DISCUSSION

Vernon L. Greene, Syracuse University

The important questions in public policy analysis seldom lend themselves to simple research designs. And even when designs are simple, implementation never is. Even in the rare instances where experimental designs are implemented in a large scale project, as in the case of the National Long Term Care Demonstration on which the papers under discussion here today are based, problems of design implementation and measurement often confound with treatment effects. The result is the reintroduction of sources of spurious variation into the experimental data, whose presence then requires the use of post hoc controls, or at least an explicit assessment of whether and to what extent such controls are necessary.

The present papers are based on one of the most important experimental policy studies that has been undertaken in the past fifty years. Its findings will be critical to planning efforts that address the challenges of an increasingly "older" society. Hence an adequate analysis of the limitations of the data sets it has produced, a task undertaken by the present authors, is a topic of the utmost importance. With this in mind, I would like to comment briefly on each paper.

The paper by Brown and Mossel documents the unhappy situation in which an experimental structural design is compromised by non-comparable measurement procedures in the experimental and control groups. Insofar as the variables in question are to be used as outcome or target variables, the immediate consequence, of course, is to directly confound treatment with measurement effects, rendering observed treatment-control differences ambiguous as to their source. In the case at hand, however, the variables in question are apparently to be treated as control variables in a multivariate linear model predicting target variables as a function of treatment-control assignment. If these variables are differentially biased across treatment and control groups, their inclusion will bias the estimation of treatment effects.

Given that the treatment and control groups here are randomly assigned, one may well wonder why statistical controls are necessary in this case, or, for that matter, even desirable. The logic of experimental design is intended precisely to obviate the need to introduce explicit control adjustments. This in turn relieves one of the task (usually hopeless in policy research) of correctly identifying all the variables to be controlled and specifying their relationship to the variable(s) under study.

In the case at hand, the introduction of controls, within a multiple regression/analysis of covariance model as given by the authors, commits one to a linear specification-of-convenience. Insofar as this specification is incorrect, treatment effect estimates will be biased even without the measurement problems noted by the authors. While it is not uncommon to introduce controls into experimental data in order to wring out the error sum of squares, and improve the "precision" of estimation, there is always danger that the gain in test power is achieved at the price of structural misspecification. Absent clear evidence of extraneous variation in the dependent variable, contaminating experimental data in this way seems to me nearly always ill-advised, even when the control data is sound.

One might go on to note that the data in question appear to be subject to a further problem. It seems reasonable to believe that the baseline assessment interview conducted by the channelling staff with the treatment group is not only a measurement event, but is in effect the initiation of treatment. This point is recognized by the authors in noting that one reason that channelling staff were not used to interview the control group was precisely to avoid "contaminating" (presumably, "treating") it. In this case one sees that, in the treatment group, treatment and measurement effects are immediately and inextricably confounded in the control variables. In a multiple regression-analysis of covariance context, this influence of the treatment factor on the covariate effects from the dependent variable also removes some of the effects of the treatment. When several such covariates are to be used jointly, as in the present case, the resulting bias may be arbitrarily complex. In particular, it may not do to simply adopt an approach of eliminating potential covariates which differ significantly across the groups at baseline. If the treatment and measurement effects are offsetting, for example, this may result in retaining covariates whose measurement effect bias is substantial.

By and large the authors have done an admirable job of making use of the data available to them to explore sources of bias among the potential control variables. Their approach, while ad hoc and hence subject to some technical limitations, as in treating sequential hypothesis tests as though they were independent, seems plausible on its face. One must agree also with their general comment that an experimental design is one of those things that is probably only worth doing if it is done correctly. Deliberately using non-equivalent measurement procedures across experimental treatment and control groups is to methodologically shoot one's self in the foot.

I might conclude by noting that the author's recommendations to not use variables the evidence for bias in which is virtually certain seems much too liberal, and places the burden of proof too far in the wrong direction. Retaining variables simply because one cannot conclude with probability .9 (or, for that matter, .8) that they are biased does not seem adequate protection against the problems such bias can cause, although in view of the large sample sizes, even statistically significant differences are likely to be substantively inconsequential.

The paper by Phillips highlights another key problem in experimental studies of public policy options. In the real world, policy interventions are themselves often very complex, sometimes by design, sometimes inevitably or unexpectedly. Complex interventions are particularly susceptible to uncontrollable variations in the magnitude and structure of their application to individuals and subgroups within the treatment population. Also, when the experimental treatment consists of valuable services that cannot be monopolistically supplied by the experimenter, it is inevitable that persons outside the treatment group, including those in the control group, will also avail themselves of such services. In this case, the inherent effect of utilization of such services will be to some degree attenuated across the treatment and control groups.

Phillips addresses the question of whether there are differences in levels of receipt of case management services across the treatment and control groups by comparing the proportion in each group reporting receipt of various components of such services. Regardless of the component considered, substantial and statistically significant differences were found, with the treatment group reporting a higher incidence of case management services. Thus it appears that channeling directly and/or indirectly enhances levels of case management services.

One source of concern in these findings is that the difference in proportions reported has been adjusted by regression methods for a variety of unspecified baseline characteristics. These are presumably the same data discussed by Brown and Mossel. If this is the case, for reasons given previously, the results presented here may be in doubt. Given the magnitude of the differences reported by Phillips, it seems unlikely that the overall finding (that the treatment group received more case management services) would be vitiated. Still, the specific impact estimates may be substantially biased. This is difficult to assess here, however, since the covariates used are not identified.

One sometimes has an uneasy sense of circularity in the findings. For example, if the treatment group is counting its baseline visit by a channeling staff member as an instance of case management service provision (say, a "visit to arrange services"), as appears to be the case, then it is hardly surprising that treatment group members will be more likely to be found to receive case management services. This is simply a measurement artifact. Generally, when the treatment and the target variable are one and the same thing, to find that those treated exhibit higher levels of the target variable is arguably rather circular, other things equal.

The Schorr paper concerns itself with estimating death rates in the Demonstration study population. Two apparently independent data sets are available--one based on official death records, the other on survey interviews endogenous to the project. At issue is whether these two data sets can be combined. The official records are thought to undercount, but equally for treatment and control groups. The survey records are suspected to tend to undercount in the control group more than in the treatment group. At issue is whether to combine the two data sets in the sense that an individual be counted as dead if so identified by either method.

One wonders first why it is desirable to risk contaminating the estimate or the mortality difference across the treatment and control groups by introducing the survey data, suspected of differential undercounting, at all. The official records data, for which a strong *a priori* case against differential undercounting can be made, provides an unbiased estimate of the mortality difference across the two groups, even though it underestimates the absolute rates.

Schorr is correct that combining the data sources will produce a more accurate estimate of the absolute rates, though only, it should be noted, if the only errors are assumed to be undercounts. If errors include "false positives," this is no longer necessarily true. As a practical matter, however, this seems a negligible problem.

The statistical tests used in justifying combining the data (the limiting distribution for which is z, not t) seems far too permissive in the same sense as earlier remarks on the Brown and Mossel paper. The effective decision rule is to not combine only if the evidence against so doing is overwhelming. It is not much of a discipline to resist combining only if the probability that the test statistics arise from similar populations approaches conventional significance levels. Somehow, it seems that the burden of plausibility should be in the other direction.

On balance, it seems to me that the problems of estimating treatment effects on mortality ought to be separated from those of estimating absolute mortality levels. The former is perhaps best dealt with using the official records data alone, the latter with the combined data.