of effect sizes. Using effect sizes on state tests in multiple-state studies adds an additional source of non-equivalence because state variabilities differ on NAEP.

However, I don't want to give up on meta-analytic approaches too easily. Perhaps adjusting differences in state variabilities using ratios of state standard deviations on comparable NAEP scales – or school variabilities using ratios of school standard

deviations on comparable state scales – or a combination of the two – could provide a viable approach to the problem. Alternatively, or in addition, state variability on NAEP – or school variability within a state – or a combination of the two – could be used as a conditioning variable in the meta-analytic model.

Varying operational definitions of treatment and control conditions has been discussed. If unevenness of treatment applications or control applications produce different effects, that will show up as heterogeneity of effect sizes in meta-analyses. When heterogeneity is observed, and all explanatory variables have been exhausted, a common follow-up is to look at outliers around the complex model. These may or may not identify unusual treatment or control abnormalities in specific applications. That is another way in which overly homogeneous or heterogeneous programs may be identified as outliers and dealt with.

#### Measuring Results

Finally, we have heard that it is the intent of some to gather future outcome data exclusively in terms of percents in achievement levels. That would be a move away from the sorts of scales that researchers have found most useful in the past. If anyone wants to start a grass roots effort to lobby for a metric designed for developing educational insight rather than a metric developed for policy purposes, I'll join, as would many others at this symposium. We could also try to influence the design of the database so that it facilitates links to state databases with the needed data for our research, as Gage Kingsbury suggested. Both these directions, combined with a way to study school-based program variability (independent variables) in depth, would enhance the value of the database to those who are trying to discover the best ways to educate our nation's youth.

