

Discussion: Cross-State Comparisons (using School-Level Data)¹

Inspired by the presentation

“Adjusting for Differences in Tests” by Robert L. Linn

David Thissen

The University of North Carolina at Chapel Hill

VERSION OF 12/6/2005

The School Level State Assessment Score Database (SSASD, or SLAD) contains state-specific achievement test results for each state, as well as additional (largely demographic) information. Merging data in the SSASD with data from other sources may provide the basis for analyses relating achievement test scores with a number of other variables.² The fact that almost all states are represented in the database invites comparisons across states of the results of within-state data analyses.

A significant obstacle in the path of cross-state comparisons involves the fact that (almost) all states use different state assessment systems with different score scales. Combining results across states that use different assessments requires the creation of comparable scores based on different tests. Robert Linn’s scholarly presentation on “Adjusting for

¹ Prepared for the Symposium on the Use of School-Level Data for Evaluating Federal Education Programs, Board on Testing and Assessment of the National Academies, December 9, 2005.

² In this discussion I consider statistical questions about the relations of achievement test scores with other variables as just that—the measurement of observed relations. I explicitly leave to others discussion of whether those relations should wisely be interpreted as “causal.”

Differences in Tests” reminds us that the idea of comparable scores from different tests has been considered critically by test theorists for (at least) the past forty years—and that practice has often ignored that criticism.

Professor Linn mentions two strategies for combining results across states: using standard scores or effect sizes and using a linkage of state assessments to NAEP. The materials circulated by the committee in advance of the workshop, and the presentations here, inspire comments on both of those possibilities:

Using standard scores or effect sizes for analyses of effects across states

One strategy that has been used in the analysis of SSASD data is to standardize the school means by subtracting their means and dividing by their standard deviation (i.e., McLaughlin, Bandeira de Mello, Cole, Blankenship, Hikawa, Farr, and González, 2002; McLaughlin and Drori, 1999).³ While it is common practice in test theory to set the arbitrary scale of test scores by dividing by some reference standard deviation to produce standardized scores, that is usually done with student-level scores. The metric formed by dividing by the standard deviation of school means is at the very least a source of potential confusion, and may lead to unintended results when it is used to facilitate combining statistics across states.

³ The combined results provided by Moss, Gamse, Jacob, Smith, Greene, & Kupfer (2003) and the Policy and Program Studies Service (2004), as cited by Linn (2005), are also probably based effectively on school-level standard deviations; however that is not entirely clear in those reports.

The standard deviation of school means arises from a combination of several aspects of the schooling and testing situations: One component of that standard deviation is variation in performance among students. However, another contributor to the magnitude of the standard deviation of school means is the size of the schools, and yet another contributor is the sociology of assignment of students to schools. A fourth contributor could be differential effectiveness of the schools; the identification of that contribution appears to be a goal, but I'd like to talk about the second and third parts:

I recently heard a presentation by Howard Wainer (based on results reported by Wainer and Zwerling, 2005), the main point of which was to say that recent interest in creating smaller, rather than larger, schools may be based on faulty evidence. Specifically, the fact that observed variation in school means among smaller schools is simply more variable than it is for larger schools may lead to the observation of a larger number of high-performing smaller schools (ignoring the equally large number of very low-performing smaller schools). But that is not so directly cogent here as the all-important reminder that averages have smaller standard deviations than do the original data, which is usually summarized in the first course in statistics with the formula

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad .$$

For comparisons across states of school means standardized by dividing by their own standard deviations, this means that the same differences (in student-level standard deviation units) will appear smaller in states with smaller schools than in states with larger schools.

While the effect of school size on the variation among school means is clear, the assignment of students to schools may induce some degree of unpredictability in the relation between the standard deviation of school means and the standard deviation of student test scores. In many locales,⁴ school assignments are gerrymandered and bussing is used to match (in the aggregate) an increasing number of demographic characteristics between schools. This may have the effect of making test scores within schools more variable, and school means more homogeneous, than one might expect from something like random assignment. This kind of assignment of students to schools may be very difficult to “correct for” using aggregate background variables describing the background of students in a particular school. And if different states do this kind of assignment to schools more or less, with different degrees of success, the result would be that the variation among school averages may mean one thing in one state and something very different in another state. Under these circumstances, the meaningfulness across states of scores “standardized” by dividing by between-school standard deviations is not clear.

⁴These include Chapel Hill, NC, where I have to read about this in the newspaper annually.

But to return to the (at least) four components of the standard deviation of school means, I would ask a (counterfactual) rhetorical question: What if the standard deviation of school means is (nearly) zero?

Now, I am aware that this does not happen. But it is an interesting *gedankenexperiment* to think that it could happen, and to ask how? If one had a state with very large schools, in which assignment to schools was done to maximize within-school variation in student characteristics (and minimize between-school variation), and there were either no school effects or school effects that counterbalanced the effects of student background variables, the variation among school means could become very small. The important point here is that it could be much smaller than in another state that had the same individual differences variation in achievement test performance among students, but which structured its schools differently. The net result would be that school means standardized by dividing by their standard deviations within the two states would not be particularly comparable.

A solution, if one is desired, may be to “standardize” school means with reference to the mean and standard deviation of student scores within each state. However, that (implicitly) assumes that the variation among students is the same from state to state, which is probably (to some extent) false.

Using a linkage of state assessments to NAEP to facilitate analyses of effects across states

Professor Linn's presentation described the two NRC reports *Uncommon Measures...* (Feuer, Holland, Green, Bertenthal, and Hemphill, 1999) and *Embedding Questions...* (Koretz, Bertenthal, and Green, 1999) that suggested that comparable scores could not, in general, be created by linking state assessment scores with NAEP.

To apply a subset of the two NRC reports' conclusions to the idea of creating comparable scores from scores on different states' assessments using NAEP, consider: The problem is that the idea of linking different states' assessments using NAEP results is that it is effectively an anchor-test linking design, with the tests from states A and B being linked through the administration of NAEP as an anchor test. Among other things, in order to have confidence in the results obtained with such a linking, one has to believe that the anchor test (NAEP) is the "same test" for the group of students who took NAEP and state A's test and the group of students who took NAEP and state B's test. On the surface, it is obviously the same test: It's NAEP. However, a bit below the level of appearance (what questions are asked, etc.) it may not be: One could imagine that differences between states A and B in, say, emphases on effort on high-stakes vs. low-stakes tests, and/or differences in the match with the NAEP blueprint of state A's curriculum vs. state B's curriculum,

might lead NAEP to be *effectively* a different test in state A than it is in state B.

To be more concrete: What one is trying to do in linking state A's assessment to that of state B (through NAEP results, or otherwise) is to obtain a set of linked scores in states A and B that can stand in for what one would have obtained if one had given state A's assessment in both states A and B, or state B's assessment in both states. However, one can easily imagine differences between states in motivational conditions and aspects of curricular match to the tests that would give results like:

States A and B perform roughly equally on NAEP, but if state A's assessment was administered in both states, A would score higher than B, and if state B's assessment was administered in both states, B would score higher than A. That would be an example of the linkage giving a wrong answer.

To see (in data) how such a linkage may not yield the same answer as would be obtained from the actual administration of the same test in multiple states, one would need data collected with the same test administered as the statewide assessment in more than one state. That rarely happens; but there is one publicly-available report that comes close to that situation:

Using data from the 1990 NAEP TSA, and statewide results in four states that used assessments provided by CTB, Ercikan (1997) considered linking functions between the statewide assessments and NAEP. Three of

the states used (different) versions of the Comprehensive Tests of Basic Skills, Fourth Edition (CTBS/4; CTB/McGraw-Hill, 1989), and the fourth used the California Achievement Test, Form E (CAT/E; CTB/McGraw-Hill, 1985). In order to obtain comparable linking functions, Ercikan used a pre-existing equating study among versions of the CAT and CTBS (CTB/Macmillan/McGraw-Hill, 1993) to convert all four states' scores onto the CAT/5 NCE scale. There were approximately 2500 students in the NAEP TSA samples, as compared with the statewide population results (for populations that ranged approximately 8,000 - 50,000).

Display 1 shows the relations among the state means on the CAT/5 NCE scale and on the NAEP scale, using data taken from Table 1 of Ercikan (1997). The means differed little from each other on the statewide assessments (on the CAT/5 NCE scale), but considerably more on the NAEP scale. Further, the order was not the same: State 1 had the highest average on both tests/scales, but State 2 was in the middle of the cluster on the statewide assessment and the lowest of the four states on NAEP. Ercikan (1997) noted that these differences may be due to any or all of several differences between the two sets of tests. "These differences include different testing dates, motivational differences between students taking statewide tests and the NAEP test, and content differences that result in different abilities being assessed by each test" (Ercikan, 1997, p. 156).

An Overview

It appears that the desired data for the desired analyses of educational program effectiveness would be student level achievement scores on comparable measures across states, reflecting progress toward common curricular goals.

What are available are school level means (or other summaries) on non-comparable measures across states with different curricula.

Problems arising from the fact that the data are school-level summaries instead of student level data could be solved by the development of some mechanism to include student-level results in the database. While it is clear that this presents challenges to privacy, those challenges may not be insurmountable.

The lack of comparability of achievement test results across states is another matter. At the very least, it presents challenges that may be very difficult to overcome. Most likely, it presents challenges that cannot be overcome. That may not be so bad, because it may be a problem that does not actually require a solution. Cross-state comparisons appear to add very little information beyond what is provided by description of the correlates of educational achievement within (multiple?) states. And the variation among assessments across states presents no problems for analyses conducted within-states. That is probably why Professor Linn listed first among his suggestions that “analyses can be conducted on a state by state basis.”

Using the ideas of meta-analysis to combine results across states appears odd at first glance for contexts in which the current data analysis has access to all of the data. The primary motivation for the development of methods of research synthesis has been to provide methods to combine results from many studies when the data analyst does *not* have access to the primary data, but instead is forced to work with summary statistics. However, carrying on the theme that within-state analyses using the data in the SSASD may be useful, “vote counting” (see Bushman, 1994) could be used to provide some level of aggregation of results across states without raising some of the questions about metric comparability that are raised by combining “effect sizes.”

A Proposal

Many analyses of the SSASD data rely on a series of statistical “adjustments” followed by relatively straightforward statistical modeling, and raise the question whether some suitable set of “adjusted” data might be computed from the data currently in the SSASD that would permit cross-state analysis to be conducted using relatively simple statistical models and procedures. The extent to which that can be done is not clear, and research reports that follow that strategy challenge their readers to determine the degree to which various “adjustments” solve the problems they are intended to solve, and to further determine if unresolved problems remain.

However, it is conceivable that some level of clarity might be achieved about what can and cannot usefully be done with data from the SSASD, possibly merged with other data sources, by using an extensive statistical modeling approach. To be explicit, imagine considering a relatively complete multi-level model, such as discussed by Thum (2005), that includes student level data. Such models cannot be fitted directly to the data available from the SSASD, because only school level data are available. However, one could work with the full multi-level model, computing school-level expectations from the model itself, to explicitly determine what parameters might be identified and estimable using the aggregate data from the SSASD, and what parameters would have to be dealt with by assumption (of point values or Bayesian prior distributions).

Then one would have a choice: One could conclude that the model-based analysis indicates that such a limited set of estimates could be obtained using only aggregate data that they are basically not worth the trouble, and forego further effort. Or, one could decide that the research questions *could* be satisfactorily answered using the aggregate results, and go ahead. In the latter case, assumptions about the relations between the observed aggregate data and the unobserved student-level data would have been made explicit, leaving few questions on that score.

In this modeling framework, aggregation across states would simply be another level in the model. That would be a particularly challenging level, because it would have to consider explicitly the facts that states use

different assessments to measure progress on different curricula. But challenges are what statistical data analysis is all about.

References

- Bushman, B.J. (1994). Vote-counting procedures in meta-analysis. In H. Cooper & L.V. Hedges, *The handbook of research synthesis* (pp. 193-213). New York, NY: Russell Sage Foundation.
- CTB/Macmillan/McGraw-Hill. (1993). *California Achievement Tests, Fifth Edition, Technical Bulletin 2*. Monterey, CA: Author.
- CTB/McGraw-Hill. (1985). *California Achievement Tests, Form E*. Monterey, CA: Author.
- CTB/McGraw-Hill. (1989). *The Comprehensive Tests of Basic Skills, Fourth Edition*. Monterey, CA: Author.
- Ercikan, K. (1997). Linking statewide tests to the National Assessment of Educational Progress: Accuracy of combining test results across states. *Applied Measurement in Education, 10*, 145-159.
- Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W. & Hemphill, F.C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington DC: National Academy Press.
- Koretz, D.M., Bertenthal, M.W. & Green, B.F. (1999). *Embedding questions: The pursuit of a common measure in uncommon tests*. (Report of the Committee on Embedding Common Test Items in State and District Assessments, National Research Council). Washington DC: National Academy Press.

McLaughlin, D., Bandeira de Mello, V., Cole, S., Blankenship, C., Hikawa, H., Farr, K., & González, R. (2002). *National Longitudinal School-Level State Assessment Database: Analyses of 2000/2001 school year scores*. Washington, DC: American Institutes for Research.

McLaughlin, D. & Drori, G. (1999). *School-level correlates of academic achievement: Student assessment scores in SASS public schools*. Draft, Washington, DC: National Center for Education Statistics.

Moss, M., Gamse, B., Jacob, R., Smith, W. C., Greene, D., & Kupfer, A. (2003). *Reading excellence act and school implementation and impact study: Annual report 2002-2003*. Cambridge, MA: Apt Associates Inc.

Policy and Program Studies Service. (2004). *Implementation and early outcomes of the comprehensive school reform demonstration (CSR) program*. Washington, DC: U.S. Department of Education, Doc # 2004-15

Thum, Y.M. (2005). *Designing gross productivity indicators: A proposal for connecting accountability goals, data, & analysis*.

Wainer, H., & Zverling, H.L. (2005). *Logical and empirical evidence that smaller schools do not improve student achievement*. Unpublished ms.

Display 1. The state means for four states on the CAT/5 NCE scale for statewide assessments administered in 1990, and from the 1990 NAEP TSA results; data from Table 1 of Ercikan (1997).

