Three Essays on the Modeling of Development

Dissertation Proposal

by

Tom Loughran
The H. John Heinz III School of Public Policy
Carnegie Mellon University

16 May 07

Thesis Committee

Daniel Nagin, *chair*

Mel Stephens

Amelia Haviland

Table of Contents

Introduction

This dissertation is intended for the advancement of methodology and techniques used in the modeling of development. There are three loosely tied essays of which this contribution is comprised.

The first chapter is entitled **'Finite Sample Effects in Group-Based Trajectory Models."** It analyses a very intricate and specific aspect of the broader body of work credited to Daniel Nagin regarding group-based trajectory models. These models, which are an application of finite mixture modeling, are used to model population heterogeneity in the development of various types of behavior such as physical aggression or anxiety over age or time. Estimation of the trajectory model is done via the method of maximum likelihood, which is beneficial in that it assures estimates which are asymptotically unbiased and approximately normally distributed. However, given that the samples of longitudinal data necessary for these models typically range in from n=500 to n=1,500, there may be certain finite sample complications which do not allow the asymptotic properties of these estimates to fully engage. To further complicate things, the quantities with which normal users of these models are generally interested represent non-linear transformations of the parameters estimated by maximum likelihood, meaning their behavior in finite samples is even more difficult to understand. In addressing these finite sample effects, we employ a quasi-bootstrap procedure to compare to a set of estimates from a large 'population', and we find that, even in samples as small as n=500, the model estimates of the parameters and their respective non-linear transformations are able to closely replicate these 'population' benchmarks. At the time of proposal, this chapter is

fully completed and was published in *Sociological Methods and Research*, November 2006, Volume 35.

The second chapter is entitled **"Consequences of a Violation of the Conditional Independence Assumption in Group-Based Trajectory Models."** This chapter again addresses another nuance of the group-based trajectory model, which, in particular, is often a criticism of the methodology. Specifically, in the specification of these group-based models trajectory models, the following assumption is made. Conditional upon an individual following a given trajectory, the individual's outcomes over successive periods are assumed to be independent. This assumption is a strong one, and it is made specifically in order to make the models more tractable. However, there has been no work done to investigate the behavior of these models in the presence of a failure of this conditional independence assumption. The purpose of this chapter is to explicitly investigate this issue and determine the consequences to users of these models. We again propose a Monte Carlo simulation in order to observe how the model behaves when using data simulated specifically to violate this conditional independence assumption, or in more exact terminology, we say that our simulated data has serial correlation. We then examine the results yielded by traditional model estimation in the face of this blatant assumption violation in order to quantify and assess the impact of such results on the average user.

The third chapter is entitled **"Accounting for Selection to Understand the Effects of Group Daycare on the Development of Physical Aggression."** This chapter considers the problem of estimating casual effects of a treatment on an outcome of interest, which is a problem often addressed by researchers. Rather than use more traditional, model-

based approached, the intent here is to introduce a different type of statistical methodology, propensity score matching, into the developmental literature. The application is to group daycare, and how a child entering into to daycare at various points during the first five years of life affects the development of physical aggression in the child. The literature on both group daycare and physical aggression is quite extensive. The daycare literature concedes that the problem of selection effects is a daunting one, and it must be properly addressed before any type of casual inference is possible regarding its treatment effects. It is this problem of selection – namely, children with different backgrounds and risk factors have different likelihood of entering into daycare, and thus, the effects on associated outcomes such as physical aggression are likely not transparent – which this chapter aims to address. However, there is a greater purpose to this paper rather than just the application of a useful methodology. The policy implications of this analysis are substantial. Group daycare is becoming more popular as more women are entering the workforce, and there is an unresolved debate as to the effects on children who this non-maternal care. By analyzing this application, we hope to add clarity to a blurry picture of how, exactly children's outcome are possibly affected by group daycare.

While each of these three chapter have policy implications, the first two are more technically driven, with the goal being to have the final results be taken under consideration by methodologists, practitioners and researchers when applying these methods to more policy specific problems. The third chapter, while still of technical importance, is offered as more of a stand alone piece of policy research which addresses

an important substantive topic, on top of simply trying to introduce a potentially useful method into the developmental literature.

Sociological Methods & Research
Volume 35 Number 2
November 2006 250-278

10.1177/0049124106292292
http://smr.sagepub.com
hosted at
http://online.sagepub.com

Finite Sample Effects in Group-Based Trajectory Models

Tom Loughran
Daniel S. Nagin
Carnegie Mellon University

Two desirable properties of maximum likelihood–based parameter estimates are that the estimates are asymptotically unbiased and asymptotically normally distributed. In this article, the authors test whether the asymptotic properties of maximum likelihood estimation are achieved in sample sizes typically used in applications of group-based trajectory modeling. Through empirical results generated by resampling of population data, they find that the maximum likelihood estimates obtained in group-based trajectory models still provide reasonably close estimates of their true population values and have approximately normal distributions, even when estimated with a sample size as small as $n = 500$. Furthermore, and more important for the users of these types of models, the authors find similarly good performance in the model's ability to estimate the transformed quantities of main interest: the group trajectories and mixing probabilities.

Keywords: *maximum likelihood estimation; group-based trajectory models; mixing probabilities; group trajectories*

Psychologists use the term *developmental trajectory* to describe the course of a behavior or outcome over age or time. Until about a decade ago, the two main branches of methodology for analyzing developmental trajectories were hierarchical modeling (Bryk and Raudenbush 1987, 1992; Goldstein 1995) and latent curve analysis (McArdle and Epstein 1987; Meredith and Tisak 1990; Muthén 1989; Willett and Sayer 1994). A 1993 article by Nagin and Land laid out a third alternative—group-based trajectory modeling. The group-based trajectory model is a

Authors' Note: The research was supported by generous financial support from the National Science Foundation (SES-9911370) and the National Institute of Mental Health (RO1 MH65611-01A2).

specialized application of finite mixture modeling. Using mixtures of suitably defined probability distributions, the method is designed to identify distinctive clusters of developmental trajectories within the population. Whereas the hierarchical and latent curve methodologies model population variability in growth with multivariate continuous distribution functions, the group-based approach uses a multinomial modeling strategy (Raudenbush 2001) and is designed to identify relatively homogeneous clusters of developmental trajectories.

The introduction of "canned" software for estimating group-based models has resulted in a growing body of research based on this method. At this time, there are two excellent software alternatives for estimating group-based trajectories models. One is an SAS-based procedure called Proc Traj. It is described in Jones, Nagin, and Roeder (2001) and in documentation available at www.ncovr.org. Proc Traj is designed to be inserted into the SAS software package. Once inserted, SAS treats it like any other standard SAS procedure. The other alternative is a widely used structural equation modeling software package called M-Plus, developed by Bengt Muthén, Linda Muthén, and colleagues (Muthén and Muthén 2004). Piquero (2005) reports that more than 60 published articles use group-based trajectory modeling.

The purpose of this article is to examine several technical issues that are important to make valid statistical inferences with these models. Group-based trajectory models are estimated by the method of maximum likelihood. Two desirable properties of maximum likelihood–based parameter estimates are that the estimates are asymptotically unbiased and asymptotically normally distributed (Kiefer and Wolfowitz 1956). The former property is fundamental because it implies that, but for sampling error, maximum likelihood–based parameter estimates correctly measure the population parameters they are intended to estimate. The latter property has great practical importance. It forms the theoretical basis for the use of widely employed normal-based statistical tests of parameter estimates. Here we report the results of analyses that are designed to address whether these two important asymptotic statistical properties seem to be achieved in finite samples of the sizes typically used in actual applications.

To address these issues, we first estimate a group-based trajectory model with a very large sample composed of more than 13,000 individuals. We show that the parameter estimates for this model have exceedingly small standard errors and therefore can plausibly be treated as if they equal the population values. We then repeatedly draw random samples of

three fixed sizes—500, 1,000, and 1,500—from the 13,000 individuals and estimate a group-based trajectory model for each sample. We chose these three sample sizes because they span the range of sample sizes used in most applications.

The multiple estimates of the model's parameters for each sample size provide the basis for generating an empirical approximation of each parameter's sampling distribution for that sample size. These sampling distributions allow us to assess whether the distribution is centered on our surrogate for the population value of the parameter. This assessment provides an empirical basis for judging the magnitude of the finite sample bias, if any. It also allows us to assess whether the sampling distribution is approximately normal.

We find that even for the smallest sample size of $n = 500$ observations, (1) the average of the parameter estimates composing the sampling distribution is close in magnitude to the population value based on the sample of more than 13,000, (2) the sampling distributions are indeed normal-like, and (3) the maximum likelihood estimate of each parameter's standard error for each sample condition is unbiased, in the sense that it closely approximates the standard deviation of the parameter's sampling distribution. We also find that these same properties hold for various nonlinear transformations of the parameter estimates that measure quantities of direct interest to users, such as the probability of trajectory group membership.

The Likelihood Function

As described earlier, the purpose of group-based trajectory modeling is to identify clusters of similar individual-level trajectories. These clusters form the trajectory groups. A trajectory group is described by the path of the group's expected behavior over age or time and a probability measuring the proportion of the population following this path. In this section, we briefly lay out the form of the likelihood function used to estimate these two defining features of a group-based trajectory model. For a more detailed discussion, see Nagin (1999, 2005).

Let the vector $\mathbf{Y}_i$ denote the longitudinal sequence of individual i's behavior over $t = 1, \ldots, T$ periods, and let $j = 1, \ldots, J$, denote the group. Conditional upon i being a member of group j, outcomes over successive periods are assumed to be independent. Thus, the likelihood of observing $\mathbf{Y}_i$ is $P^j(Y_i) = \prod_{t=1}^{T} p^j(y_{it})$, where $p^j(y_{it})$ is a probability density function.

Because the response variable in this analysis is a count of the number of times an individual has been arrested, the Poisson-based model was used, where the expected rate of offending for group j at time t equals λ_{jt}, and $p^j(y_{it}) = \frac{\lambda_{jt}^{y_{it}} e^{-\lambda_{jt}}}{y_{it}!}$.

Each group's trajectory is defined by the time path of the Poisson rate parameter λ_{jt}. This time path is assumed to follow a polynomial function of age or time. As developed below, our analysis is based on a four-group model. The trajectory for one group is specified to be constant over age, whereas the other three groups are specified to follow a quadratic path over age:

$$ln(\lambda_i) = \beta_0^1, \; j = 1, \tag{1}$$

$$ln(\lambda_{jt}) = \beta_0^j + \beta_1^j \text{ age}_t + \beta_2^j \text{ age}_t^2, \; j = 2, 3, 4. \tag{2}$$

We estimate $\ln(\lambda_{jt})$ rather than λ_{jt} to ensure that the predicted value of the Poisson parameter can never be negative. Observe that the parameters defining each group's trajectory are superscripted by j. This allows the time path of λ_{jt} to vary freely across groups.

For each individual, the unconditional likelihood of observing $\mathbf{Y}_i$ is $P(Y_i) = \sum_j \pi_j P^j(Y_i)$, where π_j is the unconditional probability of membership in group j. The likelihood for the entire sample of N individuals is thus the product of the individual likelihoods: $L = \prod^N P(Y_i)$.

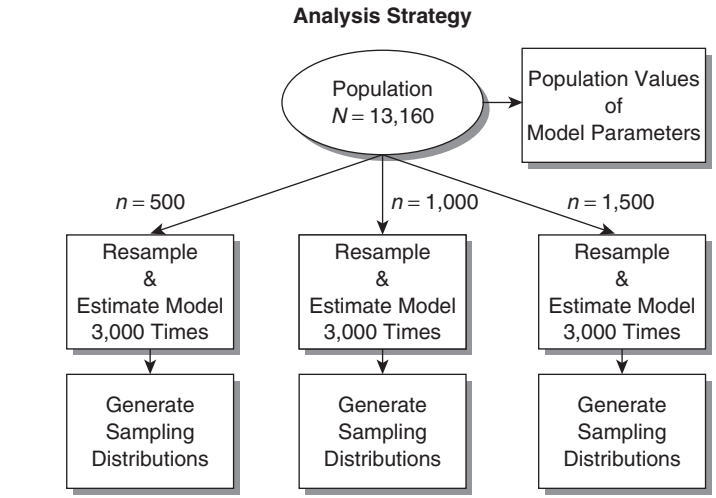
The group membership probabilities, π_j, are not estimated directly but instead are estimated by a generalized logit function:

$$\pi_j = \frac{e^{\theta_j}}{\sum_1^J e^{\theta_j}}, \tag{3}$$

where θ_1 is normalized to zero. Estimation of π_j in this fashion ensures that each such probability properly falls between zero and one.

The vectors $\beta^j = (\beta_0^j \beta_1^j \beta_2^j)$ and $\theta = (\theta_2 \theta_3 \theta_4)$ are the model parameters that are estimated directly by maximum likelihood. However, it is the trajectories, $\lambda_j = (\lambda_{j1}, \lambda_{j2}, \ldots, \lambda_{jT})$ for $j = 1, \ldots, J$, and mixing probabilities, $\pi = (\pi_1 \pi_2 \pi_3 \pi_4)$, which are nonlinear transformations of these estimated parameters, that are the quantities of actual interest to users. Thus, when considering the unbiasedness and normality of the model's parameter estimates in finite samples, we also examine these same properties for $\lambda_j = (\lambda_{j1}, \lambda_{j2}, \ldots, \lambda_{jT})$ for $j = 1, \ldots, J$ and $\pi = (\pi_1 \pi_2 \pi_3 \pi_4)$.

Figure 1
Analysis Strategy



Method

Overview

Figure 1 provides an overview of our analytical strategy. The data set that provides the empirical foundation for this analysis is composed of the arrest histories of 13,160 individuals. Because of its unusually large size, we treat this data set as if it were a population, not a sample. We first estimate a four-group trajectory model based on the data for all 13,160 individuals. The parameter estimates and their companion nonlinear transformations that calculate $\lambda_j = (\lambda_{j1}, \lambda_{j2}, \ldots, \lambda_{jT})$ for $j = 1, \ldots, J$ and $\pi = (\pi_1 \pi_2 \pi_3 \pi_4)$ are thereafter treated as the population quantities themselves. As described in the "Population Model" section, this assumption has a solid empirical justification because of the exceedingly small standard errors of the estimated parameters.

In considering sample size impacts in longitudinal modeling, two dimensions of sample size are relevant—the number of individuals who make up the sample and the number of periods for which each individual is observed. The latter dimension of sample size raises complex conceptual

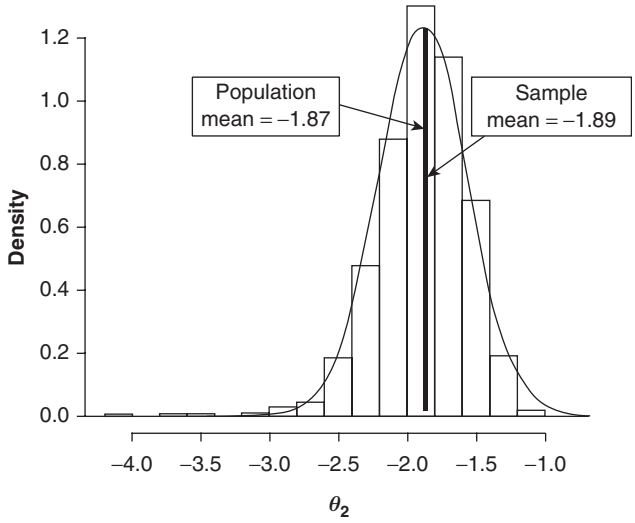
issues concerning the length of time individuals must be tracked before one can draw valid conclusions about their developmental trajectories (see, e.g., the exchange between Eggleston, Laub, and Sampson [2004] and Nagin [2004]). Here we focus on the former dimension of sample size.

To our knowledge, published applications of group-based trajectory modeling are usually based on sample sizes that range from as small as 284 to 400 individuals (Shaw et al. 2003; Nagin and Land 1993) to more than 1,000 individuals (Fergusson, Horwood, and Nagin 2000). However, the sample size in the Shaw et al. (2003) analysis ($n = 284$) is effectively larger than in Nagin and Land (1993; $n = 403$) in one very important respect. The Shaw et al. analysis examines trajectories of problem behaviors based on psychometric measures in which most individuals display at least some minimal level of symptoms. By contrast, the Nagin and Land analysis is based on the conviction histories of 403 individuals, of which about two thirds have no convictions at any age. Thus, only about 150 individuals form the basis for the three criminal trajectory groups that were the focus of their analysis. The analyses that are presented here were based on arrest histories. The incidence of arrest in this sample is somewhat higher than of conviction in the sample used by Nagin and Land, but still more than half of all individuals have no arrests. Thus, to reflect the range of sample sizes analyzed in extant applications, we conducted our analysis for sample sizes of $n = 500$, 1,000, and 1,500.

As depicted in Figure 1, we repeatedly drew random samples of $n = 500$, 1,000, and 1,500 from the population and estimated the same model for each such sample. The models were estimated with an SAS-based estimation procedure described in Jones et al. (2001). By this process, we were able to construct an empirical estimate of the sampling distribution for each of the model's parameters for sample sizes of $n = 500$, 1,000, and 1,500, respectively.

Figure 2 shows this sampling distribution for θ_2 in the $n = 1{,}000$ condition in the form of a histogram. Superimposed on the figure are two vertical bars. One reflects the population value of the parameter from the model estimated on all 13,160 individuals, and the other reflects the average of estimates for the $n = 1{,}000$ condition. The two quantities are so nearly identical that due to the resolution of the graphic, the two bars appear to be one. This suggests that for this parameter, there is little bias in a sample of size of 1,000. Also, observe that the empirical distribution is closely matched by a normal distribution with the same mean and standard deviation. The close match supports the use of conventional normal-based statistical tests. We conducted this type of analysis for each parameter

Figure 2
Histogram of θ_2, $n = 1{,}000$, With Normal Curve Overlaid



estimate and also comparable analyses for the nonlinear transformations of the parameter estimates that calculate $\lambda_j = (\lambda_{j1}, \lambda_{j2}, \ldots, \lambda_{jT})$, for $j = 1, \ldots, J$ and $\pi = (\pi_1 \pi_2 \pi_3 \pi_4)$.

Screening Runs

To generate sampling distributions of the model parameters $\beta^{\mathbf{j}}$ and θ for a given sample size condition, 3,000 random samples were drawn from the total population of 13,160 cases. Similar to the conventions of a bootstrapping procedure, we sampled with replacement. Thus, an individual's data may be used more than once in a given sample. The same four-group trajectory model was estimated for each draw, and its 14 parameters were then recorded.[1] These estimates were used to create the sampling distributions that form the basis for our analysis.

In principle, this procedure should yield 3,000 estimates of each parameter for each sample size condition. In practice, not all of the estimates were usable. Unusable estimates arose from two distinct circumstances: false convergence and convergence at a local maximum.[2]

False convergence occurred when the search procedure did not identify a solution that met convergence criteria. The SAS-based estimation procedure used in this analysis uses an integrated measure of convergence that automatically alerts the user to instances of false convergence. Such solutions are also manifestly problematic because all or most of the parameter estimates commonly have very large standard errors.

The second category of unusable solutions was the result of the estimation procedure converging to a local rather than a global optimum. The problematic status of local maximum solutions was less obvious than the false convergence solutions but still readily identifiable. Specifically, local maximum solutions revealed themselves in two ways. One involved solutions in which two trajectories were nearly indistinguishable. Most commonly, this occurred when not only the zero-order trajectory but also one of the quadratic trajectories described a trajectory of a near-zero rate of offending. The second category of local maximum convergences occurred when a quadratic trajectory was "statistically committed" to describing the near-zero offending trajectory and the zero-order trajectory was used to characterize a high-rate offending trajectory.

After screening the original 3,000 trials for the above problems, most of which were due to false convergence, we were left with 1,713 successful runs for the $n = 1{,}500$ condition, 1,633 runs for the $n = 1{,}000$ condition, and 1,613 runs for the $n = 500$ condition.

We discovered that the problem of false convergence could often be remedied by slight alterations of starting values assigned to the search procedure. Due to the level of automation necessary for running the resampling procedure, we were unable to correct the convergence problem for all iterations where a problem occurred. To make sure that our conclusions were not biased by our only using the initially "successful" runs, we took 10 random samples where false convergence occurred from each of the $n = 1{,}500$, $n = 1{,}000$, and $n = 500$ conditions, and manually altered the starting values until we were able to reach proper convergence. We then compared the results against the general sampling distribution for each condition. In each case, upon reaching a proper convergence, the results behaved similarly to those of the initially successful runs.

Data

The analysis uses data collected by Figlio, Tracy, and Wolfgang (1990) for the purpose of studying delinquency in the 1958 Philadelphia birth

Table 1
Distribution of Combined Period Offenses

	10-11	12-13	14-15	16-17	18-19	20-21	22-23	24-25
0	97.90	92.48	85.24	80.01	88.21	90.57	91.11	92.70
1	1.63	5.33	8.68	10.72	7.57	6.42	5.78	5.27
2	0.27	1.12	3.03	3.90	2.30	1.76	1.93	1.28
3	0.10	0.55	1.33	2.17	0.90	0.73	0.59	0.50
4	0.04	0.18	0.71	1.27	0.49	0.27	0.35	0.14
5	0.02	0.10	0.39	0.75	0.25	0.13	0.10	0.08
6	0.03	0.13	0.24	0.35	0.14	0.08	0.07	0.02
7	0.00	0.04	0.14	0.26	0.05	0.02	0.02	0.00
8	0.02	0.02	0.05	0.22	0.04	0.02	0.02	0.01
9	0.00	0.02	0.06	0.14	0.02	0.00	0.02	0.00
10	0.00	0.04	0.14	0.21	0.03	0.02	0.02	0.01

cohort. The data set contains longitudinal records of arrest and other police interaction of the cohort members from ages 10 through 26. For the purposes of this project, we examine only the males, which constitute 13,160 individuals, and use only the police contacts resulting in arrest.

Table 1 reports a frequency distribution of arrests by age, in two-year combined periods. Because only a very small proportion of cohort members had more than one arrest in a given year, the annual data are nearly indistinguishable from a binary process. Therefore, we combine the data into two-year periods to increase the frequency arrest counts greater than 1. All combined periods are top-coded at a maximum of 10 offenses. This top-coding rule, which affects less than .01 percent of the observation, was employed because this small contingent of outlying data points was the source of considerable model instability.

Results

Population Model

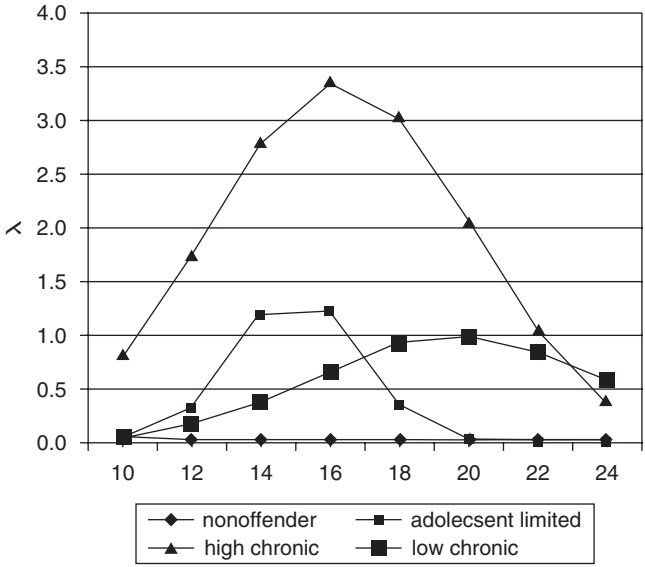
The first step in the analysis involves estimation of the population-level model. As discussed above, the population model was specified as a four-group model, with one group following a constant trajectory and the other three groups following a quadratic trajectory in age. The constant trajectory was included to model the arrest trajectory of the large percentage of subjects in the population who were either never arrested or had at most

one arrest over the age period from 8 to 25. This trajectory was intended to capture the null/nearly null arrest trajectory with a constant and small value of λ.[3] The three quadratic trajectory groups were used to model the arrest histories of individuals who were active offenders, as measured by arrest, over at least some part of the age period from 8 to 26.

Our decision to specify the population model in this fashion follows a model selection strategy outlined in Nagin (2005). This strategy begins by attempting to use the Bayesian information criterion (BIC) as the standard for deciding how many trajectory groups to include in the model. Specifically, it is recommended that the model with the preferred number of groups be the model that maximizes the BIC.

In this application, the BIC was not helpful in identifying a preferred model because over the range of models explored, BIC monotonically increased with the number of groups. We settled on the four-group model, depicted in Figure 3, for several reasons. First and foremost, the four trajectories are quite distinctive. The zero-order ''nonoffender'' trajectory group is composed mostly of individuals with no arrests. Another trajectory follows the classic hump-shaped pattern in which arrest rate peaks at age 16 and then declines to a negligible level by age 20. Moffitt (1993) calls this group ''adolescent-limited offenders.'' Another group also peaks in adolescence but at a much higher rate. Even at age 26, these individuals are offending at a material rate. We call this group ''high-rate'' chronics. The final group can be characterized as a ''low-rate chronic'' group. Their peak rate of offending at age 18 is actually less than the age 16 peak of the adolescent-limited group but, unlike the adolescent-limited group, their offending persists into their 20s. The addition of more groups does not yield any new distinctive patterns. Instead, one of the three offending groups depicted in Figure 3 is merely split. A second reason that we prefer the four-group model is that, as discussed next, its parameters are very precisely estimated. We concede that this four-group model is almost certainly not the ''true'' model. Indeed, as elaborated in Nagin (2005) and elsewhere (Nagin and Land 1993; Nagin and Tremblay 2001), the group-based trajectory should be regarded as only an approximation of a more complex underlying reality. However, the focus of this analysis is not on model selection but rather on the finite sample properties of the selected model. We can, of course, never know whether the selected model is the true model, but that uncertainty neither precludes an investigation into the statistical properties of the selected model nor obviates the utility of such an investigation. Thus, when we refer to the parameter estimates based on all the data at our disposal as the *population parameters*, we only use this

Figure 3
Four-Group Population Trajectory Model



term in the conditional sense of our having selected the four-group model as our preferred representation of the data.

Table 2 reports the estimates of the vectors β^j, which determine the trajectories, and θ, which determine the trajectory group probabilities, along with the z scores of each estimate. In each case, these z scores are very large, which implies that each of these parameters is very precisely estimated. Thus, the approximation error of treating the parameter estimates for the entire birth cohort as if they equaled the true population values is small. Table 3 reports the actual trajectories and group membership probabilities for this model, which are calculated from transformations of these population-level quantities as specified by equations (1), (2), and (3).

Bias

We turn now to an examination of the bias in the estimation of population values under the $n = 500$, 1,000, and 1,500 sample conditions. We do

Table 2
Population Parameter Estimates

Group	Parameter	Estimate	z Score
Nonoffender	Intercept	–2.02	–16.38
Adolescent limited	Intercept	–35.71	–32.34
	Linear	47.93	32.13
	Quadratic	–15.93	–31.81
High chronic	Intercept	–8.30	–29.16
	Linear	11.68	33.14
	Quadratic	–3.59	–33.25
Low chronic	Intercept	–11.33	–31.61
	Linear	11.52	29.19
	Quadratic	–2.93	–27.37
	θ_2	–1.87	21.64
	θ_3	–3.22	16.04
	θ_4	–1.96	30.18

Table 3
Population Trajectories and Group Mixing Probabilities

		λ_{iT}							
	π_j	10	12	14	16	18	20	22	24
Nonoffender	74.87	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029
Adolescent limited	11.55	0.025	0.323	1.186	1.219	0.350	0.028	0.001	0.000
High chronic	2.98	0.816	1.743	2.792	3.358	3.031	2.054	1.044	0.399
Low chronic	10.60	0.065	0.178	0.390	0.673	0.921	0.997	0.854	0.579

not use conventional hypothesis tests in making this assessment for several reasons. First, the maximum likelihood estimates are only asymptotically unbiased. In finite samples, the issue is not whether there is bias; rather, the issue is whether this bias is substantively large. Hypothesis tests are not designed to assess whether a bias is substantial. Second, we have more than 1,600 estimates of parameter values for each sample size condition. The combination of a large sample size and only modest variation in the estimates themselves gives rise to rejection of the null hypothesis that the parameter mean under each sample condition equals the population value, even when the difference is small. This is still another manifestation of our

first observation that the issue is not whether there is bias but whether it is substantively large.

Table 4 reports the mean estimates of the $\beta^{\mathbf{j}}$ and θ parameters for each of the three sample conditions. Also reported are the parameter values of the population model and the absolute percent difference between each parameter's population value and the corresponding mean under each of the sample conditions. The percent difference summary statistics indicate that the error between the population value and the mean of the estimates under the various sample conditions is generally small. One seeming exception to this conclusion involves the intercept estimate for the nonoffending group, β_0^1, which is discussed separately below.

Excluding β_0^1 from the calculation, the mean absolute percent difference calculated across all of the model's parameters declines as sample size increases. For the $n = 1{,}500$ and $n = 1{,}000$ conditions, the mean percent error is only 2.3 percent and 3.5 percent, respectively. The percent error never exceeds 10 percent for any individual parameter other than β_0^1. For the $n = 500$ condition, the mean absolute error increases to more than 7 percent, and errors in the parameters specifying the high chronic trajectory exceed 10 percent. Even these differences suggest only a modest degree of bias in sample sizes as small as 500.

However, our interest is not in the parameters $\beta^{\mathbf{j}}$ and θ per se, but rather in the trajectories of λ_t^j and the group membership probabilities that are determined by them. Because exponentiation is involved in the calculations that transform the $\beta^{\mathbf{j}}$ and θ parameters into the trajectories of λ_t^j and the group membership probabilities, respectively, seemingly small numerical differences in the untransformed parameters may lead to large differences in the transformed quantities. Also, the covariance of the parameter estimates may affect these calculations in unknown ways.

Table 5 reports the population trajectories of λ_t^j as well as the average trajectories of λ_t^j for each sample condition. Specifically, for each sample run that resulted in proper convergence, the estimates of $\beta^{\mathbf{j}}$ were transformed according to equations (1) and (2) and then averaged. The table suggests that even for $n = 500$, the bias in the estimates of λ_t^j is small to modest.

Consider first the zero-order trajectory used to model the negligible offender group. According to the population model, the mean biannual arrest rate for this group is .029 arrests per two-year period. The more than 10 percent difference between the population value of β_0^1 and the average value of β_0^1 across runs in each sample condition translated into a negligible bias in the estimate of λ for this group.

(text continues on p. 266)

Table 4
Mean Estimates and Percentage Differences of Parameters

			$n = 1,500$		$n = 1,000$		$n = 500$	
	Variable	Population	Mean	Percent Difference	Mean	Percent Difference	Mean	Percent Difference
Nonoffender	Intercept	–2.02	–2.32	13.05	–2.27	10.89	–2.32	12.98
Adolescent limited	Intercept	–35.71	–35.93	0.62	–37.12	3.82	–39.57	9.77
	Linear	47.93	48.00	0.14	49.49	3.16	52.63	8.94
	Quadratic	–15.93	–15.89	0.25	–16.36	2.63	–17.38	8.34
High chronic	Intercept	–8.30	–8.85	6.20	–9.04	8.18	–9.61	13.64
	Linear	11.68	12.39	5.71	12.66	7.70	13.44	13.09
	Quadratic	–3.59	–3.84	6.62	–3.93	8.84	–4.20	14.70
Low chronic	Intercept	–11.33	–11.32	0.10	–11.35	0.13	–11.79	3.89
	Linear	11.52	11.44	0.71	11.48	0.38	11.97	3.80
	Quadratic	–2.93	–2.90	0.78	–2.92	0.27	–3.07	4.53
	θ_1	–1.87	–1.90	1.43	–1.89	1.05	–1.86	0.25
	θ_2	–3.22	–3.11	3.58	–3.10	3.91	–3.08	4.58
	θ_3	–1.96	–1.99	1.85	–1.99	1.53	–1.95	0.11
Mean absolute percentage difference				3.16		4.04		7.59
Mean excluding β_0^1				2.33		3.47		7.14

Table 5
Mean Trajectories by Condition

		Population	$n = 1,500$		$n = 1,000$		$n = 500$	
	Age	λ_{it}	λ_{it}	Absolute Difference	λ_{it}	Absolute Difference	λ_{it}	Absolute Difference
Nonoffender	(zero-order)	0.029	0.030	0.001	0.030	0.001	0.029	0.000
Adolescent limited	10	0.025	0.030	0.005	0.029	0.005	0.034	0.010
	12	0.323	0.325	0.003	0.320	0.003	0.321	0.001
	14	1.186	1.175	0.011	1.174	0.013	1.180	0.006
	16	1.219	1.236	0.016	1.245	0.025	1.254	0.034
	18	0.350	0.374	0.023	0.371	0.020	0.360	0.010
	20	0.028	0.037	0.009	0.036	0.008	0.041	0.013
	22	0.001	0.002	0.001	0.002	0.001	0.004	0.003
	24	0.000	0.000	0.000	0.000	0.000	0.001	0.001
Mean		0.392	0.397	0.009	0.397	0.009	0.399	0.010
Percentage				2.16		2.37		2.43
High chronic	10	0.816	0.778	0.038	0.782	0.034	0.799	0.017
	12	1.743	1.680	0.063	1.690	0.053	1.720	0.023
	14	2.792	2.691	0.101	2.710	0.083	2.764	0.028
	16	3.358	3.173	0.185	3.177	0.181	3.191	0.167
	18	3.031	2.766	0.265	2.746	0.285	2.695	0.336
	20	2.054	1.808	0.246	1.787	0.267	1.738	0.315
	22	1.044	0.902	0.143	0.898	0.147	0.889	0.155
	24	0.399	0.351	0.048	0.358	0.041	0.375	0.024
Mean		1.905	1.769	0.136	1.768	0.136	1.771	0.133
Percentage				7.69		7.70		7.52

Low chronic	10	0.065	0.065	0.001	0.067	0.003	0.070	0.005
	12	0.178	0.175	0.003	0.178	0.000	0.181	0.003
	14	0.390	0.378	0.011	0.381	0.008	0.385	0.005
	16	0.673	0.649	0.024	0.652	0.021	0.656	0.018
	18	0.921	0.886	0.035	0.888	0.033	0.889	0.033
	20	0.997	0.964	0.034	0.964	0.034	0.961	0.037
	22	0.854	0.835	0.019	0.835	0.020	0.830	0.024
	24	0.579	0.579	0.000	0.580	0.001	0.578	0.001
Mean		0.582	0.566	0.016	0.568	0.015	0.569	0.016
Percentage				2.83		2.64		2.75

Table 6
Mean and Standard Deviation of Group Mixing Probabilities

		Condition			
		Population	$n = 1,500$	$n = 1,000$	$n = 500$
Nonoffender	Mean	0.749	0.747	0.744	0.734
	Standard deviation		0.032	0.039	0.048
Adolescent limited	Mean	0.116	0.115	0.117	0.121
	Standard deviation		0.025	0.030	0.040
High chronic	Mean	0.030	0.034	0.035	0.036
	Standard deviation		0.008	0.010	0.014
Low chronic	Mean	0.106	0.104	0.104	0.109
	Standard deviation		0.020	0.023	0.034

Biases in the trajectories of λ_t^j for the adolescent-limited and low-rate chronic groups were also negligible for all sample conditions. For the adolescent-limited group, the absolute difference between the population value of λ_t^j at each age and the average estimate for that age for each sample size condition never exceeds .034 arrests per biennium. The average bias over the period from ages 10 to 24 is less than .01 arrests per biennium. This discrepancy is less than 3 percent of the average rate of arrest over this period of about .4 arrests per biennium. Inspection of the discrepancies for the low-level chronic group between population values of λ_t^j and the average for the various sample conditions shows similarly small differences.

Only for the high-level chronic group are the biases nontrivial. For each sample size condition, the average absolute difference statistic suggests a bias of about .14 arrests per biennium, which is about 8 percent of the average of the biannual arrest rate for the high-chronic group. While an 8 percent bias is not inconsequential, it still seems best characterized as modest.

Table 6 reports comparisons of the population values of the group membership probabilities with the average estimate of each probability by sample condition. The population probabilities are calculated from the values of θ reported in Table 2 according to equation (3). The sample averages are based on the application of this same equation to the estimates of θ from each proper convergence. The results show that population values are nearly identical to the average from each simulation condition. This suggests a minimal level of bias in the estimates of the mixing probabilities.

Asymptotic Normality

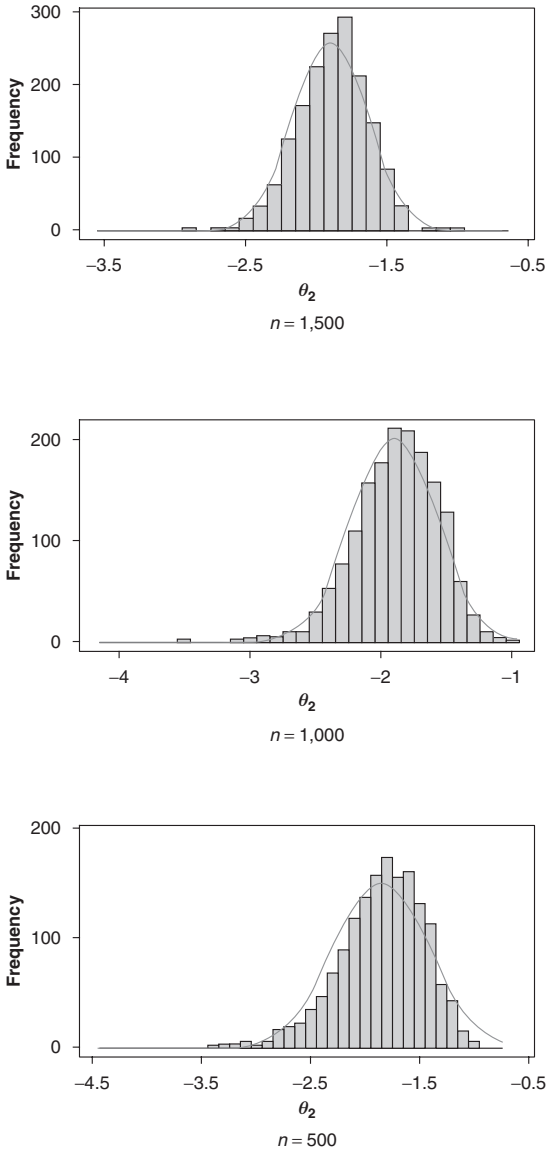
Another key property of maximum likelihood parameter estimates is their asymptotic normality. Specifically, as the size of the sample used in the model estimation increases, the distribution of each of the parameters should more closely approximate a normal distribution. Again, this is a large-sample property. For the sake of space, we only report analyses of the distribution of the three untransformed parameters comprising θ that are used in the calculation of the group mixing probabilities. The distributions of the components of each β^{j}, which specify the shapes of the trajectories, were found to be similarly normal.

As with our examination of bias, our focus is on substantive deviations from normality, not on statistical tests of normality—we already know that the distributions are unlikely to be strictly normal. The key issue is whether the assumption of normality seems to be a good approximation. This assessment is conducted in two stages. We first assess whether the empirical distribution seems to be normal in appearance. We acknowledge at the outset that this judgment is largely qualitative. The more rigorous and practically important assessment comes in the Confidence Intervals section, where we compare normal-based confidence intervals with their counterparts from the empirically generated sampling distributions. We characterize this as the more practically important assessment because our interest in the normality of the estimates stems from the centrality of the assumption of normality in conventional statistical inference.

For each of the sample size conditions, we begin the first stage assessment by examining the empirical distribution of the results generated through the resampling process. Figure 4 shows histograms of the estimate of θ_2 for $n = 500$, 1,000, and 1,500. On each graph, a normal curve is overlaid. This normal curve was generated using the sample mean and standard deviation from the estimates of θ_2 used to generate each respective histogram. Also, the distributions in each condition are properly centered about their respective population mean. Similarly, histograms for θ_3 and θ_4, omitted here for sake of space, also show sampling distributions that appear to be approximately normal.

We also examined the skew and kurtosis of the distributions. We generally found some evidence of skew and an excess of probability mass in the tails (i.e., kurtosis), which are noncongruent to the characteristics of a normal distribution. However, we reiterate our prior commentary on the importance of hypothesis testing. The issue is not whether these sampling distributions literally follow a normal distribution; rather, it is whether the

Figure 4
Distribution of θ_2 Parameters



assumption of normality seems to lead to material error in statistical inference. We directly address this issue in the Confidence Intervals section.

It is also revealing to examine how the standard deviation of the sampling distribution changes across conditions. As the sample size decreases, the standard deviation of the distribution and thereby the standard error of the parameter estimate it is characterizing should increase. For θ_2, the standard deviation is .2647 in the $n = 1{,}500$ condition, .3235 in the $n = 1{,}000$ condition, and .4294 in the $n = 500$ condition. For θ_3, the standard deviations are .2795, .3409, and .4524 in the three respective conditions. For θ_4, the standard deviations are .1991, .2360, and .3452. For each of these parameters, the standard deviations of the estimates monotonically increase as the sample size is reduced.

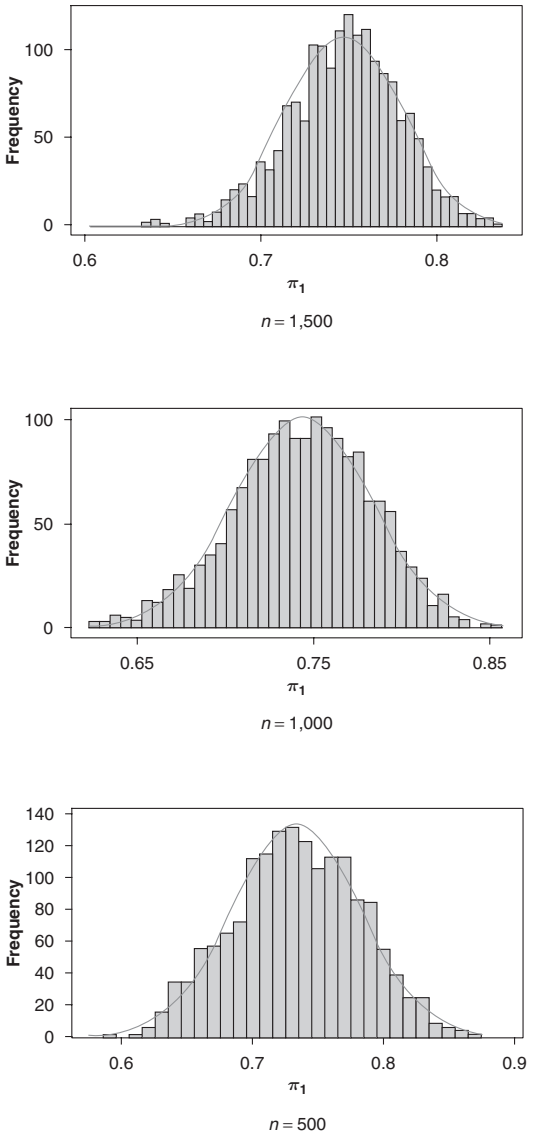
Even though the trajectories and the group mixing probabilities are nonlinear transformations of parameters estimated by maximum likelihood, the invariance property of maximum likelihood (Goldberger 1991) implies that they too are normally distributed asymptotically. Thus, we also examined the normality of the transformed quantities under the three finite sample conditions. Figure 5 shows histograms of π_1, the rare offender mixing probability, for $n = 500$, 1,000, and 1,500. As before, on each graph, a normal curve is overlaid with the same sample mean and standard deviation of the estimates of π_1 used to generate each respective histogram. In each of the three conditions, π_j looks to have an approximate normal distribution that is centered about the sample mean, with the variance increasing as the sample size decreases. Histograms for π_2, π_3, and π_4, again omitted here in the interest of space, reveal similar approximate normality in each of the three conditions.

Confidence Intervals

As discussed at the outset of the last section, the reason for our concern with the normality assumption is its central role in conventional statistical inference. We examine the adequacy of normal-based tests by first calculating confidence intervals under the assumption that the parameter estimate is normally distributed and then comparing these intervals with the empirically based counterpart intervals extracted from the distributions of estimates that result from repeated sampling from the population.

Consider the calculation of a 95 percent confidence interval for any parameter estimate, $\hat{\alpha}$, that is assumed to be normally distributed. It equals $\hat{\alpha} \pm 1.96 se(\hat{\alpha})$, where $se(\hat{\alpha})$ is the standard error of $\hat{\alpha}$. We examine the adequacy of this calculation in two stages. We first examine whether the

Figure 5
Distribution of Group 1 Mixing Probability



maximum likelihood estimate of $se(\hat{\alpha})$ appears to be an unbiased estimate of the standard deviation of $\hat{\alpha}$. We do this by comparing the average value of maximum likelihood estimates of $se(\hat{\alpha})$ for selected parameters with the standard deviation of the empirical distribution of that parameter. If the maximum likelihood estimate of $se(\hat{\alpha})$ is approximately unbiased, then these two quantities should be close in magnitude. We perform this examination for the $n = 500$ condition, as it is the smallest sample size we consider and hence the most vulnerable to bias.

Table 7 reports a comparison of the average maximum likelihood $se(\hat{\alpha})$, as calculated from a random sample from the $n = 500$ condition, against the $se(\hat{\alpha})$ generated from the $n = 500$ sampling distribution.[4] Note that the nonoffending group is omitted as the differences in the parameter β_0^1 are negligible, as previously explained. Also, note that the values reported in Table 7 are "trim" values. That is, the calculations are based on the middle 95 percent of the empirical data. We use these trim values because there are a small number of extreme solutions in which the maximum likelihood parameter estimates are drastically different than rest of the solution yet still produce trajectories that closely match those of the main body of solutions. That is, these solutions represent instances in which the estimates of β^j are wildly different from the rest of the sampling distribution but nonetheless transform into similar trajectories. In turn, the maximum likelihood standard errors associated with these parameter estimates are also much different than the main body of estimates and thus distort our estimate of the true mean standard error. It seems as if these solutions represent yet another case of convergence at a local maximum, albeit one that is harder to detect given the generally similar shapes of the trajectories produced. Since these types of instances are quite rare, simply considering only the trim values of the sampling distribution provides an adequate guard against their inclusion into the analysis. Once we consider these trim values, we find that the mean standard errors from the sampling distribution and the mean maximum likelihood standard error are generally close in magnitude.

Consider the case of the adolescent-limited trajectory, which is denoted by the superscript 2. The mean standard error of β_0^2 from the sampling distribution is 8.29, and the mean maximum likelihood standard error of β_0^2 is 8.66. Similarly, the mean standard errors of β_1^2 and β_2^2 from the sampling distribution are 11.03 and 3.67, respectively, and the mean maximum likelihood standard errors of these two parameters are, respectively, 11.38 and 3.73, which again are very close in magnitude. Also, the two sets of standard errors for the low-chronic group parameters (labeled as Group 4 in the notation) are nearly identical. For β_0^4, β_1^4, and β_2^4, the mean standard errors

Table 7
Maximum Likelihood Standard Errors Versus Sampling Distribution Standard Errors, $n = 500$ Condition

	Adolescent Limited			High Chronic			Low Chronic		
	β_0^2	β_1^2	β_2^2	β_0^3	β_1^3	β_2^3	β_0^4	β_1^4	β_2^4
Sample	8.29	11.03	3.67	2.90	3.93	1.36	3.42	4.39	1.44
MLE	8.66	11.38	3.73	1.81	2.33	0.74	3.67	4.55	1.42

Note: MLE = maximum likelihood estimation.

from the sampling distribution are 3.42, 4.39, and 1.44, respectively, and the mean maximum likelihood standard errors are 3.67, 4.55, and 1.42, respectively. The similarity is not quite so good in the case of the high-chronic group (labeled as Group 3). However, as will be shown below, even for this group, confidence interval calculations based on these maximum likelihood standard errors reasonably correspond with those extracted directly from the sampling distribution. The similarity of the maximum likelihood standard errors to the standard errors of the sampling distribution, particularly in the adolescent-limited and low-chronic group parameters, are important for inferential purposes because they suggest that confidence intervals based on the maximum likelihood standard errors will be accurate. The next step in the analysis involves a comparison of the calculation $\bar{\hat{\alpha}} \pm 1.96\bar{se}(\hat{\alpha})$, where $\bar{\hat{\alpha}}$ and $\bar{se}(\hat{\alpha})$ are, respectively, the average of the maximum likelihood–based estimates of α and its standard error, with an interval equaling the 2.5 and 97.5 percentiles of the empirically based sampling distribution. Table 8 reports 95 percent confidence intervals for each group membership probability created in the two separate ways. The quantities labeled NDB (normal distribution based) are conventional confidence intervals formed by the calculation $\bar{\hat{\alpha}} \pm 1.96\bar{se}(\hat{\alpha})$. These represent calculations a user would be able to do in a single trial, as the user will have access to the standard errors estimated by Proc Traj on each single trial.[5]

The quantities labeled as empirical represent the interval equaling the 2.5 and 97.5 percentiles of the empirically based sampling distribution. Even though a user will not have the luxury of forming confidence intervals in this manner, the similarity of the intervals in the two methods suggest that the former method is indeed an appropriate method to use. Table 9 reports 95 percent confidence intervals for calculated values of λ_t^j at ages 14 and 20. The two types of intervals were constructed in the same manner

Table 8
Ninety-Five Percent Confidence Intervals for Group MixingProbabilities, Standard and Empirical

		Nonoffenders		Adolescent Limited		High Chronic		Low Chronic	
		NDB	Empirical	NDB	Empirical	NDB	Empirical	NDB	Empirical
$n = 1,500$	Lower	0.685	0.681	0.066	0.068	0.018	0.019	0.064	0.075
	Upper	0.810	0.806	0.164	0.163	0.051	0.051	0.143	0.144
$n = 1,000$	Lower	0.668	0.663	0.057	0.061	0.015	0.017	0.059	0.070
	Upper	0.820	0.814	0.176	0.179	0.055	0.057	0.150	0.157
$n = 500$	Lower	0.640	0.640	0.043	0.045	0.009	0.014	0.042	0.057
	Upper	0.828	0.826	0.199	0.200	0.064	0.068	0.175	0.200

Note: NDB = normal distribution based.

Table 9
Ninety-Five Percent Confidence Interval for Selected Trajectories, Standard and Empirical

			Nonoffenders		Adolescent Limited		High Chronic		Low Chronic	
			NDB	Empirical	NDB	Empirical	NDB	Empirical	NDB	Empirical
$n = 1,500$	λ, age 14	Lower	0.018	0.019	0.444	0.681	1.716	1.748	0.194	0.238
		Upper	0.041	0.042	1.904	1.970	3.663	3.731	0.562	0.594
	λ, age 20	Lower	0.018	0.019	–0.065	0.009	1.111	1.003	0.566	0.552
		Upper	0.041	0.042	0.139	0.081	2.502	2.454	1.360	1.360
$n = 1,000$	λ, age 14	Lower	0.016	0.017	0.253	0.563	1.512	1.607	0.164	0.217
		Upper	0.044	0.044	2.094	2.244	3.907	4.082	0.599	0.650
	λ, age 20	Lower	0.016	0.017	–0.038	0.005	0.916	0.760	0.498	0.488
		Upper	0.044	0.044	0.110	0.100	2.659	2.612	1.429	1.442
$n = 500$	λ, age 14	Lower	0.012	0.014	–0.081	0.453	1.057	1.298	0.085	0.161
		Upper	0.046	0.047	2.441	2.799	4.472	4.754	0.684	0.759
	λ, age 20	Lower	0.012	0.014	–0.074	0.001	0.544	0.426	0.331	0.384
		Upper	0.046	0.047	0.156	0.180	2.933	2.934	1.590	1.625

Note: NDB = normal distribution based.

as the group membership probability confidence intervals above.[6] Again, the similarity of these two intervals across groups and ages reinforces the validity of using the calculated maximum likelihood standard errors in the creation of confidence intervals for a single trial.

Conclusion

In this article, we test whether the asymptotic properties of maximum likelihood estimation are achieved in sample sizes typically used in applications of group-based trajectory modeling. Through empirical results generated by resampling of population data, we find that the maximum likelihood estimates obtained in group-based trajectory models still provide reasonably close estimates of their true population values and have approximately normal distributions, even when estimated with a sample size as small as $n = 500$. Furthermore, and more important for the users of these types of models, we find similarly good performance in the model's ability to estimate the transformed quantities of main interest: the group trajectories and mixing probabilities. The behavior of these transformed quantities, which appear to also follow a normal distribution, suggests that the formation of confidence intervals using the delta-method calculation (from Proc Traj) is appropriate, as the maximum likelihood standard errors are generally very similar to the standard errors obtained from the sampling distributions of these quantities. We find that confidence intervals for both trajectories and group membership probabilities created by using typical normal-based calculations, which a user would easily be able to do, are close to the confidence intervals extracted from the more difficult to create empirical sampling distributions. These results suggest that users of this methodology can have confidence that the two key asymptotic properties of maximum likelihood estimates, unbiasedness and normality, are achieved in relatively small samples.

Our conclusions are based on a specific set of simulation conditions. The finite sample properties of the parameter estimates may be affected by many features of the model. One that we suspect is very important is the separation of the components of the mixture. Separation refers to the certainty with which the sampled units, which in our analysis are people, can be assigned to components of the mixtures—namely, the trajectory groups. We suspect that poor separation will adversely affect finite sample properties. The separation of the groups can be calibrated by the average posterior probabilities of group assignment. Specifically, after estimation,

the parameters of the model can be used to calculate the probability that the data for each individual in the estimation sample were generated by each trajectory group. This probability, called the posterior probability of group membership, can be used to assign the individuals to the group that most likely generated their data. Ideally, this probability will be close to 1 for the assigned group. In the population model, we find that the mean posterior probability of assignment to the nonoffending group for those assigned to that group is .95. Counterpart mean posterior assignment probabilities for the adolescent-limited group, the low-chronic group, and the high-chronic group are .86, .84, and .92, respectively. This suggests that our groups are indeed well separated and, by implication, that in models that are less well separated, finite sample properties may be less satisfactory. However, we note that Nagin (2005) recommends that a minimum requirement for a satisfactory model be that all mean posterior assignment probabilities exceed .7. Thus, if users adhere to this requirement, it would seem that they can proceed with some confidence that finite sample properties are good. Still, it would be desirable to explore further the separation issue. We recommend, however, that this be done with simulated, not real, data, so that the degree of separation of the mixture components can be controlled and properly evaluated. Another limiting factor in this analysis is that we examined only the Poisson-based model. Therefore, another useful extension of this work is to binary-based and censored normal-based trajectory models where the finite sample properties may be different. A final extension that we suggest involves an effort to identify the smallest feasible sample size for group-based trajectory models. Stated differently, the objective of the analysis would be to identify the point at which the maximum likelihood estimates "break." We recommend that such an examination vary not only the number of subjects in the analysis but also the number of periods over which they are observed.

Notes

1. In the relevant model, the 14 parameters include the intercept for the zero-order group, an intercept, a linear and quadratic term for each of the three quadratic groups, three estimates for the group mixing probability calculation (the first one is always zero), and a zero-inflation parameter (see note 3).

2. Another potential problem occurs in properly binning the parameters across samples. The estimation procedure does not guarantee a specified group number will always model the same trajectory group across samples. For example, in principle, what we call the high-chronic trajectory (see Population Model section) may not be captured by the same group number across runs. This complicates the automation of properly binning the parameter

estimates, which is crucial step in creating the sampling distributions. Fortunately, we found no evidence of improper binning in this application.

3. We also included a zero-inflation component in this nonoffender trajectory as still another statistical device for accounting for the overabundance of zeros in this group. See Nagin (2005) or Nagin and Land (1993) for a discussion of use of the zero-inflated Poisson model in the context of group-based trajectory modeling.

4. The calculation of $\bar{\hat{\alpha}}$ and $\bar{se}(\hat{\alpha})$ was actually based on a random sample of only 150 of the samples resulting in proper convergence. This was done for the purely practical reason that the estimation procedure is much faster when it only estimates the model's parameters and does not go through the additional step of estimating their standard errors.

5. Note that the estimates of $\bar{se}(\hat{\alpha})$ for each of the group mixing probabilities, as well as for the trajectories, cannot be estimated directly. There are several methods that can be employed to approximate these quantities. One such method, the delta-method, which is employed by Proc Traj, uses a first-order Taylor series approximation to estimate these quantities. For more details on the delta-method, see Greene (1990).

6. Note that when the interval for the trajectory crosses 0 (i.e., is negative), it is reported in Table 9 as such, even though in practice, the true parameter cannot be negative.

References

Bryk, A. S. and S. W. Raudenbush. 1987. "Application of Hierarchical Linear Models to Assessing Change." *Psychology Bulletin* 101:147-58.

———. 1992. *Hierarchical Linear Models for Social and Behavioral Research: Application and Data Analysis Methods*. Newbury Park, CA: Sage.

Eggleston, E. P., J. H. Laub, and R. J. Sampson. 2004. "Methodological Sensitivities to Latent Class Analysis of Long-Term Criminal Trajectories." *Journal of Quantitative Criminology* 20:1-26.

Fergusson, D., L. J. Horwood, and D. S. Nagin. 2000. "Offending Trajectories in a New Zealand Cohort." *Criminology* 38:525-52.

Figlio, Robert M., Paul E. Tracy, and Marvin E. Wolfgang. 1990. *Delinquency in a Birth Cohort II: Philadelphia, 1958-1988*. Philadelphia: Sellin Center for Studies in Criminology and Criminal Law and National Analysts.

Goldberger, Arthur S. 1991. *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Goldstein, H. 1995. *Multilevel Statistical Models*. 2nd ed. London: Edward Arnold.

Greene, William H. 1990. *Econometric analysis*. New York: Macmillan.

Jones, Bobby L., Daniel Nagin, and Kathryn Roeder. 2001. "A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories." *Sociological Research and Methods* 29:374-93.

Kiefer, J. and J. Wolfowitz. 1956. "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Parameters." *Ann. Math. Statist.* 27:887-906.

McArdle, J. J. and D. Epstein. 1987. "Latent Growth Curves Within Developmental Structural Equation Models." *Child Development* 58:110-13.

Meredith, W. and J. Tisak. 1990. "Latent Curve Analysis." *Psychometrika* 55:107-22.

Muthén, B. 1989. "Latent Variable Modeling in Heterogeneous Populations." *Psychometrika* 54:557-85.

Muthén, L. K. and B. O. Muthén. 2004. *Mplus User's Guide*. 3rd ed. Los Angeles: Muthén & Muthén.

Moffitt, Terrie E. 1993. "Adolescence-Limited and Life-Course Persistent Antisocial Behavior: A Developmental Taxonomy." *Psychological Review* 100:674-701.

Nagin, Daniel S. 1999. "Analyzing Developmental Trajectories: A Semi-Parametric, Group-Based Approach." *Psychological Methods* 4:139-77.

———. 2004. "Response to 'Methodological Sensitivities to Latent Class Analysis of Long-Term Criminal Trajectories." *Journal of Quantitative Criminology* 20:27-36.

———. 2005. *Group-Based Modeling of Development*. Cambridge, MA: Harvard University Press.

Nagin, D. S. and K. C. Land. 1993. "Age, Criminal Careers, and Population Heterogeneity: Specification and Estimation of a Nonparametric, Mixed Poisson Model." *Criminology* 31:327-62.

Nagin, D. S. and R. E. Tremblay. 2001. "Analyzing Developmental Trajectories of Distinct but Related Behaviors: A Group-Based Method." *Psychological Methods* 6:18-34.

Piquero, Alex R. 2005. "Taking Stock of Developmental Trajectories of Criminal Activity Over the Life Course." Working paper, University of Florida.

Raudenbush, S. W. 2001. "Comparing Personal Trajectories and Drawing Causal Inferences From Longitudinal Data." *Annual Review of Psychology* 52:501-25.

Shaw, D. S., M. Gilliom, E. M. Ingoldsby, and D. Nagin. 2003. "Trajectories Leading to School-Age Conduct Problems." *Developmental Psychology* 39:189-200.

Willett, J. B. and A. G. Sayer. 1994. "Using Covariance Structure Analysis to Detect Correlates and Predictors of Individual Change Over Time." *Psychological Bulletin* 116:363-81.

Tom Loughran is a doctoral candidate at the H. John Heinz III School of Public Policy and Management at Carnegie Mellon University, Pittsburgh, Philadelphia.

Daniel Nagin is the Teresa and H. John Heinz III Professor of Public Policy and Statistics at the Heinz School, Carnegie Mellon University, Pittsburgh, Philadelphia.

Chapter 2: *Consequences of a Violation of the Conditional Independence Assumption in Group-Based Trajectory Models*

Abstract

In the specification of group-based models trajectory models, the following assumption is made. Conditional upon an individual following a given trajectory, the individual's outcomes over successive periods are assumed to be independent. This assumption is a strong one, and it is made specifically in order to make the models more tractable. However, there has been no work done to investigate the behavior of these models in the presence of a failure of this conditional independence assumption. The purpose of this chapter is to explicitly investigate this issue and determine the consequences to users of these models. We propose a Monte Carlo simulation in order to observe how the model behaves when using data simulated specifically to violate this conditional independence assumption, or in more exact terminology, we say that our simulated data has serial correlation. We then examine the results yielded by traditional model estimation in the face of this blatant assumption violation in order to quantify and assess the impact of such results on the average user.

1. Introduction

A developmental trajectory describes the course of a behavior or outcome over age or time. Traditionally, the two most common methods for modeling and analyzing developmental trajectories were hierarchical modeling (Bryk and Raudenbush, 1987, 1992; Goldstein, 1995) and latent curve analysis (McArdle and Epstein, 1987; Meredith and Tisak, 1990; Muthen,1989; Willett and Sayer, 1994). However, Nagin and Land (1993) provide an alternative method – group-based trajectory modeling. The group-based trajectory model, a specialized application of finite mixture modeling, uses mixtures of appropriately defined probability distributions in order to identify distinctive clusters of developmental trajectories within the population. Nagin (1999, 2005) provides much detail regarding both the construction and application of this method. Whereas the hierarchical and latent curve methodologies model population variability in growth with multivariate continuous distribution functions, the group-based approach utilizes a multinomial modeling strategy (Raudenbush, 2001) and is designed to identify relatively homogenous clusters of developmental trajectories.

Two specific software packages, including a SAS-based procedure known as Proc Traj (Jones, Nagin, and Roeder, 2001) and another structural equation modeling software package called M-Plus, developed by Bengt Muthen, Linda Muthen and others (Muthen and Muthen, 2004), have made estimation of these group-based trajectory models easy and straightforward for a wide variety of users, regardless of the level of technical knowledge. As a direct consequence, in recent years, the popularity of the group-based trajectory model has grown rapidly, particularly in fields such as psychology and

criminology. Piquero (2005) reports that more than 50 published articles use group-based trajectory modeling.

Given the relative ease of estimation and the wide potential for application in developmental research, the benefits of using group-based trajectory models are apparent. However, there are criticisms. Notably, in the specification of these group-based models trajectory models, the following assumption is made. Conditional upon an individual *i* being a member of some trajectory group *j*, the individual's outcomes over successive periods are assumed to be independent. This assumption is a strong one, and it is made specifically in order to make the models more tractable. However, there has been no work done to investigate the behavior of these models in the presence of a failure of this conditional independence assumption. The purpose of this chapter is to explicitly investigate this issue and determine the consequences to users of these models. We propose a Monte Carlo simulation in order to observe how the model behaves when using data simulated to violate this conditional independence assumption. Specifically, we generate data where an individual's error terms are correlated over time, such that successive outcomes are not independent conditional on trajectory group. Technically, we say that our simulated data has serial correlation. We then examine the results yielded by traditional model estimation in the face of this blatant assumption violation in order to quantify and assess the impact of such results on the average user.

The rest of this chapter is organized as follows. **Section 2** revisits some basic results from Ordinary Least Squares estimation when serial correlation is present so as to form a logical starting point for the analysis. **Section 3** lays out the general specification of the group-based trajectory model, and shows how the model is altered after the introduction

serial correlation. **Section 4** lays out an experimental design. **Section 5** contains results, while **Section 6** offers a discussion of these results, as well as their practical consequences to users of these models.

2. Results From OLS

Before trying to determine the effects of serially dependant error terms on group-based trajectory models, we can first revisit some results from Ordinary Least Squares (OLS) in order to gain some insight. It is appropriate to draw parallels to OLS here, since at some level, if the groups are sufficiently well-separated, then it is apposite to think the model as a mixture of separate OLS regressions.

Consider the following simple model regressing some outcome *y* on some explanatory terms *x*, which we wish to estimate by OLS: $y_t = \beta_0 + \beta_1 x_t + u_t$, **Eq. (0.1),** where *t* indexes time, u_t is a random error term such that $E(u_t \mid x_t) = 0$, and β_0 and β_1 are, respectively, the intercept and slope parameters we wish to estimate**.** One of the necessary assumptions of the OLS is no serial correlation, or no autocorrelation, of error terms across periods. We may write this assumption as $Cov(u_s , u_t \mid x_s , x_t) = 0$, $s \neq t$, or, $E(u_s u_t \mid x_s , x_t) - E(u_s \mid x_s)E(u_t \mid x_t) = 0$, $s \neq t$. In other words, the assumption in OLS is that the residual term in any given time period is not correlated with the residual terms from any other time periods.

If a user is inclined to believe that an error term in one period *t* has an effect on the error term in a subsequent period, *t + 1*, then the assumption of no serial correlation is

thus violated. Specifically, we may write this as $Cov(u_s , u_t \mid x_s , x_t) \neq 0$, $s \neq t$, or, $E(u_s u_t \mid x_s , x_t) - E(u_s \mid x_s)E(u_t \mid x_t) \neq 0$, $s \neq t$.

As developed in detail below, we consider a simple case of serial correlation, a first-order autoregressive process, denoted as AR(1), where the current period error term u_t is affected by a lag of itself in the previous period (u_{t-1}). Thus, since an individual's error in the current period is systematically related to the error in the previous period, we have a clear violation of the conditional independence assumption.

OLS requires a number of assumptions hold in order to assure that it generates parameter estimates, along with associated standard errors, which are unbiased and consistent. As is the case with any assumption violation in OLS, we are interested in knowing what would happen to the parameter estimates and their associated standard errors which one would obtain from a misspecified model, or, in this particular case, a model where the error terms are correlated over time. Specifically, we would want to know if the parameter estimates obtained by using OLS are still unbiased. In other words, we would still expect to get the 'right' answer on average. In addition, we would want to know if the standard errors of these estimates are still correct. This latter point is of critical importance for drawing proper inference. Furthermore, a user may very well be interested in the t-ratios of the parameter estimates (as calculated from the standard errors), as insignificant higher-order parameters could potentially be discarded in a form of model selection, as advised by Nagin (2005)[1].

[1] For purposes of model selection, Nagin (2005) recommends allowing each group's trajectory to begin as a cubic function of age or time, and then systematically removing the highest order term from each trajectory if that term is insignificant, based on the t-ratio, which is explicitly computed using the standard error of the parameter estimate.

In OLS, the point estimate is derived without making any assumptions concerning the independence of the error term in successive periods. Therefore, even though in the presence of AR(1) the error term violates one of the assumptions of OLS, provided the conditional expectation of the error term is still zero, $E(u_t \mid x) = 0$, then the OLS estimator will still be *unbiased* in the presence of serial correlation. This is an important result, as it suggests that, even in the presence of serial correlation, we still should expect OLS to give us, on average, the 'right' answer.

The problem, however, with using OLS to when serial correlation is present centers on the standard errors of the reported estimates. As noted, even if we expect unbiased point estimates, obtaining incorrect standard errors introduce substantial problems for inferential purposes, as well as potentially be problematic in model selection as noted above. With no serial correlation, the variance of the slope coefficient in OLS can be written as:

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2} \quad \textbf{Eq. (0.2).}$$

As described by Gujarati (2003), however, in the presence of AR(1), the variance of the slope term now becomes:

$$\text{var}(\hat{\beta}_1)_{AR(1)} = \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2}\left[1 + 2\rho\frac{\Sigma(x_{it} - \bar{x}_t)(x_{it-1} - \bar{x}_{t-1})}{\Sigma(x_{it} - \bar{x}_t)^2} + 2\rho^2\frac{\Sigma(x_{it} - \bar{x}_t)(x_{it-2} - \bar{x}_{t-2})}{\Sigma(x_{it} - \bar{x}_t)^2} + \ldots + 2\rho^{n-1}\frac{(x_{i1} - \bar{x}_1)(x_{in} - \bar{x})}{\Sigma(x_{it} - \bar{x}_t)^2}\right]$$

Eq. (0.3).

Furthermore, if we think of *x* as itself having a coefficient of autocorrelation *r*, then Equation 0.3 reduces to the following:

$$\text{var}(\hat{\beta}_1)_{AR(1)} = \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2}\left(\frac{1 + r\rho}{1 - r\rho}\right) = \text{var}(\hat{\beta}_2)_{OLS}\left(\frac{1 + r\rho}{1 - r\rho}\right) \quad \textbf{Eq. (0.4)}.$$

In the case of a trajectory model, where *x* is an individual's age or time, r = 1 since the values are perfectly correlated, which further reduces the 'bias' factor in Equation 0.4 to the following:

$$\text{var}(\hat{\beta}_1)_{AR(1)} = \text{var}(\hat{\beta}_2)_{OLS}\left(\frac{1+\rho}{1-\rho}\right) \quad \textbf{Eq. (0.5).}$$

Notice that this 'bias' factor $(1 + \rho)/(1 - \rho)$ increases the larger the degree of autocorrelation, and, in the case where there is no autocorrelation, it collapse to 1.

It is important to note that the correct variance of the estimator in the presence of serial correlation depends on both the degree of autocorrelation of the error terms, as well as the autocovariances of the x terms in the model. Furthermore, the 'true' variance will exceed the incorrect variance generated by OLS. Operationally, this means that when using OLS in the presence of serial correlation, a user will likely understate the standard errors of the estimates, and perhaps arrive at erroneous inferences since, with smaller standard error leading to smaller confidence intervals, they would be more likely to reject a true null hypothesis. This will lead to the conclusion that parameter estimates are more precise than they actually are. There will be a tendency to reject the null hypothesis when, in fact, it should not be reject (Type I error).

In summary, while the above framework is suitable when dealing with individual OLS regressions, this chapter extends this general discussion into the specifics of group-based trajectory models. Specifically, as these models are an application of finite mixture modeling, the basic framework above does not necessarily predict what will happen if the components of the mixture (i.e., the individual trajectories) are indeed poorly separated. That is, if there is a reasonable degree of ambiguity about an individual's membership in a given trajectory group, due inherently to the fact that the groups are not clearly

separated, it is not entirely clear that these results still hold. Additionally, the basic model can be changed in order to add a time-varying covariate to the trajectory group, which may further complicate the matter. Furthermore, there are other parameters of the group based model, such as the group mixture probabilities, which we want to investigate in the presence of serial correlation and cannot under the above framework. Thus, we feel that our simulation method, which is developed below, is an appropriate manner in which to investigate the behavior of these components of the group based trajectory model.

3. The Group Based Trajectory Model

As described in the introduction, the purpose of group-based trajectory modeling is to identify clusters of similar individual-level trajectories. These clusters form the trajectory groups. A trajectory group is described by the path of the group's expected behavior over age or time and a probability measuring the proportion of the population following this path. In this section we briefly lay out the form of the likelihood function used to estimate these two defining features of a group-based trajectory model. For a more detailed discussion, see Nagin (1999, 2005).

3.1 The Likelihood Function

Let the vector $\mathbf{Y_i}$ denote the longitudinal sequence of individual *i*'s behavior over $t=1,..,T$ periods and let $j=1,..,J$, denote the group. Conditional upon *i* being a member of group *j*, outcomes over successive periods are assumed to be independent. The consequences of a violation of this assumption are explicitly what this paper aims to

assess. Under this assumption, however, the likelihood of observing $\mathbf{Y_i}$ is $P^j(Y_i) = \prod^T p^j(y_{it})$, where $p^j(y_{it})$ is a probability density function. While the generalized form of the model allows for this probability function to be specified as Poisson, binary logit or censored normal, based on the nature of the response variable, in this application we will concentrate exclusively on the normal model. For our purposes, we will also ignore the censoring, as it unnecessarily complicates things, and we may proceed without a loss of generality.

Each group's trajectory is defined by the time path of the some outcome y_{it}. This time path is assumed to follow a polynomial function of age or time. As developed below, our analysis is based on a simple two group model. The trajectory for one group is specified to be constant over age whereas the other group is specified to follow a linear path over time:

$y_{it} = \beta_0^1,\ j=1$ **Eq. (1.1)**

$y_{it} = \beta_0^j + \beta_1^j\ t,\ j=2.$ **Eq. (1.2)**

Notice that the parameters defining each group's trajectory are superscripted by *j*. This allows the time path of y_{jt} to vary freely across groups. Also note that Equations 1.1 and 1.2 can be specified to include additional time-varying covariates, **x**:

$y_{it} = \beta_0^1 + \gamma^1\mathbf{x}_{it},\ j=1$ **Eq. (1.3)**

$y_{it} = \beta_0^j + \beta_1^j\ t + \gamma^2\mathbf{x}_{it},\ j=2.$ **Eq. (1.4).**

For each individual, the unconditional likelihood of observing $\mathbf{Y_i}$ is $P(Y_i) = \sum_j \pi_j P^j(Y_i)$, where π_j is the unconditional probability of membership in

group *j*. The likelihood for the entire sample of N individuals is thus the product of the individual likelihoods: $L = \prod^{N} P(Y_i)$.

The group membership probabilities, π_j, are not estimated directly, but instead are estimated by a generalized logit function:

$$\pi_j = \frac{e^{\theta_j}}{\sum_1^J e^{\theta_j}} \qquad \textbf{Eq. (2.1),}$$

where θ_1 is normalized to zero. Estimation of π_j in this fashion ensures that each such probability properly falls between zero and one. Note that Equation 2.1 can be easily modified so as to allow the probability of group membership to depend on additional risk factors, **z**, which may be added into the logit framework:

$$\pi_j = \frac{e^{\theta_{0,j}+\theta_{1,j}\mathbf{z}}}{\sum_1^J e^{\theta_{0,j}+\theta_{1,j}\mathbf{z}}} \quad \textbf{Eq. (2.2),}$$

3.2 Violation of the Conditional Independence Assumption

As mentioned, we consider a violation of the conditional independence assumption in the case when an individual's error terms in successive periods are serially correlated with each other. In other words, the error term in some period *i* is actually a function of the lagged error term plus another random disturbance. In practice, this would correspond to a large shock in a given period having a persistent effect on an individual's outcomes in future periods. We now explore this case in detail.

Consider the following two-group, censored-normal model with one flat and one rising trajectory:

$$y_i = \beta_0^1 + \varepsilon_{it} + \delta\varepsilon_{i,t-1} \qquad \textbf{Eq. (3.1)}$$

$$y_i = \beta_0^2 + \beta_1^2 t + \varepsilon_{it} + \delta\varepsilon_{i,t-1} \qquad \textbf{Eq. (3.2),} \text{ where}$$

$$\varepsilon_{it} \sim i.i.d.N(0,1) \qquad \textbf{Eq. (3.3),} \text{ and}$$

$$0 < \delta < 1.$$

In this model, we demonstrate a violation of the conditional independence assumption, as any case where the parameter $\delta > 0$ positive serial correlation would exist (we ignore cases of negative serial correlation for the scope of this analysis). This is a case of AR(1) where δ is the parameter of autocorrelation. The user who ignores this serial correlation and accepts the assumption of conditional independence as true would estimate the following model:

$$y_i = \alpha_0^1 + \nu_{it} \qquad \textbf{Eq. (4.1)}$$

$$y_i = \alpha_0^2 + \alpha_1^2 t + \nu_{it} \qquad \textbf{Eq. (4.2).}$$

This model ignores the serially dependant structure of the error term. A large, unobservable shock in any given period could have a potentially persistent effect on the outcome under the influence of the lagged error term in the current period's error. Also, the variance of the error term is larger than the variance of the error represented in Equation 3.3.

In the case described above, the user who accepts the conditional independence assumption would be estimating an incorrectly specified model. The question then becomes to what degree the α parameter estimates (in Equations 4.1 and 4.2), which the user would estimate, are the same as the 'true' parameters (the β parameters in corresponding Equations 3.1 and 3.2), and moreover, how do the maximum likelihood standard errors of these parameters reported by the model in Equations 4.1 and 4.2 compare to the 'true' standard errors of the structural β parameters. These two questions form the foundation of our experimental analysis.

4. Method

4.1 Overview of Approach

We start with a baseline, two group model, consisting of one flat (time-constant) trajectory and one rising trajectory. We systematically introduce serial dependence into randomly generated data in order to test for both biases in the parameter estimates, as well as the effect on the precision (standard errors) of these estimates.

We base our approach on simulated data randomly generated under known parameters. Two considerable benefits of simulation (as opposed to using real-world data) in this case are readily apparent. First, simulating data allows us to know the 'true' model parameters, making all comparisons of the subsequent model estimates legitimate. Second, and more importantly, simulating data allows us to control the degree of serial dependence that exists in the data, and hence manipulate it experimentally.

In quantifying the biases, we concentrate on three main groups of model parameters, with which users of group-based trajectory models are typically concerned, namely 1) the

trajectory-group specific shape parameters (intercepts and linear time trends), 2) the logit coefficients associated with group risk factors, and 3) coefficients on additional time-varying covariates. We repeat iterations of this basic experiment allowing various characteristics, including the sample size, the number of periods and the degree of serial correlation, all to vary.

4.2 Experimental Design

We begin by generating random data according to the data generating processes defined in Equations 3.1-3.3 which include serial correlation. We begin by generating N = 500 observations, split between the two trajectory groups, since this is the likely sample size a typical user of this model will encounter. Also, it is arguable that the number of periods for which the data are tracked, T, is more critical in this aspect than are the number of observations. We begin with T = 4, which again is a typical number of periods a user may have available, and $\delta = .3$, which is a small yet noticeable degree of autocorrelation.

We also wish to test how other factors including sample size, number of periods and the degree of autocorrelation δ affect the results. In subsequent trials, we vary N and T, as well as also systematically increasing the value of δ, from .3 to .5 to .7. For each of the experimental conditions, the same base, two-group, normal model with one flat and one-rising trajectory described above is used. **Table 1** presents a summary of each of the 12 experimental conditions.

The parameters of each of the models we consider are estimated by a direct maximum likelihood procedure available in SAS, known as Proc Traj (Jones, Nagin, and Roeder,

2001). For each experimental condition, we use a quasi-bootstrap procedure where we generate data according to the same model b = 100 times and retain the individual parameter estimates, along with the associated maximum likelihood standard errors, after each trial. The benefits of implementing this bootstrap procedure are two-fold. First, this allows us to control for finite sample effects due to random variation in the data. Second, and more importantly, doing this enables us to create an empirical distribution for each of the parameters, which, as discussed shortly, is a critical step in discovering the 'true' standard errors of each of the estimates. **Figure 1** provides an overview of the analytic strategy.

4.2 Measurement of Bias

Given that we know the 'true' model parameters prior to estimation, we compare the observed values of each parameter, which we obtain from the means of the sampling distributions, directly to these 'true' values in order to detect evidence of bias in the point estimates. The biases in the standard errors, however, require an additional step for comparison. We take the mean of the maximum likelihood standard errors for each parameter estimate as the 'incorrect' result obtained directly from the model. In order to find a proper estimate of the 'true' parameter, the bootstrap proves to be an invaluable tool. We simply take the standard deviation of the sampling distribution of each individual parameter as an empirical estimate of the true standard error. Then we compare these two quantities in order to detect the level of bias.

Another issue we choose to examine is the effect on estimation of the overall group mixing probabilities. In specifying the population parameters for the data simulation, we

allow the separation in the trajectories to be sufficiently muddled so as not to allow the classification to be too distinct[2]. This is important in that it is likely a situation which a typical user may face in practice. While this may potentially introduce another confounding factor into the estimation process, it is important to simulate conditions which are actually likely to be faced by an average user. However, given that we do know the 'true' values of the parameters, we are able to examine each condition, using the same method, but with no serial correlation (that is, purely random error). If the bias under these circumstances is small or zero, then we can feel confident that the separation issue is responsible for no additional bias.

5. Results

6. Discussion

[2] We define the degree of separation to be a case where the mean posterior probability of group classification is less than .9. For details on the calculation of posterior classification probabilities see Nagin (2005).

References

Bryk, A. S., and S. W. Raudenbush. 1987. "Application of Hierarchical Linear Models to Assessing Change." *Psychology Bulletin,* 101: 147-158.

Bryk, A. S., and S. W. Raudenbush. 1992. *Hierarchical Linear Models for Social and Behavioral Research: Application and Data Analysis Methods.* Newbury Park, Calif.: Sage Publications.

Goldstein, H. 1995. *Multilevel Statistical Models* 2nd ed. London: Edward Arnold.

Jones, Bobby L., Daniel Nagin and Kathryn Roeder. 2001 "A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories," *Sociological Research and Methods*, 29: 374-393.

Gujarati, Damodar N., 2003. *Basic Econometrics*, 4th Ed., McGraw Hill, New York.

McArdle, J. J., and D. Epstein. 1987. "Latent Growth Curves Within Developmental Structural Equation Models." *Child Development.,* 58: 110 - 113.

Meredith, W., and J. Tisak. 1990. "Latent Curve Analysis." *Psychometrika*, 55: 107-122.

Muthén, B. 1989. "Latent Variable Modeling in Heterogeneous Populations." *Psychometrika*, 54: 557-585.

Muthén, L. K. and B. O. Muthén. 2004. *Mplus User's Guide. Third edition*. Los Angeles, CA: Muthén & Muthén.

Nagin, Daniel S. 1999. "Analyzing Developmental Trajectories: A Semi-parametric, Group-based Approach," *Psychological Methods*, 4: 139-177.

Nagin, Daniel S. 2005. *Group-Based Modeling of Development*. Forthcoming.
Nagin, D. S., and K. C. Land. 1993. "Age, Criminal Careers, and Population Heterogeneity: Specification and Estimation of a Nonparametric, Mixed Poisson Model." *Criminology*, 31: 327-362.

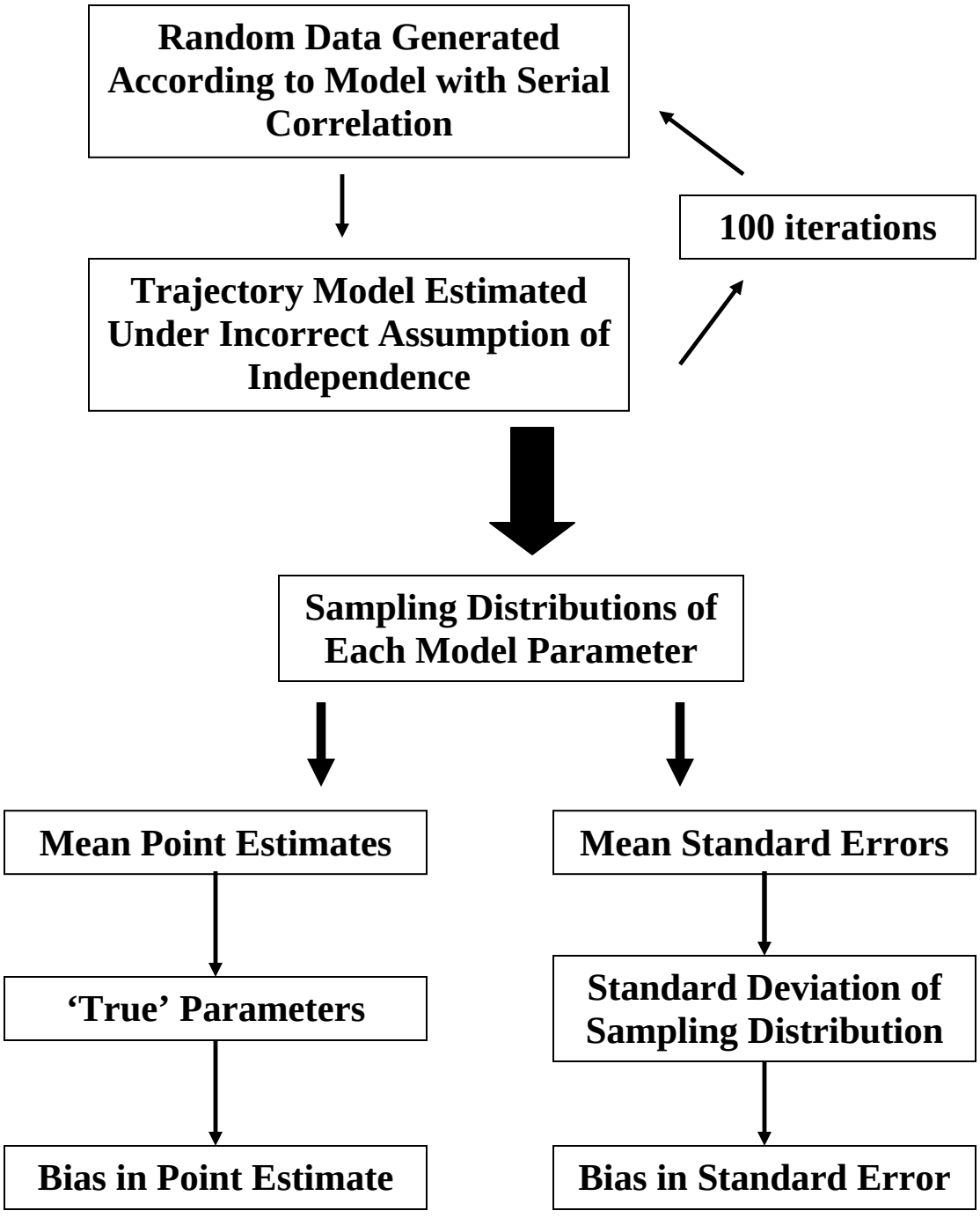
Piquero, Alex R. 2005. "Taking Stock of Developmental Trajectories of Criminal Activity Over the Life Course." Paper Presented at National Institute of Justice Workshop on Longitudinal Studies.

Raudenbush, S. W. 2001. "Comparing-personal Trajectories and Drawing Causal Inferences from Longitudinal Data." *Annual Review of Psychology*, 52: 501-25**.**

Table 1. Summary of Experimental Conditions

Condition	N =	T =	δ =
1	500	4	0.3
2	500	6	0.3
3	1,000	4	0.3
4	1,000	6	0.3
5	500	4	0.5
6	500	6	0.5
7	1,000	4	0.5
8	1,000	6	0.5
9	500	4	0.7
10	500	6	0.7
11	1,000	4	0.7
12	1,000	6	0.7

Figure 1. Overview of Analytic Strategy



Chapter 3: *Accounting for Selection to Understand the Effects of Group Daycare on the Development of Physical Aggression*

Abstract

The developmental course of physical aggression in children may be substantially altered by their receiving some form of non-maternal care, such as group daycare. While attending group daycare is now commonplace, its associated effects on a child's behavior are not well understood and the debate regarding this issue is still highly unresolved. Central to this debate is an important problem with all of the known studies of children's mental health and daycare – the inability to account for the effect of selection in daycare choice. Several factors, including child's age of entry, hours per week spent in daycare and quality of daycare, are choices made by the parents or caregivers that reflect their personal situations and the availability of daycare. Furthermore, some of the factors which may be associated with the need for daycare may themselves also be associated with higher levels of physically aggressive behavior in children. In order to control for such vital selection effects, this analysis uses a propensity score-based approach, which considers the likelihood that a child has entered daycare based on an observed vector of covariates, including variables for parental behavior and risk, familial and parenting conditions, various prenatal factors, and child-specific factors. We then match children who enter daycare to those children who do not, conditional on them having similar likelihood to enter, to produce our estimate of the treatment effect of daycare on later physical aggression levels.

1. Introduction

The developmental course of physical aggression in children may be substantially altered by their receiving some form of non-maternal care, such as group daycare. While attending group daycare is now commonplace (NICHD, 1997b), its associated effects on a child's behavior are not well understood and the debate regarding this issue is still highly unresolved (NICHD, 1993). Some studies suggest that group daycare may increase metal health problems such as physical aggression because it is a hectic environment for children which impedes basic attachments and contributes to the risk for social maladjustment (Belsky, 2001, 2002). Other studies have found that daycare of high quality perhaps reduces the risk of such mental health problems among children facing unfavorable familial situations (NICHD, 1996, 1999a, 2002).

Central to this debate is an important problem with most of the known studies of children's mental health and daycare – the inability to account for the effect of selection in daycare choice. Several factors, including child's age of entry, hours per week spent in daycare and quality of daycare, are choices made by the parents or caregivers that reflect their personal situations and the availability of daycare. Furthermore, some of the factors which may be associated with the need for daycare, such as a living in a single parent home, may themselves also be associated with higher levels of physically aggressive behavior in children. Thus, the comparison of children who do and do not attend day care involves a contrast of two groups that are inherently different in important and pre-existing ways. This problem prevents developmental psychologists from making appropriate causal inference about the effects of daycare on important child outcomes.

This paper directly addresses the issue of selection effects associated with daycare and corrects for them by using a statistical methods know as propensity score matching (Rosenbaum and Rubin, 1983). Our goal is, by way of creating suitable comparisons, is to determine a causal effect of entering daycare at various ages on the development of physical aggression in the first five years of life.

1.1 Background of Daycare and the Mental Health in Children

The National Institute of Child Health and Human Development (NICHD) Child Care Research Network has conducted extensive research concerning non-maternal child care (http://secc.rti.org/). Out of this large, multi-site, longitudinal study, there have been many studies examining the effects of non-maternal care on mental health problems in children. Much of the research on daycare and associated outcomes has to do with attachment (Belsky and Fearon, 2002, 2002a; Dallaire & Weinraub, 2005; McCartney et al, 2004; NICHD 1997a, 2001, 2006) or anxiety and emotional problems and social adjustment (NICHD, 1998, 2001, 2003, 2003a). However, there is some evidence examining daycare and the development of physical aggression (NICHD, 2004; Campbell et al, 2006). Children who spent more time in a non-maternal care environment were found less likely to initiate conflict when playing with other children (O'Brien et al, 1999). However, to the best of our knowledge, there is no definitive work on the causal effects of entering daycare on the development of physical aggression in children.

1.2 Selection Issues in Daycare Analysis

The NICHD Child Care Research Network provides a comprehensive overview of the factors associated with early entry into daycare (NICHD, 1997). Notably, this study concludes that economic factors, such as lower family income and maternal employment, were most consistently associated with early entry into non-maternal care in children, along with lower maternal educational attainment and number of children in the family. This study also suggests that some maternal personality factors, such as extraversion and agreeableness, were critical in the timing of entry decision. Given that these external factors materially affect the likelihood that a child enters into non-maternal care, we can infer that there could be drastic selection biases which arise when trying to quantify the treatment effects of daycare on outcomes.

While the bulk of previous research suggests there is an association between attending quality daycare at a young age and lower levels of physical aggression, most research concedes that understanding the selection factors associated with NMC must be understood before any attempts can be made to quantify its effects on child development (NICHD, 1997). However, several studies do try to address these selection effects for the purpose of casual inference on different child outcomes. Some associations between childcare quality and several positive outcomes in children at age 4 ½ were found remain significant even after controlling for several familial and child-specific factors (NICHD, 2003c), yet this same study failed to find a dose-response relationship between the quality and outcomes. Another study (NICHD, 1999) uses regression-based adjustments to attempt to control for selection when evaluating the effects of child care on maternal sensitivity and child engagement. Borge et al (2004) create a family risk index using occupational level, maternal education, and family functioning to test whether any

difference in physical aggression by childcare status might reflect social selection rather than social causation. This paper differentiates itself from these previous studies, by proposing the use of propensity score matching, an alternative statistical method to the previous model-based attempts, in order to specifically account for selections and evaluate causal effects of daycare on an outcome. In another context, Dehejia and Wahba (1999) show the profound benefits of propensity score methods, as they are able to nearly replicate prior experimental results of the treatment effect of labor-training programs on future wages of workers using non-experimental control data. As the assignment to treatment (i.e., entering daycare) is in this care non-random, dependant on child-specific factors, and thus, non-experimental, the benefits of using such estimators which can potentially mimic experimental results are invaluable.

1.3 The Development of Physical Aggression

Before attempting to address selection biases in estimating a causal effect of daycare on physical aggression, it is important to understand factors which may constitute a risk for higher levels of physical aggression in children. Fortunately, much is known about the development of physical aggression based on prior research. Developmental research has shown that the genesis of physically aggressive behavior begins early in life (Keenan and Shaw, 1994; Tremblay, 2004; Tremblay et al, 2004). The onset of physical aggression occurs in most children by the end of age two and, in most children, declines with age (Tremblay, 1999, 2004). Most physical aggression is inhibited by the age of school entry (Cairns et al 1989; Broidy et al, 2003). Multiple studies (Cote et al, 2006; Nagin and Tremblay, 2006, Baillargeon, 2007) conclude that the development of physical

aggression is dependant on sex, as boys tend to become more aggressive than girls by the time of school entry. Low socio-economic status (SES) is also an important risk factor (Keenan and Shaw, 1994; Cote et al, 2006). Various parenting and familial factors could potentially be important in the development of physical aggression in their children. Keenen and Shaw (1994) show some parenting factors, including rearing practices and parental psychopathology, are correlated with childhood behavior problems. Del Vecchio and O'Leary (2006) also suggest differing parenting styles could lead to different levels of aggression in young children. Cote et al (2006) report that hostile and ineffective self-reported parenting strategies were risk factors for membership in a higher physical aggression trajectory group. Stormshak et al (2000) find that distinct parenting practices may be associated with type and profile of a child's disruptive behavior problems. Maternal factors such as depression may also be critical in the development of physical aggression (Jaffee et al, 2001; Cote et al, 2006). Nagin and Tremblay (2001) show that maternal characteristics, notably low educational attainment and teenage onset of childbearing, distinguish those who persist in high levels of physical aggression among kindergarten-aged boys who display high levels of physical aggression.

In summary, there are multiple factors which may be simultaneously affecting both the development of physical aggression in children as well as the likelihood that that child enters daycare. Thus, any basic associations relating physical aggression to daycare may, in fact, be due simply to the confounding of these factors, instead of being generated by a causal mechanism. Therefore, any attempt at causal inference with respect to the effects of daycare on physical aggression must properly account for such observable characteristics in order to eliminate the biases associated with them, since, if

these same factors are also capable of predicting who receives non-maternal care in the relevant developmental period, they could be potentially confounding our estimate of the treatment effect. More specifically, in our case, these differences we observe in physical aggression among children who enter daycare at different stages could be due entirely to selection.

1.4 Overview of Method

In order to control for such vital selection effects, this analysis uses a propensity score-based approach, based on the work of Rosenbaum and Rubin (Rosenbaum, 2002; Rosenbaum and Rubin, 1983). Specifically, we estimate the likelihood that a child has entered daycare based on an observed vector of covariates, which includes variables for maternal and paternal behavior and risk, familial and parenting conditions, various prenatal factors, and child-specific factors. We then match children who enter daycare to those children who do not, conditional on them having similar likelihood to enter, to produce our estimate of the treatment effect of daycare on later physical aggression levels. We examine entry into daycare at each of three ages: before five months, between six and 17 months and between 18 and 60 months. We also consider children in the sample who do not enter daycare prior to 60 months. Our aim is to test for an age of entry effect as well as a duration effect. In addition to its rich set of covariates, this dataset also contains measurements of physical aggression at 17, 30, 42, 54, 60 and 72 months, which are used as both controls and outcomes.

1.5 Content Organization

This chapter is organized as follows. **Section 2** introduces the sample used in this analysis, and examines differences, in both the development of physical aggression as well as several important observable child-specific and familial risk factors, between children who begin daycare at different times as well as those who do not enter daycare at all. **Section 3** discusses the use of the propensity score-based matching methodology used in the analysis and highlights the success of the methodology in creating groups suitable for comparison by showing measures of balance on the observable covariates. **Section 4** provides the results and analysis of these results. **Section 5** offers a discussion and conclusions.

2. Physical Aggression and Daycare

2.1 Data

A total of 2008 children were followed annually from five to 72 months of age to assess the developmental course of physical aggression. Mothers were interviewed 6 times, when their child was 5, 17, 30, 42, 54 and 60 months old. The infants were from a population sample (N=2,223) representing 5 month olds in the Canadian province of Québec in the fall of 1997 and the spring of 1998. In addition to tracking physical aggression over time, this sample also provides a rich set of baseline characteristics, measured at 5 months, including data on temperament, maternal and paternal behavior prior to birth, and familial characteristics. It is the richness of these pre-intervention covariates, measured prior to a child's entry into daycare, which allows us to use the propensity score methodology in order to form suitable matches. **Table 1** presents the demographic characteristics of the sample.

2.2 Measuring Physical Aggression

To assess physical aggression, we selected items from rating scales that were used in a previous study with this data (Tremblay et al, 2004), which suggests that a three-item scale is sufficient to reliably assess physical aggression in children. Mothers were asked to rate their child on a frequency scale indicating whether the child never (0), sometimes (1), or often (2) exhibits physical aggression. The three items were 1) hits, bites, kicks, 2) fights, and 3) bullies others. Thus, scores on this three-item scale may range from 0 to 6. Assessments were made for each child (with some attrition) at 17, 30, 45, 54, 60 and 72 months.

2.3 Physical Aggression and Daycare

Figure 1 shows the trends of mean physical aggression for four different groups – children who entered daycare before five months, children who entered daycare between six and 17 months, children who entered after 18 months and children who do not enter daycare before five years of age. At each of the measurement periods, these trends would seem to indicate that, not only is entry into daycare associated with lower levels of physical aggression, but they also suggest that earlier exposure to daycare is, in fact, better in the long term development of physical aggression.

Regarding those children who never enter daycare in the first five years, their mean PA level is higher than all other groups at 30, 42, 54 and 60 months. Furthermore, the difference between the mean PA levels of this group and the others actually becomes more pronounced over time. While mean PA peaks for each of the groups at 42 months and begins to decline thereafter, the rate of decline is much slower in the children who do

not attend daycare compared to the other groups. There are statistically significant differences in mean PA levels at ages 54 months and 60 months between children who do not attend daycare and those who begin both prior to five months and those who begin between six and 17 months (all *p*-values < .01). The mean PA levels at 54 and 60 months for those not attending daycare are also higher compared to those who begin after 18 months. The differences are marginally significant at 54 months and highly significant (at $\alpha = .05$) at 60 months. While the mean PA scores for those children who do not attend daycare are higher than those in children who do attend daycare in each period after 30 months, the mean PA level at 17 months for those who never attend is not much different than the means of the other groups and, in fact, is slightly lower compared to the mean of the children who begin after 17 months. Of course, in terms of treatment status, there is no difference at the 17 month measurement between children who begin daycare between 18 and 60 months and those children who do not enter daycare at all. Since the children who do begin after 18 months tend to have lower PA throughout the rest of the developmental period, this further suggests daycare may have a reducing effect of children's aggression levels.

The mean aggression levels for children who begin daycare before five months are nearly identical to the PA levels of those who begin between six and 17 months. At age 54 and 60 months the mean PA level of those children beginning between six and 17 months is trivially higher than those of the children beginning prior to five months. However, there is a pronounced difference at the 30 month period, where the mean PA score of those children beginning prior to five months is actually *higher* than that of the children beginning between six and 17 months. The difference is also marginally

statistically significant (*p*-value = .062). This result could indicate that, despite their similarities in outcomes nearer to the end of the developmental period, there may be important preexisting differences between these two groups of children prior to their respective entries into daycare. The similarities of outcomes at later ages, however, could indicate that at such early ages, there is no age of entry effect for beginning daycare.

Those children who enter daycare after 18 months (but prior to 60 months) tend to have higher mean levels of PA than those children who enter into daycare between six and 17 months in each of the observation periods. These differences are significant at 17, 30 (*p*-value < .05) and 54 months (*p*-value < .01). These children also tend to have higher mean PA than children who begin prior to 5 months, save for the 30 month measurement period when the mean for the children who begin prior to 5 months is trivially higher. The mean difference between these two groups is statistically significant (*p*-value < .01) at 54 months, but not at 60 months.

On its face, the above comparisons would suggest that there may be causal mechanisms of daycare which tend to reduce the levels of PA in children. Furthermore, it appears that *when* a child enters into daycare is as important to the development of physical aggression as *if* the child does. The fact that the decline of PA in each of the groups is more rapid in children who begin daycare earlier could be reflective of a timing effect, or, more specifically, an age of entry effect, signaling that entry into daycare may be more advantageous at various stages of development than others. In summary, these comparison trends would seem to indicate that, not only is entry into daycare associated with lower levels of physical aggression, but they also suggest that earlier exposure to daycare is, in fact, better in the long term development of physical aggression.

2.4 Confounding Factors in Daycare Entry and Physical Aggression

As discussed above, the main problem in comparing the physical aggression outcomes across children who do or do not enter daycare at different times centers on the idea that there are other, confounding factors which may affect each child's likelihood of entry. Furthermore, these factors may not only be associated with the daycare entry decision, but also with the child's physical aggression levels. Thus, the comparison of children who do and do not attend day care involves a contrast of two groups that are dissimilar in important ways. In formal statistical parlance, because there is no random assignment to the treatment condition of daycare, the treatment and control groups are not comparable. In order to try to establish evidence of selection effects which potentially contaminate the basic comparisons in Section 2.3 of the effects of daycare entry on the development of physical aggression, and in turn to mitigate these effects for the purposes of causal inference, we consider several blocks of predictor variables which may be associated with the decision to enter into daycare and/or the development of physical aggression.

Family Income

One important variable which may govern the issue of selection is family income (NICHD, 1997). It may be the case that children from lower income families demand more daycare, since the lack of income forces both parents into the labor force. Similarly, it could also be that higher income families are able to defer sending both parents to work, and hence avoid or at least delay sending a child for non-maternal care. Conversely, the demand for daycare could be higher among the higher-income families, as they are the ones most likely to be able to afford the expense of daycare. As

mentioned, there is also evidence that lower socioeconomic status (SES) could be a risk factor for higher physical aggression. We observe family income at each of the six observation periods as an ordinal value ranging from 1 (annual family income < $10,000 Canadian) to 9 (annual family income > $80,000 Canadian).

Maternal Risks

Nagin and Tremblay (2001) suggest some maternal risk factors may be associated with higher levels of physical aggression in boys. Some of the same factors could also be related to the decision to send children to daycare. Mothers who are depressed, for instance, may choose to send their children to daycare earlier than those mothers who are not depressed. Mothers who are more highly educated may be more likely to enter the labor force sooner after giving birth, and in turn, be more likely to have a child entering daycare at a younger age. We use an indicator variable for maternal depression and an ordinal variable measuring educational attainment of the mother. In addition to these two factors, we consider, in the form of indicator variables, the additional risks of whether a mother has a history of antisocial behavior, and if the mother began child-bearing as a teenager.

Familial Factors

A child's home situation may be highly associated with the likelihood that the child enters daycare. Less stable family environments may play a role in the need for daycare, as well as the development of mental health problems in children (cite). We consider an overall measure of family dysfunction. Also, as noted, certain parenting styles may be

more indicative of the need (or lack thereof) for some form of non maternal care. We consider six self-reported maternal parenting measures – self-efficacy, perceived impact, coerciveness, affection, overprotection, and parental perception – which may be influential of the decision to send a child into a non maternal care environment.

Other Factors

Though there is not necessarily evidence to suggest that they affect the development of mental health problems, we also consider some paternal risks, as there is evidence that personality traits of the father may affect non-maternal care decisions (NICHD, 2000). We include indicator variables for paternal depression and antisocial history. Also, we examine some prenatal factors, such as alcohol and cigarette use by the mother during pregnancy, and premature birth, all in the form of indicator variables. The inclusion of these factors in the analysis is for general completeness, as opposed for a more theoretical justification.

2.5 Differences in Predictors by Daycare Entry Status

Table 2 reports mean income levels over time for families conditional on daycare entry timing. For children who enter daycare before five months, there is a statistically significant difference in mean family income compared to children who enter after 18 months and children who do not enter. Specifically, the children entering daycare before five months and between six and 17 months come from wealthier families on average, suggesting that family income may play an important role in the parent's decision to start their child into daycare.

Table 3 shows percentages of four key maternal predictors, conditional on timing of daycare entry – depression, antisocial history, childbearing as a teenager, and education. The percentage of children with depressed mothers in essentially the same for children who begin daycare before five months compared to the children who begin between six and 17 months, yet it is higher in both children who begin after 18 months and those who do not enter daycare at all. Similarly, levels of education are slightly higher, yet similar, in the before five months and between six and 17 month groups, yet are much lower in the other two groups. The percentage of mothers who began childbearing as a teenager in the six to 17 month group (14%) more than doubles in the after 18 month group (30%) and the non-entry group (29%). It is also higher in the before five month group (21%), yet not as high as the later and non-entry groups. There are also differences, although slight, between the four groups in terms of mother's antisocial histories.

As far as familial factors, there tend to be lower levels of family dysfunction, on average, in families of the two groups where children are sent to daycare before 17 months, compared to children entering after 18 months and those not entering at all. There are also important differences in parenting styles among the fours groups. The later the child enters daycare, the more overprotective the mother tends to be on average (the mothers who do not send their child to daycare at all have the highest reported levels of overprotection, on average). There are also differences in mother's perceived parental impact on the children, maternal perception, perceived affection and self-efficacy. There do not seem to be many dramatic differences between the paternal factors, or prenatal conditions (i.e., mother's drinking and smoking).

In summary, it appears that those children who enter daycare at different times, as well as those children who do not enter daycare at all, are substantively different in important ways. These differences may not only be contributing to a child's daycare entry status, but also to the developmental course of physical aggression in the children. In other words, given the preexisting differences between these groups, a comparison of mean physical aggression between the groups is not tantamount to uncovering a causal effect of daycare on physical aggression, as some groups may be more risk for higher levels of physical aggression than others in the first place. Thus, we proceed with our matching strategy in order to create groups more appropriate for comparison.

3. Method

3.1 Propensity Scores

The propensity score represents the probability that an individual receives some treatment conditional on a vector of observed covariates (Rosenbaum, 2002). Rosenbaum and Rubin (1983) show that, conditional on two individuals, one treated and one control, having an identical propensity score, the difference in treatment status becomes independent of all observable characteristics, **x**. The idea is to estimate a propensity score for each individual, and in turn use this estimate as a method for creating balance on key covariates that may be confounding the treatment effect estimate. We then match subjects who have very similar estimated propensity scores, but differ on treatment status, in order to measure the treatment effect. This process serves to removes the systematic part of the treatment allocation, leaving whatever differences remain as hopefully close to random. Thus, after balancing on the estimated propensity score, we

will hopefully have constructed two groups, one treated, one control, which are now appropriate for comparison, since the only difference between subjects who are treated and those who are not is simply random assignment of treatment due to chance. One caution is since we only attempt to create balance on observable characteristics - in the language of Rosenbaum, we eliminate overt biases – there still may be unobservable factors which contribute to the treatment allocation process and thus bias our results, even after our attempts at correction. These so-called hidden biases are a potential problem to our conclusions, and we address these specifically below.

In our case, we consider entry into daycare at each age to be different 'treatments'. We then use blocks of covariates – family income, maternal, familial and others – to estimate a propensity score that each child will receive treatment (i.e., begin daycare) at each of the age designations. We then match children with similar propensities for entering daycare at a certain age, based on whether they entered at the time or not, in order to create a treatment and control group which differ on only a hopefully random assignment of daycare entry. These two groups, then, allow us to estimate a true treatment effect of daycare on physical aggression. The large number of covariates is useful here since the case for balance between treated and non-treated groups in such observational data is most convincing when it is achieved over a broad range of covariates. Similarly, the balance argument is predicated on these covariates being strongly related to the treatment. To strengthen this argument, in addition to the covariates discussed above, we are also able to control for lagged values of the outcome variable, physical aggression. Balancing on prior physical aggression is very important, since, as described above, the development of physical aggression begins very early in

life, and those who still display high levels at 54 and 60 months, almost certainly displayed some physical aggression early on, meaning matching on the prior level is critical.

3.2 Estimating the Propensity Score

The propensity score is the conditional probability of receiving some treatment given some observed covariates **x**. In order to estimate $e(\mathbf{x})$, the propensity score for each individual, we simply estimate a binary logit model, using the treatment status as the dependant variable as a function of the observed covariates **x**. In our case, we use each of aforementioned predictors – pre-intervention income, sex, maternal, familial, paternal, and prenatal risks – as explanatory terms in the model specification, as they each could reasonably be associated with the likelihood a child is entered into daycare at a particular age. The predicted probability from this model for each individual, $\hat{e}(\mathbf{x})$, is thus that individual's estimated propensity score. All of the estimates of the treatment effect are then driven by this estimated propensity score.

3.3 Matching/Stratification Using the Propensity Score

After estimating the propensity score for each individual, we then check that our specification for the propensity score model is valid, as we check for balance among each of the covariates. In other words, we can check that the distribution of the observed covariates is the same for individuals with the same propensity score. Based on this estimated propensity score, we then create a matched pair consisting of a treated individual and a control individual who have the same (or very similar) propensity scores

in order to create a counterfactual. If our estimated propensity score model is correct, then two individuals with the same propensity score who differ only by their observable treatment status (i.e., one treated and one control), should theoretically differ by a then random assignment of treatment status. That is, after conditioning on propensity score, the treatment status and covariates **x** are independent. This counterfactual in effect answers to question of what the outcome of treated individual would be had they *not* received the treatment, which, of course, is unobservable to the researcher.

3.4 Matching Strategy

As mentioned, we consider four unique treatment conditions: 1) entry into daycare before five months, 2) entry between six and 17 months, 3) entry between 18 and 60 months, and 4) no daycare entry prior to 60 months. The outcomes of interest are physical aggression at 54 and 60 months. **Figure 2** provides an overview of the strategy. We examine each of the six pair-wise combinations of treatment groups in separate matching analyses. The treatment effect, denoted τ_k, represents the effect on the outcome by entering daycare in the time period on the top row compared to entering daycare in the time period of the column. For instance, τ_1 denotes the effect on later physical aggression by entering daycare before five months as opposed to entering daycare between six and 17 months. As noted, our aim is to test for both a duration effect as well as an age of entry effect. We now explicitly outline these two effects.

Testing the hypotheses H_o: $\tau_1 = \tau_2 = \tau_3$ or H_o: $\tau_4 = \tau_5$ would constitute a test of a *duration* effect, that is, longer exposure to the daycare treatment would be different than a more abbreviated exposure. Specifically, the alternatives H_1: $\tau_1 < \tau_2 < \tau_3$ or H_1: $\tau_4 < \tau_5$

(assuming that the treatment effect is negative, or, in other words, entry into daycare causes physical aggression at 54 and 60 months to decline) would mean that being in daycare for a longer duration would have a reduce later physical aggression in such children to a higher degree, and consequently be more beneficial.

Alternatively, testing the hypotheses H_o: $\tau_1 = \tau_4 = \tau_6$ or H_o: $\tau_2 = \tau_5$ would constitute a test of an *age of entry* effect, that is, if the age of first exposure to the daycare treatment is important. Specifically, the alternatives H_1: $\tau_1 < \tau_4 < \tau_6$ or H_1: $\tau_2 < \tau_5$ (again assuming all treatment effects to be negative) would mean that, for those children who have been in daycare for similar lengths of time, entering daycare at an earlier age would be more beneficial, as it would reduce later physical aggression to a higher degree.

Finally, and perhaps most importantly, a series of simple tests that each of the individual treatment effects are singularly equal to zero is highly important as well. If any of these effects are, in fact, found to be zero, then this would suggest that, while there may be observable differences in the physical aggression outcomes at later ages among children who enter daycare at different times, these differences are not attributable to the causal effects of daycare, but rather due entirely to selection effects.

4. Results

4.1 Balance Among Different Treatment Condition

As we reiterate, we do not know the true propensity score for any individual; we *estimate* it based on an appropriate specification of the logit model. The result of conditional independence of the treatment and observable characteristics depends on the fact that we know the true propensity score, and therefore, we can achieve balance on the

observable characteristics. In order to ensure that we are, in fact, using the proper propensity score, we must check each of the observed covariates for balance.

4.2 Matching Results

4.3 Sensitivity to Hidden Bias

As noted above, even if we can correct for all of the observable bias due to **x**, that is, we can properly balance on all observable covariates, the propensity score is still powerless to correct for any additional hidden biases that may exist. Therefore, in this section, we conduct a sensitivity analysis to gage, if there is such a hidden bias, how it might qualitatively affect our results. Note that we do not attempt to rule out the existence of hidden bias. Simply, we aim to answer if there was an unobservable characteristic that was potentially affecting the qualitative results of our analysis, what would such a characteristic look like, and thus, is it plausible that such a factor exists

5. Discussion

References

Belsky, J. (2001).”Emanuel Miller Lecture Developmental Risks (Still) Associated with Early Child Care.” *Journal of Child Psychology and Psychiatry*, 42 (7), 45-59.

Belsky, J. (2006). “Early child care and early child development: Major findings from the NICHD Study of Early Child Care.” *European Journal of Developmental Psychology*, 3, 95-110.

Belsky, J. and Fearon, R.M.P.. (2002). “Infant-Mother Attachment Security, Contextual Risk and Early Development: A moderational Analysis.” *Development and Psychopathology*, 14, 293-310.

Belsky, J. and Fearon, R.M.P. (2002a). “Early Attachment Security, Subsequent Maternal Sensitivity, and Later Child Development: Does Continuity in Development Depend Upon Continuity of Caregiving?” *Attachment and Human Development*, 3, 361-387.

Borge, Anne I.H., Michael Rutter, Sylvana Côté and Richard E. Tremblay (2004). “Early childcare and physical aggression: differentiating social selection and social causation.” *Journal of Child Psychology and Psychiatry*, V.45, Issue 2: 367-376.

Baillargeon, Raymond H.; Zoccolillo, Mark; Keenan, Kate; Côté, Sylvana; Pérusse, Daniel; Wu, Hong-Xing; Boivin, Michel; Tremblay, Richard E. (2007). “Gender Differences in Physical Aggression: A Prospective Population-Based Survey of Children Before and After 2 Years of Age.” *Developmental Psychology*. Jan. Vol 43,(1), 13-26.

Broidy, L. M., Nagin, D. S., Tremblay, R. E., Brame, B., Dodge, K., Fergusson, D., et al. (2003). “Developmental trajectories of childhood disruptive behaviors and adolescent delinquency: A six site, cross national study.” *Developmental Psychology*, 39, 222–245.

Cairns, R. B., Cairns, B. D., Neckerman, H. J., Ferguson, L. L., & Gariépy, J. L. (1989). “Growth and aggression: 1. Childhood to early adolescence.” *Developmental Psychology*, 25, 320–330.

Campbell, Spieker, Burchinal, Poe, and NICHD ECCRN. (2006). “Trajectories of aggression from toddlerhood to age 9 predict academic and social functioning through age 12.” *Journal of Child Psychology and Psychiatry*, 47 (8), 791-800.

Côté, S., Vaillancourt, T., LeBlanc, J. C., Nagin, D. S., & Tremblay, R. E. (2006). “The development of physical aggression from toddlerhood to pre-adolescence: A nation-wide longitudinal study of Canadian children.” *Journal of Abnormal Child Psychology*, 34, 68–82.

Dallaire, D.H. & Weinraub, M.. (December 2005). “Predicting children's separation anxiety at age 6: The contributions of infant-mother attachment security, maternal sensitivity, and maternal separation anxiety.” *Attachment & Human Development*, 7 (4), 393-408.

Dehejia, Rajeev and Sadek Wahba, (1999). “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs.” *Journal of the American Statistical Association*, Vol. 94, No. 448, 1053-1062.

Del Vecchio, Tamara and Susan G. O’Leary (2006). “Antecedents of Toddler Aggression: Dysfunctional Parenting in Mother-Toddler Dyads.” *Journal of Clinical Child and Adolescent Psychology*, Vol. 35, No. 2, 194-202.

Jaffee, S., Caspi, A., Moffitt, T. E., Belsky, J., & Silva, P. (2001). “Why are children born to teen mothers at risk for adverse outcomes in young adulthood? Results from a 20-year longitudinal study.” *Development and Psychopathology*, 13, 377–397.

Keenan, K., & Shaw, D. S. (1994). “The development of aggression in toddlers: A study of low-income families.” *Journal of Abnormal Child Psychology*, 22, 53–77.

McCartney, K., Owen, M.T., Booth, C., Vandell, D.L., & Clarke-Stewart, K.A. (2004). “Testing a maternal attachment model of behavior problems in early childhood.” *Journal of Child Psychology and Psychiatry*, 45, 765-778.

Nagin, Daniel S. and Tremblay, R. E. (1999). “Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and non violent juvenile delinquency.” *Child Development*, 70, 1181–1196.

Nagin, Daniel S. and Richard E. Tremblay, (2001). “Parental and Early Childhood Predictors of Persistent Physical Aggression in Boys From Kindergarten to High School.” *Archives of General Psychiatry*; 58:389-394.

NICHD Early Child Care Research Network. (1993).”Child-care debate: Transformed or distorted?” *American Psychologist*, 48, 692-693.

NICHD Early Child Care Research Network. (1996). “Characteristics of infant child care: Factors contributing to positive caregiving.” Early *Childhood Research Quarterly*, 11, 269-306.

NICHD Early Child Care Research Network. (1997). “Familial factors associated with the characteristics of nonmaternal care for infants.” *Journal of Marriage and the Family*, *59*, 389-408.

NICHD Early Child Care Research Network. (1997a). “The effects of infant child care on infant-mother attachment security: Results of the NICHD Study of Early Child Care.” *Child Development*, 68, 860-879.

NICHD Early Child Care Research Network. (1997b). “Child care in the first year of life.” *Merrill-Palmer Quarterly*, 43, 340-360.

NICHD Early Child Care Research Network. (1998). “Early child care and self-control, compliance and problem behavior at twenty-four and thirty-six months.” *Child Development*, 69, 1145-1170.

NICHD Early Child Care Research Network. (1999). “Child care and mother-child interaction in the first three years of life.” *Developmental Psychology*, 35, 1399-1413.

NICHD Early Child Care Research Network. (1999a). “Child outcomes when child care center classes meet recommended standards for quality.” *American Journal of Public Health*, 89, 1072-1077.

NICHD Early Child Care Research Network. (2000). “Factors associated with fathers' caregiving activities and sensitivity with young children.” *Journal of Family Psychology*, 14. 200-219.

NICHD Early Child Care Research Network. (2001). “Child care and children's peer interaction at 24 and 36 months: The NICHD Study of Early Child Care.” *Child Development*, 72, 1478-1500.

NICHD Early Child Care Research Network. (2001). “Child care and family predictors of preschool attachment and stability from infancy.” *Developmental Psychology*, 37, 847-862.

NICHD Early Child Care Research Network. (2002). “Child-care structure>process>outcome: Direct and indirect effects of child-care quality on young children's development.” *Psychological Science*, 13, 199-206.

NICHD Early Child Care Research Network. (2003). “Social functioning in first grade: Associations with earlier home and child care predictors and with current classroom experiences.” *Child Development*, 74, 1639-1662.

NICHD Early Child Care Research Network. (2003a). “Does amount of time spent in child care predict socioemotional adjustment during the transition to kindergarten?” *Child Development*, 74, 976-1005.

NICHD Early Child Care Research Network. (2003c). “Does quality of child care affect child outcomes at age 4 ½?” *Developmental Psychology*, 39, 451-469.

NICHD Early Child Care Research Network. (2004). “Trajectories of physical aggression from toddlerhood to middle childhood: Predictors, correlates, and outcomes.” *SRCD Monographs*, 69 (4, 278), 1-146.

NICHD Early Child Care Research Network. (2006). “Infant-mother attachment: Risk and protection in relation to changing maternal caregiving quality over time.” *Developmental Psychology*, 42 (1), 38-58

O'Brien, Marion, Carolyn Roy, Anne Jacobs, Mery Macaluso, Vicki Peyton (1999). "Conflict in the Dyadic Play of 3-Year-Old Children." *Early Education and Development*, Vol. 10, No. 3, 289-313.

Tremblay, R. E. (2004). "The development of human physical aggression: How important is early childhood?" In L. A. & Leavitt, D. M. B. Hall (Eds.), *Social and moral development: Emerging evidence on the toddler years* (pp. 221–238). New Brunswick, NJ: Johnson and Johnson Pediatric Institute.

Tremblay, R. E., and Nagin, Daniel S. (2005). "The developmental origins of physical aggression in humans." In R. E. Tremblay, W. H. Hartup, & J. Archer (Eds.), *Developmental origins of aggression.* (pp. 83–106) New York: Guilford Press.

Tremblay, R. E., Nagin, D. S., Séguin, J. R., Zoccolillo, M., Zelazo, P. D., Boivin, M., et al. (2004). "Physical aggression during early childhood: Trajectories and predictors." *Pediatrics*, 114, 43–50.

Rosenbaum, P. R. (2002) *Observational Studies* (2nd ed.). New York: Springer-Verlag.

Rosenbaum, P. and D. Rubin. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70: 41-55.

Stormshak, EA, Bierman, KL, McMahon, RJ, Lengua, LJ (2000) "Parenting practices and child disruptive behaviour problems in early elementary school." *Journal of Clinical Child Psychology*, 29: 17-29.

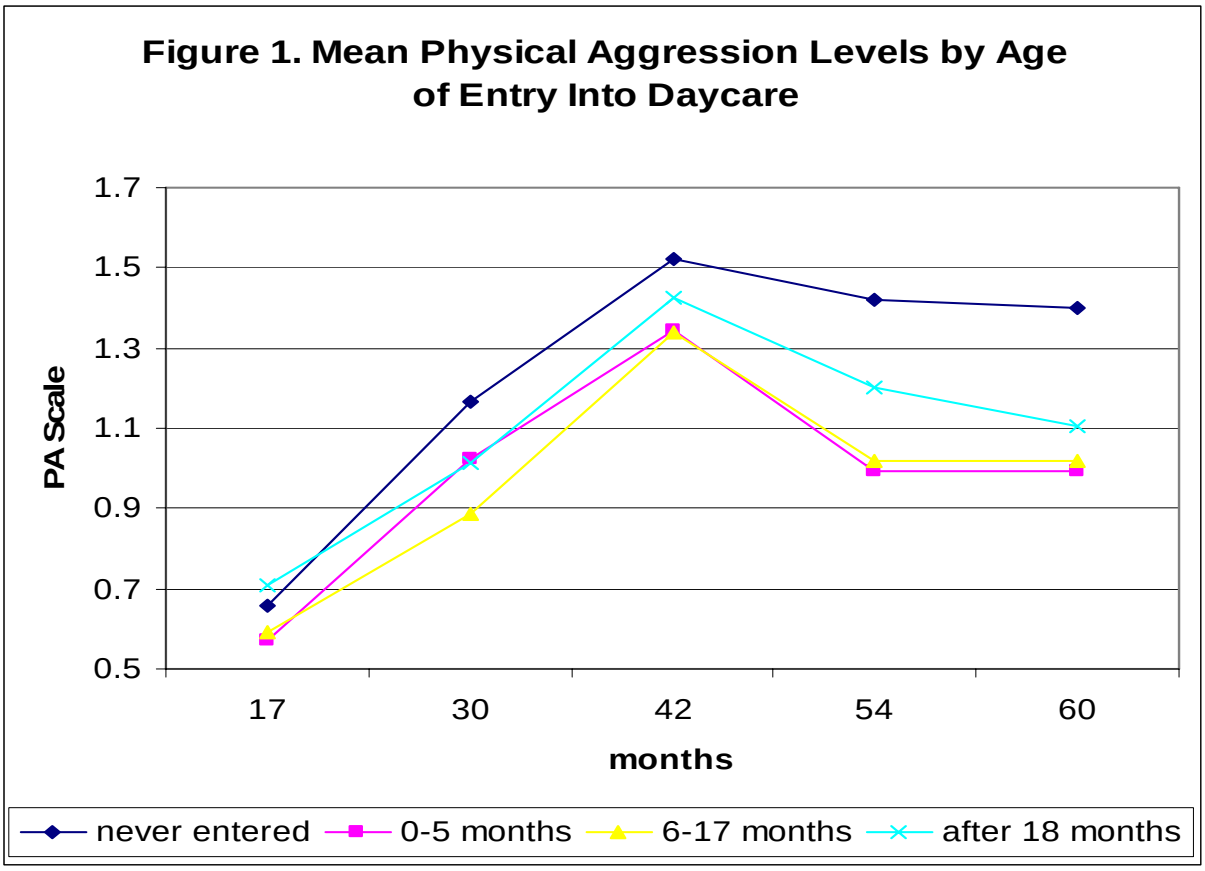


Figure 1. Mean Physical Aggression Levels by Age of Entry Into Daycare
PA Scale
1.7
1.5
1.3
1.1
0.9
0.7
0.5
17
30
42
54
60
months
never entered
0-5 months
6-17 months
after 18 months

Figure 2: Overview of Matching Strategy

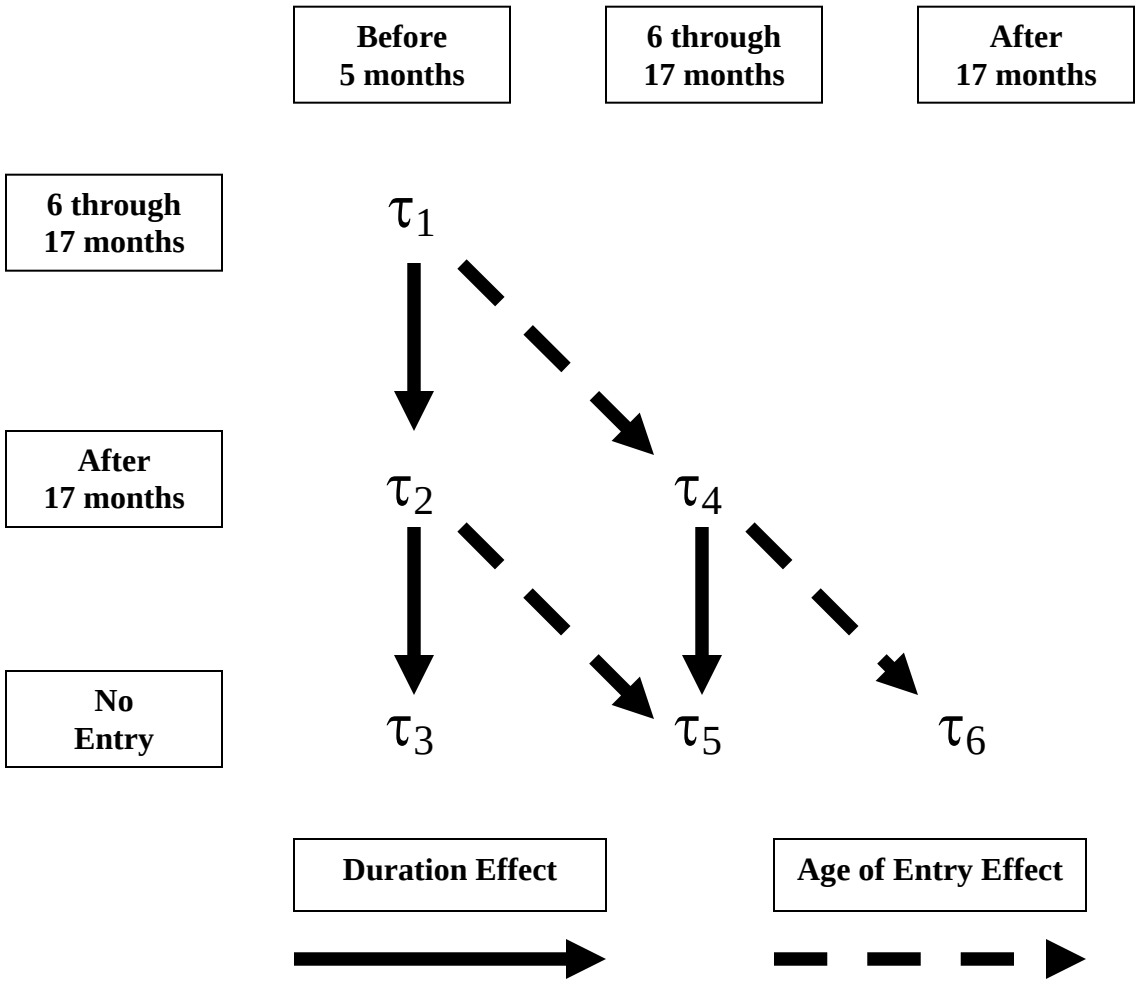


Table 1. Demographic Characteristics

	mean	s.d.
Male	0.51	0.50
First child?	0.58	0.49
age of mother at childbearing	29.32	5.22
no. of siblings	0.83	0.91
2 biological parents present?	0.92	0.27

Table 2. Mean Family Income Over Time by Daycare Entry

	never entered		0-5 months		6-17 months		after 18 months	
	mean	**s.e.**	**mean**	**s.e.**	**mean**	**s.e.**	**mean**	**s.e.**
5 months	4.84	0.22	6.04	0.12	6.58	0.07	4.97	0.09
17 months	5.16	0.21	6.51	0.12	6.77	0.06	5.15	0.09
30 months	5.41	0.21	6.77	0.12	7.13	0.06	5.34	0.09
42 months	5.77	0.21	7.05	0.12	7.30	0.06	5.67	0.09
54 months	5.75	0.21	7.25	0.11	7.37	0.06	5.89	0.09
60 months	5.81	0.21	7.19	0.12	7.42	0.06	6.10	0.09

Table 3. Means of Predictors by Daycare Entry

	never entered			0-5 months			6-17 months			after 18 months		
	n	**mean**	**sd**	**n**	**mean**	**sd**	**n**	**mean**	**sd**	**n**	**mean**	**sd**
self-efficacy	95	9.12	0.97	271	8.98	0.88	936	8.94	0.90	602	9.03	1.01
maternal impact	98	8.10	2.23	271	8.67	1.62	940	8.61	1.65	605	8.04	2.09
coerciveness	95	0.99	1.17	271	1.02	1.18	936	1.05	1.21	602	1.09	1.23
maternal affection	98	9.54	1.11	273	9.74	0.49	938	9.71	0.53	607	9.68	0.80
overprotection	97	5.31	2.08	273	4.08	1.94	938	4.36	2.12	606	5.28	2.30
maternal perception	98	7.82	1.77	271	8.24	1.47	935	7.99	1.59	606	7.78	1.95
family functioning	101	1.80	1.43	277	1.53	1.41	955	1.61	1.42	614	1.84	1.50
maternal depression	101	0.28	0.45	275	0.19	0.39	932	0.20	0.40	606	0.26	0.44
antisocial mother	96	0.14	0.34	264	0.20	0.40	918	0.19	0.39	598	0.23	0.42
teenage mother	96	0.29	0.46	266	0.21	0.41	915	0.14	0.35	589	0.30	0.46
maternal education	101	3.24	2.09	272	4.63	2.11	940	4.81	2.04	615	3.49	2.10
antisocial father	89	0.30	0.46	239	0.34	0.48	828	0.33	0.47	507	0.36	0.48
paternal depression	89	0.26	0.44	239	0.27	0.45	827	0.25	0.43	507	0.26	0.44
prenatal smoking	101	0.29	0.45	272	0.27	0.44	939	0.21	0.41	610	0.30	0.46
prenatal drinking	101	0.34	0.47	272	0.41	0.49	939	0.41	0.49	610	0.30	0.46
premature birth	102	0.06	0.24	274	0.05	0.21	941	0.04	0.19	615	0.06	0.23