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: $Y_{it}$, 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 $Y_{it}$ 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

$$Y_{it} = \pi_{2i} \cdot (t - 12)^2 + \epsilon_{it} \quad \text{(within-subjects)},$$

$$\pi_{2i} = \beta_0 + \beta_1 \cdot \text{group}_i$$

$$+ \beta_2 \cdot \log(X_i) + \beta_3 \cdot \text{gender}_i + U_i \quad \text{(between-subjects)}.$$  

Here $\pi_{2i}$ represents the acceleration in vocabulary growth for child $i$; $\epsilon_{it}$ 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 $\epsilon_{it}$
and $U_i$ are regarded as Gaussian random variables with mean 0 and variances $\sigma^2$ and $\tau$, respectively. The authors report a variety of standard errors (SEs) 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 SEs 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 $\tau$ and the $\pi_{2i}$ 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**,
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,

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

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 \( SEs \) 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 \( SEs \) 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 children’s 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

\[
Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \overline{X}_.) + r_{ij} \quad \text{(Level 1)},
\]

\[
\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}, \quad \beta_{1j} = \gamma_{10} \quad \text{(Level 2)},
\]

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; \(\overline{X}_.\) 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 \(\sigma^2\) and \(\tau_{00}\), 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 $\bar{Y}_E - \gamma_{01}(\bar{X}_E - \bar{X}_..) = 2.96$ and $\bar{Y}_C - \gamma_{01}(\bar{X}_C - \bar{X}_..) = 2.78$ 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 $\mu_0j$ 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 $\rho = .08$). 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 $\gamma_{01}$ 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>
<thead>
<tr>
<th>Class</th>
<th>1</th>
<th>1</th>
<th>15</th>
<th>15</th>
<th>16</th>
<th>16</th>
<th>22</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child</td>
<td>1</td>
<td>11</td>
<td>1</td>
<td>14</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>Group</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>E</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Y if E</td>
<td>2.6</td>
<td>3.0</td>
<td>2.8</td>
<td>3.6</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Y if C</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>3.0</td>
<td>2.9</td>
<td>1.9</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>2.5</td>
<td>2.4</td>
<td>3.5</td>
<td>3.3</td>
<td>2.7</td>
<td>3.0</td>
<td>2.1</td>
<td>2.5</td>
</tr>
</tbody>
</table>

TABLE 1
Display of the Cognitive Strategies in Writing Project data in factual-counterfactual form
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
Inference and Hierarchical Modeling

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>
<thead>
<tr>
<th>Exchangeability of Sampled and Unsampled Units in Target Population</th>
<th>Strong Ignorability of Treatment Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficult to Justify</td>
<td>Justifiable</td>
</tr>
<tr>
<td>Specific Causal Inference</td>
<td>General Causal Inference</td>
</tr>
<tr>
<td>Calibration Inference</td>
<td>Sampling Inference</td>
</tr>
<tr>
<td>General Causal Inference</td>
<td>Justifiable</td>
</tr>
</tbody>
</table>
Inference and Hierarchical Modeling

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

\[
\text{calibration inference} < \left( \begin{array}{c} \text{specific causal inference} \\ \text{sampling inference} \end{array} \right) < \text{general causal inference}. \tag{3}
\]

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 $Y_q$ 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

$$ Y_{ij} = \beta_0 + \beta_1 X_{ij} + \gamma_j + e_{ij} \quad \text{(Level 1)}, $$

$$ \gamma_j = u_j \quad \text{(Level 2)}, $$

130
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 $\gamma_j$, 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 $[\bar{Y}_j - (\beta_0 + \beta_1X_j)]$ 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 $\bar{Y}_j$, 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 $\bar{Y}_j$ 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 $\bar{Y}_j$, 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:
Draper

- **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
Inference and Hierarchical Modeling

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>
<thead>
<tr>
<th>Study</th>
<th>Aspirin</th>
<th>Placebo</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of Patients</td>
<td>Mortality Rate (%)</td>
<td>No. of Patients</td>
</tr>
<tr>
<td>UK-1</td>
<td>615</td>
<td>8.0</td>
<td>624</td>
</tr>
<tr>
<td>CDPA</td>
<td>758</td>
<td>5.8</td>
<td>771</td>
</tr>
<tr>
<td>GAMS</td>
<td>317</td>
<td>8.5</td>
<td>309</td>
</tr>
<tr>
<td>UK-2</td>
<td>832</td>
<td>12.3</td>
<td>850</td>
</tr>
<tr>
<td>PARIS</td>
<td>810</td>
<td>10.5</td>
<td>406</td>
</tr>
<tr>
<td>AMIS</td>
<td>2267</td>
<td>10.9</td>
<td>2257</td>
</tr>
</tbody>
</table>
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,

\[ y_i | \theta_i \overset{\text{indep}}{\sim} N(\theta_i, V_i) \]  
\[ \theta_i | \mu, \tau^2 \overset{\text{IID}}{\sim} N(\mu, \tau^2) \]

in which \( y_i \) is the mortality difference in study \( i \) and the \( V_i \) (the squared SEs 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 \( \tau \) comes out about 1.5\% here, leading to noticeably wider 95\% central interval estimates for the "true effect" \( \mu \) 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, 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 \( \tau \) 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 looks 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 findings3) 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
Draper

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

\[
\begin{align*}
  y_j &= X_j \beta_j + e_j \quad \text{(Level 1)}, \\
  \beta_j &= W_j \gamma + u_j \quad \text{(Level 2)},
\end{align*}
\]

where \( y_j \) is the \( n_j \times 1 \) vector of outcomes for, say, the students in, say, school \( j \), \( X_j \) is an \( n_j \times P \) matrix of (known) student-level predictors at school \( j \), \( W_j \) is a \( P \times K \) matrix of (known) school-level predictors, the components of \( e_j \) are IID \( N(0, \sigma_e^2) \), the \( u_j \) are \( P \)-variate normal with mean vector \( 0 \) and covariance matrix \( T \), 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 \( e_j \)), but the inclusion of the school-level latent variables \( u_j \) 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 MLEs 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 \( \beta_j \) 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 MLEs 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 \( \gamma \) and the variance/covariance unknowns \( \sigma^2 \) and \( T \) 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 MLEs 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 SEs 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 MLEs 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 \( \tau \) and the eight solid lines track the usual shrinkage estimates \( W_i y_i + (1 - W_i) \hat{\mu} \) of the “true” treatment effects \( \theta_i \) of Programs A–H as a function of \( \tau \), where \( \hat{\mu} \) is the grand mean \( \Sigma_{i=1}^8 W_i y_i / \Sigma_{i=1}^8 W_i \) (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 \( \tau \) as its posterior distribution with a diffuse prior, Rubin’s fully Bayesian analysis implies a summary value of \( \tau \) of about 7 SAT verbal points. However the MLE of \( \tau \) 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:
Draper

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

\[
(y_j | \beta_j, \sigma^2) \sim N_{n_j}(X_j \beta_j, \sigma^2 I_{n_j}) \quad \text{(Level 1)},
\]

\[
(\beta_j | \gamma, \tau) \sim N_p(W_j \gamma, \tau) \quad \text{(Level 2)},
\]

\[
(\gamma, \tau, \sigma^2) \sim p(\gamma)p(\tau)p(\sigma^2) \quad \text{(prior)},
\]

The goal is computation and marginalization of the posterior distribution $p(\beta, \gamma, \tau, \sigma^2 | y)$, 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 $p(\sigma^2 | y, \beta, \gamma, \tau)$, $p(\beta | y, \gamma, \tau, \sigma^2)$, $p(\gamma | y, \beta, \tau, \sigma^2)$, and $p(\gamma | y, \tau, \beta, \sigma^2)$ 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 $\varepsilon_{it}$ 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 $\varepsilon_{it}$ 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
Inference and Hierarchical Modeling

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

1By 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.
2Actually, 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.
3Rodriguez & 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


Draper


**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.