Journal of Educational and Behavioral Statistics
Summer 1995, Vol. 20, No. 2, pp. 239–240

Not Much Disagreement, It Seems

Jan de Leeuw
University of California, Los Angeles
Ita G. G. Kreft
California State University, Los Angeles

The overall friendliness of the discussion is both surprising and welcome. We do not want to cast a pall on all this sweetness and light, but it should not confuse the reader into thinking that all participants basically agree. Further analysis reveals some very deep philosophical disagreements on various levels—on the general foundations of data analysis and statistics, on the role of models, on the importance of hierarchical models, and on projected or desired future developments in this area.[1] We shall try to limit our final remarks to comments and criticisms directly aimed at our paper.

Morris. In the third section of his discussion Morris uses the NCES questions. Nothing he says is critical of what we say, however. At some point, Morris states that we "provide and recommend simple, not fully efficient methods for estimating two-level models." This is somewhat misleading, because with appropriate (Swamy) weights, the two-step estimates are indeed efficient for estimation of the regression coefficients γ_{st}. The estimates of the variance components may not be efficient, but efficiency in the variance component context is problematic anyway, because bias is so pervasive.

Goldstein. Goldstein seems to feel that the route towards solving some of the problems with simple multilevel models is to introduce more complicated ones. He advocates modeling of Level 1 variances, nonlinear components, and measurement errors. For some reason, he refuses to acknowledge that we think this growing of the model will lead to disaster eventually, and is not necessarily wise if the estimation problems for the simplest possible model are still problematic.

The new version of Goldstein's program is called MLn and handles up to 15 levels. We are still not sure if he is really serious, or if this is some kind of Monty Python–type exaggeration.

Mason. We continue to disagree with Mason on the fixed-X assumption. The story is, indeed, that under repeated sampling we fix the covariate combinations. The opposition, however, will immediately say "How?" It seems to us that right there we lose our case.

Raudenbush. The final *two* sentences of our abstract say that multilevel models are not always necessary and that traditional techniques sometimes perform better. This does not violate the distinction between choice of model and choice of method, although we could have been more clear. *Traditional* in this context means weighted or unweighted ordinary least squares, and the

Henderson unbiased estimates of the variance components. This combination gives efficient estimates of the first-level regression coefficients. Any additional iterating is, as far as we can see, making sacrifices to the Maximum Likelihood God. It is not true that more intensive computation is the only price to pay. Our paper emphasizes that we also give up unbiasedness, familiarity, easy diagnostics, computational ease in establishment of sampling distributions, worries about convergence, and possibilities to do more extensive resampling.

The first paragraph of Raudenbush's reply sounds like a sales pitch. All hierarchical research in educational statistics before the discovery of RC models is condemned. Strong characterizations such as "mismatch between reality and statistical model," "misestimated precision," and "weak power" are enough to make us deeply ashamed for ever using traditional statistical methods. We think these strong characterizations are misleading. We already know that from the point of view of the practitioner, the differences between "traditional" and "modern" methods of fitting models are small, often tiny. Of course, assuming a hierarchical model implies that the corresponding formulas for the standard errors should be used.

We also think the admonitions about power and misestimated precision are rather premature. They are true, to some extent, if the hierarchical model is indeed the correct one, but even in that case it is so far unclear to what extent they are true. Raudenbush refers to unpublished research by some of his graduate students, but as far as we know these results are all conditional on the RC model. Clearly, if we generate data with one model, then analysis based on that model will produce better results than analysis based on another model. This is only conclusive if one assumes a priori that the hierarchical linear model is appropriate. If the model is not questioned, then obviously all other models, and the techniques based on them, are wrong.

Conclusion

It is clear from the contributions to this discussion that RC models are useful, maybe even very useful, but they cannot be used as the only possible starting point in hierarchical situations. Neither can we allow reviewers of journals, or gurus in strategic positions, or people selling computer software, to dictate the appropriate technique for a given situation. No statistical technique can replace a precise formulation of a question, and a precise study of the peculiarities of the data we are using to answer that question.

Note

[1]Our own recent thinking has been influenced by some discussions of our article with John Tukey. It was too late to incorporate John's suggestions in the article, unfortunately.