Journal of Educational and Behavioral Statistics
Winter 1995, Vol. 20, No. 4, pp. 307–335

The Estimation of School Effects

Stephen W. Raudenbush
Michigan State University
J. Douglas Willms
University of New Brunswick

Key words: *multilevel data, hierarchical linear models, school effects, school evaluation, causal inference*

The increasing public demand to hold schools accountable for their effects on student outcomes lends urgency to the task of clarifying statistical issues pertaining to studies of school effects. This article considers the specification and estimation of school effects, the variability of effects across schools, and the proportion of variation in student outcomes attributable to differences in school context and practice. We present a statistical model that defines two different types of school effect: one appropriate for parents choosing schools for their children, the secondfor agencies evaluating schoolpractice. Studies of both types of effect are viewed as quasi-experiments posing formidable obstacles to valid causal inference. A multilevel decomposition of variance within and between schools has important and perhaps counterintuitive implications for school evaluation. The potential for unbiased estimation depends on the type of effect under consideration because the two types of school effect have markedly different data requirements. Commonly used estimators of each effect are shown to be biased and, in some cases, inconsistent. Analyses of survey data from Scotland illustrate the recommended techniques. We conclude with a brief discussion of the role of school evaluation in a broader agenda of research in support of school improvement.

Investigators have conducted numerous studies aimed at identifying effective schools, determining which practices are related to their effectiveness, and assessing the magnitude and stability of school contributions to student outcomes (see reviews by Good & Brophy. 1986; Gray, 1989; Heyns, 1986; Murnane, 1975; and Rowan, Bossert. & Dwyer, 1983). Although the studies are usually described as studies of school effects, different studies actually embody two quite different conceptions of a "school effect." First, the term

This research was supported by the Centre for Educational Sociology, University of Edinburgh, through grants from the Scottish Education Department and the UK Economic and Social Research Council; by the Canadian Social Sciences and Humanities Research Council; and by the Spencer Foundation.

may refer to the effect on a student outcome of a particular policy or practice, such as the effect of reducing a student-teacher ratio or the effect of adopting a school-wide peer tutoring program. Second, a "school effect" may be the extent to which attending a particular school modifies a student's outcome. The first conception is embedded within the second in that particular policies or practices contribute to the summary effect of the school on each student.

In this article, we examine the second conception of a school effect. This conception underlies current policy initiatives in a number of countries that aim to hold individual schools accountable for their contributions to student learning. Such policy initiatives take a wide variety of forms. In Britain, publication of examination results for each school is intended to help parents select schools for their children as part of a larger initiative to introduce market mechanisms into the education sector (Echols, McPherson, & Willms, 1990). In the United States, some state governments have provided cash bonuses to schools producing unusually favorable test results (Mandeville & Anderson, 1987; Stephenson, 1986). In Thailand, during the early 1980s, the Ministry of Education ranked the country's 72 provinces in terms of mean test scores. It conferred prestige on officials in high-ranking provinces and exhorted officials in low-ranking provinces to improve (Wheeler, Raudenbush, & Pasigna, 1989).

Each of these reforms raises questions about fair and scientifically valid approaches to conceiving and estimating school effects. In every case, a chorus of criticism has confronted the reformers. In Britain, critics argued that school mean examination results, unadjusted for school differences in student background, seriously distorted the public's perceptions of the relative effectiveness of schools competing for students. Similar criticisms in Thailand encouraged the ministry to modify its accountability system to rely on provincial *gain* scores in test results, rather than simple cross-sectional provincial averages. Evaluators in the United States have advocated the use of school-level residuals-discrepancies between school mean outcomes and school means predicted on the basis of student background-to represent school effects (Stephenson, 1986). Concerns about the validity of the methods used to compare schools have been embedded in more general criticisms about the costs of monitoring and its effect on the morale and autonomy of teachers (Willms, 1992). Key issues in the debate have also included the choice of outcome variables and procedures of measurement, including, for example, the use of standardized tests versus more "authentic" forms of assessment such as essays, longer-term projects, and demonstrations (Wolf, Bixby, Glenn, & Gardner, 199 1). Because our purpose is to consider issues of statistical estimation, we shall avoid measurement issues and assume that a reasonably reliable and valid measure of a socially desired outcome is available.

Researchers have debated for more than two decades the appropriate method for adjusting for student background in estimating school effects (e.g., Dyer, Linn, & Patton, 1969; Marco, 1974). A common approach has

been to regress school mean outcome scores on the school means of one or more background variables. The school effect, or what is sometimes called the "value added" (McPherson, 1992), is each school's residual from the regression. Others have suggested the aggregation of residuals from student-level regression models. Aitkin and Longford (1986) considered alternative approaches to the estimation of school effects and advocated the method of maximum likelihood based on a multilevel, hierarchical model. Other researchers have demonstrated how similar multilevel models could be used successfully to estimate the effects of school policies and practices (Goldstein, 1987; Raudenbush & Bryk, 1986). Our purpose here is not to set out yet another method and build a case for it. Rather, it is to bring some order to the debate by specifying a general model for school effects and clarifying the meaning associated with different types of estimates.

The remainder of this article is organized in six sections. The first defines two types of school effect that are conceptually distinct and appeal to different audiences. We consider the problem of estimation from the standpoint of recent thinking on causal inference in nonexperimental studies. Subsequent sections present the general model for estimating school effects, describe techniques for estimation, apply these techniques to data from 20 Scottish secondary schools, and evaluate commonly used estimation techniques. The conclusion considers the role school evaluation can play in a broader research agenda supporting school improvement.

Types of School Effects

In our view, the problem of estimation cannot be solved without first clarifying the difference between two types of school effect implicit in the design of school accountability reforms. Both types of effect involve the difference between the performance of a child in a particular school and the performance that might have been expected if that child had been in some other setting. The two effects differ in the alternative setting used as a standard for evaluating the child's performance. The choice of ***comparison setting*** is critically different for different uses of the school effect information. Following Willms & Raudenbush (1989), we refer to these as ***Type A*** and ***Type B*** effects.

Type A Effects

The Type A effect is the difference between a child's actual performance and the performance that would have been expected if that child had attended a "typical school." The notion of a typical school can be clarified by imagining an experiment in which a block of J students of identical background and aptitude are assigned at random to the J schools under evaluation. The Type A effect is then the discrepancy between a given student's performance and the average performance of students in the block.

The Type A effect is the effect parents generally consider when choosing

one of the *J* schools for their child. Parents might send their child to the school producing the hugest Type A effect, regardless of whether that school's effectiveness derives from the superb practice of its staff, from its favorable student composition, or from the beneficial influence of the social and economic context of the community in which the school is located. But it would clearly be unfair to reward school staff purely on the basis of their Type A effects, given that the staff is only partly responsible for those effects.

Type B Effects

The Type B effect is designed to isolate the effect of a school's practice. We conceive of ***school practice*** broadly to include administrative leadership, curricular content, utilization of resources, and classroom instruction, but we distinguish school practice from ***school context,*** which includes school-level factors that are exogenous to the practices of the school's administrators and teachers. Such contextual factors include the social and economic characteristics of the community in which the school is located and the demographic composition of the student body. These factors may promote a normative environment among parents and peers that promotes or undermines academic learning, quite independently of staff efforts or skill. Although school context may facilitate or inhibit effective school practice, the two are conceptually distinct,' and schools with similar contexts may vary in practice.

The Type B effect, then, is the difference between a child's performance in a particular school and the performance that would have been expected if that child had attended a school with identical context but with practice of "average" effectiveness. Our notion of average effectiveness can be made precise by imagining an experiment in which *J* schools having identical contexts are first assigned to treatment levels that vary in terms of practice. Next, a block of *J* students of identical background and aptitude are assigned at random to these schools. The Type B effect is the discrepancy between a given student's performance and the average performance of students in the same block.

The Type B effect is the effect school officials consider when evaluating the performance of those who work in schools. A school with an unfavorable context could produce a large Type B effect through the effort and talent of its staff. The school would rightly earn the respect of school evaluators even though parents shopping for a large Type A effect might not want to choose that school.

If the hypothetical experiments described above could be implemented, Type A and Type B effects could readily be estimated without bias. In reality, however, studies of school effects are quasi-experiments, and estimation requires some attempt to identify and control for exogenous covariates that are confounded with the "treatment" provided by the school.

The Logic of Causal Inference in School Evaluations

Current statistical theory regarding causal inference in nonexperimental studies (e.g., Holland, 1986; Rosenbaum & Rubin, 1983; Rubin, 1978) identifies as essential elements in a causal study two *sets* (a set of treatments to be evaluated and a population of units to be assigned to treatments) and two *random variables* (the first indicates treatment group assignment; the second is the outcome for a unit under each treatment). Application of this framework to school effects clarifies the substantial inferential challenges associated with current school evaluations.

Treatments. Holland (1986) draws a sharp distinction between *treatments* and *attributes.* A treatment must be capable of manipulation and its effect can be conceived only in relation to alternative treatments to which the unit could plausibly have been assigned. Thus, the social background of a student is an attribute[2] whereas a method of instruction is a treatment. Similarly, the socioeconomic context of the school may be regarded as an attribute whereas the pupil-teacher ratio, grade structure, and curriculum are treatments. Thus, our Type B effect results from a complex mix of treatments (school policies and practices) found in a given school.

The status of the Type A effect as a treatment effect is less obvious. From the point of view of the school administrator, the Type A effect appears influenced by a mixture of attributes (e.g., the social context of the school) and treatments. From the standpoint of parents choosing a school, however, the sum total of educative influences in a school, including the attributes of peers attending that school, may be regarded as a treatment. Thus, parents can manipulate peer attributes by transferring their child to a different school. The Type A effect thus results from an even more complex mix of treatments than does the Type B effect because it includes peer and community norms in addition to school policies and practices.

Both Type A and Type B effects result from mixes of treatments that can change over time without the investigator's knowledge. Although Willms and Raudenbush (1989) gave evidence of stability of school effects in their data, reforms such as school choice plans that promote student mobility may create a moving target for statistical inference.

Units. In studies of Type A effects, the units are students assigned to treatments, which are schools. In studies of Type B effects, life is more complicated. First, schools of varying context are assigned to the mix of practices that produce the Type B effect. Next, students are assigned to schools.

Treatment assignment, possible outcomes, and strong ignorability: Type A. The Holland-Rubin theory of causation' postulates for each student not only the outcome observed after attending a given school, but also a set of unobservable outcomes, one for each of the schools not attended by that student-that is, the set of outcomes that would have been observed had that

student attended each other school. In our case the model postulates J outcomes for each student, one for each school under study. The Type A effect for that student is then the discrepancy between that student's observed response and the mean of all J possible responses of that student. Holland defines as the "Fundamental Problem of Causal Inference" the reality that only one outcome per student will be observed; thus the Type A effect for each individual is not estimable. However, a randomized study would facilitate unbiased inference about the average effect of a treatment for a subclass of students or all students. Randomization achieves this end by insuring that treatment assignment is independent of the set of J outcomes for each student.

Given the impossibility of randomization in standard school evaluations, is unbiased inference possible? Rosenbaum and Rubin (1983) show that unbiased inference is possible in nonrandomized studies when treatment assignment is "strongly ignorable," that is, when the J outcomes for a student are conditionally independent of treatment assignment given a set of covariates, x. A special case arises when treatment assignment is strongly ignorable and the probability of assignment to a given treatment is a linear function of x; in this case a covariance adjustment can produce an unbiased estimator of the treatment effect (Rosenbaum & Rubin, 1983, Corollary 4.3). We shall use this principle in estimating Type A effects.

Treatment assignment, possible outcomes, and strong ignorability: Type B. As in the case of Type A effects, strongly ignorable assignment of students to schools is required for valid inference in the case of Type B effects. In addition, however, one must consider the assignment of schools of varied context to the Type B treatment, that is, the mix of school practices associated with a given school. This requirement poses a threat to valid inference not present in the case of the Type A effect. The next section considers this problem in detail.

Data Requirements for Estimating Type A and Type B Effects

Clearly, students are not randomly assigned to schools; rather, school membership is determined by nonrandom processes including socioeconomic and racial residential segregation (Massey & Denton, 1988). These processes ensure that schools will vary on the intake characteristics of their students. To achieve strongly ignorable treatment assignment in the case of Type A effects, the researcher must consider covariates related to the outcome that affect the propensity of a student to attend a given school. Researchers who make this effort, particularly those who obtain valid and highly reliable premeasures of achievement or aptitude in addition to key indicators of social background, can compute defensible estimates of Type A effects.[4] Unfortunately, as we shall discuss, even when the relevant data have been collected, appropriate statistical analyses have been rare.

Data requirements for unbiased estimation of Type B effects are far more daunting. The Type B effect is the effect of school practice as distinct from

school context. The problem is that differences in school context must be assumed related to school practice. To achieve strongly ignorable treatment assignment requires not only that student assignment to schools be strongly ignorable as in the case of the Type A effect, but also that a school's assignment to Type B treatments be strongly ignorable. The problem is that most studies of school effects do not measure the practices that produce the Type B effect. It is far more difficult to adequately measure school practice than to obtain good measures of family background and prior student aptitude or achievement. Because the Type B treatment assignment of each school is unknown, the selection of schools into such treatments cannot be studied, and the investigator is left to rely on the implausible assumption that schools of varying context are assigned at random to elements of school practice.

A conclusion of our investigation, then, will be that the technical requirements for producing credible Type A effect estimates, though substantial, are much more modest than are the technical requirements for producing credible Type B effect estimates. Not surprisingly, consistent estimation of the variance attributable to Type A effects is more readily within reach than is consistent estimation of the variance of Type B effects. However, we shall demonstrate that it is often possible to place brackets (upper and lower bounds) on the variance of the Type B effects, and the distance between such brackets can be viewed as one indicator of the tenability of the Type B effect estimates themselves.

Before considering procedures to estimate Type A and Type B effects and their variances, it is necessary to postulate a model that relates these effects and the influence of student background attributes, including prior academic achievement and ability. The model enables us to set out the conditions for strongly ignorable treatment assignment and consistent estimation and to delineate the sources of variation in outcomes that lie between and within schools.

A Statistical Model for School Effects

We envisage the outcome variable (Y) as arising from the influence of school practice **(*P*),** school context (C), student background (S), and random error (e) according to the additive model

$$Y_{ij} = \mu + P_{ij} + C_{ij} + S_{ij} + e_{ij}, \qquad (1)$$

where Y_{ij} = the outcome for student i in school j; μ = the grand mean of Y; P_{ij} = the effect of school practice (including, for example, school resources, organizational structure, and instructional leadership) on student i in school j; C_{ij} = the contribution of school context (including, for example, the mean socioeconomic level of the school's students and the unemployment rate of the community [Raffe & Willms, 1989]); S_{ij} = the influence of measured student background variables (including, for example, pre-entry aptitude or socioeconomic status); and e_{ij} = a random error term, including unmeasurable

sources of a particular student's outcome, assumed statistically independent of P, C, and S with zero mean and homogeneous variance $\sigma^2(e)$.

The model could be readily expanded to include an interaction effect between school practice and school context. Though such effects might be of interest, their inclusion would shed no additional light on the basic principles we seek to educe; thus, for simplicity's sake, we shall not include these interactions.

An important feature of this model is that the influence of school practice (P) and context (C) is allowed to vary across students within a school. This feature of the model allows for the possibility that a school characteristic will modify the distribution of outcomes within a school; that is, there is no assumption that a school has a uniform effect on all who attend it. Technically, this means that the model can include both main effects of school-level variables and interactions between school- and student-level variables. Hence we may write

$$P_{ij} = P_j + (PS)_{ij} \text{ and } C_{ij} = C_j + (CS)_{ij},$$

where P_j and C_j represent the main effects of school characteristics, and $(PS)_{ij}$ and $(CS)_{ij}$ represent interaction effects with student background. We shall require that within schools $(PS)_{ij}$ and $(CS)_{ij}$ have zero means and variances $\sigma^2(PS)$ and $\sigma^2(CS)$, respectively. Between schools, P_j and C_j, in turn, have zero means and variances $\tau(P)$ and $\tau(C)$, respectively.

The contribution of student background can be partitioned into within- and between-school components:

$$S_{ij} = S_j + (S_{ij} - S_j),$$

where S_j is the mean student background contribution in school j. These two components are statistically independent. We shah require that S_j and $S_{ij} - S_j$ have zero means and variances $\tau(s)$ and $\sigma^2(s)$, respectively.

Models for the Two Types of School Effects

We define two types of school effects described earlier. The first is the Type A effect, denoted

$$A_{ij} = P_{ij} + C_{ij}, \tag{2}$$

which includes effects emanating from school context (e.g., community norms and peer influence) as well as practice. As mentioned, this effect is of special importance to parents who wish to choose the optimal school for their child. By comparing schools' Type A effects, a parent could compare the expected "value added" to a students' outcome without knowing the relative contributions of context and practice.

The second effect is the Type B effect, denoted

$$B_{ij} = P_{ij}, \tag{3}$$

which includes only the effect of practice. This effect is of special importance to administrators who wish to hold school-site personnel accountable for schooling outcomes, or to policymakers who wish to discover how schools can be modified to improve outcomes.

The Type A and Type B effects can also be written in terms of main effects and interactions. For instance,

$$A_{ij} = A_j + (AS)_{ij},$$

where

$$A_j = P_j + C_j,$$

and

$$(AS)_{ij} = (PS)_{ij} + (CS)_{ij}.$$

Sources of Variation in School Effects

Estimation of the variation in Type A and Type B effects is important for several reasons. The magnitude of variation in Type A effects is an indicator to parents of the stakes in choosing among schools. Thus, if this variance were zero, choosing among a set of schools would have no consequences for the expected outcome of a child; whereas, if the variance of the Type A effects were large, such choices would be important. Similarly, the magnitude of variation in the Type B effects is a measure against which one can assess the importance of school differences in practice as determinants of student outcomes. Administrators and policymakers would be interested in estimates of the variation of both types of effects, because the variations are indicators of the extent of inequality produced by the schooling system.

The assumptions of our model imply that each school's mean, μ_j, is given by

$$\mu_j = \mu + A_j + S_j.$$

Hence, it follows that the between-group variances (denoted by τ) and within-group variances (denoted by σ^2) are

$$E(\mu_j - \mu)^2 = \tau(A) + \tau(S) + 2\,\mathrm{cov}(A, S), \tag{4}$$

and

$$E(Y_{ij} - \mu_j)^2 = \sigma^2(S) + \sigma^2(AS) + 2\text{cov}(S, AS) + \sigma^2(e), \qquad (5)$$

where

$$\tau(A) = \tau(P) + \tau(C) + 2\ \text{cov}(P, C),$$
$$\text{cov}(A, S) = \text{cov}(P, S) + \text{cov}(C, S),$$
$$\sigma^2(AS) = \sigma^2(PS) + \sigma^2(CS) + 2\ \text{cov}(PS, CS), \quad \text{and}$$
$$\text{cov}(S, AS) = \text{cov}(S, PS) + \text{cov}(S, CS).$$

Table 1 displays the sources of variation in student outcomes under our model. There are several notable features in how the variation is partitioned. First, the student background differences contribute to both between- and within-school variation. Second, school practice and school context each contribute to both the between-school variation (through the main effects) and the within-school variation (through interactions with student background). Third, both the between-school and the within-school variation include components of covariation between school effects and student background.

The consequences of the model for understanding how school effects influence student outcomes are fundamental. Suppose that a researcher partitioned the total variation in student outcomes into two components: variation within schools and variation between schools. One might assume that the amount of between-school variation puts an upper limit on the variation of school effects (either Type A or Type B). However, this assumption would be fallacious, because either type of school effect can influence within-school variation by interacting with student background. Willms and Chen (1989) found, for example, that the extent of between-classroom segregation of high- and low-ability students in a sample of 26 Israeli primary schools was related to variation in the size of the achievement gap between two major ethnic groups. This effect of school practice was manifest in student-level, not school-level, variation.

Moreover, the variation attributable to either type of school effect could, in principle, be larger than the overall variation between schools for another reason. It is possible that the covariation between either type of school effect and student background effects is negative. For example, a school district might have effective compensatory policies, such that the achievement levels of disadvantaged students were boosted considerably. In this case there would be a negative correlation between the practice effect and the student background effect. Raudenbush, Eamsukkawat, Di-Ibor, Kamali, and Taoklam (1993) provide an example from research on primary schools in Thailand. The agency administering most rural primary schools in that country assigns newly trained teachers to teach for several years in the most remote, rural schools. These schools are attended by some of the most disadvantaged

students in the society; yet, the newly trained teachers tend to be comparatively highly trained and effective. Hence, the researchers found a significant negative correlation between student background and instructional effectiveness.

Thus, Table 1 shows that potentially large components of variation, both between and within schools, may be ambiguous. Some of the variation between schools may arise from the covariation between student background and either school context or school practice. Similarly, part of the variation within schools may result from the covariation between student background effects and interaction effects involving school-level and student-level variables.

Ambiguity again arises in attempting to partition variation due to Type A effects into components associated with school practice and school context. Context and practice may covary. If they do, estimation of variation attributable to context or practice requires specification of both in the model. Even if such specification is complete, part of the variation of the Type A effects will be attributable to context-practice covariation and therefore cannot be assigned uniquely to either.

TABLE I

Decomposition of variation between and within schools for the school effects model $Y_{ij} = \mu + A_j + S_j + (AS)_{ij} + (S_{ij} - S_j) + e_{ij}$

Source	Variation
Between schools	$\tau(A) + \tau(S) + 2\ cov\ (A, S)$
Type A effects	$\tau(A)$
Practice	$\tau(P)$
Context	$\tau(S)$
Practice-Context covariation	$2\ cov\ (A, S)$
Student Background	$\tau(S)$
Type A, Student Background covariation	$2\ cov\ (A, S)$
Practice-Background covariation	$2\ \text{cov}\ (P, S)$
Context-Background covariation	$2\ \text{cov}\ (C, S)$
Within schools	$\sigma^2(S) + \sigma^2(AS) + 2\ cov\ (S, AS) + \sigma^2(e)$
Student Background	$\sigma^2(S)$
Type A by Student Background	$\sigma^2(AS)$
Practice by Student Background	$\sigma^2(PS)$
Context by Student Background	$\sigma^2(CS)$
Covariation of PS, CS	$2\ cov\ (PS, CS)$
Covariation. S, Type A by S	$2\ cov\ (S, AS)$
Covariation S, PS	$2\ cov\ (S, PS)$
Covariation S, CS	$2\ \text{cov}\ (S, CS)$
Error	$\sigma^2(e)$

Techniques for Estimation

Type A Effects

Equations 1 and 2 suggest two alternative approaches to the estimation of Type A effects. First, these effects could be estimated by addition:

$$\hat{A}_{ij} = \hat{P}_{ij} + \hat{C}_{ij} \quad . \tag{6}$$

This approach would require that the model given by Equation I be fully specified; that is, variables representing school practice, school context, and student background would have to be measured and included in the model in order to guarantee that $\hat{P}_{ij}$ and $\hat{C}_{ij}$ were unbiased. Though educational indicator systems will sometimes include good measures of student background (e.g., prior measures of achievement and demographic characteristics) and school context (e.g., the socioeconomic level of the community), they usually fail to specify adequately the contributors to the school practice effect (e.g., school organization, leadership, and instructional quality). Constructs pertaining to school quality are generally more difficult to define and measure, and the relevant data are comparatively expensive to collect (Willms, 1992, chapter 6).

An alternative approach is to estimate Type A effects by subtraction because

$$A_{ij} = Y_{ij} - \mu - S_{ij} - e_{ij}. \tag{7}$$

Clearly, A_{ij} will be estimated without bias only if S_{ij} is estimated without bias. Indeed, this subtraction method has been the method used in nearly all studies of what we are calling Type A effects. Estimation of S_{ij} will be unbiased if two conditions hold. First, relevant data on student background must be collected. Second, the statistical estimation procedure must be unbiased. Our illustrative example (below) shows that achieving unbiased statistical estimation requires careful thinking (see "Some Biased Estimators of School Effects" below).

Type B Effects

Like Type A effects, Type B effects can be estimated by either addition or subtraction; for Type B effects, however, both student background and school context must be estimated without bias. The estimator is, in general,

$$\hat{B}_{ij} = \hat{P}_{ij} = Y_{ij} - \hat{\mu} - \hat{C}_{ij} - \hat{S}_{ij}, \tag{8}$$

which shows that both S_{ij} and C_{ij} must be estimated without bias. The problem is that unless C_{ij} is orthogonal to P_{ij}, failure to specify P_{ij} will render the estimate of C_{ij}—and therefore the estimate of the Type B effect-biased and

inconsistent. The direction of bias is unknown unless the direction of the correlation between C_{ij} and P_{ij} is known.

When the correlation between effective practice and school context is positive, estimates of Type B effects will be biased against schools with advantaged contexts. This would occur, for example, if schools with advantaged contexts also tended to have better teachers, more effective school leaders, and more resources than schools with less advantaged contexts. In this case, an evaluation based on Type B effects computed via subtraction as in Equation 8 would attribute the superior outcomes of such schools too much to their advantaged contexts (e.g., the socioeconomic context of the community from which the school draws its students) and too little to the effective educational practices in those schools.

When the correlation between effective practice and context is negative, an evaluation based on such Type B effect estimators will be biased against schools with less advantaged contexts. Such a correlation could occur, for instance, under a compensatory policy in which resources were directed preferentially to disadvantaged schools.

We see, then, that consistent estimation of Type B effects requires gathering information on school organization, management, climate, instructional practice, and other relevant features of school functioning. In essence, it is impossible to know how effective a school's practice is without a theory of what makes school practice effective and a measurement strategy that leads to reliable and valid measures of the dimensions of school practice. However, by assuming that school context and practice are, in general, positively correlated (after controlling for S_{ij}), it is possible to place a lower bound on the effectiveness of practice in more advantaged schools and an upper bound on the effectiveness of practice in less advantaged schools. Moreover, the next section shows that estimates of school practice based on subtraction (Equation 8) can lead to brackets (lower and upper bounds) on the variance of these effects if one assumes that context and practice are positively correlated (controlling S_{ij}).

The Variance of School Effects

Investigators often wish to summarize the importance of school differences by estimating the proportion of variance in the outcome attributable to school effects. Just as it is possible to estimate the Type A effects without bias, even if school practice is unspecified, it is possible to estimate the variance, $\tau(A)$, of those effects without bias using appropriate multilevel statistical techniques. However, it is not possible to estimate the variance of the Type B effects consistently when practice is unmeasured except in the unlikely case that practice and context are orthogonal. Suppose that one derives an estimator $\hat{\tau}(B)$ of the Type B effects that is consistent under the assumption that school context and practice are orthogonal. If this assumption is false, the large-sample expectation of this estimator will set a lower bound for the

true variance, regardless of the direction of the correlation. Moreover, if one is willing to assume that the correlation between school context and practice is positive, then the large-sample expectations of Type A and Type B effects variance bracket the true Type B effects variance; that is,

$$P[\hat{\tau}(B) < \tau(B) < +(A)] \rightarrow 1 \qquad (9)$$

as the sample size of schools increases. These principles are illustrated in the following example.

An Example for One Scottish Local Authority

We analyze data from 5,054 students attending 20 secondary schools in Fife Education Authority. The outcome measure is the score on the "O-grade" English examination, which was taken by the majority of pupils in their fourth year of secondary school. Success in these examinations is an important determinant of postsecondary employment (Raffe, 1984). and people who do well on these examinations are more likely to remain in school beyond the compulsory period. For simplicity of exposition, we initially employ a single covariate, the reading score derived from a test administered at the end of primary 7, just prior to students' entry to secondary school. The fourth-year English scores were standardized to have a mean of zero and a standard deviation of one for the entire sample (see Willms & Raudenbush, 1989, for details on how these data were scaled). The primary 7 reading scores had a mean of 100 for the education authority, and a standard deviation of 14.26. These scores were centered (but not standardized) by subtracting 100 from each individual's score.

Uniform Effects Model

We consider first the statistical model

$$Y_{ij} = a + \beta_w X_{ij} + \gamma_c \bar{X}_j + u_j + e_{ij} \qquad (10)$$

where Y_{ij} is the fourth-year English outcome for student i in school j; X_{ij} is the primary 7 reading test score for that student; $\bar{X}_j$ is the school sample mean reading test score for school j; u_j is the random effect of school j, assumed normally distributed with mean of zero and variance τ; and e_{ij} is the student-level random error, assumed identically, independently, and normally distributed with mean zero and a variance σ^2. With these definitions in mind, the regression coefficients take on clear definitions: a is the education authority mean English score; β_w is the pooled, within-school regression coefficient relating primary 7 reading to secondary 4 English; and γ_c is the so-called "contextual effect" of prior reading, that is, the influence of attending a school having a mean reading score of X_j after controlling for one's own reading score.

An orthogonal reparamaterization of the model makes the definition of the contextual coefficient precise:

$$Y_{ij} = a + \beta_w(X_{ij} - \bar{X}_j) + \beta_b\bar{X}_j + u_j + e_{ij}. \quad (11)$$

Here Y_{ij}, u_j, and e_{ij} are defined as in Equation 10, but the student-level predictor is now $X_{ij} - \bar{X}_j$, that is, student reading deviated around the school mean; this deviation is orthogonal to the school mean, $\bar{X}_j$. This rescaling preserves the meaning of the parameters a and β_w. However, the school-level coefficient is now β_b, the between-school slope, that is, the regression of the school mean outcome on the school mean reading. The relationship between β_b and γ_c is such that $\gamma_c = \beta_b - \beta_w$, enabling us to see that the contextual coefficient, γ_c, is defined in this model as the difference between the between- and within-school regression coefficients (Burstein, 1980).

The contextual coefficient, γ_c, has been the topic of considerable discussion in educational sociology. Willms's (1986) review indicates that positive contextual coefficients have been found in many countries. Underspecification of student-level predictors can lead to spuriously non-null contextual coefficients (Hauser, 1970); however, when student and school practice effects are adequately specified, the contextual coefficient represents the contribution of an aggregated student characteristic to student outcomes over and above the effect of that characteristic measured at the student level. For example, one might find the expected outcome of attending a school of high average prior reading to be higher than the expected outcome of attending a school of low average prior reading, if one's own prior reading is held constant. In essence, the student's able classmates provide a context that facilitates student learning.[5]

In the case of our data, as is the case with most data used to construct indicators of school effects, no information has been collected on school practice. Rather, effects of school practice are conceived to be the unobservable part of the school effect that remains after removing the contributions of student background and context. Hence, the terms of the general model (Equation 1) relate to the terms of Equation 10 as follows:

$$\mu = a; \qquad P_{ij} = u_j; \qquad C_{ij} = \gamma_c\bar{X}_j; \qquad \text{and } S_{ij} = \beta_w X_{ij}. \quad (12)$$

It follows that the Type A effect is

$$A_{ij} = \gamma_c\bar{X}_j + u_j = Y_{ij} - a - \beta_w X_{ij} - e_{ij}. \quad (13)$$

Equation 13 implies that A_j takes on the same value for every student in school j ($A_{ij} = A_j$); and so we refer to this as a *uniform effects model* in that the model assumes that a school has a uniform effect on each student. Averaging Equation 13 within each school shows that

$$A_j = \bar{Y}_j - a - \beta_w \bar{X}_j - \bar{e}_j, \tag{14}$$

where $\bar{Y}_j$ and $\bar{e}_j$ are school sample means of Y_{ij} and e_{ij}, respectively. Hence, A_j can be estimated without bias if a and β_w are estimated without bias.

Estimates of Type A effects. Unbiased and asymptotically efficient estimates of the regression parameters in Equation 10 can be readily obtained by any of the recently developed maximum likelihood algorithms for nested random effects models reviewed by Kreft, de Leeuw, and Kim (1990). The program HLM 3.0 of Bryk, Raudenbush, and Congdon (1994) yielded the following results: $\hat{\alpha}$ = -0.097 *(SE* = .032); $\hat{\beta}_w$ = *0.052 (SE* = .00067); $\hat{\gamma}_c$ = .011 *(SE* = .0085). An unbiased estimate of each school's Type A effect is therefore given by

$$\hat{A}_{ij} = \bar{Y}_j - \hat{\alpha} - \hat{\beta}_w \bar{X}_j, \tag{15}$$

which in our case yields

$$\hat{A}_j = \bar{Y}_j + 0.097 - 0.052\ \bar{X}_j.$$

(Recall that prior reading had a grand mean of zero in the education authority.) The estimator given by Equation 15, though simple, is, unfortunately, rarely used in practice as we shall see in the section "Some Biased Estimators of School Effects" below.

Estimates of Type B effects. By similar logic, assuming X and u to be orthogonal, consistent estimators of the Type B effects are

$$\hat{B}_j = \hat{u}_j = \bar{y}_j - \hat{\alpha} - \hat{\beta}_b \bar{X}_j, \tag{16}$$

where β_b is defined in Equation 11. We do not, however, expect X and u to be orthogonal. When they are not, our point estimate of B_j will be biased and inconsistent, having large-sample expectation

$$E(\hat{B}_j) = B_j - \beta_{Xu} \bar{X}_j, \tag{17}$$

where β_{Xu} is the regression coefficient relating the outcome u_j to the predictor $\bar{X}_j$

Maximum likelihood variance estimation. The variances of the Type A and Type B effects under our model are

$$\tau(A_j) = \gamma_c^2 \operatorname{var}(\bar{X}_j) + \tau \tag{18}$$

$$\tau(B_j) = \tau.$$

Substituting maximum likelihood (ML) estimates for the parameters will

yield efficient estimators $+(A)$ and $\hat{\tau}(B)$ of these variances. It can readily be shown that $?(A)$ will be consistent even if school context $(\bar{X}_j)$ and school practice (u_j) are correlated. However, regardless of the direction of the correlation between school context and school practice, Equation 18 underestimates the true variance of the school effects; the large-sample expectation of this estimate is

$$E[\hat{\tau}(B)] \approx \tau(B)(1 - \rho^2_{\bar{X}u}), \tag{19}$$

where $\rho_{\bar{X}u}$ is the population correlation between $\bar{X}_j$ and u_j. Thus, in large samples, $\hat{\tau}(B)$ sets a lower bound for $\tau(B)$. Unless the correlation between $\bar{X}_j$ and u_j is negative, the upper bound occurs when the context effect, $\gamma_c \bar{X}_j$ is 0, in which case $\tau(B) = \tau(A)$. *Thus,* assuming $\bar{X}_j$ and u_j to be positively correlated, it is reasonable to infer that $\hat{\tau}(B) < \tau(B) < \hat{\tau}(A)$ when samples are large.

Nonuniform Effects Model

Our general model allows for the possibility that the effect of a school varies from child to child within that school. So far, however, the illustrative analyses have been based on the assumption that school effects are independent of child background-in effect, that a school has a uniform effect on the children who attend it. A simple example of an analysis assuming nonuniform effects follows. In this case, a school effect is computed for each child in the sample. The logic of estimation remains unchanged. However, examining the distribution of school effects does change. We recommend looking at the distribution conditional on particular values of variables describing student background.

Returning now to the Scottish data, we elaborate our model to include a randomly varying coefficient for students' prior reading scores. The model becomes

$$Y_{ij} = \alpha + \beta_w X_{ij} + \gamma_c \bar{X}_j + u_{0j} + u_{1j} X_{ij} + e_{ij}, \tag{20}$$

and we now assume

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right]. \tag{21}$$

Our definitions are now

$$S_{ij} = \beta_w X_{ij}; \qquad Cj = \gamma_c \bar{X}_j; \qquad \text{and } P_{ij} = u_{0j} + u_{1j} X_{ij}. \tag{22}$$

The primary difference with the former model is that P_{ij} now varies from

student to student within schools. Estimates of the school effects follow the earlier logic

$$\hat{A}_{ij} = Y_{ij} - \hat{\alpha} - \hat{\beta}_w X_{ij}, \qquad (23)$$
$$\hat{B}_{ij} = \hat{A}_{ij} - \hat{\gamma}_c \bar{X}_j.$$

The previous cautions about interpreting the Type B estimates still apply.

Because the school effect varies as a function of prior reading scores, we define the variance of the school effects conditional on a given value of $X_{ij} = X_o$:

$$\text{var}(A_{ij} \mid X_{ij} = X_o) = \gamma_c^2 \text{ var}(\bar{X}_j) + \text{var}(B_{ij} \mid X = X_o), \qquad (24)$$
$$\text{var}(B_{ij} \mid X_{ij} = X_o) = \tau_{00} + \tau_{11} X_o^2 + 2\tau_{01} X_o.$$

ML variance estimation substitutes ML estimates for the parameters of Equation 24. The relevant estimates for our example are (again we simplify the model by including only one covariate, primary 7 reading) $\hat{\alpha}$ = -0.100 (SE = .034); $\hat{\beta}_w$ = 0.051 (*SE* = .00064); $\hat{\gamma}_c$ = .0028 (*SE* = .0064); τ_{00} = .02082; τ_{01} = .00066; τ_{11} = .00003. Even though the estimate of τ_{11} appears small, it is statistically significant, which indicates that there are statistically significant differences between schools in their English/prior reading regression coefficients.

Some Biased Estimators of School Effects

Aggregated Pupil-Level Residuals

To estimate school effects, many investigators have specified a simple regression model of the form

$$Y_{ij} = a + \beta_t X_{ij} + \epsilon_{ij} \qquad (25)$$

and estimated the parameters using ordinary least squares (OLS). β_t in this model is the regression slope for the full sample, without regard to the nested structure of the data. The errors ϵ_{ij} are assumed independent with constant variance. The estimates of school effects are the means for each school of the pupil-level residuals:

$$\hat{EFFECT}_j = \bar{Y}_j - \hat{\alpha} - \hat{\beta}_t \bar{X}_j. \qquad (26)$$

This method was used in some of the first studies of school effects (e.g., Coleman et al., 1966), and corresponds to Method 1 of Aitkin and Longford (1986). The method assumes uniform effects of schools. However, even when

the uniform effects model holds, this method produces an uninterpretable blend of Type A and Type B effects, which can be seen using an identity recognized by Alwin (1976):

$$\hat{\beta}_t = \eta^2\hat{\beta}_b + (1 - \eta^2)\hat{\beta}_w. \tag{27}$$

Here β_b and β_w are the between-school and within-school slopes (defined earlier), respectively, and η^2 is the proportion of the total variance in X_{ij} that lies between schools, that is,

$$\eta^2 = \sum_{j=1}^{J} n_j \bar{X}_j^2 / \sum_{j=1}^{J} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2.$$

Substituting Equation 27 into Equation 26, we get

$$\hat{EFFECT}_j = \bar{Y}_j - \hat{\alpha} - \eta^2\hat{\beta}_b\bar{X}_j - (1 - \eta^2)\hat{\beta}_w\bar{X}_j. \tag{28}$$

Consider two extreme possibilities:

(1) $\eta^2 \rightarrow 0$; that is, all schools have similar mean scores on X, and therefore nearly all variation in X is within schools.[6] In this case, $\hat{\beta}_b$ is accorded essentially no weight in Equation 28 and, in any case, converges to β_w so that the estimates of school effects are identical to our estimates of Type A effects (see Equation 14).

(2) $\eta^2 = 1$; that is, all of the variance in X is between schools. This can occur only if all pupils within each school have the same score on X. In this case, the weight assigned to β_w becomes zero and the aggregated residuals become estimates of Type B effects (see Equation 16). These two extreme conditions do not occur in most natural settings, although η^2 can be close to zero in school districts serving predominantly rural areas. Willms (1983) estimated η^2 for socioeconomic status (SES) to be 0.25 for U.S. secondary schools in 1980, based on the High School and Beyond data. His estimate of η^2 for Scotland in 1980, based on a similar composite measure of SES, was 0.22, and Scotland's secondary system was one of the most comprehensive schooling systems in Europe at that time (Willms, 1986). We expect η^2 for SES to range from 0.20 to 0.40 for most developed countries, and to be higher in some developing countries.

Finally, a third possibility can arise in systems where there are no contextual effects:

(3) $\gamma_c = \beta_b - \beta_w \simeq 0$. In this case, the effects based on aggregate residuals become identical to the Type A effects. This can be the case in some schooling systems. In our example, there was no statistically significant contextual effect of prior reading scores on English examination results, but that is not generally the case.

In most cases, therefore, aggregated residuals will yield a blend of the Type A and Type B estimates and therefore a biased estimate of both effects. The extent of the bias depends on (a) which effect is of interest, (b) the extent of between-school segregation, and (c) the size of contextual coefficient.

Estimates Based on Residuals From Means-on-Means Regression

This method entails aggregating data to the school level and regressing school mean outcomes on school means of the covariates. The estimate of the school effect is the residual from the school-level regression:

$$\hat{EFFECT}_j = \bar{Y}_j - \hat{\alpha} - \hat{\beta}_b \bar{X}_j. \tag{29}$$

This estimator has the same structure as our Type B effect estimator (Equation 16) and is unbiased under the condition that school context and school practice are uncorrelated. However, this method yields less efficient estimates than does Equation 16. While Equation 16 uses information about the covariation between X and Y at the student level to reduce the within-school variance, this method, based on aggregation, ignores this information and therefore increases the sampling variance of the estimator.

Estimates Based on Banding

In the California assessment program, schools are rank-ordered on an index of SES. A band of 100 schools is then set for each school by taking the 50 schools ranking above and the 50 schools ranking below the particular school. The estimate of a school's effect is then that school's rank for its mean outcome score among the 100 schools in its band. This method is appealing because it involves simple calculation and is easily explained to policymakers. This type of estimate is really a form of Type B effect, and, like the estimates based on means-on-means regressions, it is biased when practice and context are correlated.

Biased but n-Consistent Estimates of β_w

One might use a hierarchical linear model and ML estimation to estimate the following model:

$$Y_{ij} = a + \beta X_{ij} + Uj + e_{ij}. \tag{30}$$

By summing over pupils within each school and rearranging the equation we obtain

$$u_j = \bar{Y}_j - a - \beta \bar{X}_j - \bar{e}_j. \tag{31}$$

Thus, u_j appears similar to A_j as defined by Equation 14. However, β is not

the same as β_w, and therefore the estimate of β obtained by ML estimation of Equation 31 is not identical to the $\hat{\beta}_w$ obtained when the contextual effect term, $\gamma_c \bar{X}_j$, is included in the model. Again, these estimates will be unbiased only if no contextual effect is present. If a contextual effect is present, these estimates will be biased. However, the estimates will be n-consistent. That is, as the sample size per school increases, they will converge to the Type A effect.

Source of the bias. To understand the problem of estimating Type A effects in this way, suppose the data were balanced so that every school had the same sample size of students. Then the OLS and ML estimates of β_w and β_b (Equation 11) and their sampling variances would be identical, with

$$\hat{\beta}_w = \frac{W_{xy}}{W_{xx}}, \qquad \text{var}(\hat{\beta}_w) = \frac{\sigma^2}{W_{xx}}, \tag{32}$$

and

$$\hat{\beta}_b = \frac{A_{xy}}{A_{xx}}, \qquad \text{var}(\hat{\beta}_b) = \frac{\Delta}{A_{xx}}, \tag{33}$$

where

$$\begin{aligned} W_{xy} &= \sum_{j=1}^{J} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(Y_{ij} - \bar{Y}_j), \qquad W_{xx} = \sum_{j=1}^{j} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2, \\ A_{xy} &= \sum_{j=1}^{J} \bar{X}_j \bar{Y}_j, \qquad A_{xx} = \sum_{j=1}^{J} \bar{X}_j^2, \qquad \Delta = \tau^2 + \frac{\sigma^2}{n}. \end{aligned} \tag{34}$$

Suppose now that we wish to estimate β in Equation 30 by means of ML. From now-standard formulas for the fixed effects in a two-level hierarchical linear model, the ML estimators, given known τ^2 and σ^2, are

$$\begin{aligned} \hat{\beta} &= \frac{[\text{var}(\hat{\beta}_b)]^{-1}\hat{\beta}_b + [\text{var}(\hat{\beta}_w)]^{-1}\hat{\beta}_w}{[\text{var}(\hat{\beta}_b)]^{-1} + [\text{var}(\hat{\beta}_w)]^{-1}}, \\ \hat{\alpha} &= \bar{Y}. \end{aligned} \tag{35}$$

Of course, the variances will not be known, in which case the ML estimator of β is Equation 35 with ML variance estimators substituted. Equation 35 provides insight into the logic of ML estimation theory for the hierarchical model. Suppose that no contextual effect exists, that is, that $\beta_b = \beta_w = \beta$. Then both β_w and $\hat{\beta_b}$ are unbiased estimates of β. The ML estimate of β is then the precision-weighted average of $\hat{\beta}_w$ and $\hat{\beta}_b$. When no contextual effect

is present, this precision-weighted average is optimal for squared error loss because it makes full use of the information in the data regarding this unknown parameter. However, when a contextual effect is present, the precision-weighted average is a biased estimator of the desired coefficient, β_w. Specifically (with known variances),

$$\text{Bias}(\hat{\beta}) = \frac{[\text{var}(\hat{\beta}_b)]^{-1}}{[\text{var}(\hat{\beta}_b)]^{-1} + [\text{var}(\beta_w)]^{-1}} \gamma_c. \tag{36}$$

Now if we set

$$D_x^2 = \frac{\sum_{j=1}^{J} \bar{X}_j^2}{J} \quad \text{and} \quad S_x^2 = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2}{nJ}, \tag{37}$$

we can reexpress the bias as

$$\text{Bias}(\hat{\beta}) = \frac{1}{n} \left(\frac{\frac{D_x^2}{\Delta}}{\frac{S_x^2}{\sigma^2} + \frac{D_x^2}{n\Delta}} \right) \gamma_c, \tag{38}$$

which shows that although the ML estimator of β is biased, it is n-consistent. As the number of students per school increases, the bias becomes negligible. In our Scottish data, with an average of about 250 students per school, with most of the variation in prior reading scores lying within schools (so that the precision of $\hat{\beta}_w$ greatly exceeds that of $\hat{\beta}_b$), and with a nonsignificant contextual coefficient, the bias is negligible: the estimate of β (based on Equation 43) was 0.052 179, while the unbiased estimate of β_w (Equation 23) was 0.052 112.

Biased but Accurate Estimation via Empirical Bayes

It is not necessarily true that biased estimators are inferior to unbiased estimators. If minimizing a loss function is the goal, the biased estimator may be superior. The empirical Bayes estimator of school effects provides an example. For simplicity, we illustrate this superiority in the case of the uniform effects model, though the same logic applies in the nonuniform effects case.

The previously recommended estimates of B_j and A_j in our uniform effects example are given by

$$\hat{B}_j = \bar{Y}_j - \hat{\alpha} - \hat{\beta}_b \bar{X}_j = \hat{u}_j$$

$$= B_j + \bar{e}_j - [\hat{\alpha} - \alpha + (\hat{\beta}_b - \beta_b)\bar{X}_j] \tag{39}$$
$$= B_j + \bar{e}_j + o(J^{-1}).$$

$$\hat{A}_j = \bar{Y}_j - \hat{\alpha} - \hat{\beta}_w\bar{X}_j = \hat{u}_j + \hat{\gamma}_c\bar{X}_j$$
$$= A_j + \bar{e}_j - [\hat{\alpha} - \alpha + (\hat{\beta}_w - \beta_w)\bar{X}_j] \tag{40}$$
$$= A_j + \bar{e}_j + o(J^{-1}).$$

Thus, given the realized values of u_j, j = 1, . . . , J, $\hat{A}_j$ and $\hat{B}_j$ are unbiased and each has limiting (large-J) conditional variance

$$\text{var}(\hat{A}_j \mid u, X) \approx \text{var}(\hat{B}_j \mid u, X) \approx \frac{\sigma^2}{n_j}, \tag{41}$$

where u = $(u_1, \ldots, u_J)^T$. Alternative estimators are the posterior conditional means

$$B_j^* = u_j^* = E(u_j \mid \sigma^2, \tau^2, Y) = \lambda_j\hat{u}_j \tag{42}$$

and

$$A_j^* = E(A_j \mid \sigma^2, \tau^2, Y) = u_j^* + \hat{\gamma}_c\bar{X}_j, \tag{43}$$

where

$$\lambda_j = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n_j}} \tag{44}$$

can be viewed as the reliability of $\hat{u}_j$ as an estimate of u_j. Applications of such empirical Bayes estimators in educational research are discussed by Raudenbush (1988). These estimators are conditionally biased with

$$\text{Bias}(u_j^* \mid u_j) = -(1 - \lambda_j)u_j. \tag{45}$$

The limiting (large-J) conditional variance is

$$\text{var}(u_j^* \mid u_j) \approx \frac{\lambda_j^2\sigma^2}{n_j}. \tag{46}$$

It is instructive to compute the (large-4 asymptotic relative efficiency of

the two estimators of u_j, that is, the ratio of the large-J expectations of their mean squared errors of estimation, which reduces in the balanced case to

$$\frac{E[MSE_{u_j^*} \mid u_j]}{E[MSE_{\hat{u}_j} \mid u_j]} \approx \frac{E\left[\sum_{j=1}^{J} (u_j^* - u_j)^2 \mid u_j\right]}{E\left[\sum_{j=1}^{J} (\hat{u}_j - u_j)^2 \mid u_j\right]}$$

$$\approx \frac{\sum_{j=1}^{J} E[\text{Bias}^2(u_j^* \mid u_j) + \text{var}(u_j^* \mid u_j)]}{\sum_{j=1}^{J} [\text{var}(\hat{u}_j \mid u_j)]} \qquad (47)$$

$$\approx \lambda.$$

Thus, the asymptotic efficiency of the empirical Bayes estimator relative to that of the conventional estimator, using expected mean squared error loss as a criterion, is equal to λ, the reliability of the latter. As this reliability increases, the two estimates, and hence their efficiency, converge. When the reliability is small, either because within-school samples are small or because little variance in u_j exists between schools, the empirical Bayes estimator dominates the conventional estimator.

Illustrative *example based on empirical Bayes.* We extended the model to include measures of SES and gender in addition to primary 7 reading. These variables were both significant predictors of secondary 4 English results, and therefore potentially improve our estimates of Type A effects. Models that include both SES and prior achievement are relatively well specified; the addition of more covariates (e.g., other prior achievement measures) has a negligible effect on the estimates of the Type A effects (see Willms, 1992, chapter 7). Because no context effect was discernible, we based Type A estimates on Equation 35. The empirical Bayes estimates for each of 20 schools are shown in Figure 1.

Standard errors of the empirical Bayes estimator. With variances known and *J* large, the posterior variance of the school effect u_j given the data, Y, is given by

$$\text{var}(u_j \mid Y) = \tau(1 - \lambda_j). \qquad (48)$$

However, with variances unknown and *J* finite (as in the case of our example with 20 schools that we present below), no closed-form expression is available for this variance. Therefore, we estimate the uncertainty associated with the empirical Bayes estimator via the bootstrap (Efron & Gong, 1983; Efron & Tibshirani, 1986; Hinkley, 1988). Bootstrapping mimics repetitions of sam-

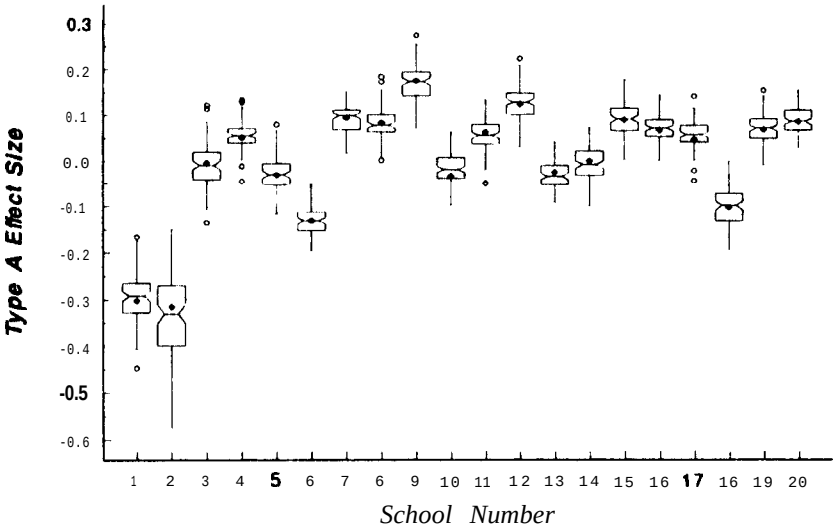

FIGURE *I. Box plots showing variability in estimates of Type A effects for 20 schools, based on 100 bootstrap replications*

pling by drawing a large number of samples, with replacement, from the achieved sample. Sampling variation of an estimator is then examined directly by estimating the desired estimator for each of the bootstrap samples. In our example, we drew a separate bootstrap sample from each of the schools in the study and estimated the empirical Bayes estimators. This was done 100 times to produce the sampling distributions of the estimates. The sampling distributions of these estimates, based on 100 bootstrap replications, are shown with box plots in Figure 1.[7]

Summary and Policy Implications

Our inquiry supports the following recommendations for researchers aiming to estimate the effects of particular schools:

(1) *Decide what effect is of interest.* Our Type A effect includes the effects of school practice and the contextual influence of wider social and economic factors that lie outside the immediate control of teachers and administrators; this is the effect relevant to parents choosing schools for their children. Our Type B effect, which includes only the effects of school practice, is relevant to school staff who desire an indication of their school's performance and to administrators interested in accountability.

(2) Consider the logical conditions for valid causal inference. Studies of school effects are quasi-experiments. Valid inference requires that treatment assignment be strongly ignorable in the sense of Rosenbaum and Rubin (1983) and as elaborated earlier in the case of Type A and Type B effects. Strong ignorability is a viable, though not easily achievable, goal in the

case of Type A effects because investigators will commonly have access to information on student background characteristics related both to the propensity to attend each given school and to the possible outcome of attending each school. Causal inference is much more problematic in the case of Type B effects because the treatment-school practice-is typically undefined so that the correlation between school context and school practice cannot be computed. Thus, even if the assignment of students to schools were strongly ignorable, the assignment of schools to treatments could not be.

(3) Match the estimation procedure to the effect of interest. The two effects require different procedures for estimation (even assuming school context and practice to be orthogonal). Appropriate estimates of both Type A and Type B effects can bracket the Type B effect if the investigator is willing to assume school context and practice to be positively correlated.

(4) Assess the comparability of the schools and the sensitivity of results to choice of covariates. Schools will be comparable only if their covariate distributions exhibit considerable overlap. The more the schools differ on the covariates, the more sensitive effect estimates will be to choice of covariates High sensitivity is an indicator that the estimates are not trustworthy.

(5) Consider the possibility that a school will influence different students differently. The assumption that a school has a uniform effect on all who attend it is hard to defend theoretically. Methods described in this article can incorporate varying effects of schools depending on student background.

This inquiry also has consequences for research design in studies aiming to support school improvement. Our discussion of Type B effects shows that one cannot know how effective a school's practice is without a theory of what makes school practice effective. Such a theory and appropriate measures, though difficult to collect, supply a foundation for studying the contributions of school practice. However, even then, the basis for causal inference will be fragile, requiring that both the assignment of schools to practices and the assignment of students to schools are strongly ignorable. The implication is that the research agenda for school improvement should include controlled field trials in which schools serving similar students are randomly assigned to alternative treatments that have shown promise in nonexperimental research. It appears that a spectrum of research designs-ranging from large-scale surveys exploring multiple predictors of student outcomes to highly controlled trials testing sharply focused hypotheses—is needed to provide a solid empirical basis for school evaluation and reform.

Notes

[1] Willms (1992, chapter 6) points out that the boundary between practice and context variables is not well defined. For example, a variable like "parental press for academic achievement" (ostensibly a "context" variable) might be affected considerably by a school staff through its homework policies and involvement with parents, or might be affected only minimally. Nevertheless, the distinction is useful conceptu-

ally, and the blurry boundary between context and practice only reinforces the claim that "the estimation of school effects is not an exact science" (Willms. 1992, p. x).

[2] Student social background can be manipulated by broad social policy such as income redistribution, but not typically by educational policy or practice.

[3] Holland (1986) describes this as "Rubin's theory" of causation.

[4] The schools must have substantially overlapping distributions on the covariates, x, to achieve strong ignorability.

[5] We have chosen for simplicity to ignore the measurement error of the sample mean X_j as a measure of the latent true mean of X within school j. The reliability of this sample mean is $\tau_x/[(\tau_x + \sigma_x^2)/n_j]$, where τ_x is the between-school variance and σ_x^2 is the within-school variance on X. Experience shows that for n_j as small as 25 per school this reliability typically exceeds .90. For our data it was .94. Because sample sizes tend to be much larger than 25 in most evaluations of school effectiveness, we have chosen to ignore this source of error in the interest of keeping the discussion focused on essential issues.

[6] Consider the extreme case in which students are assigned at random to schools, so that the intraschool correlation on X is zero. Then the large sample expectation of η^2 is of order $1/n$, where n is the typical sample size per school.

[7] Box plots summarize graphically the distribution of a set of scores. The median is shown by the horizontal line within the box, and the 75th and 25th percentiles are shown by the top and bottom of the box. The upper and lower vertical lines extend from the box to the maximum and minimum values respectively, except for outliers, which are shown as individual points. Outliers are values that are more than 1.5 times the interquartile range away from the upper or lower quartiles. See McGill. Tukey, and Larsen (1978) for a full discussion of box plots.

References

Aitkin, M., & Longford. N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A, 149*(1), *1-43.*

Alwin, D. F. (1976). Assessing school effects: Some identities. *Sociology of Education,* 49, 294–303.

Bryk, A. S.. Raudenbush, S. W., & Congdon, R. (1994). *Hierarchical linear modeling with the HLM/2L and HLM/3L programs.* Chicago: Scientific Software International.

Burstein, L. (1980). The analysis of multi-level data in educational research and evaluation. *Review of Research in Education, 8, 158-233.*

Coleman, J. S., Campbell, E. Q., Hobson, C. F., McPartland, J., Mood, A. M., Weinfeld, F.D., & York, R. L. *(1966). Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Dyer, H. S., Linn, R. L., & Patton, M. J. (1969). A comparison of four methods of obtaining discrepancy measures based on observed and predicted school means on achievement tests. *American Educational Research Journal, 6, 591-605.*

Echols, F., McPherson, A. F., & Willms, J. D. (1990). Choice among state and private schools in Scotland. *Journal of Education Policy, 5*(3), *207-222.*

Efron, B.. & Gong, G. (1983). A leisurely look at the bootstrap, jackknife, and cross-validation. *American Statistician, 37,* 36–48.

Efron. B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures. *Statistical Science, I, 54-77.*

Goldstein, H. *I.* (1987). *Multilevel models in educational and social research.* London: Oxford University Press.

Good, T. L.. & Brophy. J. E. (1986). School effects. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 570–602). New York: Macmillan.

Gray, J. (1989). Multilevel models: Issues and problems emerging from their recent application in British studies of school effectiveness. In D. R. Bock *(Ed.), Multilevel analyses of educational data* (pp. 127–145). Chicago: University of Chicago Press.

Hauser, R. M. (1970). Context and Consex: A cautionary tale. *American Journal Of Sociology, 7.5, 645-654.*

Heyns, B. (1986). Educational effects: Issues in conceptualization and measurement. In J. G. Richardson *(Ed.), Handbook of theory and research for the sociology of education* (pp. 305-340). Westport, CT: Greenwood Press.

Hinkley, D. V. (1988). Bootstrap methods. *Journal of the Royal Statistical Society, Series B, 50, 321-337.*

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81, 396, 945-960.*

Kreft, I. G., de Leeuw. J.. & Kim, K.-S. *(1990). Comparing four different statistical packages for hierarchical linear regression: GENMOD, HLM, ML2, and VARCL* (CSE Tech. Rep. 3 1 1). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Mandeville, G. K., & Anderson, L. W. (1987). The stability of school effectiveness indices across grade levels and subject *areas. Journal of Educational Measurement,* 24, 203-216.

Marco, G. L. (1974). A comparison of selected school effectiveness measures based on longitudinal data. *Journal of Educational Measurement, II, 225-234.*

Massey, D. S., & Denton, N. A. (1988). Suburbanization and segregation in U.S. metropolitan *areas. American Journal of Sociology, 94, 592-626.*

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *The American Statistician, 32(I),* 12-16.

McPherson, A. F. *(1992). Measuring added value in schools* (NCE Briefing No. 1). London: National Commission of Education.

Murnane, R. J. (1975). *The impact of school resources on the learning of children.* Cambridge, MA: Ballinger Publishing Co.

Raffe, D. (1984). School attainment and the labour market. In D. Raffe *(Ed.), Fourteen to eighteen: The changing pattern of schooling in Scotland* (pp. 174-193). Aberdeen: Aberdeen University Press.

Raffe, D., & Willms. J. D. (1989). Schooling the discouraged worker: Local-labour-market effects on educational participation. *Sociology, 23, 559-581.*

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics, 13,* 85-l 16.

Raudenbush. S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59,* I-17.

Raudenbush. S. W.. Eamsukkawat, S., Di-Ibor, I., Kamali. M., & Taoklam, W. (1993). On-the-job improvements in teacher competence: Policy options and their effects on teaching and learning in Thailand. *Educational Evaluation and Policy Analysis, IS,* 379-297.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Riometrika,* 17, 41-55.

Rowan, B., Bossert, S. J., & Dwyer, D. C. (1983). Research on effective schools: A cautionary note. *Educational Researcher; 12*(4), 24–31.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6, 34-58.*

Stephenson, R. S. (I 986, April). *Is the school an appropriate unit of merit?* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Wheeler, C. W., Raudenbush, S. W.. & Pasigna, A. (1989). *Policy initiatives to improve primary school quality in Thailand: An essay on implementation, constraints, and opportunities for educational improvement* (BRIDGES Report Series No. 5). Cambridge, MA: Harvard Institute for International Development.

Willms, J. D. (1983). *Achievement outcomes in public and private high schools.* Unpublished doctoral dissertation, Stanford University.

Willms. J. D. (1986). Social class segregation and its relationship to pupils' examination results in Scotland. *American Sociological Review, 51,* 224-241.

Willms, J. D. (1992). *Monitoring school performance: A guide for educators.* Washington, DC: Falmer.

Willms, J. D., & Chen, M. (1989). The effects of ability grouping on the ethnic achievement gap in Israeli elementary schools. *American Journal of Education,* 97, 237-257.

Willms. J. D., & Echols, F. H. (1992). Alert and inert clients: The Scottish experience of parental choice. *Economics of Education Review, II. 339-350.*

Willms. J. D., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26, 209-232.*

Wolf, D., Bixby, J., Glenn, J., III, & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17,* 31–74.

Authors

STEPHEN W. RAUDENBUSH is Professor of Measurement and Quantitative Methods, College of Education, Michigan State University, 461 Erickson Hall, East Lansing, MI 48824. He specializes in multilevel and longitudinal statistical methods.

J. DOUGLAS WILLMS is Professor and Research Chair, Faculty of Education, University of New Brunswick, Fredericton, New Brunswick, Canada E3B 6E3. He specializes in sociology of education and research methods.

Received September 22, 1993
Accepted August 12, 1994