Assessment of Treatment Effects Using Latent Variable Modeling: Comments on the New York School Choice Study

Bengt Muthén

Booil Jo

University of California, Los Angeles

C. Hendricks Brown

University of South Florida [*]

Version 1, December 18, 2002

[*] The research of the first author was supported under grant K02 AA 00230 from NIAAA. The research of the second and third authors was supported by NIDA and NIMH Grant MH40859. We thank Chen-Pin Wang for research assistance, Joyce Chappell for graphical assistance, and the members of the Prevention Science Methodology Group and the Fall ED299A class for helpful comments. The email address of the first author is bmuthen@ucla.edu.

1 Introduction

The article by Barnard, Frangakis, Hill and Rubin (BFHR) is timely in that the Department of Education is calling for more randomized studies in educational program evaluation (see the discussion of the "No Child Left Behind" initiative, e.g. Slavin, 2002). BFHR can serve as a valuable pedagogical example for successful sophisticated statistical analysis of randomized studies. Our commentary is intended to provide additional pedagogical value to benefit the planning and analysis of future studies, drawing on experiences and research within our research group.[1]

BFHR provides an exemplary analysis of the data from an imperfect randomized trial that suffers from several complications simultaneously: noncompliance, missing data in outcomes, and missing data in covariates. We are very pleased to see their application of cutting edge Bayesian methods for dealing with these complexities. In addition, we believe the methodological issues and the results of the study have important implications for design and analysis of randomized trials in education and for related policy decisions.

BFHR provides results of the New York City school choice experiment based on one year achievement outcomes. With the planned addition of yearly follow-up data, growth models can provide an enhanced examination of causal impact. We discuss how such growth modeling can be incorporated and provide a caution that applies to BFHR's use

[1] The "Prevention Science Methodology Group" (PSMG; www.psmg.hsc.usf.edu), co-PI's Brown and Muthén, has collaborated over the last 15 years with support from NIMH and NIDA.

of only one posttest occasion. In our commentary we also consider the sensitivity of the latent class ignorability assumption in combination with the assumption of compound exclusion.

2 Longitudinal Modeling Issues

BFHR focuses on variation in treatment effect across compliance classes. This part of the commentary considers variation in treatment effect across a different type of class based on the notion that the private school treatment effect might very well be quite different for children with different achievement development.[2] To study such a "treatment-baseline interaction" (or, rather "treatment-trajectory interaction"), we will switch from BFHR's pretest-posttest analysis framework, essentially a very advanced ANCOVA-type analysis, to the growth mixture modeling framework of Muthén, Brown et al. (2002). An underlying rationale for this modeling is that individuals at different initial status levels, and on different trajectories, may benefit differently from a given treatment. ANCOVA controls for initial status as measured by the observed pretest score. Unlike the observed pretest score, the latent variable of initial status is free of time-specific variation and measurement error.

The focus on longitudinal aspects of the New York School Choice study (NYSCS)

[2] Also of interest is potential variation in treatment effects across schools, with respect to both the public school the child originated in and the private school the child was moved to, but this multilevel aspect of the data will be left aside here for lack of space.

is both substantively and statistically motivated. First, treatment effects may not have gained full strength after only a one-year stay in a private school.[3] Second, the use of information from more timepoints than pretest and posttest makes it possible to identify and estimate models that give a richer description of the normative development in the control group and how the treatment changes this development.

Consider three types of latent variables for individual i. The first, C_i, refers to BFHR's compliance principal strata. The next two relate to the achievement development as expressed by a growth mixture model: T_i refers to trajectory class and η_i refers to random effects within trajectory class (within-class model is a regular mixed effects model). Unlike the latent class variable C_i, the latent class variable T_i is a fully unobserved variable as is common in latent variable modeling (see, e.g. Muthén, 2002a). Consider the likelihood expression for individual i, using the [] notation to denote probabilities/densities,

$$[C_i, T_i | X_i]\ [\eta_i | C_i, T_i, X_i]\ [Y_i | \eta_i, C_i, T_i, X_i]\ [U_i | \eta_i, C_i, T_i, X_i]\ [R_i | Y_i, \eta_i, C_i, T_i, X_i], \qquad (1)$$

where X_i denotes covariates, U_i denotes a compliance stratum indicator (with C_i perfectly measured by U_i in the treatment group for never takers and perfectly measured

[3] The NYSCS currently has data from three follow-ups, i.e. providing repeated measures data from four grades. Although BFHR used percentile scores which do not lend themselves to growth modeling, a conversion to "scale scores" (i.e. IRT-based, equated scores) should be possible, enabling growth modeling. Unfortunately, educational research traditionally uses scales that are unsuitable for growth modeling, such as percentile scores, normal curve equivalents, and grade equivalents (for a comparison in a growth context, see Seltzer, Frank & Bryk, 1994). Hopefully, this tradition can be changed.

in the control group for always takers, other group-class combinations having missing data), and R_i denotes indicators for missing data on the repeated measures outcomes Y_i (pre- and post-treatment achievement scores). This type of model can be fitted into the latent variable modeling framework of the Mplus program (Muthén & Muthén, 1998-2002; Technical Appendix 8), which has implemented an EM-based maximum-likelihood estimator.[4] As a special case of (1), conventional random effects growth modeling includes η_i, but excludes C_i and T_i and assumes MAR so that the last term in (1) is ignored. Growth mixture modeling (Muthén & Shedden, 1999; Muthén, Brown et al., 2002; Muthén & Muthén, 1998-2002) includes η_i and T_i. BFHR includes C_i, but not T_i (or η_i), and includes the last term in (1), drawing on latent ignorability of Frangakis and Rubin (1999). Muthén and Brown (2002) studies latent ignorability related to T_i in the last term of (1). In randomized studies, it would be of interest to study C_i and T_i classes jointly because individuals in different trajectory classes may show different compliance and missingness may be determined by these classes jointly.

If data have been generated by a growth mixture model with treatment effects varying across trajectory classes, what would pretest-posttest analysis such as that in BFHR reveal? To judge the possibility of such treatment-trajectory interaction in the NYSCS, we considered several recent applications of growth mixture modeling that have used T_i to represent qualitatively different types of trajectories for behavior and achievement scores on children in school settings. Drawing on these real-data studies, two growth mixture scenarios were investigated.[5] For simplicity, no missing data on the outcome or

[4] For related references, see the Mplus web site www.statmodel.com.

[5] A detailed description of these real-data studies and scenarios and their parameter values are given

pretest is assumed and C_i classes are not present. In a 3-class scenario, the treatment effect is only noteworthy for a 70% middle class, assuming that the low class membership (10%) hinders individuals from benefiting from the treatment and assuming that the high class (20%) does not really need the treatment. The achievement development in the 3-class scenario is shown in the left panel of Figure 1 while the right panel shows the corresponding posttest (y_2) - pretest (y_1) regressions. The lines denoted ANCOVA show a regular ANCOVA analysis allowing for an interaction between treatment and pretest (different slopes). In the 3-class scenario, the ANCOVA interaction is not significant at $n = 2000$ and the treatment effect in the middle class is underestimated by 20%, while overestimated in the other two classes. In a 2-class scenario (not shown here), where the treatment is only noteworthy for individuals in the low class (50%), ANCOVA detects an interaction that is significant at the NYSCS sample size of $n = 2000$, but underestimates the treatment effect for most children in the low class (at the low-class average pretest value of zero, the treatment effect is underestimated by 32%).

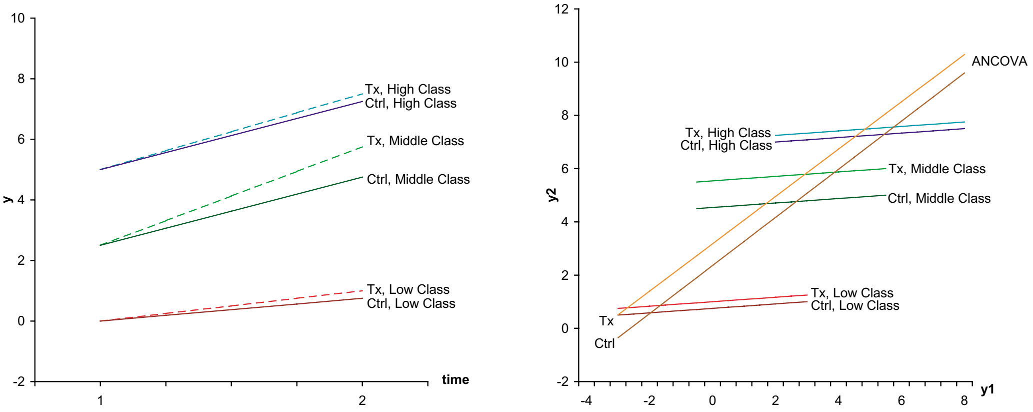

Figure 1: Growth Mixture Modeling Versus Pretest-Posttest Analysis.

in Mplus Web Note #5 at www.statmodel.com/mplus/examples/webnote.html.

Although the NYSCS children are selected from low-performing schools[6], there may still be sufficient heterogeneity among children in their achievement growth to make a treatment-trajectory interaction plausible. The 3-class scenario is possible, perhaps with more children in the low class relative to the other two classes. If this is the case, the ANCOVA analysis shown in Figure 1 suggests a possible reason for BFHR's finding of low treatment effects. The empirical studies and the results in Figure 1 suggest that future program evaluations may benefit from exploring variation in treatment effects across children characterized by different development. Using data from at least two post-treatment time points (three time points total), the class-specific treatment effects generated in these data can be well recovered by growth mixture modeling.[7] A more flexible analysis is obtained with more post-treatment time points. An improved design for the determination of the latent trajectory classes is to use more than one pre-treatment time point so that the trajectory class membership is better determined before the treatment starts.

3 Compound Exclusion and Latent Ignorability

Based on the ideas of principal stratification (Frangakis & Rubin, 2002) and latent ignorability (Frangakis & Rubin, 1999), BFHR successfully demonstrates that the complexities of educational studies can be better handled under more explicit and flexible

[6] The average NYSCS math and reading percentile rankings are around $23 - 28$.

[7] Monte Carlo simulation results are given in Mplus Web Note #5 at www.statmodel.com/mplus/examples/webnote.html.

sets of assumptions. Although we think the structural assumptions they employed are reasonable in the NYSCS, we would like to add some thoughts on the plausibility of two assumptions, considering more general situations.

Compound exclusion (CE) is one of the key structural assumptions in identifying principal effects under latent ignorability. However, the plausibility of this assumption can be questioned in practice (Frangakis et al., 2002; Hirano et al., 2000; Jo, in press, 2002; Shadish, Cook, & Campbell, 2002; West & Sagarin, 2000). In the NYSCS, it seems realistic to assume that winning a lottery has some positive impact on always-takers. However, it is less clear how winning a lottery will affect never-takers. One possibility is that winning a lottery has a negative impact on parents, because they fail to benefit from it. Discouraged parents may have a negative influence on children's test scores or response behaviors. This negative effect may become more evident if noncompliance is due to parents' low expectation or lack of interest in their children's education. Another possibility is that winning a lottery has a positive impact on children's test scores or response behaviors. For example, parents who are discouraged by being unable to send their children to private schools even with vouchers may try harder to improve the quality of existing resources (e.g., in the public schools their children attend) and be more motivated to support their children to improve their academic performance. Given these competing possibilities, it is not easy to predict whether and how CE is violated.

Depending on the situation, causal effect estimates can be quite sensitive to violation of the exclusion restriction in outcome missingness (Jo, 2002b), which is less known than the impact of violating exclusion restriction in observed outcomes (Angrist et al., 1996; Jo, 2002). The implication of possible violation of CE and its impact is that the relative benefit of models assuming latent ignorability (LI) and standard ignorability (SI) depends on degrees of deviation from CE and SI. Identification of causal effects under LI relies on the generalized (compound) exclusion restriction (i.e., both on the outcomes and missingness of outcomes), whereas identification of causal effects under SI relies on the standard exclusion restriction (i.e., only on the outcomes). Therefore, in some situations, the impact of deviation from CE may outweigh the impact of deviation from SI, resulting in more biased causal effect estimates in models assuming LI than in models assuming SI (Jo, 2002b). For example, if SI holds, but CE is seriously violated (say 20% increase in the response rate due to treatment assignment for compliers and 15% increase for never-takers), causal effect estimates and the coverage probability assuming LI and CE can drastically deviate from the true value and the nominal level. This type of violation does not affect models assuming SI and the standard exclusion restriction. To empirically examine the plausibility of SI, LI, and CE, it will be useful to do sensitivity analyses of models imposing different combinations of these assumptions. As the authors point out, this investigation can be conducted by relaxing compound exclusion (e.g., Frangakis et al., 2002; Hirano et al., 2000), or by employing alternative structural assumptions (e.g., Jo, in press). More research is necessary to examine the efficiency of these alternative models, and to explore factors associated with insensitivity of LI models to violation of compound exclusion.

4 Conclusion

Causal inferences of the type BFHR provide are a dramatic improvement over the existing literature now available on the question of whether school choice will produce better achievement outcomes for children in an urban public school system. The randomized lottery provides an exceptionally powerful tool to examine the impact of a program, far more useful than observational studies that have causal change intertwined hopelessly with self-selection factors. Statisticians are just now investigating variations in such principal strata analyses, that is, those involving latent classes formed as a function of randomized trials involving intervention invitations (such as vouchers), encouragement designs, and field trial designs involving more than one randomization (Brown & Liao, 1999). The latent categories in this paper, which BFHR labels complier, never-taker, always taker, and defier, represent only one type of design. There are other terms that may be more relevant to the scientific questions underlying trials where subjects are randomly assigned to different levels of invitation (e.g. Angrist & Imbens, 1995), or different levels of implementation. Such trials not only have great potential for examining questions of effectiveness, sustainability, and scalability, but they require terms more consistent with adherence than compliance. Again, we congratulate the authors to an important addition to the methodological literature which we predict will have lasting impact.

Additional References

Angrist, J. D., & Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association, 90*, 431-442.

Brown, C.H. & Liao, J. (1999). Principles for designing randomized preventive trials in mental health: An emerging developmental epidemiologic perspective. *American Journal of Community Psychology*, special issue on prevention science, 27, 673-709.

Jo, B. (in press). Estimating intervention effects with noncompliance: Alternative model specifications. Forthcoming in the *Journal of Educational and Behavioral Statistics*.

Jo, B. (2002). Model misspecification sensitivity analysis in estimating causal effects of interventions with noncompliance. *Statistics in Medicine, 21*, 3161-3181.

Jo, B. (2002b). Sensitivity of causal effects under ignorable and latent ignorable missing-data mechanisms. In preparation. Available at www.statmodel.com/mplus/examples/jo/

Muthén, B. (2002a). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*, 81-117.

Muthén, B., Brown, C.H. (2002). Non-ignorable missing data in a general latent variable modeling framework. Draft.

Muthén, B., Brown, C.H., Masyn, K., Jo, B., Khoo, S.T., Yang, C.C., Wang, C.P., Kellam, S., Carlin, J., & Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics, 3*, 459-475.

Muthén, L. & Muthén, B. (1998-2002). *Mplus User's Guide.* Los Angeles, CA: Muthén & Muthén.

Muthén, B. & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*, 463-46.

Seltzer, M.H., Frank, K.A. & Bryk, A.S. (1993). The metric matters: the sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis, 16*, 41-49.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Slavin, R.E. (2002). Evidence-based education policies: transforming educational practice and research. *Educational Researcher, 31*, 15-21.

West, S. G., & Sagarin, B. J. (2000). Participant selection and loss in randomized experiments. In L. Bickman (Ed.), *Research Design* (pp. 117-154). Thousand Oaks, CA: Sage.