Supplementary Analyses for the Article "Assessment of Treatment Effects Using Latent Variable Modeling: Comments on the New York School Choice Study"

Bengt Muthén
University of California, Los Angeles
*

Mplus Web Notes: No. 5
Version 1, December 20, 2002

*The research was supported under grant K02 AA 00230 from NIAAA. I thank Joyce Chappell and Booil Jo for graphical assistance, and the members of the Prevention Science Methodology Group including Hendricks Brown and Booil Jo for helpful comments. The email address is bmuthen@ucla.edu.

1 Introduction

Muthén, Jo and Brown (in press) gives a commentary on Barnard, Frangakis, Hill and Rubin (in press). This web note provides a brief background and gives the parameter values for the investigation of growth mixture modeling in Muthén, Jo and Brown (in press). Monte Carlo simulation studies corresponding to these models are also reported on.

Muthén and Curran (1997) introduced the idea of controlling for initial status in growth when assessing treatment effects on development over time. The argument is that individuals at different initial status levels may benefit differently from a given treatment. Unlike the observed pretest score used in analysis of covariance, the initial status is free of time-specific variation and measurement error. Muthén-Curran used multiple-group latent growth curve modeling in a randomized intervention study to show that initially more aggressive children in a classroom study benefited from the intervention in terms of lowering their trajectory slope. The Muthén-Curran technique is not, however, able to capture a non-monotonic intervention effect that exists for children of medium-range aggression, but is absent for the most or least aggressive children. In contrast, such a non-monotonic intervention effect can be handled using growth mixture modeling (Muthén & Shedden, 1999, Muthén, Brown et al., 2002). In this way, growth mixture modeling offers considerably more flexibility, not only controlling for initial status but also for development over time in the sense of trajectory class membership. An experimental version of the Mplus program (Muthén & Muthén, 1998-2002) to be released in 2003 allows such models to be fitted also in cluster samples, such as with students observed within schools.

The Muthén, Brown et al. (2002) growth mixture modeling for randomized trials can be expressed as follows for linear growth. It is assumed that the outcome y is in the same metric at the different time points, as would be obtained via IRT-based equating. The random effects are allowed to have different distributions for individuals belonging to different trajectory classes and for different intervention status. For trajectory class k,

$$y_{ti} = \eta_{0i} + \eta_{1i}\, a_{ti} + \epsilon_{ti}, \tag{1}$$

$$\eta_{0i} = \beta_{0k} + \zeta_{0i}, \tag{2}$$

$$\eta_{1i} = \beta_{1k} + \gamma_{1k}\, Z_i + \zeta_{1i}, \tag{3}$$

where η_{0i} is the pre-treatment initial status random effect (unaffected by treatment), η_{1i} is the growth rate random effect, a_{ti} is grade scored $0, 1, \ldots,$ Z_i is a 0/1 treatment dummy variable, and γ_{1k} is the class-varying treatment effect on the growth rate. The idea behind this model is that there is a growth model for normative development in the control group. When adding the treatment group to the analysis, only a few more parameters are used to represent change in the normative development. As pointed out in Muthén (2002), in some applications the classes have a real content motivated by substantive theories, whereas in other applications the model simply provides a more flexible way to represent individual differences.

Some recent applications of growth mixture modeling provide a background for the artificial data studies below. Muthén, Brown et al. (2002) found that a randomized behavioral intervention to reduce aggressive behavior in classrooms grade 1-7 had the largest effect for the children who were in a high-aggressive trajectory class (15%). Muthén (2002) found a class of students (19%) who had very poor mathematics development in grades 7-10, who were disengaged in school and significantly more likely to drop out of high school. Muthén, Khoo, Francis, and Boscardin (2002) found a failing class (10%) in word recognition development grades 1 -2. As background for this commentary, an analysis of reading development grades 2-5 was undertaken on scale scores (California Achievement Test) from a sample of about a thousand children drawn from Baltimore public schools (Dolan et al., 1993). Three trajectory classes emerged: a failing class (6%); a class developing normally (71%); and a class of children developing extremely well (23%).

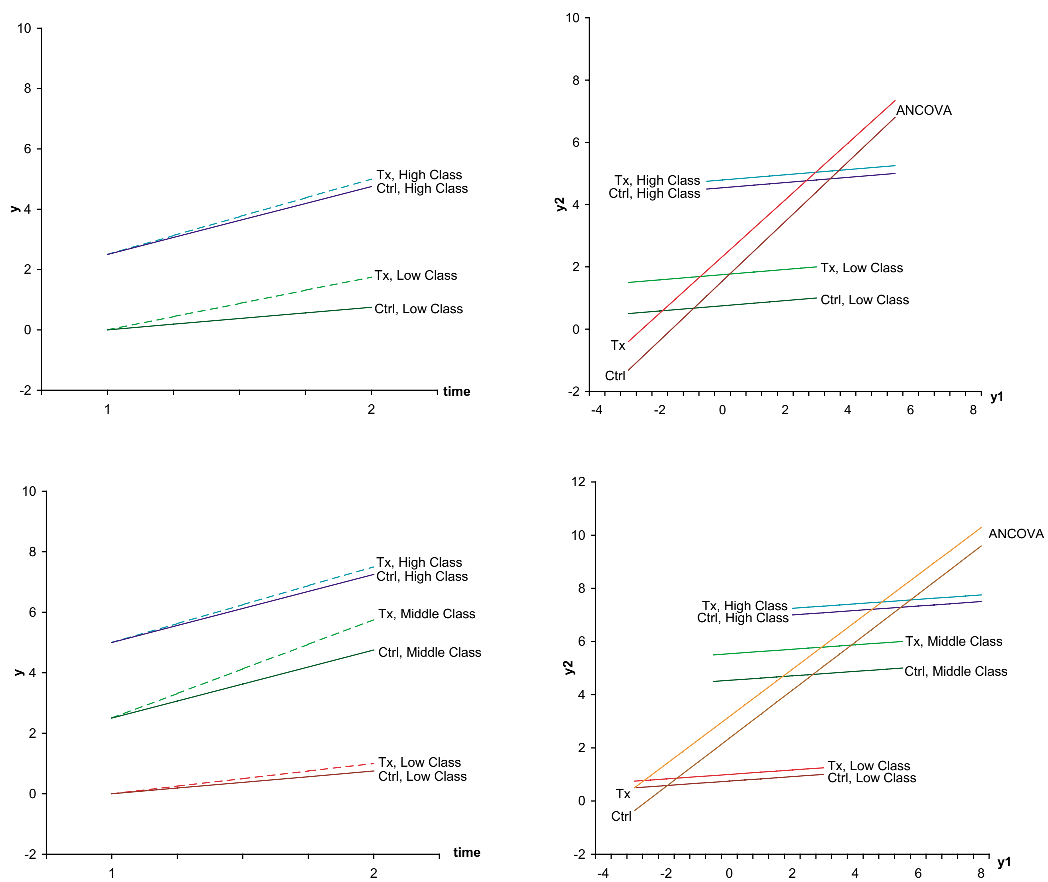

Figure 1: Growth Mixture Modeling Versus Pretest-Posttest Analysis.

2 Two Population Studies

The achievement development considered in the two scenarios of Muthén, Jo and Brown (in press) are shown in the left-most panels of Figure 1. In the first scenario, shown in the top panels, a 2-class model is considered with a 50% low class that shows only 1/4 standard deviation normative growth (control group growth) in the outcome means, along with a high class that shows 3/4 SD normative growth. For the low class the treatment effect size (posttest standardized mean difference) is small to medium, 0.33. For the high class the effect size is only 0.08. In the second scenario, shown in the bottom panels, a 3-class model is considered with a 10% low class with 1/4 SD normative growth and effect size 0.08, a 70% middle class with 3/4 SD normative growth and effect size 0.33, and a 20% high class with 3/4 SD normative growth and effect size 0.08. The class separation in the random effects means is about 1/2 SD. Perfect randomization is assumed in the sense that Z_i is uncorrelated with η_{0i} in (2). The residual variances for the random effects and the outcomes are the same across treatment groups and trajectory classes. In all cases, the R^2 in the pretest and posttest outcomes is 0.75. The growth rate variance is 1/5 of the initial status variance, a ratio commonly seen in practice. The R^2 in the growth rate as a function of the treatment dummy variable is 20% in the class showing a 0.33 treatment effect size. The design is balanced. The parameter values are given in the Mplus Monte Carlo input in the Mplus Web Note section for this web note at www.statmodel.com/mplus/examples/webnote.html (see Study A and Study B, which add a second posttest occasion). In the right-most panels of Figure 1, the lines for each class are drawn from -1 to +1 SD away from the class mean of the pretest y_1. The ANCOVA lines are obtained from the Monte Carlo analysis described below, using 10 replications with a sample size of $100,000$.

3 Monte Carlo Simulations Using Mplus

The Mplus Web Note section for this web note shows Mplus Monte Carlo simulations for growth mixture modeling with the parameter values used in the Figure 1 scenarios.[1] All cases use $n = 2000$, the sample size of Barnard et al. (in press), and a 50-50 treatment - control group split. Two posttest time points are considered to make the model identified, i.e. three time points total (pretest-posttest analysis is considered in the next two sections). Study A considers the 2-class scenario and Study B considers the 3-class scenario. 500 Monte Carlo replications are used.[2]

In the Mplus Monte Carlo runs, the initial status is labelled i and the growth rate is labelled s. The output shows good coverage for the treatment effect (labelled "s on x"). For the 3-class model the results improve when going from 3 to 4 time points (not shown). The power to reject zero treatment effect is around $0.97 - 0.99$ for the class with

[1]For a summary of the Mplus language, see www.statmodel.com/mplus/language.html.

[2]These analyses can be performed using the free Mplus demo version.

the large effect, while the power is low to detect the smaller effects (see Mplus output column labelled % Sig Coeff). Other scenarios are easily investigated by simple changes in the Mplus input, for example including covariates that predict class membership, using more than one pre-treatment time point, using piecewise growth modeling, adding missingness predicted by covariates and class, or adding compliance classes. Model misspecification can be studied by letting the Monte Carlo data generating model be different from the analysis model.

4 Using Only Pretest-Posttest Information

An interesting question is if the treatment-trajectory interaction of the kind illustrated by the two scenarios above can be detected using the common design of having only two time points, pretest and posttest.

Treatment-trajectory interaction can be explored by a lowess plot (Cleveland, 1979) of posttest on pretest for the treatment and control groups. This procedure was applied to the Baltimore reading data in Brown (1993), where a randomized mastery learning intervention was introduced after a Fall first grade assessment. Relating the first grade Spring test to the Fall pre-intervention test, a beneficial intervention effect was found for children in the middle of the pretest score range. An ANCOVA analysis indicated no interaction and a beneficial intervention effect throughout the range of low and high achieving children. In the two scenarios above, however, the lowess plots are very similar to ANCOVA and do not give a clear indication of the interaction.

ANCOVA mixture modeling is possible in Mplus. For these two scenarios, however, mixture ANCOVA is not able to recover the parameter values well. In contrast, growth mixture modeling gives some interesting findings. To provide a background for this, consider first a conventional single-class growth model such as in (1), (2), and (3). Modify this model to include a possible failure of randomization when $\gamma_0 \neq 0$,

$$y_{ti} = \eta_{0i} + \eta_{1i}\, a_{ti} + \epsilon_{ti}, \tag{4}$$

$$\eta_{0i} = \beta_0 + \gamma_0\, Z_i + \zeta_{0i}, \tag{5}$$

$$\eta_{1i} = \beta_1 + \gamma_1\, Z_i + \zeta_{1i}, \tag{6}$$

where $V(\epsilon_t) = \theta_t$, $V(\zeta_0) = \psi_{00}$, $V(\zeta_1) = \psi_{11}$, $Cov(\zeta_1, \zeta_0) = \psi_{10}$. With $a_{ti} = 0, 1, \ldots$, η_{0i} is the pre-treatment initial status, and the growth model implies

$$Pretest: \quad y_{1i} = \eta_{0i} + \epsilon_{1i}, \tag{7}$$

$$Posttest: \quad y_{2i} = \eta_{0i} + \eta_{1i} + \epsilon_{2i}, \tag{8}$$

$$= \beta_1 + \gamma_1\, Z_i + 1\, \eta_{0i} + \zeta_{1i} + \epsilon_{2i}. \tag{9}$$

Comparing (9) to ANCOVA regression of posttest (y_2) on pretest (y_1) and treatment dummy (Z), ANCOVA amounts to using a covariate y_1 that measures the true covariate η_{0i} with error ϵ_{1i} in (7).

If ζ_0 is uncorrelated with ζ_1 so that ζ_1 is uncorrelated with η_{0i} in (9), the bias in the regression coefficient for the covariate that is not measured with error (Z) is obtained by the general expression for regression with error in one covariate (Carroll, Gallo & Gleser, 1985; Carroll, Ruppert & Stefanski, 1995, pp. 25-26),

$$\gamma_{1,ANCOVA} = \gamma_1 + \gamma_{y2|\eta_0} \, (1 - \lambda) \, \gamma_{\eta_0|Z}. \tag{10}$$

Here, (9) gives

$$\gamma_{y2|\eta_0} = 1, \tag{11}$$

the regression coefficient for y_2 on the true covariate η_{0i} in (9), λ is the conditional reliability

$$\lambda = \frac{V(\eta_{0i}|Z)}{V(\eta_{0i}|Z) + \theta_1} = \frac{\psi_{00}}{\psi_{00} + \theta_1} \tag{12}$$

and

$$\gamma_{\eta_0|Z} = \gamma_0, \tag{13}$$

the regression coefficient for η_{0i} on Z. Note that if treatment and pretest are uncorrelated, $\gamma_0 = 0$ so that $\gamma_{\eta_0|x} = 0$, resulting in no γ_1 treatment effect bias despite the unreliability in the pretest.

How does growth modeling perform with only two time points? The growth model in (4), (5), (6) gives the covariance matrix $V(y_2, y_1, Z)'$:

$$\begin{pmatrix} \theta_1 + \psi_{00} + \gamma_0^2 \sigma_{ZZ} & & \\ \psi_{00} + \psi_{10} + \gamma_0 \gamma_1 \sigma_{ZZ} & \theta_2 + \psi_{00} + \psi_{11} + \psi_{10} + (\gamma_0^2 + \gamma_1^2)\sigma_{ZZ} & \\ \gamma_0 \sigma_{ZZ} & (\gamma_0 + \gamma_1)\sigma_{ZZ} & \sigma_{ZZ} \end{pmatrix}. \tag{14}$$

Excluding the just-identified means, there are 6 pieces of information and 8 parameters. To identify the model, 2 restrictions have to be applied, e.g. $\psi_{11} = 0$ (growth rate variance), $\psi_{10} = 0$. Irrespective of this, the treatment effect is identified and is correctly estimated:

$$\gamma_1 = (\sigma_{y_2,Z} - \sigma_{y_1,Z})/\sigma_{ZZ} = ((\gamma_0 + \gamma_1)\sigma_{ZZ} - \gamma_0 \sigma_{ZZ})/\sigma_{ZZ}. \tag{15}$$

This reduces to the well-known gain score estimator using treatment and control group means,

$$\gamma_1 = E(y_2)_T - E(y_2)_C - [E(y_1)_T - E(y_1)_C]. \tag{16}$$

Note that standard ANCOVA without interaction multiplies the last term in brackets by the common slope for the treatment and control groups. The growth model uses the same 3 variables (y_2, y_1, and Z) as ANCOVA, but differently, without regressing on the fallible y_1, instead regressing both y_2 and y_1 on Z. (14) shows that this γ_1 estimator is not affected by unreliability of y_1 ($\theta_1 > 0$), failure of randomization ($\gamma_0 \neq 0$), or correlated growth factor residuals ($\psi_{10} \neq 0$). This can be taken as an argument in favor of using growth modeling instead of ANCOVA. For related discussions of gain score analysis versus ANCOVA, however, see Laird (1983).

5 Monte Carlo Simulations Using Only Pretest-Posttest Information

It is interesting to study growth mixture modeling with two time points using the same approach as for the single-class growth modeling above, namely by the model misspecification of fixing $\psi_{11} = 0$ and $\psi_{10} = 0$. Using this approach in the two scenarios above, "population runs" were made using 10 replications with $n = 100,000$. This produced the interesting result of treatment effect estimates very close to the true values. Unlike the single-class model, the residual variance misspecification by necessity influences the determination of the classes. In these scenarios, however, the class formation is to a large extent driven by the means of the initial status factor and the growth rate, and the misspecification effect is not large in a large sample. Once the classes have been approximately correctly formed, the correct within-class estimation of (15) is given from (14) despite misspecified residual variance.

Monte Carlo runs with $n = 2,000$ and 500 replications give less good average point estimates, but show reasonably good coverage for the treatment effects in all classes, except the high class in the 3-class scenario. The Mplus outputs for the 2- and 3-class scenarios are presented as Study C and Study D in the Mplus Web Note section. More research is needed to explore how well this procedure works in general, but it appears that perhaps some rough guidance can be obtained from growth mixture modeling with only two time points, at least in large samples. The use of data from more than one posttest occasion is, however, strongly recommended to strengthen the analysis.

References

Barnard, J., Frangakis, C.E., Hill, J.L & Rubin, D.B. (in press). A principal stratification approach to broken randomized experiements: A case study of school choice vouchers in New York City. Forthcoming in *Journal of the American Statistical Association.*

Brown, C. H. (1993). Analyzing preventive trials with generalized additive models. *American Journal of Community Psychology, 21*, 635-664.

Carroll, Gallo & Gleser (1985). Comparison of least squares and errors-in-variable regression, with special reference to randomized analysis of covariance. *Journal of the American Statistical Association, 80*, 929-932.

Carroll, Ruppert & Stefanski (1995). *Measurement Error in Nonlinear Models.* Chapman & Hall.

Dolan, L. J., Kellam, S. G., Brown, C. H., Werthamer-Larsson, L., Rebok, G. W., Mayer, L. S., Laudolff, J., Turkkan, J., Ford, C., & Wheeler, L. (1993). The short-term impact of two classroom-based preventive interventions on aggressive and shy behaviors and poor achievement. *Journal of Applied Developmental Psychology, 14*, 317-345.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74*, 829-836.

Laird, N. (1983). Further comparative analyses of pretest-posttest research designs. *The American Statistician, 37*, 329-330.

Muthén, B. (2002). Statistical and substantive checking in growth mixture modeling. Forthcoming in *Psychological Methods.*

Muthén, B. & Curran, P. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods, 2*, 371-402.

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, B., Jo. B. & Brown, C.H. (in press). Assessment of treatment effects using latent variable modeling: Comments on the New York School Choice Study. Forthcoming in *Journal of the American Statistical Association.*

Muthén, B., Khoo, S.T., Francis, D. & Kim Boscardin, C. (2002). Analysis of reading skills development from Kindergarten through first grade: An application of growth mixture modeling to sequential processes. In S.R. Reise & N. Duan (eds), *Multilevel Modeling: Methodological Advances, Issues, and Applications* (pp. 71-89). Mahaw, NJ: Lawrence Erlbaum Associates.

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.