Journal of Quantitative Criminology, Vol. 15, No. 1, 1999

Assessing Stability and Change in Criminal Offending: A Comparison of Random Effects, Semiparametric, and Fixed Effects Modeling Strategies

Shawn Bushway,[1,2] Robert Brame,[2] and Raymond Paternoster[2,3]

An important theoretical problem for criminologists is an explanation for the robust positive correlation between prior and future criminal offending. Nagin and Paternoster (1991) have suggested that the correlation could be due to time-stable population differences in the underlying proneness to commit crimes (population heterogeneity) and/or the criminogenic effect that crime has on social bonds, conventional attachments, and the like (state dependence). Because of data and measurement limitations, the disentangling of population heterogeneity and state dependence requires that researchers control for unmeasured persistent heterogeneity. Frequently, random effects probit models have been employed, which, while user-friendly, make a strong parametric assumption that the unobserved heterogeneity in the population follows a normal distribution. Although semiparametric alternatives to the random effects probit model have recently appeared in the literature to avoid this problem, in this paper we return to reconsider the fully parametric model. Via simulation evidence, we first show that the random effects probit model produces biased estimates as the departure of heterogeneity from normality becomes more substantial. Using the 1958 Philadelphia cohort data, we then compare the results from a random effects probit model with a semiparametric probit model and a fixed effects logit model that makes no assumptions about the distribution of unobserved heterogeneity. We found that with this data set all three models converged on the same substantive result—even after controlling for unobserved persistent heterogeneity, with models that treat the unobserved heterogeneity very differently, prior conduct had a pronounced effect on subsequent offending. These results are inconsistent with a model that attributes all of the positive correlation between prior and future offending to differences in criminal propensity. Since researchers will often be completely blind with respect to the tenability of the normality assumption, we conclude that different estimation strategies should be brought to bear on the data.

KEY WORDS: criminal offending; stability; change; random effects models; semiparametric models; fixed effects models.

[1]The authors contributed equally to this article.
[2]The University of Maryland, National Consortium for Violence Research.
[3]To whom correspondence should be addressed at Department of Criminology, University of Maryland, College Park, Maryland, USA.

0748-4518/99/0300-0023$16.00/0

1. INTRODUCTION

In many models of criminal offending a measure of previous criminal conduct is included as an explanatory variable in recognition of the adage that "the best predictor of future behavior is prior behavior." The stability of behavior over time is not an observation criminologists have monopolized—economists have found that those currently unemployed are more likely than the employed to be out of work at some future point in time (Phelps, 1972), accident researchers have long observed that those who have had accidental injuries in the past are more likely than chance alone to have others in the future (Arbous and Kerrich, 1951; Banks, 1977; Greenwood and Wood, 1919), and psychologists have noted that psychological disorders experienced in the past are a good predictor of subsequent bouts of emotional distress (Fischer *et al*., 1984; Robins, 1966, 1978). Having observed this pattern, criminologists, like their colleagues in other disciplines, wish to account for such observed regularities in behavior. Why is criminal offending fairly stable over time?

Following Heckman (1981), Nagin and Paternoster (1991) have suggested two general processes which could account for the observed stability in criminal conduct over time. The first of these processes implicates time-stable differences between individuals in their latent tendency to commit crimes. According to this proneness explanation, individuals vary in the probability with which they will commit crime at all points in time because they differ with respect to some risk factor (impulsivity, criminal propensity, or an antisocial trait, etc.) that is established early in life and remains, at least relatively, stable over time. Since this process attributes continuity in offending over time to persistent differences between individuals in a latent criminal risk factor, it has been termed a *population heterogeneity* explanation (Heckman, 1981; Hsiao, 1986; Nagin and Paternoster, 1991). Quite simply, population differences in a predisposition or proneness to commit crimes leads to differences in offending at all subsequent points in time. Any observed positive correlation between past and future offending therefore, is simply due to sample selection rather than causality, and is a variant of the problem of omitted variable bias (Nagin and Paternoster, 1991, p. 166). More generally, the correlation between current offending and events that are themselves the product of time-stable individual differences is spurious rather than causal. The population heterogeneity explanation of continuity in criminal offending is illustrated in Fig. 1.

A population heterogeneity explanation is compatible with a number of criminological theories. For example, Wilson and Herrnstein (1985, p. 209) have articulated a theory of crime that is a catalogue of individual differences of "enduring personal characteristics." To them, some persons

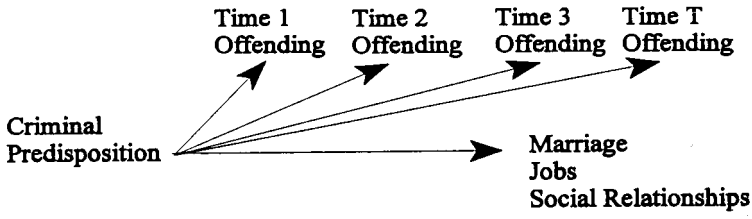

Fig. 1. Population heterogeneity explanation of continuity in offending over time.

have a higher propensity to commit crime because they are impulsive, are difficult to condition, and have a high discount rate (are short-sighted). These differences are formed very early in life and they have diverse manifestations later in life (poor school performance, alcoholism and drug addiction, and unemployment). Others have explained variations in offending with time stable individual differences in biological characteristics, such as central nervous system (Moffitt and Lynam, 1994; Raine, 1997) or neurotransmitter dysfunctions (Berman *et al.*, 1997).

Perhaps the most well-known population heterogeneity explanation in criminology today is Gottfredson and Hirschi's (1990) general theory of crime. In their theory, crime and other self-destructive behaviors are due to individual differences in self-control. Self-control is a person's ability to appreciate and consider the long-term consequences of their actions. It is formed early in life as a product of socialization experiences within the family and is relatively time-stable thereafter. Persons low in self-control are unable to resist the short-run temptations offered by crime and other acts that, like crime, provide immediate and easy gratification (unemployment, sexual promiscuity, drug abuse).

While the specific source may differ, in each of these population heterogeneity explanations, continuity in criminal offending is due primarily to preestablished differences in offending propensity, such that a person's experiences in later life have no causal impact on criminal offending. As is true for the relationship between past and future offending, any observed correlation between such later life experiences and criminal conduct is merely spurious.

A second explanation for the observed continuity in criminal offending over time argues that persons who commit criminal offenses are substantially transformed by their experiences in such a manner as to weaken existing restraints and/or strengthen existing incentives to commit crimes in the future. In this view, previous criminal conduct has a genuine causal effect on future criminality and has been termed a *state dependence* explanation

(Nagin and Paternoster, 1991). Unlike the population heterogeneity explanation, the state dependence explanation of the continuity in offending over time emphasizes contagion and the dynamic interplay between the actions of offenders and their environment. That is, prior offending affects future offending because it can destroy marriages and jobs and bring offenders into closer affiliation with like-minded other offenders. For those who commit crimes, therefore, things can get materially worse and they may become more enmeshed in offending. Similarly, finding the right partner or job may materially improve the lot of a prior offender, and such changes in their lives may lead to short- or long-term desistance from crime. The state dependence explanation for the positive correlation between past and future criminal behavior is shown in Fig. 2.

As Nagin and Paternoster (1991, pp. 166–167) have noted, the state dependence explanation is also congenial with a number of theories of crime. For example, criminal behavior can damage a person's social bond by weakening conventional attachments, by eroding conventional commitments or aspirations, and by undermining moral restraints (Hirschi, 1969). Consistent with social learning theory, criminal conduct may lead one into closer affinity with deviant others, and the values and norms they support (Akers, 1985). Consistent with the process described by labeling theorists, criminal conduct can create "problems of adjustment" that are responded to by additional, secondary deviance (Lemert, 1972). Criminal acts can also lead to conflict with teachers and parents and have other consequences that both generate new sources of strain and abrade previously successful adaptations to strain (Agnew, 1992).

Although we have discussed the population heterogeneity and state dependence explanations as if they were rival hypotheses, they are not incompatible processes. Indeed, one can easily believe that continuity in criminal behavior may be due to a mixture of both differential proneness to crime (population heterogeneity) and contagion (state dependence). Sampson and Laub's (1993, 1995, 1997; Laub and Sampson, 1993) theory of age-graded informal controls is such a mixed model that combines both

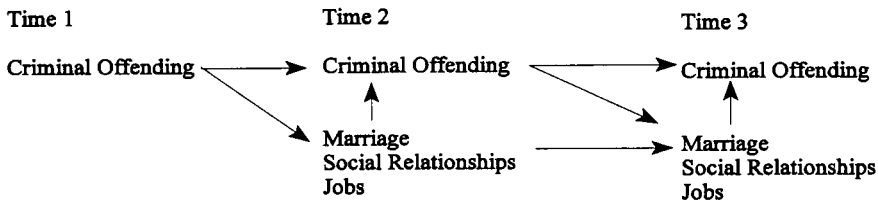

Fig. 2. State dependence explanation of continuity in offending over time.

explanations. Their position is that while there is differential initial proneness to crime, the commission of criminal acts has negative causal effects on the lives of offenders, consequences that may lead to additional crime (Sampson and Laub, 1997, pp. 144–145):

> Invoking a state dependence argument (Nagin and Paternoster, 1991), our theory incorporates the causal role of prior delinquency in facilitating adult crime through a process of "cumulative disadvantage." ... We emphasize a developmental model where delinquent behavior has a systematic attenuating effect on the social and institutional bonds linking adults to society.... The cumulative continuity of disadvantage is thus not only a result of stable individual differences in criminal propensity, but a dynamic process whereby childhood antisocial behavior and adolescent delinquency foster adult crime through the severance of adult social bonds.

Unlike a pure population heterogeneity or state dependence explanation, therefore, mixed models like Sampson and Laub's include both. While not hostile to the importance of individual differences, mixed models presume that change and later life events do, nonetheless, matter. This mixed model is shown in Fig. 3.

The current debate in criminology between the predominance of continuity and change in offending (Gottfredson and Hirschi, 1990, 1995; Nagin and Paternoster, 1991, 1994; Sampson and Laub, 1993, 1995) can, therefore, be understood as a debate over population heterogeneity vs. state dependence explanations. A pure population heterogeneity explanation implies that continuity in criminal offending is due entirely to time-stable differences in a latent proneness to crime and that later life events such as marriages or jobs will have no effect on offending after controlling for sources of criminal propensity. A pure state dependence argument implies that continuity in offending is due entirely to a process of contagion, where criminal behavior increases the probability of future criminal acts by reducing inhibitions and strengthening incentives to crime. A mixed model implicates both proneness and contagion. A critical test of three positions, therefore,

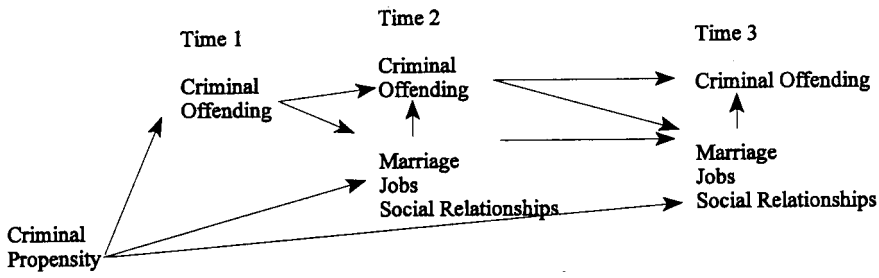

Fig. 3. Mixed population heterogeneity and state dependence explanation of offending.

would require assessing the impact of dynamic factors on crime *after* a careful calibration and control for sources of individual heterogeneity.

The strategy for those investigating population heterogeneity and state dependent effects, therefore, has been to measure individual differences in criminal propensity with available data. To do this, researchers have constructed indices of criminal propensity comprised of measured individual factors (for example, IQ, daring personality, attitude toward deviance, poor parental child-rearing behavior, parental criminality) that have been thought to be related to criminal behavior (Nagin and Paternoster, 1991; Sampson and Laub, 1993; Nagin and Farrington, 1992a, b; Paternoster and Brame, 1997). There are, of course, two problems with this strategy: (1) there is no agreed-upon understanding of exactly what the elements of such an index of criminal propensity might be, and (2) even if such agreement did exist, it is unlikely that available data sources would have measured indicators of all or even many such elements. As a result, while some sources of heterogeneity are observed and measured, researchers could not be confident that they have captured all such between-individual differences in criminal propensity. In response to this, researchers have tried to incorporate unobserved sources of persistent heterogeneity in their statistical models.

The first attempt to control for unobserved sources of criminal propensity was with so-called "random effects" models. Random effects models decompose the error term into two components, one of which reflects time-stable differences across individuals (unobserved persistent heterogeneity). With such a random effects probit model, Nagin and Paternoster (1991) found that prior criminal behavior had an effect on subsequent behavior even after controlling for observed and unobserved sources of criminal propensity. These findings were inconsistent with a pure population heterogeneity theory. Subsequent to this, random effects models were used by Nagin and Farrington (1992a, b); Paternoster and Brame (1997); Sampson and Laub (1993); and Paternoster *et al*. (1997), and in each case the reported findings could not be squared with a pure population heterogeneity explanation—controlling for both observed and unobserved heterogeneity, prior criminal offending had an effect on current criminal offending.

Although the random effects model provided a tractable way to control for unobserved heterogeneity, Nagin and Paternoster (1991, p. 169) cautioned about the fragility of their findings. They observed (p. 183) that such random effects models presume that unobserved heterogeneity is normally distributed in the population and that "[r]esults can be very sensitive to distributional assumptions about the nature of the heterogeneity." In view of the possible sensitivity of results to the distributional assumptions about unobserved heterogeneity, Nagin and Land began to develop an alternative

modeling strategy (Nagin and Land, 1993; Land *et al*., 1996; Land and Nagin, 1996). In their semiparametric mixed Poisson model, no parametric assumption is made about the distribution of persistent unobserved heterogeneity; instead, the heterogeneity is nonparametrically approximated in model estimation.[4] This flexibility comes at a cost: we have to assume that unobserved individual differences are drawn from a discrete (multinominal) probability distribution. If unobserved individual differences are, in fact, drawn from a continuous distribution, there will be some misspecification bias.

There is, of course, another, though far more blunt approach to the treatment of persistent unobserved heterogeneity. A fixed-effects strategy, though not yet commonly employed by criminologists, models unobserved heterogeneity explicitly as a time-constant intercept term for each individual in the sample. Essentially, this individual-specific intercept term captures *all* individual effects that are constant over time—no other assumptions about persistent unobserved heterogeneity are necessary. This approach takes out any time-constant individual effect (i.e., gender, race, intelligence, criminal propensity) without explicitly specifying its substance. The fixed effects approach uses a constant for each individual to absorb all individual-specific effects so that any observed effect for a dynamic factor must, by definition, be independent of stable criminal propensity. Indeed, it is independent of any stable individual characteristic.

Essentially, then, the Nagin and Paternoster (1991) paper highlighted the possible sensitivity of a random effects modeling strategy. To the extent that persistent unobserved heterogeneity can be assumed to be normally distributed, the random effects approach to studying state dependence and population heterogeneity seems reasonable. If unobserved criminal propensity is not normally distributed, then estimated structural coefficients from random effects models may be biased. The distributional sensitivity issue was addressed by Nagin and Land by introducing a nonparametric alternative, although they did not directly confront the sensitivity issue originally raised by Nagin and Paternoster. In addition to the nonparametric approach developed by Nagin and Land, there is a fixed effects strategy that rather bluntly deals with time-stable individual differences by including a constant for each individual. We can think of these three approaches to the treatment of unobserved persistent heterogeneity (random effects,-nonparametric, fixed effects) as lying on a continuum—the random effects approach is the most restrictive model because it treats unobserved

[4]Nagin and Land referred to their model as a *semi*parametric model because it "combines a parametric specification of the regression component of the model with a non-parametric specification of the error term" (Land and Nagin, 1996, p. 170).

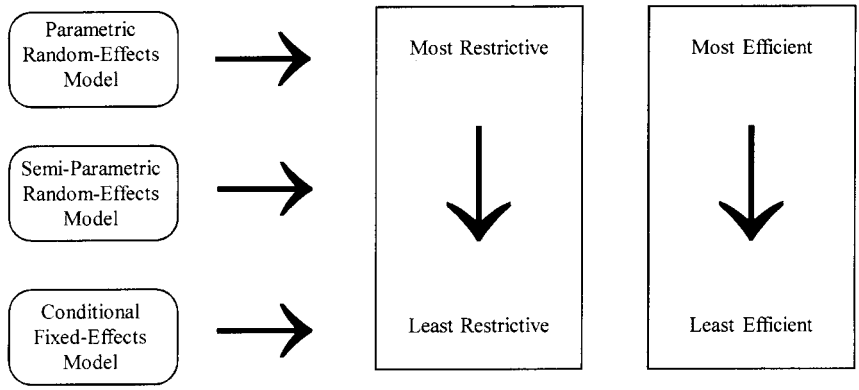

Fig. 4.

heterogeneity as part of the error term and normally distributed, the nonparametric model treats heterogeneity as part of the error term but assumes nothing about the shape of its distribution, and the fixed effects model simply absorbs all individual-specific effects with unique intercept terms. In addition, as we will demonstrate later, these models also lie on an efficiency continuum, with the fixed effects model having the least efficient estimator and the random effects model providing the most efficient estimator (see, Fig. 4).

With these different estimation strategies to an important substantive problem as a backdrop, we have two objectives we wish to accomplish in this paper: (1) to explain in as clear a manner as possible the assumptions of each statistical model, especially the fixed effects approach, because most criminologists will not be familiar with it, and (2) to examine the state dependence/population heterogeneity explanations by comparing the results obtained from these three different modeling strategies on a simple data set. Ultimately, our argument will be that all three strategies employed together can provide useful theoretical and methodological insights that could not be obtained by the use of only one strategy. In other words, our position is that rather than choosing among random effects, semiparametric, and fixed effects strategies, all methods should be used when the data allow. If different statistical models that make different assumptions lead to similar conclusions, researchers can be more confident that the story they are telling is a resilient one. In addition, an important insight can be gained about the sensitivity of findings to distributional assumptions by comparing the results of different statistical models that make different distributional assumptions about unobserved persistent heterogeneity.

The paper is organized as follows. First, in order, we discuss the assumptions underlying the random effects, semiparametric, and fixed

effects models. Second, we demonstrate the utility of using complementary statistical strategies using the 1958 Wolfgang Philadelphia cohort data.

2. MODELING CONTINUITY AND CHANGE IN CRIMINAL OFFENDING

2.1. The Random Effects Probit Model

In order to sort out the effects of state dependence and individual heterogeneity in a longitudinal sequence of binary outcomes, it is necessary to observe the set of outcomes and the sources of the heterogeneity—the time-constant factors which might contribute to the offending behavior. Unfortunately, in criminology research, it is not possible to measure all of the time-constant factors which contribute to an individual's offending behavior. Models which do not include relevant measures of individual heterogeneity are subject to the problem of omitted variable bias. As a result, one might infer that the state dependent effect (a causal impact of prior offending) is larger than it is in reality. In response to this problem, Nagin and Paternoster (1991) suggested that methods be used to control for *unobserved* individual heterogeneity. These methods should allow for more valid inference about the relative plausibility of the state dependent and individual heterogeneity explanations described above.

Nagin and Paternoster (1991) first proposed using the so-called random effects probit model to address this problem. This model divides the error term into two components, a random error component and an individual-specific, time-constant component. This can be represented by the following equation:

$$y_{it}^{*} = \delta(t) + \gamma y_{it-1} + \varepsilon_{it}, \qquad \varepsilon_{it} = \alpha_i + \nu_{it} \tag{1}$$

where y_{it}^{*} is a latent variable for the ith person in the panel, $\delta(t)$ is a coefficient that captures the amount of change in y_{it}^{*} associated with a unit change in time, γ is the parameter which measures the state dependent effect, ε_{it} is the overall error term, α_i is the individual specific component, and ν_{it} is the purely random normal disturbance with zero mean and unit variance. The observed outcome, y_{it} is equal to 1 if the latent construct $y_{it}^{*} > \tau$ and 0 otherwise (τ is an arbitrary cutoff point). This equation explicitly recognizes that some unmeasured elements of the model will not be truly random but will instead be fixed for a given individual over time (reflected in α_i). With this specification in hand, the crucial task is to estimate the parameters of the model such that the fixed component of the error term (α_i) is not allowed to bias the estimate of the state dependent

effect, γ. This task is accomplished by assuming that the fixed component of the error term is associated with a known mixing distribution.[5]

More specifically, Hsiao (1986, p. 169) asserts that the probability mass function for this model using the set of T binary outcomes, y_{it}, associated with an individual whose unobserved heterogeneity is captured by the variable α_i, is given by

$$pr[y_i|\alpha_i] = \prod_{t=1}^{T} F[y_{it}|y_{it-1}, \alpha_i]$$
$$= \prod_{t=1}^{T} \Phi[(2y_{it} - 1)(\delta(t) + \gamma y_{it-1} + \alpha_i)] \quad (2)$$

where $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. Since α_i is unobserved, this model cannot be estimated as written. Statisticians have found that they can address this problem by integrating α_i out of the multivariate outcome distribution. This integration can be performed provided we make the following two assumptions: (1) α_i is drawn from a normal mixing distribution with zero mean and variance, σ^2, and (2) α_i does not contribute to the lag of the first observation in the sequence of observations, y_{i0}.[6] The latter assumption is known as the "initial conditions assumption" (see, e.g., Nagin and Paternoster, 1991; Nagin and Farrington, 1992a) and simply means that the process must be observed at its beginning. When both assumptions are met, we can consistently estimate δ and γ using the following probability mass function (Hsiao, 1986, p. 169):

$$pr[y_i|\alpha_i] = \prod_{t=1}^{T} \int \Phi[(2y_{it} - 1)(\delta(t) + \gamma y_{i,t-1} + \alpha_i)]\, d(\alpha_i) \quad (3)$$

The advantage of this procedure is that we are able to integrate the individual heterogeneity, α_i, out of the probability mass function. As a practical matter, we can then interpret our probit coefficients as though the individual heterogeneity did not exist. The quadrature methods described by Butler and Moffitt (1982) are used to evaluate the integral.

[5]The mixing distribution refers to the underlying distribution of individual heterogeneity in the population. Specifically, it refers to the particular functional form which the researcher has chosen to represent this unobserved feature of the population from which the sample was drawn. In the current case, the mixing distribution is continuous and normal. Intuitively, the mixing distribution can be thought of as the formula by which the unobserved time-stable traits are distributed in the population.

[6]In the current model, no other factors besides lagged y are included as explanatory factors. In more sophisticated models with exogenous right-hand side variables, researchers must also assume that α_i is not correlated with any of these exogenous factors.

The assumption about initial conditions is arcane and can be rather difficult to understand in terms of the statistical model. Those analyzing trend data must, nonetheless, directly confront this issue because violation of this assumption can lead to a positively biased estimate of the coefficient on the lag of the dependent variable, γ. Such bias could lead to incorrect inference about the importance of state dependence (Hsiao, 1986). In this paper, we sidestep the issue of initial conditions in order to concentrate on the concerns raised by the first assumption concerning the normal distribution of individual heterogeneity in the population. We do this by performing our analysis on a dataset in which the initial conditions assumptions are satisfied, the 1958 Philadelphia cohort collected by Wolfgang. The sample has arrest data starting at birth. Of course, much criminological research is performed on datasets for which this assumption is in fact not satisfied. Because of the importance of the initial conditions issue, we deal with it separately in another paper (Brame *et al*., 1998).

The assumption that the random effects, α_i, are normally distributed provides the basis for a convenient algorithm to evaluate the integral in Eq. (3). Assuming that $\alpha_i \sim N(0, \sigma^2)$, rather than some other distribution, is a technical advantage because the methods for integration under the assumption that heterogeneity is normally distributed are well developed (see, e.g., Greene, 1997, pp. 896–898). Yet it is possible that this normality assumption is violated in most criminology research. It is axiomatic that criminal offending follows a skewed distribution (Nagin and Land, 1993). Hence, it is possible that criminal propensity is similarly skewed.[7]

The fact that an assumption is violated, however, does not necessarily mean that the model cannot provide a reasonably good estimate of the parameters of interest. Indeed, it was generally thought that the specification of the distribution of individual heterogeneity in the population would not have a large impact on estimates of structural parameters (Heckman and Singer, 1984). On the other hand, Maltz (1994) has convincingly argued that this type of general belief in the robustness of the statistical model can lead to an undesirable situation in which individual researchers never check the validity of their modeling assumptions.

In 1984, Heckman and Singer challenged the general belief in the distributional robustness of random effects models when they studied this problem carefully in the context of single-spell duration dependence models.[8]

[7]But Osgood and Rowe (1994) have observed that a skewed criminal offending distribution does not necessarily imply a skewed criminal propensity distribution.

[8]These models try to predict the amount of time spent in a particular spell, such as a spell of unemployment. Researchers have tried to explain why people who have been unemployed for a long time seem to stay unemployed. One explanation (state dependence) is that something about unemployment causes an individual to become more unemployable, and hence it becomes increasingly unlikely that the individual will leave the spell of unemployment. The second explanation (individual heterogeneity) simply says that not all people are the same,

They integrated the random effects, α_i, out of the model based on the assumption that they were drawn from one of three distributions in the population: (1) normal, (2) log normal, and (3) gamma.[9] In each case, the coefficients of substantive interest led to dramatically different conclusions. As they pointed out, "*(A)d hoc* specifications of model unobservables critically affect the empirical estimates achieved from 'structural' duration models" (Heckman and Singer, 1984, p. 276).

Nagin and Paternoster (1991, p. 183) responded to this specific result by warning in the conclusion to their paper that the "(r)esults (using the random effects model) can be very sensitive to distributional assumptions about the nature of the heterogeneity." Subsequently, Nagin and Land (1993) implemented the semiparametric model proposed by Heckman and Singer (1984) and described below. Although a random effects model, their semiparametric approach imposes no parametric assumptions on the functional form of the individual heterogeneity. Instead, the mixing distribution can be viewed as multinominal (i.e., a categorical variable). The only restriction placed on the mixing distribution is that it is discrete rather than continuous. Each category within this multinominal mixture can then be viewed as a point of support for the distribution of the α_i. Essentially what the model does is estimate a separate intercept, or point of support, for as many distinct groups as can be identified in the data. Within this framework, each individual has some nonzero probability of being assigned to each point of support.

Although statisticians often prefer nonparametric methods to parametric ones *a priori*, we believe that the dismissal of random effects models on these grounds may have been premature. Rather than asking if different functional form assumptions provide different answers using the same data, we think the more interesting question is whether or not the classic random effects model can provide reasonable estimates even if the true distribution is not normal.

As a first cut at this question, we performed a simulation experiment based on the process implied by Eq. (1). There were two parts to the experiment. In the first part, we generated the heterogeneity, α_i, from a highly positively skewed lognormal distribution. In the second part, we again generated the heterogeneity from a skewed lognormal distribution, but in this

and as a result, some people are more motivated, employable, etc., than others. These higher achievers will become reemployed sooner than the lower achievers. Eventually, only the low achievers are still unemployed, which explains why those with long spells are less likely to become reemployed.

[9]At present, it is necessary to develop special programs to estimate the models using assumptions other than the normal distribution. These models have more difficult likelihood functions than the traditional models that assume normality.

Table I. Monte Carlo Simulation Results

Parameter	True value	Average estimate	SD	% bias
Part 1: Log normal heterogeneity with strong positive skew				
$\delta(1)$	−3.0	−2.132	0.112	28.9
$\delta(2)$	−2.9	−2.085	0.109	28.1
$\delta(3)$	−2.8	−1.997	0.095	28.7
$\delta(4)$	−2.7	−1.899	0.097	29.7
$\delta(5)$	−2.6	1.789	0.088	31.2
γ	0.3	0.395	0.105	31.7
Part 2: Log normal heterogeneity with moderate positive skew				
$\delta(1)$	−3.0	−2.049	0.121	31.7
$\delta(2)$	−2.9	−1.956	0.106	32.6
$\delta(3)$	−2.8	−1.847	0.099	34.0
$\delta(4)$	−2.7	−1.747	0.101	35.3
$\delta(5)$	−2.6	−1.646	0.083	36.7
γ	0.3	0.311	0.131	3.7

part, the skewness of the heterogeneity was reduced in magnitude compared to that in the first part. In each case, the standard random effects probit model was used to estimate the parameters. For each part of the experiment, we generated 100 data sets, with 1000 cases in each data set followed for $T = 5$ periods of time.[10] Heckman and Singer's (1984) results implied that the performance of the standard random effects probit model will be poor when the parametric assumptions upon which it is based are violated. In examining the results of the first part of our simulation experiment, we found this to be the case. Even after controlling for unobserved heterogeneity, the random effects probit estimator yielded biased estimates for all structural parameters in the model. The second part of our experiment, however, suggests that the magnitude of the bias is related to the amount of skew in the distribution of the heterogeneity. In this part of the experiment, our parameter estimates were still biased for the time-trend effects, $\delta(t)$, that we included in the model but the estimate of the state dependent effect, γ, was very close to the value used to generate the data. The results are given in Table I and a more detailed description of our simulation protocol appears in Appendix A.

These results confirm that violation of the distributional assumptions of the random effects model can be problematic. They also suggest, however, that the model provides an accurate estimate for the state dependent

[10]Our simulation protocol generates offense frequency distributions which are highly skewed, like those found in many criminological data sets. Furthermore, we allowed for increasing rates of offending over time by incorporating a linear time trend into the data generating process.

effect when the skew is moderate. A natural way to proceed, then, would be to perform a simulation experiment over the range of skewness thought to occur "in nature." Unfortunately, we do not have a good sense of the amount of skewness in criminal propensity. Heckman and Singer (1984) have proposed a very appealing strategy in the face of such ignorance. In view of the strong parametric assumption of the random effects model, Heckman and Singer (1984, p. 309) recommend that researchers explicitly compare the results of this estimator with a less restrictive, nonparametric maximum-likelihood estimator (NPMLE):

> The NPMLE can be used to check the plausibility of any particular parametric specification of the distribution of unobserved variables. If the estimated parameters of a structural model achieved from a parametric specification of the distribution of unobservables are not "too far" from the estimates of the same parameters achieved from the NPMLE proposed in this paper, the econometrician would have much more confidence in adopting the particular specification of the mixing measure. Development of a formal test statistic to determine how far is "too far" is a topic for the future.

In subsequent sections of this paper, we adopt this "compare and contrast" strategy advocated by Heckman and Singer. We propose to test the robustness of the random effects probit model on actual data by comparing the results from this model with those from the less restrictive semiparametric probit and fixed effect logit models. These latter two models should provide estimates that are less biased a priori because they provide for a more general structure for unobserved individual heterogeneity. If the random effects model provides similar answers, then we have reason to think either that the distributional assumptions are not violated or that the violations do not affect the integrity of the random effects model.

It could be argued that, regardless of the results, we should automatically move to these less restrictive models and discard the more restrictive random effects model. This preference is hindered by the fact that these less restrictive models impose their own constellation of costs in terms of computational complexity, efficiency loss, and/or modeling limitations which might prevent their widespread application.[11] We believe that the best approach would be to estimate two or more alternative models simultaneously. Such an approach would avoid overreliance on a particular model's assumptions or conceptualization of unobserved heterogeneity. Further, if different estimators yield similar answers, researchers can be more certain that the results are not an artifact of a particular model specification, and, no longer quite so worried about bias, they can confidently rely on the

[11]Nagin and Land are currently developing "canned" software which will facilitate easier use of these methods with count data.

results from the most efficient model specification. If the results are very different, the researcher can decide to use the results from the more conservative model with fewer assumptions. In Section 4, we discuss a formal method which will help researchers decide which model to use when the results are ambiguously different.

In the next two sections, we describe the semiparametric *probit* model and the fixed effects *logit* model with lagged dependent variables. We believe that neither model has been previously estimated in the criminological literature. The semiparametric probit model is still a random effects model, but it no longer restricts the mixing distribution to be normal. Instead it assumes that the distribution of unobserved persistent heterogeneity is discrete. The fixed effects logit model makes no assumptions at all about the distribution of individual heterogeneity in the population.

2.2. The Semiparametric Probit Model

In this section, we consider another modeling strategy that can be brought to bear on the problem of analyzing the relationship between criminal offending activity at different points in time. This approach, initially developed for criminology by Nagin and (Nagin and Land, 1993; Land and Nagin, 1996; Land *et al*., 1996), uses a finite mixture of discrete probability distributions to study longitudinal sequences of *event count* outcomes. After considering some of the basic features of this approach, we extend Nagin and Land's finite mixture strategy to the problem of investigating longitudinal sequences of *binary* outcomes. This allows us to compare the model directly with the random effects probit and avoids the additional complexity of event count data.

In the conventional random effects probit model we are concerned about whether we can obtain valid inferences about the relationship between offending behavior at different discrete time periods when the distribution of stable individual differences is not normal. As Nagin and Land (1993) have pointed out, estimators based on finite mixtures do not make assumptions about the shape of the distribution of individual heterogeneity in the population, sometimes called the mixing distribution. Consequently, inferences based on the finite mixture estimator may be more accurate than inferences based on the standard random effects probit estimator when the distribution of stable individual differences is not normally distributed.

The probability mass function for the finite mixture model for a sequence of binary outcomes is given by

$$pr(y_i|\pi, \gamma, \delta, \alpha) = \sum_{j=1}^{K} \pi_j \left[\prod_{t=1}^{T} (y_{it} pr(y_{it}|\alpha_j) + (1 - y_{it})(1 - pr(y_{it}|\alpha_j))) \right] \quad (4)$$

where π_j is the unconditional probability of membership in point of support (j). The probability mass for individual i, at time t, and point of support j is given by

$$pr(y_{it}|\alpha_j) = \Phi(\alpha_j + \delta(t) + \gamma y_{it-1}) \tag{5}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, α_j is the intercept for point of support j, δ is a vector of $t-1$ binary time period indicators that uniquely identify each specific discrete period in the data for each individual, and γ captures the effect of the previous period's activity on the contemporary binary outcome.

While the finite mixture model described above provides a very flexible way to approach the problem of controlling for unobserved heterogeneity, it also has some weaknesses. First, if the true mixing distribution is continuous, then a finite mixture estimator is a clear misspecification. Nevertheless, as Nagin and Land (1993) have noted, even if the true mixing distribution is continuous, we can still obtain useful inferences from finite mixtures because they will approximate a continuous mixture in large samples and the approximation will improve as the number of observations grows. With finite numbers of observations, however, it is important to keep in mind that the estimator only approximates a continuous mixture. Thus, the estimator assumes that, within a discrete point of support of the mixing distribution, all individuals are exchangeable with each other. To the extent that this is not literally true, there will be some bias in the parameter estimates obtained with this approach.

A second weakness becomes prominent in the case where longitudinal sequences of binary outcomes are under investigation. In our experience, we have not been able to identify parameter estimates associated with finite mixtures unless we have data (1) with at least seven discrete periods of outcomes and (2) where there are some periods in which the prevalence of event occurrence is greater than 12–15%. Minimum conditions for identification still need to be derived; thus, when the two above conditions hold, we know the model is identified but we are uncertain about what the minimum conditions for identification actually are. We are continuing to work on this problem.

2.3. The Fixed Effects Model

In this section, we present the fixed effects logit model for the analysis of offending participation. Unlike the previous two models, the fixed effects model no longer conceptualizes the unobserved heterogeneity as part of the error term. Instead, the fixed effects approach models unobserved heterogeneity explicitly as a time constant intercept for each individual in the

sample. The model can be written as follows:

$$y_{it}^{*} = \alpha_i + \gamma y_{it-1} + \nu_{it} \tag{6}$$

where α_i is no longer an error component but rather a separate intercept for each individual. This is known as an autoregressive logistic regression model.

In general, this intercept is estimated for each individual. The intercept will absorb all individual-specific factors which are constant over each wave of the panel data. As a result, any significant state-dependent impact for the model must be, by definition, independent of the unobserved heterogeneity. This type of unequivocal control for unobserved heterogeneity as it is revealed in the data would provide a very strong test of the state dependent hypothesis.

Implementation of this model is made difficult by the discrete nature of the dependent variable and the limited number of periods usually available in panel data. To see why, consider the likelihood function for the fixed effect logit model, as given by Maddala (1987, p. 328):

$$pr(y_{it} | \alpha_i, y_{it-1}) = \frac{\exp(\alpha_i + \gamma y_{it-1})}{1 + \exp(\alpha_i + \gamma y_{it-1})} \tag{7}$$

where y_{it} is 1 if the ith individual commits a crime during this period and 0 otherwise, α_i is a time constant intercept for each individual i which reflects unmeasured sources of heterogeneity, and γ is the parameter that captures the effect of prior offending activity. Estimating this model by maximizing the likelihood function based on Eq. (7) will not provide consistent estimates of α_i unless the panel has a large number of time periods because each period provides a unique observation of α_i. Most microlevel criminological data do not have the required large number of time periods. As a result, estimates of γ, which are conditional on α_i, are also not consistent in cases with small T. In a continuous framework, this problem might be dealt with by differencing the data or taking differences from the within individual mean. Such a procedure will eliminate α_i for each individual from the model when the equations for Y^* in adjacent periods are subtracted from one another and allow for consistent estimates of γ. However, this procedure is not possible in the discrete, nonlinear framework.

As a nonlinear alternative, Chamberlain (1984) suggests working with a conditional likelihood function. The appropriate conditioning quantities are the total number of offenses, Σy_{it}, and the offense status in the last year of observation, y_{iT}. Initial conditions are dealt with by conditioning on the first year of observation, y_{i1}.[12] Conditioning on these three quantities

[12]In our framework, initial conditions are not a concern since we will be using a data set in which initial conditions are satisfied. This redundancy should not affect the results.

together sweeps out the fixed effects and allows for the consistent estimation of γ. Formally, he writes the conditional probability mass function as

$$pr\left(y_i \left| \sum_{t=1}^{T} y_{it}, y_{i1}, y_{iT} \right.\right) = \frac{\exp(\gamma \sum_{t=2}^{T} y_{it} y_{it-1})}{\Sigma_{d_k \in B_i} \exp(\gamma \sum_{t=2}^{T} d_t d_{t-1})} \tag{8}$$

This equation is complicated enough to require detailed explanation. Like the other models we consider in this paper, the probability of observing the ith individual's joint distribution of the outcomes $y_i = (y_{i1}, \ldots, y_{iT})$ is the central focus. Unlike the other models, however, this joint probability is conditional on three quantities: (1) the sum of the outcomes over all periods in the sequence (Σy_{it}), (2) the outcome in the first period of the sequence (y_{i1}), and (3) the outcome in the last period of the sequence (y_{iT}). We refer to the set of all possible combinations of these three quantities as a triple, B_j, where $j = 1, 2, \ldots, J$ [with $J = (T-1) \times 4$]. Each of the J triples is comprised of all joint outcomes, ($y_{i1}, \ldots, y_{iT}$), which share the same values for each of the three quantities in the triple. Each such longitudinal sequence is denoted $d_k = (d_1, \ldots, d_T)$, for the $k = 1, 2, \ldots, K$ possible sequences that comprise the triple, B_j. Note that not all triples have the same number of sequences, so K is free to vary across each of the J triples.

The actual equation is comprised of two parts. The kernel of the numerator is the sum of the products of adjacent outcomes within the sequence for the ith individual multiplied by γ. The conditional maximum-likelihood estimate of γ captures the effect of prior offending activity on future offending activity. The denominator is the sum of the exponentiated sums of the products of adjacent outcomes multiplied by γ for all K longitudinal sequences that comprise the triple, B_j, to which the numerator sequence belongs.

For example, suppose we wanted to estimate the joint probability of the outcome [0, 0, 1, 1] in a $T = 4$ period panel, where 0 means that the individual did not offend in the period and 1 means that the individual did offend in the period. We first need to identify the triple to which this outcome set belongs. For this outcome, $\Sigma_t y_{it} = 2$, $y_{i1} = 0$, and $y_{i4} = 1$. There is only one other outcome which belongs to this triple: [0, 1, 0, 1]. The sum of the product of the adjacent terms for [0, 0, 1, 1] is just $(0 \times 0 + 0 \times 1 + 1 \times 1) = 1$. The sum of the product of the adjacent terms for [0, 1, 0, 1] is just $(0 \times 1 + 1 \times 0 + 0 \times 1) = 0$. Thus, an individual with a [0, 0, 1, 1] outcome contributes $\exp(\gamma \times 1)/[\exp(\gamma \times 0) + \exp(\gamma \times 1)]$ to the conditional likelihood function. A detailed explanation for the case where $T = 5$ is presented in Appendix B.[13]

[13]We have developed SAS computer programs which will estimate fixed effects logit models of this form for up to 10 waves of panel data. We will be happy to make these programs available upon request.

One unique feature of this model is that not all individuals will contribute to the conditional likelihood function. Only in the case where γ does not cancel out of the conditional likelihood function associated with an individual's outcome sequence will that individual contribute to the likelihood function. For example, in the case with four periods, only those individuals who offend according to the following 4 patterns (of a total of 16 offending patterns) contribute to the likelihood function: [1100], [0011], [1010], and [0101]. In Corcoran and Hill's (1985) data with five time periods, only 7% of a sample of 1251 observations or 89 people actually contribute to the likelihood function. Although Chamberlain claims that the results are consistent regardless of the number of observations that actually contribute to the likelihood function, Maddala (1987) nonetheless points to this small sample problem as one of the weaknesses of the method. At the very least, more cases should lower the standard error of the estimates.[14]

The main weakness in this model is that the conditioning exercise which eliminates the fixed effect does not work when other variables are present. Since other factors can be included in the absence of the lagged endogenous variable, one is faced with a choice: either include other factors and omit the lagged endogenous variable or include the lagged endogenous variable and omit all other factors. The other two models described in this section do not force the researcher to make such a choice. Therefore, while this model does an excellent job of bluntly controlling for unobserved heterogeneity by including a fixed effect in the original specification (and then eliminating the need to actually estimate the fixed effect), its usefulness for identifying the effects of prior offending (state dependence) in the presence of heterogeneity in criminal propensity is limited relative to the other methods presented.

One very useful feature of the fixed effects model, however, is that it can provide an upper bound on the state dependent effect, before other time varying factors are included in the model. This simple autoregressive model can also be estimated with the random effects and semiparametric models to determine the impact of the more restrictive assumptions about the individual heterogeneity made by these latter two models.

The paper proceeds by estimating just such a simple autoregressive model on the 1958 Philadelphia cohort data with all three methods described in this section. If the results are similar across all three models, then we have reason to believe that the parametric assumptions are not very restrictive. We also estimate a slightly more complicated model with an intercept and six time dummies. We omit the time dummy for the first period. The time dummies allow for global shifts in offending propensity over

[14]One way to increase the number of cases is to increase the number of time periods.

the period of observation (Gottfredson and Hirschi, 1990). Although the above discussion did not include time-varying explanatory variables besides lagged offending, the random effects and semiparametric models generalize easily. This comparison will help to determine if the comparison made in the simple case generalizes to more realistic models.

3. DATA

We base our analysis on data from the 1958 Philadelphia birth cohort study (Tracy *et al*., 1990). The data are comprised of the 13,160 males who were born in Philadelphia in 1958. We characterize the longitudinal offending sequence of each of these males with a $1 \times T$ vector, where $T = 7$,[15] $y_i = [y_{i,t=1}, y_{i,t=2}, \ldots, y_{i,t=7}]$. Each of the $y_{i,t}$ is coded 1 if the individual has a police contact for criminal activity during that period and 0 otherwise. We define the seven periods in the following way: Period 1 = age 6 to 8, Period 2 = age 9 to 11, Period 3 = age 12 to 14, Period 4 = age 15 to 17, Period 5 = age 18 to 20, Period 6 = age 21 to 23, and Period 7 = age 24 to 26. For each individual the longitudinal offending sequence, y_i, identifies the specific periods within the seven-period panel in which that individual had at least one police contact for criminal activity.

Figure 5 presents the proportion of the 13,160 males who had at least one police contact for each of the seven periods. Corroborating much prior research (see, e.g., Hirschi and Gottfredson, 1983; Blumstein *et al*., 1986; Nagin and Land, 1993), this descriptive analysis suggests that participation in criminal activity starts at very low levels in childhood and rises steadily to a peak of nearly 25% in the 15–17 age band. After that point, participation steadily declines. In the next section, we turn to a description of our analysis results.

4. RESULTS

4.1. Normal Random Effects Probit and Semiparametric Probit

For this paper, we are interested primarily in the following substantive question: What is the relationship between prior and future criminal offending when the effects of stable individual differences have been taken into account? A finding that prior offending has no effect on future offending once unobserved persistent heterogeneity is controlled would be difficult to reconcile with a pure state dependence explanation. Similarly, a finding that

[15]Past research used a maximum of five periods. We used seven periods, which we felt increased the flexibility of this approach. Sensitivity analysis suggests that the following analysis is robust to the number of periods used.

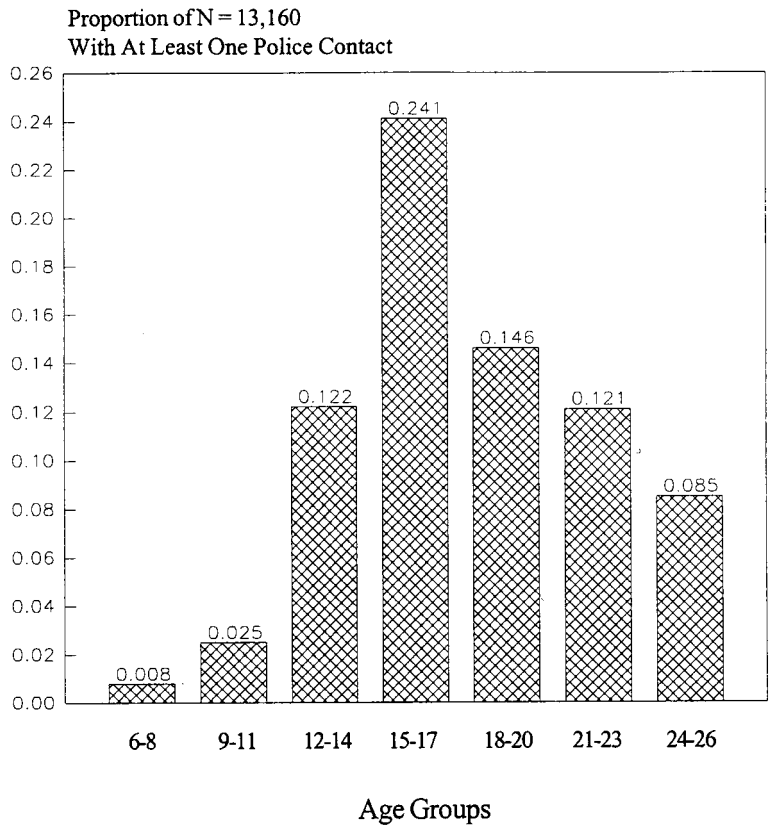

Fig. 5. Prevalence of offending by age for 1958 Philadelphia birth cohort males ($N = 13{,}160$).

prior offending has a substantial effect on future offending even after controlling for criminal propensity would be equally difficult to square with a pure population heterogeneity explanation. We are also interested in whether different methods for answering this question lead us to different sets of conclusions.

To address these questions, we estimate the parameters of five models. Model 1, presented in columns 2 and 3 in Table II, is of the form

$$y_{it}^{*} = \delta_0 + \gamma y_{it-1} + \varepsilon_{it} \tag{9}$$

where the error term, ε_{it}, is given by $\varepsilon_{it} = \alpha_{it} + \nu_{it}$, with $y_{it} = 1$ when $y_{it}^{*} > 0$ and $y_{it} = 0$, otherwise. The parameter estimates associated with this model

Table II. Investigation of State Dependence with Normal Random Effects Probit Model

	Model 1. No trend controls		Model 2. Trend controls	
Parameter	Parameter estimate	$\|t\|$ ratio	Parameter estimate	$\|t\|$ ratio
Overall intercept	–1.514	160.43	–2.924	64.68
Time trend effects				
$\delta(1)$			0	—
$\delta(2)$			0.517	10.60
$\delta(3)$			1.463	32.42
$\delta(4)$			1.961	43.39
$\delta(5)$			1.390	30.37
$\delta(6)$			1.314	28.90
$\delta(7)$			1.065	23.27
State dependent effect				
γ	1.052	48.13	0.611	25.23
$\text{Cov}(\varepsilon_{it}, \varepsilon_{it+1})$				
ρ	0.120	7.16	0.331	18.88
log-likelihood	–27,802.03		–25,312.56	

include a population intercept term, δ_0, an estimate of the effect of offending in y_{it-1}, γ, and an estimate of the correlation between the collection of ε_{it} for the same individual, ρ. The maximum-likelihood estimate of the correlation parameter, ρ, is calculated by dividing the variance of α_i by the variance of ε_{it} (see Nagin and Paternoster, 1991; Nagin and Farrington, 1992a).

The results associated with Model 1 point to several conclusions. First, the estimate of ρ (.120) is relatively small but significantly different from zero. This provides evidence of persistent unobserved heterogeneity. Second, the estimate of γ is positive, significantly different from zero, and substantively quite large (1.052). The best way to interpret this estimate is to compare $pr[y_{it}|\alpha_i,\ y_{it-1}=1]$ to $pr[y_{it}|\alpha_i,\ y_{it-1}=0]$. In this case, $pr[y_{it}|\alpha_i,\ y_{it-1}=1]=\Phi[\delta_0+\gamma]=0.322$, while $pr[y_{it}|\alpha_i,\ y_{it-1}=0]=\Phi[\delta_0]=0.065$. Thus, individuals who offended in the previous period are almost five times more likely to offend in the current period as individuals who did not controlling for individual differences.

Since Fig. 5 shows a pronounced time trend, we speculated that it would be important to incorporate some form of time trend control. Model 2 is simply Model 1 with a dummy variable for $T-1$ time periods in the analysis (the effect of the dummy variable for $t=1$ is absorbed into the intercept term). The time dummies provide a very general kind of control over population shifts in the probability of offending over time. Table II presents the parameter estimates associated with Model 2.

The results associated with Model 2 produce two important conclusions. First, the estimated heterogeneity in the $\alpha(i)$, as measured by ρ, apparently increases to 0.331 from 0.120 as a result of controlling for time trends. Second, the estimate of γ is somewhat lower in Model 2 than in Model 1 (0.611 vs 1.052). As an indication of the effect that this estimate implies, we compare predicted probabilities conditional on a particular time period. First, we calculate $pr[y_{it=4} = 1|\alpha_i, y_{it=3} = 1] = \Phi[\delta_0 + \delta_4 + \gamma] = 0.362$. Next we calculate $pr[y_{it=4} = 1|\alpha_i, y_{it=3} = 0) = \Phi[\delta_0 + \delta_4] = 0.168$. The estimates associated with Model 2, therefore, suggest that, at $t = 4$, individuals who offended in the previous period are twice as likely to offend in the current period as individuals who did not. In short, the effect of introducing a control for time trend is to attenuate the estimated effect of prior offending activity.[16]

We now turn to an investigation of the semiparametric random effects probit model. Recall that this model has the same form as the normal random effects probit model with one exception: we now assume that the α_i are drawn from $j = 1, 2, \ldots, K$ discrete categories. Our specification allows us to estimate each of the α_j and the proportion of the population to which each α_j estimate applies. Table III presents Model 3, which is a semiparametric random effects probit model with no time controls. In this model, we were able to identify two discrete categories for the α_j. Thus, we have $\alpha_{j=1}$ and $\alpha_{j=2}$. The parameter estimates further suggest that the estimate of $\alpha_{j=1}$ applies to 62.3% of the population, while the estimate of $\alpha_{j=2}$ applies to the remaining 37.7%. Interestingly, the estimated effect of prior offending in this model is virtually identical to the effect that we estimated for the normal random effects probit Model 1 (1.035 vs 1.052).

Table III also presents the estimates associated with Model 4. This specification is similar to Model 3 with the exception of allowing for a time trend in the probability of offending activity. Under Model 4, we were able to identify three discrete categories for the α_j.[17] Thus, the estimate of $\alpha_{j=1}$

[16]For comparison, we calculated these probabilities for other time periods as well.

	T_2	T_3	T_5	T_6	T_7
Probability of offending in period T_j					
Offended in previous period	0.036	0.198	0.178	0.159	0.106
Did not offend in previous period	0.008	0.072	0.063	0.054	0.032
Ratio of probabilities	4.50	2.75	2.82	2.94	3.31

[17]This yields an interesting source of common ground between the normal and the semiparametric random effects probit models. Recall that in the normal random effects probit model, controls for time trends significantly increased our estimate of ρ. In the semiparametric random effects probit model, we were able to move to a $K = 3$ point of support model when we controlled for time trends, whereas only a $K = 2$ point of support model was estimable when time trends were ignored. In both models, then, we have evidence of increased unobserved heterogeneity when time trends are held constant.

Table III. Investigation of State Dependence with Semiparametric Random Effects Probit Model

	Model 3. No trend controls		Model 4. Trend controls	
Parameter	Parameter estimate	$\|t\|$ ratio	Parameter estimate	$\|t\|$ ratio
Random effects				
$\alpha_{(j=1)}$	–1.796	36.33	–3.235	27.08
$\alpha_{(j=2)}$	–1.050	27.85	–2.240	7.15
$\alpha_{(j=3)}$			–1.609	8.11
Support mass				
$\pi_{(j=1)}$	0.623	13.30	0.663	6.71
$\pi_{(j=2)}$	0.377	—	0.251	5.21
$\pi_{(j=3)}$			0.086	—
Time trend effects				
$\delta(1)$			0	—
$\delta(2)$			0.509	10.39
$\delta(3)$			1.449	32.04
$\delta(4)$			1.944	41.79
$\delta(5)$			1.380	29.04
$\delta(6)$			1.299	27.84
$\delta(7)$			1.049	22.33
State dependent effect				
γ	1.035	49.25	0. 608	23.72
log-likelihood	–27,792.80		–25,302.06	

applies to 66.3% of the population, while the estimates of $\alpha_{j=2}$ and $\alpha_{j=3}$ apply to 25.1 and 8.6% of the population, respectively. The most interesting aspect of Model 4, however, is its agreement with the results of the normal random effects probit model with time controls (Model 2). Specifically, the estimated effect of prior offending in Model 4 is virtually identical to the estimated effect of prior offending in Model 2 (0.608 vs 0.611). For this specific analysis, then, the normal and semiparametric random effects probit models lead to the same substantive conclusions about the relationship between prior and future criminal behavior after controlling for stable unobserved individual differences.

4.2. Model Comparison

What we have found thus far is that the semiparametric probit model yields estimates that square precisely with the standard random effects probit estimator that makes strong parametric assumptions about the distribution of unobserved persistent heterogeneity. The problem of deciding whether to select the semiparametric effects probit model or the standard

normal random effects probit model may be important in some situations. As noted earlier, the basic framework for this decision is established by Heckman and Singer (1984). In the case when the estimates obtained from the semiparametric model are similar to those obtained from the normal random effects model, they advocate choosing the random effects model because of its more efficient estimates and more parsimonious structure. In the case when the estimates differ, Heckman and Singer advocate adopting the semiparametric estimator because it approximates any distribution of unobserved heterogeneity. Since the normal random effects probit model does not have this virtue, it seems quite sensible to place more confidence in the semiparametric estimator when the two estimators lead to different conclusions.

This discussion highlights the natural tension between the normal random effects model and the semiparametric estimators. The normal model is more parsimonious but it also makes stronger assumptions about variables that cannot be observed. The semiparametric model does not make the same strong assumptions about unobservable variables, but it requires estimation of more parameters, is less efficient, and it is somewhat more difficult to interpret. In light of this tension, it would be useful to have a formal model selection strategy which would provide researchers with guidelines about when to choose one method over the other when the estimates are neither obviously similar (as in our case) nor obviously dissimilar. This need is acknowledged by Heckman and Singer, who state that "(D)evelopment of a formal test statistic to determine how far (apart) is "too far" is a topic for the future (Heckman and Singer, 1984, p. 309).[18] In what follows we present a formal test which should help researchers choose between the semiparametric and the normal random effects methods.

The main problem is that the normal probit and semiparametric models are nonnested, meaning that one model is not a more general specification of the other. Model selection strategies for frequentist (a.k.a. classical) statistical comparisons of nonnested models are not well developed. Bayesian model selection strategies, on the other hand, are better developed but they are somewhat controversial [see, for example, the recent exchange between Raftery (1996) and Gelman and Rubin (1996)]. Furthermore, all of the statistical analyses presented in this paper are based on frequentist, not Bayesian, ideas. Despite these issues, we believe that there is considerable value in using Bayesian model selection strategies for this problem. Here, we explore one approach that is easy to implement and interpret.

The selection strategy we propose involves the calculation of the so-called Bayesian information criterion (BIC) (see, e.g., Schwarz, 1978; Kass

[18]We are unaware of any recent attempts by these researchers to propose a formal test statistic.

and Raftery, 1995). This approach is rare in criminology but not unprecendented. Land *et al*. (1996) used the BIC to identify the appropriate number of components in a mixing distribution of a semiparametric Poisson model. What makes our problem different is that we are not comparing two semiparametric models, but rather we are comparing a semiparametric model with a normal random effects model.

To illustrate this approach, suppose that we have two models in the model space. For convenience, we denote these two models M_1 and M_2, respectively, and the data on which the parameters of the model are estimated is denoted D. The posterior probability for M_1 is given by

$$p(M_1|D) = \frac{p(D|M_1) * p(M_1)}{[p(D|M_1) * p(M_1)] + [p(D|M_2) * p(M_2)]}$$

A similar operation is required for calculating $p(M_2|D)$. Wasserman (1997, pp. 6–8) details the calculations that are required for this equation. Unfortunately, some of these calculations are quite difficult. As Kass and Raftery (1995) show, a useful approximation—when one is willing to assign equal *a priori* probability to each of the two models—can be obtained by

$$p(M_1|D) \approx \frac{\exp(\hat{q}_1)}{\exp(\hat{q}_1) + \exp(\hat{q}_2)}$$

$$p(M_2|D) \approx \frac{\exp(\hat{q}_2)}{\exp(\hat{q}_1) + \exp(\hat{q}_2)}$$

where

$$\hat{q}_1 = \hat{l}_1 - \frac{d_1}{2}\log(n)$$

$$\hat{q}_2 = \hat{l}_2 - \frac{d_2}{2}\log(n)$$

where l_j is the log of the likelihood function associated with M_j, d_j is the number of parameters estimated under M_j, and log( ) denotes the logarithm to the base e (Kass and Wasserman, 1995).

Now the assumption underlying the BIC is that the best way to select between two candidate models is to choose the one that has the highest posterior probability (when the priors are equal). From Wasserman (1997), a convenient way to investigate this is to calculate the ratio

$$B_{12} = \frac{p(M_1|D)}{p(M_2|D)}$$

and if the ratio is larger than 1.0, we know that the posterior for M_1 is

Table IV. Jeffrey's Scale of Evidence for the Interpretation of Bayes Factors [i.e., exp(BIC)][a]

$\exp(\text{BIC}) \approx B_{ij}$	Interpretation
$B_{ij} < 0.1$	Strong evidence for Model j
$0.1 < B_{ij} < 0.33$	Moderate evidence for Model j
$0.33 < B_{ij} < 1.0$	Weak evidence for Model j
$1.0 < B_{ij} < 3.0$	Weak evidence for Model i
$3.0 < B_{ij} < 10.0$	Moderate evidence for Model i
$B_{ij} > 10.0$	Strong evidence for Model i

[a]Table adapted from Wasserman (1997, p. 7).

greater than the posterior for M_2. If, on the other hand, the ratio is smaller than 1.0, we know that the posterior for M_2 is greater than the posterior for M_1. From the above discussion, we also know that this ratio is difficult to calculate. As Wasserman (1997) shows, however, it is easy to use the approximations to calculate

$$B_{12} \approx \left[\frac{\exp(\hat{q}_1)}{\exp(\hat{q}_1) + \exp(\hat{q}_2)} \right] \div \left[\frac{\exp(\hat{q}_2)}{\exp(\hat{q}_1) + \exp(\hat{q}_2)} \right]$$

and the natural logarithm of this ratio, the BIC, can be calculated by

$$\log(B_{12}) \approx \hat{l}_1 - \hat{l}_2 + \frac{d_2 - d_1}{2} \log(n)$$

which can then be exponentiated to recover B_{12}, as described by Kass and Raftery (1995). Table IV, adapted from Wasserman (1997, p. 7), presents a chart showing the interpretation of the magnitude of B_{12}. Evaluation of the BIC provides a useful strategy for guiding model selection in these sorts of situations. Specifically, Wasserman (1997, p. 14) observes that if the actual process that generates the data is in the model space, BIC evaluation will lead to the correct model choice as the sample size grows large. In addition, he notes that if the process that generates the data is included in both M_1 and M_2, the posterior probability for the more parsimonious model will approach 1.0, while the posterior probability for the less parsimonious model will approach 0.0.

We evaluated the BIC for our problem where M_1 was the semiparametric probit model and M_2 is the normal random effects probit model. After exponentiating the approximation to $\log(B_{12})$, we obtained $B_{12} \approx 0.024$. Based on Table IV, this result suggests that the posterior probability of the normal random effects probit model is much greater than the posterior probability of the semiparametric probit model. If we were forced to choose between the two specifications, then, the BIC would guide us to choose the

Table V. Investigation of State Dependence with the Conditional Fixed Effects Logistic Model[a]

	Model 5. No trend controls	
Parameter	Parameter estimate	$\|t\|$ ratio
State dependent effect		
γ	1.591	38.61
log-likelihood	–4627.28	

[a]We can transform the logistic state dependent effect so that it is measured in "probit" units by dividing it by 1.6 or 1.7 (different methodologists recommend different denominators). When we do this, we obtain approximations of 0.994 and 0.936, respectively.

normal random effects probit model over the semiparametric model. This is as expected. Recall that the estimates for the coefficient on lagged y were very similar (0.611 vs 0.608, respectively) and the normal random effects model is more efficient than the semiparametric model. Although further work on optimal strategies for model selection in this setting would be valuable, our sense is that BIC provides useful information for choosing between different estimators.

4.3. Fixed Effect Logits

Our final analysis focuses on the conditional fixed effects logit model. Table V presents the results of this analysis (Model 5). Model 5 produces one parameter estimate only, γ. As in the random effects probit models, γ captures the effect of prior offending on future offending after controlling for stable individual differences. But there are some important differences as well.

First, the conditional fixed effects logit model does not allow all individuals to contribute to the likelihood function. Instead, only individuals with outcome patterns that keep γ in the probability mass function for the outcome set, y_i, actually contribute to the likelihood function. In this case, only 2,547 of the 13,160 individuals in the analysis had outcomes that allowed them to contribute to the likelihood function. While this does not induce any bias into the estimate of gamma, there is some loss of efficiency. In exchange for this loss of efficiency, we gain very strong nonparametric control over persistent unobserved heterogeneity. Second, at its current stage of development, there is no way to introduce any controls for time trends (or, for that matter, anything else) into the model. Thus, with a fixed effects logit specification, we are confined to an analysis that does not include any covariates except prior offending activity.

The results obtained from estimating γ in Model 5 suggest, once again, that there is a pronounced relationship between prior and future criminal behavior once stable individual differences have been controlled. Since γ is measured in logistic distribution units rather than normal distribution units, its magnitude is not directly comparable to the estimates obtained in the probit models. If we divide the logit-based estimate of γ by 1.6, however, we get an approximation to the probit-based estimate of γ (see, e.g., Amemiya, 1981). In this case, the transformation yields an approximation of $\gamma \approx 0.994$. Since Model 5 does not include controls for time trends, we compare it to the normal random effect probit (Model 1) and the semiparametric probit (Model 3) ($\gamma = 1.052$ and $\gamma = 1.035$, respectively). All three of these models converge on virtually the same conclusion about the strength of the relationship between prior and future criminal behavior net of stable individual differences. Unfortunately, there is no formal test by which we can compare the three models.[19] However, we believe that the fixed effect model provides a nice baseline upon which to establish the other two models, which have greater flexibility. If the semiparametric and fixed effect models are very similar in the simplest models, we then have reason to believe that the semiparametric model's assumption about a discrete mixing distribution is not binding. This should not change as the model specification becomes more complex. In a sense, this comparison goes one step beyond Heckman and Singer, because it provides a way to loosen one additional assumption.

5. CONCLUSIONS

There is an ongoing stream of research initiated in 1991 by Nagin and Paternoster which attempts to measure the role of state dependence in the offending process, while controlling for observed and unobserved persistent heterogeneity. This research has utilized the random effects probit model, which the authors noted at the time could be highly sensitive to distributional assumptions about the mixing distribution (the way unobserved criminal propensity is distributed in the population). This observation had been originally made by Heckman and Singer (1984). While Heckman and Singer proposed alternative, semiparametric models, now estimated in the field of criminology by Nagin and Land (1993), we do not believe that their original intent was to advocate against the use of fully parametric models. Rather, they suggested in their paper that researchers could increase their confidence in fully parametric models by comparing the results obtained

[19]The BIC will not work in this case because the fixed effects model has a logistic distribution, rather than a normal distribution like the other two models.

with less restrictive nonparametric (or semiparametric) models. This paper is based on that philosophy. We compare the results from a fully parametric random effects probit model with those from a semiparametric probit model with panel data from the 1958 Philadelphia cohort. Furthermore, we compared these two sets of results with those from a model which makes no assumptions whatsoever about the mixing distribution—the fixed effects logit model. The analysis produced several key findings.

First, and most important, all five models we estimated using the three methods suggest that there is a strong positive relationship between prior and future criminal behavior even after stable unobserved individual differences have been held constant. Several major contemporary theories of crime (see, e.g., Wilson and Herrnstein, 1985; Gottfredson and Hirschi, 1990) identify stable individual differences as the primary source of variation in criminal activity over the life span. It is difficult to reconcile such theoretical explanations with the results we obtained in this analysis. These substantive findings lead us to conclude that some kind of state dependence process is at work, at least in the 1958 Philadelphia cohort data.

Second, given the first finding, it is also important to note that the *magnitude* of the observed effect is much smaller than what we would have measured in the absence of controls for unobserved heterogeneity. We believe that this evidence provides strong support for the claim that any future work on panel models which aims to estimate accurately the magnitude of the impact of time-varying components on offending must make use of methods which control for stable unobserved differences.

Third, the use of any one method to "control" for stable unobserved differences might reasonably be viewed with some suspicion. If we had used only the normal random effects probit model, we could have been criticized for ignoring the possibility that differences in crime-proneness are not normally distributed. If we had used only the semiparametric random effects probit model, we could have been criticized for ignoring the possibility that differences in crime-proneness are drawn from a continuous rather than a discrete probability distribution. If we had used only the conditional fixed effects logit model, we could have been criticized for limiting the model unnecessarily and throwing away a large proportion of our data. Instead, we used all three methods. And all three methods generated virtually identical estimates of the relationship between prior and future criminal activity after controlling for persistent unobserved heterogeneity. In attacking the data with three approaches we are reasonably confident that the substantive conclusions discussed in the preceding paragraph are robust.

If the results had differed, the safe approach would have been to rely most heavily on the results reported with the most general model estimated. Because both the fixed effect and the semiparametric methods place fairly restrictive demands on the data (because they make fewer parametric

assumptions), researchers will occasionally find themselves unable to estimate one or both models. In the case when only the random effects model can be estimated, the results should be interpreted with a fair degree of caution. Although we have illustrated this multiple-method approach with a dichotomous dependent variable, it can be generalized to models with count data and continuous data as well. The same approach can also be taken in models without a lagged dependent variable.

Fourth, on a more technical note, our analysis suggests that controls for time trends are particularly important whenever one wishes to investigate the sources of the relationship between prior and future criminal activity. Since we are not currently able to estimate fixed effect models that allow for time trends, we are confined to random effects models. The two random effects models that we specified provided evidence that (1) time trend controls allow the estimator to identify greater levels of heterogeneity in crime proneness and (2) models with time trend controls yield attenuated estimates of the effect of prior criminal activity on future criminal activity. Clearly, then, it is possible for very general temporal shifts in the probability of offending activity to masquerade as genuine state dependence effects. Fortunately, the addition of time controls to random effects models is a very simple task.

Fifth, we discovered some reasons to believe that this pattern of results will not be obtained in every case. For example, we conducted a Monte Carlo simulation study to examine the effect of a nonnormal distribution of crime proneness on the accuracy of conclusions drawn from the normal random effects probit model. We found that increasing levels of positive skew in the distribution of crime-proneness were associated with increased bias in the estimated effect of prior criminal activity on future criminal activity. Since there is no reason to believe, *a priori*, that the results of our substantive analyses are generalizable beyond the specific data set that we used, we think that multiple-method strategies for investigating questions such as the one addressed here (as well as other questions that involve the study of longitudinal data) are necessary.

Finally, we were able to implement a formal Bayesian model comparison technique to compare the semiparametric probit model with the random effects probit model. This tool should help researchers make intelligent choices between models when the results are more ambiguous than those reported in this paper. The application of Bayesian modeling techniques to compare nonnested models is an interesting approach that should be explored further in future research.

APPENDIX A: SIMULATION PROTOCOL

We conducted two sets of simulations to evaluate the performance of the random effects probit estimator that assumes normal unobserved

heterogeneity. For both sets of simulations, we generated 100 data sets with $i = 1, 2, \ldots, N = 1000$ observations for $t = 1, 2, \ldots, T = 5$ discrete time periods. The data generating process is

$$y_{it}^{*} = \delta_t + \gamma y_{it-1} + \alpha_i + \nu_{it}$$

where the α_i were drawn from the lognormal distribution. To construct the lognormal variates, we exponentiated draws from the normal distribution. The lognormal distribution can be usefully viewed as the natural logarithm of a corresponding normal distribution with some mean, μ, and variance, σ^2.

For the first set of simulations, we used the unit normal as the corresponding distribution for the lognormal variate, α_i. In this case, the mean of a lognormal variate corresponding to a unit normal variate is $\exp((0+1)/2) = 1.649$ and the standard deviation of the lognormal variate is $(\exp(0+2) - \exp(1))^{1/2} = 2.161$. The skewness associated with α_i for the first set of simulations is given by[20]

$$(\exp(1) + 2) \times (\exp(1) - 1)^{1/2} = 6.185$$

For the second set of simulations, we used the normal with zero mean and 0.5 standard deviation as the corresponding distribution. In this case, the mean is $\exp((0+0.25)/2) = 1.133$ and the standard deviation is $(\exp(0+0.5) - \exp(0+0.25)^{1/2} = 0.604$. The skewness associated with α_i for this set of simulations is $(\exp(0.25) + 2) \times (\exp(0.25) - 1)^{1/2} = 1.750$.

We generated the ν_{it} from the unit normal distribution and simulated a linear time trend by setting $\delta_{t=1} = -3.0$, $\delta_{t=2} = -2.9$, $\delta_{t=3} = -2.8$, $\delta_{t=4} = -2.7$, $\delta_{t=5} = -2.6$. The state dependent effect, γ, was set equal to 0.3. At each time period, each observation was assigned $y_{it} = 0$ if $y_{it}^{*} \leqq 0$ and $y_{it} = 1$ if $y_{it}^{*} > 0$. In order to meet the initial conditions assumption, y_{i0} was set to zero for all $i = 1, 2, \ldots, N$ observations.

APPENDIX B: THE FIXED EFFECTS LOGIT ESTIMATOR

The heart of the fixed effects logit model involves conditioning on the group of sufficient statistics, Σy_t, y_{i1}, and y_{iT}. The conditioning collects the outcomes into sets, or triples, which share the same values of the sufficient statistics. Then, we can assign the probability of a given outcome given that

[20]Note that the skewness of any symmetric distribution is zero.

the outcome belongs to a given triple. We illustrate this procedure with a series of steps.

Step 1

It helps to identify the number of triples in the outcome set. There will be $4 \times (T-1)$ triples. For example, when $T = 5$, there will be 16 triples, or unique combinations of sufficient statistics. Then we can identify each of the triple of outcomes by its unique set of sufficient statistics. This is easiest to do in table form. Create three columns, one for each statistic—Σy_t, y_{i1}, and y_{iT}—and $4(T-1)$ rows. Each statistic has a set of possible outcomes, which need to be considered in combination with the other statistics. They are as follows: $\Sigma y_t \in \{0, \ldots, T\}$, $y_{i1} \in \{0, 1\}$ and $y_{iT} \in \{0, 1\}$, where 0 indicates no offense and 1 indicates an offense. Start with Σy_t and map the possible values of y_{i1} and y_{iT} for each sum. The triples follow some consistent patterns.

(A) $\Sigma y_t = 0$: no offenses were committed. There is only one possible combination of y_{i1} and y_{iT} which will sum to 0, that is, $y_{i1} = 0$ and $y_{iT} = 0$.

(B) $\Sigma y_t = 1$. There are three triples in this group—the triple for those outcomes which have both $y_{i1} = 0$ and $y_{iT} = 0$, the triple for those outcomes which have $y_{i1} = 0$ and $y_{iT} = 1$, and the triple for those outcomes which have $y_{i1} = 1$ and $y_{iT} = 0$. Note that there will be no triple for those outcomes which have both $y_{i1} = 1$ and $y_{iT} = 1$, since the sum of all outcomes is only 1.

(C) $\Sigma y_t = n$, where $1 < n < T - 1$. There are four possible triples of outcomes in each case—the triple for those outcomes which have both $y_{i1} = 0$ and $y_{iT} = 0$, the triple for those outcomes which have $y_{i1} = 0$ and $y_{iT} = 1$, the triple for those outcomes which have $y_{i1} = 1$ and $y_{iT} = 0$, and the triple for those outcomes which have both $y_{i1} = 1$ and $y_{iT} = 1$.

(D) $\Sigma y_t = T - 1$. As in the case where $\Sigma y_t = 1$, there are only three triples. The missing triple in this case will be the situation where $y_{i1} = 0$ and $y_{iT} = 0$, since it is not possible to have this situation and $\Sigma y_t = T - 1$.

(E) $\Sigma y_t = T$. This triple mirrors the case where $T = 0$, where instead of all zeros, there are now all ones—an offense was committed in each period. The only possible triple is the case where $y_{i1} = 1$ and $y_{iT} = 1$.

Example: $T = 5$

Σy_t	y_{i1}	y_{iT}
0	0	0
1	0	0
1	1	0
1	0	1
2	0	0
2	1	0
2	0	1
2	1	1
3	0	0
3	1	0
3	0	1
3	1	1
4	1	0
4	0	1
4	1	1
5	1	1

Step 2

In this step we want to identify the total number of outcomes in each triple. The key statistic to compute is the number of "free" offenses (f). Free offenses are the number of offenses remaining after subtracting the number of offenses that occur in either the first or the last period from the total number of offenses. For example, if $\Sigma y_t = 1$ and $y_{i1} = 1$ and $y_{iT} = 0$, there are no free offenses. However, if $\Sigma y_t = 3$ and $y_{i1} = 1$ and $y_{iT} = 0$, then there are two free offenses.

Go through the table and compute the number of free offenses. Then, in the next column, compute the number of outcomes by the following rules.

(A) All cases that have no free offenses will have only one outcome. The offenses are all assigned to either the first or the last period, so the $T-2$ middle periods will have no offenses. All cases for which $f = T-2$ will also have only one outcome because the $T-2$ middle periods must all have an offense.

(B) All other triples will have more than one outcome which satisfies the sufficient statistics. Combinatorics can help determine a priori how many possible outcomes there are. The formula is simply $(T-2)!/f!(T-2-f)!$, where f is the number of free offenses. For example, consider the case when $T = 5$, $\Sigma y_t = 3$, $y_{i1} = 1$, and $y_{iT} = 0$. There are two free offenses since one of the total three offenses is committed to occur in the first period. The remaining two

offenses can occur in any of the $T-2(3)$ middle time periods which are not determined by the sufficient statistics. So in this case, there will be $(5-2)!/2!1! = 3!/2! = 3$ possible outcomes. In fact, in the case when $T = 5$, if the triple is not limited to 1 outcome, there must be three outcomes. As shown above, if there are two free outcomes, there are three outcomes. If there is one free outcome, then once again, there will be three possible outcomes, since $3!/1!2! = 3$.

As a check on this step, count the total of outcomes to make sure they sum to the expected total, which is simply 2^T. For example, when $T = 5$, there will be $2^5 = 32$ total outcomes.

Example $T = 5$

Σy_t	y_{i1}	y_{iT}	f	No. outcomes
0	0	0	0	1
1	0	0	1	3
1	1	0	0	1
1	0	1	0	1
2	0	0	2	3
2	1	0	1	3
2	0	1	1	3
2	1	1	0	1
3	0	0	3	1
3	1	0	2	3
3	0	1	2	3
3	1	1	1	3
4	1	0	3	1
4	0	1	3	1
4	1	1	2	3
5	1	1	3	1
Total				32

Step 3

Simply calculate each of the unique outcomes for each triple. Start with the triples which have only one outcome. These are straightforward. Next, move to the triples with more than one outcome. Simply write down all possible unique orderings of the offenses for the "free" middle periods. As T gets large, this can be tedious, but knowing how many unique outcomes to expect should make the process easier. Then calculate the likelihoods according to eq. 8 in the text.

Example $T = 5$

Σy_t	y_{i1}	y_{iT}	f	No. outcomes	Outcome(s)	$\Sigma y_t y_{it-1}$	Prob(outcome$\|\Sigma y_t, y_{i1}, y_{iT}$) $= \exp(\gamma \Sigma y_t y_{it-1})/D$[a]
0	0	0	0	1	00000		1
1	0	0	1	3	01000, 00100, 00010	0, 0, 0	1[b]/D, 1/D, 1/D $D = 3$
1	1	0	0	1	10000		1
1	0	1	0	1	00001		1
2	0	0	2	3	01100, 00110, 01010	1, 1, 0	$\exp(\gamma)/D$, $\exp(\gamma)/D$, 1/D $D = 2\exp(\gamma) + 1$
2	1	0	1	3	11000, 10100, 10010	1, 0, 0	$\exp(\gamma)/D$, 1/D, 1/D $D = \exp(\gamma) + 2$
2	0	1	1	3	01001, 00101, 00011	0, 0, 1	1/D, 1/D, $\exp(\gamma)/D$ $D = 2 + \exp(\gamma)$
2	1	1	0	1	10001		1
3	0	0	3	1	01110		1
3	1	0	2	3	11100, 10110, 11010	2, 1, 1	$\exp(2\gamma)/D$, $\exp(\gamma)/D$, $\exp(\gamma)/D$ $D = \exp(2\gamma) + 2\exp(\gamma)$
3	0	1	2	3	00111, 01101, 01011	2, 1, 1	$\exp(2\gamma)/D$, $\exp(\gamma)/D$, $\exp(\gamma)/D$ $D = \exp(2\gamma) + 2\exp(\gamma)$
3	1	1	1	3	11001, 10101, 10011	1, 0, 1	$\exp(\gamma)/D$, 1/D, $\exp(\gamma)/D$ $D = 2\exp(\gamma) + 1$
4	1	0	3	1	11110		1
4	0	1	3	1	01111		1
4	1	1	2	3	10111, 11011, 11101	2, 2, 2	$\exp(2\gamma)/D$, $\exp(2\gamma)/D$, $\exp(2\gamma)/D$ $D = 3\exp(2\gamma)$
5	1	1	3	1	11111		1

[a] D is just the sum of the $\exp(\gamma \Sigma y_t y_{it-1})$ for each triple of outcomes.
[b] $\exp(0) = 1$.

ACKNOWLEDGMENTS

Support for this research was provided by a grant from the National Consortium for Violence Research, Carnegie–Mellon University, Pittsburgh, PA. We would like to thank John Engberg, Larry Wasserman, Daniel Nagin, and seminar participants at the National Consortium for Violence Research for their comments and suggestions.

REFERENCES

Agnew, R. (1982). Foundation for a general strain theory of crime and delinquency. *Criminology* 30: 47–88.

Akers, R. L. (1985). *Deviant Behavior: A Social Learning Approach*, *Third Edition*, Wadsworth, Belmont, CA.

Amemiya, T. (1981). Qualitative response models: A survey. *J. Economic Lit.* 19: 483–536.
Arbous, A. G., and Kerrick, J. E. (1951). Accident statistics and the concept of accident proneness. *Biometrics* 7: 340–432.
Banks, W. W. (1977). The relationship between previous driving record and driver culpability in fatal, multiple-vehicle collisions. *Accident Anal. Preven.* 9: 9.
Berman, M. E., Kavoussi, R. J. and Coccaro, E. F. (1997). Neurotransmitter correlates of human aggression. In Stoff, D. M., Breiling, J., and Maser, J. D. (eds.), *Handbook of Antisocial Behavior*, Wiley: New York.
Blumstein, A., Cohen, J., Roth, J. A., and Visher, C. A. (eds.). *Criminal Careers and ''Career Criminals,''* 2 Vols., National Academy Press, Washington, DC.
Brame, R., Bushway, S., and Paternoster, R. (1998). On the use of panel research designs and random effects models to investigate static and dynamic theories of criminal offending. Working paper, University of Maryland, Department of Criminology and Criminal Justice.
Butler, J. S., and Moffitt, R. (1982). A computationally efficient quadrature procedure for the one-factor multinominal probit model. *Econometrica* 50: 761–764.
Chamberlain, G. (1984). ''Panel data.'' In Griliches, Z., and Intriligator, M. (eds.), *Handbook of Econometrics*, *Vol. II*, North-Holland, Amsterdam.
Corcoran, M., and Hill, M. S. (1985). Reoccurrence of unemployment among young adult men. *J. Hum. Resources* 20: 165–183.
Fischer, M., Rolf, J. E., Hasazi, J. E., and Cummings, L. (1984). Follow-up of a pre-school epidemiological sample: cross age continuities and predictions of later adjustment with internalizing and externalizing dimensions of behavior. *Child Develop.* 55: 137–150.
Gelman, A., and Rubin, D. B. (1996). ''Comment on Raferty.'' Marsden, P. V. In (ed.), *Sociological Methodology, Vol. 25.* Basil Blackwell, Oxford.
Gottfredson, M., and Hirschi, T. (1990). *A General Theory of Crime*. Stanford University Press, Stanford, CA.
Greene, W. (1997). *Econometric Analysis*, third edition. Macmillan Publishing Company, New York.
Greenwood, M., and Wood, H. W. (1919). *A Report on the Incidence of Industrial Accidents Upon Individuals With Special Reference to Multiple Accidents*. London: British Industrial Fatigue Research Board (No. 4).
Heckman, J. J. (1981). Statistical models for discrete panel data. In Manski, C. F., and McFadden, D. (eds.), *Structural Analysis of Discrete Data With Econometric Applications*, MIT Press, Cambridge.
Heckman, J. J., and Singer, B. (1984). Econometric duration analysis. *J. Econometrics* 24: 63–132.
Hirschi, T. (1969). *Causes of Delinquency*. University of California Press, Berkeley.
Hirschi, T., and Gottfredson, M. (1995). Control theory and the life-course perspective. *Studies on Crime and Crime Prevention: Biannual Review* 4: 131–142.
Hirschi, T., and Gottfredson, M. (1983). Age and the explanation of crime. *Am. J. Sociol.* 89: 552–584.
Hsiao, C. (1986). *The Analysis of Panel Data*. Cambridge University Press, New York, NY.
Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90: 377–395.
Land, K. C., and Nagin, D. S. (1996). Micro-models of criminal careers: A synthesis of the criminal careers and life course approaches via semiparametric mixed Poisson regression models with empirical applications. *J. Quant. Crim.* 12: 163–191.
Land, K. C., McCall, P. L., and Nagin, D. S. (1996). A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models: With empirical applications to criminal careers data. *Sociol. Meth. Res.* 24: 387–442.

Laub, J. H., and Sampson, R. (1993). Turning points in the life course: Why change matters to the study of crime. *Criminology* 31: 301–326.

Lemert, E. M. (1972). *Human Deviance, Social Problems, and Social Control (Second Edition)*. Prentice-Hall, Englewood Cliffs, NJ.

Maddala, G. S. (1987). Limited dependent variable models using panel data. *J. Econ. Res.* 22: 307–338.

Maltz, M. D. (1994). Deviating from the mean: The declining significance of significance. *J. Res. Crime Delinq.* 31: 434–463.

Moffitt, T. E., and Lynam, D. (1994). The neuropsychology of conduct disorder and delinquency: Implications for understanding antisocial behavior. In Fowles, D., Sutker, P., and Goodman, S. (eds.), *Psychopathology and Antisocial Personality: A Developmental Perspective: Vol. 18. Progress in Experimental Personality and Psychopathology Research*, Springer, New York.

Nagin, D. S., and Farrington, D. P. (1992a). The stability of criminal potential from childhood to adulthood. *Criminology* 30: 235–260.

Nagin, D. S., and Farrington, D. P. (1992b). The onset and persistence of offending. *Criminology* 20: 501–523.

Nagin, D. S., and Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology* 31: 327–362.

Nagin, D. S., and Paternoster, R. (1991). On the relationship of past and future participation in delinquency. *Criminology* 29: 163–190.

Nagin, D. S., and Paternoster, R. (1994). Personal capital and social control: The deterrence implications of a theory of individual differences in criminal offending. *Criminology* 32: 581–606.

Osgood, D. W., and Rowe, D. C. (1994). Bridging criminal careers, theory, and policy through latent variable models of individual offending. *Criminology* 32: 517–554.

Paternoster, R., and Brame, R. (1997). Multiple routes to delinquency? A test of developmental and general theories of crime. *Criminology* 35: 49–84.

Paternoster, R., Dean, C. W., Piquero, A., Mazerolle, P., and Brame, R. (1997). Generality, continuity, and change in offending. *J. Quant. Crim.* 13: 231–266.

Phelps, E. (1972). *Inflation Policy and Unemployment Theory: The Cost Benefit Approach to Monetary Planning*. Macmillan, London.

Raftery, A. E. (1996). Bayesian model selection in social science research. In Marsden, P. V. (ed.), *Sociological Methodology, Vol. 25*. Basil Blackwell, Oxford.

Raine, A. (1997). Antisocial behavior and psychophysiology: A biosocial perspective and a prefrontal dysfunction hypothesis. In Stoff, D. M., Breiling, J., and Maser, J. D. (eds.), *Handbook of Antisocial Behavior*. Wiley, New York.

Robins, L. (1966). *Deviant Children Grown Up*. Williams and Wilkins, Baltimore.

Robins, L. (1978). Sturdy childhood predictors of adult antisocial behavior. *Psychol. Med.* 8: 611–622.

Sampson, R. J., and Laub, J. H. (1993). *Crime in the Making: Pathways and Turning Points Through Life*. Harvard University Press, Cambridge, MA.

Sampson, R. J., and Laub, J. H. (1995). Understanding variability in lives through time: Contributions of life-course criminology. *Studies on Crime and Crime Prevention: Biannual Review* 4: 143–158.

Sampson, R. J., and Laub, J. H. (1997). A life-course theory of cumulative disadvantage and the stability of delinquency. In Thornberry, T. P. (ed.), *Developmental Theories of Crime and Delinquency*. Transaction Publishers, New Brunswick, NJ.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6: 461–464.

Tracy, P., Wolfgang, M. E., and Figlio, R. M. (1990). *Delinquency Careers in Two Birth Cohorts*. Plenum, New York.

Wasserman, L. (1997). Bayesian model selection and model averaging. Working paper. Department of Statistics, Carnegie–Mellon University, Pittsburgh, PA.

Wilson, J., and Herrnstein, R. (1985). *Crime and Human Nature*. Simon and Schuster, New York, NY.