

WATER UNDER THE BRIDGE

A Response to Bingham, Heywood, and White

AURELIO JOSE FIGUEREDO

JOHN HETHERINGTON

LEE SECHREST

University of Arizona

Increasingly, powerful computers and statistical packages now offer researchers a wide range of statistical analyses capable of handling large numbers of cases and variables. Along with this cornucopia of statistical possibilities, however, certain problems emerge. Foremost among these is the possibility that researchers will end up using statistical analyses with which they have little familiarity, such as with respect to underlying assumptions and functional limitations. The use of unfamiliar statistical packages and procedures may, in turn, lead to inappropriate statistical analyses and unwarranted conclusions. The resulting problems can be very serious indeed, especially when conclusions are relevant to public policies.

A recent example of what we think is an all-too-common problem is the article, "Evaluating Schools and Teachers Based on Student Performance: Testing an Alternative Methodology," by Bingham, Heywood, and White (1991), recently published in this journal. Although presented by Bingham et al. as a new method of determining the influence of teachers and schools on student achievement, the statistical methods actually employed by them result in conclusions that are almost certainly wrong and that could result in very bad policy decisions. Their analyses are inherently unable to provide reliable and valid results because of a failure to recognize the statistical limitations of the procedures they followed. Because it may not be immediately obvious to the entire readership of this journal how the particular

AUTHORS' NOTE: *The authors wish to thank Will Shadish for comments on an earlier draft and all the members of the Evaluation Group for Analysis of Data for generally helpful discussions.*

EVALUATION REVIEW, Vol. 17 No. 1, February 1992 40-62

© 1992 Sage Publications, Inc.

statistical errors made by Bingham et al. might ultimately lead to bad policy decisions, we have also taken it upon ourselves to explicate as fully as possible within the constraints of brevity the major policy implications of the possibly divergent results of the alternative statistical methodologies that may be used in this context.

WHAT THEY DID

Bingham et al. selected a school district representative of many of the larger school systems in the United States and tried to identify most of the variables that they believed influence academic achievement among fifth-grade elementary students. Selecting from areas such as individual and family characteristics of the students, classroom and school characteristics, and past academic performance of individual students, Bingham et al. attempted to identify all prior variables believed to influence current academic achievement. To be more exact, they hoped that by beginning with a very lengthy list of diverse variables, they might have captured all the variance attributable to factors existing prior to the entry of the student into his or her indexed fifth-grade classroom. This variance is the "water under the bridge" referred to in the title of this present article. If that were true, then what happened subsequently to the student during the fifth grade would be "value added" (or subtracted) by the particular school and teacher involved. Bingham et al. listed 56 of these predictor variables as a "sample" of the variables used in the analysis.

A multiple regression/correlation analysis was then employed in two broad steps. The first of these two steps was the statistical removal of individual student aptitudes from standardized test scores to obtain a measure of the residual "teacher/school effect." The second step was a causal analysis of these residuals to determine and represent the contribution of the educational system to student achievement. In estimating these teacher/school effects, dummy codings were employed to represent schools and classrooms within schools.

In the first step, three main sets of predictor variables were identified: student characteristics beyond the influence of schooling, natural ability of students, and peer variables. These predictors were entered into two multiple regression analyses¹ using the two available, current standardized test scores (either the Iowa Test of Basic Skills or the Scott, Foresman Reading Series) as dependent variables. From the results of these first two equations, predicted scores were computed for each student, and the predicted score was

then subtracted from the observed score, yielding a residual score that Bingham et al. called the teacher/school effect. Bingham et al. believed that by removing differences in individual student aptitude, as indicated by the wide array of student variables, the residual scores would represent any additional contribution (value added) of the educational system, the so-called teacher/school effect, to student achievement.

Owing to limitations in computational resources, separate stepwise regression analyses were first run on various subsets of the data. For example, each of the three predictor variable sets described earlier was initially pre-screened in separate stepwise regression analyses for each of the two standardized tests, yielding six different predictive equations. The three predictive equations for each of these mutually exclusive and exhaustive variable sets were then combined into composite equations containing only the predictors from each variable set accounting for the most variance in standardized test scores. The composite equation having the highest squared multiple correlation was selected as the final equation by Bingham et al. and reported in detail. This final equation included 45 predictor variables and resulted in a squared multiple correlation of .62.

In the second step, Bingham et al. proceeded to analyze the resulting residuals from the final predictive equation. Because these residuals contained both school and teacher influences, they needed to separate the two. Hence the residuals were averaged across the 103 elementary schools included in the sample. They then employed another stepwise regression procedure, in which reading scores of individual students were used as the dependent variable and dummy variables for elementary schools were used as independent variables. The final predictive equation had a squared multiple correlation of .13 and included 50 of the 103 elementary school dummy variables as statistically significant predictors. That is, 50 of the 103 elementary schools appeared to have statistically significant effects on children's residualized reading scores. These reading scores were taken as proxies for achievement measures.

Then the 50 school-dummy variables were entered into the final equation of the first step, along with the 45 significant predictor variables, and another stepwise regression was run. Nine of the predictor variables, and 44 of the school dummy variables were omitted from the resulting equation. The final residual scores computed from this latest equation now were taken to represent the effect of the teacher, background and school effects having, presumably, been removed. It needs to be noted that the teacher effect actually must be taken as a classroom effect because the data set did not permit distinguishing between individual teachers, such as when teachers

changed jobs during the year. Based on a selection criterion of at least 15 observations per classroom, Bingham et al. were able to analyze the residuals averaged across students for 133 classrooms. By establishing a standard deviation for the average classroom residuals, they were then able to identify classrooms that deviated more than 2 standard deviations from the overall average classroom residual.

In sum, Bingham et al. argued that the classroom effect represents the ability of the instructor to increase the actual student scores above the score predicted by the equation, controlling for effects of the individual schools and factors over which schools or teachers do not supposedly have any direct control, such as individual student characteristics.

THE LOGIC OF THE APPROACH

To understand the rationale behind these analyses, it is necessary to review the basic logic underlying the general "value-added" approach. Because this basic logic was not clearly presented in the original article, the present exposition is based on our best reconstruction of the reasoning that was used. It is based on an explicitly additive model, where

$$\text{Scholastic Achievement} = (\text{Teaching Effects}) + (\text{Nonteaching Effects}).$$

This is no more than a restatement of the fundamental premise of logic and thus hardly controversial. The crux of the matter is to identify what the teaching effects are as opposed to the nonteaching effects in school achievement. Presumably, the most direct way to determine this might be to compare otherwise equivalent groups of children who, respectively, have and have not had the benefit of formal schooling. Our system of universal compulsory education, however, is not conducive to this kind of study.

On the other hand, it is not acceptable to attribute all development that occurs during schooling, whether universally administered or not, to the effects of teaching. Many developmental trends, such as pubertal maturation or getting progressively taller with age, do not seem to benefit from formal instruction. In the absence of a truly uneducated control group, therefore, the only alternative option is to try to relate variance in school achievement to variance in teaching. Any trend found to be independent of this variance in teaching can be confidently characterized a "nonteaching effect."

This is similar to the approach commonly used in behavioral genetics, where a trait that develops reliably, regardless of environmental variation, is deemed to possess developmental homeostasis, or biological preparedness. Previous theories of development sought to distinguish the "innate" from the

“acquired” characteristics of an individual. Because all living things must develop within *some* environment, it was not possible to define “innate” as traits that developed in the *absence* of an environment. Rather, the term could be defined as traits that developed reliably in spite of significant *variation* in the environment.

This study did not directly measure variance in teaching, however, but attempted to isolate the additive teaching effects from the *nonteaching* effects by first identifying and subtracting away the *nonteaching* effects. As previously stated, *nonteaching* effects were identified as factors over which a teacher could have no control. Of these, certain specified categories, such as Individual Student Characteristics and Family Characteristics, are obvious candidates. Others, such as Classroom Characteristics, School Characteristics, and Past Academic Performance, are more problematic until one considers that what was being predicted by this study was specifically *fifth-grade* teaching effects. The basic equation must, therefore, be rewritten as follows:

$$\text{Scholastic Achievement} = (\text{Fifth-Grade Teaching Effects}) \\ + \text{Non}(\text{Fifth-Grade Teaching Effects}).$$

This can be further broken down as

$$\text{Scholastic Achievement} = (\text{Fifth-Grade Teaching Effects}) \\ + (\text{Non-Fifth-Grade Teaching Effects}) \\ + \text{Fifth-Grade Nonteaching Effects} \\ + \text{Non-Fifth-Grade Nonteaching Effects}).$$

The preceding equation allows for prior non-fifth-grade teaching effects as well as true *nonteaching* effects. Furthermore, not all the effects of education are due to the teachers themselves but are perhaps attributable to school environments, school administrators, other students, and local financial resources, such as the local tax base and supplementary state support and regulation. Therefore, teaching effects can be decomposed into teacher and school effects. Also, substituting “Pre-Fifth-Grade” for “Non-Fifth-Grade,” since that is clearly what is meant, gives us

$$\text{Scholastic Achievement} = \text{Fifth-Grade Teacher Effects} \\ + \text{Fifth-Grade School Effects} \\ + \text{Pre-Fifth-Grade Teacher Effects} \\ + \text{Pre-Fifth-Grade School Effects} \\ + \text{Fifth-Grade Nonteacher Effects} \\ + \text{Fifth-Grade Nonschool Effects} \\ + \text{Pre-Fifth-Grade Nonteacher Effects} \\ + \text{Pre-Fifth-Grade Nonschool Effects}.$$

This expansion now allows for four separate teaching effects (i.e., the combinations of fifth and pre-fifth grade crossed with teacher and school effects), only the first of which is of immediate interest. Although this study attempted to distinguish hierarchically between the first two effects (i.e., fifth grade combined teaching effects into fifth-grade teacher effects and fifth-grade school effects), it made little effort to partition the remainder of the confounded effects, except by least squares estimation in a stepwise multiple regression. By this logic, we see that the first of the two broad steps used in this study was apparently intended to statistically remove all but the first two effects, although perhaps not to distinguish very well between them. The second of the two broad steps was clearly intended to hierarchically partition the remainder among those first two effects.

WHY THIS IS WRONG

We want to make clear that a good bit of the work presented by Bingham et al. is to be admired; we do not want in any way to discourage attacks on large problems with large data sets. Their efforts are nearly heroic in terms of mastering problems involved in constructing and managing such large data files, and they showed a good bit of imagination in figuring out how to get around limitations of computer hardware and software. A list of more than 500 variables for more than 10,000 cases is formidable under any circumstances. We appreciate the sensitivity they showed to missing data problems and think that their solution of using mean values as estimates, thus not influencing regression coefficients, and adding dummy variable coding to signify missing data on a per-case and per-variable basis is admirable. We also applaud the reporting of nonstandardized regression weights, which are readily interpretable in their data set and which permit ready detection of anomalies, such as those we point to in this article. We believe that Bingham et al. deserve high marks for their efforts. We do not think that their study deserves to be singled out for criticism; what they did is all too common. We hope that it will become less so.

Unfortunately, the procedure used by Bingham et al. has several flaws that undermine any conclusions based on the reported results. These flaws concern the specification, estimation, and statistical testing of the variables that they claimed influence student performance. We address each of these important tactical components in turn and later discuss possible alternative strategic approaches to the problems raised.

MEASUREMENT

The measured criterion variables were scores on two tests: the Iowa Test of Basic Skills and the Scott, Foresman Reading Series. Because neither of these tests is of inherent interest and because the aim of the elementary education system is obviously broader than improving performance on standardized tests (or “semistandardized” tests, in the case of the Scott, Foresman, as Bingham et al. note), we conclude that the dependent variable of true interest is a *latent* achievement/performance variable. Given that the reliabilities of the measured variables are unlikely to exceed .8, particularly over time, and that the correlation between the two measured reading variables is .64, we may conclude that the latent variable is not measured without considerable error and that the measured variables do not exhaust the meaning of the latent variable. Bingham et al. themselves noted that at least some of the effects of schools and teaching may be lagged over time and not captured in an immediate measure.

An issue of paramount importance is how the regression residuals of these two variables were conceptualized and therefore interpreted. In light of the preceding considerations, the residualized variables should be treated with considerable caution at the empirical level, with more caution at the conceptual level, and with extreme caution in the context of any policy decisions (a point to which we return later). At both the empirical and conceptual levels, though, the regression residuals must be treated as variance otherwise unexplained and determined to some *unknown* degree by characteristics of the teacher.

SPECIFICATION

A critical error in the use of the multiple regression technique by Bingham et al. is the assumption that the residual—the variance not accounted for by the equation—can be interpreted as a measure of the teacher’s influence on student performance. They argued that if they had correctly identified all of the variance not attributable to teaching by means of the predictor variables, what is left over, by default, must be teaching variance. Unfortunately, for their case, that is not true. The residual may contain teaching variance—and indeed it almost certainly does—but it also contains the variance of all unspecified variables *and* error.

One example of a source of error would be the occurrence of such “life events” in these fifth-grade children as parental strife, economic hardship, and personal illness. All such variables might well affect performance in the

fifth grade and not be predicted by prior history. Such effects would be part of the residual attributed in this study to "teaching." Neither the possibility of variance attributable to unspecified variables nor that representing error can simply be assumed away to allow for ease of computation and explanation. This problem is, to some extent, admitted by Bingham et al. Limiting the discussion of their procedures to this one ambiguity, though, obscures the more massive problems left unmentioned. Moreover, it does not help much to admit that the problem exists and then to ignore it in any subsequent discussions. Bingham et al. suggested that over time any errors in value-added estimates will cancel out, but that would be true only for truly random errors, not for systematic biases nor confounded effects.

A more insidious specification problem plagues the methodology used by Bingham et al. when they attempt to specify "factors over which teachers do not have control" (p. 192). The correct identification of these variables is essential to their methodological approach. Although most of the variables identified as not under teacher control are probably not, in fact, under fifth-grade teacher control, as when they are considerably prior in time or outside the school context, that may not be true for all variables in their equation(s) meant to remove nonteaching variance. To the extent that teacher have any influence over selection of students into their classes or even to the extent that selection is biased in any way, what may appear to be "prior" variables may actually represent teacher influence. Some teachers may exert influence, perhaps subtle, in order to get certain students into their classes (as in gender favoritism) or to exclude others (as in racial bias). Even more subtle might be tendencies to assign siblings to the same teacher, which would have the effect of confounding background variables with teacher effects. That a student had an "exceptional education flag" (special education) for fifth grade is regarded as a background variable not under the influence of the fifth-grade teacher, as is number of years a student was in fifth grade. Both of these variables may, however, reflect in some degree the views and behavior of the fifth-grade teacher. Both variables have negative weights, meaning that students characterized by exceptional education flags or by having been retained in fifth grade previously tend to do less well than otherwise predicted.

Another variable at the fifth-grade level potentially susceptible to teacher influence is enrollment in the free lunch program. Oddly, that variable has a positive regression weight, meaning that pupils in the free lunch program tend to do better. It is entirely possible that teachers differ in the extent to which they try to identify children in need of and eligible for free lunch and get them enrolled into that program.

These problems are magnified greatly for analyses in which Bingham et al. tried to identify outstanding schools, that is, schools that produce value added over that presumed to be beyond their control. Many of the variables in the prediction equation(s) are obviously under the control of schools, especially to the extent that pupils remain in the same school over the years prior to the fifth grade. Whether a student is Black or White or male or female is certainly beyond the school's influence. However, the age of a child when he or she enters a grade or number of years that a child is kept in a particular grade is directly influenced by the school staff. Whether a student receives an exceptional education flag is a decision made within the school. Thus the residual school variance can only be interpreted as variability in performance unrelated to what the schools have already done for or to the child.

Furthermore, this problem of specification is not removed by the use of a residualization procedure because residualization only measures the deviation from the mean of teacher effects and school effects. Hence it does not remove any systematic biases inherent in the educational system as a whole. Thus if certain groups are disproportionately compromised by factors inherent in the American educational system, as a whole, this analysis will blame the victims rather than the system.

ESTIMATION

Although widely accepted as a useful statistical tool, the multiple regression/correlation analysis used for both the initial residualization and the final predictive equation is fraught with dangers in estimating effect sizes when one uses a large number of predictor variables in the linear equation.

The first of the two broad steps previously outlined was putatively only used for the initial residualization on nonteaching influences. The results of these analyses were, nevertheless, presented for detailed interpretation, the "general tenor" of which was deemed consistent with both "intuition and many previous results." We later address the appropriateness of the use of these procedures for purposes of residualization of nonteaching influences. For the present, however, we limit ourselves to the examination of the claim that these results are either directly interpretable or intuitively reasonable.

For clarity of exposition, the 56 predictors reported by Bingham et al. are listed in Table 1 by the general categories that were reportedly used. Again, this is our rational reconstruction. Although we believe that we followed their stated organizational principles as faithfully as possible, we do not know for certain to which of these general categories the individual predictors were

TABLE 1: Predictor Variables and Means

<i>Predictor</i>	<i>Mean</i>
Individual student characteristics	
Whether or not student is Black	-5.67
Whether or not student is White	2.18
Whether or not student is male	1.45
Age as of September 1 in year that student attended second grade	3.55
Age as of September 1 in year that student attended fifth grade	-7.40
Number of years that student was in fifth grade	-4.72
First year that student was in second grade	-4.13
Dummy for "first year that student was in second grade"	-4.07
First year that student was in fourth grade	4.77
Student had exceptional flag for third grade	4.16
Student had exceptional flag for fourth grade	5.00
Student had exceptional flag for fifth grade	-4.47
Percentage of years that student has exceptional educational flag out of all years that student has valid "exed" data	-5.16
Grade-K school type (specialty/neighborhood)	-1.78
Family characteristics	
Student received free or reduced school lunches during second grade	2.91
Student received free or reduced school lunches during third grade	1.69
Student received free or reduced school lunches during fourth grade	1.60
Student received free or reduced school lunches during fifth grade	1.80
Mean score on a number of variables indicating subsidized lunch	-9.85
Dummy for "student's parents received food stamps/AFDC when student was in fifth grade"	9.80
Percentage of families that had income below poverty level in 1979 for census tract in which the student lived during fifth grade	-4.16
Dummy for "1979 median income for families with children under 18 in census tract in which student lived during fifth grade"	-6.68
Dummy for "number of parent student lived with as of the last year in fifth grade"	1.23
Percentage of other second-grade students in school who would have only one parent at home by the end of fifth grade	5.41
Dummy for "number of schools student attended in second grade"	-3.05
Dummy for "mean number of schools per grade"	2.56
Classroom characteristics	
Average score of other second-grade students at school on the fifth-grade Iowa Test of Basic Skills (reading-pr)	-0.177
Average score of other fifth-grade students at school on the second-grade Iowa Test of Basic Skills (reading-pr)	0.586
Average score of other fifth-grade students at school on the second-grade Iowa Test of Basic Skills (reading-ge)	-0.231

(continued)

TABLE 1 Continued

<i>Predictor</i>	<i>Mean</i>
Average score of other fifth-grade students at school on the fifth-grade Iowa Test of Basic Skills (math-pr)	-1.60
Average score of other fifth-grade students at school on the fifth-grade Iowa Test of Basic Skills (math-ge)	3.94
Average score of other fifth-grade students at school on the fifth-grade Iowa Test of Basic Skills (reading-ge)	1.24
Dummy for "average score of other fifth-grade students at school on the fifth-grade Iowa Test of Basic Skills (reading-ge)"	-4.75
Average score on Scott, Foresman reading test of other third-grade students at school	-4.24
Average score on Scott, Foresman reading test of other fifth-grade students at school	-4.83
Average number of years in grade for others in second grade	-7.63
Percentage of other fifth-grade students at school who had an exceptional educational flag during the year that student was in the fifth grade	-15.36
Percentage of other fourth-grade students at school who are White	-6.34
Percentage of other second-grade students at school who are Hispanic	-3.77
Percentage of other fifth-grade students at school who are Black	3.37
Percentage of other students receiving free lunch at school during the year that student was in third grade	-7.98
School characteristics	
Average score on fifth-grade school lunch variable of other fifth-grade students at school	10.09
Dummy for "schoolwide pupil stability during the first year that student was in kindergarten"	-1.64
Dummy for "percentage of teachers at school with less than 2 years' experience during year that student was in third grade"	-2.06
Percentage of minority teachers at school during year that student was in second grade	-6.74
Past academic performance	
Second-grade Iowa Test of Basic Skills (reading-ge)	3.99
Second-grade Iowa Test of Basic Skills (reading-pr)	0.282
Second-grade Iowa Test of Basic Skills (math-ge)	2.25
Dummy for "second-grade Iowa Test of Basic Skills (math-pr)"	7.01
Scott, Foresman reading test, second grade	-3.05
Scott, Foresman reading test, third grade	1.10
Scott, Foresman reading test, fourth grade	10.45
Dummy for "Scott, Foresman reading test, fourth grade"	2.66
First-grade attendance	-0.036
Second-grade attendance	-0.038
Number of years that student was in third grade	-2.08

assigned by Bingham et al. Such complete agreement in classification, however, is not essential to the present argument.

First, it is highly unlikely that the large number of predictors were statistically independent of one another. When two or more variables are highly correlated, the statistical estimation method of squared error minimization used in multiple regression is incapable of sorting out the independent effects of each of them on the dependent variable. This condition is referred to as multicollinearity (Pedhazur 1982) and results in highly unstable regression coefficients. As a result of multicollinearity, regression coefficients may actually be higher, lower, or in a completely different direction. Useful diagnostic procedures for detecting multicollinearity do exist (Pedhazur 1982; Cohen and Cohen 1983), but Bingham et al. did not refer to them and probably did not make use of them.

When the variance between two variables is shared or partially confounded, as it is with most variables in social science research, one is not able to obtain pure, independent estimates of the effect of each variable on another one. Multiple regression procedures ascribe to the "first" variable all of the variance that it shares with others. The "second" variable, therefore, gets a much smaller increment of variance than it would have had had it been entered first in the equation. The "second" variable will then have a reduced or sometimes even negative estimated effect size even though it may actually have a positive effect. Especially when correlations with another variable are nearly equal for two predictors, regression coefficients will be unstable across samples or analyses as first one variable then the other happens to have the highest bivariate correlation.

Some examples of what multicollinearity can do to coefficients can be seen by viewing the listing of regression weights reported by Bingham et al. For example, a student who was older than average when attending second grade was expected to be 3.55 points above the fifth-grade grand mean, but a student older than average when attending fifth grade should score 7.40 points below the mean. Another example is that four variables reflecting students' involvement in free lunch programs in Grades 2 through 5 had positive coefficients, but a summary variable reflecting involvement in the free lunch program had a compensatory negative coefficient of about the same magnitude as the sum of the coefficients for the 4 separate years. Another example is that missing data for an item pertaining to the number of schools the pupil attended in second grade had a negative coefficient, but missing data for estimating the mean number of schools per grade had a positive coefficient.

Consider the finding that the percentage of other fourth-grade students who are White has a coefficient of -6.34 and the percentage of other fifth-grade students who are Black has a coefficient of 3.37 . Yet the coefficient for being a White student rather than non-White is 2.18 . Together, those findings would imply that being a White student in a heavily Black school is predictive of good performance. That seems unlikely. Another artifact resides in the finding that a third-grade flag for exceptional education has a coefficient of 4.16 and fourth-grade flag 5.00 , whereas a fifth-grade flag has a coefficient of -4.47 . A regression artifact is one possible explanation because a student would have been flagged in third or fourth grade because of unusually low performance (in second or third grade) but could be expected to have improved (regressed to the mean) in the intervening time. A fifth-grade flag is a sign of an unexpectedly low performance in fourth grade or early fifth grade, without time for regression to be detected having elapsed.

Correlated variables fed indiscriminately into a multiple regression equation result in highly unstable and uninterpretable regression coefficients. Because the least squares estimation procedure is unable to determine the independent effect of each variable, coefficients that should be similar may be reported as either positive or negative. Furthermore, a problem lies in the interpretation of these unstable coefficients in regard to policy decision. For example, examine two of the variables in the past academic performance category: first- and second-grade attendance. Although the coefficients are small ($-.036$ and $-.038$, respectively), both of them were included into the final equation as significant variables. Not many educators would believe that missing school in the first and second grades could be beneficial to one's fifth-grade academic performance. What kind of interpretation could we make from this information? Should we suggest that children miss as many classes as possible to ensure that their fifth-grade academic performance is enhanced?

TESTING

The final category of shortcomings, testing for statistical significance, is as important as the first two. A substantial problem in using a large number of predictor variables in a multiple regression analysis is the possibility that coefficients may be reported as significant as a result of "alpha slippage." Alpha slippage, sometimes known as capitalization on chance, can occur when testing large numbers of variables or subjects. Even at an alpha level of $.01$, 1 in every 100 statistical tests will, on average, be significant by chance. Bingham et al. listed only 56 of over 500 original predictor variables

used in the initial analysis and reported a sample size of 10,654 students. It would be reasonable to assume that some capitalization on chance occurred in this analysis, but its overall effect would be difficult to estimate. We can guess, though, that for an alpha of .05 (presumably; Bingham et al. did not state that value), about 25 of 500 tests should exceed that value if tested against only one criterion. We cannot be certain how many statistical tests were actually done nor how many might have been significant but disregarded for some reason. But if 25 of the 56 obtained significant findings were spurious, it would be interesting to know which ones they are. Unfortunately, it is not possible to determine this from the results.

THE ILLOGIC OF THE APPROACH

For the basic logic used by Bingham et al. to be valid, several fundamental assumptions must be simultaneously true. First, the set of predictors identified as nonteaching effects must be exhaustive of the true nonteaching effects variance. Second, this set of predictors must not contain any true teaching variance (i.e., fifth-grade teacher and school effects). Third, the dummy codes representing the fifth-grade teacher and school effects within that teaching variance must be statistically significant and of appreciable magnitude.

Regarding the first of these assumptions, there is no way of knowing whether this subset of 56 predictors (or even whether the original set of over 500 predictors, for that matter) included all the factors relevant to scholastic achievement. The fact that 56 of them were found statistically significant does not necessarily imply that all causally relevant variables were included. Owing to the great theoretical redundancy of these measures, as further indicated by the fact that these 56 predictors were presumed to account for all the significant variance of the over 500 original predictors, it is quite possible that other relevant causes might have been overlooked in this excess of concentration on a very restricted range of substantive topics.

Regarding the second of the necessary assumptions, many of the listed predictors clearly contained true fifth-grade teaching variance. Some of these, already mentioned, may contain teacher effects, such as "Number of years student was in fifth grade" and "Student had exceptional flag for fifth grade." Others, not specifically mentioned earlier, may contain true fifth-grade school effects, such as the various subscales of the "Average score of other fifth grade students at school on the Iowa Test of Basic Skills." If the educational system is deemed to have no control over these outcomes, then we may already have our most important answer for public policy.

Regarding the third of the necessary assumptions, recall that in their analysis of fifth-grade school effects, Bingham et al. reported that only 50 of 103 elementary school dummy codes were statistically significant, yielding a squared multiple correlation of .13. This means that .13 of the residual variance was accounted for by these schools. Because this residual variance constituted only .38 of the total (1.00 minus a squared multiple correlation of .62), this means that the dummy-coded school effects were responsible for only about .05 of the variance in student achievement. We certainly hope that this was because of the inclusion of many school effects in the prior predictors and that this represents an underestimate of the benefits of elementary education.

No information was given about the statistical significance of the so-called fifth-grade teacher effects, that is, the regression residuals as residualized yet again by the earlier equation. Without the statistical protection of testing and rejecting the "omnibus null" hypotheses on the entire set of coded teacher effects, selecting teachers who are 2 standard deviations above the average is no guarantee that they are not there by chance. Indeed, it is almost inevitably the case that a certain proportion of cases will appear there by chance. If these effects are any smaller than the dummy-coded school effects, as a proportion of the total variance, then we have little confidence that this remaining variance can possess any practical importance. As it stands, if prior predictors accounted for .62 and school effects for an additional .05, we may have already exceeded the reliable variance of the achievement tests used as dependent variables. For the reliable variance of a test to exceed a total of .67, the reliability coefficient of that test must exceed .82, which is not very likely for this kind of standardized achievement test at these grade levels. This means that the remaining residuals, ascribed by Bingham et al. to fifth-grade teacher effects, may represent no more than measurement error. Again, we hope that this seemingly small teacher effect resulted from the inclusion of some teacher effects in the prior predictors and that the true influence of teachers was underestimated in this study.

In summary, it appears that all three of the fundamental assumptions necessary to the basic logic of this approach were systematically violated. First, the prior predictors selected may not have accounted for either all the true nonteaching variance or all the true pre-fifth-grade teaching variance. Second, the prior predictors selected may, nonetheless, have included a substantial amount of true fifth-grade teaching variance, including both fifth-grade "teacher effects" and "school effects." Third, the remaining "school effects," although statistically significant, were small in absolute magnitude, and the doubly residualized "teacher effects" were of undeter-

mined statistical significance and probably even smaller in absolute magnitude. These problems are inherent in the basic design, or flawed logic, of the study, assuming that the operational procedures are correct. Add to this the problems in the practical execution of the study, in both parameter estimation and hypothesis testing, as well as the resulting unreasonableness and uninterpretability of the reported results, and you have conclusions that do not inspire great confidence as a basis for the formulation of educational policy.

WHAT SHOULD BE DONE

In view of the multiple and very serious methodological problems that we have identified regarding the specification, estimation, and testing of these regression models, we believe that the statistical procedures advocated by Bingham et al. are in their present form not usable for their intended purposes. There remain, therefore, two major options for what to do with this kind of information: first, attempt to salvage whatever more limited, but perhaps still useful, functions can be adequately served from the results of the procedures that were actually used; and second, recommend what better procedures can instead be used to address the originally intended functions of these results. Although these two options are not mutually exclusive, each is addressed in turn.

SALVAGING THE PRESENT RESULTS

The basic strategy most suitable for salvaging some useful information from the proposed procedures is to switch to a much more conservative interpretation of these results. First of all, in the empirical specification of the "prior predictors," that is, those factors presumably outside the control of fifth-grade teachers, it is important to clearly distinguish between prediction goals and explanation goals (see Pedhazur 1982; Cohen and Cohen 1983). The approach to the selection of these variables used by Bingham et al. can *only* be justified as a pure "mindless" prediction scheme. For these purposes, it is not necessary to know which final 50 of the initial 500 or so predictors were empirically selected by the stepwise regression procedures, provided that they could be shown to capture all the significant variance of the original set. It is also not necessary to know precisely how that proportion of significant variance in tested fifth-grade student achievement was partitioned among them. It is therefore neither necessary nor, in fact, desirable to either

report or attempt to interpret the potentially misleading partial regression weights assigned to them by simultaneous least squares estimation.

In other words, the initial regression analyses can tell us nothing more than that a certain sampling of prior predictors, within specified substantive domains, was capable of accounting for .62 of the variance in fifth-grade scores on certain specified achievement tests and that this presumably extraneous variance had been removed from any further analysis. On the other hand, this finding may represent useful information for policymakers by providing a rough estimate of how much of the variance in these achievement tests apparently is not normally influenced by fifth-grade educational interventions. Although, because of massive capitalization on chance, this proportion is likely to be a considerable overestimate, readily available statistical corrections, such as adjusted R^2 , might be used to obtain more reasonable estimates.

For future potential users of the proposed procedures, ill advised as they might be, it is only necessary to specify the theoretical bounds of that sampling domain as criteria for inclusion in future sample-specific prediction models, provided that domain is assumed to include all potential outside influences on tested fifth-grade student achievement. The specific identities of the particular predictors that were empirically selected by stepwise regression in this study are highly unlikely to be generalizable across independent samples.

The correct interpretation of the regression residuals from these prediction models is that they are substantially closer to measuring true fifth-grade "teacher effects" than "school effects." Only a few of the predictors listed were identified as potentially under the influence of fifth-grade teachers, and these could readily be eliminated from future prediction models, whereas the vast majority of the predictors listed were at least partially under either the prior or present control of the elementary school attended. This is not to say that the student is not partially responsible for his or her own prior achievement; the elementary school attended must also share at least partial responsibility. From the perspective of the individual fifth-grade teacher, pre-fifth-grade student achievement may, indeed, represent "water under the bridge." But, from the perspective of the institution offering all prior grades of elementary education, there must be an overall accountability for the entire elementary education process and not just some terminal "value added" to a past record of failure.

Furthermore, even for the limited use of these residuals as measures of fifth-grade teacher effects, it must be shown that any residual differences between classrooms are significantly different from each other beyond the

level of both measurement error and random chance. Some quantitative estimate must also be made of both the relative and absolute magnitudes of these differences. Otherwise, we might be inadvertently establishing a randomized "teacher lottery" for dispensing rewards intended to reinforce performance and recognize merit. The random distribution of either undeserved or disproportionate rewards might do more damage to teacher morale than leaving things in their present unfortunate condition. It might add grievous insult to injury to compound generalized maltreatment with repeated instances of obvious injustice.

Of course, even the most conservative interpretations of the results of these modeling procedures are still committed to the basic assumptions of the "value added" approach: the arithmetic additivity of all effects. Multiple regression typically assumes that the predictors have only main and direct effects on the criterion variables. This condition is highly implausible. More often, significant predictors have either mediator or moderator effects, meaning indirect and interactive joint effects, or both. If there are any true moderator or mediator effects that are not explicitly included in the model, then the model is misspecified, and the model predictions, however carefully they are interpreted, are simply wrong.

DEVELOPING MORE LOGICAL ALTERNATIVE APPROACHES

The influence of multicollinearity and alpha slippage, although difficult to estimate, cast considerable doubt on the validity and reliability of the final coefficients that were obtained by the several rounds of stepwise multiple regression performed. In view of the large sample size (over 10,000 cases) used in this study, one may surmise that the principal function served by the stepwise regression procedures applied was not merely one of testing the predictors for statistical significance. Even with over 500 predictors, and especially given the probable alpha slippage associated with screening that many variables, rejecting the null hypothesis is all but guaranteed with such large sample sizes unless all the predictors screened represented unusually poor choices by the researchers.

Thus the primary function of the empirical selection procedures applied may be logically surmised to have been one of data reduction, or of selecting that smaller subset of predictors that could best serve as adequate proxies for the much larger set of relevant variables within the specified domain. With numerical prediction rather than theoretical explanation as the immediate goal, the precise identities and interpretations of these proxy variables were

thus not as important as their sheer predictive capacity as proxies for the broader domain of relevant variables that they were empirically selected to represent.

We propose that this data reduction function could be better served not by mindlessly constructing a ridiculously large multiple regression equation using exploratory techniques of model specification but by employing a common-factor modeling technique. These multivariate methods capitalize on the problems of multiple regression by using the common variance, or multicollinearity, between predictor variables to construct hypothetical common factors, or "latent variables," of which the measured variables are only manifest indicators. In fact, Bingham et al. had begun the theoretical background work in this area by dividing the predictor variables into a priori categories, such as "Individual Student Characteristics" and "Classroom Characteristics." Hence a large number of variables could be concisely represented by a few discriminable common factors, or latent variables, thus reducing the problem of alpha slippage and avoiding the multicollinearity problems of multiple regression while providing a parsimonious explanation of the phenomena. Thus an interpretable explanation need not be sacrificed to mere prediction as a goal of this analysis, and conceptual virtue can be made out of computational necessity.

By the use of latent-variable causal modeling, factor-analytic structural equation models could be constructed to predict the two achievement test outcomes. Indeed, the two achievement tests that were used as separate criterion variables could themselves be used as indicators of a common hypothetical construct representing student achievement. Another benefit of common-factor modeling is that certain indicators that are clearly codetermined by more than one casual influence can be modeled as "factorially complex," or indicators of more than one hypothetical construct. For example, pre-fifth-grade student performance measures could be represented as a function of both student *and* school characteristics rather than simplistically ascribed solely to one or the other of these factors. This would help address the thorny problem of distinguishing student from school effects in prior scholastic performance.

The problem of estimating teacher effects also has a better and more straightforward solution. If Bingham et al. were truly interested in the effects of teachers on student performance, they would have included information regarding teachers in the list of predictor variables. Variables such as length of tenure and level of education would be pertinent information to specify for a "teacher characteristics" common factor, or latent variable. Residuals are truly a foul brew of variance; it is much better (and safer) to estimate

directly the effects that one is interested in estimating rather than to rely indirectly on the variance that was not accounted for by the rest of the model as an indicator of the effects of interest.

CONCLUSION

Prior to any statistical analysis, careful consideration of the assumptions and limitations of any particular statistical technique should occur. It is too easy to allow the computer to do all the thinking and then simply report the output. Before pursuing any statistical methodology, the implications of its use to the methodological design and the integrity of the conclusions should be considered. Alternative statistical techniques that are applicable should be compared for the advantages and disadvantages of each. Finally, one needs to be realistic in reporting the limitations of the conclusions.

THE IMPOTENCE OF UNAIDED EMPIRICISM

In undertaking their monumental analysis without guidance from any theory at all, or at least no explicated theory, Bingham et al. doomed themselves to the tyranny of numbers. They had no basis for deciding that any one result was better than or preferable to any other. Consequently, they ended up with many empirical findings in the form of regression weights that simply do not make sense and that are, in some instances, likely artifacts. Many of the findings would not be likely to replicate at all in other samples, other grades, or for other outcome measures, let alone replicate in other school systems. The many variables tested guaranteed that a substantial number of the "significant" findings are probably chance occurrences that would not be replicable at all.

Application of even rudimentary theory, scarcely more than common sense, would have kept Bingham et al. from taking seriously and publishing the long list of predictors for which they claimed statistical significance. "Theory is method" (Lipsey 1990), to some extent, and judicious application of theory would have helped to rule out some findings, most notably those stemming from multicollinearity problems. A more conceptual approach would likely have avoided most artifactual and multicollinearity problems; at least they would have been more obvious. By prior consideration of the latent constructs underlying the observed, measured variables, initial clustering of variables would have been achieved. These clusters could then have

been refined by statistical analysis (such as factor analysis), and the matrix of variables would then have been condensed or reduced both substantially and productively.

For example, it is likely that one resulting factor would have been the family economic circumstances for each pupil. A second factor might then have been the residual variance in neighborhood economic circumstances not related to the individual family variable. In all probability, many more of the initial variables would have been preserved in the measurement models as contributing to the accuracy of the measurement of latent traits even though those variables individually were not significantly associated with the test score criteria. What would surely have emerged would have been a reduced model with relatively few terms forming a more comprehensible "theory" of value-added elementary education.

It is also possible that by using individual regression techniques, such as growth curve analysis, one might have been able to develop a useful measure of change in family or student characteristics over time, for example, so that a downward trajectory for one student could have been distinguished from the flat or upward trajectory of another student in the same immediate circumstances. We believe that such analyses would have made better use of the longitudinal data that were apparently available rather than relying on fifth-grade performance alone as the critical outcome. However, because this alternative approach is qualitatively different from what Bingham et al. had in mind, we will not discuss it further here.

POLICY IMPLICATIONS

With respect to policy, there may be less in the approach taken and the results achieved by Bingham et al. than meets the eye. They apparently believe that their method may be used to identify "best and worst" teachers in schools. Presumably, the method might also be used to develop plans for differential intensity of supervision, additional training, or even compensation, such as merit pay. For such policy decisions, the true effect size becomes a greatly important issue. For example, if .20 of the variance in student performance was residual and therefore potentially attributable to quality of teaching, that seems substantial. But note that there is no way of knowing how much of that residual was pure measurement error.

Whether a school system has a policy of allocating .20 of the variance in teacher pay to merit or only, say, .05, would likely loom large in teachers' eyes. Presumably, one would want the *variance* in teacher pay to be related to the *variance* in teacher performance. If the variance in actual performance

was for instance, .10, and the standard deviation in teachers' pay was \$3,000, allocating .10 of the variance in pay to merit would result in a reduction in the standard deviation to only about \$2,850, that is, the standard deviation based on other variables such as seniority and formal training. It would be hard to make a case for allocating more variation to merit pay than is found for meritorious performance. Similarly, if the variance actually attributable to teacher adequacy is small, that is, most of the residual is actually error variance, the cost-effectiveness of such remedies as closer supervision or additional training would be vitiated.

Bingham et al. admitted that their residual measure of teaching may include other variance, including random error. They believed therefore that over time, that error will be negligible because from one year to the next it will tend to cancel out. That is true, but it also has two important policy implications. First, the fact that errors cancel out only over time means that any attempt to reward teachers (or schools) for value-added increments to student performance should also be fairly long. Rewards would not be very immediate and might lose whatever force they have. A second policy implication is that rewards probably should not be incremental in nature: teachers should receive performance bonuses rather than increments to their base salaries. An "error" resulting in the increase of a teacher's base salary by, say, \$1,000 does not cancel out unless that increment could in some manner be taken away fairly quickly, not a likely scenario in most school systems.

We understand, and sympathize with, the plight of Bingham et al. in not having measures at the individual teacher level. That is most unfortunate, but it also tends to make any use of the findings in policy a rather blunt instrument for improving teacher performance and education. A fairly simple task would be to follow up on these findings to determine the characteristics of teachers (and schools) who are high or low value-added types. But until those characteristics are known, the idea of rewarding teachers for performance that they may not be able to understand themselves and that others may not be able to model successfully is not going to be a very satisfactory policy situation.

More insidious political implications may also derive from this particular construction of the problem. Treating prior student scholastic achievement as so much "water under the bridge" and rewarding teachers for essentially "making the best of a bad job" thereafter encourage a fatalistic attitude toward elementary education. Focusing on the apparently trivial amounts of "value added" by individuals to what is becoming a collective national disaster distracts the attention of policymakers from the massive and radical system-wide reform that is clearly needed. Using the proposed statistical procedures,

students who possess characteristics somehow predictive of poor performance in the current educational system are "written off" by multiple regression, so that no one is ever accountable for their expected failure. This practice could permit a wholesale evasion of responsibility on the part of the educational establishment that should neither be countenanced nor camouflaged in statistical technicalities.

In short, we believe that the policy implications of the findings of Bingham et al. are questionable at best and maybe dismissable altogether. Our belief is that work of the kind they have undertaken is potentially useful in understanding the educational process more generally, in identifying points at which the system is weak, and, eventually, in improving it at the system level. We are dubious in the extreme about its value in identifying either teachers or schools as outstanding.

NOTE

1. We are simplifying somewhat the description of what was actually done in order to facilitate our own discussion.

REFERENCES

- Bingham, R. D., J. S. Heywood, and S. B. White. 1991. Evaluating schools and teachers based on student performance: Testing an alternative methodology. *Evaluation Review* 15(2): 191-218.
- Cohen, J., and P. Cohen. 1983. *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Lipsey, M. W. 1990. Theory as method: Small theories of treatment. In *Research methodology: Strengthening causal interpretations on experimental data*, edited by L. Sechrest, E. Perrin, and J. Bunker, 33-51. Washington, DC: National Center for Health Services Research.
- Pedhazur, E. J. 1982. *Multiple regression in behavioral research: Explanation and prediction*. New York: Holt, Rinehart & Winston.

All three authors are members of the Evaluation Group for Analysis of Data at the University of Arizona, where Aurelio Jose Figueredo is Assistant Professor of Psychology, John Hetherington is a doctoral student in psychology, and Lee Sechrest is Professor of Psychology.