ISER Working Paper Series

What you don't see can't hurt you? Panel data analysis and the dynamics of unobservable factors

Mónica Hernández
School of Health and Related Research
University of Sheffield

Stephen Pudney
Institute for Social and Economic Research
University of Essex

No. 2011-13
May 2011

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

www.iser.essex.ac.uk

Non-technical summary

One of the biggest problems in applied economic and social research is causality. We often find very strong empirical associations in survey data, but it is rarely possible to show conclusively that these associations represent cause-and-effect. A good example of this is in illicit drug use: nearly all heroin and cocaine users report having taken cannabis at an earlier stage of their drug 'careers', but it is very hard to determine whether cannabis use *causes* subsequent use of hard drugs. A particular problem is that there may be underlying personal factors – psychological characteristics, beliefs, access to health information, parental and peer group influence, etc – which we cannot observe in surveys and which act as 'confounding variables'. For example, someone may have a particular psychological inclination towards self-gratification, which makes him or her more likely to use cannabis and also more likely to use heroin. It may be the underlying psychological weakness that causes the empirical association between cannabis and heroin, rather than cannabis-taking *causing* an increased risk of initiation into heroin use.

In areas like illicit drug use where long-term human experiments are infeasible and unethical, researchers have to rely on observational studies which collect information on large numbers of individuals over time. Statistical analysis then takes account of underlying unobserved confounding variables by assuming that they remain constant over time. This makes it possible to infer the nature of an individual's latent characteristics from his or her past history and then 'purge' their current behaviour of the influence of these confounding variables, to reveal the true causal relationship.

These statistical methods are often regarded as best practice, but they rest on the strong assumption that unobserved personal characteristics are constant over time. This assumption is particularly strong in the context of the behaviour of young people, who are undergoing complex developmental and socialisation processes, and it conflicts with many of the ideas of developmental psychology and sociology.

The aim of this paper is to question the assumption of time-invariant latent personal characteristics and explore the consequences for research findings if those characteristics change over time, while the researcher wrongly assumes them constant. Using a statistical model of crime and drug-taking behaviour by people aged 10-19 observed over a four-year period 2003-6 we show that, even if cannabis has no causal impact on the risk of initiation into other drugs, conventional statistical methods tend to show a spurious positive effect of cannabis. Our conclusion is that many research studies suggesting a significant causal "gateway" from early cannabis use into later drug problems are inherently unreliable because they rest on a questionable assumption which is rarely tested.

What you don't see can't hurt you? Panel data analysis and the dynamics of unobservable factors *

Mónica Hernández

School of Health and Related Research

University of Sheffield

Stephen Pudney

Institute for Social and Economic Research

University of Essex

This version May 10, 2011

Abstract: We investigate the consequences of using time-invariant individual effects in panel data models when the unobservables are in fact time-varying. Using data from the British Offending Crime and Justice panel, we estimate a dynamic factor model of the occurrence of a range of illicit activities as outcomes of young people's development processes. This structure is then used to demonstrate that relying on the assumption of time-invariant individual effects to deal with confounding factors in a conventional dynamic panel data model is likely to lead to spurious "gateway" effects linking cannabis use to subsequent hard drug use.

Keywords: Panel data; Dynamic factor models; Individual effects; Illicit drugs; Crime; Gateway effect

JEL classifications: C33, I18, K42

Contact: Steve Pudney, ISER, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK; tel. +44(0)1206-873789; email spudney@essex.ac.uk

*This work was supported with funding from three sources: the Open Society Institute (OSI-ZUG) (project "Two innovations in drugs policy: a cost-benefit analysis"), the European Research Council (project no. 269874 [DEVHEALTH]) and the ESRC Research Centre on Micro-Social Change (award no. RES-518-285-001). It was carried out in part while Pudney was a Faculty Visiting Scholar in the Melbourne Institute and Department of Economics at the University of Melbourne.

1 Introduction

Standard models for panel data analysis, such as fixed effects and random effects regression, include a time-invariant unobservable factor which serves as the primary link between successive observations on the same individual. This class of models has become so standard in applied economics as to be almost a fetish, going unchallenged despite the very strong time-invariance restriction it embodies. The panel data literature has focused heavily on the random versus fixed effects question which centres on the possibility of correlation between unobserved effects and the observed covariates, but much less attention has been paid to the time-invariance assumption. Despite this lack of concern, there is good reason to question the latter assumption. Evidence from psychology, sociology and economics suggests that many of the psychological characteristics and cognitive and non-cognitive capacities which these unobservables purport to represent are not fixed for all time, particularly for young people undergoing the complex and dynamic processes of development and socialisation.

One reason for the near-universal reliance on time-invariant individual effects is that, in the single-equation context of conventional panel data econometrics, it is impossible to identify completely general time-varying individual effects separately from other sources of unobserved variation in the dependent variable, such as measurement error, and it can be difficult to generalise the model in practice. For example, although the property of equicorrelated residuals implied by the time-invariance assumption can be relaxed by allowing for an additional autoregressive error component, it is often difficult to achieve good identification in practice (Calzolari and Magazzini 2009).

This single-equation approach to modeling can be contrasted with a multi-indicator approach where, with multiple dependent variables acting as indicators of a common set of unobservables, it is possible to identify the time-series structure of those unobservables much more clearly. The multi-indicator dynamic factor approach also matches more closely a view

of behavioural processes which is common throughout the social sciences. We envisage personal and family background, circumstances and events shaping the development of a set of individual preferences and capacities, which in turn influence the complex patterns of behaviour that we observe as outcomes. These underlying circumstances and events might include parental social class and parenting norms, neighbourhood characteristics, school quality and resources and events like dissolution of the parental partnership and possible repartnering. The personal attributes that develop in response to the changing environment might include cognitive and noncognitive capacities, preference attributes such as risk aversion and time preference, other psychological characteristics such as self-control and independence, and information or perceptions about opportunities and constraints on behaviour. For research involving analysis of conventional questionnaire-based survey data, these personal characteristics are often largely unobserved and must be treated as latent factors, detectable indirectly by their influence on behavioural outcomes. Recent examples of this approach in economics include Heckman et al (2006) and Cunha and Heckman (2007). The approach can be seen as generalising work on developmental trajectories by Nagin and Tremblay (1999, 2001), which used discrete latent class mixture models and by Pudney (2003) on crime and drug use, which used common time-invariant effects in a multi-indicator analysis. It also generalises a vast body of work using static common factor models (see Bollen 1989 for an exposition). The idea of analysing a fundamental developmental process rather than investigating a simple causal impact of one variable on another is very much in the spirit of an important strand of research in criminology (Sampson and Laub 2003, 2005).

Time-varying unobserved effects are a possibility in virtually all panel data applications, but we focus on a case where latent dynamics is particularly likely to be a concern: the involvement of young people in illicit activity. Negative outcomes like crime and drug abuse are often seen as pathological aspects of development of secondary interest to 'normal' economic and social functioning of individuals. However, for many people these negative outcomes

are as real and important as positive outcomes like education and employment. For example, in England at the time our panel data commenced, the net rate of enrolment in higher education was around 35% for 20 year olds (HESA 2003) whereas the proportions of 19-21 year-olds reporting some past involvement in crime or illicit drug use were each just over 50%.[1] Illicit and anti-social behaviour has frequently been studied econometrically, using the assumption of time-invariant individual effects to identify and control for the confounding effect of persistent unobservable heterogeneity and, following concerns raised by Pudney (2010), we use this particular application of panel data methods to examine the dangers of the time-invariance assumption.

The language of genetics is frequently used to interpret unobservable factors ("genetic endowments") and, implicitly, to justify the assumption of time-invariance. Recent research in neuroscience does give some basis for this idea (Mayer and Höllt 2005, Erickson 2007). For example, inherited vulnerability to alcohol dependency is well-established empirically and may be linked to specific genes (Dick and Bierut 2006); it has been suggested (Anthony et al 1994) that perhaps as many as 9% of people may have some predisposition to cannabis dependency, possibly stemming from dysregulation of cannabinoid receptors in the mesolimbic dopamine system, which might conceivably have some genetic origin. However, the contribution of genetic factors to dysfunctional and anti-social behaviour remains largely unknown scientific territory and it is highly likely that the genetic influences that do exist interact in complex ways with the social and physical environment in determining behaviour and may themselves be modified by the environment through epigenetic processes. Whatever role genetics may have, it is surely too naive to represent that role by an additive time-invariant 'effect'.

The paper has three main objectives: first, in section 2, we use simple theoretical arguments to show the possibility of serious bias, giving rise to distorted inferences about

[1] Author's calculation from the 2003 Offending, Crime and Justice Survey. Proportions reporting involvement within the last year are lower but still substantial: 24% (crime) and 33% (drugs).

the causal pathways that lead to serious illicit behaviour, when an invalid assumption of time-invariant individual effects is imposed on the analysis. Second, in sections 3 and 4, we develop an estimation procedure for a class of multi-indicator dynamic models and apply the method to panel data reflecting a wide range of problem behaviours among the 10-19 age group. We also explore the implications of the model for the impact of individuals' personal characteristics and circumstances on their trajectories of illicit behaviour. Third, in section 5, we use this estimated dynamic factor model as the basis for a Monte Carlo simulation to confirm empirically the poor performance of some common regression and GMM estimators when the time-invariance assumption is invalid. We conclude in section 6, arguing that the traditional econometric approach embedding time-invariant individual effects in single-equation models should be questioned much more than it is at present and, in the specific field of illicit drugs research, that much of the applied research literature on causal 'gateway' effects linking cannabis use to more serious problems later in life is open to criticism on these grounds.

2 Spurious association in models with individual effects

If latent factors evolve over time but are wrongly assumed to be time-invariant, the result is likely to be misspecification bias. Individual effects models work by 'controlling for' unobservables assumed to be constant over time. Loosely speaking, this means that, in analysing outcomes at time t, observations from other periods are used to infer the level of the unobservable individual effect which can then be 'stripped out' of the relationship generating the period t outcome. Time-invariance of the individual effect is clearly necessary if this procedure is to succeed in removing the unobserved confounder, and we would generally expect there to be some degree of spurious association if the estimation procedure fails to account for the whole of the common unobserved effect.

A simple model of the consumption of illicit drugs will make this clear. The gateway hypothesis (MacCoun and Reuter 2001, Kandel 2002) asserts that consumption of cannabis causes an increase in the future risk of hard drug consumption. A wide range of statistical methods has been used in attempts to estimate a causal gateway effect, including duration analysis (Van Ours 2003), discrete-time transition models (Pudney 2003), and instrumental variable estimation of static (Beenstock and Rahav 2002) and dynamic (DeSimone 1998, Kenkel et al 2001) regression models. We focus on dynamic regression methods, but the general point being made here is likely to apply also to other dynamic modeling techniques.

2.1 A spurious gateway effect

Let C_{it} and H_{it} be individual i's consumption of cannabis and hard drugs respectively in period $t = 1...T$. Consider a one-factor dynamic model, specified as follows.

$$C_{it} = \beta X_{it} + Q_{it} + \varepsilon_{it} \tag{1}$$

$$H_{it} = \gamma X_{it} + \lambda Q_{it} + \eta_{it} \tag{2}$$

where X_{it} is a vector of exogenous covariates, Q_{it} is the individual's latent state of personal development and socialisation, and ε_{it} and η_{it} are classical random disturbances, assumed to be mutually independent. Thus, there is no causal gateway effect here, only a correlation generated by the common influences X_{it} and Q_{it}.[2]

Assume Q_{it} is generated by a dynamic regression structure:

$$Q_{it} = \rho Q_{it-1} + \delta Z_{it} + u_i + \nu_{it} \tag{3}$$

Equation (1) can be lagged one period and re-expressed as $Q_{it-1} = C_{it-1} - \beta X_{it-1} - \varepsilon_{it-1}$. Now substitute this for the term Q_{it-1} in (3) and substitute the result into (2) to give the

[2]The dynamic latent factor models used by Heckman and others do not generally involve lagged indicators in the measurement equations. Each of the equations (1)-(2) can be extended by including the lagged dependent variable to represent substance-specific habit effects, which would imply the inclusion of second-order lags of C_{it} and H_{it} in (4). This is not generally done in regression models of the gateway effect; it would greatly complicate the analysis without materially affecting the basic argument we make here.

following dynamic regression relationship for hard drug use:

$$H_{it} = \lambda\rho C_{it-1} + \lambda\delta Z_{it} + \gamma X_{it} - \lambda\rho\beta X_{it-1} - \lambda\rho\varepsilon_{it-1} + \lambda u_i + \lambda\nu_{it} + \eta_{it} \tag{4}$$

Estimation of a single-equation model with C_{it-1} used as an explanatory covariate is sometimes used to model a causal gateway linking past use of cannabis to the risk of current use of hard drugs. After 'controlling for' the individual effect u_i, the coefficient of C_{it-1} is typically interpreted as a direct measure of the causal gateway effect. In this case, such an interpretation would suggest that there exists a positive gateway effect $\lambda\rho$, despite the fact that the true causal gateway effect is precisely zero in this case.

However, the situation is more complicated, since a conventional regression of H_{it} on C_{it-1}, Z_{it}, X_{it} will not give a consistent estimate of the spurious effect $\lambda\rho$ either. There are three sources of deviance: (i) the lagged covariates X_{it-1} are not usually included in the estimated model; (ii) there is a positive correlation between the residual component u_i and the lagged variable C_{it-1}; and (iii) there is a negative correlation between the residual component $-\varepsilon_{it-1}$ and C_{it-1}.

Thus, the results of a dynamic analysis of the relationship between current hard drug use and previous cannabis use will depend on the magnitude of the spurious gateway coefficient $\lambda\rho$, the extent to which the omitted X_{it-1} can be proxied by the included covariates Z_{it} and X_{it}, and the relative sizes of the variances of the hard drug-specific unobservable η and the common persistent factor u. Note that conventional strategies using transformation to eliminate u_i or instrumental variables to deal with endogeneity of C_{it-1} will not, in general, deliver consistent estimates of the true (zero) causal gateway coefficient.

2.2 A special case: fixed-effects regression with an AR(1) factor

To demonstrate the nature of the bias in a more concrete way, suppose Q_{it} is generated by the following special case of the autoregressive process (3):

$$Q_{it} = \rho Q_{it-1} + u_i + \nu_{it} \tag{5}$$

where u_i represents an unobserved fixed endowment with mean 0 and variance σ_u^2. For simplicity, assume the process $\{Q\}$ is stationary, so the initial condition Q_{i0} is randomly distributed according to the following relation.

$$Q_{i0} = \frac{u_i}{1-\rho} + \sum_{j=-\infty}^{0} \rho^{-j} \nu_{ij} \tag{6}$$

An analyst mistakenly assumes that persistent unobservable effects are time-invariant, and attempts to estimate a causal cannabis-hard drug gateway effect by regressing H_{it} on X_{it} and C_{it-1}, after removing the 'fixed effects' by subtracting individual-specific means. He or she will find significant evidence of a gateway effect if there is a positive correlation between $H_{it} - \bar{H}_i$ and $C_{it-1} - \bar{C}_i^l$, where $\bar{H}_i$ is person i's mean heroin consumption over the sample period $1...T$ and $\bar{C}_i^l = T^{-1} \sum_{t=0}^{T-1} C_{it-1}$ is mean lagged cannabis consumption. If Q_{it} is truly time-invariant, $H_{it} - \bar{H}_i$ and $C_{it-1} - \bar{C}_i^l$ are independent conditional on the X_{it}, so fixed effects regression gives a consistent estimate of the (zero) gateway effect, as $n \to \infty$ with T fixed. On the other hand, if Q is time-varying:

$$C_{it} - \bar{C}_i^l = \beta(X_{it} - \bar{X}_i) + Q_{it} - \bar{Q}_i + \varepsilon_{it} - \bar{\varepsilon}_i \tag{7}$$

$$H_{it} - \bar{H}_i = \gamma(X_{it} - \bar{X}_i^l) + \lambda(Q_{it} - \bar{Q}_i^l) + \eta_{it} - \bar{\eta}_i^l \tag{8}$$

Consequently, if $Q_{it} - \bar{Q}_i$ and $Q_{it-1} - \bar{Q}_i^l$ are positively correlated, there will be a spurious estimated gateway effect.

For any period $t \in \{2..T\}$, the covariance $Cov\left(Q_{it} - \bar{Q}_i\,,\; Q_{it-1} - \bar{Q}_i^l\right)$ can be written:

$$\begin{aligned} Cov\left(Q_{it} - \bar{Q}_i\,,\; Q_{it-1} - \bar{Q}_i^l\right) \;=\; & Cov\left(Q_{it}, Q_{it-1}\right) - T^{-1} \sum_{s=0}^{T-1} Cov\left(Q_{it}, Q_{is}\right) - T^{-1} \sum_{s=1}^{T} Cov\left(Q_{it-1}, Q_{is}\right) \\ & + T^{-2} \sum_{s=1}^{T} \sum_{j=0}^{T-1} Cov\left(Q_{is}, Q_{ij}\right) \end{aligned} \tag{9}$$

The model (5) implies:

$$Cov\left(Q_{is}, Q_{it}\right) = \frac{\sigma_u^2}{(1-\rho)^2} + \frac{\rho^{|t-s|}\sigma_\nu^2}{1-\rho^2} \tag{10}$$

and therefore $Q_{it} - \bar{Q}_i$ and $Q_{it-1} - \bar{Q}_i^l$ have covariance:

$$Cov\left(Q_{it} - \bar{Q}_i\,,\; Q_{it-1} - \bar{Q}_i^l\right) = \frac{\sigma_\nu^2}{1-\rho^2}\left\{\rho - T^{-1}\sum_{s=1}^{T}\rho^{|t-1-s|} - T^{-1}\sum_{s=0}^{T-1}\rho^{|t-s|} + T^{-2}\sum_{s=1}^{T}\sum_{j=0}^{T-1}\rho^{|s-j|}\right\} \tag{11}$$

Figure 1 plots the quantity $T^{-1}\sum_{t=1}^{T} Cov\left(Q_{it} - \bar{Q}_i\,,\; Q_{it-1} - \bar{Q}_i^l\right)/\sigma_\nu$ against ρ for various panel lengths T. As one might expect, the longer is the observation window, the more serious is the potential bias from misspecifying the time-varying latent factor as invariant. For the moderately large positive values of the autoregressive parameter ρ which are empirically plausible, there is a positive covariance between the residual term in the within-group transformed equation for H_{it} and the lagged term $Q_{it-1} - \bar{Q}_i^l$. In this simple illustrative model, only for the $T = 2$ case is there is no possibility of a spurious positive gateway effect – indeed, in very short panels there is a real possibility of an equally spurious negative gateway effect.

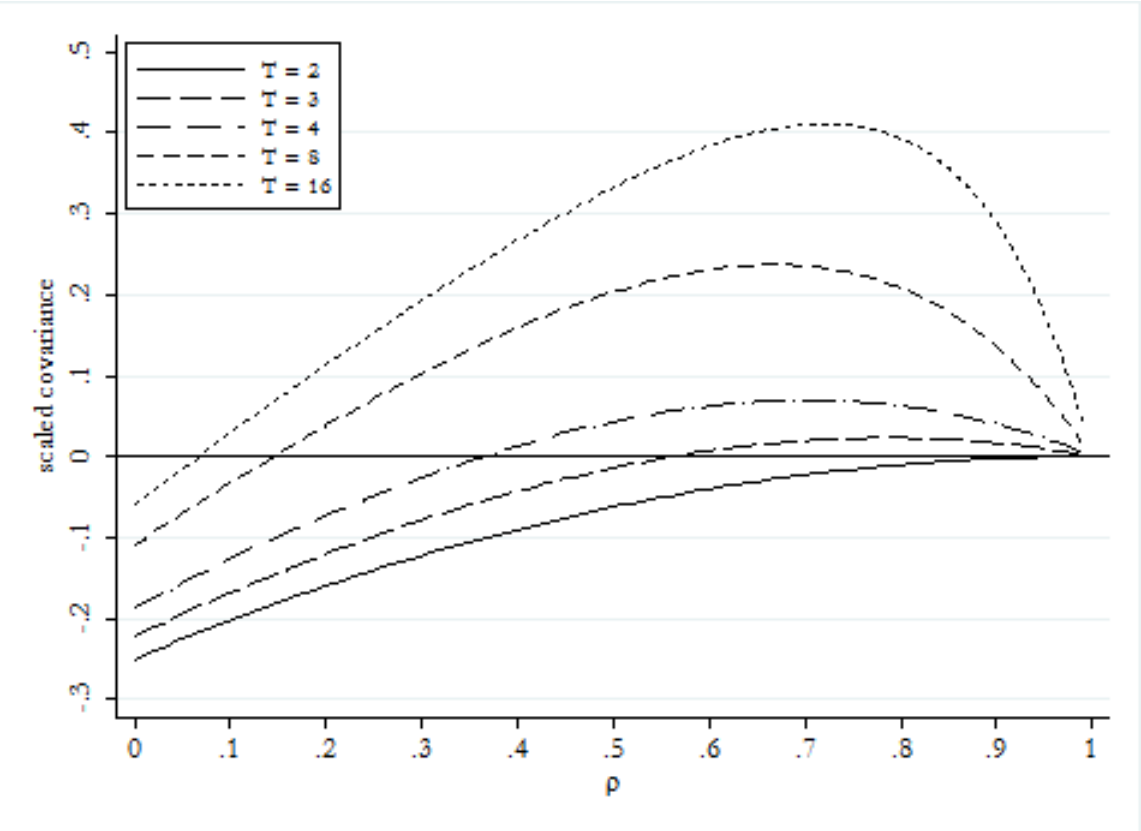


Figure 1 Autocovariance of the within-group transformed latent factor

3 A dynamic factor model

The model of the previous section serves to illustrate the source and likely nature of bias but it is too simple for application, so we now develop and apply a more general dynamic factor model. For any given individual, time t is measured as years of age from a fixed origin of $t = 0$. The latent factors presumed to underlie behaviour are arranged in a vector $\boldsymbol{q}_t$, which evolves over time according to a linear stochastic process:

$$\boldsymbol{q}_t = \boldsymbol{A}\boldsymbol{q}_{t-1} + \boldsymbol{B}\boldsymbol{z}_t + \boldsymbol{u} + \boldsymbol{\varepsilon}_t \ , \qquad t = 1, 2, ... \tag{12}$$

where $\boldsymbol{z}_t$ is a vector of observed variables representing the individual's changing social and economic environment, $\boldsymbol{u}$ is a vector of unobserved factors which are completely persistent over time, $\boldsymbol{\varepsilon}_t$ is a vector of transient unobserved factors and $\boldsymbol{A}$ and $\boldsymbol{B}$ are coefficient matrices. The vectors $\boldsymbol{q}_t$, $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}_t$ are R-dimensional and $\boldsymbol{z}_t$ is $k_z \times 1$.

The initial condition of this latent development process relates to period 0 (defined as age 10 for each individual in our application), and is specified as:

$$\boldsymbol{q}_0 = \boldsymbol{G}\boldsymbol{z}_0 + \boldsymbol{\eta} \tag{13}$$

where the covariate vector $\boldsymbol{z}_0$ explaining the initial state may contain a different collection of variables than $\boldsymbol{z}_1, \boldsymbol{z}_2, ...$ The vector $\boldsymbol{\eta}$ is the unobservable component of $\boldsymbol{q}_0$.

The M-dimensional vector of developmental outcomes at time t is $\boldsymbol{y}_t = (y_t^1 ... y_t^M)$. These might take various forms: continuous variables, binary and categorical variables, event counts, etc. For the sake of specificity, we assume that they are ordinal indicators and that an ordered probit structure is adequate to capture the conversion of the underlying continuous development process into discrete observed outcomes. Moreover, to avoid unimportant complications relating to normalisation of latent indicators, we assume that each ordinal indicator has at least three possible levels, thus excluding binary variables. The dimension M is likely to exceed greatly the number of latent factors, R. We allow for three

types of influence on the outcomes $\boldsymbol{y}_t$: the individual's general state of development and social attunement (represented by $\boldsymbol{q}_t$); observable factors $\boldsymbol{x}_t$ specific to particular outcome types (such as local enforcement provision or drug availability and price); and unobservable transient factors $\boldsymbol{\zeta}_t$ (such as randomly-arising opportunities for illicit gain). The linear index which drives the ordered probit model is:

$$\boldsymbol{y}_t^* = \boldsymbol{C}\boldsymbol{q}_t + \boldsymbol{D}\boldsymbol{x}_t + \boldsymbol{\zeta}_t \tag{14}$$

$$y_t^m = j \qquad \text{iff} \quad \Gamma_{j-1}^m \leq y_t^{*m} < \Gamma_j^m \,, \qquad j = 1...J_m \,,\ m = 1...M \tag{15}$$

where $\left\{\Gamma_j^m\right\}$ is a set of threshold parameters normalised with $\Gamma_0^m = -\infty$ and $\Gamma_{J_m}^m = \infty$ for each m, where J_m is the number of response categories for the mth variable in the vector $\boldsymbol{y}$. For simplicity, we have assumed that the same set of outcome variables is observed at every wave, so that the coefficient matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ are not time-subscripted.

We assume that the unobservables $(\boldsymbol{\eta}, \boldsymbol{u}, \boldsymbol{\varepsilon}_1 ... \boldsymbol{\varepsilon}_T, \boldsymbol{\zeta}_1 ... \boldsymbol{\zeta}_T)$ are jointly normally distributed, independently of all observed covariates, with a zero mean vector. Each pair of unobservables is independent, except for $\boldsymbol{\eta}$ and $\boldsymbol{u}$, which have covariance $\boldsymbol{\Sigma}_{\eta u}$. In addition, we assume that the variance matrix $\boldsymbol{\Sigma}_{\zeta\zeta}$ is diagonal, implying that all contemporaneous dependence between outcomes at time t is due to the effect of $\boldsymbol{q}_t$ and $\boldsymbol{x}_t$ which, in combination, are responsible for all serial dependence in the process $\{\boldsymbol{y}_t\}$. Write this underlying combination of factors as $\tilde{\boldsymbol{y}}_t = \boldsymbol{C}\boldsymbol{q}_t + \boldsymbol{D}\boldsymbol{x}_t$. The conditional period-t outcome probability is then $Pr\left(\boldsymbol{y}_t | \boldsymbol{q}_t, \boldsymbol{x}_t\right) = p_t\left(\boldsymbol{y}_t \,|\, \tilde{\boldsymbol{y}}_t; \boldsymbol{\Gamma}\right)$, where $p_t(.)$ is the multiple ordered probit probability function, known up to a set of parameters $\boldsymbol{\Gamma}$. Under the conditional independence assumption:

$$p_t\left(\boldsymbol{y}_t \,|\, \tilde{\boldsymbol{y}}_t; \boldsymbol{\Gamma}\right) = \prod_{m=1}^{M} \left[\Phi\left(\Gamma_{y_t^m - 1}^m - \tilde{y}_t^m\right) - \Phi\left(\Gamma_{y_t^m}^m - \tilde{y}_t^m\right)\right] \tag{16}$$

In Appendix 1, we examine this model in detail and demonstrate identification.

3.1 The simulated likelihood function

Assume initially that all the explanatory covariates are observed (or can be constructed) at each time period $t = 0, 1, 2, ...$ for every individual and let $\boldsymbol{q}_0$ be the initial state of the latent development process. We then observe $\boldsymbol{y}_t$ for a sequence of periods $\tau ... \tau + k$, where $\tau \geq 0$ and $k \leq 3$, since OCJS observations begins at age 10 or later and lasts for a maximum of four waves. In Appendix 1 we derive the covariance matrix $\boldsymbol{V}^*$ of the observation-period realisation $\boldsymbol{y}^* = (\boldsymbol{y}^*_\tau ... \boldsymbol{y}^*_{\tau+k})$. Then $\boldsymbol{y}^*$ can be decomposed as $\boldsymbol{y}^*_t = \boldsymbol{\mu} + (\boldsymbol{I} \otimes \boldsymbol{C})\boldsymbol{H}\boldsymbol{\lambda}$ where $\boldsymbol{\lambda} \sim N(\boldsymbol{0}, \boldsymbol{I})$, $\boldsymbol{H}$ is a matrix satisfying $\tilde{\boldsymbol{V}} = \boldsymbol{H}\boldsymbol{H}'$ and $\boldsymbol{\mu}_t$ and $\tilde{\boldsymbol{V}}$ (defined by equations (22)-(26) of Appendix 1) are the conditional mean vector and covariance matrix of the sequence $\boldsymbol{y}^*_\tau ... \boldsymbol{y}^*_{\tau+k}$. The likelihood for a given sample individual observed over $t = \tau ... \tau + k$ is:

$$L = E_\lambda \left\{ \prod_{t=\tau}^{\tau+k} p_t(\boldsymbol{y}_t | \boldsymbol{\mu}_t + (\boldsymbol{I} \otimes \boldsymbol{C})\boldsymbol{H}\boldsymbol{\lambda}; \boldsymbol{\Gamma}) \right\} \tag{17}$$

where E_λ denotes the expectation with respect to the $N(\boldsymbol{0}, \boldsymbol{I})$ density of $\boldsymbol{\lambda}$.

The dimension of the integral defining this distribution is $(k+1)R$, which is independent of the number of outcome variables that are used. Nevertheless, the dimensionality is too high for conventional quadrature algorithms to be used and we instead replace the expectation in (17) by the mean over a set of S pseudo-random draws $\boldsymbol{\lambda}^{(1)} ... \boldsymbol{\lambda}^{(S)}$. The results presented below are based on $S = 200$ replications, with antithetic variance reduction used to improve simulation precision. Numerical optimisation is performed using the simulated annealing global optimisation algorithm (Goffe et al 1994) as implemented in Gauss by E.G.Tsionas to produce a starting point for a quasi-Newton algorithm implemented in the Gauss MAXLIK routine.

4 Estimates of the factor model

4.1 Data: the Offending, Crime and Justice Survey

The Offending, Crime and Justice Survey (OCJS) was commissioned by the Home Office with the objective of providing a base for measuring prevalence of offending behaviour and drug use in the general population. Specifically, the survey covers the general public aged between 10 and 65, living in private households in England and Wales. The initial core sample of the 2003 survey consists of 10,085 respondents, interviewed in the period January-July. A subset of these respondents were then re-interviewed in three further waves over 2004-6 to generate a four-wave panel. Some of these panel respondents dropped out of the survey and some new respondents were added after the 2003 wave, so that the sample is unbalanced, both in terms of entry into and exit from the panel. The survey is mainly carried out with computer-assisted personal interviewing (CAPI), but a computer-assisted self-administered questionnaire (CASI) is used for sensitive areas like illicit activity. Relevant questions establish the frequency of criminal activity, anti-social behaviour and drug use within the last 12 months. See Murphy and Roe (2007) for a full description of OCJS methods and questionnaire content.

Our chronology takes period 0 to be age 10 for each individual. This has the practical advantage that it coincides with the lower age limit for inclusion in the OCJS sample that we use in the empirical application and it also corresponds to the age of criminal responsibility in England and Wales, from which the OCJS sample is drawn. The sample used for estimation consists of the set of OCJS respondents who were: (i) aged 16 or under for at least one wave of interviewing; (ii) gave valid responses to all questions relating to the nine forms of illicit behaviour; (iii) gave no mutually inconsistent answers to the sequence of repeated questions on age, gender and ethnicity; (iv) did not ever claim to have used the fictitious drug 'semeron' (included in the questionnaire as a check on data quality). In a small number of cases, there

was non-monotonic attrition involving unit non-response at a wave within the individual's period of participation in the panel; in such cases, observations following re-entry to the panel were discarded to ensure that a continuous set of observations was available for all sample members. The resulting estimation sample contains 3,026 individuals, appearing in an average of 2.7 waves per individual and spanning the age range 10-19.

We use nine trinary indicators of illicit activity, reflecting the frequency of criminal activity, drug use, anti-social behaviour and truancy during the 12 months preceding the interview, For the two categories (violence/damage and harder drugs) constructed from multiple survey questions, frequency is defined as the maximum of the reported frequencies for each constituent. Table 1 gives the distribution of each indicator in the sample used for estimation. Although each separate form of illicit activity is relatively rare in the sample, when viewed as a group they are highly prevalent: in only 59% of interviews do respondents report no illicit activity of any kind, and only 39% of the 3,026 individuals appearing in the panel reported no illicit activity at any interview. This sparse pattern of activity shows clearly the importance of taking a broad view of behaviour using multiple indicators, rather than focusing analysis on one or two specific forms of activity as is usual in the econometric literature.

Table 1 Distribution of outcomes: sample proportions, pooled sample

	Frequency in previous year (%)		
Outcome	None	Occasional[1]	Frequent[2]
Cannabis	91.3	3.4	5.3
Minor property crime	88.6	8.2	3.2
Noisy/rude behaviour	82.9	11.1	6.0
Other illicit drugs[3]	98.0	1.0	1.1
Serious property crime	98.4	1.2	0.4
Property damage or violence	81.9	12.2	5.9
Nuisance to neighbours	88.9	8.3	2.8
Graffiti	94.9	3.2	1.9
Truancy[4]	90.4	7.1	2.5

[1] *drugs*: < once a month last year; *crime*: once/twice last year; *anti-social behaviour/truancy*: 1-4 times last year.
[2] *drugs*: once a month or more; *crime*: more than twice last year; *anti-social behaviour/truancy*: > 4 times last year;
[3] poppers, amphetamines, ecstasy, LSD, cocaine, crack, heroin. [4] mean for those aged 16 or under

The OCJS questionnaire allows us to construct covariates representing parental social class and attitudes, family disruption, history of family trouble with the law and school discipline. These covariates, appearing in the vector $\boldsymbol{z}_t$ which drives the latent development process, are defined in detail in Table A1 of Appendix 2. The OCJS does not interview parents or teachers, so these covariates are based on the young person's perception and recall.

Although driven by the same latent development process, different outcomes may have different age profiles, for several possible reasons. There may be age differences in access: for example, truancy is only possible below the minimum school leaving age, and age limits on entry to bars and purchase of alcohol restrict binge drinking and associated crimes of violence and damage among under-18s. Such restrictions also make hard drugs more difficult to obtain by the young. Relative drug prices may also matter, since younger people are perhaps less able to afford the more expensive drugs like heroin and cocaine. To take account of these differences in access, we enter age as an outcome-specific covariate in $\boldsymbol{x}_t$. Note that any age effect in the latent development process cannot be separately identified and we resolve this by excluding age from $\boldsymbol{z}_t$. Consequently, the estimated age effects represent both the effect of ageing on the development process and its influence on the availability of different forms of illicit activity.

A practical difficulty is that the covariate process $\{\boldsymbol{z}_t\}$ is only partially observed since, with the exception of gender, we do not have observations on any of the covariates in $\boldsymbol{z}_t$ prior to the year of entry into the panel, τ. To deal with this, we impute the missing values for past $\boldsymbol{z}_t$ by setting each of them equal to the earliest observed value for the relevant covariate.

4.2 Parameter estimates

In models like (12)-(14), selection of the appropriate number of factors is not straightforward. Our main purpose in this paper is to investigate the issue of potential bias in conventional

panel data models involving a single individual effect, so a 1-factor specification is both parsimonious and natural. In fact, attempts to add a second factor to the structure were not successful, with no improvement in fit by the AIC or BIC likelihood criteria. MSL estimation was implemented using 200 pseudo-random replications with antithetic variance reduction.

The estimates are presented in Tables 2 and 3 below. The final empirical specification is the result of simplifying a less parsimonious model on the basis of coefficient significance tests. The initial specification involved additional variables including religious affiliation, various neighbourhood characteristics and more detailed versions of some of covariates appearing as simple dummy variables in the final specification. Given the long computational run times for this model, we have not pursued the process of testing down to its conclusion and, consequently, there remain several insignificant coefficients in the final specification. Table 2 shows the coefficients $\boldsymbol{A}$ and $\boldsymbol{B}$ in the dynamic model and $\boldsymbol{G}$ in the initial conditions model for the unobserved factor, together with the intra-class correlation, $\sigma_u^2/(\sigma_u^2+\sigma_\varepsilon^2)$. The dynamic factor process shows a high degree of autocorrelation, indicating the existence of slowly-decaying deviations from the time-invariance that is assumed in conventional panel data methods.

Table 2 Parameter estimates: latent process

Covariate	Current q_t	Initial q_0
Lagged q	0.653***	
	(0.034)	
Family trouble with law	0.454***	
	(0.081)	
Female	-0.159***	-1.173***
	(0.051)	(0.413)
Not two parents	0.278***	0.9879**
	(0.066)	(0.466)
Stepfather	0.068	2.098***
	(0.080)	(0.635)
Parental interest	-0.339***	0.018
	(0.043)	(0.450)
School discipline weak	0.357***	
	(0.042)	
Parental social class 1-3		-0.222
		(0.338)
Intra-class correlation	0.238***	
	(0.055)	

Table 3 shows the estimated factor loadings ($\boldsymbol{C}$) relating the behavioural indicators to the latent developmental state, and also the age coefficients ($\boldsymbol{D}$) which determine the baseline profile of behavioural risk. There are two important features of the estimates in Table 3. First, there are significant loadings for all indicators, underlining the potential advantages of a multi-indicator approach over a single-equation econometric analysis. A second striking feature is the diversity in the age coefficients, which suggests that there are large differences in the life stage at which different problem behaviours become salient. This is a complicating issue for causal analysis, since some behaviours may tend to be observed earlier than others simply because of differences in the ease of access to, of cost of, the behaviour in question: precedence in time does not necessarily imply causal ordering.

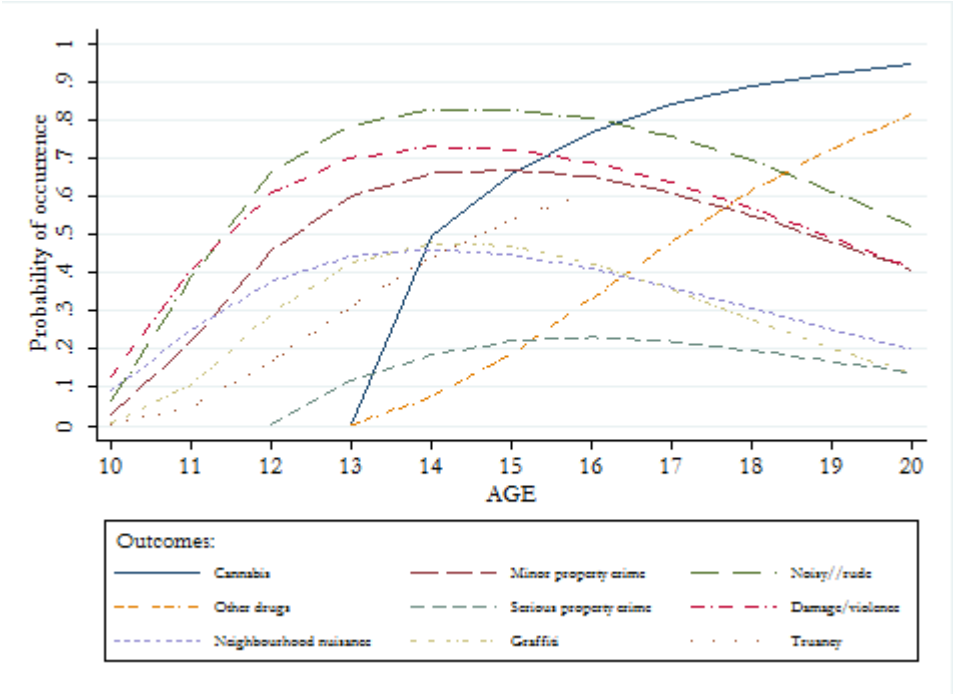
Table 3 Parameter estimates: outcomes

Covariate	Latent q_t	Age
Cannabis	1	2.997***
	(-)	(0.982)
Minor property crime	0.444***	-2.336***
	(0.035)	(0.452)
Noisy/rude behaviour	0.492***	-2.902***
	(0.037)	(0.494)
Other illicit drugs	1.117***	5.592***
	(0.150)	(1.395)
Serious property crime	0.612***	-2.006***
	(0.079)	(0.759)
Property damage or violence	0.401***	-2.683***
	(0.030)	(0.409)
Nuisance to neighbours	0.321***	-2.389***
	(0.027)	(0.354)
Graffiti	0.697***	-4.435***
	(0.065)	(0.783)
Truancy	0.330***	0.894**
	(0.029)	(0.393)

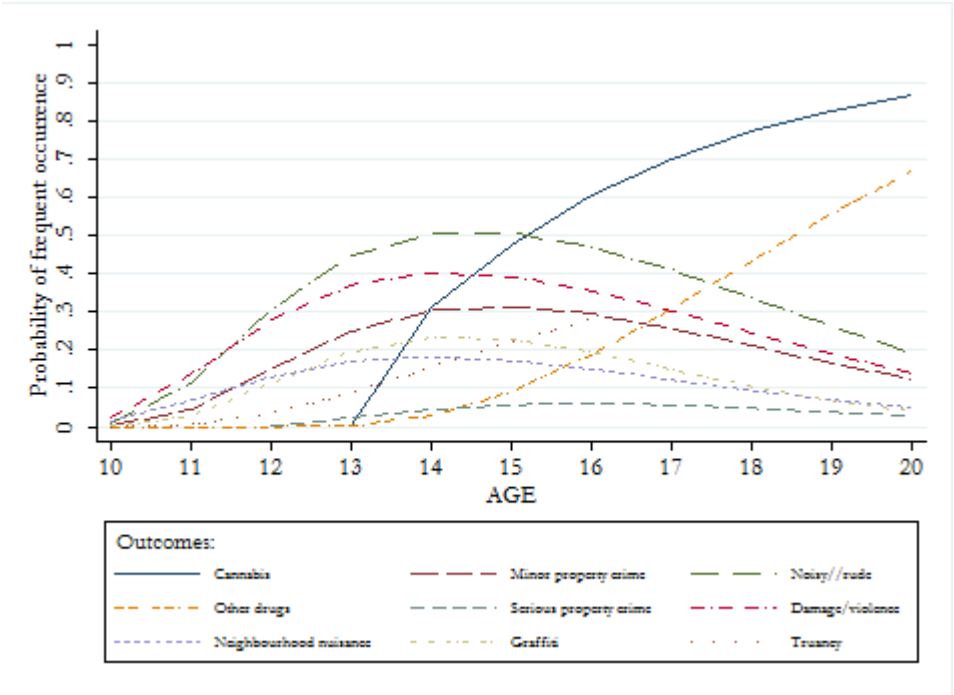
4.3 Dynamic properties

The dynamic properties of the model cannot easily be deduced directly from the parameter estimates, so we use dynamic simulation to illustrate the implications of these estimates. The implied developmental trajectories are calculated in the form of a sequence of probabilities of illicit activity for two hypothetical individuals with unfavourable and favourable observable characteristics. The former is male, from a non-professional/managerial social class, has a family history of trouble with the police, has divorced or separated natural parents and a step-father, and perceives little parental or school discipline. The latter is female, from a high-social class family with no history of trouble with the police, and perceives strong parental and school discipline. We hold all characteristics constant through time, set unobserved random errors at their mean values of 0, and calculate at each age from 10 to 20 the probability of (i) any occurrence of each specific type of illicit behaviour and (ii) occurrence at high frequency.

The risk profiles are plotted in Figures 2 and 3 for the unfavourable and favourable cases respectively. Despite the fact that all outcome indicators are strongly reflective of the latent factor, there is a clear difference between the age profiles for drug use and most other types of illicit behaviour: drug use tends to develop later and to rise strongly with a leveling-off around age 20. In contrast, crime and anti-social behaviour are characteristic of younger people, with peak incidence around age 14 and a steady decline after that age. The exception is truancy, which rises steadily to age 16, when compulsory schooling comes to an end.

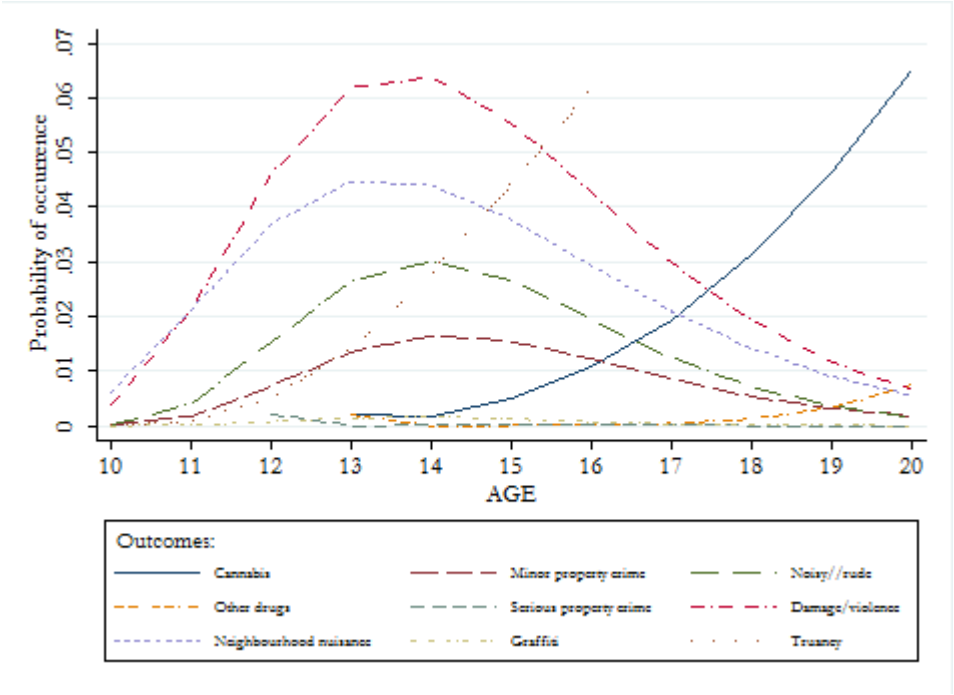


(i) any occurrence

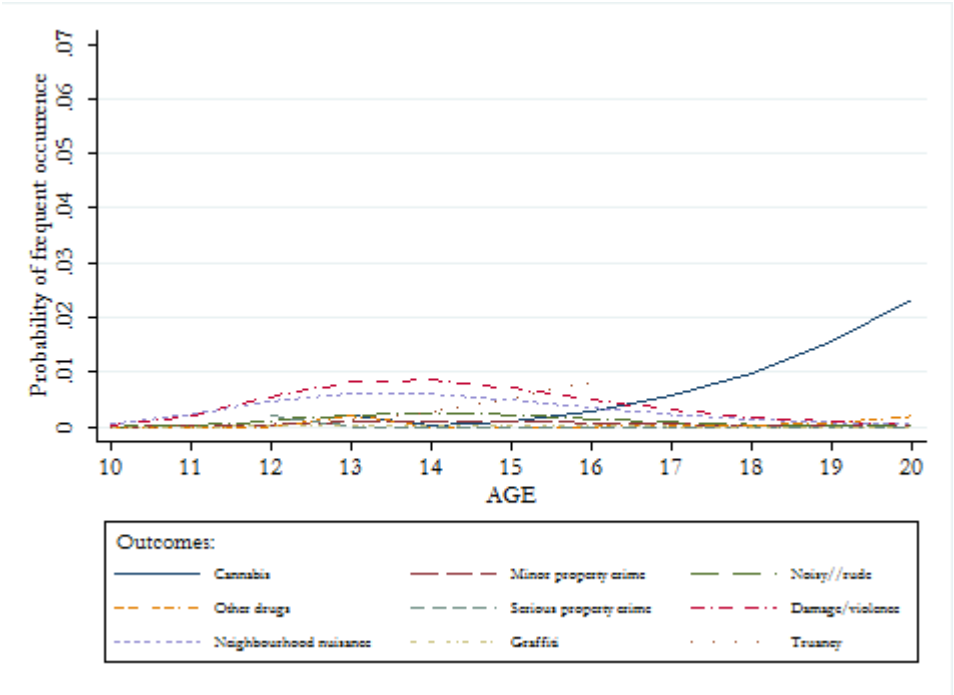


(ii) frequent occurrence

Figure 2 Age-specific probabilities of illicit or antisocial behaviour for individuals with unfavourable characteristics



(i) any occurrence



(ii) frequent occurrence

Figure 3 Age-specific probabilities of illicit or antisocial behaviour for individuals with favourable characteristics

5 Misspecification of the latent factor: the gateway effect

We now treat the dynamic factor model of the previous section as the true data generation process and use it to investigate the consequences of estimating a model broadly representative of the single-equation approach that is typical of econometric panel data analysis involving time-invariant individual effects. This representative misspecified model is:

$$H_{it} = \alpha H_{it-1} + \beta C_{it-1} + \boldsymbol{\gamma}\boldsymbol{w}_{it} + u_i + \varepsilon_{it} \tag{18}$$

where C_{it} and H_{it} are the cannabis and 'other drugs' outcome variables, $\boldsymbol{w}_{it} = (\boldsymbol{x}_{it}, \boldsymbol{z}_{it})$ is the full vector of covariates used in the 'true' dynamic factor model and u_i and ε_{it} are the time-invariant and time-varying residual terms.

The Monte Carlo simulation model works as follows. For each of a series of replications indexed by $r = 1...R$, generate $(\eta_0^r, u_i^r, \varepsilon_{it}^r, \boldsymbol{\zeta}_{it}^r)$ as a pseudo-random draw from the appropriate multivariate normal distribution and then:

Step 1 Generate the initial value $q_{i0}^r = \boldsymbol{G}\boldsymbol{z}_0 + \eta_0^r$ for each $i = 1...n$.

Step 2 Generate developmental sequences of length T_i as $q_{it}^r = Aq_{it-1}^r + \boldsymbol{B}\boldsymbol{z}_t + u_i^r + \varepsilon_{it}^r$, where T_i is the age (measured from an origin at 10 years) when individual i was last observed in the panel.

Step 3 Construct the latent behavioural indicators $\boldsymbol{y}_{it}^{*r} = \boldsymbol{C}q_{it}^r + \boldsymbol{D}\boldsymbol{x}_{it} + \boldsymbol{\zeta}_{it}^r$

Step 4 Construct the observable ordinal indicators $\boldsymbol{y}_{it}^r$ from the latent $\boldsymbol{y}_{it}^{*r}$ using the relevant threshold values.

Step 5 Apply standard panel data estimators of the model (18) to the continuous latent data $\boldsymbol{y}_{it}^{*r}$ and the ordinal data $\boldsymbol{y}_{it}^r$

In this way, we examine the performance of the following four standard estimation methods, of which two are known to give inconsistent estimates in dynamic models fitted to short panels and two, based on the generalised method of moments (GMM) were specifically developed for dynamic modeling. All four of these methods are designed for analysis of continuous rather than discrete variables, but are quite commonly used for discrete variables in the applied research literature. To abstract from the possible biases induced by using regression-type methods on ordinal variables, we apply them in two variants of the experiment: first using the simulated ordinal data on C_{it} and H_{it}; and second using the continuous latent variables which underlie C_{it} and H_{it}.

(*i*) *Fixed effects regression* involves least-squares regression applied to (18) after transformation to deviations from individual means. This within-group transform eliminates the persistent effect u_i but also transforms the time-varying residual to the form $\varepsilon_{it} - \bar{\varepsilon}_i$ and thus introduces between-period residual correlation. Consequently, even in a well-specified model, there is correlation between the transformed residual and lagged dependent variable, causing inconsistency as $n \to \infty$ in a short panel. (Nickell 1981).

(*ii*) *Random effects regression* uses a weighted combination of within- and between-individual variation, by applying least-squares regression to a transformed version of (18) where each variable is converted to a quasi-mean difference of the form $y_{it} - \theta_i \bar{y}_i$, where θ_i is a constant (depending on panel length T_i and estimates of $var(u_i)$ and $var(\varepsilon_{it})$), chosen to achieve efficiency in the classical panel data regression model. Since the quasi-difference transform does not eliminate u_i from the residual nor preserve the assumed serial independence of $\{\varepsilon\}$, there are two sources of inconsistency when the method is used to estimate a dynamic regression, even when the regression is well-specified.

(*iii*) *GMM estimation in differences* (Arellano and Bond 1991) uses time differencing of (18) to eliminate u_i. The differenced equation has a moving average residual process $\varepsilon_{it} - \varepsilon_{it-1}$ which, for a correctly-specified model of the form (18), would be uncorrelated with

the instruments $\{H_{it-2}, H_{it-3}, ...; C_{it-2}, C_{it-3}, ...; \boldsymbol{w}_{it}\}$. The Arellano-Bond estimator then minimises a GMM criterion function embodying these moment conditions. We also use a collapsed version of the instrument set, eschewing the period-specific use of lagged H and C to avoid excessive numbers of instruments. We use a two-stage robust version of GMM that incorporates a finite-sample correction due to Windmeijer (2005), as implemented in the Stata procedure *XTABOND2* (Roodman 2009).

(*iv*) *GMM estimation in differences and levels* (Arellano and Bover 1995) extends the simple Arellano-Bond differenced model by including moment conditions relating to the levels version of equation (18) as well as its differenced form. The additional moment condition is of the form $E\left(\Delta H_{it-1}[u_i + \varepsilon_{it}]\right)$, which adds valuable information otherwise ignored by the GMM estimator, but which is valid in the context of model (18) only under the strong additional assumption that the deviation of the initial observation on H from the long-run equilibrium value is uncorrelated with u_i and that C_{it-1} is also uncorrelated with u_i (Roodman 2009).

Table 4 summarises the results of these simulations, carried out using 500 replications. The first panel is based on direct analysis of the continuous latent indicators, rather than the discrete responses observed in actual survey data, and reveals clear positive biases for all of the six estimators. As might be expected, random effects regression performs worst, having the largest mean bias and a 100% rejection rate for the asymptotic t-test of the hypothesis of a zero gateway coefficient against the alternative of a positive effect. Fixed effects regression produces a smaller positive mean bias, but still rejects the zero gateway effect in 95% of replications. The four variants of the GMM estimator, although designed to avoid dynamic biases in panel data models, display a positive mean bias similar to that of crude fixed effects regression, and the t-test for the zero gateway effect has rejection rates ranging from 69% to 96%. The Arellano-Bond differences-only version of GMM is more affected by sampling

error than the Arellano-Bover level+differences variant, and this contributes to its slightly lower probability of rejecting the zero gateway hypothesis.

Table 4 Simulation results for the gateway parameter β

	Fixed effects regression	Random effects regression	GMM differences only full[a]	GMM differences only collapsed[b]	GMM levels+differences full[c]	GMM levels+differences collapsed[d]
			Continuous data			
Mean $\hat{\beta}$	0.096	0.309	0.088	0.102	0.111	0.108
Standard deviation $\hat{\beta}$	0.023	0.017	0.038	0.040	0.033	0.033
Proportion significant*	0.994	1.000	0.694	0.800	0.962	0.944
			Discrete data			
Mean $\hat{\beta}$	0.063	0.156	-0.076	-0.099	0.047	0.051
Standard deviation $\hat{\beta}$	0.023	0.018	0.056	0.060	0.032	0.034
Proportion significant*	0.950	1.000	0.000	0.000	0.428	0.474

No. of instruments: [a] 27; [b] 15; [c] 37; [d] 19; * Significantly greater than zero, using a 1-tailed 95%-level Wald test

The second panel of Table 4 is based on application of the estimators to simulated data in the discrete form observed in the OCJS. Although the regression and GMM estimators are all designed for analysis of continuous data, it has become common practice to apply these methods to discrete panel data, since this avoids the complexity of nonlinear modeling and retains the possibility of using fixed effects and differencing to eliminate unobservables. It is often argued that ignoring discreteness in this way makes little difference to the results obtained. Our simulations clearly contradict this. Note that, because of differences in scale between the discrete indicators and their latent counterparts, mean biases cannot be compared across the two panels of Table 4, but estimation precision, as measured by the coefficients of variation of the simulated coefficients, definitely declines as a consequence of the lower information content of the discrete indicators. The difference-only version of the GMM estimator now behaves quite perversely, with a negative mean bias and a zero rejection rate for the null hypothesis of a zero gateway coefficient.[3] In contrast, the Arellano and Bover version of GMM again has a positive mean bias only slightly smaller than that of fixed effects

[3]However, there would be rejection rates of 38% and 33% for the collapsed and full instrument sets respectively, if we were to use a *two-sided* 95%-level test.

regression and continues to reject the zero gateway hypothesis in an uncomfortably large proportion (43-47%) of cases. The reduction in the rejection rate relative to fixed-effects regression is due to the greater extent of sampling variation so, if a spuriously significant gateway effect were avoided in practice, it would mainly be due to statistical imprecision rather than lack of bias – hardly a triumph for econometric modeling.

6 Conclusions

Panel data are valuable because they give us more scope for establishing causal relationships than do simple cross-sections. They also allow us to deal with persistent unobservable factors that might otherwise generate spurious associations. In this study, we have examined the issue of causal association in the presence of persistent unobservables, focusing specifically on the important case of the gateway hypothesis, which holds that involvement in low-level drug abuse (such as smoking cannabis) causes a rise in the risk of subsequent more serious drug abuse. The standard method of estimating gateway effects is to use panel data to allow observation of current cannabis use and subsequent use of other drugs and control for unobservable factors which are assumed constant over time.

In this paper we have done three things. First, we have argued that, rather than investigating a simple impact of one behavioural outcome (cannabis use) on another (subsequent hard drug use), we should allow the possibility that both outcomes are expressions of the same underlying developmental process which governs the way that behavioural decisions are made. This is line with much of the literature on human development in psychology, sociology and, more recently, in economics.

Second, we have estimated a dynamic latent variable model of nine forms of illicit and antisocial behaviour, using OCJS panel data for 10-19 year-olds in England and Wales in 2003-6. The estimated model displays a coherent pattern of dynamics underlying this wide

range of behaviours, with a common unobservable factor that displays a high degree of persistence – but not complete invariance – over time.

Third, we have demonstrated theoretically that, if the latent factors which underlie behaviour are evolving, rather than constant, over time, then standard methods based on random or fixed effects analysis are likely to give a biased picture of the pattern of causation, typically resulting in spurious empirical 'gateway' effects which have little connection with true causal mechanisms. Using the estimated dynamic factor model as a 'true' data generating process in a Monte Carlo study, we have confirmed empirically the theoretical prediction of bias causing standard panel data methods which incorporate time-invariant unobservables to indicate a completely spurious 'causal' gateway between cannabis and harder drugs.

A true causal gateway may or may not exist in reality but, in our view, standard econometric methods of panel data analysis are not adequate to identify it reliably.

References

[1] Anthony, J. C., Warner, L. A. and Kessler, R. C. (1994). Comparative epidemiology of dependence on tobacco, alcohol, controlled substances and inhalants: basic findings from the National Comorbidity Survey, *Experimental and Clinical Psychopharmacology* **2**, 244-268.

[2] Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations, *Review of Economic Studies* **58**, 277-297.

[3] Arellano, M. and Bover, O. (1995). Another look at instrumental variable estimation of error-component models, *Journal of Econometrics* **68**, 29-51.

[4] Bollen, K.A. (1989), *Structural equations with latent variables*, New York: Wiley.

[5] Calzolari, G. and Magazzini, L. (2009). Poor identification and estimation problems in panel data models with random effects and autocorrelated errors, Università di Verona, Dipartimento di Scienze economiche, Working Paper no. 53.

[6] Cunha, F. and J. Heckman (2007), The technology of skill formation, *American Economic Review* **97**, 3147.

[7] DeSimone, J. (1998). Is marijuana a gateway drug?, *Eastern Economic Journal* **24**, 149-164.

[8] Dick, D. M. and Bierut, L. J. (2006). The genetics of alcohol dependence, *Current Psychiatry Reports* **8**, 151-157.

[9] Ericksson, C. K. (2007). *The Science of Addiction. From Neurobiology to Treatment*. New York: Norton.

[10] Goffe, W. L., Ferrier, G. D. and Rogers, J. (1994). Global optimization of statistical functions with simulated annealing, *Journal of Econometrics* **60**, 65-99.

[11] Heckman, J. J., Stixrud, J., Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior, *Journal of Labor Economics* **24**, 411-482.

[12] ESA (2003). *Higher Education Statistics for the UK 2002/3*, London: Higher Education Statistics Agency.

[13] Kandel, D. B. (2002). *Stages and Pathways of Drug Involvement*, Cambridge: Cambridge University Press.

[14] Kenkel, D., Mathios, A. D. and Pacula, R. L. (2001). Economics of youth drug use, addiction and gateway effects, *Addiction* **96**, 151-164.

[15] Kershaw, C., Nicholas, S. and Walker, A. (2008), Crime in England and Wales 2007/8. Findings from the British Crime Survey and police recorded crime. London: Home Office Statistical Bulletin 07/08.

[16] MacCoun, R. J. and Reuter, P. (2001). *Drug War Heresies. Learning from Other Vices, Times and Places*, Cambridge: Cambridge University Press.

[17] Mayer, P. and Höllt, V. (2005). Genetic disposition to addictive disorders: current knowledge and future perspectives, *Current Opinion in Pharmacology* **5**, 4-8.

[18] Murphy, R. and Roe, S. (2007). Drug misuse declared: findings from the 2006/07 British Crime Survey. England and Wales. London: Home Office Statistical Bulletin 18/07.

[19] Nagin, D. S. and Tremblay, R. E. (1999) Trajectories of boys' physical aggression, opposition and hyperactivity on the path to physically violent and non-violent juvenile delinquency. *Child Development* **70**, 1181-1196.

[20] Nagin, D. S. and Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods* **6**, 18-34.

[21] Nickell, S. J. (1981). Biases in dynamic models with fixed effects, *Econometrica* **47**, 1249-1266.

[22] Pudney, S. E. (2003). The road to ruin? Sequences of initiation to drugs and crime in Britain, *Economic Journal* **113**, C182-C198.

[23] Pudney, S. E. (2010). Drugs policy - what should we do about cannabis?, *Economic Policy* **25**, 165-211.

[24] Roodman, D. (2009) A note on the theme of too many instruments, *Oxford Bulletin of Economics and Statistics* **71**, 135-158.

[25] Sampson, R. J. and Laub, J. H. (2003) Life-course desisters? Trajectories of crime among delinquent boys followed to age 70. *Criminology* **41**, 555-592.

[26] Sampson, R. J. and Laub, J. H. (2005) (Special Editors). Developmental criminology and its discontents: Trajectories of crime from childhood to old age. *Annals of the American Academy of Political and Social Science* **602**, November.

[27] Van Ours, J. C. (2003). Is cannabis a stepping-stone for cocaine? *Journal of Health Economics* **22**, 539-554.

[28] Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators, *Journal of Econometrics* **126**, 25-51.

Appendix 1: Identification

The latent process (12) implies that, for any period $t \geq 1$, $\boldsymbol{q}_t$ can be expressed as:

$$\boldsymbol{q}_t = \boldsymbol{A}^t \boldsymbol{q}_0 + \boldsymbol{S}_t \boldsymbol{u} + \sum_{s=1}^{t} \boldsymbol{A}^{t-s} \left(\boldsymbol{B} \boldsymbol{z}_s + \boldsymbol{\varepsilon}_s \right) \tag{19}$$

where $\boldsymbol{S}_t$ is the t-element sum $\boldsymbol{I} + \boldsymbol{A} + ... + \boldsymbol{A}^{t-1}$.

This implies the following set of linear indexes underlying the observed sequence of outcomes:

$$\boldsymbol{y}_0^* = \boldsymbol{C}\boldsymbol{G}\boldsymbol{z}_0 + \boldsymbol{D}\boldsymbol{x}_0 + \boldsymbol{C}\boldsymbol{\eta} + \boldsymbol{\zeta}_0 \tag{20}$$

$$\boldsymbol{y}_t^* = \boldsymbol{C}\boldsymbol{A}^t\boldsymbol{G}\boldsymbol{z}_0 + \sum_{s=1}^{t} \boldsymbol{C}\boldsymbol{A}^{t-s}\boldsymbol{B}\boldsymbol{z}_s + \boldsymbol{D}\boldsymbol{x}_t + \boldsymbol{C}\boldsymbol{S}_t\boldsymbol{u} + \boldsymbol{C}\boldsymbol{A}^t\boldsymbol{\eta} + \sum_{s=1}^{t} \boldsymbol{C}\boldsymbol{A}^{t-s}\boldsymbol{\varepsilon}_s + \boldsymbol{\zeta}_t \,, \quad t \geq 1 \tag{21}$$

The joint distribution of $\boldsymbol{y}_\tau^* ... \boldsymbol{y}_{\tau+k}^*$ conditional on the whole realisation of observed covariates is multivariate normal with mean vector containing elements of the form $\boldsymbol{\mu}_t = E\left(\boldsymbol{y}_t^* | \boldsymbol{Z}, \boldsymbol{X}\right)$:

$$\boldsymbol{\mu}_t = \begin{cases} \boldsymbol{C}\boldsymbol{G}\boldsymbol{z}_0 + \boldsymbol{D}\boldsymbol{x}_0 & t = \tau = 0 \\ \boldsymbol{C}\boldsymbol{A}^t\boldsymbol{G}\boldsymbol{z}_0 + \sum_{s=1}^{t} \boldsymbol{C}\boldsymbol{A}^{t-s}\boldsymbol{B}\boldsymbol{z}_s + \boldsymbol{D}\boldsymbol{x}_t & t > \tau = 0 \text{ or } t \geq \tau > 0 \end{cases} \tag{22}$$

Let $\tilde{\boldsymbol{V}}$ be the conditional covariance matrix of the sequence $\tilde{\boldsymbol{y}}_t ... \tilde{\boldsymbol{y}}_{t+k}$, constructed from the following terms:

$$\tilde{\boldsymbol{V}}_{00} = \boldsymbol{\Sigma}_{\eta\eta} \tag{23}$$

$$\tilde{\boldsymbol{V}}_{tt} = \boldsymbol{A}^t\boldsymbol{\Sigma}_{\eta\eta}\boldsymbol{A}^{t\prime} + \boldsymbol{S}_t\boldsymbol{\Sigma}_{uu}\boldsymbol{S}_t' + \sum_{s=1}^{t} \boldsymbol{A}^{t-s}\boldsymbol{\Sigma}_{\varepsilon\varepsilon}\boldsymbol{A}^{t-s\prime} + \boldsymbol{A}^t\boldsymbol{\Sigma}_{\eta u}\boldsymbol{S}_t' + \boldsymbol{S}_t\boldsymbol{\Sigma}_{u\eta}\boldsymbol{A}^{t\prime} \qquad \text{for } t > 0 \tag{24}$$

$$\tilde{\boldsymbol{V}}_{0t} = \boldsymbol{\Sigma}_{\eta\eta}\boldsymbol{A}^{t\prime} + \boldsymbol{\Sigma}_{\eta u}\boldsymbol{S}_t \qquad \text{for } t > 0 \tag{25}$$

$$\tilde{\boldsymbol{V}}_{t,t+s} = \boldsymbol{A}^t\boldsymbol{\Sigma}_{\eta\eta}\boldsymbol{A}^{t+s\prime} + \boldsymbol{S}_t\boldsymbol{\Sigma}_{uu}\boldsymbol{S}_{t+s}' + \sum_{j=1}^{t} \boldsymbol{A}^{t-j}\boldsymbol{\Sigma}_{\varepsilon\varepsilon}\boldsymbol{A}^{t+s-j\prime} + \boldsymbol{A}^t\boldsymbol{\Sigma}_{\eta u}\boldsymbol{S}_{t+s} + \boldsymbol{S}_t\boldsymbol{\Sigma}_{u\eta}\boldsymbol{A}^{t+s\prime} \qquad \text{for } t, s > 0 \tag{26}$$

The residual autocovariances of the latent regressions for $\boldsymbol{y}_t^*$ are:

$$\boldsymbol{V}_{tt}^* = \boldsymbol{C}\tilde{\boldsymbol{V}}_{tt}\boldsymbol{C}' + \boldsymbol{\Sigma}_{\zeta\zeta} \qquad \text{for } t \geq 0 \tag{27}$$

$$\boldsymbol{V}_{t,t+s}^* = \boldsymbol{C}\tilde{\boldsymbol{V}}_{t,t+s}\boldsymbol{C}' \qquad \text{for } t \geq 0,\ s > 0 \tag{28}$$

Make the following assumptions, which are sufficient to permit a simple constructive demonstration of identification but considerably stronger than necessary.

A1 Scale normalisation For any indictor y^m (with $J_m > 2$), set $\Gamma_1^m = 0$ and $\Gamma_2^m = 1$ with the corresponding diagonal element of $\boldsymbol{\Sigma}_{\zeta\zeta}$ left unrestricted. This is not the normalisation conventionally used for the ordered probit model, but it is observationally equivalent to it and has the advantage that the variance parameters can be left unrestricted.

A2 Non-collinearity The vector $\{\boldsymbol{G}\boldsymbol{z}_0, \boldsymbol{z}_1\}$ and each of the vectors $\{\boldsymbol{z}_t, \boldsymbol{x}_t\}$ (for any period $t \geq 0$) has a positive definite covariance matrix.

A3 Coefficient rank The coefficient matrices $\boldsymbol{C}$ and $\boldsymbol{G}$ are of rank R. Note that if both $\boldsymbol{C}$ and $\boldsymbol{G}$ were of lower rank, the model could be reconstituted as a full rank system with a smaller number of factors, constructed as linear combinations of the original ones.

A4 Normalisation of factor loadings The number of outcome indicators exceeds the number of latent factors ($M > R$) and the $M \times R$ matrix of factor loadings is normalised to be of the form $\boldsymbol{C} = \begin{pmatrix} \boldsymbol{I}_R \\ \boldsymbol{C}_2 \end{pmatrix}$ for a suitable ordering of the outcome variables. The 'basis' indicators are all non-binary, so that $J_m > 2$ for $m = 1...R$.

A5 Serial dependence of latent factors Every element of $\boldsymbol{q}_{t-1}$ feeds forward into at least one element of $\boldsymbol{q}_t$, implying that all columns of $\boldsymbol{A}$ have at least one non-zero element.

We state without proof the following proposition:

Proposition Consider any system of "seemingly unrelated" ordered probit equations:

$$\boldsymbol{y}^* - \boldsymbol{w} = \boldsymbol{\Psi}\boldsymbol{\xi} + \boldsymbol{\upsilon} \tag{29}$$

$$y_t^m = j \qquad \text{iff} \quad \Gamma_{j-1}^m \leq y_t^{*m} < \Gamma_j^m \,, \qquad j = 1...J_m \,,\ m = 1...M \tag{30}$$

where the displacement vector $\boldsymbol{w}$ is observed. The coefficient matrix $\boldsymbol{\Psi}$ and residual covariance matrix $\boldsymbol{\Sigma}_{\upsilon\upsilon}$ can be estimated consistently in a randomly-sampled cross-section if the covariates $\boldsymbol{\xi}$ are strictly exogenous and have a positive definite variance matrix. The residual autocovariances $\boldsymbol{V}_{s,t}^*$ can also be estimated consistently, by estimating the set of cross-section ordered probits simultaneously, allowing for cross-equation residual correlation.

To establish identification in the most straightforward way, consider individuals who are first observed at $t = 0$ and for at least two further periods. In period 0 we can estimate a system of cross section ordered probit regressions of $\boldsymbol{y}_0$ on $(\boldsymbol{z}_0, \boldsymbol{x}_0)$. The proposition, together with inspection of (20) reveals that this identifies the coefficient matrices $\boldsymbol{CG}$ and $\boldsymbol{D}$ and the covariance matrix $\boldsymbol{V}_{00}^*$, given the non-collinearity assumption A1. By assumption A4, the first R rows of $\boldsymbol{CG}$ are equal to $\boldsymbol{G}$, which is therefore identified. The second block of $M - R$ rows are equal to $\boldsymbol{C}_2\boldsymbol{G}$ and, since $\boldsymbol{G}$ has rank R, these can be solved uniquely for $\boldsymbol{C}_2$, which is therefore identified. In period 1, a system of cross section ordered probit regressions of $\boldsymbol{y}_1$ on $(\boldsymbol{G}\boldsymbol{z}_0, \boldsymbol{z}_1)$ with displacement vector $\boldsymbol{w} = \boldsymbol{D}\boldsymbol{x}_1$ identifies the coefficient matrices $\boldsymbol{CA}$ and $\boldsymbol{CB}$. The first R rows of these are $\boldsymbol{A}$ and $\boldsymbol{B}$, which are therefore also identified.

With $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ and $\boldsymbol{D}$ identified from two waves of data, it remains to establish the identifiability of the covariance structure $\boldsymbol{\Sigma}_{\eta\eta}, \boldsymbol{\Sigma}_{uu}, \boldsymbol{\Sigma}_{u\eta}, \boldsymbol{\Sigma}_{\varepsilon\varepsilon}, \boldsymbol{\Sigma}_{\zeta\zeta}$ from the residual covariances of these cross-section multi-equation ordered probit models. Let $(\boldsymbol{V})^+$ denote the $R \times R$ principal submatrix of any matrix $\boldsymbol{V}$. Then:

$$(\boldsymbol{V}_{00}^*)^+ = \boldsymbol{\Sigma}_{\eta\eta} + (\boldsymbol{\Sigma}_{\zeta\zeta})^+ \tag{31}$$

and, since $(\boldsymbol{\Sigma}_{\zeta\zeta})^+$ is diagonal, the off-diagonal elements of $\boldsymbol{\Sigma}_{\eta\eta}$ are identified directly from $(\boldsymbol{V}_{00}^*)^+$. Now consider the sub-diagonal elements of the whole of $\boldsymbol{V}_{00}^*$ written in vectorised form:

$$\boldsymbol{S}\text{vec}\left((\boldsymbol{V}_{00}^*)^+\right) = \boldsymbol{S}\left[\boldsymbol{C}\otimes\boldsymbol{C}\right]\text{vec}\left(\boldsymbol{\Sigma}_{\eta\eta}\right) \tag{32}$$

where $\text{vec}(\boldsymbol{\Sigma}_{\eta\eta})$ is the operator that stacks the rows of $\boldsymbol{\Sigma}_{\eta\eta}$ into a column vector and $\boldsymbol{S}$ is the $M(M-1)/2 \times M^2$ matrix of 1s and 0s that selects the sub-diagonal elements from $\text{vec}\left(\boldsymbol{V}_{00}\right)$. Note that $\boldsymbol{S}\text{vec}\left(\boldsymbol{\Sigma}_{\zeta\zeta}\right) = \boldsymbol{0}$. Now write $\text{vec}\left(\boldsymbol{\Sigma}_{\eta\eta}\right) = \boldsymbol{T}_1\boldsymbol{d} + \boldsymbol{T}_2\boldsymbol{s}$, where $\boldsymbol{d}$ is the vector of diagonal elements of $\boldsymbol{\Sigma}_{\zeta\zeta}$, $\boldsymbol{s}$ is the vector of subdiagonal elements and $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ are corresponding matrices of 1s and 0s. Then:

$$\boldsymbol{S}\left[\boldsymbol{C}\otimes\boldsymbol{C}\right]\boldsymbol{T}_1\boldsymbol{d} = \boldsymbol{S}\left\{\text{vec}\left((\boldsymbol{V}_{00}^*)^+\right) - \left[\boldsymbol{C}\otimes\boldsymbol{C}\right]\boldsymbol{T}_2\boldsymbol{s}\right\} \tag{33}$$

It follows that $\boldsymbol{d}$ can be recovered uniquely if $\text{rank}\left(\boldsymbol{S}\left[\boldsymbol{C}\otimes\boldsymbol{C}\right]\boldsymbol{T}_1\right) \geq R$. Since $\boldsymbol{S}, \left[\boldsymbol{C}\otimes\boldsymbol{C}\right]$ and $\boldsymbol{T}_1$ are of rank $M(M-1)/2, R^2$ and R respectively, the diagonal elements of $\boldsymbol{\Sigma}_{\eta\eta}$ are identified provided $M(M-1)/2 \geq R$, which is satisfied whenever $M > R$ or $M = R > 2$.

Then observe from (27) and (28) that the following identity holds:

$$\boldsymbol{\Sigma}_{\zeta\zeta}^+\boldsymbol{A}' = \boldsymbol{V}_{02}^+ - \boldsymbol{V}_{01}^+(\boldsymbol{I}-\boldsymbol{A})' + \boldsymbol{V}_{00}^+\boldsymbol{A}' \tag{34}$$

Since $\boldsymbol{\Sigma}_{\zeta\zeta}^+$ is diagonal and every row of $\boldsymbol{A}'$ has at least one non-zero element, the first R variance parameters on the diagonal of $\boldsymbol{\Sigma}_{\zeta\zeta}$ are identified by (34).

With $\boldsymbol{\Sigma}_{\eta\eta}$ and $\boldsymbol{\Sigma}_{\zeta\zeta}^+$ determined, the remaining covariance parameters can be identified as follows:

$$\boldsymbol{\Sigma}_{\eta u} = \boldsymbol{V}_{01}^+ - \boldsymbol{\Sigma}_{\eta\eta}\boldsymbol{A}' \tag{35}$$

$$\boldsymbol{\Sigma}_{uu} = \boldsymbol{V}_{12}^+ - \boldsymbol{V}_{11}^+\boldsymbol{A}' - \boldsymbol{A}\boldsymbol{\Sigma}_{\eta u} \tag{36}$$

$$\boldsymbol{\Sigma}_{\varepsilon\varepsilon} = \boldsymbol{V}_{11}^+ - \boldsymbol{A}\boldsymbol{\Sigma}_{\eta\eta}\boldsymbol{A}' - \boldsymbol{\Sigma}_{uu} - \boldsymbol{A}\boldsymbol{\Sigma}_{\eta u} - \boldsymbol{\Sigma}_{\eta u}'\boldsymbol{A}' - \boldsymbol{\Sigma}_{\zeta\zeta}^+ \tag{37}$$

$$\boldsymbol{\Sigma}_{\zeta\zeta}^{++} = \boldsymbol{V}_{00}^+ - \boldsymbol{C}_2\boldsymbol{\Sigma}_{\eta\eta}\boldsymbol{C}_2' \tag{38}$$

where $\boldsymbol{\Sigma}^{++}$ denotes the submatrix consisting of the last $M-R$ rows and columns of $\boldsymbol{\Sigma}$.

Appendix 2: Construction of covariates

Table A1 Definitions and sample properties of covariates

Covariate	**Construction of variable**	**Mean**
Age	(current age - 10)/10	0.403
Female	dummy: reference category = male	0.488
Family trouble with law	dummy: reference category = no sibling has been in trouble with the police	0.066
Not two parents	(age < 17 only) dummy: reference category = currently brought up by natural mother + natural father	0.230
Stepfather	(age < 17 only) dummy: reference category = currently brought up by natural mother + step-father	0.102
Parental interest	number of items from following list that parents would mind about: fighting, graffiti, truancy, cannabis smoking	3.27
School discipline weak	(age < 17 only) number of items from following list: no clear rules on behaviour; easy to truant; violence against teachers	0.378
Parental social class 1-3	dummy: household reference person has NS-SEC social class 1-3	0.430