DOCUMENT RESUME

ED 387 514 TM 023 741

AUTHOR Henry, Neil W.
TITLE On Applying Statistical Methods: Losing Local Control.
PUB DATE 18 Apr 95
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995). Comments on an article, "Losing Local Control," by H. J. Walberg and H. J. Walberg, III, in "Educational Researcher," v23 n5 p19-26, 1994.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; *Data Collection; Educational Finance; Elementary Secondary Education; Minority Groups; Models; *Research Methodology; School Districts; *School District Size; *Statistical Analysis
IDENTIFIERS *Missing Data

ABSTRACT

Although the field of statistics is receiving increasing recognition in the media and everyday life, concern continues about the quality of statistical education and statistical analysis and the application of statistical methods. This paper focuses on three areas of concern: (1) the quality of data, especially missing data; (2) the interpretation of statistical models; and (3) the rhetoric of statistical inference. An article by H. J. Walberg and H. J. Walberg, III, "Losing Local Control," is used to illustrate some pitfalls in relating student achievement to school and district size, educational funding, and minority enrollment. Issues of selection and the application of a statistical model are apparent in the study. It is essential that the supporting arguments presented in a study match the data analysis. In the study cited, the multiple regression analysis of state data cannot be regarded as providing an interpretable model. (Contains 10 references.) (SLD)

ON APPLYING STATISTICAL METHODS: LOSING LOCAL CONTROL

Neil W. Henry

Departments of Sociology and Anthropology
and Mathematical Sciences

Virginia Commonwealth University

Richmond, VA 23284-2014
email: nhenry@vcu.edu

Presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, April 18, 1995

ON APPLYING STATISTICAL METHODS: LOSING LOCAL CONTROL

Neil W. Henry
Virginia Commonwealth University

Statistics is a growth industry. Enrollment in introductory statistics courses in colleges and universities is at an all-time high, and is still growing. The media regularly report on their front pages the statistical results of surveys and experiments that would have been buried in the fine print 25 years ago. Statistical software that put complex statistical calculations and high quality graphics at the disposal of just about anyone with a computer is available and getting less expensive by the week.

At the same time that the language of statistics is being folded into everyday discourse, however, there continues to be controversy about the quality of statistical education and statistical analysis, and especially about the way that statistical methods are applied. This is particularly critical when they are used to buttress arguments about social policy. Three areas of concern will be discussed in this paper:

- Quality of data, especially "missing" data;
- Interpretation of statistical models;
- The rhetoric of statistical inference.

I begin by describing critiques provided by Howard Wainer, David Freedman and Donald McCloskey, and examine a recent paper by Walberg and Walberg to illustrate some of the ways in which the methods of statistics may be misapplied.

In a series of papers Wainer has investigated the use of aggregated SAT scores for the evaluation of public programs. Wainer (1992 [1989]) contains a good summary of his arguments. The data analyses he criticizes were carried out with state-level data, and the mean SAT score was a key dependent variable. It is generally accepted that the high school students who take the SAT

are a self-selected sample of all the students, and that the self-selection process itself varies from state to state. The mean score reported thus will be a misleading and flawed measure of the quality of educational outcomes. Quality of outcomes is, of course, what analysts are concerned with when they correlate mean SAT with variables such as average per-pupil expenditures.

Researchers have tried to overcome this inherent bias by various kinds of adjustments, sometimes based on explicit causal models. That is, they can consider what characteristics of students and their local environments lead some to take the SAT and others not to take it. Wainer pointed out that two different teams of researchers who had taken this route produced adjusted mean SAT rankings of the states that correlated .5: "not accurate enough for most practical purposes (i.e., comparing states' performance with an eye toward drawing inferences about the relative efficacy of their different education policies)" (Wainer, 1992: 141).

Highly aggregated data such as statewide mean SAT scores poses serious problems for interpreters of statistical relations. This is true even when the bias of self-selection is not present. A theory that motivates the specification of a particular statistical model at the level of the individual student may not be defensible at an aggregated level, such as the school, school district or state. For instance, other things being equal, spending more money on an individual's education may improve the quality of the student's performance. This individual level proposition's truth or falsity need not be reflected in the size of the correlation between mean per-pupil spending and mean test scores at the state level however. For example, if different states choose to spend dollars on students of different ability, the resulting improvements in quality need not be the same. Conversely, if expenditures had no effect on individuals, but states with different proportions of students of different ability chose to spend different amounts of money per student, an aggregate correlation would be observed even though there were no causal relationships.

In the sociological literature the error of thinking that correlations at the aggregate level imply corresponding relationships at the individual level was identified and dubbed "the ecological fallacy" by W. S. Robinson (1950) half a century ago. In one of his earliest works on causal inference Hubert Blalock warned:

> *In shifting from one unit of analysis to another we are very likely to affect the manner in which outside and possibly disturbing influences are operating on the dependent and independent variables under consideration. (Blalock, 1964: 98)*

David Freedman has been a harsh critic of the use of causal modeling methodology by social scientists. It is important to realize that his position is not merely that complex structural equations models do not adequately represent substantive theory, but that even results obtained by "ordinary" multiple regression analysis should be viewed with skepticism. Freedman (1992 [1987], with commentary) provides an excellent statement of his position. He contrasts descriptive and structural models: the former "passes a curve through a data set [which] may help in understanding the data set, or summarizing it, or explaining it to someone else", while the latter "involves an empirical committment to a theory about how the data were generated." He goes on:

> *In my opinion, the confusion between descriptive and structural models pervades the social-science scholarly literature of the past 20 years, and has distorted the research agenda of a generation. In the end, this confusion might easily destroy the idea of scientific right and wrong. (Freedman, 1992: 123)*

Because of the complexity of many path models, especially those containing latent variables, it is easy to make the mistake of thinking that empirical estimation of their parameters provides a strong test of the underlying substantive theory. This will only be the case when the equations of the model actually represent processes that are substantively defensible. The same can be said about a causal interpretation of a single multiple regression equation. A common textbook interpretation of a regression coefficient is that "an increase in X_1 of one unit, with X_2 held

constant, causes a change in Y equal to the value of the coefficient of X_1 ." (See, for instance, McClendon, 1994: 98.) When based on non-experimental, cross-sectional data, such an interpretation goes substantially beyond the evidence in the data. If there were such an effect, then the regression coefficient would provide an unbiased estimate of its magnitude. On the other hand, there need not exist such an effect (i.e. the stipulated change may not occur) even though the regression coefficient can always be calculated.

McCloskey's 1990 book is a stimulating and entertaining analysis of the rhetoric of economic writing and his insights are applicable to many other disciplines. He states that scientists seem to have two ways of understanding things: "either by way of a metaphor [model] or by way of a story [history]." He finds both in the field of economics, and considers physics to be dominated by models and biology by stories. The insight that links his criticism of economists to Freedman's criticism of sociologists is contained in the following quote:

> *Economists spend a lot of time worrying whether their metaphors [models] meet rigorous standards of logic. They worry less whether their stories [stylized facts] meet rigorous standards of fact. The choice to have high standards of logic, low standards of fact, and no explicit standards of metaphor and story is itself a rhetorical one. (McCloskey, 1990: Ch. 1)*

To relate this to the use of multiple regression and other statistical methods, contrast the high degree of emphasis on efficient estimation of model parameters and demands for statistical significance with the often unfounded but unchallenged interpretation of statistical estimates as parameters of a causal process.

The article by Walberg and Walberg (1994) is an interesting combination of literature review and data analysis. They summarize books and papers that support the idea that small organizations, more specifically small schools, are more effective than larger ones. They point out that over the past 50 years the average size of schools and of school districts has increased sharply, along with per-pupil educational expenditures and the state-funded share of those school expenditures. They assert "that these trends were in exactly the wrong direction" if improving

student achievement were the goal of American schooling. Their empirical contribution is to relate state-level average student achievement to school and district size, state share of educational funding, per-student expenditures on education and percent minority enrollment. Measurements are for the 1989-90 school year.

In their analysis Achievement is defined to be the mean score of students on the National Assessment of Educational Progress 8th grade mathematics proficiency test. In contrast to the SAT scores discussed by Wainer, the NAEP is "conducted with random state public school samples" (p. 22), so self-selection does not seem to pose a problem. The authors conclude that "on average, states with large enrollments and large schools and states that pay more of the costs of education tend to have the lowest achievement." To support this they display bivariate scatterplots of these variables with Achievement, and the associated Pearson correlation coefficients. The discussion of these graphs and statistics is clear and unambiguous, although there are a few minor problems with the presentation. The graph of Achievement with District Size is plotted with a logarithmic scale for District Size and a fitted line, while the discussion indicates that all statistics were calculated with the raw scores. Also, the values of Achievement for some states seem to change slightly from one graph to another. IA, for example, is shown as higher than MT in Figures 1 and 3, but below MT in Figures 2 and 4.

Selection of another kind has occurred, however. Only 37 states and the District of Columbia had NAEP data on mathematic proficiency, and only these 38 cases are discussed. The article does not name the omitted states, though they can be identified indirectly from labels on the scatterplots. There is no mention of where these states stood on the several variables which were available, and conclusions are not qualified in any way with respect to the non-random sample. It is not surprising that some data are "missing." What is surprising is that the authors do not seem to care about the fact, even enough to make an assertion (which a high percentage of readers would agree with) to the effect that the omitted states would not alter the qualitative conclusions.

Wainer (1992) emphatically declares that the reasons why cases are missing may be informative to the main purpose of a study. He advises researchers with missing test scores to use other test scores to clarify their intuition about what is going on, and to perform sensitivity analyses (Wainer, 1992: 204-5). Either approach would have strengthened the trustworthiness of Walberg and Walberg's conclusions.

As already noted, the mean school district size was so skewed that a logarithm plot of this variable with Achievement was displayed. Two cases had mean district sizes over 100,000. Only 3 cases had between 10,000 and 100,000 students per district (the largest of these is about 40,000). At the other extreme, three states averaged close to 300 students per district. Because of this very great spread (mean district size is reported to be 9,141 and the standard deviation 25,862) the size of the correlation coefficient computed on the raw data is greatly exaggerated by the two largest cases. These two outliers, Hawaii and the District of Columbia, appear quite influential even on the logged scale.

There is good reason to drop these two cases from the data analysis, on theoretical not merely statistical grounds. Hawaii and DC are both single school district units. Research that emphasizes the distinction between local control and state control of education finds the two levels confounded. The "local" bureaucracy is not distinguishable from the state's; and the local funding load is not distinguishable unless revenues can be broken down by source (property tax, income and sales tax, for example). Walberg and Walberg defend in a footnote their decision to retain DC in the study (though they drop it when relating achievement to state share of spending because of the unique federal contribution to its budget). The authors are mute in regard to Hawaii, however. In contrast John Meyer, et al. (1994 [1988]: 195) in their analysis of educational bureaucratization and centralization, omitted Hawaii, noting that "Hawaii has only one school district, making it impossible to calculate separate figures for state and local revenues." Thus, in this study, 99.9% of Hawaii's funding is allocated to the state. DC's position on the outcome

variable, a full standard deviation below its nearest neighbor, means that it can be highly influential in statistical summaries of relationships with Achievement.

The Walbergs describe average district size and average school size as policy variables, subject to manipulation by those who wish to produce higher levels of achievement from their average student. In contrast they refer to minority percentage and per-pupil expenditure as "control variables" even though both have increased substantially over the half-century whose trends they discuss. Statewide minority percent varies from .9% to 95.3% in this group of 38, with a mean of 22%. This variable, which the authors insist on referring to as an indicator of SES (see their footnote 4), turns out to have a correlation of -.74 with Achievement. It thus could be said to account for over half of the variation in prediction of mean Achievement. Omitting Hawaii and DC from the analysis would have decreased the strength of this association substantially.

Walberg and Walberg state that these bivariate relationships "merely illustrate the findings for individual states" (p.24). They turn to multiple regression analysis to "provide the significance tests for statistically controlled comparisons" (p.24). Their conclusion seems modest and is stated descriptively rather than causally (Freedman might approve): "Other things being equal, states with larger districts and larger schools and that pay a greater share of public school costs do worse in achievement" (p. 22). A picky reader might note that since two equations were estimated, one without district size and one without school size, they should have substituted "or" for "and" between those two variables. Their reason for not including the two variables together was that the two variables were correlated .36, which was deemed "moderate collinearity." This is curious reasoning, since their two so-called control variables, Minority percentage and State share of spending, have even higher correlations with one of the size variables (Minority&School size r = .42; State share & District size r = .41). Few statisticians would endorse their excuse.

Multiple R squares are .59 and .64 for the two equations. Since the Minority percentage explains about .54 of the variation in Achievement, the Size and State share variables are not adding very much to our ability to predict state Achievement levels. Nevertheless, the coefficients have the predicted sign and are more than twice the size of their nominal standard errors.

Later a slightly revised version of the conclusion is stated: "Thus, the results suggest that, other things being equal, states with larger average size schools or districts achieve significantly less well on average" (p.24). It is time to consider use of that "S" word. Because the phrases "significance tests" and "statistically significant" have been used elsewhere in the article, and P-values have been presented in the tables, it is safe to imagine that the authors mean to invoke the authority of classical hypothesis testing. But what can be the justification for such an interpretation here? Certainly not finite population random sampling: there was no indication that these 38 cases should be considered a random sample of the 50 states plus DC. Besides, in such a case finite population corrections would have had to be applied to all inferential statistics.

Is there a model involved that contains a random variable? There might be if a causal model corresponding to the regression equation(s) was being developed. Here the rhetoric of the 3/4 of the text that does not deal with new data must be examined. In those sections we find over and over references to change over time: trends, shifts, growth. There is a clear implication in this text that if school district consolidation had not occurred the achievement of students today would be higher than it is. Of course there is no evidence, direct or indirect, provided for that "story." It is nevertheless a good story, a well argued and persuasive history. The statistical model applied to these data is largely irrelevant to that story, however. We could see data like this if the story were true or if it were false.

The analyses of educational funding and bureaucratization by Scott and Myers (1994:Ch. 8-9)

and their associates provide an interesting contrast. They too perform multiple regression analyses of state (and school district) level data. In the table that is closest in style to Walberg and Walberg's empirical work they present regressions of size and complexity variables on enrollment and funding variables (Table 9.4, p. 196). They explicitly acknowledge that they are exploring a causal model, but the nature of their variables matches up well with their narrative. They are dealing with a simpler problem, perhaps: money goes into the system and jobs get produced. They aren't concerned with the educational process itself. As a result their models are more plausible than Walberg and Walberg's as a description of the important aspects of the production process.

Furthermore, by examining data that documents relationships at several points in time they gain added insight into the process. An interesting quote illustrates the complexity of data analyses carried out in support of complex arguments: "The effects of state centralization on numbers of schools and school districts, which were supported in the cross-sectional analyses, do not appear longitudinally" (Meyers, et al., 1994: 198). The same result might appear if we were able to track aggregated student achievement over time: we simply do not know.

In closing I'll return to the issue of the level of analysis. School districts vary widely in size within many states. For example, New York State's mean district size is near the median for the country, well below 10,000. The New York City district, however, "with 945,000 students, is larger than the enrollments of 37 states" (Walberg and Walberg, 1994: 20). Obviously they find the size of this district appalling, and believe that it is harmful to the educational process. Yet a major change in its structure (dividing it into 10 autonomous districts) would have little effect on the *state's* average district size and no necessary effect on average school size. These are the "independent" variables in their regression analyses. Once again, the supporting arguments presented in the article do not match the data analysis. They could support an analysis where the unit was the school. They could even support an analysis at the district level, since student

achievement might be tied to greater community involvement in small districts (Meyer et al., 1994: 188). The multiple regression analysis of state data, however, cannot be regarded as providing an interpretable model.

NOTE: There appear to be some inconsistencies in the correlations and regression coefficients reported by Walberg and Walberg. I have been unable to reproduce, even approximately, some of the regression coefficients from the correlations and standard deviations in the article. The discrepancies are strongest with respect to the coefficient for expenditures and for school size.

REFERENCES

[1] Blalock, Hubert M. , 1964. *Causal Inferences in Nonexperimental Research*. Chapel Hill: University of North Carolina Press.

[2] Freedman, David A. , 1992 [1987]. "As Others See Us: A Case Study in Path Analysis," in Shaffer (1992). Originally published in the *Journal of Educational Statistics*, Summer, 1987.

[3] McClendon, McKee J., 1994. *Multiple Regression and Causal Analysis*. Itasca, IL: F.E. Peacock.

[4] McCloskey, Donald N., 1990. *If You're So Smart: The Narrative of Economic Expertise*. Chicago: University of Chicago Press.

[5] Meyer, John W., W. R. Scott, D. Strang, and A. L. Creighton, 1994 [1988]. "Bureaucratization Without Centralization: Changes in the Organizational System of U.S. Public Education, 1940-1980." Chapter 9 in Scott and Meyer (1994).

[6] Robinson, W. S. , 1950. "Ecological Correlations and the Behavior of Individuals," *American Sociological Review*, 15, 351-357.

[7] Scott, W. Richard and John W. Meyer, 1994. *Institutional Environments and Organizations*. Thousand Oaks CA: SAGE Publications.

[8] Shaffer, Juliet (ed.), 1992. *The Role of Models in Nonexperimental Social Science: Two Debates*. Washington: AERA and ASA.

[9] Wainer, Howard, 1992 [1989]. "Eelworms, Bullet Holes, and Geraldine Ferraro: Some problems with Statistical Adjustment and Some Solutions." In Shaffer (1992). Originally published in the *Journal of Statistical Education*, Summer, 1989.

[10] Walberg, Herbert J. and H. J. Walberg III, 1994. "Losing local control", *Educational Researcher*, 23, 5, 19-26