

# Group-based Trajectory Modeling Extended to Account for Nonrandom Participant Attrition

Sociological Methods & Research

40(2) 367–390

© The Author(s) 2011

Reprints and permission:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0049124111400041

<http://smr.sagepub.com>



**Amelia M. Haviland<sup>1</sup>, Bobby L. Jones<sup>2</sup>,  
and Daniel S. Nagin<sup>3</sup>**

## Abstract

This article reports on an extension of group-based trajectory modeling to address nonrandom participant attrition or truncation due to death that varies across trajectory groups. The effects of the model extension are explored in both simulated and real data. The analyses of simulated data establish that estimates of trajectory group size as measured by group membership probabilities can be badly biased by differential attrition rates across groups if the groups are initially not well separated. Differential attrition rates also imply that group sizes will change over time, which in turn has important implications for using the model parameter estimates to make population-level projections. Analyses of longitudinal data on disability levels in a sample of very elderly individuals support both of these conclusions.

## Keywords

attrition bias, longitudinal data, trajectory groups, finite mixture models

---

<sup>1</sup>RAND Corporation, Pittsburgh, PA, USA

<sup>2</sup>University of Pittsburgh, Pittsburgh, PA, USA

<sup>3</sup>Carnegie Mellon University, Pittsburgh, PA, USA

## Corresponding Author:

Daniel S. Nagin, Heinz College Carnegie Mellon University, Pittsburgh, PA, 15213, USA

Email: [dn03@andrew.cmu.edu](mailto:dn03@andrew.cmu.edu)

## Introduction

One of the greatest challenges of collecting longitudinal data is minimizing participant attrition. In some cases attrition is unavoidable because participants die. Death is a common source of attrition in medical studies of individuals with serious disease, disabilities, or advancing age. Death may also be a nontrivial source of attrition in long-term studies of individuals with high risk for antisocial behavior. Eggleston, Laub, and Sampson (2004) analyze data based on the classic Glueck and Glueck (1950) longitudinal study of 500 males who had been incarcerated as juveniles. Based on data assembled by Laub and Sampson (2003) tracking these individuals until age 70, Eggleston et al. (2004) report that nearly 50 percent had died by age 70. In a more contemporary sample followed over a shorter time frame, 39 of the initial 1,517 participants of the Pittsburgh Youth Study died prior to age 30, mostly due to homicide (Loeber and Farrington forthcoming). More generally, antisocial behavior and psychopathology are risk factors for poor health (Farrington 1995; Graham 2003; Joukamaa 1998; Laub and Vaillant, 2000; Nieuwbeerta and Deerenberg 2004; Odgers et al. 2007; Timonen et al. 2003), which in turn may contribute to higher mortality rates in criminal and mentally ill populations. These same populations are also more likely to drop out of longitudinal studies either because participants simply cannot be located or because they are unwilling to continue their participation.<sup>1</sup>

If the nonrandom attrition of study participants is closely tied to the behavioral patterns or outcomes that are the object of the investigation, biases of unknown size and sign may be induced.<sup>2</sup>

When the longitudinal measurements on certain categories of individuals are censored due to the individuals' terminating their participation in the study or truncated due to their death, the opportunity to measure interesting trajectories of behaviors or outcomes may be lost. Here we use the terminology *truncation by death* for any situation in which the outcome of interest is not defined after attrition in comparison to censoring due to terminated study participation where the outcome is defined but not observed; see Zhang and Rubin (2003) for elaboration of this important distinction. For example, high mortality rates of individuals with a specific medical condition may make it impossible to predict what functioning levels would be if the condition could be remediated. Similarly, because prison sentences are often lengthy for individuals who commit serious violent crimes, the extended confinement of such individuals makes it difficult to test predictions about what their criminal trajectories would have been absent extended confinement. Finally, in certain circumstances, for example, when attrition is due to death, the reasons for attrition are a substantively interesting topic in their own right.

The purpose of this article is to demonstrate an extension of group-based trajectory modeling (Nagin 1999, 2005; Nagin and Land 1993; Nagin and Tremblay 1999; Roeder, Lynch, and Nagin 1999) that accounts for nonrandom participant attrition that varies across trajectory groups. In this context, our interest is in exploring issues such as how the modeling extension affects trajectory group shapes and sizes and model-based predictions about outcomes among the nontruncated population. The article is organized as follows. The second section lays out the model extension. The third section reports simulation results on the effects of dropout on trajectory group shapes and sizes and on projections of nontruncated population averages from model estimates. The fourth section illustrates an application of the model extension to a large China-based longitudinal study of disability level in an elderly population that has a very high mortality rate. The fifth section concludes.

We begin with an important caveat about the aim of this article. Participant attrition is a specific example of missing or truncated data, a topic about which there is a vast literature. The literature focusing specifically on the issue of the effects of nonrandom participant attrition on longitudinal analysis is far smaller, but this topic too has received considerable scholarly attention (cf. Brown 1990; Diggle and Kenward 1994; Little 1995; Molenberghs and Verbeke 2005; Zhang and Rubin 2003). To our knowledge, this article is the first to visit this issue in the context of group-based trajectory modeling, or what Muthén (2001) refers to as *latent class growth analysis*. Our aim is not to advance the theoretical literature on the modeling of nonrandom participant attrition. As will be seen, our approach is very conventional and addresses a specific type of nonrandom attrition. Instead, our purpose is to provide researchers who use group-based trajectory modeling with a practical tool for examining whether there is evidence of differences in attrition rates across trajectory groups and whether those differences might be materially affecting conclusions deriving from a model that assumes the sample attrition occurs at random. It is also our objective to provide a tool for analyzing whether differences in attrition rates across trajectory group are associated with observable participant characteristics.

## Model

In this section we lay out our approach to generalizing the group-based trajectory to accommodate participant attrition.

### *The Basic Model Without Dropout*

Using finite mixtures of suitably defined probability distributions, the group-based approach for modeling developmental trajectories is intended to provide

a flexible and easily applied method for identifying distinctive clusters of individual trajectories within the population and for profiling the characteristics of individuals within the clusters. Thus, 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. Technically, the group-based trajectory model is an example of a finite mixture model. Maximum likelihood is used for the estimation of the model parameters. The maximization is performed using a general quasi-Newton procedure (Dennis, Gay, and Welsch 1981; Dennis and Mei 1979).

The fundamental concept of interest is the distribution of outcomes conditional on age (or time); that is, the distribution of outcome trajectories denoted by  $P(Y_i|Age_i)$ , where the random vector  $Y_i$  represents individual  $i$ 's longitudinal sequence of behavioral outcomes and the vector  $Age_i$  represents individual  $i$ 's age when each of those measurements is recorded.<sup>3</sup> The group-based trajectory model assumes that the population distribution of trajectories arises from a finite mixture of unknown order  $J$ . The likelihood for each individual  $i$ , conditional on the number of groups  $J$ , may be written as

$$P(Y_i|Age_i) = \sum_{j=1}^J \pi^j \cdot P(Y_i|Age_i, j; \beta^j), \quad (1)$$

where  $\pi^j$  is the probability of membership in group  $j$ , and the conditional distribution of  $Y_i$  given membership in  $j$  is indexed by the unknown parameter vector  $\beta^j$ , which among other things determines the shape of the group-specific trajectory. Typically the trajectory is modeled with a polynomial function of age. For given  $j$ , conditional independence is assumed for the sequential realizations of the elements of  $Y_i$ ,  $y_{it}$ , over the  $T$  periods of measurement. Thus, we may write

$$P(Y_i|Age_i, j; \beta^j) = \prod_{t=1}^T p(y_{it}|age_{it}, j; \beta^j), \quad (2)$$

where  $p(\cdot)$  is the distribution of  $y_{it}$  conditional on membership in group  $j$  and the age of individual  $i$  at time  $t$ .<sup>4</sup> For the applications reported in this analysis,  $p(\cdot)$  is either the normal or censored normal distribution, but in general there is no restriction on the form of  $p(\cdot)$ .

### *The Basic Model Extended to Include Dropout*

The basic trajectory model accommodates missing data due to participant attrition whether due to truncation or censoring or to intermittent missed

assessments by assuming the data are defined but unobserved and are missing at random. For longitudinal latent class mixture models, this implies that conditional on the observed outcome values (as well as any additional covariates) attrition is independent of the unobserved outcomes. Conditional on observed outcomes and covariates, it thus assumes attrition and latent trajectory group membership are independent. As the basic model does not allow for trajectory probabilities to change over time, the estimated trajectory probabilities can be interpreted as those for the full population, without any missing data or attrition.

The proposed generalization relaxes the assumption that latent trajectory group membership is independent of attrition and allows for estimation of additional quantities relevant when attrition is present. The attrition process is directly modeled and allows for variation across trajectory groups. The generalization permits dropout probability to vary as a function of observed outcomes prior to dropout as well as other covariates. Thus, the trajectory-specific attrition rates and the resulting trajectory group probabilities over time for the remaining population after attrition are estimated. The extension applies to nonrandom attrition due either to censoring or truncation. However, the interpretation of the model's parameters crucially depends upon this distinction, a point we develop more fully in the third and fourth sections. The additional estimated quantities in the extended model are likely to be of particular interest when attrition is due to death or more generally an event causing the outcome to be undefined (i.e., truncated), as this would focus interest on patterns in the remaining population.

In the discussion that follows we use the term *missing* to describe instances of attrition due either to censoring or truncation. The basic trajectory model is generalized to account for nonrandom attrition as follows.<sup>5</sup> The elements of  $Y_i, y_{it}$ , are redefined to incorporate attrition information. As before,  $y_{it}$  equals its realized value prior to dropout. In the period of dropout and thereafter,  $y_{it}$  is designated as missing. The revised likelihood specifies the joint probability of the realized values of  $y_{it}$  (and the probability that they are observed), prior to dropout and their missingness thereafter.

Suppose participants are measured over a total of  $T$  measurement occasions. Let  $w_{it} = 1$  if individual  $i$  has dropped out by  $t \leq T$  and 0 otherwise,  $\tau_i =$  the period  $t > 1$  that individual  $i$  drops out and  $T + 1$  if individual  $i$  does not drop out, and  $\theta_t^j =$  probability of dropout in period  $2 \leq t \leq T$  given membership in group  $j$ .<sup>6</sup> Thus, by construction, the  $w_{it}$  are dependent over time as  $w_{it} = 0$  in a given period implies  $w_{it} = 0$  in any prior periods. Similarly,  $w_{it} = 1$  in a given period implies  $w_{it} = 1$  in any subsequent periods.

To account for dropout at  $\tau_i < T + 1$  and the attendant data censoring or truncation, equation (2) must be altered. For each period up to  $\tau_i$  for which there is data, the probability of the observed data given membership in group  $j$  is  $p(y_{it}|w_{it}=0, age_{it}, j; \beta^j)(1 - \theta_t^j)$ . Multiplying across the  $\tau_i - 1$  periods prior to dropout for which there is data, the probability becomes  $\prod_{t=1}^{\tau_i-1} p(y_{it}|w_{it}=0, age_{it}, j; \beta^j)(1 - \theta_t^j)$ . Finally, to account for the missingness due to dropout from period  $\tau_i$  and onward, this probability must be multiplied by the probability of dropout at  $\tau_i$ ,  $\theta_{\tau_i}^j$ . Thus, equation (2) in its more general form is

$$P(Y_i|age_i, j; \beta^j, \theta_{\tau_i}^j) = \left[ \prod_{t=1}^{\tau_i-1} p(y_{it}|w_{it}=0, age_{it}, j; \beta^j)(1 - \theta_t^j) \right] \theta_{\tau_i}^j \quad (3)$$

Equation (3) is substituted into equation (1) to form the unconditional likelihood for individual  $i$ .

Our model extension allows  $\theta_t^j$  to vary by trajectory and within trajectory across time. The model also allows for specification of  $\theta_t^j$  as a function of observed covariates. Let  $x_i$  denote measured psychosocial characteristics of individual  $i$  and  $H_t$  denote a vector composed of realizations of  $y_{it}$  through  $\tau_i - 1$ , namely,  $y_{i1}, y_{i2}, \dots, y_{i\tau_i-1}$  and of any other time varying-covariate through  $\tau_i - 1$ , such as age, that might be associated with dropout probability. Then  $\theta_t^j$  can be expressed as  $\theta_t^j(x_i, H_t, t; \gamma_j)$  where  $\gamma_j$  are trajectory group-specific parameters of the drop-out process. Here we specify  $\theta_t^j(x_i, H_t, t; \gamma_j)$  to follow the binary logit function:

$$\theta_t^j(x_i, H_t, t; \gamma_j) = 1 / (1 + e^{f^j(x_i, H_t, t; \gamma_j)}), \quad (4)$$

where  $f^j(x_i, H_t, t; \gamma_j)$  is specified to be linear in its parameters.

We note that equation (3) allows for estimation of differential drop-out rates by trajectory group over time. Equation (4) allows for additional modeling of attrition at the individual level to estimate the relationship between dropout and observed values of covariates or observed outcomes within trajectory. With or without modeling dropout at the individual level, the extended model assumes that attrition is independent of unobserved outcomes (if defined) within trajectory group.

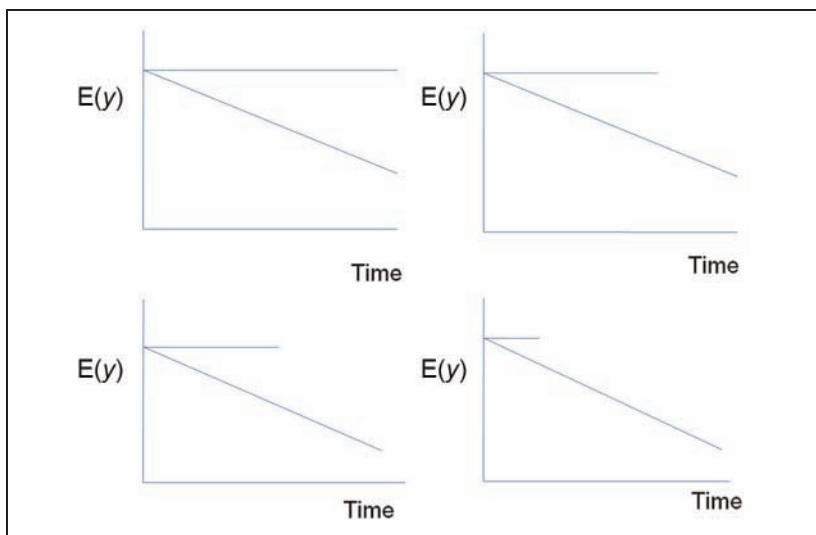
Statistics and econometrics use different terminology for describing types of missing data. Both are useful for describing our approach. The work of Rubin (1976) and Little and Rubin (1987) has shaped the terminology in statistics. They distinguish three types of missing data: (1) data that are missing

completely at random, (2) data that are missing at random conditional on observed covariates, and (3) nonignorable missing data where the process generating missing values depends on unobserved covariates such as the value of what the outcome variable would have been had it been observed (if defined). In the econometrics literature (cf. Heckman 1976; Manski 1995), missing data are treated as a form of selection. The second category of missing data in the Little and Rubin (1987) taxonomy is described as selection on observables or exogenous selection and the third category as endogenous selection.

Within trajectory group our approach to modeling attrition is an example of what economists would call an exogenous selection model. Also, because the trajectory groups are closely related to the concept of principal strata as defined by Frangakis and Rubin (2002), the proposed model is a special case of “latent ignorability” in the sense that missingness is assumed ignorable within the latent strata. Whether this is a reasonable approach has to be judged on a case-by-case basis, a point that is emphasized in Little (1995). However, as trajectory groups are latent and not observed, the extended model addresses a particular form of endogenous selection (or nonignorable missing data) in which dropout is a function of observed outcomes, observed covariates, and unobserved trajectory group membership. In cases where the trajectory groups are well separated, observed outcomes values will indicate trajectory membership with little uncertainty and the dropout will effectively be exogenous (or missing at random).<sup>7</sup> In cases where the trajectory groups are not well separated, we would expect the base model that does not address the endogenous missing data mechanism to be susceptible to bias. The first simulation in the next section explores this very circumstance.

## Simulation Results

In this section we report the results of simulations that were designed to explore the effect of trajectory-specific participant attrition on trajectory shapes and sizes and predictions of nontruncated population averages based on model parameter estimates. We examine these effects for the base model without additional modeling for dropout and for the extended model. Our aims were to (1) explore the nature of the biases that are induced by the failure to account for endogenous dropout, (2) ascertain the degree to which these biases are averted with the application of the extended model, and (3) determine the ability of the extended model to accurately estimate the additional dropout parameters. However, we did not undertake a wide-ranging exploration of circumstances where dropout might induce biases in group-based trajectory



**Figure 1.** Two-group model with different drop probabilities

models. That would be beyond the scope of this article. Instead, our investigation focused on exploring the situation where there is a small group that is disproportionately prone to dropout. Of particular interest was how this specific form of dropout might inhibit one's ability to detect and accurately estimate the size of the small group. Examples of such groups are the most antisocial in a study of criminal development, the most mentally ill in a study of the developmental trajectories of psychopathology, or the most frail in the study of the developmental course of disability in an elderly population.

We anticipate that the model without the dropout extension will be most affected in this regard when groups are initially not well separated and there are large differences in the rate of dropout between groups. Figure 1 illustrates this situation for a simple two-group model. In the initial period both groups have the same expected outcome value. Thereafter, the expected outcome value for one group declines linearly whereas the expected outcome value of the other group remains unchanged. Further suppose that dropout is limited to the initially small group with an unchanging expected value. Successive panels in the figure are meant to represent the effect of having successively fewer assessment periods to estimate the size and course of the constant trajectory. Because as time goes on the trajectories for the two groups become increasingly separated, more assessment periods will make



the groups easier to detect statistically. Conversely, as the number of assessment periods declines, the separation between the groups diminishes, which will make identification of the trajectory groups ever more difficult. Moreover, because dropout is restricted to the constant trajectory, it is this group whose size will be ever more difficult to detect. In the extreme where there is only one period of data for the constant group, its trajectory is indistinguishable from the declining group.

To examine these issues we generated data according to the following two group model:

1. Trajectory group characteristics:

$$\text{Group 1 : } y_{it}^1 = 10 + \epsilon; \pi^1 = .2$$

$$\text{Group 2 : } y_{it}^2 = 10.5 - .5t + \epsilon; \pi^2 = .8$$

$$T = 1, 2, \dots, 6$$

where  $y_{it}^1$  and  $y_{it}^2$  are the simulated outcomes of for the  $i$ th member in period  $t$  for groups 1 and 2, respectively;  $\pi^1$  and  $\pi^2$  are, respectively, the probability of membership in groups 1 and 2; and  $\epsilon$  is an independent normally distributed random variable with  $\mu = 0$  and  $\sigma = 1$ .

2. The one feature that varies across simulations is the dropout probability per period for group 1, which ranged from 0 to .4 by increments of 0.05. Dropout probability was always equal to 0 for group 2. Dropout within group 1 was random across members and remained constant at the specified value across period. However, because dropout probability differs across groups, it is nonrandom at the population level.
3. Sample size: At the outset of all simulations there were a total of 2,000 members, 400 from group 1 and 1,600 from group 2. Because dropout was limited to group 1, group 2 members all had six periods of simulated measurements whereas the measurement series ends at the period of dropout for group 1 members who attrite.
4. For each group 1 dropout probability, 100 simulations were run.

Several observations about the design of the simulation are as follows: We limited dropout to the small group, group 1, to mimic the reality that dropout is often highest among the smaller groups. We purposely designed the simulations so that the groups had equal means at  $t = 1$  and diverge after that,

with a smaller high group and a larger decreasing group. Because the groups are not well separated at the outset of the simulation, group membership initially will be highly uncertain. Thus, the group identity of the group 1 members who drop out early will be highly uncertain in the model without the dropout extension. By contrast, in the model with the dropout extension, dropout is informative about group membership. Indeed, because dropout is limited to group 1, dropout is a certain signal of membership in that group. Had we allowed dropout in group 2, the signal of dropout would not definitively establish group membership, but the principle that dropout is informative of group membership if dropout rates differ across groups still holds. The strength of that signal will depend on the difference in rate across group. This is why we varied the dropout rate for group 1. Finally, dropout is conditional on trajectory group membership but not on individual covariates or on observed (or unobserved in the case of defined but missing observations) outcome values within trajectory group. Due to this particular form of dropout, we would expect both the models with and without the dropout extension to have unbiased estimates of the mean trajectory shapes.

Table 1 summarizes the results of this simulation study. The headings "Without dropout" and "With dropout" refer to whether the models were estimated without or with the dropout extension. All models were estimated with the SAS-based procedure Proc Traj (Jones and Nagin 2007; Jones, Nagin, and Roeder 2001). As in the basic model, maximization of the likelihood is performed using a general quasi-Newton procedure. The entries under the headings "Model without dropout" and "Model with dropout" are averages over the 100 simulations for each group 1 dropout probability.

At each design point (group 1 per period dropout rate) of the simulation, group 1 expected number of assessment periods and total dropout probability on or before period 6 are also reported. Even for a per period dropout probability of only .05, more than 20 percent of group 1 members drop out from period 2 through period 6 (recall there is no dropout in period 1, which corresponds to the baseline assessment). Across the entire sample (pooling groups 1 and 2), this corresponds to a total dropout rate of 5 percent, which is about equal to the rate in long-term longitudinal studies that are most successful in retaining participants. The expected number of assessments per group 1 participant, however, remains high at 5.3. For drop-out rates of 10 percent or more, group 1 retention rates begin to plummet rapidly. For a probability of .15 or more, less than half of group 1 members are retained for six periods.

Turning now to the effect of dropout on model parameter estimates, not reported in the table are statistics on the effect on trajectory shape parameter

**Table 1.** Simulation Results: Group 1 and Group 2 Initially not Well Separated

Group 1 per period dropout probability	Expected group 1 assessment periods	Probability of group 1 dropout on or before period 6	Model without dropout		Model with dropout	
			Group 1 probability estimate ( $\pi^1$ )	Percentage bias	Group 1 probability estimate ( $\pi^1$ )	Dropout probability estimate Percentage bias
0	6.0	0	.200	0.0	.200	0.0
.05	5.3	.226	.171	-14.5	.199	-0.5
.10	4.7	.410	.146	-27.0	.199	-0.5
.15	4.2	.556	.122	-39.0	.200	0.0
.20	3.7	.672	.100	-50.0	.199	-0.5
.25	3.3	.762	.079	-60.5	.200	0.0
.30	2.9	.832	.061	-69.5	.199	-0.5
.35	2.6	.884	.046	-77.0	.199	-0.5
.40	2.4	.922	.034	-83.0	.199	-0.5

estimates. Even at the highest drop-out rate there was no evidence of material bias in these parameter estimates either in the model with dropout or without the dropout extension. Specifically, both types of models correctly identified each trajectory group's common starting point of 10 with high precision. The slope for the declining trajectory was accurately estimated and the lack of change in group 1 was also identified, albeit with less precision.

Estimates of the initial size of group 1,  $\pi^1$ , however, were far less immune to bias. As dropout probability increased across simulations, the model without the dropout extension produced increasingly biased estimates of the initial size of group 1. The average estimate of  $\pi^1$  declined steadily as group 1 dropout probability increased, from .171 for a dropout probability of .05 to .034 for a dropout probability of .4. Even for a group 1 dropout probability of .1,  $\pi^1$  is underestimated by 27 percent. By contrast, the model with the dropout extension does a far better job of estimating both  $\pi^1$  and the dropout probability. There is no evidence of bias, and the average estimate of both parameters is very close to the true value for each value of the dropout rate.

We next illustrate a more subtle impact of unequal drop-out rates across trajectory groups. In the model with the dropout extension, the group membership probabilities,  $\pi^j$ , are properly interpreted as the size of group  $j$  at baseline, period 1. With differential dropout across trajectory groups, the relative sizes of the trajectory groups contributing to the pool of nonmissing data will change over time. Groups with higher drop-out rates will become relatively smaller than at baseline compared to the lower drop-out groups. For example, in the simulation model with a drop-out rate of .4 for group 1, by period 6 more than 90 percent had dropped out whereas all of group 2's members remain. Thus, by period 6 nearly all the nonmissing values of  $y$  were contributed by group 2. Let  $\tilde{\pi}_t^j$  denote the size of the nonattriting members of group  $j$  at time  $t$  and recall that  $\theta^j$  denotes the per period dropout probability of group  $j$ . For  $t = 1$   $\tilde{\pi}_t^j = \pi^j$ . For  $t > 1$ , and assuming constant drop-out rates across time within group,<sup>8</sup>

$$\tilde{\pi}_t^j = \frac{\pi^j (1 - \theta^j)^{t-1}}{\sum_j \pi^j (1 - \theta^j)^{t-1}} \quad (5).$$

Observe that in equation (5) each  $\pi^j$  is weighed by  $(1 - \theta^j)^{t-1}$ . Note that  $(1 - \theta^j)$  measures the probability of not dropping out in each period  $t$ . Thus,  $\pi^j$  is weighted by group  $j$ 's probability of survival through period  $t$ ,  $(1 - \theta^j)^{t-1}$ . Consequently, the time path of  $\tilde{\pi}_t^j$  will depend on group  $j$ 's survival probability relative to other groups. Groups with relatively high

survival probabilities will increase in relative size and those with lower survival probabilities will decline in relative size.

The proper interpretation of  $\tilde{\pi}_t^j$  depends crucially upon whether the missingness is due to censoring or truncation (e.g., death). If missingness is due to censoring,  $\tilde{\pi}_t^j$  should be interpreted as the expected proportion of group  $j$  members who remain in the *sample* at time  $t$ , whereas if missingness is due to truncation, then  $\tilde{\pi}_t^j$  should be interpreted as the expected proportion of the group  $j$  members who remain in the *population* at time  $t$ .

This distinction is crucial to making predictions about the expected value of  $y_t$  when missingness is due to truncation versus censoring. When missingness is due to censoring, under the admittedly strong assumption that  $E(y_t^j)$  is the same for the missing and nonmissing observations,  $E(y_t) = \sum_j \pi^j y_t^j$ , where the reader is reminded that  $\pi^j$  measures the size of the trajectory group at baseline. However, this relationship does not apply when attrition is due to truncation. Because group membership probabilities change over time for those who do not attrite, this change must be taken into account in translating trajectory group-specific predicted values of the outcome of interest,  $y_t^j$ , into a prediction about the (nonmissing) expected value of  $y_t$  at the population level. Specifically, when missingness is due to truncation,  $E(y_t) = \sum_j \tilde{\pi}_t^j y_t^j$  not  $\sum_j \pi^j y_t^j$ .

To illustrate this point we estimate a modified version of the aforementioned simulation in which we set the mean of the constant trajectory at 12.5. We do this so that even at baseline, the trajectories are well separated. For this simulation, missingness should be interpreted as resulting from truncation. Table 2 reports calculations of  $E(y_t)$  for this model with  $\theta_1^1 = 0.2$  for  $t$  between 2 and 6. The column heading "Average  $y$ " reports the average value across the 100 simulations of the sample averages of the  $y_t$  among those who hadn't dropped out (i.e., died) each period. In the model without dropout the average estimated value of  $\pi^1$  across the 100 simulations was .185, which is only modestly smaller than its true value .2. The modest bias in the  $\pi^1$  estimate is consistent with the aforementioned discussion about the importance of separation rather than size in gaining reasonable estimates of group size at baseline in models that do not account for dropout. Also, as with the results in simulation 1, the trajectory shape parameters in the model without dropout were close to their true values. Notwithstanding, if the model parameter estimates are used to predict the average value of  $y_t$ , the model without adjustments for dropout gives an estimate of the average values as if there were no nonrandom dropout due to truncation. As time goes on, the predictions have an increasingly upward bias relative to the average values among those who have not dropped out because the calculations do not take into account the

**Table 2.** Simulation 2: Predicting Population Averages With and Without Adjustments for Dropout

Period	Average $y$	Model without dropout	Model with dropout	
		Predicted $y$	$\hat{\pi}_t^1$	Predicted $y$
0	10.5	10.5	.200	10.5
1	10.0	10.1	.166	10.0
2	9.48	9.67	.137	9.48
3	8.95	9.26	.112	8.95
4	8.41	8.84	.092	8.42
5	7.87	8.43	.075	7.87

ever smaller representation of group 1 with its “high chronic” trajectory who remain in the population. Consider next the predictions of the model with adjustments for dropout. In this model,  $\pi^1$  is estimated almost exactly, .199. The other model parameters are similarly close to their true values, including the groups 1 and 2 dropout rate estimates. Averaged across the 100 simulations, they are, .200 and .000, respectively. The estimates of  $\hat{\pi}_t^1$  based on the estimated values of  $\pi^1$  and the drop-out rate starts at .2 in period 1 but steadily decline thereafter. By period 6,  $\hat{\pi}_5^1$  equals .075. Also reported are the average predicted values of  $y$ , which are nearly identical to the values reported in the “Average  $y$ ” column.

In the next section we illustrate the substantive importance of accounting for differential attrition due to truncation across trajectory groups in predicting initial trajectory group sizes and population level averages over time.

## Application

In this section we report the results of an application of the dropout model extension to data from the National Institutes of Health–supported Chinese Longitudinal Healthy Longevity Survey (CLHLS), a four-wave survey conducted in randomly selected counties and cities in 22 Chinese provinces, which together account for more than 85 percent of the total population of China. The baseline data were collected in 1998. At that time, one of the main aims of the project was to collect individual health and disability data on the oldest segments of the country, and as such, the baseline targeted individuals aged 80 and older, with an oversampling of extremely old people, specifically, nonagenarians and centenarians. A weighting scheme that

accounts for the oversampling is included in the data. The baseline sample included 8,805 individuals between ages 80 and 105. Additional waves were conducted in 2000, 2002, and 2005.

The focus of our analysis is the developmental course of disability in this very elderly sample. Although additional functional status items were added in later waves, the baseline sample included only items from Katz's original Activities of Daily Living (ADL) scale (Katz et al. 1963). Specifically, individuals were asked whether, on their own and without assistive devices, they could conduct the following tasks: bathing, dressing, toileting, getting up from a bed and chair, and eating. When the interviewee was unable to respond to these ADL items due to cognitive and other health problems, proxy respondents were utilized. In the common conceptualization of the disablement process, the inability to conduct one of these ADL tasks independently is considered to be a disability (Freedman, Martin, and Schoeni 2002; Verbrugge and Jette 1994). The greater number of inabilities, the more severe the disability.

Not surprisingly, with such an elderly study population, mortality rates were high over the years of data collection. By the latest available wave of data collection in 2005, 68.9 percent of study participants had died. Analyzing data such as these pose many statistical challenges that are beyond the scope of this article. One is related to the fact that the participant age range at baseline is nearly four times larger than the total period of assessment, seven years. This combined with the high mortality rates creates enormous selection-related challenges in combining data for individuals who, for example, were age 85 or younger at baseline with individuals who were, for example, 95 or older at baseline. Not only is there no overlap in their assessment ages, 47.6 percent of those who were 85 and younger in 1998 were deceased by 2005, well before reaching what would have been their 95th birthday. Clearly, the older individuals at baseline were more robust than their comparatively younger counterparts at the same age. Thus, for our application we restrict our analysis to individuals who were similar in age at baseline, 90 to 93 years old. We also exclude individuals who dropped out of the study for reasons other than death. Thus, our application focuses solely on missingness due to truncation.<sup>9</sup>

Table 3 reports summary data on the age 90 to 93 cohort. At baseline there were 1,078 individuals in this restricted sample. By 2005, 89 percent had died. As would be expected, average ADL counts of the living rose, from 0.8 in 1998 to 1.3 in 2005. As with the simulation analysis, Proc Traj, updated to estimate group-specific dropout models, was used for model estimation. Proc Traj provides two distributional alternatives that are possible candidates for modeling the ADL counts, the censored normal and the

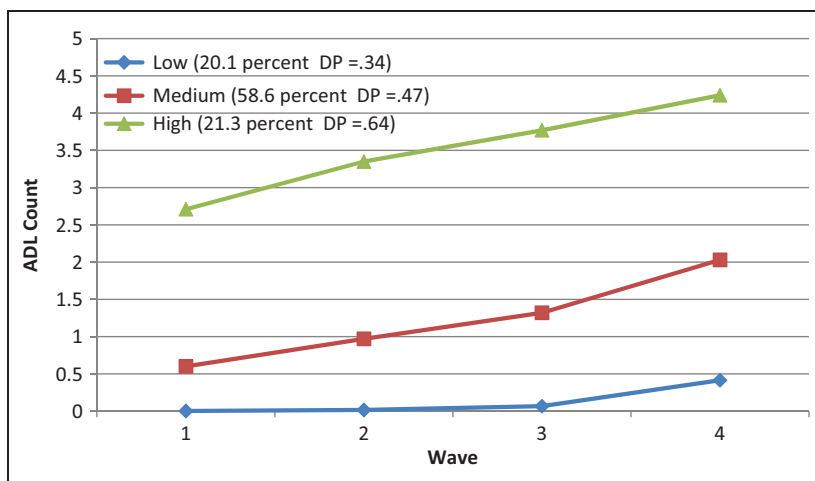
**Table 3.** Predicted Population Average Activities of Daily Living (ADL) Counts From the Models With and Without Dropout

Period	Average ADL count	Model without dropout		Model with dropout				
		Predicted ADL count	Percentage error	Predicted ADL count	Percentage error	$\hat{\pi}_t^1$	$\hat{\pi}_t^2$	$\hat{\pi}_t^3$
1998	0.84	0.91	8.3	0.93	10.7	.201	.586	.213
2000	1.05	1.19	13.3	1.07	1.9	.254	.600	.146
2002	1.16	1.42	22.4	1.17	0.9	.309	.593	.097
2005	1.26	1.89	50.0	1.58	25.4	.366	.571	.063

zero-inflated Poisson. We used the former because the maximum ADL count is five and a sizeable number of study participants are at or near that cap. The censored normal option is specifically designed to accommodate scales in which there are sizeable clusters of data at the scale's theoretical minimum and maximum. While the Poisson option by definition accommodates the scale minimum of zero, its implementation in Proc Traj does not allow for censoring at a maximum value.

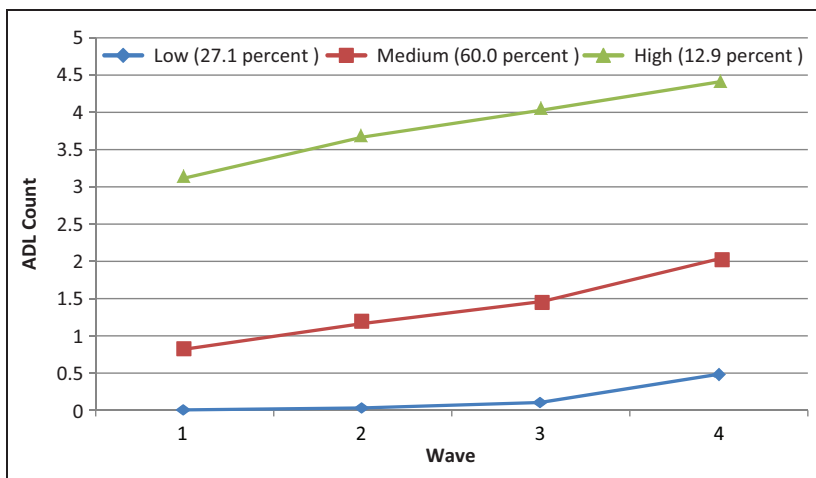
Based on the Bayesian Information Criterion, a three-group model was found to be the best model. Figures 2 and 3, respectively, depict the estimated trajectories from models with and without the dropout extension. For the model with the dropout extension, the dropout model is specified in its most basic form—constant across assessment occasions. The trajectories in the models with and without dropout are very similar, a result that mirrors the findings from the simulations analysis. Each model specification estimates a low but rising trajectory of ADL counts starting at about 0 at baseline and rising to about .4 in 2005. Each also has a medium rising trajectory starting at about .5 and finishing at an ADL count of about 1.7. Finally, each has a high and rising trajectory starting at about 2.6 and ending near the ADL count maximum of 5. The initial size of the high rising group, however, differs materially between the two models—12.9 percent in the model without the dropout extension and 21.3 percent in the model with the dropout extension. The larger initial size estimate of the high increasing group in the model with dropout is again what we would expect from the first set of simulation results. The estimated dropout rate due to death of the high group is 64 percent per assessment occasion; by 2005, only 3 individuals out of 160 from this group were still alive. By contrast, the drop-out rates for the low and medium trajectory groups, while high, are considerably smaller, 34 percent and 47 percent, respectively.





**Figure 2.** Activities of daily living (ADL) trajectory model with dropout

Table 3 is the counterpart to Table 2 from the simulation section. It reports projections of the population average ADL count from 1998 to 2005 for the models with and without the dropout extension. Its purpose is to reiterate the lesson of Table 2. The misspecified model not accounting for dropout will produce biased estimates of  $E(y_t)$  because it does not take into account the changing size of the trajectory groups among the relevant population—the living. Specifically, for the model without the dropout extension, these will be projections of the population average ADL count as if there were no difference in death rates across trajectory groups.<sup>10</sup> For the model with the dropout extension, these will be projections of the population average ADL count among those still living. In the baseline period, the model without dropout overpredicts the actual average ADL count among those living by 8.3 percent. The initially modest overprediction steadily grows to 50 percent by 2005. The growing prediction error reflects the differential mortality rates across the trajectory groups. This can be seen in the estimates of  $\hat{\pi}_t^j$  in the model with the dropout extension, which are also reported in Table 3. At baseline the high trajectory is estimated to include 21.3 percent of the 90 to 93 cohort. However, due to the very high mortality rate in the high trajectory group relative to the low and medium groups, by 2005 the high trajectory is estimated to include only 6.3 percent of the cohort's population who remain alive. By contrast, the representation of the low trajectory group grows from an estimated size of 20.1 percent in 1998 to 36.6 percent by



**Figure 3.** Activities of daily living (ADL) trajectory model without dropout

2005. When the  $\tilde{\pi}_t^j$  estimates are used to weight the predicted trajectories by group, the model does a much better job of predicting the actual average ADL counts by period. The error at baseline is slightly larger than the model without dropout, 10.7 percent versus 8.3 percent. However, in the later assessment periods the dropout adjusted model does far better than the model not adjusted for dropout. In 2000 and 2002 the errors are 2 percent or less. In 2005 the error is material, 25 percent, but far smaller than the 50 percent error in the model without dropout.

We next illustrate the introduction of covariates as predictors of probability of trajectory group membership and probability of dropout. We do this by expanding the basic model with dropout to allow both of these probabilities to be a function of the indicator variable *female* and the scale variable *health*, self-reported health status at baseline. The variable *health* is measured on a 4-point scale with higher values indicating a poorer self-reported health status. Let  $x_i$  denote a vector of these two covariate characteristics for individual  $i$ . We also allow probability of dropout in period  $t$ ,  $\theta_t^j$ , to vary as a function of the ADL count in the prior period,  $y_{it-1}$ . As described in the second section,  $\theta_t^j(x_i, y_{it-1})$  is specified to follow the binary logit function (equation (4)), and as discussed elsewhere (cf. Nagin 1999, 2005; Roeder et al. 1999),  $\pi^j(x_i)$  is specified to follow a multinomial logit model.

Table 4 reports the parameter estimates for  $\pi^j(x_i)$  and  $\theta_t^j(x_i, y_{it-1})$ . The results show that both sex and health status at baseline are significant predictors of trajectory group membership. Those with poorer self-reported health

**Table 4.** The Effect of Covariates on Trajectory Group Membership Probabilities and Within Group per Period Mortality Probability

Variable	Coefficient estimate	t score
Panel A: Trajectory group membership probability		
Group 2		
Constant	-7.834	-7.22
Health	2.000	6.46
Female	0.317	0.63
Group 3		
Constant	-1.670	-2.38
Health	0.470	2.37
Female	1.089	4.51
Panel B: Per period mortality probability		
Group 1		
Constant	-0.740	-2.46
ADL Count t-1	0.879	1.06
Health	0.217	1.63
Female	-0.362	-2.04
Group 2		
Constant	3.250	1.54
Activities of daily living (ADL) count t - 1	-0.909	-4.31
Health	0.617	1.95
Female	-0.968	-1.68
Group 3		
Constant	-1.104	-3.45
ADL count t - 1	0.243	4.30
Health	0.278	2.71
Female	0.004	0.02

status at baseline were more likely to follow one of the two higher ADL trajectories than to follow the low ADL trajectory, a result that is not surprising. Females were also significantly more likely to follow the high ADL trajectory compared to the low trajectory. Gender, however, was not a predictor of membership in the middle ADL trajectory compared to the low trajectory. Thus, these results imply that at *baseline* the low trajectory was disproportionately populated by individuals in better health and that conditional on health status, the high ADL trajectory was disproportionately populated by women.

Consider now the predictors of dropout probability, which in this case is the probability of death. For all trajectory groups the coefficient estimate for

the health variable is positive, which implies that individuals in poorer health at baseline were more likely to drop out. For a one-tailed test this association is statistically significant for the middle (odds ratio [OR] = 1.85) and high (OR = 1.32) trajectory groups and nearly significant for the low trajectory group. The positive association of the variable health with probability of dying implies that as time passes, the middle and high trajectory groups are increasingly populated by individuals who were in better health at baseline. For the low and middle trajectories, the females are less likely to drop out, although for a two-tailed test this relationship is only significant at the .05 level for the low trajectory (OR = 0.70). This relationship implies that with passing time, the low trajectory will increasingly be populated by women. Finally, the ADL count in the prior period is a significant predictor of mortality in the moderate and high trajectory groups. In the high trajectory group, higher ADL counts in the prior period are associated with a heightened risk of death in the current period, the expected result (OR = 1.28). However, in the middle ADL trajectory group, the association is unexpectedly negative (OR = 0.40). We can offer no explanation for this result.

## Conclusion

In this article, we have extended the group-based trajectory model to account for a specific form of nonrandom participant attrition and have explored the effects of the model extension on both simulated and real data. The analyses of simulated data establish that estimates of trajectory group size as measured by group membership probabilities can be badly biased by differential attrition rates across groups if the groups are initially not well separated. Differential attrition rates also imply that group sizes will change over time, which in turn has important implications for using the model's parameter estimates to make nonattriting population-level projections. The analyses of real data, the Chinese Longitudinal Healthy Longevity Survey, demonstrate that the correction of the bias noted in the simulation can lead to substantively different results about the sizes of trajectory groups defined by activities of daily living and that the dropout model extensions can provide additional insight into the changing composition of these groups over time.

Much research remains to be done on this model extension. Broadening the range of simulations to examine the effects of endogenous selection would be useful. Here we have in mind simulating the effects of estimating the model under the assumption that dropout is a function of  $y_{it-1}$  when in fact it is a function of what the realized value of  $y_{it}$  would have been had it been observed. Such a drop-out process may more realistically characterize situations where dropout

is voluntary. For example, a participant in a job training program may become lost to follow-up if she moves to take a well-paying job or becomes discouraged by the prospects of finding a well-paying job even with continued training. It would also be useful to compare models with and without the dropout extension in analyses of data with lower drop-out rates than in the CLHLS.

### Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article:

The research was supported by generous financial support from the National Science Foundation (SES-0647576) and the National Institute of Mental Health (RO1 MH65611-01A2).

### Notes

1. An analysis based on the Montreal Longitudinal Study of Boys illustrates this reality. An analysis by Nagin and Tremblay (1999) applied group-based trajectory modeling to teacher assessments of physical aggression from ages 6 to 15. The analysis identified a small group of boys, composing about 4 percent of the sampled population, who followed a trajectory of chronic physical aggression. More than one-third had no assessments at age 17, whereas only 14 percent of those following a trajectory of rare physical aggression had no assessments at the same age.
2. Notably Eggleston, Laub, and Sampson (2004) explore such biases in the context of group-based trajectory modeling. They find, not surprisingly, that the manner in which the absence of criminal activity following death is coded (i.e., as 0 offenses or missing data) materially affects results. Their findings lend support to the value of the extension to group-based trajectory modeling developed in this article.
3. Trajectories can also be defined by time (e.g., time from treatment).
4. See chapter 2 of Nagin (2005) for a discussion of the conditional independence assumption.
5. Intermittent missing data continue to be modeled under the assumption that the data are missing at random. Thus, to simplify the presentation of the extended model, we assume no intermittent missing data.
6. For an individual to be in the study he or she cannot have dropped out in period 1. Thus, the probability of dropout is necessarily zero in period 1 for all trajectory groups.
7. The term *separation* refers to the distinctiveness of the trajectories comprising the mixture. If the probability of overlap in outcomes across trajectories is always

very small, for example, because they are always far apart, the trajectory groups would be said to be well separated. If the groups are very well separated, individuals could be sorted among the trajectory groups by visual inspection of the data with few or no classification errors. In this circumstance, a mixture model would be unnecessary for undertaking the analysis because individuals could be stratified by trajectory group *ex ante*.

8. For drop-out rates that vary over time within trajectory group, replace  $(1 - \theta^j)^{t-1}$  with  $\prod_{k=1}^{k=t-1} (1 - \theta_k^j)$  in both the numerator and denominator.
9. As with most longitudinal data sets, censoring was also present in these data. Among the 90 to 93 cohort, 18.6 percent were lost to follow-up. For the purposes of this application, we listwise deleted these individuals. This of course has implications for the representativeness of the sample among the living. While not pursued in this article, the model extension proposed here could be further extended to model both forms of missingness simultaneously and hence incorporate them each in appropriate ways. As an alternative, multiple imputation techniques could be applied for those with censored as opposed to truncated data when both are present if the researcher is willing to make the missing at random assumption.
10. Or, nonsensically, the estimates can be obtained by assigning activities of daily living (ADL) rates to those who had died, based on their trajectory group membership.

## References

- Brown, C. H. 1990. "Protecting Against Nonrandomly Missing Data in Longitudinal Studies." *Biometrics* 46:143-57.
- Dennis, J. E., D. M. Gay, and R. E. Welsch. 1981. "An Adaptive Nonlinear Least-squares Algorithm." *ACM Transactions on Mathematical Software* 7:348-83.
- Dennis, J. E. and H. H. W. Mei. 1979. "Two New Unconstrained Optimization Algorithms Which Use Function and Gradient Values." *Journal of Optimization Theory and Applications* 28:453-83.
- Diggle, B. P. and M. G. Kenward. 1994. "Informative Drop-out in Longitudinal Data Analysis." *Applied Statistics* 43:49-72.
- Eggleson, Elaine, John H. Laub, and Robert Sampson. 2004. "Methodological Sensitivities to Latent Class Analysis of Long-term Criminal Trajectories." *Journal of Quantitative Criminology* 20:1-26.
- Farrington, D. P. 1995. "Crime and Physical Health: Illnesses, Injuries, Accidents, and Offending in the Cambridge Study." *Criminal Behavior and Mental Health* 5:261-78.
- Frangakis, C. E. and D.B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58:21-9.
- Freedman, V. A., L. G. Martin and R.F. Schoeni. (2002). "Recent Trends in Disability and Functioning Among Older Adults in the United States." *Journal of the American Medical Association* 288:3137.

- Glueck, S. and E. Glueck. (1950). Unraveling Juvenile Delinquency. *Juvenile Court Judges Journal* 2:32.
- Graham, A. 2003. "Post-prison Mortality: Unnatural Death Among People Released From Victorian Prisons Between January 1990 and December 1999." *Australian and New Zealand Journal of Criminology* 36(1):94-108.
- Heckman, J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5:475-92.
- Jones, B. L. and D. S. Nagin. 2007. "Advances in Group-based Trajectory Modeling and a SAS Procedure for Estimating Them." *Sociological Methods and Research* 35:542-72.
- Jones, B., D. S. Nagin, and K. Roeder. 2001. "A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories." *Sociological Research and Methods* 29:374-93.
- Joukamaa, M. 1998. "The Mortality of Released Finnish Prisoners; a 7 Year Follow-up Study of the WATTU Project." *Forensic Science International* 96:11-19.
- Katz, S., A. B. Ford, R. W. Moskowitz, B. A. Jackson and M. W. Jaffe. (1963). "Studies of Illness in the Aged." *Journal of the American Medical Association* 185:914.
- Laub, J. H. and R. J. Sampson. 2003. *Shared Beginnings, Divergent Lives: Delinquent Boys to Age 70*. Cambridge, MA: Harvard University Press.
- Laub, J. H. and G. E. Vaillant. 2000. "Delinquency and Mortality: A 50-year Follow-up Study of 1,000 Delinquent and Non-delinquent Boys." *American Journal of Psychiatry* 157:96-102.
- Little, R. J. 1995. "Modeling the Drop-out Mechanism in Repeated-measures Studies." *Journal of the American Statistical Association* 90:1112-121.
- Little, R. J. A. and D. B. Rubin. 1987. *Statistical Analysis With Missing Data*. New York: John Wiley.
- Loeber, R. and D. P. Farrington. Forthcoming. *Who Will Kill and Who Will Be Killed? Development of Young Homicide Offenders and Victims*. New York: Springer.
- Manski, C. F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.
- Molenberghs, G. and G. Verbeke. 2005. *Models of Discrete Longitudinal Data*. New York: Springer.
- Muthén, B. O. 2001. "Second-generation Structural Equation Modeling With a Combination of Categorical and Continuous Latent Variables: New Opportunities for Latent Class/Latent Curve Modeling." Pp. 289-322 in *New Methods for the Analysis of Change*, edited by A. Sayers and L. Collins. Washington, DC: American Psychological Association.
- Nagin, D. S. 1999. "Analyzing Developmental Trajectories: A Semi-parametric, Group-based Approach." *Psychological Methods* 4:139-77.
- Nagin, D. S. 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. 1999. "Trajectories of Boys' Physical Aggression, Opposition, and Hyperactivity on the Path to Physically Violent and Nonviolent Juvenile Delinquency." *Child Development* 70:1181-196.
- Nieuwbeerta, P. and I. Deerenberg. 2004 "Crimineel leven: vroeg en niet-natuurlijk sterven." *Bevolkingstrends* 52(3de kwartaal):52-59.
- Odgers, C. L., A. Caspi, R. Poulton, H. Harrington, M. Thompson, J. M. Broadbent, N. Dickson, M. R. Sears, B. Hancox, and T. E. Moffitt. 2007. "Prediction of Adult Health Burden by Conduct Problem Subtypes in Males." *Archives of General Psychiatry* 64:476-84.
- Roeder, K., K. Lynch, and D. S. Nagin. 1999. "Modeling Uncertainty in Latent Class Membership: A Case Study in Criminology." *Journal of the American Statistical Association* 94:766-76.
- Rubin, D. B. 1976. "Inference and Missing Data." *Biometrika* 63:581-92.
- Timonen, M., T. Mäkityrö, J. Miettunen, H. Hakko, P. Zitting, J. Kinnunen, and P. Räsänen. 2003. "Somatic Morbidity and Criminality: The Northern Finland 1966 Birth Cohort Study." *Forensic Science International* 132:68-75.
- Verbrugge, L. M. and A. M. Jette. (1994). "The Disablement Process." *Social Science & Medicine* 38:1-14.
- Zhang, Junni L. and Donald B. Rubin. 2003. "Estimation of Causal Effects via Principal Stratification When Some Outcomes Are Truncated by 'Death.'" *Journal of Educational and Behavioral Statistics* 28:353-68.

## Bios

**Amelia M. Haviland** is a full statistician at RAND. Her research focuses on causal inference in observational studies and longitudinal data analysis with applications in health, crime, and labor economics. Publication outlets include *Health Services Research*, *Developmental Psychology*, *Criminology*, and the *Journal of Human Resources*.

**Bobby L. Jones** is a statistician for the Center for Research on Health Care Data Center, University of Pittsburgh. Recent publications include "Advances in Group-based Trajectory Modeling and a SAS Procedure for Estimating Them" in *Sociological Methods and Research* (2007) with Daniel Nagin.

**Daniel S. Nagin** is the Teresa and H. John Heinz III Professor of Public Policy and Statistics, Heinz School of Public Policy and Management, Carnegie Mellon University. His research focuses on the evolution of criminal and antisocial behaviors during the life course, the deterrent effect of criminal and noncriminal penalties on illegal behaviors, and the development of statistical methods for analyzing longitudinal data. He is the author of *Group-based Modeling of Development* (2005).