



## Better Ways to Compare Schools?

Harvey Goldstein

*Journal of Educational Statistics*, Vol. 16, No. 2. (Summer, 1991), pp. 89-91.

Stable URL:

<http://links.jstor.org/sici?sici=0362-9791%28199122%2916%3A2%3C89%3ABWTCS%3E2.0.CO%3B2-9>

*Journal of Educational Statistics* is currently published by American Educational Research Association.

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/aera.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

---

JSTOR is an independent not-for-profit organization dedicated to and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

## COMMENTARY

---

### Better Ways to Compare Schools?

Harvey Goldstein

*Institute of Education—University of London*

The term *school effectiveness* is found throughout the educational literature, and there is even an annual international conference devoted solely to this topic. The laudable aim of discovering which factors are associated with the success or failure of schools clearly is popular with policy makers and administrators as well as researchers.

Much of the activity in this area is concerned with achievement testing, and some of the most interesting work is based on the use of multilevel models. These models are interesting because they attempt to remain faithful to the hierarchical structure of educational data where, for example, pupils are nested within schools and schools within boards or authorities. Applications of these models have been successful in studying the way in which group differences (e.g., between ethnic minorities) vary from school to school; the way it is possible to measure *value added*, or progress, during schooling by using suitable measures of intake achievement; and the way achievement changes over time using repeated measurements on students. Other work has looked at how schools change over time using repeated measurements of cohorts of students.

The application of these models to multivariate data has shown how multiple matrix designs can be modeled efficiently and how efficient estimates can be obtained for generalizability models. A detailed account of these models with applications can be found in Raudenbush and Bryk (1986) and Goldstein (1987). Some of the current development work is concerned with the formulation of multilevel loglinear models, multilevel time series models, multilevel structural equation models, and random cross classifications, and publications in these areas are beginning to appear.

All of this is exciting because it provides us with a general tool for the valid analysis of educational data and promises to yield new insights. Indeed, it promises this for all observational data in the human sciences and especially in areas such as complex sample surveys, where direct modeling of the data structure is now possible. In education, the intelligent application of these models promises a deeper understanding of the processes of schooling and the determinants of achievement. With suitably large samples of schools, it should be possible to study how both student and school characteristics play

## Commentary

a part and, in particular, how the learning environment contributes to observable educational outcomes.

Nevertheless, one particular use of these models in school effectiveness studies is beginning to cause concern. Consider the simple variance components model:

$$y_{ij} = \alpha + \beta x_{ij} + u_j + e_{ij},$$

where the outcome  $y_{ij}$  is an achievement score and  $x_{ij}$  is a pretest for student  $i$  in school  $j$ . The  $u_j$  and  $e_{ij}$  are random variables with zero means and are referred to as *school* and *student level residuals* respectively.

In school effectiveness studies, following a multilevel analysis, a predicted or posterior estimate is obtained for each school's residual that is an estimate of each  $u_j$ . Depending on the model fitted, this will be interpreted as a school mean in the case of a variance components model or as, say, an ethnic group difference for a school in a model where the coefficient of a dummy variable for an ethnic group is treated as a random coefficient, varying randomly across schools. Having estimated, for example, such a mean for each school, these will be labeled as school *effects* and then treated as if they were legitimate descriptors of performance. In the most extreme case, these effect estimates will be published, and schools will be ranked on that basis.

It must be said that, confronted with a choice between a ranking of schools, based upon such multilevel school effects with adjustment for intake achievement, or a ranking, based not upon a multilevel analysis but on raw school mean outcome scores, most of us would probably settle for the former. Its superiority, however, is severely circumscribed, and it has some severe limitations that ought to make us think very carefully before using it.

First, and most obviously, these effects often will have relatively large standard errors so that it is not possible statistically to separate the estimates. In some cases, perhaps only the extreme estimates can be separated out, with the result that only a very crude ordering is achievable. Anyone who wants to produce school rankings should be concerned with this. Appropriate measures will be required to avoid having users draw unwarranted conclusions.

Second, the term *school effect* itself is somewhat misleading. It is, after all, merely an estimate of a school level residual, conditional on the model fitted. If we happen to have omitted an important variable from the model, its subsequent inclusion will generally change the residual estimates. Likewise, simply carrying out a nonlinear transformation of the response or explanatory variables can alter the residual ranking. Often, where we have choices between assumptions, we cannot rely on the data to guide a choice because it may not be very extensive. Thus, as in all statistical models, the estimates we obtain are sensitive to the assumptions we make, and this will

tend to be more important for residual estimates than for the estimates of the fixed and random parameters in the model. Again, there is a clear obligation on the part of the data analyst to carry out sensitivity analyses to study the effects of changing assumptions.

My own view is that school effect estimates from multilevel models should be treated the same way we treat the results of a screening test. An extreme value may indicate a real effect or problem that a researcher, an administrator, or a school might then wish to follow up. Such a follow-up would seek to explain why, for instance, a particular school seemed to be performing badly or why a school, that seemed to perform well with one kind of student from some curriculum subjects, seemed to perform badly for other kinds of students in other curriculum areas. Used in such a diagnostic way and as a research tool for exploring the nature of effectiveness, such data can be useful. If disseminated publicly as a list of schools that can be ranked from good to bad, such data can be harmful and misleading.

As with many new techniques that promise a substantial advance in understanding, multilevel modeling is not a panacea. Its power is limited, and it is most certainly not a magic wand that will allow us automatically to make definitive pronouncements about differences between individual schools.

### References

- Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*, 1–17.
- Goldstein, H. (1987). *Multilevel models in educational and social research*. New York: Oxford University Press; London: Griffin.

## LINKED CITATIONS

- Page 1 of 1 -



*You have printed the following article:*

### **Better Ways to Compare Schools?**

Harvey Goldstein

*Journal of Educational Statistics*, Vol. 16, No. 2. (Summer, 1991), pp. 89-91.

Stable URL:

<http://links.jstor.org/sici?sici=0362-9791%28199122%2916%3A2%3C89%3ABWTCS%3E2.0.CO%3B2-9>

---

*This article references the following linked citations. If you are trying to access articles from an off-campus location, you may be required to first logon via your library web site to access JSTOR. Please visit your library's website or contact a librarian to learn about options for remote access to JSTOR.*

## **References**

### **A Hierarchical Model for Studying School Effects**

Stephen Raudenbush; Anthony S. Bryk

*Sociology of Education*, Vol. 59, No. 1. (Jan., 1986), pp. 1-17.

Stable URL:

<http://links.jstor.org/sici?sici=0038-0407%28198601%2959%3A1%3C1%3AAHMFSS%3E2.0.CO%3B2-1>