Journal of Educational and Behavioral Statistics
Summer 1995, Vol. 20, No. 2, pp. 115–147

Inference and Hierarchical Modeling in the Social Sciences

David Draper
University of Bath, UK

Key words: *causal inference, education policy, inferential limitations, meta-analysis, multilevel models, school effectiveness*

Hierarchical models (HMs; Lindley & Smith, 1972) offer considerable promise to increase the level of realism in social science modeling, but the scope of what can be validly concluded with them is limited, and recent technical advances in allied fields may not yet have been put to best use in implementing them. In this article, I (a) examine 3 levels of inferential strength supported by typical social science data-gathering methods, and call for a greater degree of explicitness, when HMs and other models are applied, in identifying which level is appropriate; (b) reconsider the use of HMs in school effectiveness studies and meta-analysis from the perspective of causal inference; and (c) recommend the increased use of Gibbs sampling and other Markov-chain Monte Carlo (MCMC) methods in the application of HMs in the social sciences, so that comparisons between MCMC and better-established fitting methods—including full or restricted maximum likelihood estimation based on the EM algorithm, Fisher scoring, and iterative generalized least squares—may be more fully informed by empirical practice.

Much of the data gathered in the social sciences to answer scientific and decision-making questions has a *nested* or *hierarchical* character. Examples in fields as disparate as economics, education, and health policy come immediately to mind:

- the multistage cluster sampling employed by the U.S. government's main instrument for estimating local and national unemployment rates, the Current Population Survey (e.g., Bureau of the Census, 1978), in which random samples are taken at each of the state, area, and (city) block levels;

I am grateful to Leigh Burstein, Carol Fitz-Gibbon, David Freedman, Harvey Goldstein, Sander Greenland, David Lane, Dennis Lindley, and Michael Seltzer for helpful discussions, comments, and references. I owe particular thanks to Steve Raudenbush for a detailed and interesting critique, to which I have attempted to respond on all matters of fact but few matters of interpretation and emphasis, in the interest of a more vigorous discussion. Membership on this list does not imply agreement with the ideas expressed here, nor are these people responsible for any errors that may be present.

- the natural grouping of information relevant to the study of educational effectiveness (e.g., Bryk & Raudenbush, 1992) into variables gathered at the student, class, school, and district levels; and
- the measurement of quality of hospital care (e.g., Draper et al., 1990; Rogers et al., 1990) obtained from samples of patients chosen from each of a number of sampled hospitals, themselves perhaps drawn from a sample of geographic areas.

For decades, quantitative workers in the social sciences have taken advantage of the hierarchical character of such data at the design stage of their investigations (e.g., Deming, 1947), using the *multilevel* organization of the population of interest to guide the data gathering. One might have expected that anything playing such a central role in the design must also have been accurately represented in the analysis, but surprisingly, until about 10 years ago, hierarchical analyses that made effective use of the nested data structure were the exception rather than the rule in much social science research, and there are egregious examples of underestimated uncertainty assessments arising from a failure to account for the homogeneity within levels of the hierarchy exhibited by cluster samples (see, e.g., Kish, 1957, for a summary of this problem). Why did day-to-day empirical work lag behind the perception of correct practice for so long?

The reason was mainly the constraints of technique. Although standard analysis of variance methods dating back to the 1920s (e.g., Fisher, 1925; see Scheffé, 1956, for some of the history) have long provided partial answers to some of the questions posed by some kinds of data gathered in a fully nested manner, the general formulation of the hierarchical linear model was not given until the early 1970s (Lindley & Smith, 1972), and the fitting of such models in something approaching full generality proved elusive until the introduction of the EM algorithm (Dempster, Laird, & Rubin, 1977) later in that decade. Since then, a variety of alternative fitting methods have been developed—including full or restricted maximum likelihood based on Fisher scoring (Longford, 1987) or iterative generalized least squares (Goldstein, 1986), and Gibbs sampling and other Markov-chain Monte Carlo (MCMC) methods (e.g., Smith & Roberts, 1993), although this last approach has not yet caught on fully in multilevel modeling in some of the social sciences, including education—and the number of applications of hierarchical models (HMs) is burgeoning.

Multilevel Data Analysis Before HMs

Historically popular methods for analyzing multilevel data that preceded HMs include what de Leeuw (1992) calls *disaggregation* and *aggregation* techniques. In the former—for instance, in a study of student performance with a four-level nesting structure (schools, teachers, classes, and individuals)—one might attempt an ordinary least squares (OLS) regression in which

"teacher, class, and school characteristics are all assigned to the individual, and the analysis is done on the individual level." This is unsatisfactory because the implied covariance matrix of the performance outcome fails to capture the within-school, within-teacher, and within-class homogeneity one would expect the data to exhibit through positive intracluster (intraclass) correlations. In the latter—for instance, with the *means on means* or *ecological regression* approach (see, e.g., the discussion of Aitkin & Longford, 1986)—one attempts to avoid this problem by aggregating across units at the lower levels of the hierarchy and then building linear models for the aggregates. But this runs afoul of the aggregation bias problem, long familiar to econometricians (e.g., Judge et al., 1988) and other quantitative workers, in which aggregate relationships typically appear stronger than they would at the individual level, where predictions of the greatest relevance to policy must be made.

Advantages of HMs

Hierarchical models offer at least three clear advantages, both conceptual and technical, over these and other methods for the analysis of multilevel data in the social sciences:

(a) As noted by many authors (e.g., Goldstein, 1987; Burstein, Kim, & Delandshere, 1989), HMs provide a natural environment within which to express and compare theories about structural relationships among variables at each of the levels in the organizational or sampling hierarchy.

(b) In sharp contrast with standard regression methods applied to observations made with cluster sampling, the fitting of HMs yields better calibrated uncertainty assessments in the presence of positive intracluster correlations of a magnitude typical in social science data (e.g., Scott & Holt, 1982; Longford, in press).

(c) HMs offer an explicit framework in which to express similarity (*exchangeability*) judgments (e.g., Draper, Hodges, Mallows, & Pregibon, 1993), in order to combine information across units (such as students or schools) to produce accurate and well calibrated predictions of observable outcomes.

However, as with any other methodological advance, there are limits to what may be validly concluded on the basis of a hierarchical analysis—examples that overstep this boundary have already begun to appear—and there is always room for potential technical improvements. In this article, I (a) focus on issues of interpretation of multilevel analyses in education, arguing that the level of explicitness in the scope of inferential conclusions drawn from HMs needs to be raised; (b) examine the causal implications of the use of HMs in school effectiveness studies and meta-analyses; and (c) conclude with some remarks on the value of comparative study of the various fitting methods in current or potential use in hierarchical modeling, with an

emphasis on contrasting the methods that are currently most popular with Gibbs sampling and other MCMC methods.

Interpreting Hierarchical Analyses

As an example of where I am headed in this section, consider the HM analysis presented by Huttenlocher, Haight, Bryk, Seltzer, and Lyons (1991) on the effects of parental speech on early childhood vocabulary growth. The data for this analysis were gathered in the following way.

> Parents who were full-time caregivers were recruited through newspaper ads from a relatively educated middle-class, urban community. For all children [studied], this caregiver was the mother. There were two groups of parent-child pairs. Each group contained 11 children (6 boys and 5 girls). The groups varied somewhat in the conditions in which subjects were observed. Subjects in Group 1 were seen every 2nd month for 5 hr. Five children were 14 months [old] at the beginning of the study, and 6 children were 16 months; all children were 26 months at its conclusion. Subjects in Group 2 were seen every 4th month for 3 hr., beginning when children were 16 months and continuing until they were 24 months. . . . Children and their mothers were observed during children's typical daily activities. . . . The written transcript, including all utterances produced by the child and directed at the child, was completed later from an [audio or video] tape recording.

In their modeling work, the authors chose to define the exposure of the children to speech from their mothers by measuring the total number of words the mothers directed to their children in the 3-hour observation period at 16 months. One may visualize the data gathered in this way as a rectangular array with 22 rows and 10 columns: Y_{ti}, the vocabulary size for child i at t months ($t = 14, 16, \ldots, 26$, with missing values at four of these time points for Group 2); X_i, the exposure for child i; and dummy variables for group membership and gender. Preliminary data analysis indicated that the relationship between Y_{ti} and t was roughly quadratic, with near-zero coefficients for the constant and linear terms when time was measured forward from 1 year of age, and that it was useful to work with the exposure variable on the log scale.

Among other models the authors fit the growth-curve HM

$$\begin{aligned} Y_{ti} &= \pi_{2i} \cdot (t - 12)^2 + \epsilon_{ti} && \text{(within-subjects)}, \\ \pi_{2i} &= \beta_0 + \beta_1 \cdot \text{group}_i \\ &\quad + \beta_2 \cdot \log(X_i) + \beta_3 \cdot \text{gender}_i + U_i && \text{(between-subjects)}. \end{aligned} \tag{1}$$

Here "π_{2i} represents the acceleration in vocabulary growth for child i," ϵ_{ti} is the "deviation of child i from his/her growth trajectory at time t," U_i "represents a unique effect for child i on the acceleration parameter," and the ϵ_{ti}

and U_i are regarded as Gaussian random variables with mean 0 and variances σ^2 and τ, respectively. The authors report a variety of standard errors (*SE*s) and p values computed on the basis of this model—for instance, the estimated coefficient $\hat{\beta}_2$ for log (exposure) in the between-subjects level of model (1) comes out 0.89 with an *SE* of 0.36 ($p < .05$)—and speak in a rather general way about the lessons learned from these inferential results, for example, "In summary, our data strongly suggest that the number of word learning trials to which a child is exposed is an important factor in the acquisition of vocabulary items." But what meaning, if any, may be attached to such p values—and to the *SE*s the authors quote for the parameter estimates arising from the fitting of model (1) above—and what is the valid scope of inferences drawn from this model with these data?

The predictive approach to inference, and its interpretive advantages. In answering this question, it is useful to consider the perspective provided by the approach to inference, based on the prediction of observable quantities, advocated by de Finetti (e.g., 1974–1975) and developed by Lindley (e.g., 1972), Geisser (e.g., 1993), and others. Within this approach, the only inferential elements with objective reality are data values X you have already observed and data values Y you have not yet observed. Inference is then the process of quantifying your uncertainty about future observables Y on the basis of things you know, including the X values and the context in which they were obtained. Informally one might call X the data you have and Y the data you wish you had; with this terminology, a statistical model supporting your inference is a mathematical story that links these two data sets.

Parameters, such as τ and the π_{2i} in model (1) above, may come up in this story as placeholders for particular kinds of uncertainty on the way to prediction of observables, but in many cases (see, e.g., Lane, 1986) they have no objective reality of their own, and do not receive anywhere near as much emphasis as they get in other inferential approaches. By focusing on things that you can see rather than things you can't, this outlook has the advantage of readily available calibration information on the quality of your inferences: in educational research you can make predictions, with uncertainty assessments, for a number of students and schools, and compare their actual outcomes with what you thought they would be. If you miss by a lot more than you thought you would in a lot of these predictions, you are out of calibration, and need to revise your uncertainty assessments. This may be contrasted with, say, confidence statements about unobservable parameters—how do you know when they're right?

In practice in the social sciences, the data you have and the data you wish you had may differ from each other in three main ways:

(a) Problems of *measurement error* arise when you are trying to quantify something elusive, such as intelligence, and you are not sure if you got it right. The data you have then consist of one or more scales, say, that you hope measure the "underlying construct" of interest, and the data you wish

you had are the actual values of that construct, if only you knew how to measure it well. In this case, X can fall short of Y in at least two ways, identified by the concepts of *reliability* (X is an unbiased but possibly noisy estimate of Y) and *validity* (X may be biased for Y). This is an important topic, both rather generally in social science research (see, e.g., Freedman, 1983, for a sharp critique of such research based largely on this issue) and particularly in education, where Burstein (e.g., 1980) and others have expressed concern at the application of complex analytic methods to problems relying on measures of key constructs (such as the difficulty level of the material taught, in studies of student performance) of unclear validity.

(b) Problems of a *counterfactual* nature arise with experimental data when you are trying to quantify the effects of a particular cause, such as a new teaching method, and the outcome for each student in the experiment may be observed for only one setting of the supposedly causal factor, new versus standard method, say. Here, the data you have for the treatment (control) students is the outcome they exhibited under the new (standard) method, and the data you wish you had is the outcome the same students would have exhibited if, instead, they had been taught with the standard (new) method. This model dates back at least to Neyman (1923/1990)—and in the special, and almost certainly false, case in which the two outcomes for each person are identical (though differing from person to person), to Fisher (1935)—and has been extensively developed over the last 15 years or so by Rubin (1978), Holland (1986), and others. It provides a good example of the value of de Finetti's approach in clarifying what people mean when they talk about causal inference (see, e.g., Sobel, in press) and what must be assumed to support such inference.

(c) Problems of a *sampling* nature (e.g., Cochran, 1978) arise when there is a finite population of subjects of direct scientific or policy interest (such as all sixth-grade students enrolled in California public schools in the fall of 1993, or all eighth-grade math teachers in New York as of March 1, 1994) and, typically for reasons of cost, you are able to obtain data only on a subset of the population. In this case, the data you have is the information on the sampled individuals and the data you wish you had is the corresponding information on the unsampled people.

All three of these ways in which X and Y differ may, in turn, be thought of as special cases of the general *missing data* problem, addressable—at least, in principle—by imputation methods (e.g., Rubin, 1987; Little & Rubin, 1987). See Little and Schenker (in press) for a recent review of this approach.

A Taxonomy of Inferential Strength in Statistical Modeling

When inferential examples, in the social sciences in general and in educational research in particular, are examined from the predictive point of view, four kinds of inference—of varying strength and scope of generalizability—are discernible, which may be termed *calibration inference, specific causal*

inference, general causal inference, and *sampling inference.* In the remainder of this section, I examine the implications of this taxonomy for hierarchical modeling, although the discussion applies rather generally to the use of stochastic models in the social sciences.

Calibration Inference

The lowest level of inferential strength and generalizability of results arises when attempts are made to model data that are neither experimental in character nor sampled, from the population P of most direct scientific or decision-making relevance to the question at hand, in a way that supports straightforward exchangeability assumptions about how the sampled and unsampled units are likely to be similar and how they are likely to differ. Freedman, Pisani, Purves, & Adhikari (1991) call such data *samples of convenience*; a slightly less pejorative name for them might be *uncertain exchangeability (UE) samples.*

The Huttenlocher et al. (1991) example above would seem to fall into this category. From various conclusions of unqualified scope drawn in that paper (e.g., "The present study provides the first direct evidence that amount of exposure is important to vocabulary growth"), the authors appear to have quite a broad population in mind, and yet the data consist of 22 mother-child pairs from families living in a single "relatively educated middle-class urban community" (which I will refer to here as Chicago for discussion purposes) who responded to newspaper ads. Also, several statements headed in the direction of causal inference are made, for example,

> To evaluate the substantive significance of the relation between exposure and acceleration ($\hat{\beta}_2 = .89$), we note that the raw frequency of mothers' speech in our sample ranges from approximately 700 to 7000 words, . . . a difference of 2.30 units in the log metric. Controlling for differences in gender and group, π_{2i} is expected to be $2.30 \cdot 0.89 = 2.09$ units larger for a child whose mother speaks 7000 words than a mother who speaks only 700 words. . . . This translates into a difference in child vocabulary of $2.05 \cdot (16 - 12)^2 = 33$ words at 16 months, 131 words at 20 months, and 295 words at 24 months.

However, the data are purely observational in character, and little or no information on potential confounding factors (hereafter PCFs)—such as the amount of nonverbal communication between the children and mothers—is available to permit any adjustment for the effects of these variables.

Some (e.g., Freedman et al., 1991) would say that no inferential conclusions are possible with UE samples—and from the predictive viewpoint it does seem difficult to identify the as yet unobserved data values to which Huttenlocher et al.'s (1991) analysis refers—but there is a limited form of inference that I nevertheless find both justifiable and somewhat useful in this case. The point has been made (see, e.g., Kahneman, Slovic, & Tversky, 1982; Diaconis, 1985) that people are quite good at identifying interesting patterns in data—so

good, in fact, that they are capable of finding them even when they are not really there to be found, in the sense that the apparent pattern would fail to materialize in attempts to validate the results of the data-gathering activity by repeating it. It is arguable that we need some form of *calibration* inference to restrain our enthusiasm in the search for scientific relationships. Indeed, this was the original motivation for significance tests, and—given that estimates of quantities of direct scientific or policy relevance, together with uncertainty assessments for those estimates, are typically much more meaningful than p values—it is essentially the only worthwhile use of such tests (see, e.g., Oakes, 1990, for a thorough account of the misuse of significance tests in the social sciences).

Two forms of calibration inference in routine use are procedures based on *hypothetical sampling models* and *permutation tests*, both due to Fisher (1925, 1935). In hypothetical sampling inference

> the [data] values (or sets of values) before us are interpreted as a random sample [from] a hypothetical infinite population of such values as may have arisen in the same circumstances. The distribution of this population will be capable of some kind of mathematical specification, involving a certain number, usually few, of parameters. (Fisher, 1925)

You then proceed to work out a sampling distribution for estimates of those parameters, in effect by calculating all possible values the parameter estimates could take on across hypothetical replications of the sampling experiment that produced your data; the standard deviation of this distribution is the *SE* Fisher would have you quote as a measure of your uncertainty about the values of the hypothetical population parameters, and this *SE* becomes the denominator in z-ratios that lead to p values for tests of null hypotheses about those parameters.

The trouble with this formulation applied to UE samples is that the hypothetical population corresponding to the observed sample and the population P of real interest will often not be the same. Survey sampling specialists (e.g., Cochran, 1978) call the former the *sampled* population and the latter the *target* population, and emphasize that differences between them lead to invalid inferences about the target population. I do not find standard errors computed from UE samples meaningful, and I do not see that the parameters in a model such as (1) above have any meaning when such models are applied to UE samples, except as technical intermediaries that aid in the prediction of future observables (e.g., vocabulary sizes at ages beyond those observed for the children in Huttenlocher et al.'s [1991] data set, an activity those authors do not emphasize). The point is that Huttenlocher et al. did not write their article as if they were interested only in what would happen if you repeatedly ran newspaper ads in Chicago and recruited 22 mother-child pairs each time, but—without an argument supporting exchangeability of the people in their study with other people, as yet unnamed—that is the only population to which their parameter estimates and *SE*s are of direct inferential relevance.

In permutation inference, which arises most naturally when comparing two groups on a single outcome, you condition on the observed data and consider all possible ways in which the observations might be rearranged among the two groups, computing a summary such as the difference in group means for each possible permutation; a p value may then be calculated by asking how often differences as large as the one observed or larger would occur. When he introduced this procedure, Fisher (1935) had in mind experimental situations involving random assignment to the two groups, but (e.g., Freedman & Lane, 1983) the calculation may be performed no matter how the data in the groups were obtained, and even when no causal or sampling inference is justifiable, the resulting p value does seem to have some calibrative value. The idea is that life is short, there is not enough time to investigate all the interesting-looking relationships, and so perhaps we should focus on the ones that seem likely to show up again if we go out and get more data.

With moderate to large samples, p values based on comparisons of means will tend to be similar with the permutation and hypothetical sampling approaches, essentially by virtue of the Central Limit Theorem (e.g., Welch, 1937), so that a somewhat roundabout justification of the normal-theory p values calculated by Huttenlocher et al. (1991) may be attempted: there is probably some sort of calibration-style permutation test that their normal-theory tests are trying to approximate. In the absence of better sampling or causal motivation for their data, however, I find no scientific meaning in the parameter estimates and *SE*s Huttenlocher et al. report, and the strongest interpretation I am able to make of their p values is calibrative. It is clear that we have learned something about child development for people outside Chicago, but (see, e.g., Holland, 1989) without judgmentally estimating what might be called a variance component for nonexchangeability between sampled and unsampled units in this broader population, and a variance component for the effects of unmeasured PCFs—quantities that are unaddressed by Huttenlocher et al.'s data—it is difficult to quantify just what has been learned more broadly, either causally or externally to the 22 mother-child pairs in their study.

For other recent examples of what appear to be inferential hierarchical analyses of UE samples, see Bryk and Frank (1991), Bryk and Raudenbush (1987; 1989, sections 4.2 and 5.1; 1992, chapters 4, 5, and 8), Fitz-Gibbon (1991), Goldstein (1987, sections 2.3 and 4.4; 1989), and Raudenbush and Bryk (1989, section 5.3). It is possible that stronger conclusions are supported by some of these studies, but so little space is devoted in them to the origins of the data analyzed and the valid scope of their findings that it is hard to tell.

Some of the papers and books mentioned in this section are methodological in character, and in such work people sometimes sidestep the question of what kind of inference they are making by calling the modeling "illustrative," but (e.g., Draper, 1987) if you use UE samples in your examples and draw what look like substantive conclusions, you risk misleading your audience

about the scope of the "illustrative" findings, and in any case an unfortunate precedent is set for the substantive papers that will later apply the methodology.

Few discussions of the relationship between the sampled and target populations are to be found in the recent use of HMs in educational research; an exception is Longford (1991b), who gives a critique of the difficulties encountered in trying to make the two populations coincide in pretest studies.

Causal Inference

A higher level in the ladder of inferential strength arises when investigators interested in the effects of particular causes—such as a novel teaching method or a new way to allocate educational funding—gather data on some subjects, using either a controlled experiment or a well-conducted observational study design, and build a causal inference model. An example is provided by the Cognitive Strategies in Writing Project (Englert et al., 1988), examined by Bryk and Raudenbush (1992). This study

> sought to improve childrens' writing and to enhance their self-perceptions of academic competence through a variety of strategies. The outcome variable was a measure of perceived academic self-competence (mean 2.92, *SD* 0.58) for which a pretest, denoted X_{ij}, served as the covariate. The study involved 256 children in 22 classrooms in a standard two-group design, with 15 experimental and 7 control classrooms. Because teachers administered the treatments to intact classrooms, we have, in classical terms, a nested or hierarchical design: students are nested within classrooms and classrooms are nested within two treatment groups. (Bryk & Raudenbush, 1992, p. 96)

To these data Bryk and Raudenbush fit the HM

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{..}) + r_{ij} && \text{(Level 1)}, \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} W_j + u_{0j}, \qquad \beta_{1j} = \gamma_{10} && \text{(Level 2)}, \end{aligned} \tag{2}$$

in which Y_{ij} is the self-perceived competence of child i in classroom j at the end of the experiment; W_j is a classroom-level dummy variable for experimental/control status; $\bar{X}_{..}$ is the pretest grand mean, which came out 2.86 here; and the r_{ij} and u_{0j} are "errors," regarded as Gaussian random variables with mean 0 and variances σ^2 and τ_{00}, respectively. The estimate of the treatment effect in this model comes out $\hat{\gamma}_{01} = .19$, with an *SE* of .10,[1] which is statistically significant at the .05 level if you do a one-tailed test (although probably not practically significant: the difference between adjusted experimental and control means is only about 6% of the control mean, and is less than a third of the overall between-child *SD*). Causal conclusions are definitely in the air here (" ... [the] experimental children developed a significantly higher perceived self-competence than did the control children"), but causal for whom, and under what implicit assumption(s) not yet articulated?

The counterfactual framework mentioned in the previous subsection is clarifying. Table 1 gives a visualization of the data from this perspective, with plausible numerical values for the observed outcomes and pretest scores, and question marks for the counterfactual outcomes. Actually, there are two counterfactual stories operating here: in addition to wondering about what the experimental (control) students' outcomes would have been if they had been in the control (experimental) group, the presence of the pretest X in model (2) means that Bryk and Raudenbush would also like to know what the observed experimental and control means on the outcome Y would have been if the mean X values in the two groups had been the same instead of differing slightly ($\bar{X}_E = 2.90$, $\bar{X}_C = 2.80$).

In this case, the HM estimates from model (2) fill in the question marks—or at least provide guesses for their averages within the experimental and control groups, if these groups had been the same on average on the pretest—by computing the adjusted means $\bar{Y}_E - \hat{\gamma}_{01}(\bar{X}_E - \bar{X}..) = 2.96$ and $\bar{Y}_C - \hat{\gamma}_{01}(\bar{X}_C - \bar{X}..) = 2.78$ as in the usual analysis of covariance. The only essential difference, in fact, from the usual ANCOVA is that the presence of the "random" classroom effects u_{0j} in (2) above accounts sensibly for the within-class homogeneity exhibited by the students on the academic competence outcome (the conditional intracluster correlation for Y given X here is about $\hat{\rho} = .08$). However, the adjusted means are good estimates of the averages across the question marks in the two groups only if the groups are similar on variables likely to be strongly correlated with Y (after adjusting for X)—that is, if they are similar on PCFs, such as the amount of encouragement the students got at home while the experiment was in progress. Why should this be so?

One reason would be random treatment allocation (or, at least, randomization would tend to promote balance on the PCFs), but Bryk and Raudenbush fail to mention whether the experimental and control classrooms were assigned at random. In the absence of random assignment, the use of $\hat{\gamma}_{01}$ as an estimate, for the children in Table 1, of the effect on self-perceived academic competence caused by the intervention rests on an implicit assumption of what Rosenbaum and Rubin (1983) call *strong ignorability* of the

TABLE 1
Display of the Cognitive Strategies in Writing Project data in factual-counterfactual form

Class	1	...	1	15	...	15	16	...	16	22	...	22
Child	1	...	11	1	...	14	1	...	9	1	...	13
Group	E	...	E	E	...	E	C	...	C	C	...	C
Y if E	2.6	...	3.0	2.8	...	3.6	?	...	?	?	...	?
Y if C	?	...	?	?	...	?	3.0	...	2.9	1.9	...	3.1
X	2.5	...	2.4	3.5	...	3.3	2.7	...	3.0	2.1	...	2.5

treatment allocation mechanism conditional on the pretest X: the assumption (for the students in the table) that Y and experimental status are independent given X. If you do not believe this assumption—for instance, if you have a PCF, not yet in model (2), in mind that is likely to be influential after X is accounted for—then there is no reason for you to regard the issue of the effect caused by the intervention for the students in Table 1 as settled. Model (2) has correctly inflated the SE of $\hat{\gamma}_{01}$, to account for within-classroom homogeneity, in relation to the value it would have had assuming independence of all 256 student outcomes, but it has modified neither $\hat{\gamma}_{01}$ nor its SE to account for uncertainties in the validity of causal inferences based on the raw group means adjusted only for X.

Specific versus general causal inference. Moreover, even if either treatment assignment were random in this experiment or you were willing to assume strong ignorability given the pretest, the topic of scope of inferences based on $\hat{\gamma}_{01}$ remains to be addressed. Under random assignment or strong ignorability, model (2) shows that something causal is going on for the 256 students in Table 1 (although the effect is not very big), but what about other students? From a policy point of view, it is nice to know that there exists a group of students for whom the Cognitive Strategies in Writing Project intervention makes a modest difference, but without some effort to relate these 256 children to the broader collection of other students to whom the intervention might be offered, it is not clear how much has been learned about whether the program should be tried elsewhere.

The distinction between specific causal inference (there is something causal going on for the people in this experiment) and general causal inference (there is something causal going on for everybody in the target population) is particularly forceful in medical research, where—to judge from articles in leading journals over the past 10 years—it is nevertheless routinely ignored. A typical recent example is a study by Ahmed, Garrigo, and Danta (1993), in which 12 patients who presented at Mount Sinai Medical Center in Miami Beach with exercise-induced asthma during the last half of 1992 (and who, in addition, were selected on the basis of a number of intake criteria) were randomized into three groups of size 4; one group received heparin, another cromolyn sodium, and the third a placebo. A statistically significant difference between the groups was found, in favor of heparin, and both the abstract ("Inhaled heparin prevents exercise-induced asthma without influencing histamine-induced bronchoconstriction") and the discussion section of the article read as if there is no longer any uncertainty about how everybody with this ailment should be treated.

This pitfall clearly also arises in educational research, and the use of HMs—whether the studies giving rise to the data to which the HMs are fit are randomized or not—does not avoid it. In particular, making the classroom effects random rather than fixed in model (2) does not convert a UE sample of classrooms into a random sample from the target population of real policy

interest. A parallel comment applies to the use of random variables in Huttenlocher et al.'s (1991) model (1), above: Regarding the "errors" at both levels of model (1) as realizations of random variables does not make the 22 mother-child pairs any more or less "randomly" drawn from the target population than they would have been if you had treated the lack-of-fit terms in model (1) in some other way.

For other recent examples of specific causal inference with HMs in observational studies ("quasi-experiments") involving educational interventions, see Raffe (1991) and Jacobsen (1991).

Sampling Inference

The fourth entry in the inferential hierarchy arises when sufficient good fortune and money are available to sample representatively from the target population and build a sampling model. There are two cases to distinguish, according to whether or not causal inferences are desired instead of (or in addition to) sampling inferences; in the sampling world, the distinction is roughly that between *descriptive* and *analytical* surveys (Cochran, 1978). A recent example involving HMs that had the potential to include both kinds of inference is the study conducted by Lockheed and Longford (1991), who examined school- and student-level factors associated with successful mathematics achievement in Thailand. The data, from the IEA Second International Mathematics Study, were gathered with a two-stage sampling design employing stratification at the first stage (the primary sampling units were national education regions) and clustering at the second (the cluster units were schools, chosen randomly to produce a 1% sample of all eighth-grade mathematics classrooms in each region). One class was selected at random per school, and all students enrolled for the entire academic year in the chosen class became part of the sample, resulting in a data set containing 32 variables—13 at the student level, 5 pertaining to school characteristics, 4 to the teacher, 9 to the classroom, and 1 to the region—measured on a total of 4,030 students in 99 schools. The choice of one class from each school confounded the school, teacher, and classroom levels in the design, so that Lockheed and Longford simply referred to measures at those levels as "group" variables.

Sampling inferences, such as estimation of the average number of students per math class in Thailand (about 44) or the proportion of students in the country with access to a calculator at home (roughly 31%), were not the main emphasis in Lockheed and Longford's study, but could have been based on HMs that reflected the sampling design, with fixed effects for the regions and random effects for the classes and students to accurately estimate the relevant intracluster correlations. In the analytic part of their work, Lockheed and Longford fit a variety of multilevel models with fixed and random coefficients to find characteristics at the student and group levels that were associated with student achievement in mathematics, noting (for instance)

that large school size was positively associated with high achievement, and high levels of teacher time spent on maintaining order in the classroom were negatively associated with math learning. Lockheed and Longford concluded with the important point that even though their sample was representative of eighth-grade mathematics education in Thailand, the process by which students were assigned to treatment groups defined by factors such as school enrollment and teachers' disciplinary styles was observational, so that it would be rash to regard these associations as causal.

> Identifying the variables associated with higher outcome scores [in an observational study] does not offer a direct answer to the principal question of a development agency about the distribution of its resources to a set, or continuum, of intervention policies in an educational system. Without any prior knowledge of [that system], any justification for an intervention policy based [only] on the results of regression (or variance component) analysis, or even of structural modeling (LISREL), has no proper foundation.

Additional recent examples of HMs based on what appear to be representative samples from policy-relevant target populations in education include Bryk and Raudenbush (1989, section 3.2), Lee and Smith (1991), Paterson (1991), Raudenbush and Bryk (1989, section 4.2), and Zuzovsky and Aitkin (1991). I have been unable to find any instances of general causal inference—in which a representative sample from the target population is enrolled in a randomized controlled experiment—in the recent education literature on the use of HMs.

The Value of Explicitness in Inferential Conclusions

Table 2 summarizes this section by displaying the four kinds of inference in the taxonomy above in a two-by-two array, with rows defined by the

TABLE 2
Types of inference supported by various sampling and design assumptions

		Strong Ignorability of Treatment Assignment	
		Difficult to Justify	Justifiable
Exchangeability of Sampled and Unsampled Units in Target Population	Difficult to Justify	Calibration Inference	Specific Causal Inference
	Justifiable	Sampling Inference	General Causal Inference

sampling plan and columns indexing the experimental design. In the language of evaluation research in, say, psychology (e.g., Campbell & Stanley, 1966), the rows of this table correspond to different levels of *external validity* and the columns to varying levels of *internal validity.* There is a partial ordering, in inferential strength, of the cells of the table of the form

$$\begin{matrix}\text{calibration}\\ \text{inference}\end{matrix} < \left(\begin{matrix}\text{specific}\\ \text{causal}\\ \text{inference}\end{matrix} , \begin{matrix}\text{sampling}\\ \text{inference}\end{matrix} \right) < \begin{matrix}\text{general causal}\\ \text{inference}\end{matrix} . \qquad (3)$$

Random assignment of units to the groups (sampled, unsampled) and (treatment, control) is sufficient, but not necessary, to land you in the "justifiable" row and column in Table 2, respectively; well-designed nonrandom sampling plans and observational studies can also achieve these distinctions, at least judgmentally, by measuring and appropriately adjusting for all relevant PCFs.

I have considered the taxonomy of inferential strength in this table by way of arguing that an increase in clarity about the scope of valid inferential conclusions supported by complicated analyses, including those based on HMs, would be a net gain, not just in education but quite generally in the social sciences. What I have in mind, quite literally, would be for people to start explicitly saying in their papers—ideally both in the abstract and in the body of the article, including the discussion section—which kind of inference they are trying to do. In addition to improved communication of exactly what has been learned in a given study and what remains to be discovered by future studies, this greater explicitness might actually increase the rate at which uncertainty about the population-wide effects of social policy interventions declines, by promoting a larger funding emphasis on the sorts of experimental and sampling designs that are most effective in decreasing such uncertainties: controlled experiments and randomized sampling plans, with (a) stratification on known PCFs at the top of the design, (b) clustering as needed for reasons of cost, and to study the effects of the interventions at different levels of the organizational hierarchy into which the intervention must fit, and (c) randomization at the bottom of the design to balance the unknown PCFs.

There is nothing revolutionary about strong designs of this type, and their greater expense over the empirically more prevalent class of observational studies and UE samples is an obstacle to be reckoned with in the social sciences, but other fields in which both data-gathering strategies have been tried—such as medicine (see, e.g., Freedman et al., 1991)—have amply demonstrated that the average number of retrospectively valid causal conclusions per monetary unit is higher with the costlier designs. Ethical considerations sometimes preclude the use of controlled experiments, as with studies of the causal link between smoking and adverse health outcomes in humans,

but successes in areas as different as health policy (e.g., Brook et al., 1983) and criminal justice (e.g., Ares, Rankin, & Sturz, 1963) have shown that the field for social experimentation is wide, and in any case there is nothing unethical about conducting a fully representative sampling study. An increased degree of candidness about the inferential limitations of observational studies and UE samples could shift the funding balance in the social sciences toward stronger designs, just as it did in medicine decades ago.

The Use of HMs in School Effectiveness Studies

The last 10 years have seen an increase in the attention paid to the quality with which public and commercial institutions carry out their mandates (e.g., Box, 1994). In education this has taken the form of increased interest in measuring the effectiveness of schools and teachers at promoting learning. One result in Great Britain, for example, has been a call for the publishing of "league tables," ranking schools in each area of the country by the achievements of their students on standardized tests at the end of each year. As mentioned by Goldstein (1992) and others, the British government initially proposed to do this without adjusting in any way for the achievement levels of the students upon entry to the schools, but more recently (e.g., "League Table," 1994) a greater willingness to measure the "value added" by each school, through a comparison of input and output achievement levels, has emerged. A number of authors have noted that HMs can play an important part in analyses of this type. My main points in this section are that the value added by HMs in school-effectiveness studies would be increased by a stronger attempt to tell an explicit causal story about the outputs of the HM analyses that drive policy choices, and that cohort standardization and other biostatistical concepts already in use in hospital effectiveness studies can help in this attempt.

The *Guardian* value-added survey ("From the Raw to the Refined," 1993; Goldstein & Thomas, 1993) provides a good example within which to examine the connections between HMs and causal inference. In early 1993, Goldstein and Thomas requested the participation of almost all secondary-education institutions in England, Wales, and Northern Ireland in a major school effectiveness study, but were able to achieve a school-level response rate of only about 15%. This produced a sample of 29,985 students in 425 institutions, from which data of two kinds were gathered: a baseline score X_{ij} created by aggregating a set of standardized tests for student i in institution j, and a later achievement score Y_{ij} for the same student, obtained in a similar manner from a different set of tests. The simplest HM that captures the flavor of their analyses has the form

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 X_{ij} + \gamma_j + e_{ij} \qquad \text{(Level 1)}, \\ \gamma_j &= u_j \qquad\qquad\qquad\qquad\quad \text{(Level 2)}, \end{aligned} \tag{4}$$

in which the u_j and e_{ij} are regarded as realizations of Gaussian random variables and the second level of the HM is like a regression with no school-level predictors. Goldstein and Thomas use this model to produce shrinkage estimates of the *school effects* γ_j, standard errors for these estimates, and a series of plots of intervals of the form $\hat{\gamma}_j \pm c \cdot \hat{SE}(\hat{\gamma}_j)$, created to facilitate school rankings (with the multiplier c chosen so that "for all possible pairs of comparisons, the average significance level is 5%"). In this model, the $\hat{\gamma}_j$ take the form of a weighted average of the usual ANCOVA school-level residuals $[\bar{Y}_{\cdot j} - (\hat{\beta}_0 + \hat{\beta}_1 \bar{X}_{\cdot j})]$ and the grand mean of those residuals (zero), with weights $(1 - \hat{B}_j)$ and $\hat{B}_j$, respectively, where $\hat{B}_j$ is the usual random-effects shrinkage factor for school j.

Despite concerns about nonexchangeability of Goldstein and Thomas's schools with the rest of the policy-relevant population (which would, among other things, have implications for the quality of inferences about β_1, and therefore about the process of adjustment for baseline scores), it is unquestionable that this methodology represents a noticeable improvement over both the league tables rankings previously released by the British government (based only on the raw school means $\bar{Y}_{\cdot j}$, which, on average, were pulled about 75% of the way back toward the grand mean by model (4) in Goldstein and Thomas's analysis, meaning that the $\bar{Y}_{\cdot j}$ provide a quite unstable set of predictions of what would be expected at the sampled schools next year) and results obtainable by treating the school effects as fixed. But what estimated causal effects, if any, do the $\hat{\gamma}_j$ actually correspond to?

This situation is exactly analogous to the health policy problem of trying to measure quality of care at the hospital level by comparing a hospital's observed mortality rate with what you would have expected given how sick its patients were when they arrived (e.g., Daley et al., 1988; Longford, 1991a; Thomas, Longford, & Rolph, 1992), except that the outcome variable in the health policy problem is dichotomous rather than continuous. In the case of hospital mortality rates (e.g., Draper, 1994), it has proven useful to make a link between analyses like that of Goldstein and Thomas and standardization methods motivated by counterfactual considerations, and the analogy is so strong that it seems worthwhile to do so in school effectiveness studies, as well.

From the point of view of experimental design, the process by which students wind up in particular schools in Goldstein and Thomas's data is observational, with achievement score Y as the quantitative outcome, school S (qualitative, at 425 levels) as the supposedly causal factor (SCF), and baseline score X as a quantitative PCF. To obtain a better estimate of the causal effect of S on Y in the presence of confounding from X than that based on the raw school means $\bar{Y}_{\cdot j}$, you have to compare the factual data set (Y; S, X) with your estimate of the counterfactual data set obtained by holding one of (S, X) constant and changing the other. Here this amounts to guessing at one or the other of the following counterfactuals:

- *standardization to the national cohort* (hold S constant, change X): How would this school have done if its group of students had been different (e.g., average) at entry instead of whatever they were?
- *standardization to the school cohort* (hold X constant, change S): How did the rest of the country do on average with students like those at this school?

Because of nonlinearities in a given school's effect on learning (e.g., on a common 100-point pre/post scale, a particular school may take students scoring 20 on intake and bring them up on average to 30, but may only raise children starting at 60 to an average of 65), the answers to these two questions may not lead to the same school assessment conclusions.

Given the shrinkage character of the $\hat{\gamma}_j$, it is not at all clear which, if either, of these two counterfactual stories is, in effect, estimated by model (4). Sorting this out would be a net gain, because the various players involved in the school assessment drama have different utility functions corresponding to different counterfactuals. The British government, for example, presumably wants to know which are the best and worst schools, to use the former as case studies and to figure out how to improve the latter; for this purpose, an analysis that standardizes to the national cohort would probably be easiest, because it directly holds the PCF of differential student ability at intake constant in the ranking it produces. Many individual schools, on the other hand, do not expect their intake cohorts to differ much from year to year; for each of them, standardization to their own cohort is probably of most direct relevance. A particular family, to take a third example, will want to know what differences they might expect to see if they were to send their child to each of the quite small number of schools that are feasible for them on geographical and/or cost grounds. For this purpose, standardization to their child (and others exchangeable with him/her) would be the ideal.

In closing this section, it is probably worth emphasizing the limitations of HM (or any other) analyses of data like those collected by Goldstein and Thomas, in order that school effectiveness studies not be oversold. First, the observational nature of the data and the small yearly school-level sample sizes will sharply limit subanalyses—for example, of the kind just mentioned as ideal for parent-level decision making—to broad exchangeability classes. Goldstein and Thomas, for example, produced separate results for each of the three groups (low, middle, high) defined by the first and third quartiles of the X distribution, and this may be about as fine as you can cut it.

Second, even after careful modeling, school-level rankings based on models such as (4) carry large uncertainty bands, about which lists of point estimates like the original raw British government league tables are negligently silent. When distributions across the categories (definitely better than, definitely worse than, uncertain) are computed for the schools in Goldstein and Thomas's sample—by making pairwise comparisons between each school and all the

other schools—and these distributions are averaged, the results are approximately (15%, 15%, 70%), meaning that the bands are too wide to locate most schools more accurately than somewhere in the middle of a large gray area.

As noted by Goldstein (1991), value-added rankings at their best should probably be used only diagnostically, as warning flags indicating where to focus attention in seeking causal explanations, which is like the best uses of hospital mortality rates so far.

> Multilevel modeling is not a panacea. Its power is limited, and it is most certainly not a magic wand that will allow us automatically to make definitive pronouncements about differences between individual schools.

Hierarchical Models and Meta-Analysis

HMs and meta-analysis (e.g., Glass, McGaw, & Smith, 1981) arrived on the social science scene at about the same time, and no wonder: the former is such a natural technical tool for implementing the latter that meta-analysis may be said to have been waiting for HMs to come along. Quantitative research synthesis is now an integral part of disciplines as varied as education and medicine, with the use of hierarchical models to capture between-study variability commonplace. In this section, I examine an interpretive point, arising in the use of HMs for this purpose, which has normative implications for allocation of research effort and resources. A separate technical point will come up in the section below on fitting HMs.

Consider the meta-analysis presented by Goodman (1989) and reexamined by Draper, Gaver, et al. (1993) on data from six controlled clinical trials of the effect of aspirin on mortality for patients who had survived a heart attack. Table 3 gives the data from the six studies; it may be seen that the first five trials were in good agreement with each other and with the view that aspirin causes a decline in mortality of about 2.3 percentage points (a 20% drop from 11.5%, the composite placebo mortality for the first five trials), but it is also clear that the sixth and largest trial—the Aspirin Myocardial Infarction

TABLE 3
Number of patients and mortality rate from all causes, for 6 trials comparing the use of aspirin and placebo by patients following a heart attack

	Aspirin		Placebo		Comparison	
Study	No. of Patients	Mortality Rate (%)	No. of Patients	Mortality Rate (%)	Diff (%)	*SE* of Diff (%)
UK-1	615	8.0	624	10.7	−2.8	1.7
CDPA	758	5.8	771	8.3	−2.5	1.3
GAMS	317	8.5	309	10.4	−1.8	2.3
UK-2	832	12.3	850	14.8	−2.6	1.7
PARIS	810	10.5	406	12.8	−2.3	2.0
AMIS	2267	10.9	2257	9.7	+1.2	0.9

Study (AMIS)—was in sharp disagreement with the other experiments. Three natural questions arise: (a) Why did AMIS get such different results? (b) If the answer to (a) is uncertain, what should be done next to reduce this uncertainty? (c) What therapy should be recommended to heart attack patients while we are waiting for the answer provided by (b)?

To move toward answers to these questions, Goodman used standard empirical Bayes methods to fit a Gaussian random-effects two-level model,

$$\begin{aligned} (y_i \mid \theta_i) &\overset{\text{indep}}{\sim} N(\theta_i, V_i) \qquad \text{(Level 1)}, \\ (\theta_i \mid \mu, \tau^2) &\overset{\text{IID}}{\sim} N(\mu, \tau^2) \qquad \text{(Level 2)}, \end{aligned} \tag{5}$$

in which y_i is the mortality difference in study i and the V_i (the squared *SE*s in Table 3) are assumed known. Level 1 of this model is not hard to justify, at least approximately, on large-sample causal inference grounds; Level 2 embodies a prior judgment of exchangeability of the effects of aspirin on mortality across the patient cohorts and treatment protocols in the six studies.

The maximum likelihood estimate of τ comes out about 1.5% here, leading to noticeably wider 95% central interval estimates for the "true effect" μ than those obtained from the usual fixed-effects model that assumes $\tau = 0$ ($(-3.2, +0.2)$ versus $(-2.1, +0.2)$). Adding Level 2 to the hierarchy has definitely improved the fit,[2] but because there are no study-level predictors in Goodman's model, there has been no increase in causal understanding in moving from a fixed-effects to a random-effects formulation, which means that model (5) is, at best, only part of the story.

Epidemiologists (e.g., Greenland, 1993) have begun to note that, although the use of (5) is certainly better than pretending that τ is 0 in the presence of substantial between-study heterogeneity, random-effects models that simply describe the heterogeneity rather than attempt to explain it can actually promote an antiscientific attitude of indifference to the cause of the study-level discrepancies. Indeed, a caricature of (acceptable statistics, unacceptable science) involves {tossing the heterogeneity into the Level 2 error term, verifying that the unexplained bit does look independent and identically distributed (IID) according to some standard distribution, and going on to the next problem feeling like the job was well done}, rather than, say, carefully reading the articles documenting the studies—and perhaps also interviewing the principal investigators—to identify how the protocols and patients were different (Greenland claims that useful information of this type is almost always fairly straightforward to obtain) and to then include these differences as Level 2 predictors. Model (5) provides a better answer to the short-run third question posed above than that offered by the fixed-effects formulation, but it is silent on the long-run (and ultimately more important) first and

second questions. A full answer to all three questions involves the use of model (5) as a kind of placeholder on the way to full causal understanding, while the answers to the first and second questions are sought. Note that this perspective implies a different use of research time and money than that employed to answer the third question alone.

For a social science example of a much more satisfying meta-analysis—in which (a) model (5) is fit to 19 experiments estimating teacher expectancy effects on students' observed IQ scores and substantial between-study heterogeneity is noted, and (b) much of that heterogeneity is shown to disappear when the number of weeks of student-teacher interaction prior to the experiment is accounted for—see Bryk and Raudenbush (1992, chapter 7).

Fitting HMs in Education: Why So Little Comparative Study and MCMC?

Turning now to the fitting of HMs, a Babel of options has arisen in the last 7 or 8 years, creating an embarrassment of apparent riches for the potential user and raising questions of choice, many of them still lacking fully satisfying answers. The main options are all one form or another of full or restricted maximum likelihood (FML, REML), including the EM algorithm, as implemented in Bryk, Raudenbush, Seltzer, and Congdon's (1988) program HLM, and in Wong and Mason's (1989) REML program GENMOD; iterative generalized least squares (IGLS), as carried out by Goldstein's (1987) REML programs ML2 and ML3; and Fisher scoring, as implemented in Longford's (1987) FML program VARCL. These programs can easily produce somewhat different answers on the same data set.

There are two curiosities in how the subject of fitting HMs has evolved so far. First, although an excellent study of GENMOD, HLM, ML2, and VARCL—documenting their design philosophies, implementation details, underlying models, software routines, data formats, user friendliness, and idiosyncrasies, and comparing their results on several data sets—has been available for the last few years (Kreft, de Leeuw, & Kim, 1990), the definitive investigation, probably involving extensive simulation, in which the four approaches and implementations are compared to known truth, has yet to see the light of day. A recent simulation study by Rodríguez & Goldman (1993) of the use of ML3 and VARCL to fit random-effects logistic regression models (which contains some disquieting findings[3]) is a good beginning, but considerable comparative territory of great relevance to applied practice remains unexplored.

Second, during the same period in which general-purpose implementations of FML and REML have arrived and have begun to receive widespread use in practice, a leading alternative to these methods—a fully Bayesian treatment of hierarchical models, involving Gibbs sampling or other MCMC techniques—has been under intense development in the statistics community and

yet has received little attention in many of the social sciences, particularly education.

The first of these curiosities should take care of itself over the next few years as investigators like Rodríguez and Goldman add more pieces to the comparative puzzle. The second curiosity is more interesting for what it reveals about the present state of the old Bayesian/frequentist dichotomy, as follows.

Using Seltzer, Wong, and Bryk (1993) and Seltzer (1993) as a starting point, consider the Gaussian two-level HM

$$\begin{aligned} \mathbf{y}_j &= \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{e}_j \qquad \text{(Level 1)}, \\ \boldsymbol{\beta}_j &= \mathbf{W}_j\boldsymbol{\gamma} + \mathbf{u}_j \qquad \text{(Level 2)}, \end{aligned} \tag{6}$$

where $\mathbf{y}_j$ is the $n_j \times 1$ vector of outcomes for, say, the students in, say, school j, $\mathbf{X}_j$ is an $n_j \times P$ matrix of (known) student-level predictors at school j, $\mathbf{W}_j$ is a $P \times K$ matrix of (known) school-level predictors, the components of $\mathbf{e}_j$ are IID $N(0, \sigma^2)$, the $\mathbf{u}_j$ are P-variate normal with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{T}$, and there is independence across j at both levels of the model.

Since the advent of least squares 200 years ago, it has been standard in the regression framework of Level 1 in model (6) to include what might be thought of as individual-level latent variables (the $\mathbf{e}_j$), but the inclusion of the school-level latent variables $\mathbf{u}_j$ in Level 2 (to increase the realism of the modeling, by accounting for between-school heterogeneity of the regression relationship in Level 1) creates problems for standard frequentist methods such as ML. The likelihood function in this fully Gaussian formulation can be written down in closed form (although this is not true for versions of model (6) with Bernoulli outcomes), but the *MLE*s must be searched for iteratively.

Each of the existing fitting approaches tries to solve this problem in its own way, and each approach has its pluses and minuses:

- The EM algorithm produces full or restricted ML estimates by regarding the fitting of model (6) as a missing data problem, in which the $\boldsymbol{\beta}_j$ play the role of missing data. A big plus for EM is that it always converges, no matter what starting values it is given; a big minus is that it often reaches the *MLE*s sufficiently slowly that users may become impatient and stop it before it has converged. In fact, there is good evidence from Kreft, de Leeuw, & Kim (1990) that this has been happening with the package HLM. A natural improvement would be to switch over from EM to Newton-Raphson iterations after a while; perhaps future releases of the EM packages will try this.
- IGLS and Fisher scoring are closely related methods that treat the estimation of the regression coefficients $\boldsymbol{\gamma}$ and the variance/covariance unknowns σ^2 and $\mathbf{T}$ as separate but related problems, alternating the estimation of each using current guesses for the other. These methods

have the advantage of much quicker convergence than EM, and VARCL's implementation of Fisher scoring has as an additional plus: the ability to base its iterations on the Gaussian sufficient statistics rather than the raw data, so that it can accommodate quite large data sets. A minus for these approaches is that convergence to the global *MLE*s is not guaranteed unless the starting values are sufficiently good.

The principal drawback of all three of these methods, though, is generic to the reliance on the maximum of the likelihood function when this may be unwise. An example is provided by Rubin's (1981, 1989) meta-analysis of data gathered by Alderman & Powers (1979) to assess the effectiveness of high school coaching programs that attempt to raise the verbal Scholastic Aptitude Test (SAT) scores of enrolling students. The data arose from parallel randomized experiments at $k = 8$ high schools (labeled A–H) on a total of 559 students, with an average of about 30 treatment and 40 control students at each school. The estimated treatment effects ranged across the eight schools from about -1 to about $+28$ points, with *SE*s for these estimates on the order of 9–16 points (to assess the practical significance on these differences, SAT verbal scores in college-bound students might average about 600 with an *SD* of about 75 points).

In addition to conducting a fully Bayesian analysis of these data based on model (5) above, Rubin also examined the usual empirical Bayes solution that produces inferences about the Level 2 regression parameter(s) (in this case, just an intercept term) by conditioning on the *MLE*s of the variance parameters. He summarized the empirical Bayes fitting of model (5) in a splendid plot, reproduced below as Figure 1, in which the dotted line is the marginal likelihood for τ and the eight solid lines track the usual shrinkage estimates $W_i y_i + (1 - W_i)\hat{\mu}$ of the "true" treatment effects θ_i of Programs A–H as a function of τ, where $\hat{\mu}$ is the grand mean $\Sigma_{i=1}^{k} W_i y_i / \Sigma_{i=1}^{k} W_i$ (which here came out to about 8 SAT verbal points) and $W_i = 1/(V_i + \tau^2)$. It is evident from this plot that there is some heterogeneity in the effects of the coaching programs: viewing the marginal likelihood of τ as its posterior distribution with a diffuse prior, Rubin's fully Bayesian analysis implies a summary value of τ of about 7 SAT verbal points. However the *MLE* of τ is zero, which poorly summarizes the evidence for heterogeneity.

The problem is that the marginal likelihood functions for variance parameters in hierarchical models fitted to data sets with small numbers of Level 2 units may be highly skewed, as in Figure 1. The main solution to this problem in general with model (6) is the one Rubin adopted with model (5): a fully Bayesian analysis that propagates the uncertainty in the variance parameters through to the uncertainty about the regression parameters. This approach rewrites model (6) and adds a layer to the hierarchy for the specification of prior information:

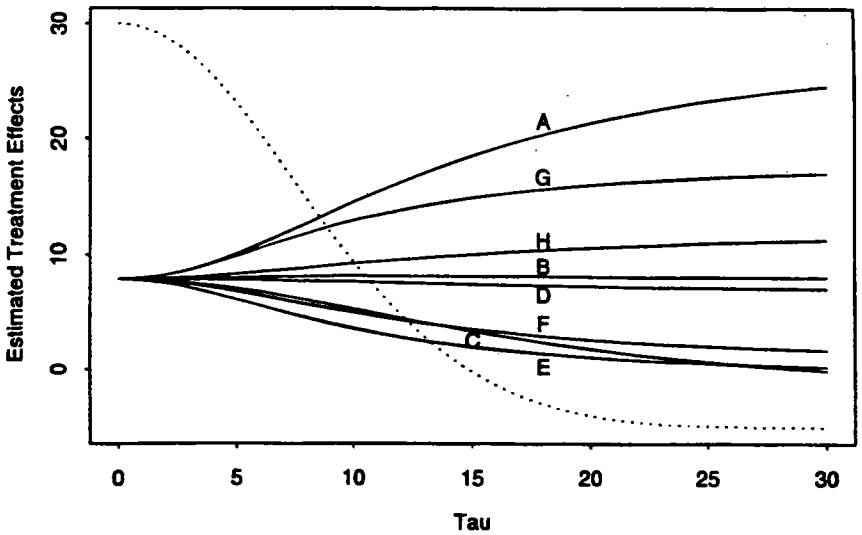

FIGURE 1. *Shrinkage estimates of coaching effects in the Alderman & Powers (1979) study as a function of the parameter τ, which quantifies heterogeneity across the eight experiments (from Rubin, 1981)*

$$\begin{aligned} (y_j|\beta_j, \sigma^2) &\sim N_{n_j}(\mathbf{X}_j\beta_j, \sigma^2\mathbf{I}_{n_j}) \quad \text{(Level 1)}, \\ (\beta_j|\gamma, \mathbf{T}) &\sim N_P(\mathbf{W}_j\gamma, \mathbf{T}) \quad \text{(Level 2)}, \\ (\gamma, \mathbf{T}, \sigma^2) &\sim p(\gamma)p(\mathbf{T})p(\sigma^2) \quad \text{(prior)}, \end{aligned} \tag{7}$$

The goal is computation and marginalization of the posterior distribution $p(\beta, \gamma, \mathbf{T}, \sigma^2|y)$, and—as with the three ML methods—no closed-form solution is possible.

Four approaches to fitting model (7) are available, at least in principle: methods based on Laplace approximations (e.g., Tierney & Kadane, 1986), quadrature (e.g., Naylor & Smith, 1982), Monte Carlo integration with importance sampling (e.g., Hammersley & Handscomb, 1964), and simulation from the posterior distribution, for instance, using MCMC techniques such as Gibbs sampling (e.g., Smith & Roberts, 1993). Of these, the most naturally tailored to the hierarchical structure of model (7) is Gibbs: the idea is to iteratively sample from the four conditional posterior distributions $p(\sigma^2|y, \beta, \gamma, \mathbf{T})$, $p(\beta|y, \gamma, \mathbf{T}, \sigma^2)$, $p(\mathbf{T}|y, \gamma, \beta, \sigma^2)$, and $p(\gamma|y, \mathbf{T}, \beta, \sigma^2)$ all of which do have closed-form expressions with standard, relatively flexible choices for the prior level in (7). The main disadvantages of Gibbs sampling are that it is highly computationally intensive and that it can be hard to figure out if it has converged, although decent diagnostics are finally beginning to emerge (e.g., Tanner, 1993). Its main advantages are that you get more accurate uncertainty assessments for the fixed-effects regression parameters in model (7), and that it is easy to answer just about any relevant question

using simple descriptive summaries of the simulated draws from $p(\beta, \gamma, T, \sigma^2 | y)$.

I conjecture that 10 years from now the industry standard in fitting hierarchical models will be one or another of the fully Bayesian methods, probably some sort of MCMC, to avoid the problems that maximum likelihood runs into with highly skewed marginal likelihood functions for the variance parameters. (In advocating fully Bayesian solutions, Rubin, 1989, offers the rather vague advice that "As problems become harder, it becomes more important to be more fully Bayesian," but it is the skewness of the likelihood function—and the sharp underpropagation of uncertainty about the regression parameters resulting from pretending that the posterior uncertainty about the variance parameters is zero—that are the real diagnostics for the need to be fully Bayesian here.) This will create an apparently new problem of its own—specification of the prior distribution at the top level of the hierarchy—and, indeed, it is possible that investigators have been avoiding Gibbs because of this. However, this problem is not really new, since all of the existing methods are, in effect, fully Bayesian with one implicit prior or another (quite possibly not a very sensible prior, at that), and, in any case, the way you deal with prior specification in practice is through (a) detailed subject-matter study followed by (b) sensitivity analysis, both of which are part of good analyses already. This statement is a bit facile in making the elicitation of informative prior distributions in complicated problems sound easier than it actually is, but considerable progress has already been made (e.g., Kadane, Dickey, Winkler, Smith, & Peters, 1980) and, in any case, increasing the average level of experience in the analytic community in this important activity will be a net gain.

Discussion

Multilevel models have unquestioned usefulness in the social sciences, particularly in education, for at least two good reasons. First, such models permit the direct framing of theories about the effects of structural change at each of a variety of levels in the educational hierarchy, and second, HMs at last offer the promise of routine and accurate adjustments to the standard uncertainty assessments based on simple random sampling, when the data are gathered in a hierarchical fashion in the presence of large intracluster correlations. However, the use of these models represents a net increase in the complexity of statistical modeling in the social sciences, and whenever such an increase has occurred in the past (see, e.g., Gould, 1981, on factor analysis), it has opened up the possibility of interpretive confusion and overstatement of what may be validly concluded from a given body of evidence.

I have used this article in part as an occasion to question the uncritical use of statistical models—not just HMs, and not just in education—with data from observational studies and samples whose exchangeability with the unsampled portion of the target population is uncertain. It is possible to

interpret this as an overly harsh attack on both hierarchical methods and observational data, but that is not my intent. Perhaps it would be good in closing to clarify the scope of my criticism, and the role I believe each of these things—multilevel models, and the design and analysis of observational studies and UE sampling plans—may constructively play in social science research.

Stochastic models for "nonstochastic" data? The main point I have tried to make about the modeling of observational data is that it would be a net gain for investigators to make a greater effort to justify the fitting of models involving random variables with such data. When the data-gathering process, either in experimental design or in sample surveys, involves the planned introduction of randomization, it is both natural and appropriate in the modeling of the resulting data—from either a frequentist or a Bayesian viewpoint—to regard the observations as the realizations of random variables, and ideally to identify the parameters in such models as population values of scientifically relevant quantities that may be estimated from the data. But what is the logical basis for the use of stochastic models with observational studies and UE samples?

For concreteness, consider again the growth-curve analysis employing model (1) by Huttenlocher et al. (1991) of data on mothers and children from 22 families in the Chicago area, self-selected by responding to newspaper ads. If you had gathered these data yourself, you would have every right to fit a parabola to each child's vocabulary growth curve, to let ϵ_{ti} stand for the lack of fit of such a parabola for child i at time t, to notice that the quadratic slopes from least squares do not appear to be the same for each child, to attempt to relate these slopes linearly to some function of the amount the child's mother spoke to him or her, and to call the lack of fit of this second equation U_i. But what would give you the right to regard the ϵ_{ti} and U_i as realizations of IID random variables?

The usual frequentist answer involves thinking of the data as randomly sampled values from a population, either (a) literally chosen at random from the collection of people of direct scientific or policy relevance, which would both justify what I have called sampling inference and animate the parameter estimates in model (1) and their standard errors with clear scientific meaning; or (b) hypothetically chosen from the collection of all possible data values you could get if you were to repeat your data-gathering activity forever, which would justify calibration inference arising from Fisher's hypothetical sampling model as described above, but would lend relevance only to the p value(s) at the heart of that inference. In the frequency interpretation of probability, the first of these options is not available with UE samples, leaving calibration as the only justifiable mode of inference with such data for frequentists.

In the Bayesian paradigm, you are always free to use probability to quantify your uncertainty about things you do not know for sure, so the Bayesian

story looks more promising as a basis for the "random variables" part of the phrase "IID random variables" above; and the device of de Finetti's Theorem (e.g., Draper, Hodges, et al., 1993) allows you to pass from exchangeability assumptions to IID. But just because you judgmentally assert exchangeability of sampled and unsampled units or of treatment and control people (apart from treatment status) does not make your assertion any more correct, when the available data are gathered in a nonrepresentative fashion from the target population or the treatment and control people differ with respect to relevant PCFs. Being a Bayesian permits you more freedom in attempting to move out of the top left corner of Table 2 with observational data and UE samples, but it does not provide a guarantee that any such movement is justified.

The overall point here is that conceptual issues about the *meaning* of inferential outputs come first, and have often been bypassed in the rush to complex modeling in social science research. This has negative consequences of both a process and an outcome character: the reification of hypothetical population parameters (process), and the overstatement of scope of findings (outcome). I am sure that the investigators mentioned above in the sections "Interpreting Hierarchical Analyses" and "A Taxonomy of Inferential Strength in Statistical Modeling" are keenly aware of the limitations of their data, and consequently of their results, but they do not always communicate these limitations effectively. If teaching and methodological research in statistics were sufficiently good, readers of these papers would be able to figure out the limitations for themselves, but it is probably unwise to count on this.

Making good use of observational data. I am not against the retrospective analysis of data from UE samples and observational studies when that is all one has. What I am against is (a) loose summaries of what has been learned from analyzing them, and (b) a prospective over-reliance on them when stronger designs are possible. When they are not possible (e.g., as noted by Goldstein [1994, personal communication], a randomized trial at one level of an educational hierarchy can often be no more than an observational study at other levels of the hierarchy), the emphasis shifts to identifying, measuring, and adjusting for all relevant PCFs to better justify the assumption of strong ignorability.

The other good use for observational data is in generating causal theories, to be subjected later to verification or falsification with stronger research designs. Raudenbush (1993, personal communication) identifies the tradeoff between representative sampling and concern for intensive measurement, and suggests that "Moving from nationally representative data with 'thin' measurement to local data with 'thick' measurement and then back to nationally representative data with new insights about measurement is a sensible way for science to evolve." Problems arise in this approach only when people forget to note that the results of exploratory analyses are tentative, and when nobody bothers to do the confirmatory follow-up studies. Both of these situations are disturbingly commonplace.

Prediction is key. When it is working best, the scientific method has a built-in feedback loop. You (a) formulate a theory, (b) use it to generate testable predictions of observable quantities, together with uncertainty assessments for those predictions, and (c) see how close these predictions come to observable reality. If the fit is bad, you modify the theory; if it is good, you think of a more stringent predictive test. Either way, you go back to the second step. With unfortunate encouragement from the majority of statistical practice and pedagogy over the last 50 years, the social sciences have lagged behind other disciplines (e.g., astronomy) in the application of this formula, and education is no exception. In the 25 or so substantive articles, book chapters, and books on HMs in education I studied while preparing this article, prediction was mentioned rarely and almost always only in passing.

HMs provide a natural technical framework for generating predictive distributions that can help to validate or falsify educational theories: in longitudinal analyses, by guessing at the future and waiting for it to unfold; in internal validation of cross-sectional analyses, by setting aside data values and trying to predict them from the rest; in external validation of cross-sectional studies, by predicting what will happen in the next sampled classroom. There is a golden opportunity for HM software developers and users to tighten up the feedback loop by shifting some of the emphasis from inference about unobservable parameters to prediction of observables on scales of direct educational relevance. What will become of this opportunity?

Notes

[1]By comparison, the *SE* from an OLS model that does not include random classroom effects is .074, a value that is misleadingly small because it does not account for the positive intracluster correlation arising from the nesting of students within classes.

[2]Actually, model (5) is not fully satisfactory, either: AMIS is so discrepant that it is necessary to give up either normality or unconditional exchangeability in attempting an analysis like that based on this model.

[3]Rodríguez & Goldman found "that the estimates of fixed effects and variance components produced by [VARCL and ML3] are subject to very substantial downward bias when the random effects are sufficiently large to be interesting."

References

Ahmed, T., Garrigo, J., & Danta, B. S. (1993). Preventing bronchoconstriction in exercise-induced asthma with inhaled heparin. *New England Journal of Medicine, 329,* 90–95.

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society, Series A, 149,* 1–42.

Alderman, D., & Powers, D. (1979). *The effects of special preparation on SAT-Verbal scores* (Research Rep. No. 79-1). Princeton, NJ: Educational Testing Service.

Ares, C. E., Rankin, A., & Sturz, H. (1963). The Manhattan bail project. *NYU Law Review, 38,* 67–95.

Box, G. (1994). Statistics and quality improvement. *Journal of the Royal Statistical Society, Series A, 157,* 209–229.

Brook, R. H., Ware, Jr., J. E., Rogers, W. H., Keeler, E. B., Davies, A. R., Donald, C. A., Goldberg, G. A., Lohr, K. N., Masthay, P. C., & Newhouse, J. P. (1983). Does free care improve adults' health? Results from a randomized controlled trial. *New England Journal of Medicine, 309,* 426–434.

Bryk, A. S., & Frank, K. (1991). The specialization of teachers' work: An initial exploration. In S. W. Raudenbush & J. D. Willms (Eds.), *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective.* (pp. 185–201). San Diego, CA: Academic Press.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147–158.

Bryk, A. S., & Raudenbush, S. W. (1989). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 159–204). San Diego, CA: Academic Press.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., Seltzer, M., & Congdon, R. T. (1988). *An introduction to HLM: Computer program and users' guide.* Chicago: University of Chicago.

Bureau of the Census. (1978). *The current population survey: Design and methodology* (Tech. Paper No. 40). Washington, DC: Census Bureau, Department of Commerce.

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

Burstein, L., Kim, K.-S., & Delandshere, G. (1989). Multilevel investigations of systematically varying slopes: Issues, alternatives, and consequences. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 233–276). San Diego: Academic Press.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Boston: Houghton-Mifflin.

Cochran, W. G. (1978). *Sampling techniques* (3rd ed.). New York: Wiley.

Daley, J., Jencks, S., Draper, D., Lenhart, G., Thomas, N., & Walker, J. (1988). Predicting hospital-associated mortality for Medicare patients: A method for patients with stroke, pneumonia, acute myocardial infarction, and congestive heart failure. *Journal of the American Medical Association, 260,* 3617–3624.

de Finetti, B. (1974–1975). *Theory of probability* (Vols. 1–2). New York: Wiley.

de Leeuw, J. (1992). Series editor's introduction to hierarchical linear models. In A. S. Bryk & S. W. Raudenbush, *Hierarchical linear models: Applications and data analysis methods* (pp. xiii–xvi). Newbury Park, CA: Sage.

Deming, W. E. (1947). *Some theory of sampling.* New York: Wiley.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1–38.

Diaconis, P. (1985). Theories of data analysis: From magical thinking through classical statistics. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Exploring data tables, trends, and shapes* (pp. 1–36). New York: Wiley.

Draper, D. (1987). On exchangeability judgments in predictive modeling, and the

role of data in statistical research [Comment on "Prediction of future observations in growth curve models," by C. R. Rao]. *Statistical Science, 2,* 454–461.

Draper, D. (1994). *Input-output screening for quality assessment in health and education* (Statistics Research Report 94:05). University of Bath, Bath, UK.

Draper, D., Gaver, D. P., Goel, P. K., Greenhouse, J. B., Hedges, L. V., Morris, C. N., Tucker, J. R., & Waternaux, C. (1993). *Combining information: Statistical issues and opportunities for research.* Alexandria, VA: American Statistical Association.

Draper, D., Hodges, J. S., Mallows, C. L., & Pregibon, D. (1993). Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society, Series A, 156,* 9–38.

Draper, D., Kahn, K. L., Reinisch, E. J., Sherwood, M. J., Carney, M. F., Kosecoff, J., Keleer, E. B., Rogers, W. H., Savitt, H., Allen, H., Reboussin, D., & Brook, R. H. (1990). Studying the effects of the DRG-based Prospective Payment System on quality of care: Design, sampling, and fieldwork. *Journal of the American Medical Association, 264,* 1956–1961.

Englert, C. S., Raphael, T. E., Anderson, L. M., Anthony, H. M., Fear, K. L., & Gregg, S. L. (1988). *A case for writing intervention: Strategies for writing informational text.* East Lansing: Michigan State University, Institute for Research on Teaching.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh, UK: Oliver & Boyd.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh, UK: Oliver & Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* Edinburgh, UK: Oliver & Boyd.

Fitz-Gibbon, C. T. (1991). Multilevel modeling in an indicator system. In S. W. Raudenbush & J. D. Willms (Eds.), *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective* (pp. 67–83). San Diego, CA: Academic Press.

Freedman, D. (1983). Statistics and the scientific method (with discussion). In W. M. Mason & S. E. Fienberg (Eds.), *Cohort analysis in social research: Beyond the identification problem.* New York: Springer.

Freedman, D. A., & Lane, D. (1983). Significance testing in a nonstochastic setting. In P. J. Bickel, K. Doksum, & J. L. Hodges, Jr. (Eds.), *A festschrift for Erich L. Lehmann* (pp. 185–208). Belmont, CA: Wadsworth.

Freedman, D., Pisani, R., Purves, R., & Adhikari, A. (1991). *Statistics* (2nd ed.). New York: Norton.

From the raw to the refined: Value-added, 1993. (1993, November 20). *Guardian Education Supplement.*

Geisser, S. (1993). *Predictive inference: An introduction.* New York: Chapman & Hall.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73,* 43–56.

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

Goldstein, H. (1989). Models for multilevel response variables, with an application to growth curves. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 107–125). San Diego, CA: Academic Press.

Goldstein, H. (1991). Better ways to compare schools? *Journal of Educational Statistics, 16,* 89–91.

Goldstein, H. (1992). Statistical information and the measurement of educational outcomes [Editorial]. *Journal of the Royal Statistical Association, Series A, 155,* 313–315.

Goldstein, H., & Thomas, S. (1993). Guardian *A-level analysis 1993: Technical Report.* London: Institute of Education, University of London.

Goodman, S. N. (1989). Meta-analysis and evidence. *Controlled Clinical Trials, 10,* 188–204.

Gould, S. J. (1981). *The mismeasure of man.* New York: Norton.

Greenland, S. (1993). *A critical look at some popular meta-analytic methods.* Manuscript submitted for publication.

Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo methods.* London: Chapman & Hall.

Holland, P. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association, 81,* 945–970.

Holland, P. (1989). Discussion of "Fisher scoring algorithm for variance component analysis of data with multilevel structure," by N. T. Longford. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 311–317). San Diego, CA: Academic Press.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27,* 236–248.

Jacobsen, S. (1991). The relationship between kindergarten screening measures and grade three achievement. In S. W. Raudenbush & J. D. Willms (Eds.), *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective* (pp. 167–183). San Diego, CA: Academic Press.

Judge, G. G., et al. (1988). *Introduction to the theory and practice of econometrics* (2nd ed.). New York: Wiley.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S., & Peters, S. C. (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association, 75,* 845–854.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridgé, UK: Cambridge University Press.

Kish, L. (1957). Confidence limits for clustered samples. *American Sociological Review, 22,* 154–165.

Kreft, I. G. 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. 311). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Lane, D. (1986). Comment on "Some subjective Bayesian considerations in the selection of models," by B. M. Hill. *Econometric Reviews, 4,* 253–258.

League table revamp on cards. (1994, March 25). *Times Educational Supplement.*

Lee, V. E., & Smith, J. B. (1991). Sex discrimination in teachers' salaries. In S. W. Raudenbush & J. D. Willms (Eds.), *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective* (pp. 225–247). San Diego, CA: Academic Press.

Lindley, D. V. (1972). *Bayesian statistics, a review.* Philadelphia: Society for Industrial and Applied Mathematics.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B, 34,* 1–41.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Little, R. J. A., & Schenker, N. (in press). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *A handbook for statistical modeling in the social and behavioral sciences.* New York: Plenum Press.

Lockheed, M. E., & Longford, N. (1991). School effects on mathematics achievement gain in Thailand. In S. W. Raudenbush & J. D. Willms (Eds.), *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective* (pp. 131–148). San Diego, CA: Academic Press.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika, 74,*817–827.

Longford, N. T. (1991a). Logistic regression and random coefficients. *Computational Statistics and Data Analysis,* forthcoming.

Longford, N. T. (1991b). Searching for multivariate outcomes in education. In S. W. Raudenbush & J. D. Willms (Eds.), *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective* (pp. 115–130). San Diego, CA: Academic Press.

Longford, N. T. (in press). Random coefficient models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *A handbook for statistical modeling in the social and behavioral sciences.* New York: Plenum Press.

Naylor, J. C., & Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics, 31,* 214–225.

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments: Essay on principles (with discussion). *Statistical Science, 4,* 465–480. Original work published 1923.

Oakes, M. (1990). *Statistical inference.* Chestnut Hill, MA: Epidemiology Resources.

Paterson, L. (1991). Trends in attainment in Scottish secondary schools. In S. W. Raudenbush & J. D. Willms (Eds.), *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective* (pp. 85–99). San Diego, CA: Academic Press.

Raffe, D. (1991). Assessing the impact of a decentralised initiative: The British Technical and Vocational Education Initiative. In S. W. Raudenbush & J. D. Willms (Eds.), *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective* (pp. 149–166). San Diego, CA: Academic Press.

Raudenbush, S. W., & Bryk, A. S. (1989). Quantitative models for estimating teacher and school effectiveness. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 205–232). San Diego, CA: Academic Press.

Rodríguez, G., & Goldman, N. (1993, July). *An assessment of estimation procedures for multilevel models with binary responses.* Paper presented at the Workshop on Multilevel Analysis at the RAND Corporation, Santa Monica, CA.

Rogers, W. H., Draper, D., Kahn, K. L., Keeler, E. B., Rubenstein, L. V., Kosecoff, J., & Brook, R. H. (1990). Quality of care before and after implementation of the DRG-based Prospective Payment System: A summary of effects. *Journal of the American Medical Association, 264,* 1989–1994.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies. *Biometrika, 70,* 41–55.

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

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics, 6,* 377–400.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin, D. B. (1989). Some applications of multilevel models to educational data. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 1–17). San Diego, CA: Academic Press.

Scheffé, H. (1956). Alternative models for the analysis of variance. *Annals of Mathematical Statistics, 27,* 251–271.

Scott, A. J., & Holt, D. (1982). The effects of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association, 77,* 848–854.

Seltzer, M. H. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational Statistics, 18,* 207–235.

Seltzer, M. H., Wong, W. H., & Bryk, A. S. (1993). *Bayesian analysis in applications of hierarchical models: Issues and methods* (Tech. Rep. No. 114). Los Angeles: University of California.

Smith, A. F. M., & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov-chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B, 55,* 3–23.

Sobel, M. E. (in press). Causal inference in the social and behavioral sciences. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *A handbook for statistical modeling in the social and behavioral sciences.* New York: Plenum Press.

Tanner, M. A. (1993). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions.* New York: Springer-Verlag.

Thomas, N., Longford, N. T., & Rolph, J. E. (1992). *A statistical framework for severity adjustment of hospital mortality rates* (N-3501-HCFA). Santa Monica, CA: RAND Corporation.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association, 81,* 82–86.

Welch, B. L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika, 29,* 21–52.

Wong, G. Y., & Mason, W. M. (1989). *Ethnicity, comparative analysis, and generalization of the hierarchical normal linear model for multilevel analysis* (Research Rep. No. 89–138). Ann Arbor: University of Michigan, Population Studies Center.

Zuzovsky, R., & Aitkin, M. (1991). Curricular change and science achievement in Israeli elementary schools. In S. W. Raudenbush & J. D. Willms (Eds.), *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective* (pp. 25–36). San Diego, CA: Academic Press.

Author

DAVID DRAPER is a member of the Statistics Group, School of Mathematical Sciences, University of Bath, Claverton Down, Bath, Avon BA2 7AY, England; dd@maths.bath.ac.uk. He specializes in Bayesian statistics, hierarchical models, and model uncertainty.

Journal of Educational and Behavioral Statistics
Summer 1995, Vol. 20, No. 2, pp. 190–200

Hierarchical Models for Educational Data: An Overview

Carl N. Morris
Harvard University

Key words: *maximum likelihood, random effects, multilevel models, multilevel model checking*

The use of hierarchical models in statistical applications, and for educational data, is a promising but still underutilized approach. However, because these models are more complicated than many standard methods, it is important that we, as users and developers, not rush to use them before we understand them. We emphasize here, in support of the views on hierarchical models expressed in the 3 preceding papers by Draper, by Rogosa and Saner, and by de Leeuw and Kreft, the need to not diminish hard thinking about data and iterative model checking when fitting hierarchical models, the need for more and better software, the need to test methods to assure their proper calibration, and the need to produce supporting materials to aid analysts and users of hierarchical modeling methods.

The RAND conference convened researchers knowledgeable of and interested in the theory and application of hierarchical models, especially for education data. Their concerns involved the proper use and interpretation of such models, and the availability and quality of software.

Hierarchical models are extremely promising tools for data analysis. As hierarchical modeling has become better understood and better supported, its applications have proliferated in such diverse fields, besides education, as health and medicine, quality assurance, demography, and remote sensing. Hierarchical models are the source of considerable excitement now, with the computing environment making their use widespread in this decade. Nevertheless, there is the danger that these fascinating but complex models will be oversold before their performance is adequately understood, and that their nominal operating characteristics may not be achieved in particular applications.

The author is grateful to the Center for Advanced Study in Behavioral Sciences (CASBS) at Stanford University for hosting his sabbatical, 1993–1994, and to the NSF Grant SES 9022192 at CASBS. This work was supported also by AHCPR Grant HS 07118-02 at Harvard Medical School, which aids the development of hierarchical models and their applications to medical and health policy data.

I also appreciate comments made by Cindy Christiansen, Ree Dawson, and Phil Everson.

Hierarchical model is an umbrella term that I use here for two separate statistical objectives described by one multilevel model. The first objective concerns inferences about the model's *structural parameters,* which govern the Level 2 distributions (*Level 2* is defined in Equation 2, below). These are also called the *random effects,* the *mixing distribution,* the *random coefficients,* or the *hyperparameters,* from various statistical perspectives. The second objective concerns inferences about the *individual parameters.* The individual parameters are the quantities to be estimated in empirical Bayes inference, in methods for borrowing strength, for Stein estimation, and for multiparameter inference. Both inference problems are addressed in the "hierarchical Bayes" literature, a term adopted and emphasized by Lindley and Smith (1972).

The theory for inferences about the structural parameters tends to be less complicated than for the individual parameters, and there is more commercially available software for that purpose. Educational applications of hierarchical models—for example, those discussed at the RAND conference—tend to emphasize inferences about the structural regression parameters. Because the theory for structural parameter inference is easier, although not easy, the models considered sometimes can be more complicated than those adopted for individual parameter inferences.

I began my interest in hierarchical models jointly with Brad Efron, originally because we were inspired by the now celebrated work of Charles Stein, and our work resulted in extensions of Stein's formulation to cover more applications (e.g., Efron & Morris, 1975). Now my interest stems from wanting to facilitate a more general perspective of statistics required to make hierarchical models work in practice. The original Stein setting does not allow making certain inferences from hierarchical models—for example, interval estimates—and it inhibits applications to distributions other than the Normal. Still, a great practical advantage of Stein's perspective is that it avoids large sample asymptotics in k, the number of individual parameters, by establishing exact operating characteristics for fixed-size samples—for example, minimaxity for summed squared-error loss functions—and it encourages the search for optimum estimators. The calculations in James and Stein (1961) show that when k is moderate or small, as is common, including in some of the example applications at this conference, maximum likelihood and other large sample techniques provide inaccurate inferences. Examples are provided below to illustrate the bias for small k toward overstating precision in hierarchical models. That is, the estimates of the k individual parameters are biased, and the precisions calculated by statistical packages that are based on maximum likelihood are overstated when k is small. The problem of biased variance estimates is diminished for estimating the structural parameters (random effects), but even there, variability is understated.

Because maximum likelihood procedures are used widely for variance component estimation in the commercially available hierarchical modeling

software, testing and evaluation of methods for analysis of hierarchical models is needed to determine when the software provides estimates that have approximately the operating characteristics claimed. Such tests will almost certainly show for all the existing packages that estimating means is fairly robust, but that variance estimates are too small, and that confidence intervals cover insufficiently. These concerns will be magnified as the dimension p of the unknown covariance matrix of the structural parameters increases (only the case of $p = 1$, when τ^2 is a real number, is discussed here or by the conference participants), and they will almost assuredly worsen if the number of levels of the hierarchical model increases beyond the two levels discussed here. Little such testing has been reported, despite Stein's legacy.

Robustness to departures from the assumed model also needs to be understood better. In particular, model checking methods must be integrated into software. Appropriate diagnostic methods are analogous to procedures already used standardly for regression modeling, but they must be extended to check with data the additional assumptions made when fitting hierarchical models.

The Model and an Example

The main hierarchical models treated in the papers by Draper and by Rogosa and Saner address Normally distributed data with regression models for the data at Level 1, and another regression model governing the distributions of the individual parameters at Level 2 (cf. Draper's section "Fitting HMs in Education . . . ," de Leeuw & Kreft's section "Multilevel Models," Rogosa & Saner's section "Straight-Line Growth Curve Formulation"). Many applications can be handled by a special case that has been widely researched by denoting sufficient statistics as the observed data for different individuals (e.g., classrooms or schools), the Level 1 units.

At Level 1, the data Y_i follow Normal distributions

$$Y_i \sim N(\theta_i, V_i) \quad \text{indep.} \quad i = 1, \ldots, k \tag{1}$$

where k is the number of individuals, θ_i the true individual parameter, and V_i the known variance of individual i. In practice, the variances might be unknown, but with V_i estimated quite accurately from the within-individual sum of squares, or as σ^2/n_i, with σ^2 estimated by pooling all the data, and n_i the number of observations for the ith individual.

Level 2 of this model specifies the unknown distributions for θ_i, which also are assumed Normally distributed.

$$\theta_i \sim N(\beta' \mathbf{x}_i, \tau^2) \quad \text{indep.} \quad i = 1, \ldots, k. \tag{2}$$

In this case, $\mathbf{x}_i$ is an r-dimensional vector of covariates specific to the ith individual, including any constant term. The regression coefficients for this

structural model are denoted $\beta = (\beta_1, \ldots, \beta_r)'$, and τ^2 is the between-groups variance. β and τ^2 are the hyperparameters, or the random effects. In many educational applications the main objective is estimating these random effects, but there are other important applications in which estimating the vector $(\theta_1, \ldots, \theta_k)$ of unknown individual parameters is the main interest. By combining (1) and (2) we derive

$$Y_i \sim N(\beta' \mathbf{x}_i, V_i + \tau^2) \quad \text{indep.} \quad i = 1, \ldots, k. \tag{3}$$

From the likelihood function for this distribution the unknown β and τ^2 can be estimated, provided k is at least $r + 1$. Methods for estimating these unknown random effects include method of moments, maximum likelihood, and Bayesian methods. When k is large, any consistent method may be used, but for small values of k the estimation may be delicate and care must be taken to use accurate methods.

To estimate the individual parameter θ_i, one has the conditional distribution for known β and τ^2:

$$\theta_i | Y_i, \beta, \tau^2 \sim N[(1 - B_i)Y_i + B_i\mu_i, V_i(1 - B_i)] \tag{4}$$

where $B_i = V_i/(V_i + \tau^2)$ and $\mu_i = \beta' \mathbf{x}_i$. Here B_i is the *shrinkage factor.* Values of B_i near zero indicate little shrinkage for that individual component and, therefore, little benefit to using a hierarchical model. Values of B_i near one provide nearly full shrinkage to the mean μ_i, so that familiar weighted least squares regression methods can be used to approximate the analysis. Values of B_i not near the two extremes, zero and one, give results that are substantially different from those given by standard regression methods, and thereby justify the use of hierarchical modeling methods. Thus, the individual values B_i, or an average of their values, serve as a diagnostic to decide when the hierarchical model must be fitted.

This notation and a fuller description of the analysis for this Normal model is provided in Morris (1983b). A special case, of key theoretical importance, occurs for equal variances $V_i = V$ (James & Stein, 1961). While the assumptions of equal V_i are too restrictive, rarely applying to real data, the rat data discussed by Draper fits this model.

In this special case, the shrinkage constant B (the shrinkage values are equal for equal variances) is estimated by maximum likelihood as $k \cdot V / S$, subject to this value not exceeding one, where S is the residual sum of squares of the Y_i values, taken around their fitted mean. The James-Stein estimator uniformly dominates the maximum likelihood estimator, multiplying the maximum likelihood estimator by $(k - r - 2)/k$. This factor, always less than unity, accounts for a bias due to using maximum likelihood in this case. The bias is small when k is large, as in some of the examples, but it is quite

severe in Draper's and in Rogosa and Saner's discussions of growth curves with $k = 22$ and $k = 10$, respectively. For example, the Rogosa and Saner rat data has $k = 10$ and $r = 2$, so the *mle* of the shrinkage factor is nearly double the best estimate.

Moreover, most estimates use the *mle* of the variance $V_i(1 - B_i)$ in (4) as the estimate of variance for the hierarchical modeling estimate of θ_i. Not only does using an overly large estimate of B_i cause underestimation of this variance, but even more importantly, this formula $V_i(1 - B_i)$ is valid only when the hyperparameters β and τ^2 are known. Additional terms need to be tacked on otherwise (see Morris, 1983b).

Let us now consider an example (Bryk & Raudenbush, 1991) for estimating individual effects, for which the model (1)–(4) is assumed to apply. These are the teacher expectancy data (Bryk & Raudenbush, chap. 7), which involve $k = 19$ classroom studies. The 19 effect sizes $Y_1, \ldots, Y_{19}$ are assumed distributed as (1). A covariate, the number of weeks of prior teacher-student contact, is available, as used in (2). When analyzed by HLM, or any other commercially available package based on maximum likelihood methods, the variances reported for the standard errors of the effect estimates θ_i, when the covariate is not used, are given by (4). All variances reported are too small, substantially so, as shown next.

An alternative and preferable method that acknowledges and provides for uncertainty due to estimating the hyperparameters is one based on a fully Bayesian model, essentially that described by Draper: assigning flat distributions to the hyperparameters β and τ^2. (Using a flat distribution on $[0, \infty)$ for τ^2 provides proper posterior distributions, and gives good properties in the equal variance situation, where the method can be studied theoretically [cf. Morris, 1983a].) This method, which makes the posterior density proportional to the likelihood for τ^2, as in Figures 1 and 2, provides results quite similar to those in Morris (1983b). When no covariate is used, the variances for this preferred method range from 39% to 197% larger in the 19 cases than those obtained using maximum likelihood. Use of the covariate provides further shrinkage. In this case, the preferred analysis provides variance ratios for the individual estimates that range from 2% to 535% larger, with a median of 217%. That is, correct variances are triple the values given by maximum likelihood estimation.

Estimating random effects for these data with $k = 19$ leads to less underestimation of variances, but it still is quite noticeable. When no covariate is used, the preferred method provides a variance that is 71% larger than the *mle* does for the mean μ. When the covariate is included, the variances of the estimates of β_0 and β_1 (2) are reported as 15% and 18% higher by the preferred analysis than the corresponding values for the *mle*.

Figure 1 shows the likelihood function, based on (3), for the unknown variance τ^2, for these 19 observations and fitting the covariate. Even though $\hat{\tau} = 0$ is the modal value, substantial likelihood is present for values of τ up

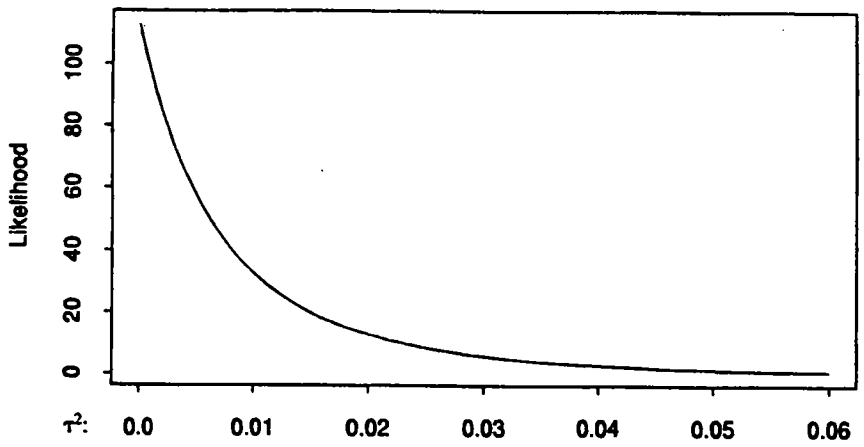

FIGURE 1. *Likelihood graph (restricted likelihood, β integrated out) for τ^2, teacher expectancy data (Bryk & Raudenbush, 1991). Covariate fitted,* mle *is $\hat{\tau} = 0.00$. All the likelihood lies to the right of 0, which indicates that variation is underassessed by the* mle.

to 0.15. The modal value, which is the maximum likelihood estimate, understates the likely values of τ^2. Figure 2 is the corresponding likelihood, fitted by assuming no covariates. Note again that much more of the likelihood, as indicated by area under the curve, lies to the right of the mode. These graphs are the same as two equivalent ones published by Raudenbush and Bryk (1985), except that we also think of these graphs as posterior densities on τ^2. Draper's Figure 1 provides a second such example with $k = 8$ (the

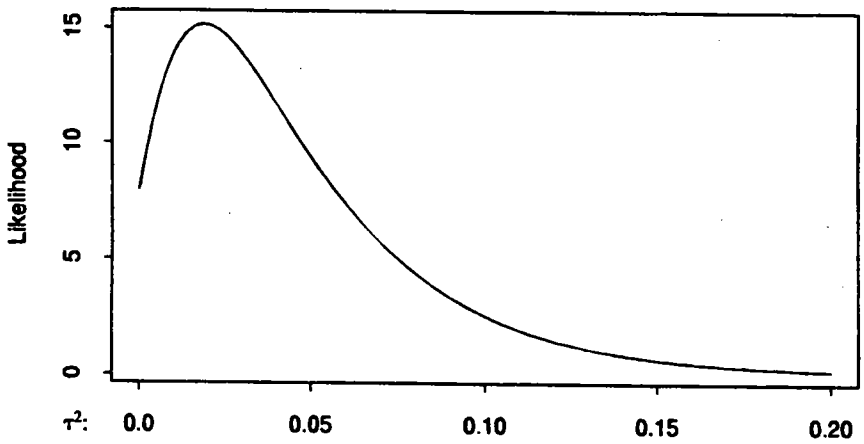

FIGURE 2. *Likelihood graph (restricted likelihood, β integrated out) for τ^2, teacher expectancy data (Bryk & Raudenbush, 1991). Without covariate,* mle *is $\hat{\tau} = 0.14$. Most of the likelihood lies to the right of 0.14, which suggests that variation is underassessed by the* mle.

example was originally presented in Rubin, 1981) where the likelihood for τ is maximized at zero, but larger values, in that case up to $\tau = 15$, are quite plausible. These examples show that maximum likelihood estimation for estimating variance components often provides estimates that understate τ^2, the between-individuals variance, and therefore that understate variances.

If the purpose of hierarchical modeling is to give a more adequate measure of uncertainty and, especially, to provide better variances of estimates, then we still have not adequately done this when we use maximum likelihood methods and the number of individuals k is small. The commercially available hierarchical modeling packages provide more general analyses than those just described for (1)–(2), including when each observation $Y_i = (Y_{i1}, \ldots, Y_{ip})$ is a p-dimensional vector. See, for example, the models described in Draper's section "Fitting HMs in Education . . . " and de Leeuw and Kreft's section "Multilevel Models." The concerns about severe biases for small k that have been raised here are likely to be more severe for higher dimensions when p exceeds 1. More testing is needed to show, when the models are correct, that the procedures being provided have approximately the operating characteristics (coverage probabilities, mean squared error, etc.) suggested by the values provided nominally. It is easy to see that this does not happen for small k with maximum likelihood methods.

Discussion of Three Papers

The conference was convened partly to bring together researchers involved in a range of hierarchical modeling topics, including the proper use and appropriate applications of hierarchical modeling. A key impetus was provided by the National Center for Educational Statistics (NCES), which asked a range of questions about hierarchical models, four of which are addressed by de Leeuw and Kreft. These questions imply that considerable interest and attention is being paid to hierarchical modeling methods. That must please those of us who have helped to foster its development and implementation, but we must be concerned that such methods can be oversold, or are so complicated that researchers will lose sight of practical issues. Draper reminds us of these points with his discussion of observational data and of the different levels of their scientific validity. The first concern in applied statistics is to gain a deep understanding of the data and their relationships before cramming them into some package that provides sophisticated answers. If the data lack the basic information needed, no analysis, however cleverly done, can provide accurate answers.

In their paper, de Leeuw and Kreft have provided simple and helpful answers to four of the questions (relisted here in the Appendix) posed by NCES. Further comment follows.

Question 1 presumes fitting of a "correct" model. Obviously, use of an ill-fitting hierarchical model that recognizes two levels but that fails to capture the essential distributions and relationships in the data could be harmful. The

value of fitting a two-level model can be partly measured, as mentioned in the preceding section, by computing the typical amount of shrinkage, B_i. Note that $1 - B_i$ is akin to the reliability, as Rogosa notes in his Appendix C. While the conference and the accompanying papers have focused almost entirely on Normal data, hierarchical models cover other distributions, too, including Poisson, Binomial, etc. These extensions would be needed for such data. Besides commercial software, other researchers have been developing free software for hierarchical analyses. Examples in the Splus program include methods for the Normal, Poisson, and Binomial distributions now available from Harvard (you may find out about the anonymous ftp request by writing morris@stat.harvard.edu), and there is an Splus program for the Normal distribution by W. DuMouchel available by E-mail at dumouch@bayes.cpmc.columbia.edu. These programs account for all levels of uncertainty in the models, including uncertainty in the hyperparameters.

Question 2 corresponds here, for example, to the equivalence of Equations 3–4 to Equations 1–2. As de Leeuw and Kreft note, different versions are valuable for different reasons. I call (1)–(2) the "descriptive model," because it is most valuable for thinking about applications. But (3)–(4), called the "inferential model," is a mathematically equivalent version that enhances the role of making appropriate inferences, also called "estimation" by de Leeuw and Kreft. De Leeuw and Kreft provide and recommend simple, not fully efficient methods for estimating two-level models, as do Rogosa and Saner with their SFYS (smart first-year student) example. These approaches are motivated by the need for simply understood results that assure one that a more complicated analysis is valid, or that make it possible to substitute methods based on standard software for more complicated ones. Draper takes a very different view by predicting that Markov-chain Monte Carlo (MCMC) methods will be widely preferred and dominant within a decade. I do not know what direction we are headed in, but I have preferred something more intermediate, by favoring iterative methods that compute fully efficient estimates, but which converge quickly enough so that most analyses can be obtained in a few seconds, in today's computer environment. This is easily done for the model (1)–(4), but it becomes much more difficult to do for even the more general Normal models considered by de Leeuw and Kreft and by Draper. MCMC methods have several awkward features. They currently require enormous computing time, and they do not replicate upon recalculation because they use simulation techniques. Of course, computers will become much faster, but even then many applications, especially for simpler hierarchical models, can be handled adequately by the simpler models. Another advantage of faster programs is that they can be used repeatedly to calibrate by simulation the operating characteristics of hierarchical modeling methods. This is an extremely important undertaking that has thus far been mostly ignored.

Question 3 has been answered in several ways by de Leeuw and Kreft. I

note again that the average shrinkage factor is a crucial summary that governs the reduction in variance (see (4)), in addition to being the amount of shrinkage. The variance ratio (maximum V_i/minimum V_i) is another important diagnostic because hierarchical models work better when this ratio is small (near 1, which is the Stein setting). To know whether any model works well requires the use of model checking methods.

Little has been said about model checking for hierarchical models in these papers. Standard model checking methods can be used to validate the Level 1 model, but the new requirement for hierarchical modeling is to validate the Level 2 model (Equation 2). This can be done, subject to the validity of (1), by making tests based on the model (3), when relates the data to the hyperparameters. In particular, as Draper emphasizes, there are crucial exchangeability judgments embraced in (2) that sometimes are not carefully considered in applications, see Morris (1983b) for more on this.

Usually, good applied modeling is conducted iteratively, as one fits a model and then checks it against the data. That iterative process is no different for hierarchical models than for more familiar cases; it is more complicated simply because it involves checking the Level 2 model.

Question 4 is answered with an emphatic yes by de Leeuw and Kreft, and by Rogosa and Saner. I agree that more than HLM software is needed, partly because of the failure of maximum likelihood methods for small data sets when k, the number of individuals, is not large. (Note that it is irrelevant if n_i, the number of observations that comprise each of the individual data sets, is large; it is k that matters.) Beyond that, Rogosa and Saner note some specific problems with the reliability of the HLM software. I will add to those concerns by saying that we purchased HLM Version 3.0 in September, 1993. Though it performed as expected for the examples we tried with $p = 1$, it failed in our first example for $p > 1$. The program, which is based on EM methods, did not cause the likelihood function to increase monotonically at each step, as the EM method is required to do. Thus, we were not confident of its convergence. Moreover, the printed results included a nonsymmetric correlation matrix with numerous correlations not between -1 and 1. When this example was reported to HLM, we were notified that a newer version would not have these errors. That version, 3.01, was recently provided, and though we have not checked it thoroughly, it does not have the problem just mentioned for the same data.

Summary and Recommendations

I am not sure what the limits of the topic "statistics" are, but uncertainty and its quantification are at the core. Hierarchical modeling is an approach that enhances our ability to assess uncertainty. I offer the following recommendations as we strive together to make hierarchical modeling a widely available and useful tool.

(1) Fitting hierarchical models is no substitute for thinking hard about data

and its structure, and for intimate involvement in understanding and in the careful analysis of all components of the data. As beneficial as hierarchical modeling may be, it must not be allowed to distract the analyst's attention away from gaining a basic understanding of the data.

(2) Fitting hierarchical models must, like any other form of statistical analysis, be combined with model checking efforts.

(3) For small k, maximum likelihood estimation can be distorted. In particular, hierarchical modeling methods must account for uncertainty in hyperparameter estimates, which maximum likelihood estimation ignores. Bayesian methods provide a natural way to do this. However they are derived, the resulting inferential methods must be checked to insure when and whether they perform well in repeated sampling.

(4) There still is much software development to be done to support hierarchical modeling. It must be extended to cover a variety of distributions, it must provide formal and graphical model checking support, it must be accurate and error free, and the operating characteristics must be tested and calibrated to the nominal values.

(5) More research is needed on the power analysis for studies that will use hierarchical methods. This information and methods for experimental design and survey design need to be available to those proposing and planning such studies.

(6) We must train users to recognize when multilevel features are present in their data. They must be sensitive to the additional hierarchical model requirements, including, especially, exchangeability in Level 2 of the model. Journal and grant referees also need this information to know better if hierarchical models have been used properly, and to good effect. More books, courses, and user-friendly software will help to insure proper interpretation of analytical results and the inferences being made.

APPENDIX

Questions posed by NCES and discussed by de Leeuw and Kreft

Question 1: Is some form of hierarchical linear model always preferable when conducting analysis with independent variables from two levels of a hierarchical data set?

Question 2: Some analysts are more comfortable presenting HLM results in terms of a combined model, i.e., a single regression equation containing interaction terms. Others prefer to discuss the coefficients without recourse to a single regression equation. Are the two approaches equally valid?

Question 3: Most discussion of HLM results centers on the individual coefficients: the betas and gammas. There is, of course, some interest in the overall measures, such as the proportion of variance explained. What is the best way to obtain and present overall measures when using HLM?

Question 4: Are there alternatives to the HLM software that NCES should consider using?

References

Altman, D. G., & Goodman, S. N. (1994). Transfer of technology from statistical journals to the biomedical literature. *Journal of the American Medical Association, 272,* 129–132.

Bryk, A. S., & Raudenbush, S. (1991). *Hierarchical linear models for social and behavioral research: Applications and data analysis methods.* Newbury Park, CA: Sage.

Efron, B., & Morris, C. N. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association, 70,* 311–319.

James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 361–379). Berkeley: University of California Press.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, B34,* 1–41.

Morris, C. N. (1983a). Parametric empirical Bayes confidence intervals. In G. E. P. Box, T. Leonard, & C.-F. Wu (Eds.), *Scientific inference, data analysis, and robustness* (pp. 25–50). New York: Academic.

Morris, C. N. (1983b). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association, 78,* 47–55.

Raudenbush, S., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10,* 78–98.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics, 6,* 377–401.

Author

CARL N. MORRIS is Professor and Chair, Department of Statistics, Harvard University, Science Center, Cambridge, MA 02138; morris@hustat.harvard.edu. He specializes in statistics.

Journal of Educational and Behavioral Statistics
Summer 1995, Vol. 20, No. 2, pp. 201–204

Hierarchical Data Modeling in the Social Sciences

Harvey Goldstein
Institute of Education, University of London

The last 10 years of active research in the area of hierarchical, multilevel data modeling has brought problems as well as benefits. The three conference papers reflect well both the potentialities of the new procedures and some of the dangers we need to guard against. As in all statistical modeling of the real world, our inferences are no better than the data upon which they are based and the adequacy of the assumptions we are prepared to make.

The paper by de Leeuw and Kreft sounds some useful warnings, and I will discuss that one first. The paper by Rogosa and Saner focuses in detail on a repeated measures application and one software package, and asks questions about the usefulness of the available analysis procedures. I shall have some general remarks about ways of handling repeated measures data, but leave comments about the HLM software to Professor Raudenbush to respond to. The paper by Draper is concerned with causal inference and ways in which this can be strengthened by using multilevel models. He also places these models in their historical context, and his discussion of competing estimation procedures raises some interesting topics for future research.

de Leeuw and Kreft

This paper considers the relatively simple linear two-level model with a continuous response variable. It provides a useful introduction by taking the reader from a series of separate equation regressions to a random coefficients model. The authors are right to emphasize the need to provide interpretational guidance for users, but, in my view, tend to exaggerate some of the difficulties. For example, the Level 2 covariance matrix of random coefficients can be used to provide estimates of the between-school variance as a function of the predictor variables, and this can be plotted to give insights into how, say, the school level variation changes with social background or gender. In addition, by calculating posterior means of the coefficients for each school, the individual (estimated) school relationships can be plotted—remembering, of course, that these are "shrunken" estimates.

While the use of a relatively simple model has advantages, it ignores some interesting extensions. It is a pity that the authors, having got as far as considering a two-level random coefficient model, do not discuss, for example, the modeling of the Level 1 variance. In many educational data sets we find heteroscedasticity at Level 1. Thus, boys tend to have higher variances for test scores than girls, and in a longitudinal study one often finds that those students with low pretest scores have smaller variance on a posttest

score than those with high pretest scores. Indeed, in some cases, fitting complex variation at Level 1 considerably improves the overall explanatory power of the model and the stability of other parameters. It is also the case, of course, that there is now considerable interest in nonlinear multilevel models, especially generalized linear models for proportions and count data, but I shall return to that below.

The distinction drawn between simple noniterative estimation procedures and iterative maximum likelihood (ML) or restricted maximum likelihood (REML) is now, I think, rather artificial. The standard advantages of ML or REML in terms of efficiency are important and the computational penalty is not usually very severe. The simpler methods are, however, sometimes more robust—a property shared with the iterative generalized estimating equation (GEE) approach (Liang & Zeger, 1986). This property may be useful, for example, when we suspect that multivariate Normality does not hold, but for most social scientists it is the structure of the model which requires explication rather than the details of the estimation procedure.

I am, of course, delighted that the authors, in their final section on software, speak well of the flexibility of the ML3 software. This flexibility was designed from the outset because we wished to have an open general system that could easily incorporate new developments. This has allowed us to add facilities, such as the ability to handle random cross-classifications, measurement errors, and nonlinear, especially generalized linear, models, as the relevant estimation theory has been developed. This is currently coming to fruition in the form of the next, many-level version, MLn.

There is a danger, and this paper reminds us of it, that multilevel modeling will become so fashionable that its use will be a requirement of journal editors, or even worse, that the mere fact of having fitted a multilevel model will become a certificate of statistical probity. That would be a great pity. These models are as good as the data they fit: they are powerful tools, not universal panaceas.

Rogosa and Saner

Repeated measures data constitutes a very good example of a situation in which a two-level model is really essential, because most of the variation typically is at the higher level. The literature on fitting repeated measures data, especially from growth studies, has a long history (see, for example, Goldstein, 1979) and its formulation as a two-level model immediately solves a great number of outstanding problems.

One of these is that previous models, based upon a multivariate formulation, were able to handle only measurements made at discrete times, possibly with some missing responses. In the two-level formulation, this requirement is completely unnecessary and we can have any pattern and number of repeated measurements per individual, including individuals who contribute only one measurement, and obtain fully efficient (ML or REML) estimates using any

of the existing multilevel software packages. It is a pity, therefore, that the authors stick with discrete time data sets, because their conclusions about comparisons among estimation procedures rely heavily on the fact that their example data sets are highly balanced.

The authors make a useful point about data description and presentation. Nevertheless, after fitting a two-level model we can estimate residuals (posterior means) and plot their standardized values in a number of ways, which generally will be more reliable than the simple OLS plots when the number of measurements per individual is small. The authors are also right to point to the little work that has been done on study design.

Finally, it is worth pointing out that the basic two-level repeated measures model can be extended in a number of useful directions. At the Institute of Education, we have recently completed work on fitting models where the Level 1 residuals have an additional time series structure, which often occurs in growth data with measurements taken close together in time (Goldstein, Healy, & Rasbash, 1994). The models can also be extended to multivariate responses and can be used to provide efficient methods for growth prediction (Goldstein, 1995).

Draper

David Draper's discussion of justifiable inference is clear and a further useful reminder that we should pay as much attention to the source of our data as to the methods of their analysis. The discussion of Huttenlocher's analysis, however, raises a further issue which is not discussed.

When researchers use convenience samples, they sometimes do so because they have evidence (or a view based on their professional experience) which leads them to believe that there is a close correspondence between their convenience population and the real population of interest. The problem is that this correspondence is uncertain and difficult to quantify and is often not made explicit. Yet it does sometimes happen that inferences based upon formally inadequate samples give accurate inferences or predictions—voting intention surveys are a case in point and this may be more than just luck. Of Draper's examples, some fall into this category. Among them is one on fitting growth curves to London children which I used in my book (Goldstein, 1987). This is an interesting case because I was clearly guilty of improperly contextualizing the study which produced the data. In fact, that study was one of a series of collaborative studies across Europe of which one of the intentions was to see whether growth patterns could be replicated. It turns out that in the area of child growth there is indeed a considerable uniformity of pattern across different population groups (Tanner, 1962) so that there is good reason to feel confident about the generalizability of the results. From a scientific point of view, it is the replicability of findings in very different contexts that is usually more convincing than the evidence from a single representative sample. The moral would seem to be that investigators should

be more explicit about all their sources of evidence when they attempt to produce generalizable statistical inferences. This could be added to Draper's list of desiderata in his section "The Value of Explicitness in Inferential Conclusions."

From my point of view, the main reason for producing the *Guardian* value-added survey was to counter the misuse by the British government of raw school examination results to produce league tables. The intention was to demonstrate that both adjusting for intake achievement and presenting uncertainty intervals were necessary, although not sufficient, conditions for valid comparisons. We were not primarily interested in causal inferences, although I believe that the data are adequate enough for that, and we are currently pursuing it.

The emergence of Markov-chain Monte Carlo (MCMC) methods such as Gibbs sampling is clearly very important for a wide range of estimation problems, especially where there are small numbers of units. It is not at all surprising, of course, that in Rubin's example with eight Level 2 units, the likelihood estimate of the variance is zero and that the inclusion of prior information gives a positive estimate. In a likelihood framework this emphasizes the importance of procedures such as bootstrapping, which, like MCMC methods, allows accurate assessment of parameter uncertainty.

It is interesting that Draper quotes Rodríguez's findings on the bias in estimation for multilevel models with binary responses. This has led to collaborative methodological work resulting in a considerable improvement (Goldstein, 1995) and is an example of the kind of critical evaluation of techniques which Draper emphasizes.

References

Goldstein, H. (1979). *The design and analysis of longitudinal studies.* London: Academic Press.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Griffin.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold; New York: Halstead Press.

Goldstein, H., Healy, M. J. R., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine, 13,* 1643–55.

Liang, K., & Zeger, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika, 73,* 45–51.

Tanner, J. M. (1962). *Growth at adolescence.* Oxford: Blackwell.

Author

HARVEY GOLDSTEIN is Professor, Institute of Education, 20 Bedford Way, London, WC1HOAL, England; hgoldstn@ioe.ac.uk. He specializes in the modeling of hierarchical data structures.

Journal of Educational and Behavioral Statistics
Summer 1995, Vol. 20, No. 2, pp. 205–209

Hierarchical Models and Social Sciences

Nicholas T. Longford
Educational Testing Service

Key words: *educational process, noninformative allocation, observational study, uncertainty*

The view is presented that multilevel methods are just one element in a hypothetical complete analysis of observational data on human subjects. In most contexts several sources of uncertainty, in addition to those captured by a multilevel analysis, are present, and so the confidence placed in the results of a typical multilevel analysis is unrealistically optimistic. A "software-free" analysis of longitudinal data with rectangular design is outlined. Questions posed by the National Center for Education Statistics and elaborated by de Leeuw and Kreft are briefly discussed.

Sources of Uncertainty

Multilevel models have, in recent years, provided a powerful impetus for methodological developments in statistics with orientation toward applications in social sciences. The relevance of multilevel models to several prominent problems in educational research, such as school effectiveness studies and longitudinal surveys, is well established. Researchers are invited to apply these methods by several software packages, and there is a burgeoning list of references illustrating and offering advice on their use. However, substantive products, in the form of contributions to understanding or to improvement of educational processes, are few, if any.

Multilevel methods are commonly credited with improved estimation, especially of the standard errors of the model parameters. The core of the argument about the improvement is the more realistic nature of the model in comparison with its by-now outdated alternatives. A finer issue is the extent of the improvement afforded by the multilevel models. Optimism about this issue is widespread, although, in my view, not always justified or well qualified. I believe that in any study involving students and their mental performance, there are numerous sources of uncertainty which have a nontrivial impact on the conclusions of the analysis.

By way of illustration, consider a study in which students in a number of classes are given a test at the beginning of an academic year, and another test, for the same domain of knowledge, at the end of the academic year. We are interested in the "population-average" improvement and in the differences in the mean improvement across the classes.

Bob Mislevy's comments on an earlier version of this paper are acknowledged.

First, the representation of the tested domain in the test form(s) or other measurement instruments used is bound to be imperfect because there are no well-established standards for assessment of the representation, and, in any case, the domain itself does not have an unambiguous definition. Reference to large sample size is out of place because for the relevant kind of "averaging," we need a large number of test forms. Next, instead of students' *abilities,* we can, at best, measure their *performances,* which are subject to temporal variation, and affected by motivation and other everyday influences. Further, each test has a finite length, and so it is associated with imperfect reliability, even after conditioning on performance. For comparing classes, we want to extract the *net contributions* of the instruction by teaching staff and the classroom/school environment (further uncertainties about what this means . . .); the pretest score is an important variable to condition on, but it is far from sufficient. In principle, additional conditioning (background) variables may isolate the classroom effects as certain adjusted differences. These background variables have to be such that the allocation of students to classes be conditionally noninformative. Unfortunately, it is rarely possible to assess how close we are to this state of affairs. Although each additional conditioning variable takes us closer to the noninformative assignment, it also contributes to model complexity and ill-conditioning, especially in modeling between-cluster differences.

Some of these problems would disappear if students were assigned to classes at random, or by a noninformative design. Such an allocation is unrealistic, though, and the impact of the realized allocation cannot be assessed. The elegant but dishonest approach we tend to exercise is to ignore this and other sources of uncertainty. Such dishonesty, or economy of integrity, catches up with us collectively when we realize that the conclusions of any study apply to an extremely narrow context, and that in a slightly different context, radically different conclusions have been arrived at using different measurement instruments, testing conditions, conventions for modeling, and software. Hutchison sounded the warning most eloquently in discussing the controversy over the teaching styles study of Bennett (1976):

> If the reading public interested in education, and in this I include politicians and administrators, as well as teachers and parents, are to become used to a pattern of publication of clear-cut results followed by their complete dismissal by some apparently equally eminent authority, then the credibility of any educational research and its statistical foundations will be, at least, very seriously eroded. (Hutchison, 1981, p. 443)

This criticism may invite a whole spectrum of responses. One extreme, exemplified by Freedman (1988), is to dismiss any results or conclusions of observational studies as of little practical relevance, referring to a too primitive description and incomplete understanding of the processes underlying the imperfectly recorded observations of imperfectly defined quantities. The other extreme, to ignore these problems and present an optimistic picture of

unequivocal conclusions, is strongly encouraged by the desire to deliver goods (reports, publications, or the like) and the hope, in most circumstances a realistic one, that their quality and integrity will escape close scrutiny.

I do not wish to stake out my position in this spectrum because I believe that any position is poorly informed. A way out can yet be found by the balance of a formal approach, of which multilevel methods are a component, and informal assessment of the sources of uncertainty associated with the features of the analyzed study that are not modeled formally. Although it is hard to admit to more uncertainty than what is indicated by the "correct" standard errors obtained using the "appropriate" software, we have to combat the prospect of uncertainty being calibrated by the reader ("Oh yes, the standard errors are quite small, but it is only a study"). Unbiased and efficient estimation of uncertainty about the quantity of interest is almost as important as efficient estimation of the quantity itself.

Inertia and the widespread but ill-conceived notion of statistical significance as the raison d'être of analysis in much of social science statistics stand in the way of integrity. How can an analyst admit to uncertainty and lack of significance when highly significant results are the norm in our field and imply good study design? In view of the considerable resources that were invested in the survey design and data collection, any mention of uncertainty may be met with incredulity. This reaction is, to a large extent, conditioned by the partisan and therefore questionably qualified justification for high expenditure on educational surveys.

Mislevy (in press) presents a critique of the studies comparing educational achievement across countries. Several points he raises, not all of them elaborated here, carry directly over to large-scale national or statewide surveys.

Longitudinal Data

When each subject $i = 1, \ldots, I$ is observed at times $j = 1, \ldots, J$, the observed vectors $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})$ are a random sample. Assuming normality, the distribution of this sample is fully described by its expectation $\boldsymbol{\mu} = E(\mathbf{y}_i)$ and variance matrix $\boldsymbol{\Sigma} = \text{var}(\mathbf{y}_i)$. Imposing no structure, that is, fitting the saturated model, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated straightforwardly:

$$\hat{\boldsymbol{\mu}} = \frac{1}{I}\sum_{i=1}^{I} \mathbf{y}_i$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{I-1}\sum_{i=1}^{I} (\mathbf{y}_i - \hat{\boldsymbol{\mu}})(\mathbf{y}_i - \hat{\boldsymbol{\mu}})^{\top}$$

with the sampling variance matrix for $\hat{\boldsymbol{\mu}}$ estimated by $I^{-1}\boldsymbol{\Sigma}$. Any structure on the mean $\boldsymbol{\mu}$ or the variance matrix $\boldsymbol{\Sigma}$ can be imposed by suitable averaging of the estimates from the saturated model; it is easy to see that, in essence,

this is what the Fisher scoring and generalized least squares iterations do. This approach is not wedded to covariance structures arising from multilevel models; structures arising in times series or graphical models can be fitted with equal ease, even in the presence of subject-level explanatory variables (see Longford, 1993, chap. 4).

NCES Questions

The questions posed by the National Center for Education Statistics (NCES) acknowledge the complexity of the hierarchical linear models and their software implementations. What Question 1 refers to as "forms" of the models are best perceived as conventions or ways of presenting the models. The user will do best service to the data by being acquainted with several forms, so that the models can be inspected and discussed from a variety of perspectives. This applies also to Question 2. Any reasonable model has a description in both forms; any lack of invariance should be viewed with utmost suspicion.

A particular software package may rely on one form, but that may be no reflection on the merit of the other forms or on the merit of the software. The software should not dictate the terms on which the analysis is to be performed, but rather should humbly serve the purpose of the analysis. Dedicated use of a single specialized package, even in a narrow range of applications, is a recipe for the software to impose on the analyst the questions to which it provides the answers.

The importance of the user's control over the data handling and estimation process is generally underrated. The most suitable software package is the one that affords most control, but that control cannot be effectively exercised without expertise. Some deficiencies in software can be compensated for, but, when dealing with complex problems, deficiencies in (statistical) expertise are much more difficult to overcome.

The "proportion of variation explained" is an anathema when we are unable to do any explanation. At best, we can identify some patterns in the data, but cannot infer directly (carry out a test) why these patterns arise. There are straightforward extensions of the R^2 familiar from ordinary regression. Simply, we compare the reduction of the variance component at each level due to the covariates; see Longford (1993, sec. 2.11) for details.

Software

Software packages for multilevel analysis are regarded by many users as embodiments of the implemented methodology. All too frequently I have been addressed the self-dismissing comment, "I am only a user ... who wants to do a correct analysis." Paying more attention to features of software packages than to features of methods implies predominance of technological aspects over theoretical ones in applications of statistics. I do not regard this as a desirable trend in social statistics because it promotes a trivial mechanical approach to statistical analysis.

References

Bennett, N. S. (1976). *Teaching styles and pupil progress.* London: Open Books.

Freedman, D. (1988). As others see us: A case study in path analysis. *Journal of Educational Statistics, 12,* 101–128.

Hutchison, D. (1981). Discussion of "Statistical modelling of data on teaching styles" by M. A. Aitkin, D. A. Anderson, & J. Hinde. *Journal of the Royal Statistical Society, Ser. A, 144,* 419–461.

Longford, N. T. (1993). *Random coefficient models.* Oxford, UK: Oxford University Press.

Mislevy, R. J. (in press). What can we learn from international assessments? *Educational Evaluation and Policy Analysis.*

Author

NICHOLAS T. LONGFORD is Senior Research Scientist, Educational Testing Service, 15-T, Rosedale Road, Princeton, NJ 08541; nlongford@ets.org. He specializes in multilevel models and inference about variation.

Journal of Educational and Behavioral Statistics
Summer 1995, Vol. 20, No. 2, pp. 210–220

Reexamining, Reaffirming, and Improving Application of Hierarchical Models

Stephen W. Raudenbush
Michigan State University

It has now been 14 years since Leigh Burstein's (1980) influential review of the profound inferential difficulties associated with prior quantitative inquiry on schools and classrooms. That far-ranging discussion focused on the mismatch between conventional statistical models and the realities under investigation. Models were single-level, founded on the naive assumption that persons respond independently to educational practices. This atomistic conception of social life, unsatisfying as a conceptual framework to guide inquiry, also produced a host of statistical difficulties: forced choice of unit of analysis; unnecessary trade-offs between misestimated precision and weak power; aggregation bias; and unexamined heterogeneity of regression. Burstein argued that more sensible statistical models were essential to progress in conceptualization, measurement, design, and analysis, and called for a comprehensive reform in statistical practice based on multilevel models. For similar views, see Cronbach and Webb (1975), Cronbach (1976), and Aitkin, Anderson, and Hinde (1981), among others.

This special issue of *JEBS* marks how far the field has advanced in the 14 years since Burstein's review. Motivating the special issue is a sea change in statistical analysis during that time:

- It is now routine to formulate coherent and quite general models for cross-sectional data having two levels (e.g., students within schools), three levels (students within classrooms and schools), longitudinal panel models, longitudinal models for students nested within social settings, and cross-classified models for cross-sectional data (e.g., students nested within neighborhood-by-school cells) or for longitudinal data (e.g., students migrating across social contexts such as classrooms).
- Efficient estimation procedures are now readily available and rapidly becoming computationally fast for each of the designs mentioned above, allowing for covariates at each level having fixed or random effects and unbalanced designs.
- Significant changes in streams of inquiry parallel—and are facilitated by—changes in modeling perspective. Rather than looking at schools as adding a constant to each student's knowledge, schools are viewed as modifying the entire social distribution of outcomes within them (e.g., Lee & Bryk, 1989), consistent with the best thinking in sociology of education (e.g., Bidwell & Kasarda, 1980; Barr & Dreeben, 1983). School effects are increasingly conceptualized as effects on learning

rates rather than on status, as evidenced by the planned Early Childhood Longitudinal Study sponsored by the National Center for Education Statistics.

- When the data are hierarchical, naive applications of marginal independence models (models assuming independence given only the fixed effects), which previously enjoyed a near monopoly, are no longer acceptable in the minds of journal reviewers in education, psychology, sociology, and allied fields. These reviewers now assume, as did Burstein in 1980, that we can do better. Now, in fact, we can.

The three articles before us can be only be appreciated or even understood in the context of the sea change described above. Rapid change must provoke critical reexamination. A novel approach can quickly become the new orthodoxy, drowning the critical spirit that produced it. It is a mark of the maturity of the multilevel movement that this time has come, and the spirit of reexamination is served well by the articles in this special issue. It is our loss that Leigh Burstein's untimely death has prevented his seeing this special issue and sharing his insights on it.

Statistical Inference and Scientific Judgment: A Response to Draper

We are indebted to David Draper for stimulating a needed discussion about the logical basis for statistical and scientific inference in social science. The topic of hierarchical models (HMs) serves as the occasion for this broader discussion; his criticism of particular applications using HMs could apply equally to any number of nonrandomized studies using convenience samples, regardless of analytic technique.

Methodologists differ in their beliefs about the requisite conditions for valid statistical inference. There are strict constructionists, who view statistical inference as meaningful only when a probability sample has been selected from a well-defined population. A second, broader perspective is that any collection of data is a sample from some population, and that, although the target population for inference remains tentative, statistical inference is useful. A third, Bayesian perspective views probability as subjective uncertainty about the process that produces data rather than relative frequency in a population. The notion of exchangeability (de Finetti, 1964) is often functionally equivalent to assuming the data arise from a simple random sample. The advantage is that exchangeability lays the basis for rational decision making under uncertainty even when no formal sampling mechanism is involved. In the Bayesian view, conclusions from empirical research combine new data with prior information to produce a new synthesis. If we adopt the essence of the Bayesian learning model in forming scientific judgments, we find that when the design of the study is weak, we must lean more heavily on prior information, including, for example, the stream of research of which the latest study is only one part. This learning model avoids unproductive dichoto-

mies (a scientific inference either is or is not justified), leaving a role for degrees of belief and scientific judgment.

Although Draper describes three levels of inference, when the rubber hits the road, he's a strict constructionist. Because Huttenlocher et al.'s (1991) study involves a convenience sample of mother-infant pairs and because random assignment of mothers to speech conditions was not employed, Draper writes, "I find no scientific meaning in the parameter estimates and *SE*s Huttenlocher et al. report." Elsewhere, he makes clear that he expects uncertainty about scientific conclusions to be quantified by a confidence interval (he criticizes a classroom study for failing to modify a confidence interval to reflect "uncertainty about the validity of causal inferences").

Although my own applications of hierarchical models and those of Anthony Bryk have mostly involved probability samples from well-defined populations, we both tend to be broad constructionists. To me, a statement that two groups "differ significantly" ($p < \alpha$) is a statement about the probability of obtaining a difference of a given magnitude between two groups randomly sampled from the same population. The statement that two groups "differ significantly" never *by itself* implies that this difference generalizes to a defined population, nor, contrary to Draper's assertion, should the finding of a statistically significant difference between groups by itself imply the existence of a causal relationship. A small p value or a short confidence interval (a *statistical* inference) can only supply one piece of evidence in favor of a particular *scientific* inference. Random sampling and random assignment strengthen the case. However, I believe there are many examples in educational work where intensive investigation of teaching and learning processes, even though based on convenience samples, has substantial scientific merit. Given limited resources for research, there is often a trade-off between a concern for nationally representative sampling and the concern for intensive measurement. It is too facile to say that sampling is always primary in science.

Statisticians like to think that causal inference and causal generalization can usually or should always be based on methodological grounds alone. I doubt, however, that science works this way most of the time. Rather, causal assertions, such as the assertion that smoking causes lung cancer, likely require a case based on a web of evidence from a variety of sources interpreted in the light of the best available theory. In the case of smoking, evidence has accumulated over time: correlational evidence on humans; experimental evidence on animals; mechanistic evidence based, for example, on the lung tissues of smokers and nonsmokers. Over time, the burden of proof gradually shifted from the proponents to the skeptics of the assertion that smoking caused lung cancer. The skeptics could invent alternative theories for the elevated incidence of lung cancer among smokers, but their theories were less plausible and fit less well with the entire body of evidence (which was undoubtedly made up of individually flawed studies). The case in favor of

the causal inference was more coherent theoretically and fit better with relevant streams of research.

To make an analogy with smoking and lung cancer, many of the best educational researchers are now doing research that is more like looking at lung tissue than like computing correlations between smoking and lung cancer or conducting smoking experiments on animals. When I raise with my colleagues in educational psychology the arguments Draper raises about the importance of representativeness (which I do frequently!), they often say that to study classes, kids, teachers, and processes in the depth they feel is necessary is difficult or impossible in a large probability sample.

In defense of my colleagues, I doubt if too many cell biologists take a random sample of lung cells from a nationally representative sample of citizens in order to study the mechanisms by which smoking putatively affects the probability of lung cancer. Yet, such mechanistic research has apparently been decisive in creating a consensus among experts about this causal linkage and its generalization.

We cannot wait for the perfect social science study that randomly selects subjects from a large and well-defined population and then randomly assigns subjects to treatments, employing valid measures to allow unquestionable and generalizable causal inferences. Such a vision is useful to promote better research practice, but not as a description of how social science or epidemiology has typically advanced or will typically advance (see Kuhn's [1962] discussion of the myth of the single decisive study). To adopt such rigid expectations will inevitably lead to discounting the value of each individual study, undermining the possibility of recognizing contributions of streams of necessarily imperfect social research.

Let us apply the broad constructionist view to the Huttenlocher et al. (1991) study. Although Draper, a statistician, "finds no scientific meaning" in the relationship between maternal speech and acceleration of language development found in this study, Huttenlocher, a developmental psychologist intimately familiar with theory and prior research in the area, does find scientific meaning, as do the reviewers of *Developmental Psychology.* Draper criticizes the study for failing to control nonverbal communication, but Huttenlocher and the reviewers apparently find no prior theory or evidence that such nonverbal communication is related to vocabulary development, and perhaps, given the small size of the sample, worry about overfitting the model. Draper sees as too unqualified Huttenlocher et al.'s assertion that "The present study provides the first direct evidence that amount of exposure [to maternal speech] is important to vocabulary growth"; he is perhaps concerned about the use of the word "important." In contrast, Huttenlocher et al. emphasize the word "first" in that sentence (Bryk, personal communication); they know that it will take a number of replication studies on diverse samples to strengthen confidence in its generalizability. Although I admittedly have adopted the voice of the scientist in this dialogue with the statistician (how else would

that voice be heard here?), I view the dialogue as useful. Textbook discussions of design do not end scientific disputes.

Choice of model (HM versus other) is orthogonal to one's stance as a strict versus broad constructionist; whatever the stance, one might love or hate HMs for a given study. The utility of HMs to summarize evidence in the "Cognitive Strategies in Writing" example depends on their capacity to produce a more precise estimate of an adjusted mean difference (than would be found in an aggregated analysis) with an honest confidence interval (as compared to a conventional student-level analysis assuming marginal independence). Whether such a statistical inference should be interpreted causally depends on scientific judgment which must be informed by an elaborate set of analyses and substantive considerations that have little to do with the modeling issues at hand in chapter 5 of Bryk and Raudenbush (1992). To set the record straight, no causal inference was made in the original exposition: A statement that one group of children scored higher than another is not a causal inference; such an inference requires an *explanation* of the observed difference between the two groups and not merely a recognition of its *existence.*

I found the last four sections of Draper's paper quite satisfying. Partly this response reflects self-gratification, in that Raudenbush and Willms (in press) have proposed a way of thinking about the estimation of school effects that I view as entirely consistent with Draper's section on "The Use of HMs in School Effectiveness Studies," while his section "Hierarchical Models and Meta-Analysis" is consistent with my past work on meta-analysis. Moreover, the algorithmic work Draper proposes (e.g., finding sensible combinations of EM and Fisher scoring/IGLS[1]) is well under way in a beta version of HLM. Bayesian estimation via Gibbs sampling is especially appealing when there are small numbers of higher-level units. An example is Seltzer's (1993) reanalysis of the Huttenlocher et al. data with a *t*-prior with 5 *df.* (By the way, this analysis gave *stronger* evidence of a link between maternal speech and vocabulary acceleration than was found in the original analysis). However, the Bayesian approach runs somewhat counter to the spirit of de Leeuw and Kreft's advice, to which I now turn.

de Leeuw and Kreft: Choice of Models, Methods, Algorithms, Software—and Interpretations

This article provides an exceptionally lucid response to commonly asked questions about HMs for multilevel analysis. Its essential strength is in separating choice of model from choice of estimation method, choice of method from choice of algorithm, and choice of algorithm from choice of software. It is essential for analysts to understand that each choice involves a wide set of options—wider than many have realized. Increasing the recognition of this fact alone makes the article useful. The authors adroitly explain how issues of interpretation are embedded in modeling, in particular, how

the "two-step" approach to model specification can enhance understanding of the model parameters.

I wish, however, that the article had clarified the final sentence in the abstract. Even if the researcher is interested only in the fixed regression coefficients of studies having two-level designs with large groups and small intraclass correlations (a setting that seems exceptional in educational research), it is not clear to me which "traditional techniques perform as well or better" than "multilevel models." In fact, the comparison of "traditional techniques" and "multilevel models" seems to violate the useful distinction between choice of model and choice of method of estimation. I trust that by elaborating on this concern I will not conceal my overriding applause for the article.

Under the now standard model of de Leeuw and Kreft's Equations 18 and 19, the ordinary least squares (OLS) estimator of their Equation 22 is a consistent (though not efficient) estimator of $\boldsymbol{\gamma}$. As the authors mention, it is trivial to compute consistent standard errors for their Equation 24 as

$$\text{Est. Var}(\hat{\boldsymbol{\gamma}}) = \left(\sum_{j=1}^{m} \mathbf{Z}_j^T \mathbf{Z}_j\right)^{-1} \times \sum_{j=1}^{J} \mathbf{Z}_j^T(\hat{\boldsymbol{\beta}}_j - \mathbf{Z}_j\hat{\boldsymbol{\gamma}})(\hat{\boldsymbol{\beta}}_j - \mathbf{Z}_j\hat{\boldsymbol{\gamma}})^T\mathbf{Z}_j\left(\sum_{j=1}^{m} \mathbf{Z}_j^T\mathbf{Z}_j\right)^{-1}. \quad (1)$$

This approach combines OLS point estimates with a robust "Huber-corrected" sampling variance estimate. However, following Zeger, Liang, and Albert (1988), we can compute the generalized least squares (GLS) estimate of Equation 26, or, equivalently in the case of full-rank data, Equation 27, based on efficient variance-covariance estimates, and then compute robust sampling variances:

$$\text{Est. Var}(\hat{\boldsymbol{\gamma}}) = \left(\sum_{j=1}^{m} \mathbf{Z}_j^T \hat{\mathbf{W}}_j^{-1}\mathbf{Z}_j\right)^{-1} \times \sum_{j=1}^{J} \mathbf{Z}_j^T \hat{\mathbf{W}}_j^{-1}(\hat{\boldsymbol{\beta}}_j - \mathbf{Z}_j\hat{\boldsymbol{\gamma}})\,(\hat{\boldsymbol{\beta}}_j - \mathbf{Z}_j\hat{\boldsymbol{\gamma}})^T\hat{\mathbf{W}}_j^{-1}\mathbf{Z}_j \times \left(\sum_{j=1}^{m} \mathbf{Z}_j^T \hat{\mathbf{W}}_j^{-1}\mathbf{Z}_j\right)^{-1}. \quad (2)$$

This approach also provides robust standard errors, that is, standard errors that are insensitive to assumptions about the covariance structure. However, when model assumptions are sensible, the approach yields asymptotically efficient point estimates of the fixed effects and covariance components along

with empirical Bayes estimates of β_j for each unit j, and readily extends to the rank deficient case and to three-level or cross-classified structures. The only price to pay is more intensive computation, a small price given increasingly efficient hardware and software and the asymptotic superiority and generality of generalized least squares with variance estimates (2) as compared to OLS with variance estimates (1).

It should be emphasized that even Equation 1 is based on a "multilevel model," specifically, a two-level structure. Robust standard errors for OLS estimates in a three-level setting would require a different algorithm. Thus, the estimates are not robust to misspecification of the number of levels in the structure. This should be emphasized lest readers view the endorsement of "traditional techniques" or "unweighted least squares" as support for marginal independence models.

Finally, a note on algorithms: HLM uses the Aitken accelerator (Laird, Lange, & Stram, 1987) to speed convergence. A beta version now uses Fisher scoring and EM with good results to combine the advantages these authors have skillfully identified. Apparently there are no "older" packages with which the new packages can compete, just old labels for ever-changing programs.

Rogosa and Saner: Demystifying the Demystification

Rogosa and Saner describe as "their main expository purpose" to "demystify" HM analyses of longitudinal panel data by comparing results obtained from the HLM program to those obtained from simpler approaches using comparatively simple examples. Extensive reanalyses of balanced data sets using straight-line growth models reveal that HLM gives identical results to those obtained in a two-step OLS analysis. The authors find this equivalence "surprising if not disconcerting," but I find it neither.

It is well-known that restricted maximum likelihood (REML), the method of estimation used in HLM, duplicates the standard mixed-model ANOVA results for the classical balanced experimental designs. Raudenbush (1993b) shows equivalence in the case of one-way random effects, the two-factor mixed hierarchical design (e.g., students within classes within treatments), randomized blocks design (with repeated measures on students or longitudinal panel models as a special case), and the mixed model for two-way cross-classification (e.g., children within treatments implemented at each of many day-care centers). The article shows how estimates and exact t or F tests for fixed effects and variance parameters can be recovered from the HLM output. Canonical examples are those simple data sets in the chapters of Kirk (1982) corresponding to these research designs. Complete raw data, classical ANOVA results, and HLM results are presented with detailed specification of how to reproduce the classical results using the more general approach. Similarly, the three-level model allows analysis for the three-factor hierarchical design and several more complex split-plot designs (e.g., persons changing over time and nested within classrooms that are, in turn, nested within treatments).

Crossed random effects models (e.g., Raudenbush, 1993a) map to a variety of designs having cross-classified random factors. Goldstein (1987) has shown how HMs can duplicate standard multivariate results.

Thus, HMs based on REML duplicate the familiar balanced data results for the classical designs while facilitating generalization to the more complex data characteristic of educational field research having unbalanced designs, covariates at each level (e.g., time-varying predictors), and continuous and discrete responses. These models therefore combine the virtues of the experimental design literature (which emphasizes the need to understand sources of variation in data and how these sources affect inference) with the key virtue of the general linear model (flexibility in incorporating continuous and discrete covariates in linear models).

Features of Rogosa and Saner's article that I found most valuable included an emphasis on description, graphical display, and model checking using simple diagnostics. Although these have been underemphasized in many methodological discussions, the HLM program includes a residual file that can be used for a variety of model-checking and data exploration procedures described in detail in chapter 9 of Bryk and Raudenbush (1992) and in the current HLM manual (which uses data from High School and Beyond rather than the "rat data"). Included in this file are least squares equations for each unit having full-rank data and empirical Bayes equations for all units. Rogosa and Saner's emphasis on better uncertainty estimation for variance components and random effects is also well placed. Large-sample standard errors based on the information matrix are least useful when such estimates are most needed—when the number of higher-level units is small. The bootstrap as implemented in Timepath appears to be an attractive alternative for balanced data. However, for unbalanced designs, resampling must be multilevel and is computer-intensive (Laird & Louis, 1987; see Bagakas, 1992, and Raudenbush & Willms, in press, for educational applications). Bayes estimation via Gibbs sampling (e.g., Seltzer, 1993) provides posterior distributions for parameters and functions thereof. However, the approach imposes new assumptions in the form of prior distributions, which may not be friendly to Rogosa and Saner's perspective.

There are many specific issues raised in Rogosa and Saner's article to which I could respond in detail.[2] Conditional reliability estimates and correlations between random effects, for example, have clear meanings (see Raudenbush & Bryk, 1985, p. 66, with application to studies of school differences controlling demographic background). Space limitations forbid such a detailed response. It seems more useful in any case to focus in the future on the useful issues Rogosa and Saner raise concerning description, model checking, and uncertainty estimation for variance-covariance components. In pursuing these, simulations and model comparisons are likely to be most useful when the data are unbalanced and the number of higher-level units small.

Morris's Overview

Carl Morris has made seminal contributions to the theory of hierarchical models and his comments on applications of these models in education are most helpful. While I agree entirely with his recommendations, I offer a few qualifying remarks designed to discourage readers from overgeneralizing his criticism of maximum likelihood (ML).

Morris correctly points out that when the number of Level 2 units, k, is small, inferences about individual effects, θ_i, and Level 2 regression coefficients, β, conditional on ML point estimates of the Level 2 variance, τ^2, can be misleading. He reanalyzes data from Raudenbush and Bryk (1985) to illustrate this point and recommends a Bayesian approach that fully takes into account uncertainty about τ^2. I certainly agree with this analysis and view it as entirely consistent in spirit with the purpose of Raudenbush and Bryk (1985). Rather than employing a Bayesian analysis, that article plotted the likelihoods Morris plots in his Figure 1[3] and examined the sensitivity of all inferences to likely errors of estimate of τ^2. Our conclusion was that inferences about the Level 2 regression coefficients were less sensitive than inferences about the individual effects and relatively insensitive after the covariate was added. A detailed discussion of this problem along with recommendations similar to those of Morris appears in Bryk and Raudenbush (1992, pp. 220–222). Does this imply that maximum likelihood should never be used when k is small?

The answer, in my opinion, is no. When the data are balanced, REML estimates for hierarchical designs duplicate the classical ANOVA results, giving exact F tests for the βs and for τ^2. When the data are unbalanced, the sensitivity of inferences about β to errors of estimation of τ^2 will depend on the degree of imbalance. This sensitivity is large for the teacher expectancy data in the case of no covariates not only because of the uncertainty about τ^2 but also because of the radical imbalance in Level 1 variance across the 19 studies under synthesis. My comment is not in disagreement with Morris's view, which primarily concerned inferences about the individual effects, θ_i. It is, rather, an elaboration of the conditions under which REML will mislead with respect to the Level 2 coefficients, β, the focus of interest in many educational studies. Fotiu (1989) has provided a simulation study comparing REML and Bayes estimation via Gibbs sampling and shown that inferences about β estimated via REML and based on a t reference distribution are quite robust unless the imbalance is very pronounced. Raudenbush, Cheong, and Fotiu (1994) compare REML and Bayes estimates in cross-national comparisons of reading literacy.

A final word on software: Knowing that the HLM software had been thoroughly evaluated in simulation studies by Bassiri (1988) and Fotiu (1989), and having been involved in extensively checking HLM3.0, I was concerned to read about Morris's experience with HLM3.0. An inquiry with Scientific

Software (SSI) revealed that the errors he described were introduced when SSI ported HLM3.0 to the Sun Workstation. Flawed versions of HLM were distributed to four Sun users, all of whom were subsequently sent HLM3.01, which corrected the flaw. The much more widely used PC version was not affected. This experience reinforces Morris's recommendations for the most painstaking and extensive testing of software even after seemingly minor modifications. I apologize to him and anyone else inconvenienced by this error.

Notes

[1]Fisher scoring can be shown to be mathematically equivalent to iterative generalized least squares (IGLS) in the case of normal data and normal random effects (Raudenbush, 1994).

[2]For example, Rogosa has claimed for years to have an early version of HLM that produced "wild results," but, despite repeated attempts at correspondence, I have not been able to obtain this version of the program or the data he analyzed. Nor have their been any data-destroying fires or reported fires in the Raudenbush or Bryk residences.

[3]These likelihoods are equivilant to posteriors if τ^2 is a priori uniform on the nonnegative real line.

References

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modeling of data on teaching styles. *Journal of the Royal Statistical Society, A144,* 419–461.

Bagakas, J. G. (1992). *Two level nested hierarchical linear model with random intercepts via the bootstrap.* Unpublished doctoral dissertation, Michigan State University, East Lansing.

Barr, R., & Dreeben, R. (1983). *How schools work.* Chicago: University of Chicago Press.

Bassiri, D. (1988). *Large and small sample properties of maximum likelihood estimates for the hierarchical linear model.* Unpublished doctoral dissertation, Michigan State University, East Lansing.

Bidwell, C., & Kasarda, J. (1980). Conceptualizing and measuring the effects of school and schooling. *American Journal of Education, 88,* 401–430.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods.* Newbury Park, CA: Sage.

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

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design and analysis.* Occasional paper of the Stanford Evaluation Consortium, Stanford University.

Cronbach, L. J., & Webb, N. (1975). Between and within-class effects in a reported aptitude-by-treatment interaction: Reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology, 6,* 717–724.

de Finetti, B. (1964). Foresight: its logical laws, its subjective sources. In H. E. Kyburg, Jr., & H. E. Smokler (Eds.), *Studies in subjective probability* (93–158). New York: Wiley.

Fotiu, P. R. (1989). *A comparison of the EM and data augmentation algorithms on simulated small sample hierarchical data from research on education.* Unpublished doctoral dissertation, Michigan State University, East Lansing.

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

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27,* 236–248.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences.* Belmont, CA: Wadsworth.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Laird, N., Lange, N., & Stram, D. (1987). Maximum likelihood computation with repeated measures: Application of the EM algorithm. *Journal of the American Statistical Association, 82,* 97–105.

Laird, N. M., & Louis, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association, 82,* 739–756.

Lee, V., & Bryk, A. S. (1989). A multilevel model of the social distribution of educational achievement. *Sociology of Education, 62,* 172–192.

Raudenbush, S. W. (1993a). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics, 18,* 321–349.

Raudenbush, S. W. (1993b). Hierarchical linear models as generalizations of certain common experimental design models. In L. Edwards (Ed.), *Applied analysis of variance in behavioral science* (459–496). New York: Marcel Dekker.

Raudenbush, S. W. (1994). *Equivalence of Fisher scoring to iterative generalized least squares in the normal case with application to hierarchical linear models.* Unpublished manuscript, Michigan State University College of Education, Program on Measurement and Quantitative Methods.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10,* 75–98.

Raudenbush, S. W., Cheong, Y. F., & Fotiu, P. R. (1994). Synthesizing cross-national classroom effects data: Alternative models and methods. In M. Binkley, K. Rust, & M. Winglee (Eds.), *Methodological issues in comparative international studies: The case of reading literacy.* Washington, DC: National Center for Educational Statistics.

Raudenbush, S. W., & Willms, J. D. (in press). The estimation of school effects. *Journal of Educational and Behavioral Statistics.*

Seltzer, M. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational Statistics, 18,* 207–235.

Zeger, S. L., Liang, K-Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics, 44,* 1049–1060.

Author

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

Journal of Educational and Behavioral Statistics
Summer 1995, Vol. 20, No. 2, pp. 221–227

Comment

William M. Mason
University of California, Los Angeles

My remarks are stimulated by the insightful, informative and rewarding papers of de Leeuw and Kreft, Draper, and Rogosa and Saner.

From a sociology of knowledge perspective, the intellectual history of the introduction of multilevel models appears to be running a predictable course: (1) A new statistical formulation is introduced. (2) The new formulation seems to answer a widely perceived need. (3) It is enthusiastically developed and promulgated as The Answer. (4) In response, journeyman practitioners accord the new methodology high status. Graduate students learn some version of it. So do young professors. Older ones consider how close they are to retirement before deciding how much to invest in it. Then comes the reaction: (5) It is said that the new method is like some other, previously used method in various ways. (6) Critics begin to ask, "Do we really get different answers (from the way we used to do it, and already understand)?" Typically, for any particular methodology, the answer is in some cases no, and in some, yes. In all instances, though, the new methodology has a generality that previous approaches lacked. Moreover, it provides a new way of thinking about problems. (7) Nevertheless, disaffection begins to set in. The assumptions are not always met and often seem unjustifiable in the context of a particular problem. Or perhaps the approach becomes identified with a particular school of substantive thought, and that perspective is found wanting. (8) By now, the high-water mark has been reached and the tide of enthusiasm begins to ebb. (9) But the ark comes to rest somewhere on the side of the mountain, at an elevation that differs from its former resting place, because, in fact, the new methodology has something genuine to offer, and because it continues to evolve as generalizations, extensions, and refinements of understanding are offered. It is not just the same old thing in a new package. Thus, eventually, the new methodology finds a place in the social science armamentarium.

The process I describe is impressionistically derived, but it seems to describe fairly well what has happened to many of the heralded statistical modeling innovations I have witnessed over the past 30 years as a social scientist: "regression analysis" (as it was so labeled and taught when I was a graduate student), path analysis, structural equation models more generally, the generalization to structural equation models with latent variables, loglinear models, exploratory data analysis, and, most recently, survival models with heterogeneity and multilevel models with covariance components.

De Leeuw and Kreft's contribution, and Rogosa and Saner's, suggest to me that multilevel analysis is at Stage 7 in the absorption process for a

statistical innovation. These highly knowledgeable authors, in their own unique styles, help to dampen unrestrained enthusiasm while providing insight. Rogosa and Saner demonstrate that under certain circumstances (not enumerated analytically in their paper), you can get the same answers using the HLM software package and using other kinds of software. They also point out the usefulness of "common sense" approaches to obtain checks on results obtained from programs that perform covariance component computations. It would be incorrect to assert that fixed-effect and random effect (covariance component) modeling always yield the exact same result, and I do not read Rogosa and Saner to be making that claim.

De Leeuw and Kreft argue that for many purposes, it suffices *not* to estimate a covariance component model: A fixed effects model will do. In addition, they suggest, not without evidence, that relatively simple computations will often suffice, even when one does wish to estimate a covariance component model. (De Leeuw and Kreft's comments about the EM algorithm are much appreciated. Having developed two programs that employ it, I concur.)

They argue against routine acceptance of shrinkage estimates obtained by multilevel programs that employ empirical Bayes computations. Again, the point would be to compare the shrinkage estimates not only with the macro estimates but also with the within-context estimates. For each context, there are three estimates available. We should not be blinded by the technical sophistication of a stochastic parameter formulation into uncritically accepting the empirical Bayes estimates—which depend on the macro model. Moreover, it is not enough to argue that the systematic component of the macro model is usually "good enough," because it may not be, and because the assumption that the macro errors are Gaussian may be false (when the Gaussian assumption is made, as it typically is). Other assumptions, such as independence of errors, may also be violated.

De Leeuw and Kreft are also concerned about the realism of the usual multilevel model with Gaussian errors. They focus in particular on the fixed-X assumption. This assumption is often reasonable. The story we tell is that under repeated sampling, we would fix the covariate combinations. The position I would advocate is that we in fact make or relax the fixed-X assumption rather flexibly. If, following de Leeuw and Kreft, we are willing to be flexible about the presence or absence of covariance components, why not with respect to the stochasticity of the covariates themselves? In any case, regression models with stochastic regressors and nonspherical disturbances are not exactly unknown models.

David Draper's exceptional paper is like a luscious Christmas cake, crammed full of goodies that literally fall out on the way to their ultimate destination. Where to start? I want to air a couple of ruminations stimulated by this paper. First, "Why not fully Bayesian HMs?" That was actually our starting point (Lindley & Smith, 1972). It is a safe bet that in the next 10 years, as Draper prophesies, there will be fully Bayesian multilevel software

packages—a safe bet because some researchers are already doing Bayesian multilevel analysis. But who in the substantive research community (as distinguished from professional statisticians) will be using this software? Unless journeyman practitioners in the social sciences become knowledgeable of this approach, they will either be unable to use Bayesian software, or they will use it blindly. Will the social sciences then be better off than they are now? The emergent revolution in the teaching of statistics in the past few years has been the emphasis on interactive, analytic graphics and on careful description. Examination of commonly used statistics, biometrics, and econometrics texts does not suggest that a new cohort of Bayesian scientists will soon arrive. The issue is not whether Ph.D. statisticians and a handful of statistically sophisticated social scientists are or can be Bayesian in the orientation and practice of their research. The issue is how to change what is done by practitioners, individuals who have had several statistics courses at most, and none with substantial Bayesian content.

How to change practice by changing what practitioners do can be thought of as a "supply side" problem, to borrow from the recent argot of pop political-economy. There is also a "demand side." Another musing stimulated by David Draper's paper concerns the demand side of how we "do" and report statistical inference. Two of the coauthors of the Huttenlocher et al. (1991) article discussed by Draper are none other than Anthony Bryk and Michael Seltzer, who understand more than a little about statistical inference. Although these authors will have their own explanation(s) of their inference strategy, it is not hard to guess what the editorial response would likely have been had they eschewed standard errors, or indeed done anything unconventional with respect to statistical inference, despite the self-selected nature of the sample they used. At the end of the day, reviewers and editors demand asterisks next to coefficients. Right or wrong, they want p values. Although this can be changed, the process will be slow. Those responsible for teaching statistics courses for tomorrow's reviewers and editors have their marching orders.

As useful and helpful as these papers are, and as the RAND conference itself was, those who focus on educational research might benefit from the recognition that hierarchically structured data are seen in most, if not all, disciplines. Specialized literatures have developed that deal with seemingly specific problems using discipline-specific vocabulary and particular bundles of tools. Recognizing that, to continue the debate over which algorithm to use in the Gaussian case would appear to be of second-order importance. We know how to obtain reasonable answers with Gaussian errors, relatively large numbers of observations per context, and relatively large numbers of contexts. Change any one of those conditions and our knowledge is not so complete, although there is a lot of research activity concerning the tools. Grazing our way through several literatures, here is a (nonrandom) sampling of points and questions that perhaps bear consideration:

(a) Can we develop intuition and lore for non-Gaussian cases when the n_j are relatively large and J is relatively large? My own experience suggests that complex estimation (e.g., Wong & Mason, 1985, 1991) may be unnecessary in this case. A two-step estimated generalized least squares (EGLS) approach has not been formally derived for the generalized linear model, for the large n_j, large J case, but how badly does naive two-step EGLS perform compared to more complex alternatives? And where does it break down?

(b) The hardest case is that of small n_j and large J, with non-Gaussian errors. Here there is a great deal of literature, but no closure. Much of the work is under the rubric of panel data, but the data structures are inherently multilevel, even if the vocabulary and concerns of the statistical developers are not (e.g., Hsiao, 1986).

(c) Survival models—regression models of time to event—are widely used in the social and biological sciences. It is possible with current software to estimate what amounts to a survival model with random intercept using data with a hierarchical structure (Yates, Yi, Honore, & Walker, 1987). For example, one can estimate infant mortality allowing for between-family "heterogeneity," and it is possible to carry out the estimation making no parametric assumption about the form of the heterogeneity, as well as to assume that it is Gaussian. Doing this requires that a substantial percentage of the families analyzed have more than one child, else the multilevel structure is lost and the heterogeneity becomes unidentified (in which case the value of allowing for it is the subject of contention). And although the results in the literature are mixed, a judicious conclusion is that currently it cannot be said that ignoring heterogeneity generally yields answers that are the same as those obtained when heterogeneity is allowed for. Stochastic covariate coefficients, a theoretical possibility, are of interest, and work on this case is under way.

(d) How should we deal with situations in which the first-level outcome and a second-level potential regressor are jointly endogenous? This happens frequently. For example: (a) Many parents consider the school system, and particular schools, when choosing where to live. So, although "good parents" with "good kids" make for good schools, it is also true that good schools attract good parents and good kids. (b) Government officials may choose to place health clinics or family planning clinics based on knowledge of local conditions. If clinics are placed where they are most needed, and if the resources to run the clinics are similarly distributed, then cross-sectional analysis will reveal that where investment is greatest, outcomes are least favorable. Here the issue is one of modeling, not estimation and not algorithms (e.g., Frankenberg, 1992; Rosenzweig & Wolpin, 1982; Pitt, Rosenzweig, & Gibbons, 1993). (c) The same kind of point arises in assessment of the impact of parochial schools. Parents who value a particular kind of outcome and environment elect to send their children to parochial schools. Unless the endogeneity of this social process is modeled, hierarchical models that merely include a handful of "gross" parental characteristics as covariates will provide

potentially severely biased estimates of school effects. In economics, the contemporary solution to this problem typically involves the use of instrumental variables to identify the coefficients of particular equations. When this is done, the solutions adopted often preclude the use of a standard multilevel model, because contextual variables end up being used as instruments. Some thought needs to be given to how problems that are normally conceived of as involving cross-level simultaneity can be dealt with in a multilevel perspective.

(e) In economics, at least, fixed-effect approaches are regarded as productive (e.g., Chamberlain, 1980). For example, one way to deal with the endogeneity of clinic placement is to use two (or more) waves of panel data. Under assumptions about error covariance structure that many regard as reasonable, differences across waves can solve an identification problem engendered by endogeneity and, at the very least, provide "good" estimates of first-level covariates. This approach can be thought of as applying to the case of small n_j and large J, and it can be used with binary response variables.

(f) Thoughtful scholars in several disciplines are beginning to question uncritical acceptance of context as defined by the hierarchical structure of particular data sets. Owing to my own lack of knowledge of intellectual ferment among those who do educational research, I do not know whether this is a concern in the study of school effects. But it seems reasonable to ask whether individuals in a given putative context all share the same definition of context, whether there are multiple, overlapping contexts of relevance, and whether the impact of context is distributed evenly (or in some other way) across individuals. Answers to these questions may be highly substance driven.

(g) In economics, if not in other disciplines, assumptions about the error structure are taken relatively seriously. This can help to cast new and informative light on at least one debate that multilivel modelers find consuming: whether to center the covariates within groups. In particular, suppose that on substantive grounds you cannot defend the assumption that the micro errors are uncorrelated with the macro covariates. Then one possible solution is to carry out within-group differencing of the micro covariates, thus rendering the micro disturbances orthogonal to the macro covariates. Economists think of this differencing as implementing a fixed-effects approach. But differencing can also be fit into the multilevel perspective as a way of meeting one of the basic assumptions. If you obtain different substantive conclusions depending on whether you within-group centered (and given appropriate respecification of the intercept equation in the presence of within-group centering), you may be staring at evidence of one kind of specification error: the micro errors may be correlated with the macro variables.

None of my remarks should be taken as "anti" multilevel modeling. We should use modeling strategies when and where we think they are advanta-

geous. Multilevel modeling is not a universal tool, but it can be helpful. And when is that?

(a) In my own work I find it liberating to be able to write coefficients as response variables and know that, in principle, it is possible to treat them as such in a coherent statistical framework. In theoretical development, I find it useful to derive hypotheses about coefficients sometimes using the macro equations, and sometimes using the combined mixed-model equation. I have not been able to pin down anything systematic here. Sometimes hypotheses about the partial derivatives seem more straightforwardly derived one way than the other.

(b) I find it helpful to understand the strength of the contribution of macro variables on micro outcomes through intercept and coefficient variability. If this sounds faintly like an interest in correlation, it is! Although the primary usefulness of regression depends on interpretation of regression coefficients (e.g., Kendall & Stuart, 1973), like many others I still find correlation measures useful. (Perhaps the Human Genome Project will reveal the reason why.) Moreover, I find it useful to obtain estimates of the impact of macro variables on intercepts and slopes adjusted for their sampling variances—something you can't do with ordinary least squares (OLS), and something you can't do by plugging macro variables into micro equations estimated by OLS in the case of Gaussian errors.

(c) And finally, I find it useful to work, where feasible and appropriate, in a framework of considerable generality. Even if the covariance components are not my primary interest, I still want to estimate them, because they do provide information about the problem I wish to model.

References

Chamberlain, G. (1980). The analysis of covariance with qualitative data. *Review of Economic Studies, 47,* 225–238.

Frankenberg, E. (1992). *Infant and early childhood mortality in Indonesia: The impact of access to health facilities and other community characteristics on mortality risks.* Unpublished doctoral dissertation, University of Pennsylvania.

Hsiao, C. (1986). *Analysis of panel data.* Cambridge, UK: Cambridge University Press.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27,* 236–248.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B, 34,* 1–41.

Kendall, M., & Stuart, A. (1973). *The advanced theory of statistics* (3rd ed, Vol. 2). London: Charles Griffin.

Pitt, M., Rosenzweig, M., & Gibbons, D. (1993). The determinants and consequences of the placement of government programs in Indonesia. *The World Bank Economic Review, 7,* 319–348.

Rosenzweig, M., & Wolpin, K. (1982). Governmental interventions and household behavior in a developing country: Anticipating the unanticipated consequences of social programs. *Journal of Development Economics, 10,* 209–226.

Wong, G. Y., & Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association, 80,* 513–524.

Wong, G. Y., & Mason, W. M. (1991). Contextually specific effects and other generalizations of the hierarchical linear model for comparative analysis. *Journal of the American Statistical Association, 86,* 487–503.

Yates, G., Yi, K-M., Honore, B., & Walker, J. (1987). CTM: A program for the estimation and testing of continuous time multi-state multi-spell models [Computer software]. Chicago: NORC.

Author

WILLIAM M. MASON is Professor, Department of Sociology, UCLA, 405 Hilgard Ave., Los Angeles, CA 90024-1551; mason@soc.sscnet.ucla.edu. He specializes in demography and quantitative methodology.

Journal of Educational and Behavioral Statistics
Summer 1995, Vol. 20, No. 2, pp. 228–233

Rejoinder

David Draper
University of Bath

I am grateful to the discussants for their interesting comments. With the exception of some of Raudenbush's remarks, the only real disagreement I might have with the discussion is something that did not come up in it: Nobody seemed interested in pursuing the theme of prediction of observables. Maybe in the social sciences we are not ready to come face to face with the likely low quality of many of our individual-level predictions, even though policy interventions succeed or fail one person at a time, and in spite of the clear evidence from other sciences that the pace of learning quickens when inaccurate predictions of important phenomena come to light. The general movement toward Bayesian analyses of multilevel models I have recommended here will help to put better predictive tools in practitioners' hands. With respect to software (and many other things, too), people generally do what is easy (cf. the third and fourth sections of Longford's comment), and supplying options in the HM packages to generate predictive distributions should go some distance toward making prediction of observables more routine in multilevel work.

On other topics, Mason is certainly correct that if Bayesian methods are to be used both widely and responsibly in the social sciences in the long-range future, there will have to be a sharp increase in the coverage of Bayesian techniques and outlook in statistics teaching at all levels, both inside and outside statistics departments, in the short and medium terms. To borrow his market-forces analogy, perhaps both an effort to move in this direction pedagogically and continued attempts to publish good Bayesian applied papers in social science journals (e.g., Seltzer, 1993)—papers which solve problems that are difficult to crack with other approaches—are needed, so that readers of such journals will begin to ask for more curriculum coverage of Bayesian topics.

One point of clarification with respect to Goldstein's comment about Rubin's (1981) example: It is not the incorporation of prior information that improves on maximum likelihood inference in that example, but the willingness to integrate over the marginal likelihood for τ rather than maximize over it. The dotted line in my Figure 1 is a rescaled version of the marginal posterior distribution for τ with a completely flat prior; the positive estimate of τ arises not because the likelihood information is being combined with prior information tilted away from zero, but because the *MLE* is a poor summary of the highly skewed marginal likelihood for τ.

Raudenbush misquotes me on Huttenlocher et al. (1991). To say that I

find no scientific meaning in the parameter estimates and *SE*s these authors report is not to say that I find no such meaning in the relationship between maternal speech and language development. What meaning, for instance, would Raudenbush attach to the finding that the *SE* for the log-exposure coefficient β_2 is 0.36? Because of the way the Huttenlocher et al. data were gathered, this estimate does not accurately quantify our uncertainty about β_2 in any population *P* of broad scientific interest, unless you are willing to assert that variance components for both nonexchangeability (of Huttenlocher et al.'s sample with the rest of *P*) and nonignorability (of assignment to maternal speech conditions) are close to zero. Such an assertion would require an explicit justification (involving, among other things, identification of *P*) that neither Huttenlocher et al. nor Raudenbush have fully provided.

Having said this, I am not claiming that nothing of scientific relevance has been learned by this study about the relationship between maternal speech *X* and child language development *Y*. In my terminology, Huttenlocher et al. are entitled to draw a calibration inference; an association between *X* and *Y* has been demonstrated for the 22 mother-child pairs in their study. Whether that association is causal (Raudenbush also misrepresents my views on this point) and how it would hold up in other mother-child pairs are interesting questions that have not been settled by this single study. Despite the tone of his remarks, I have no disagreement with Raudenbush about the basic nature of causal inference. Tentative causal conclusions, especially from observational studies, require both data evidence and a good story about why the data came out the way they did, and the story may later change when new information comes in. Ultimately, causality is a judgment about how the world works, not a property of the world itself. When investigators in one generation decide provisionally that *A* causes *B*, and people 20 years later conclude that *C* is really causing *A* and *B* to move in tandem, it is typically not the world that has changed, but rather our understanding of it.

Raudenbush's lung cancer example supports his argument only because it has gradually been found over time that the similarities between people in the nature of the mechanism(s) by which smoking causes lung cancer are substantially larger than the differences; that is, interactions between treatment (smoking dose) and subject-specific characteristics (potential confounding factors [PCFs]), such as gender and age, are small. Unfortunately, people arrive firmly at this sort of conclusion only retrospectively; I doubt Raudenbush would prospectively claim all such interactions to be negligible in educational interventions. There is a valuable role to be played by what he calls lung-tissue studies in education (nobody is suggesting that we wait around for his "perfect social science study"), as long as the amount learned by each such investigation is not oversold.

This same issue of presence or absence of treatment × PCF interactions comes up in Goldstein's comment about "replicability of findings in very different contexts." One way to estimate such interactions—a kind of meta-

analytic, observational-study approach—is, as Goldstein notes, for a variety of investigators to try to replicate an initial tentative causal finding with a variety of subject cohorts. Another way is to try to prospectively identify what the important PCFs are likely to be and to design a single larger study that stratifies on them, with sufficiently big samples in the cells of the treatment × PCFs grid to estimate the interactions well. The latter approach has the advantage of increased explicitness, although from the science-as-a-career point of view it seems to suffer from the drawback that only one team of investigators would get credit for the work. However, while one team would indeed have to be responsible for the meta-planning, other teams could be conducting parallel studies, one for each cell in the stratification grid, and there is room in this strategy for lots of credit to accrue to these teams as well. This is certainly not the place to try to settle the bigger-versus-smaller-science debate, to which Raudenbush also alludes, but the issue deserves more discussion than it has so far received (e.g., how should major granting agencies allocate their funds among "big" and "small" studies to maximize the rate at which scientific and policy progress unfolds in any given discipline?).

Perhaps the most useful way for me to conclude is to demonstrate concretely that Bayesian analysis of HMs is—probably for better *and* for worse, as Mason fears—about to become considerably more routine. Until recently, when you wanted to do a hierarchical Markov-chain Monte Carlo (MCMC) analysis, you had to program up, say, a Gibbs sampler yourself. Recently, a prototype version of a fairly general-purpose Gibbs-sampling package became available: the (infelicitously named) program BUGS (Gilks, Thomas, & Spiegelhalter, 1994), developed at the MRC Biostatistics Unit in Cambridge, UK (available by anonymous ftp from ftp.mrc-bsu.cam.ac.uk). As an example of the use of this program, Figure 2 presents some aspects of an MCMC meta-analysis of the teacher expectancy data from Bryk and Raudenbush (1992, chapter 7), mentioned in the fifth section of my article and examined by Morris in his comment. The data consist of effect size estimates y_i (together with their estimated standard errors, $\sqrt{V_i}$, assumed known) from $k = 19$ experiments measuring the influence of teachers' expectations on pupils' IQ scores, together with a study-level predictor x_i, the number of weeks of student-teacher contact prior to each experiment. With the addition of a prior distribution on the hyperparameters, the HM fit by Bryk and Raudenbush is (in my notation)

$$\begin{aligned} (y_i \mid \theta_i) &\overset{\text{indep}}{\sim} N(\theta_i, V_i) && \text{(Level 1)}, \\ (\theta_i \mid \alpha, \beta, \sigma^2) &\overset{\text{indep}}{\sim} N(\alpha + \beta(x_i - \bar{x}), \sigma^2) && \text{(Level 2)}, \\ (\alpha, \beta, \sigma^2) &\sim p(\alpha)p(\beta)p(\sigma^2) && \text{(prior)}. \end{aligned} \tag{8}$$

Figure 3 gives a simple program yielding inferences and predictions based on this model in BUGS, with uninformative priors on α, β, and σ^2 for

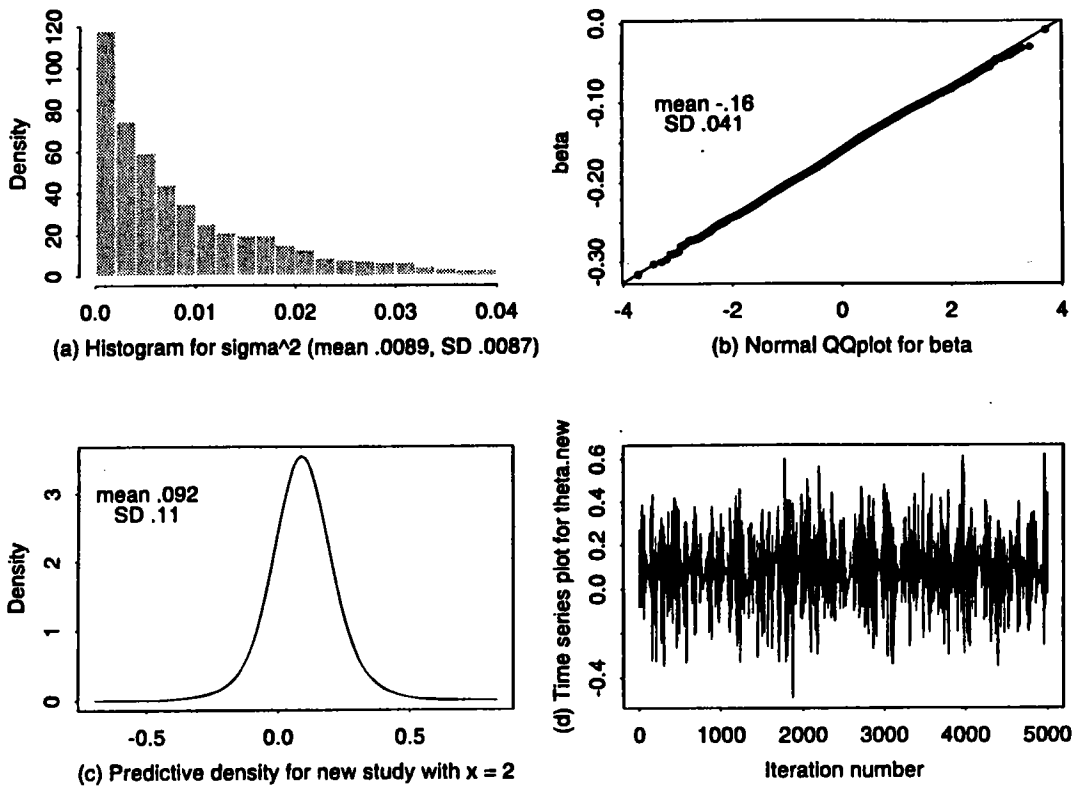

FIGURE 2. *Aspects of Markov-chain Monte Carlo (MCMC) teacher expectancy analysis*

```
model iqcov;
const
  k = 19, m = 101;
var
  u, p[m], sigmasq, tau, x.bar, alpha, beta, mu.new, theta.new,
  y[k], theta[k], precision[k], mu[k], x[k];
data in "iq-cov.dat";
inits in "iq-cov.in";                          # Prior for sigmasq is (discrete)
{                                              #   uniform on (.0002,.04), the
  u ~ dcat(p[]);                               #   region in which the
  sigmasq <- 0.0002 + 0.0398*(u - 1.0)/100.0;  #   likelihood is appreciable
  tau <- 1.0/sigmasq;                          # Precision tau = reciprocal variance
  alpha ~ dnorm(0.0,1.0E-4);                   # Uninformative priors for
  beta ~ dnorm(0.0,1.0E-4);                    #   alpha and beta
  x.bar <- mean(x[]);
  mu.new <- alpha + beta*(2.0 - x.bar);
  theta.new ~ dnorm(mu.new,tau);               # theta for new study with x = 2
  for (i in 1:k) {
    y[i] ~ dnorm(theta[i],precision[i]);
    mu[i] <- alpha + beta*(x[i] - x.bar);
    theta[i] ~ dnorm(mu[i],tau);
  }
}
```

FIGURE 3. *BUGS program to fit the hierarchical model in (8) using Gibbs sampling in the teacher expectancy study*

illustration and comparability with Morris's results. Gibbs sampling is iterative and requires a strategy for starting and stopping; I used an initial discarded ("burn-in") run of 1,000 iterations and a single long run of 5,000 iterations thereafter to obtain final results (diagnostics not presented here indicate convergence was achieved with this strategy in this case). The 6,000 total iterations took about 7 minutes on a SPARCstation 1000.

The output of the program in Figure 3 is a data set with 5,000 rows, one for each iteration after burn-in, and 23 columns, one for each quantity of interest specified by the program: α, β, σ^2, the 19 θ_i, and the effect size θ_{new} to be expected from a future study with $x = 2$. Consecutive iterations of the Gibbs sampler are, in general, correlated, in some cases highly so, but as long as convergence has been achieved, one nice thing about Gibbs is that you can regard the 5,000 numbers in any column of this data set as random draws from the marginal posterior distribution for the corresponding quantity, to be summarized in any way that seems useful. Standard summaries include means and *SD*s (note, for instance, that in agreement with Morris's results, the posterior variance for β in Figure 2b is 32% larger than the value reported by Bryk and Raudenbush from their maximum likelihood analysis); distributional summaries such as histograms (e.g., the marginal posterior for σ^2 in Figure 2a, which may be compared with Morris's Figure 1), Gaussian QQplots to assess posterior normality (as in Figure 2b), and kernel density estimates (e.g., Figure 2c, which presents the posterior predictive distribution for θ_{new}); and time series plots to assess convergence as in Figure 2d, which shows a healthy pattern with relatively little serial correlation.

There is considerably more work to be done on modeling and convergence diagnostics, specification of prior distributions in a way that does not introduce artifacts (e.g., I am indebted to Carl Morris for comments that motivated the prior for σ^2 here), and so on, but already BUGS is beginning to permit full Bayesian analysis of complicated models without the drudgery of one-off programming. It will be interesting to see how this sort of alternative to naive maximum likelihood fares in the social science marketplace in the next few years.

References

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician, 43,* 169–178.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27,* 236–248.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics, 6,* 377–400.

Seltzer, M. H. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational Statistics, 18,* 207–235.