Assessing Program Effects in the Presence of Treatment-Baseline Interactions: A Latent Curve Approach

CSE Technical Report 464

Siek-Toon Khoo and Bengt Muthén

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)/
Graduate School of Education & Information Studies
University of California, Los Angeles

December 1997

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education, and by grant K02 AA 0023001 from NIAAA, and grant 1 R21 AA10948-01A1 from NIAAA.

ASSESSING PROGRAM EFFECTS IN THE PRESENCE OF TREATMENT-BASELINE INTERACTIONS: A LATENT CURVE APPROACH[1]

Siek-Toon Khoo and Bengt Muthén
CRESST/Graduate School of Education & Information Studies
University of California, Los Angeles

ABSTRACT

The aim of this paper is to explore methods for evaluating the effects of randomized interventions in a longitudinal design. The focus is on methods for modeling the possibly nonlinear relationship between treatment effect and baseline and evaluating the treatment effect taking this nonlinear relationship into account. A control/treatment growth model formulation based on Muthén and Curran (in press) was used as the framework to assess treatment effects. Piecewise linear growth modeling was chosen to study the treatment effects during the different periods of development. A multistage analysis procedure was proposed for assessing treatment effects in the presence of nonlinear treatment-baseline interactions. To avoid biasing effects of measurement errors in the observed baseline scores, initial status factor score estimates from a latent growth model were used in this analysis. Subsets of subjects, based on the form of the nonlinear treatment-initial status interaction, were then used for treatment-control, multiple-group latent growth modeling to assess treatment effects. Standard errors of the estimates from this multistage procedure were obtained by a bootstrap approach. The methods were illustrated using data from the Prevention Research Center at the Johns Hopkins University involving an intervention aimed at improving classroom behavior, the Good Behavior Game (GBG).

1 Introduction

Preventive interventions have the potential to reduce antisocial behavior, minimize substance abuse, and enhance student learning. There is a need for proactive preventive approaches and effective programs for identifying high-risk children and for early preventive intervention.

The focus in the evaluation of program effectiveness is essentially to assess the change that has taken place in the intervention group compared with the

[1] We thank Sheppard Kellam and the Johns Hopkins Prevention Research Center for providing data for illustrations.

change that would have taken place had there been no intervention. Researchers would want to know whether a program has an impact, how strong the impact is, whether the impact is lasting and what the short-term and long-term effects are. In terms of intervention design improvement and future implementation, it would also be useful to know under what conditions the program would be effective, for whom the program is effective, and what interaction effects there are between treatment and the individual background characteristics and experiences.

Intervention effects can be more effectively assessed when subjects are randomized into treatment and control groups to ensure that the control group is as similar to the treatment group as possible. Randomized interventions are costly and not easy to design, arrange, or control. Very often they show small main effects. Instead, effects often appear as interactions due to differential response to the intervention (Brown & Liao, 1996). Interactions between treatment and individual characteristics are common. If a treatment is delivered in groups, then certain group characteristics may influence program effectiveness as well, especially if the treatment involves interactions among the group members. If a meaningful interaction effect exists, then it is important to detect it and to investigate its nature during the intervention trials in order to identify the subgroups that will benefit most from a program and the conditions under which a program works best.

Treatment-baseline interactions represent systematic differences among individuals at different baseline levels in the benefit they reap as a result of intervention. For example, high-risk children may benefit more from an intervention than children at low risk. Children at different baseline levels may require different levels of intervention, and an intervention program that is targeted for all children may not be vigorous enough for those with high-baseline and may be irrelevant for those in the low-risk group. When this happens, the treatment is only effective for individuals in the moderate baseline range. The treatment-baseline interaction in this case is relative to baseline ranges rather than individual baseline levels.

Appropriately targeted programs are more efficient and more cost effective. Interaction effects that describe differential treatment impact need to be examined carefully in order to improve program implementation and design and also to better understand the change in the underlying developmental

processes as a result of the intervention. But before a program can be appropriately targeted, program effectiveness needs to be evaluated in terms of impact and in terms of interactions with individual characteristics and conditions of implementations.

Much of the research on studying change, especially in the education field, has been based on the two-time-point, pretest-posttest design. Data with only two time points generally do not provide enough information for studying change (Bryk & Weisberg, 1977; Rogosa, Brandt, & Zimowski, 1982). It is advantageous to study change in developmental processes over extended periods. A longitudinal design with several outcome measurements collected over extended periods, ideally covering the pre-intervention period, the period during intervention, and the post-intervention period, would be more appropriate for studying developmental changes. Longitudinal designs provide information about individual patterns of change over time and make it possible to evaluate program effectiveness in terms of growth trajectories.

Recent developments in the statistical theory of random effects modeling (Bock, 1989; Bryk & Raudenbush, 1987, 1992; Goldstein, 1986; Laird & Ware, 1982; Strenio, Weisberg, & Bryk, 1983) enable more integrated and flexible approaches for the study of individual differences in change and the modeling of individual growth. Growth modeling techniques in the context of latent variable modeling (McArdle & Epstein, 1987; Meredith & Tisak, 1984, 1990; Muthén, 1991, 1993, 1997; Muthén & Curran, in press; Willett & Sayer, 1994), which combine random effect modeling with the flexibility of latent variable modeling, lend powerful tools to the study of developmental changes. Latent growth modeling can be used to model treatment effects of randomized preventive interventions in a flexible way.

In research studies where there are distinct developmental periods, it may be advantageous to model them as several stages of growth. The different periods may have different growth patterns. Intervention programs with distinct phases such as the pre-intervention period, the intervention period and the post-intervention period are examples of these. Different factors may be relevant for influencing growth in the different phases, or the influence of a covariate may vary in the different phases. Modeling the similarities and differences of the phases would provide more insightful analyses.

In studying treatment-baseline interaction, the most relevant relationship is that between the initial status and the change due to intervention. The treatment effect is the added growth over and above the growth due to natural maturation. Muthén and Curran (in press) tested a linear relationship between the treatment effect and the initial status within the growth model. When an intervention program primarily benefits individuals in a certain baseline range, then the relationship between treatment effect and baseline may not be linear for the entire range. For one thing, it is unlikely that there is treatment-baseline interaction in the group where there is no treatment effect.

The purpose of this paper is to investigate longitudinal methods for modeling the possibly nonlinear relationship between treatment effect and baseline and evaluating the treatment effect taking this nonlinear relationship into account. The Muthén-Curran method (Muthén & Curran, in press) for program evaluation will be used as the framework for assessing treatment effects. A multistage analysis procedure is proposed for assessing the treatment effects in the presence of nonlinear treatment-baseline interactions with respect to baseline groups.

The methods are investigated and illustrated using data from a randomized intervention program carried out by the Prevention Research Center at Johns Hopkins University. The Good Behavior Game is a classroom-based intervention program that was designed to reduce aggressive behavior in classrooms for elementary and middle school children. The illustration will examine the treatment effects in two developmental stages of growth. Differential treatment effects will be investigated with respect to baseline groups.

2 The Latent Curve Framework

Recent methods used in the study of individual change over time draw on statistical techniques which have been given several different terms, for example, random effects models or mixed effects models (Laird & Ware, 1982), and hierarchical linear models (HLM; Bock, 1989; Bryk & Raudenbush, 1987, 1992; Goldstein, 1986, 1995; Strenio et al., 1983).

Use of random effects models for growth modeling has been extended and applied in the context of latent variable modeling (McArdle & Epstein, 1987; Meredith & Tisak, 1984, 1990; Muthén, 1991, 1993, 1997; Willett & Sayer, 1994).

This development goes beyond the conventional auto-regressive models of structural equation modeling (see, e.g., Wheaton, Muthén, Alwin, & Summers, 1977) and offers great flexibility in dealing with multivariate outcomes, multiple processes, measurement models in the covariates, and modeling of mediational effects.

In random coefficient growth modeling, level-1 gives the within-subject model and level-2 gives the between-subject model. At level-1, the observed status Y_{it} of an individual i at time t is expressed as the sum of the individual's growth trajectory plus a residual term (ε_{it}) representing random error at time t. A basic model with linear growth can be expressed as

$$Y_{it} = \alpha_i + \beta_i x_{it} + \varepsilon_{it} \quad , t = 1, 2, ..., T \tag{1}$$

where α_i and β_i are the random growth parameters, the initial status and growth rate that vary over individuals, and x_{it} is age, time points or other time-related variables. In educational data, x_{it} usually denotes grade or testing occasion and does not vary across individuals for a given t. At level-2, each of the growth parameters from level-1 becomes an outcome variable, varying from the group mean values, α and β, as a function of some individual background variables Z_i:

$$\begin{aligned} \alpha_i &= \alpha + \gamma_\alpha Z_i + \delta_{\alpha i} \\ \beta_i &= \beta + \gamma_\beta Z_i + \delta_{\beta i} \end{aligned} \tag{2}$$

where $\delta_{\alpha i}$ and $\delta_{\beta i}$ are the residuals associated with α_i and β_i respectively, and γ_α and γ_β are the regression coefficients.

It is possible to model the growth process within the existing latent variable framework for situations where x_{it} does not vary across individuals for a given t—for example, in educational data where all the individuals under study share the same testing occasions or grade. When the time-related variable does not vary across individuals, it allows α_i and β_i to be treated as latent variables and x_t as a path coefficient. Equation (1) above can be written as

$$Y_{it} = \alpha_i + x_t \beta_i + \zeta_{it} \quad , t = 1, 2, ..., T \tag{3}$$

where ξ_{it} is the individual specific residual at time t. This essentially sets the stage for modeling growth using latent variable modeling. With $x_{t=1}$ set to 0, α_i is an individual's initial status factor and β_i the growth rate factor. In order to achieve this formulation in the latent variable context, the path coefficients for the initial status factor α_i are all set to 1. In conventional latent variable notation, the measurement part of this model is given by

$$\mathbf{Y}_i = \Lambda\eta + \varepsilon_i \tag{4}$$

where $\mathbf{Y}_i$ is a T x 1 vector of observed outcomes for the T time points, Λ a T x 2 matrix of factor loadings, η a 2 x 1 vector of latent variables representing the growth parameters and ε_i a T x 1 vector of error terms. If we take T to be 4, we have

$$\mathbf{Y}_i = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} \; ; \; \Lambda = \begin{bmatrix} 1 & x_{t=1} \\ 1 & x_{t=2} \\ 1 & x_{t=3} \\ 1 & x_{t=4} \end{bmatrix} \; ; \; \eta = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \text{ and } \varepsilon_i = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

In this formulation, the Y_{it}'s are strung out as a multivariate vector, and we have a regular single-level latent variable model where all the flexibilities of structural equation modeling can be applied. As latent variables, α_i and β_i can be modeled as endogenous or exogenous variables as in the usual latent variable modeling. To realize the full flexibilities of latent variable modeling, the y-intercept vector ν can be added to the above equation to set the stage for growth modeling with multiple indicators at each time-point or multiple-group modeling,

$$\mathbf{Y}_i = \nu + \Lambda\eta + \varepsilon_i \tag{5}$$

In a single group growth analysis with single measurement at each time point, ν can be set at zero while estimating the mean of α_i. Alternatively, the same end can be achieved by having a common intercept across time, while fixing the mean of α_i at zero. The use of the intercept ν in a multiple-group setting will be described in section 2.2.

2.1 Evaluation of Longitudinal Program Effects

Evaluation of treatment effects in the growth modeling context has been made very accessible with the general availability of sophisticated multilevel modeling software. The methodology advances in latent growth modeling using available structural equation modeling software add modeling flexibility.

In longitudinal studies with at least three waves of data, longitudinal program effects can be evaluated using growth modeling by estimating the effects of the dichotomous treatment/control variable on the growth rate. In (2) above, if Z_i is the dichotomous treatment variable, then the coefficient γ_α gives the difference of the initial status means between the control and the treatment group, while the coefficient γ_β gives the mean shift in the growth rate due to treatment. The same applies in the latent growth framework. In this way, the treatment effect is modeled as a fixed effect.

Within the latent variable growth modeling framework, Muthén and Curran (in press) describe a general formulation for evaluating randomized intervention studies. They propose a multiple-population analysis approach for the evaluation of treatment effects by introducing an extra factor in the treatment group that captures the added growth rate due to treatment. This approach, which treats the treatment effect as random rather than fixed and allows it to vary across individuals, is described in the next section.

2.2 The Two-Group (Control/Treatment) Formulation

In a randomized intervention setting, the control group acts as a proxy for the treatment group with the assumption that the treatment group would display the same growth patterns as the control group if there had not been an intervention. The treatment effect is evaluated by comparing the growth patterns of the treatment group to the normative growth patterns of the control group. The control and treatment groups are analyzed in a two-group setting. The two-group formulation treats the control group and the treatment group as two different populations analyzed in a multiple-group latent variable analysis in line with Jöreskog and Sörbom (1979). The control group is used to establish the normative growth parameters (the initial status factor and the normative growth rate factor); the treatment group has an added growth rate factor that captures treatment effect in addition to these normative growth factors. This is realized by constraining the normative growth factors (α_i, β_{Ni}) in the treatment group to

be equal to those in the control group. The added growth rate factor (β_{Ei}) that captures the treatment effect is omitted in the control group. The basic growth model can be expressed as

Control group:

$$Y_{it} = \nu + \alpha_i + x_t \beta_{Ni} + \zeta_{it} \quad , t = 1, 2, ..., T \tag{6}$$

Treatment group:

$$Y_{it} = \nu + \alpha_i + x_t \beta_{Ni} + x_{Et} \beta_{Ei} + \zeta_{it} \quad , t = 1, 2, ..., T \tag{7}$$

A nice feature of this formulation is that it models the treatment effect as a random variable that varies across individuals in the treatment group. Covariates can be introduced into the model to explain the variation in this variable (β_{Ei}), thus modeling the influence of individual characteristics in the degree of response to treatment. This is in effect modeling interaction effects between the treatment and the covariates. By virtue of the latent variable framework, it also allows the regression of the treatment effect on the initial status factor, thus modeling the treatment-baseline interaction. With the flexibility of multiple-group modeling in the latent variable framework, non-equivalent groups can be allowed to have different initial status and the difference tested for significance. This can be achieved by using the y-intercept to capture the initial status mean of the first group by having a common y-intercept across time and across groups while holding the initial status mean of the first group at zero. The initial status mean of the second group then represents the mean difference that can be tested for significance.

The "added growth due to treatment" factor β_{Ei} can be regressed on the initial status factor or other covariates to capture the interaction effects. β_{Ei} can be expressed as a function of initial status and possibly other covariates:

$$\beta_{Ei} = \gamma_\alpha \alpha_i + \gamma_1 Z_1 + \gamma_2 Z_2 + \delta_i \tag{8}$$

Here, γ_α captures the linear interaction between the initial status and the treatment effect on the growth rate. This is possible because the regression only pertains to the treatment group, thus achieving the same aim as using a dichotomous treatment/control variable. Furthermore, nonlinear interactions

with the covariates can be dealt with in the observed covariates. Nonlinear interaction between the initial status and the treatment effect cannot be easily modeled as yet using available structural equation modeling (SEM) software. This is because conventional SEM does not allow nonlinear relations involving latent variables. A multistage procedure for evaluating treatment effects in the presence of nonlinear treatment-baseline interactions is described in the next section.

3 A Multistage Procedure

Figure 1 displays four plots depicting four hypothetical situations to illustrate four possible types of treatment-baseline interactions. In all the panels, the treatment is one that is designed to bring about a decrease in the outcome variable y.

In each of these plots, for a given baseline value, it is possible to compare the y values of two individuals, one from the treatment group and one from the control group. If we take the y value on the control curve as the normative value without an intervention, the difference between the two y's will be the "treatment effect" or the "y drop" for an individual in the treatment group at that baseline level. If there is no treatment-baseline interaction, the two lines would have been running parallel, as in plot A. Here, the treatment effect is nearly the same at all baseline levels. Plot B shows that the treatment effect increases along the baseline and the increase varies linearly with baseline, indicating a linear treatment-baseline interaction. Plot C shows a treatment effect that first increases with baseline and then decreases after a certain baseline level, showing a nonlinear relationship in the interactions.

Plot D shows an interesting situation, where there is hardly any treatment effect for an individual with baseline level below the value of 2. The treatment is only effective for individuals with higher baseline, and for these individuals, the treatment effect is quite uniform. This also shows a nonlinear baseline interaction; the treatment is effective for a certain baseline range and not for the other with no treatment-baseline interactions within each range. In this case, the treatment effects would be underestimated if the treatment-baseline interaction is ignored and overestimated if the interaction is assumed linear.

Because conventional SEM does not allow nonlinear relations involving latent variables, the nonlinear interaction between the initial status and the

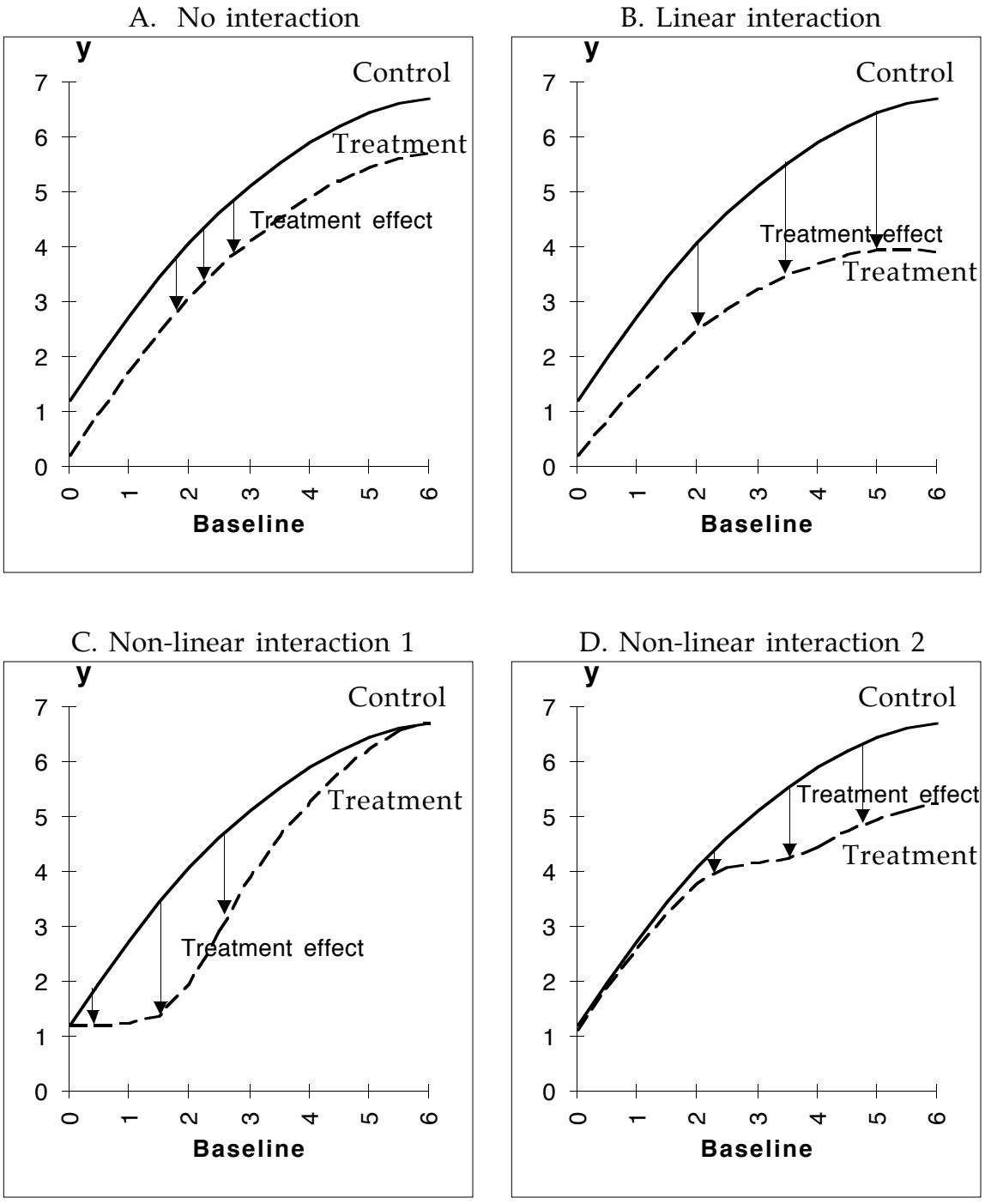

Figure 1. Treatment-baseline interactions.

treatment effect cannot be expressed in terms of a single functional form. However, most nonlinear relationships can be approximated using piecewise linear forms if the general form of the nonlinear relationship is known. A possible approach for modeling nonlinear treatment-baseline interactions would be to subset the sample along the baseline continuum into subgroups such that the interaction is linear within each subgroup. The challenge to this method is to find the appropriate cutpoints.

It is proposed here that a multistage procedure be carried out. First, a better estimation of individual baseline levels is obtained to better assess the nature of the interactions. Next, the nature of the interactions is assessed, and in the presence of nonlinear treatment-baseline interactions, the baseline groups with differential treatment effects are identified. Lastly, the treatment effect within each baseline group is evaluated.

3.1 Estimation of Baseline and Assessing the Nature of the Interactions

A class of nonparametric regression procedures called generalized additive models (GAM; Hastie & Tibshirani, 1990) can be a useful tool for exploring the presence of nonlinear relationship and its shape, for example, in identifying the nature and direction of treatment-baseline interactions. It may also be useful in the determination of cutpoints. Brown (1993) describes how GAM can be used for examining intervention effects and for assessing the potential interactions between the intervention and individual characteristics including baseline. Generalized additive models allow a dependent variable Y to change in a nonlinear fashion with each predictor X_r with

$$E(Y) = \sum f_r(X_r) \ , \qquad \text{r} = 1, 2, ..., \text{R} \tag{9}$$

where the f_r are individual smooth functions of X_r.

GAM can be used to explore the nature of the treatment-baseline interaction in preliminary analysis to study the relationship between later measurements and baseline. The outcome variable at a later time point can be used as a dependent variable fitted to a smooth function of the baseline used as the predictor as those shown in Figure 1. An advantage of using GAM is that the shape of the function in a certain baseline range is not much affected by the shape along other parts of the baseline, therefore the effects of outliers are mainly local.

Using the fallible first measurement as a covariate can, however, produce biased results as in using a single pretest measure in an ANCOVA analysis (Reichardt, 1979). The initial status as captured in a growth model would make a better proxy for baseline than a single premeasure. It is more reliable because a series of measurements across time provide more information on the initial status of an individual than just a single measurement. If the status across time is growing at a certain rate, then each of the observed outcomes across time is a function of the initial status and the growth rate. Each outcome provides information for and contributes to the estimation of the initial status and the growth rate. This is analogous to having multiple indicators contributing to the estimation of a factor in a confirmatory factor analysis, thus purging the measurement errors.

The initial status is a unobserved latent variable and cannot be used in the GAM analysis directly. The value of the initial status factor in the growth model can, however, be estimated for each individual based on the growth model using factor score estimation and then used in the GAM exploration.

Consider a vector of observed outcomes Y_i, let ν be the vector of y-intercepts, Λ the matrix of factor loadings, η_i the vector of latent variables, and ε_i the vector of error terms. In a factor analysis model with mean structure, $Y_i = \nu + \Lambda\eta_i + \varepsilon_i$, the estimation of factor scores $\hat{\eta}_i$ by the regression method is given by Lawley and Maxwell (1971) as a function of observed outcome Y_i:

$$\hat{\eta}_i = \alpha + \Psi\Lambda'(\Lambda\Psi\Lambda' + \Theta)^{-1}(Y_i - \nu - \Lambda\alpha) \tag{10}$$

where $\alpha = E(\eta)$, $\Psi = V(\eta)$, and $\Theta = V(\varepsilon)$.

This formula can be applied to estimating factor scores for the growth parameters in a growth model. The latent growth model is in effect a confirmatory factor model with a mean structure, where the Y_i's are repeated measurements of the same variable across time and Λ is structured in such a way that the factors capture the initial status and the growth rate.

How well the factor scores can be estimated is indicated by the correlations between the estimated factor scores and the true factor values. These correlations are called the factor score determinacies and can be calculated from the covariance matrix between the true and the estimated factor scores. The covariance matrix is given by

$$\mathrm{Cov}(\hat{\eta}, \eta) = \Omega = \Psi\Lambda'(\Lambda\Psi\Lambda' + \Theta)^{-1}\,\Lambda\Psi \tag{11}$$

Let ω_{jj} be the jth diagonal elements of Ω. Then for latent variable η_j,

$$\omega_{jj} = \mathrm{Cov}(\hat{\eta}_j, \eta_j) = V(\hat{\eta}_j)\,, \tag{12}$$

and the factor score determinacy is therefore given by

$$\omega_{jj} / \mathrm{SQRT}\,[V(\hat{\eta}_j)V(\eta_j)] = \mathrm{SQRT}\,[\omega_{\mathrm{jj}} / V(\eta_j)]. \tag{13}$$

The regression method for estimating factor scores is equivalent to the customary empirical Bayes (EB) estimator in growth modeling (e.g., Bock, 1989). Factor scores can also be estimated with covariate information in addition to the observed outcome. Using the regression method factor score estimate as a covariate gives regression coefficients that are consistent and unbiased (Muthén & Hsu, 1993; Tucker, 1971).

For valid comparisons of the control group and the treatment group, the estimated factor scores for both groups need to be on the same metric, that is, on the same interval scale. This is achieved when the parameter values for use in the calculation of factor scores are estimated using equal measurement structure in line with multiple-group CFA theory (Jöreskog & Sörbom, 1979; Sörbom, 1982). The measurement parameters that need to be equal across groups are the factor loadings, Λ, and the y-intercepts, ν. This is the so-called measurement invariance condition. In the growth modeling framework, this condition means that the initial status factor and the growth rate factor need to have the same meaning in both groups. In general, this condition can be achieved by using the multiple-group approach, analyzing the two groups simultaneously. The Λ is set to be the same across groups and the y-intercepts fixed at zero across time and across groups, while estimating the factor means without constraints. An alternative is to estimate the y-intercepts but constrain them to be equal across time and across groups, while holding the initial status factor mean of one group, usually the control group, to zero and estimating the initial status factor mean of the treatment group. The advantage of the second method is that the initial status factor mean of the treatment group gives the difference between the means of initial status of the two groups.

3.2 Subsetting the Sample and Multiple-Group Comparisons

If the sample can be subsetted along the baseline continuum into subgroups such that the relationship between initial status and treatment effect is linear within each subgroup, then latent growth modeling can be carried out in a multiple-group setting so that each subgroup treatment/control pair can have its within-group interaction effects modeled. For example, plot D in Figure 1 would motivate subsetting the control and treatment group by dividing at baseline value 2.

Within each pair, there are the control group and the treatment group with comparable baseline. Full growth modeling utilizing all time points can then be carried out to evaluate treatment effects comparing growth patterns of the low-baseline treatment group with the low-baseline control group, and the high-baseline treatment group with the high-baseline control group. With the growth modeling, it is then possible to see how the growth trajectories of control and treatment group individuals compare over an extended period for the different baseline ranges.

The GAM exploration can help in deciding the cutpoints in the subsetting of the sample to be used in this paired multiple-group analysis. The decision on the cutpoint may not be clearcut. Given that the GAM analysis is based on snapshots at two time points, results may depend on which time point is chosen as the dependent variable. Sensitivity analysis using different time points may need to be carried out unless the choice is substantiated by good research reasons.

The uncertainty involved in the multistage procedure, including the subjective decision on choice of cutpoint, should be taken into account in making inferences in the final growth analysis and the evaluation of treatment effects. When this is not done, it is very likely that the standard errors would be underestimated and the inference may be too optimistic (Faraway, 1992; Weisberg, 1985, p. 229). One possible solution would be to carry out a bootstrapping procedure to obtain more realistic estimates of the standard errors.

This multistage procedure, which involves paired subgroup analysis for evaluating treatment effects in the presence of nonlinear interactions between treatment and baseline, will be illustrated using the Johns Hopkins Prevention Research Center data in the next section. The illustration will include estimation of bootstrap standard errors.

4 Illustration of the Multistage Procedure

The Good Behavior Game intervention program was part of a larger preventive intervention trial completed by the Johns Hopkins Prevention Research Center. The longitudinal study started in 1985 with two years of classroom-based randomized preventive intervention in the first and second grades. Data were collected from the first grade through the eighth grade from about 1,000 students, during both the two years of intervention period and six years of post-intervention period. Schools involved were randomly assigned to either an intervention or a non-intervention condition (external control). The intervention schools were randomly assigned to one of the two intervention programs: the Mastery Learning (ML) program for improving reading proficiency or the Good Behavior Game (GBG) program that was a classroom-based intervention designed to reduce aggressive behavior in the classrooms. Within the intervention schools, children entering first grade were assigned randomly to classrooms, and one classroom per school was selected as a control classroom (internal control).

The data used for this study are pertinent to the GBG program. Children were measured in their aggression level four times during the two-year intervention period, in the fall and the spring of the first and second grades. After the intervention had ended, the children were evaluated once a year in the spring during the six-year follow-up. The primary measure of the aggressive behavior was the Teacher Observation of Classroom Adaptation-Revised (TOCA-R) scale. The scale includes items such as *breaks rules*, *fights*, *harms property* and *loses temper*.

The GBG program is aimed at the social adaptational process in the classroom related to rules and authority; the hypothesis was that social maladaptive aggressive behavioral responses were malleable through changing the social adaptational process in the classroom, and that the changes would remain in the child's coping responses to later social task demands concerned with rules and authority. Details of the intervention program and measurement used can be found in Kellam, Rebok, Ialongo, and Mayer (1994).

4.1 Previous Research Findings

Kellam et al. (1994) studied the course of aggression among the Cohort 1 male and female children. They found that the GBG had increasing effects as the

level of aggression rose in the fall of first grade, but only among the males, and only among the males at or above the median on aggression in the first grade. Analysis of covariance was used to test the impact of GBG separately by gender and then by the subpopulations of male and female at different levels of initial aggression taking the measurement at the Fall of Grade 1 as baseline and comparing the children's aggression level at Grade 6. Their results show a possible treatment-baseline interaction. They concluded that there was a treatment effect only for males who were at or above the median level of baseline aggression and that the effect increased with the increase in baseline. Their findings were based on data at two time points, comparing aggression level at Grade 1 and Grade 6.

Muthén and Curran (in press) used data from all eight time points from Grade 1 to Grade 6 for the Cohort 1 males to illustrate their two-group formulation for modeling intervention effect, comparing longitudinal growth trajectories between the control group and the treatment group. They fitted a quadratic trajectory to the data and demonstrated how linear interactions between the initial aggression status and the treatment effect can be modeled within the context of the two-group control/treatment formulation. They found significant interaction between the initial aggression level and the treatment effect. They also found significant treatment effect in the linear growth rate.

For direct comparison of results, the same dataset and measurement scale were used for the current method illustration, and the clustering of the data ignored as in the Muthén-Curran study. The subset of data used include only the Cohort 1 males and only children who stayed in the same treatment or control condition for the two years of the intervention period. There were 75 such male children in the GBG treatment group. The GBG control and ML control were pooled to form the control group with 111 male children. These were the children in the intervention schools who did not undergo any GBG treatment program.

In the GBG intervention study, it was found that the third grade was a transition period worth further investigation (Kellam et al., 1994) where mean aggression was observed to have peaked during the spring of Grade 3 with larger variance. In this illustration, a piecewise linear growth model (Bryk & Raudenbush, 1992; Seltzer, Frank, & Bryk, 1994) was considered for studying the treatment effects during the different periods of development in view of the

findings in Kellam et al. (1994). The two phases are from Grade 1 to Grade 3 and from Grade 3 to Grade 6 with a large part of the early phase coinciding with the intervention period. Using the piecewise linear growth model, the shorter and longer term intervention effects can be assessed; also, the nature and size of the treatment-baseline interactions during the two periods can be determined separately.

4.2 The Growth Model and Preliminary Analyses

The two-stage piecewise linear growth model has an initial status factor and two growth rate factors, one for each piece. Model specifications of piecewise linear growth models in the latent curve context are discussed in Khoo (1997). Figure 2 shows the path diagram of this model. The initial status factor was allowed to covary with the two growth rate factors. The input for LISCOMP program (Muthén, 1987) is available from the authors.

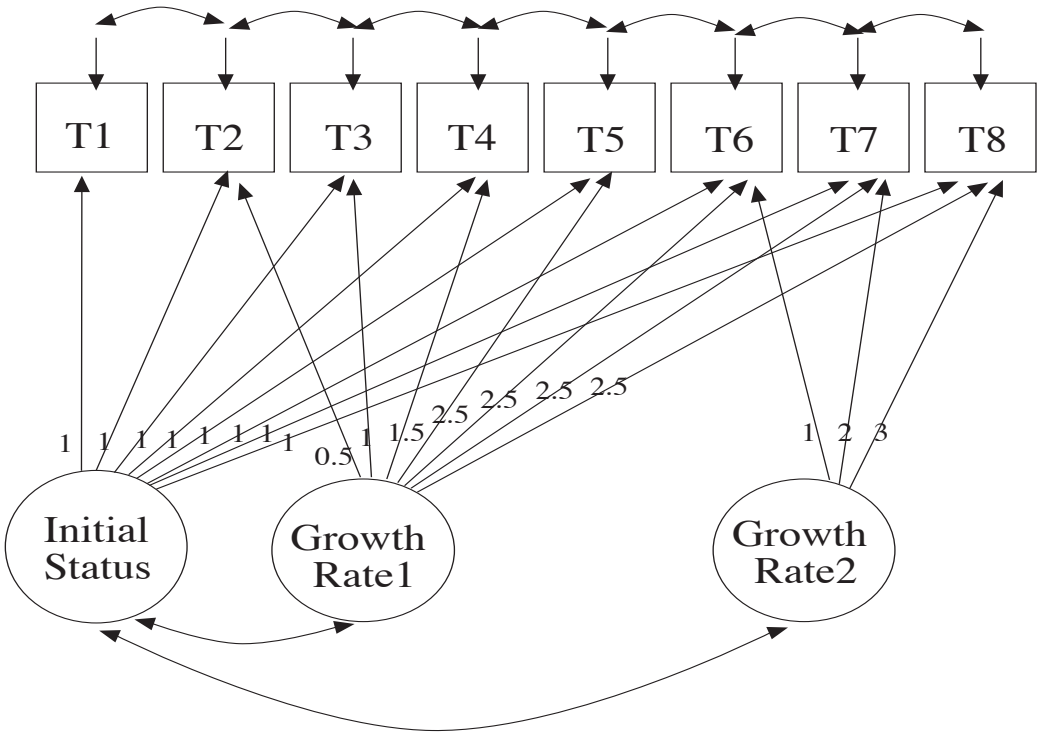

Figure 2. A two-stage piecewise linear growth model.

The data were first analyzed separately for the control group and the treatment group. This piecewise model fitted well for both groups; better for the control group, $\chi^2(21) = 25.59$, $p = 0.22$, than for the treatment group, $\chi^2(21) = 30.70$, $p = 0.08$. The early growth rate means (during intervention) are positive and significant (at 0.05 level) for both groups. For the control group, the mean aggression level remained nearly constant after the spring of the third grade whereas for the treatment group, it decreased gradually.

The two groups were analyzed jointly in a multiple-group setting, with added growth rate factors to model the treatment effects as described in Muthén and Curran (in press). The treatment group had the same three growth factors as the control groups, which established the normative growth, and two added growth rate factors to capture the growth rate differences due to intervention. The path diagram for this two-group model is shown in Figure 3. The equality of the initial status between the two groups was tested in terms of the mean and the variance of the factor. Results showed that only the equality constraint across group on the variance of the initial status factor needed to be relaxed. The added growth rate factors were regressed on the initial status factor to test for linear treatment-baseline interactions in the two different stages of growth. This model fitted reasonably well, $\chi^2(50) = 60.87$, $p = 0.14$. Estimates and standard errors are shown in Table 1 and the mean trajectories are shown in Figure 4.

There is significant variation in the initial status factor in both the control and the treatment groups. The early growth rate factor has a mean that is significantly greater than zero and a significant variance. In the later stage, from the third grade to the sixth grade, the growth rate factor mean is not significant, but there is significant variation in the growth rate. These results show that neither of the regression intercepts of the early added growth rate and the later added growth rate is significantly different from zero. Because the initial status means are set at zero for both groups, these regression intercepts are the added growth rate means, indicating that there are no overall treatment effects in the early or the later stages. In both the growth stages, the initial status influence on the added growth rate factor in the treatment group was found to be negative (-0.134, -0.098) and significant ($t = -2.949$, $t = -2.093$), showing substantial treatment-baseline interaction. The more aggressive individuals benefit more from the treatment; their aggression level decreases at a higher rate. This finding is consistent with that of Kellam et al. (1994) and Muthén and Curran (in press).

Table 1

Two-Group (Control/Treatment) Analysis: Two-Stage Piecewise Growth Model

		Control group (n = 111)			Treatment group (n = 75)	
		$\chi^2(50) = 60.87$, $p = 0.14$				
Outcome variables						
Intercept		2.014	(0.023)	=	2.014	(0.023)†
Growth factors						
Initial status						
	Mean	0.0	(*)		0.0	(*)
	Variance	0.624	(0.106)		1.148	(0.229)
Early growth rate						
	Mean	0.128	(0.033)	=	0.128	(0.033)
	Variance	0.036	(0.011)	=	0.036	(0.011)
Later growth rate						
	Mean	0.005	(0.027)	=	0.005	(0.027)
	Variance	0.034	(0.015)	=	0.034	(0.015)
Growth factor covariances						
Initial status - Early growth rate		0.0	(*)		0.0	(*)
Initial status - Later growth rate		0.0	(*)		0.0	(*)
Regression on initial status						
Added early growth rate						
	Intercept				0.006	(0.054)
	Slope				-0.134	(0.046)
Added later growth rate						
	Intercept				-0.060	(0.054)
	Slope				-0.098	(0.047)
Residual variances for outcome variables						
Time 1		0.441	(0.085)		0.441	(0.117)
Time 2		0.452	(0.076)		0.378	(0.095)
Time 3		0.417	(0.068)		0.528	(0.110)
Time 4		0.519	(0.081)		0.721	(0.134)
Time 5		0.421	(0.078)		0.641	(0.130)
Time 6		0.404	(0.078)		0.745	(0.139)
Time 7		0.272	(0.088)		0.268	(0.072)
Time 8		0.261	(0.131)		0.631	(0.146)

* Parameter is fixed in this model.
† Standard errors are given in parentheses.
= Parameter set equal across group.

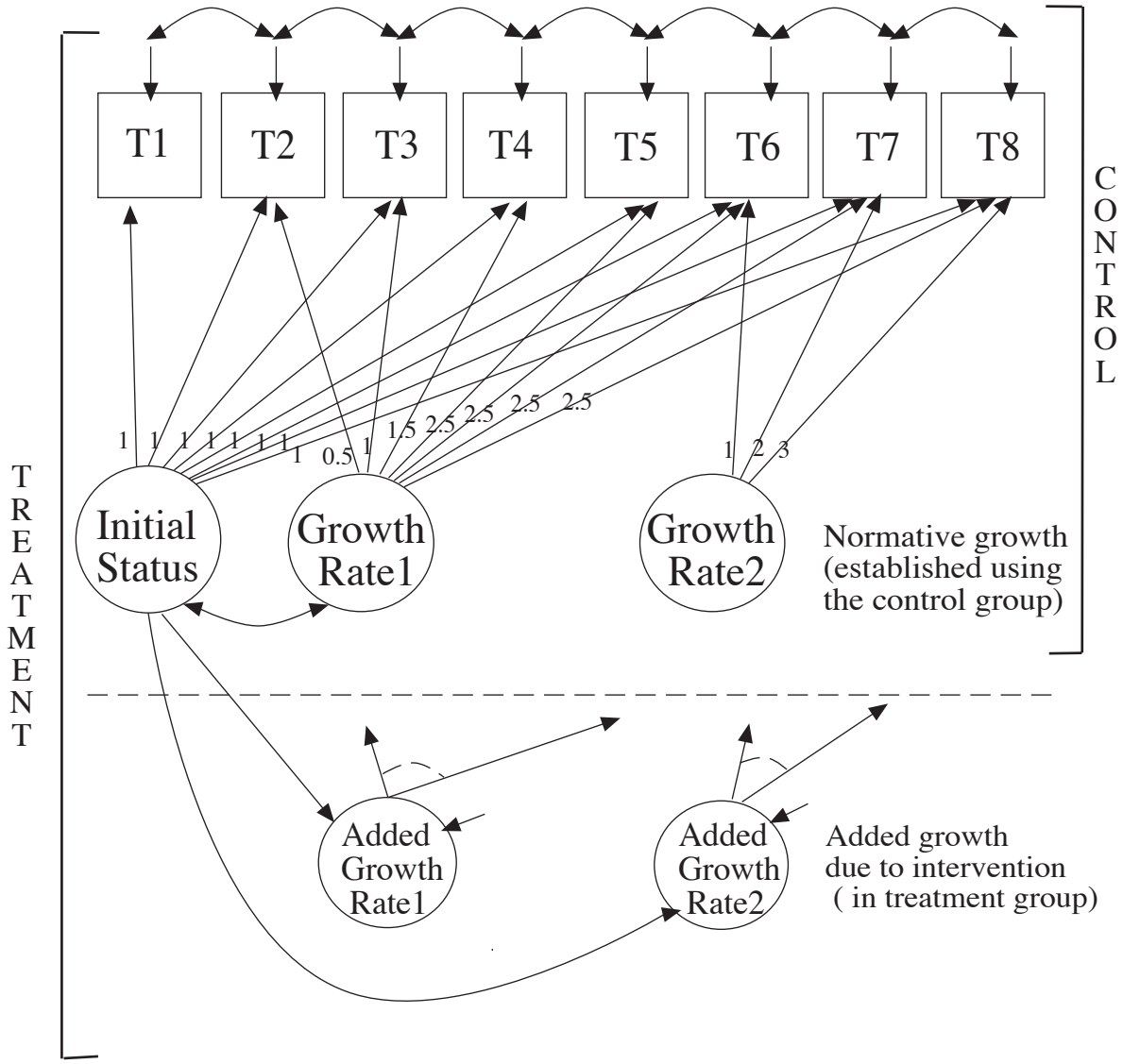

Figure 3. A two-stage piecewise linear growth model in a multiple-group (control/ treatment) setting.

As shown in Muthén and Curran (in press), it is easy to conclude that there is no overall treatment effect since the treatment effect is seen only through the interaction effect with baseline. In order to identify the subgroup who benefited substantially from the intervention and to assess the actual treatment effects in this subgroup, further analyses were carried out using the multistage procedure proposed in Section 3.

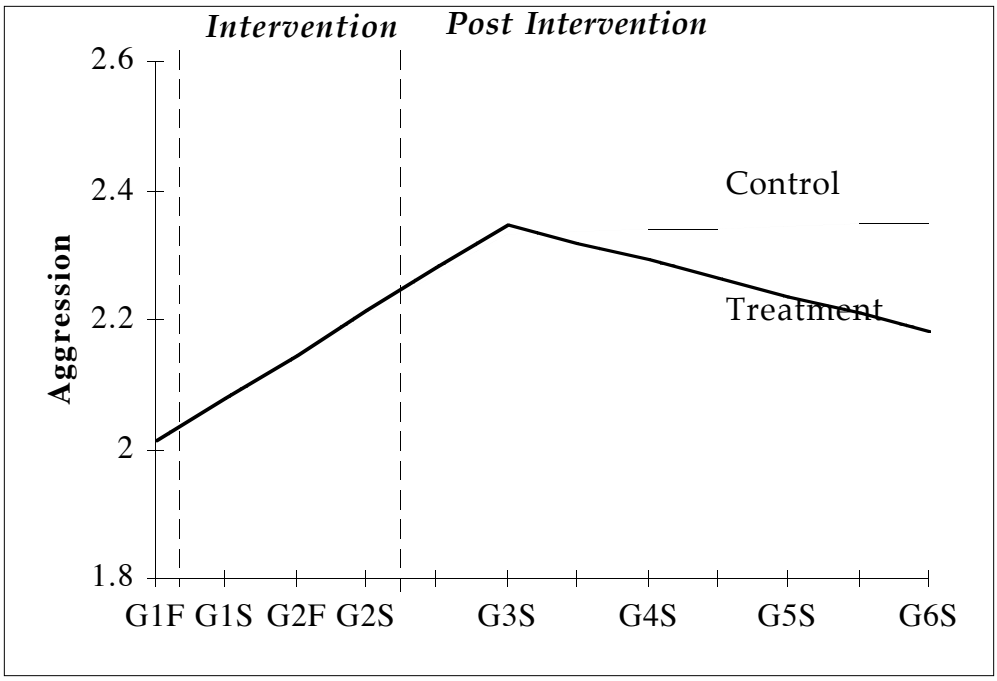

Figure 4. Mean growth trajectories (simultaneous joint analysis).

4.3 Baseline Estimation and Assessing Nature of Treatment-Baseline Interactions

Data from the first four time points were used for the estimation of baseline. Three waves of data would have been sufficient for a linear growth model to estimate the initial status, but because there were four waves of data covering the two-year intervention period, all four time points were used. The two groups were analyzed simultaneously with the same Λ and with the y-intercept fixed at zero across time and across groups. This ensures measurement invariance across the control and treatment groups and therefore puts the factor scores of the two groups on the same metric. In addition, because the control and treatment group shared the same measurement occasions and were measured under the same conditions, the error variances of each outcome variable were equated across groups. This constraint was tested to not significantly affect the fit of the model. No other parameters were constrained to be equal. There was a reasonably good model fit, $\chi^2(8) = 10.41$, $p = 0.24$. The first measurement was predicted by the model with an R^2 of 0.68 for the control group and 0.80 for the treatment group. The initial status factor scores were calculated for each individual based on the regression method using the estimated parameters and the observed outcome

variables. Factor score determinacy, which measures the correlation between the estimated factor scores and the true factor values, was 0.88 for the control group and 0.95 for the treatment group. These factor score determinacy values were considered high enough for the purpose of ranking the subjects for subsetting.

Figure 5 shows a GAM plot of aggression level at the spring of Grade 6 (Time 8) fitted to a smooth function of the factor score estimated baseline for the treatment group and the control group. The black dots on the fitted lines show individual fitted values. The 95% confidence bands are shown around the fitted lines. While the fitted line for the control group looks nearly linear with aggression at Time 8 increasing with baseline, the treatment group shows a decline in the aggression level for subjects with baseline level above the mean. With this plot, it is possible to compare the final aggression level of two individuals, one from the treatment group and one from the control group, who have the same baseline. The difference between the aggression levels will be the "treatment effect" or the "aggression drop" for an individual in the treatment group at that baseline level if we take the control aggression level as the aggression level of the individual if there had not been an intervention. If there is no treatment-baseline interaction, the two lines would have been parallel, with the "aggression drop" due to treatment nearly the same at all baseline levels. The plot shows that the "aggression-drop" is not uniform along the baseline and does not vary linearly with the increase of baseline. An effect of treatment is not apparent for children who are below the mean aggression level. A treatment effect only appeared for individuals whose baseline aggression was above the mean. This is consistent with the Kellam et al. (1994) findings, but within the high group for whom the treatment was effective, the treatment effect appears to vary nonlinearly with baseline. On the extreme right end of the plot, there are two individuals who appear to be outliers. How the growth of these individuals may affect the analysis results will be investigated.

Because the generalized additive models are based on two time points only, the treatment effects shown may be sensitive to the choice of the dependent variable. As a sensitivity check, GAM plots varying as dependent variable the outcome at the spring of Grade 1 (Time 2) to the spring of Grade 5 (Time 7) were also plotted and shown in Figure 5B and Figure 5C. From the plots, we can see that the same interaction effects start showing at Time 3, but they do not appear to be substantial until Time 7 and Time 8. The aggression curve for the treatment

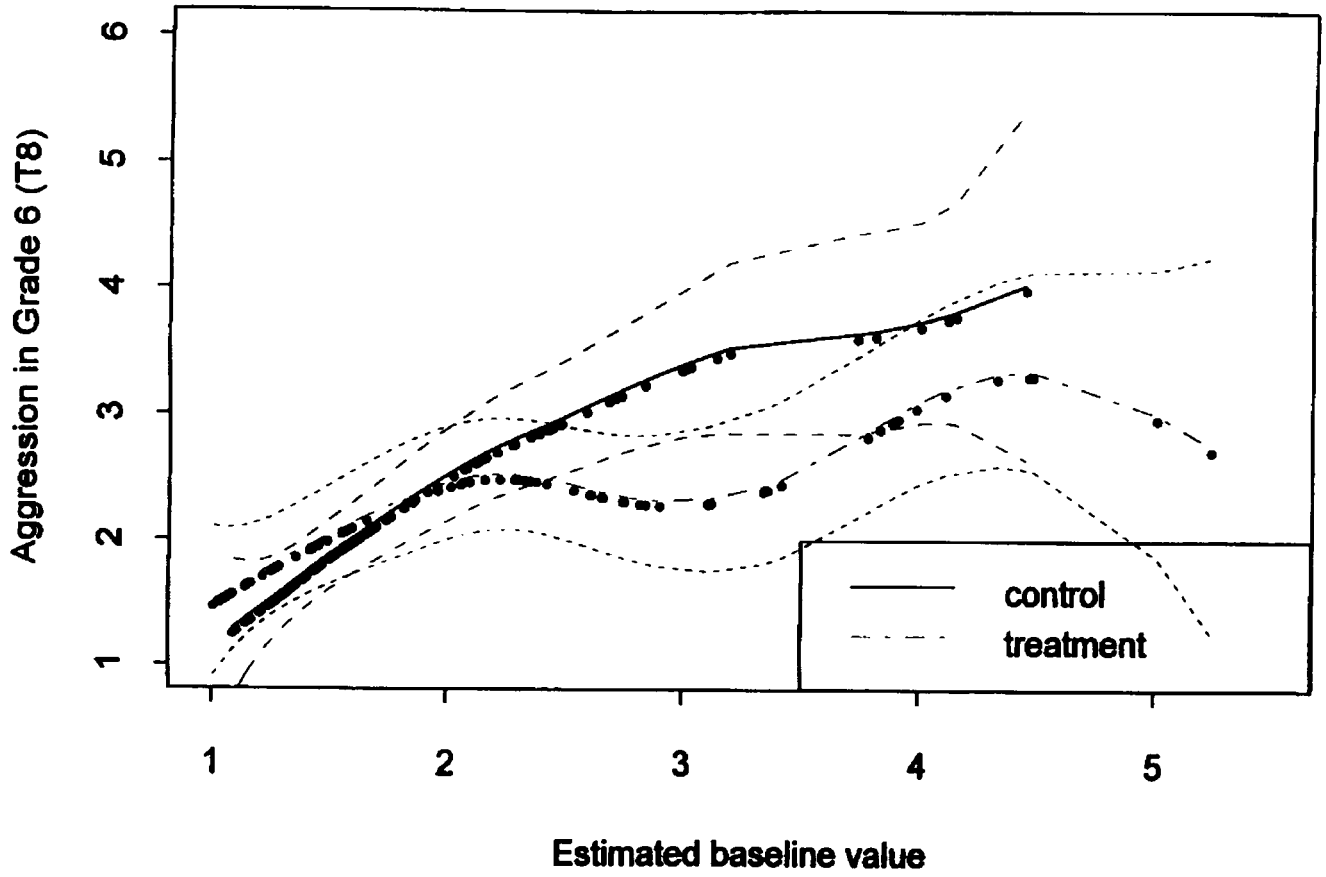

Figure 5. GAM fitted values (T8).

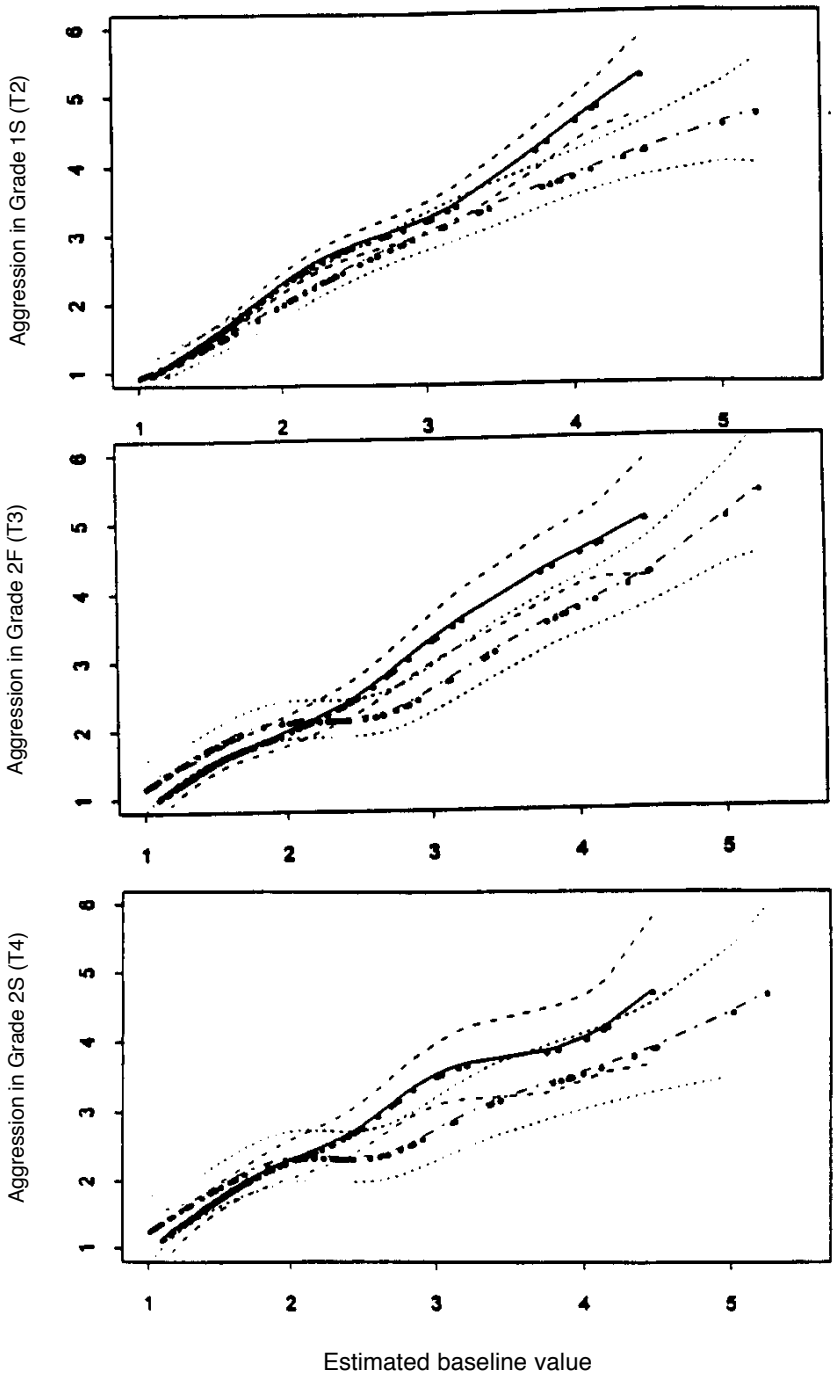

Figure 5 B. GAM fitted values (T2–T4).

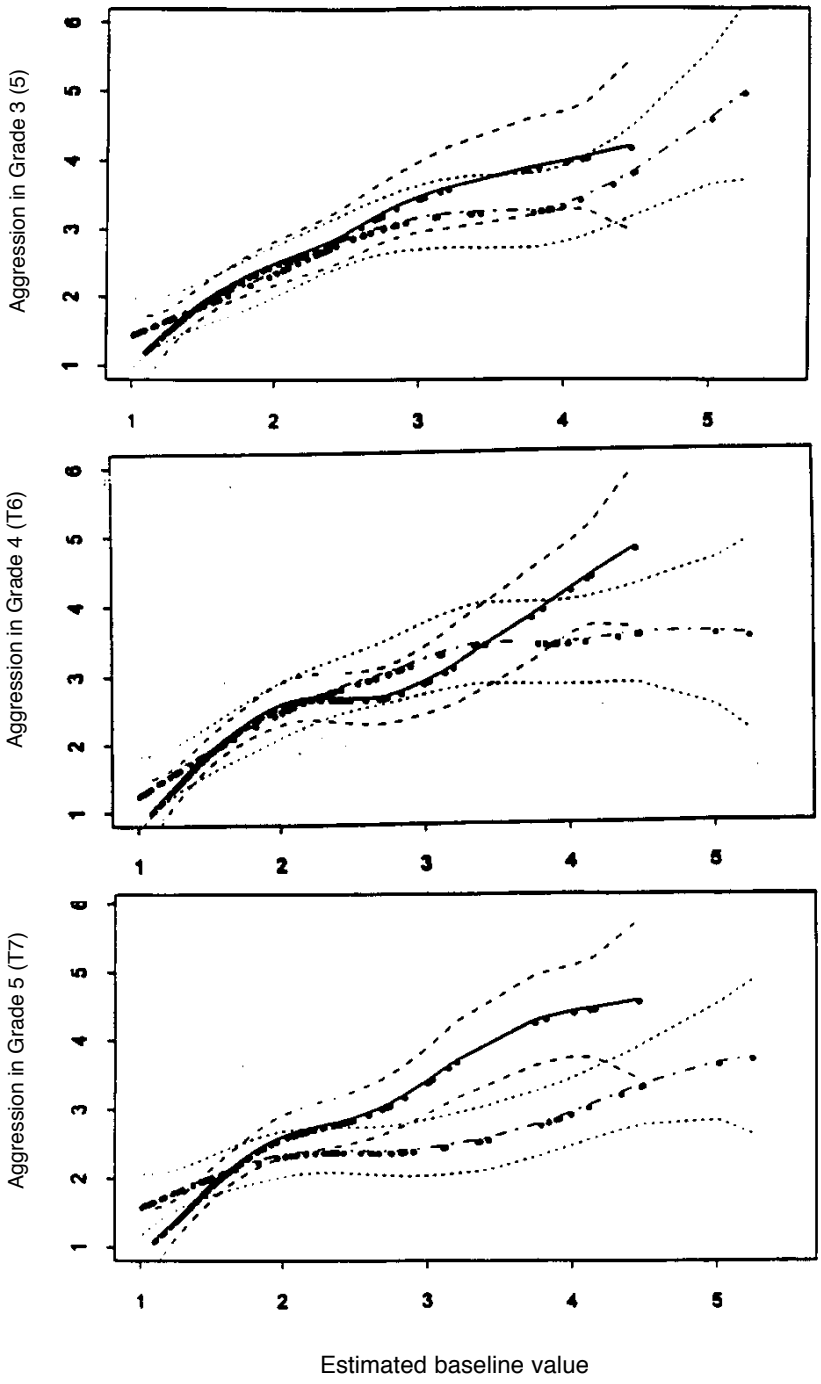

Figure 5 C. GAM fitted values (T5–T7).

group appears to drop below that of the control group at the higher end of baseline in all of the plots except for Time 5 (Grade 3) and Time 6 (Grade 4). At Time 5, the aggression level of the two outliers is high, and this brings the treatment curve up at the right hand end. At Time 6, the treatment curve is above the control curve in the middle range of baseline, though not by a significant amount. At Time 7 and Time 8, the gap between the control curve and treatment curve widens for the higher baseline range. The cutpoint decision would have been similar if using Time 7 or Time 8. Looking at the series of plots, it is noted that the estimated aggression levels of the two outliers dropped from Grade 4 and stayed low from Grade 4 through Grade 6.

4.4 Multiple-Group Growth Analyses

In order to model the nonlinear interaction within the growth model, the control and the treatment groups were both divided into the high-baseline group and the low-baseline group based on the estimated baseline and the differential response to treatment. A cutpoint was chosen based on the Figure 5 GAM plot using Grade 6 (Time 8) aggression. The point chosen was that which appeared to divide the two samples into two groups where there appeared to be treatment effects in one group and not in the other. These two groups are the high-baseline pair (42 in the control group and 41 in the treatment group) and a low-baseline pair (69 in the control group and 34 in the treatment group). The 95% confidence bands around the two GAM curves can also be taken into account in making the decision on the minimum baseline value where the treatment starts to show differential effect.

The observed means of the eight time points from the fall of Grade 1 to the spring of Grade 6 are plotted in Figure 6 for both the high- and the low-baseline groups. The data were reanalyzed based on the same two-piece linear growth model but carrying out pairs of two-group analyses where the low-baseline control group was compared with the low-baseline treatment group (Model L), and the high-baseline control group was compared with the high-baseline treatment group (Model H).

Model L started as exactly the same model as the two-group, two-stage piecewise growth model for the joint analyses of the full control and treatment groups. But this Model L has a poor fit. Each of the equality constraints was tested in turn, and it was found to be necessary to relax the equality constraints for the variances of the normative growth rate factors. The variance of the later growth

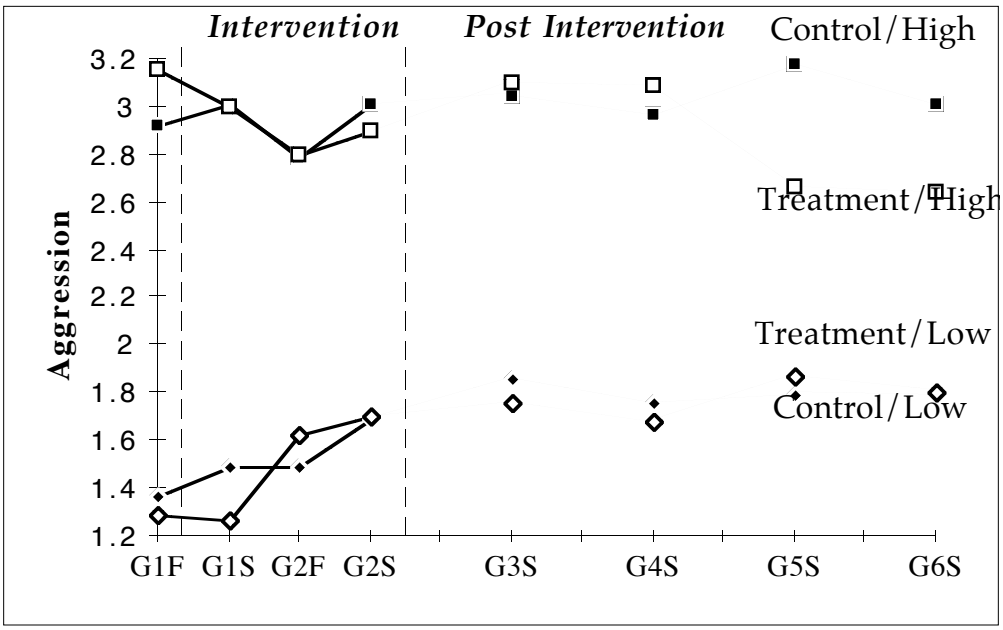

Figure 6. TOCA-R observed means over time (high- and low-baseline groups).

rate was found to be zero in the control group and was fixed at zero. Even with these equality constraints between the control group and the treatment group relaxed, Model L does not fit well, $\chi^2(49) = 74.73$, $p = 0.01$. Estimates are shown in Table 2.

Model H started as the two-group, two-stage piecewise growth model with the regression of the added growth rate factors on the initial status. The error variances for outcome variables were also equated across groups. The model has a good fit, $\chi^2(56) = 65.299$, $p = 0.19$. Estimates are shown in Table 3. The two regression slopes on the initial status are negative but not significant, showing that the treatment-baseline interaction effects are not significant anymore within the more homogenous group. The same model was analyzed without interaction effects. This model fits reasonably well also, $\chi^2(58) = 70.93$, $p = 0.12$. Estimates are shown in Table 4, and estimated mean trajectories are plotted and shown in Figure 7 together with the low-baseline groups. The mean normative growth rates are not significantly different from zero both in the early stage and in the post-intervention stage. The later normative growth rate has a greater variation than the early normative growth rate, but neither is significant in these high-baseline groups. The added early growth rate mean for the treatment group is also not significant and practically zero, but the added growth rate mean in the later piece is found to be significant ($m = 0.201$, t-value = -2.07) and negative.

Table 2

Two-Group (Control/Treatment) Piecewise Analysis: Low-Baseline Groups

		Control group ($n = 69$)			Treatment group ($n = 34$)	
		$\chi^2(49) = 74.73,\ p = 0.01$				
Outcome variables						
Intercept		1.301	(0.029)	=	1.301	(0.029)†
Growth factors						
Initial status						
	Mean	0.0	(*)		0.0	(*)
	Variance	0.0	(*)		0.033	(0.019)
Early growth rate						
	Mean	0.212	(0.040)	=	0.212	(0.040)
	Variance	0.064	(0.014)		0.027	(0.011)
Later growth rate						
	Mean	-0.018	(0.032)	=	-0.018	(0.032)
	Variance	0.0	(*)		0.008	(0.027)
Added early growth rate						
	Mean				-0.019	(0.056)
	Variance				0.0	(*)
Added later growth rate						
	Mean				0.042	(0.054)
	Variance				0.0	(*)
Residual variances for outcome variables						
Time 1		0.141	(0.024)		0.031	(0.020)
Time 2		0.226	(0.039)		0.095	(0.031)
Time 3		0.291	(0.051)		0.240	(0.063)
Time 4		0.459	(0.082)		0.233	(0.063)
Time 5		0.421	(0.096)		0.254	(0.084)
Time 6		0.329	(0.087)		0.296	(0.098)
Time 7		0.335	(0.086)		0.207	(0.124)
Time 8		0.405	(0.089)		0.359	(0.236)
Residual covariances for outcome variables						
Time 1-Time 2		0.100	(0.025)		-0.008	(0.019)
Time 2-Time 3		0.033	(0.017)		-0.003	(0.022)
Time 3-Time 4		0.281	(0.058)		0.158	(0.053)
Time 4-Time 5		0.045	(0.040)		0.033	(0.037)
Time 5-Time 6		-0.084	(0.060)		0.054	(0.068)
Time 6-Time 7		0.068	(0.066)		-0.004	(0.060)
Time 7-Time 8		0.049	(0.061)		0.032	(0.155)

* Parameter is fixed in this model.
† Standard errors are given in parentheses.
= Parameter set equated across group.

Table 3

Two-Group (Control/Treatment) Piecewise Analysis With Interactions: High-Baseline Groups

		Control group ($n = 42$)			Treatment group ($n = 41$)	
		$\chi^2(56) = 65.30$, $p = 0.19$				
Outcome variables						
Intercept		2.980	(0.105)	=	2.980	(0.105)†
Growth factors						
Initial status						
	Mean	0.0	(*)		0.0	(*)
	Variance	0.501	(0.133)	=	0.501	(0.133)
Early growth rate						
	Mean	0.027	(0.070)	=	0.027	(0.070)
	Variance	0.040	(0.019)	=	0.040	(0.019)
Later growth rate						
	Mean	0.049	(0.063)	=	0.049	(0.063)
	Variance	0.074	(0.048)	=	0.074	(0.048)
Growth factor covariances						
Initial status - Early growth rate		0.0	(*)		0.0	(*)
Initial status - Later growth rate		0.0	(*)		0.0	(*)
Regression on initial status						
Added early growth rate						
	Intercept				-0.016	(0.090)
	Slope				-0.108	(0.118)
Added later growth rate						
	Intercept				-0.194	(0.093)
	Slope				-0.171	(0.121)
Residual variances for outcome variables						
Time 1		0.659	(0.154)	=	0.659	(0.154)
Time 2		0.548	(0.115)	=	0.548	(0.115)
Time 3		0.817	(0.150)	=	0.817	(0.150)
Time 4		0.904	(0.155)	=	0.904	(0.155)
Time 5		0.718	(0.147)	=	0.718	(0.147)
Time 6		0.825	(0.163)	=	0.825	(0.163)
Time 7		0.255	(0.198)	=	0.255	(0.198)
Time 8		0.486	(0.384)	=	0.486	(0.384)

(Estimates for residual covariances not shown)

* Parameter is fixed in this model.
† Standard errors are given in parentheses.
= Parameter set equated across group.

Table 4
Two-Group (Control/Treatment) Piecewise Analysis: High-Baseline Groups

		Control group ($n = 42$)			Treatment group ($n = 41$)	
		$\chi^2(58) = 70.93$, $p = 0.12$				
Outcome variables						
Intercept		2.969	(0.104)	=	2.969	(0.104)†
Growth factors						
Initial status						
	Mean	0.0	(*)		0.0	(*)
	Variance	0.389	(0.099)	=	0.389	(0.099)
Early growth rate						
	Mean	0.030	(0.070)	=	0.030	(0.070)
	Variance	0.025	(0.020)	=	0.025	(0.020)
Later growth rate						
	Mean	0.047	(0.065)	=	0.047	(0.065)
	Variance	0.089	(0.049)	=	0.089	(0.049)
Added early growth rate						
	Mean				-0.014	(0.089)
	Variance				0.0	(*)
Added later growth rate						
	Mean				-0.201	(0.094)
	Variance				0.0	(*)
Residual variances for outcome variables						
Time 1		0.805	(0.154)	=	0.805	(0.154)
Time 2		0.597	(0.115)	=	0.597	(0.115)
Time 3		0.853	(0.150)	=	0.853	(0.150)
Time 4		0.886	(0.155)	=	0.886	(0.155)
Time 5		0.744	(0.147)	=	0.744	(0.147)
Time 6		0.846	(0.163)	=	0.846	(0.163)
Time 7		0.224	(0.198)	=	0.224	(0.198)
Time 8		0.420	(0.384)	=	0.420	(0.384)
Residual covariances for outcome variables						
Time 1-Time 2		0.380	(0.112)		-0.003	(0.128)
Time 2-Time 3		0.177	(0.091)		0.161	(0.106)
Time 3-Time 4		0.336	(0.124)		0.459	(0.131)
Time 4-Time 5		0.328	(0.122)		0.074	(0.115)
Time 5-Time 6		0.140	(0.112)		0.233	(0.140)
Time 6-Time 7		0.177	(0.121)		-0.042	(0.123)
Time 7-Time 8		-0.311	(0.247)		-0.245	(0.270)

* Parameter is fixed in this model.
† Standard errors are given in parentheses.
= Parameter set equated across group.

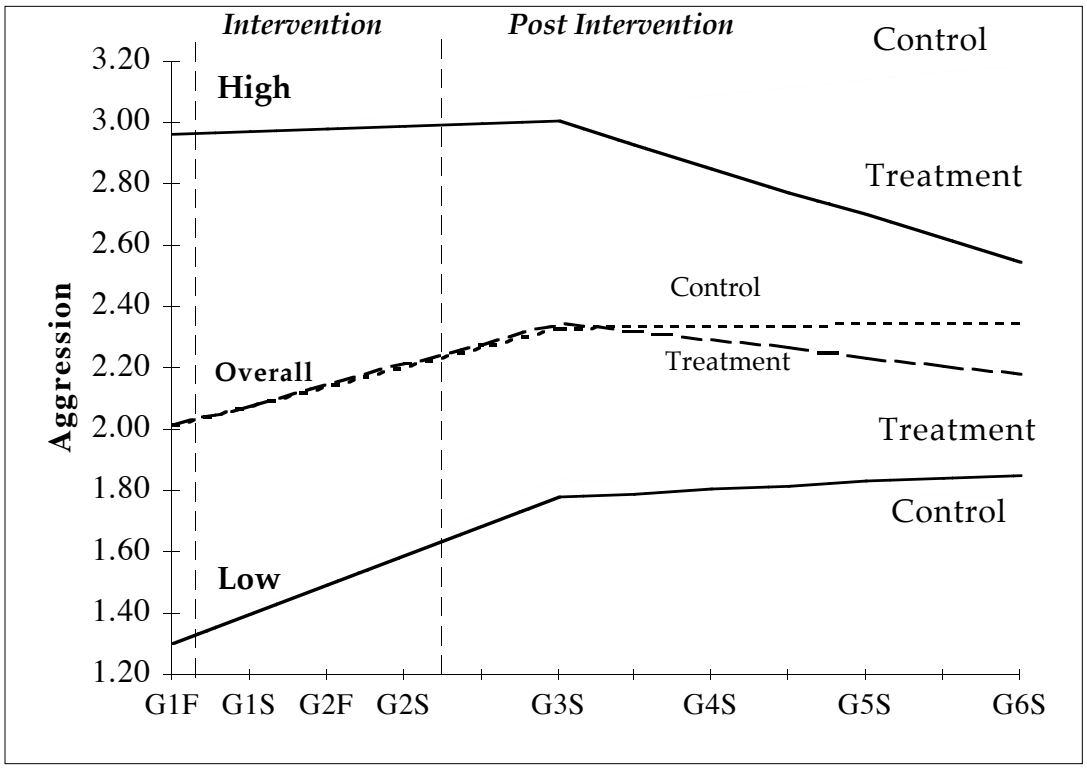

Figure 7. Growth trajectories of high- and low-baseline groups showing treatment-baseline interaction.

This final model was run with the two outliers in the treatment group deleted. The model fit did not change much, $\chi^2(58) = 71.2$, $p = 0.11$, but the added growth rate mean in the later piece that was found to be significant before is now smaller in magnitude and no longer significantly different from zero at the 5% level ($m = 0.169$, t-value = -1.92). This calls for caution in the inference.

The above analyses involve multiple steps including subjective decision on choice of cutpoints. The tests of significance, however, were based on only the last step of the analyses. It is very likely that the standard errors are underestimated and the inference too optimistic. Furthermore, the growth models were fitted using the maximum likelihood method assuming multivariate normality, while the outcome variables at the early time points are rather skewed and the multivariate normality assumptions may not be met. The maximum likelihood estimator is quite robust to non-normality, but the standard errors would be underestimated. Bootstrapping procedure was used to obtain more realistic standard errors for the estimates.

4.5 Estimating Standard Errors by Bootstrap

Two hundred bootstrap samples (B = 200) were generated by randomly sampling with replacement 111 times from the control group (n = 111) and 75 times from the treatment group (n = 75) so that sample sizes were equal to the original sample. For each of these bootstrap samples, factor score of the initial status was estimated for each individual in both the control and treatment groups. The estimation was calculated based on estimates obtained by applying the same linear growth model based on the first four time points to the bootstrap sample. The individuals in both the control and treatment groups were then ordered from low to high on their initial status factor scores. The treatment group was subsetted into two groups, the low-baseline group and the high-baseline group based on the same group ratio as the original data. The cutpoint was noted and applied to the control group to divide it into the two low- and high-baseline groups as in the treatment group.

The same two-group analysis was then carried out to compare the low-control group to the low-treatment group and the high-control group to the high-treatment group using the same model. Parameter estimates were recorded for the 200 samples, and bootstrap samples that produced inadmissible solutions were discarded and redrawn. The inadmissible solutions discarded were those that resulted in negative variances, which probably occurred due to some odd samples. There were about 20% of these out of the total number of samples drawn. Bootstrap estimates of standard errors were calculated for each parameter. Because the main interest is in the inferences to the results of the analysis of the high-baseline control/treatment pair where there may be significant findings, only the bootstraps results for the high-baseline pair are reported.

Table 5 shows the same parameter estimates as those of Table 4, but with the original standard errors replaced by the bootstrap estimates of standard errors. Comparing the two sets of standard errors, the largest differences are in the standard errors of the intercept, which captures the mean initial status for the two groups and the variance of the initial status factor. These bootstrap estimates are 0.238 and 0.133 compared to the original values of 0.104 and 0.099. These differences are to be expected because the initial status mean and variance for the high-baseline group are the quantities that will vary a lot with the decision of cutpoint. It seems that this is where the standard errors were very much underestimated if the variation due to cutpoint decision was not taken into

Table 5

Two-group (Control/Treatment) Piecewise Analysis: High-Baseline Groups (With Bootstrap Estimates of Standard Errors)

		Control group ($n = 42$)			Treatment group ($n = 41$)	
		$\chi^2(58) = 70.93$, $p = 0.12$				
Outcome variables						
Intercept		2.969	(0.238)	=	2.969	(0.238)†
Growth factors						
Initial status						
	Mean	0.0	(*)		0.0	(*)
	Variance	0.389	(0.133)	=	0.389	(0.133)
Early growth rate						
	Mean	0.030	(0.109)	=	0.030	(0.109)
	Variance	0.025	(0.022)	=	0.025	(0.022)
Later growth rate						
	Mean	0.047	(0.062)	=	0.047	(0.062)
	Variance	0.089	(0.048)	=	0.089	(0.048)
Added early growth rate						
	Mean				-0.014	(0.115)
	Variance				0.0	(*)
Added later growth rate						
	Mean				-0.201	(0.097)
	Variance				0.0	(*)
Residual variances for outcome variables						
Time 1		0.805	(0.193)	=	0.805	(0.193)
Time 2		0.597	(0.170)	=	0.597	(0.170)
Time 3		0.853	(0.186)	=	0.853	(0.186)
Time 4		0.886	(0.149)	=	0.886	(0.149)
Time 5		0.744	(0.182)	=	0.744	(0.182)
Time 6		0.846	(0.206)	=	0.846	(0.206)
Time 7		0.224	(0.211)	=	0.224	(0.211)
Time 8		0.420	(0.383)	=	0.420	(0.383)
Residual covariances for outcome variables						
Time 1-Time 2		0.380	(0.181)		-0.003	(0.202)
Time 2-Time 3		0.177	(0.131)		0.161	(0.126)
Time 3-Time 4		0.336	(0.149)		0.459	(0.140)
Time 4-Time 5		0.328	(0.197)		0.074	(0.119)
Time 5-Time 6		0.140	(0.227)		0.233	(0.145)
Time 6-Time 7		0.177	(0.164)		-0.042	(0.131)
Time 7-Time 8		-0.311	(0.327)		-0.245	(0.311)

* Parameter is fixed in this model.
† Bootstrap estimates of standard errors are given in parentheses.
= Parameter set equated across group.

account. The bootstrap standard errors are also slightly larger for some of the growth rates means, but the differences are not large enough to make a difference in the inferences made. It is noted that the standard errors for the treatment effects in this case are not affected so much by the decision of cutpoint and the bootstrap resampling. The treatment effect found significant for the second growth stage for the more aggressive boys is still significant taking the uncertainty of the multistage procedure into account.

With these results, it appears that the different baseline groups responded quite differently to treatment. The piecewise approach has the advantage of allowing the evaluation of treatment effects in the different stages. Subsetting the sample based on estimated baseline and analyzing the subgroups separately can tease out the differential treatment effects in the two groups. The fact that in this analysis, the treatment was found to be "effective" only after the third grade ought to generate questions for substantive researchers. Whether the difference between the control and the treatment group is attributable to the intervention after such a long period ought to be asked. Plausible alternative explanations ought to be explored. If the difference was really due to treatment, then there ought to be explanations for the delayed effect. A replication of the experiment may be needed in order to establish the validity of this finding.

5 Discussion

The advantage of conducting program evaluation using longitudinal data is obvious if the intervention is designed to change a developmental process. The Muthén-Curran two-group formulation allows the modeling of treatment effects as random effects that vary across the individuals, while considering normative growth established using the control group. This method of assessing treatment effects in the latent growth framework has the full advantage of structural equation modeling flexibility. Individual differences in growth and factors influencing growth due to treatment can be modeled directly. The interaction effects of treatment with background variables including baseline can be assessed easily.

The multistage procedure for modeling and assessing treatment effects was aimed at evaluating treatment effects for those most at risk in the presence of nonlinear treatment-baseline interactions. The multistage procedure also calls for better methods of obtaining realistic standard errors of estimates.

Bootstrapping the whole procedure to estimate the standard errors to some extent takes the uncertainty involved into account. Further investigations should be carried out regarding obtaining better estimates of standard errors and to evaluate the adequacy of the bootstrapping procedure in this context.

In terms of modeling longitudinal growth, treatment effects of short duration are usually lost in the modeling and treated as measurement errors. This kind of temporary effect would likely occur right after an intervention starts but disappear after a while. This was observed in the GBG data where there was a sharp drop from Time 1 to Time 2 with an upward trend after Time 2. It is very difficult to decide whether this effect is of a temporary nature and is of no consequence in the big picture, or whether it alters the trajectory of growth permanently by shifting it down. This argues for multiple measurements over at least a short period before intervention to set a preintervention trend, so that any effects of consequence can be correctly interpreted. This will also contribute to the estimation of baseline.

Usually when an intervention program has an immediate effect, we would ask whether the effect lasts over time. When there is no short-term effect and a difference appears between the control group and the treatment group after a long period, as was found in the GBG analysis, questions regarding whether the difference is attributable to the intervention ought to be asked. There may be other plausible alternative explanations for the difference. Unless there is a sound hypothesis and the mechanism for explaining why and how the effect would be delayed is in place, it is difficult to justify attributing the delayed difference in outcome to the intervention especially when the sample is small.

The implications of finding that the Good Behavior Game intervention was only effective for the more aggressive males would need careful consideration. Targeting the intervention at the more aggressive children only may not produce the desired effects. The Good Behavior Game is a team-based behavior management game that promotes good behavior by rewarding teams that do not exceed maladaptive behavior standards (Kellam et al., 1994). The classroom teacher made sure that the teams were heterogeneous when assigning each child in the class to one of three teams. This heterogeneous group composition may be necessary to bring about change in the more aggressive boys. Targeting the intervention only at the more aggressive males would result in a different design and a different intervention that ought to be tested anew. Even though

the analyses cannot be used for targeting the treatment to more aggressive individuals, they are useful in two regards. First, they make it possible to find a treatment effect that may be hidden due to the interactions of treatment with baseline. Second, they indicated for whom the treatment is effective and the sizes of the treatment effects during different stages.

When program effectiveness is observed only at the high end of baseline and the distribution of outcome variable is skewed towards the low end, there is the concern of distributional assumptions in the analysis and also the issue of outliers. Down-weighting the outliers may mean down-weighting an important case in the range of baseline where the intervention is most relevant. This may imply that an intervention that is targeted for problem behavior should probably oversample those with problems in the intervention trial.

References

Bock, R. D. (1989). *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

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

Brown, C. H., & Liao, J. (1996). *Principles for designing randomized preventive trials in mental health* (Tech. Rep.; manuscript submitted for publication). Tampa: University of South Florida, Department of Epidemiology and Biostatistics.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*, 147-158.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Bryk, A. S., & Weisberg, H. I. (1976). Value-added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics, 1*, 127-155.

Faraway, J. J. (1992). On the cost of data analysis. *Journal of Computational and Graphical Statistics, 1*, 213-229.

Goldstein, H. I. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrics, 73*, 43-56.

Goldstein, H. I. (1995). *Multilevel statistical models*. London: Edward Arnold.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman and Hall.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Kellam, S. G., Rebok, G. W., Ialongo, N., & Mayer, L. S. (1994). The course and malleability of aggressive behavior from early first grade into middle school: Results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry, 35*, 259-281.

Khoo, S. T. (1997). *Assessing interactions between program effects and baseline: A latent curve approach*. Unpublished doctoral dissertation, University of California, Los Angeles.

Laird, N. M., & Ware, H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*, 963-974.

Lawley, D. N., & Maxwell, A. E. (1971). Factor analysis as a statistical method (2nd ed.). London: Butterworths.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58*, 110-133.

Meredith, W., & Tisak, J. (1984, June). *"Tuckerizing" curves.* Paper presented at the annual meeting of the Psychometric Society, Santa Barbara, CA.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrica, 55*, 107-122.

Muthén, B. O. (1987). *LISCOMP. Analysis of linear structural equations with a comprehensive measurement model. Theoretical integration and user's guide.* Mooresville, IN: Scientific Software.

Muthén, B. O. (1991). Analysis of longitudinal data using latent variable models with varying parameters. In L. Collins, & J. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 1-17). Washington DC: American Psychological Association.

Muthén, B. O. (1993). Latent variable modeling of growth with missing data and multilevel data. In C. R. Rao & C. M. Cuadras (Eds.), *Multivariate analysis: Future directions 2* (pp. 199-210). Amsterdam: North-Holland.

Muthén, B. O. (1997). *Longitudinal studies of achievement growth using latent variable modeling* (CSE Tech. Rep. No. 412). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Muthén, B. O., & Curran, P. J. (in press). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods.*

Muthén, B. O., & Hsu, J. W. (1993). Selection and predictive validity with latent variable structures. *British Journal of Mathematical and Statistical Psychology, 46*, 255-271.

Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. D. Cook, & D. T. Campbell (Eds.), *Quasi-experimentation: Design and analysis issues for field settings* (pp. 147-205). Boston: Houghton Mifflin.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 90*, 726-748.

Seltzer, M. H., Frank K. A., & Bryk A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis*, *16*, 41-49.

Sörbom, D. (1982). Structural equations models with structured means. In K. G. Jöreskog, & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction.* Amsterdam: North Holland.

Stine, R. A. (1989). An introduction to bootstrap methods: Examples and ideas. *Sociological Methods and Research*, *8*, 243-291.

Strenio, J. L. F., Weisberg, H. I., & Bryk, A. S. (1983). Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates. *Biometrics, 39*, 71-86.

Tucker, L. R. (1971). Relation of factor score estimates to their use. *Psychometrika, 36*, 427-436.

Weisberg, A. (1985). *Applied linear regression* (2nd ed.). New York: Wiley.

Wheaton, B., Muthén, B., Alwin, D., & Summers, G. (1977). Assessing reliability and stability in panel models with multiple indicators. In D. R. Heise (Ed.), *Sociological methodology 1977* (pp. 84-136). San Francisco: Jossey-Bass.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116,* 363-381.