Pudney, Stephen

Working Paper

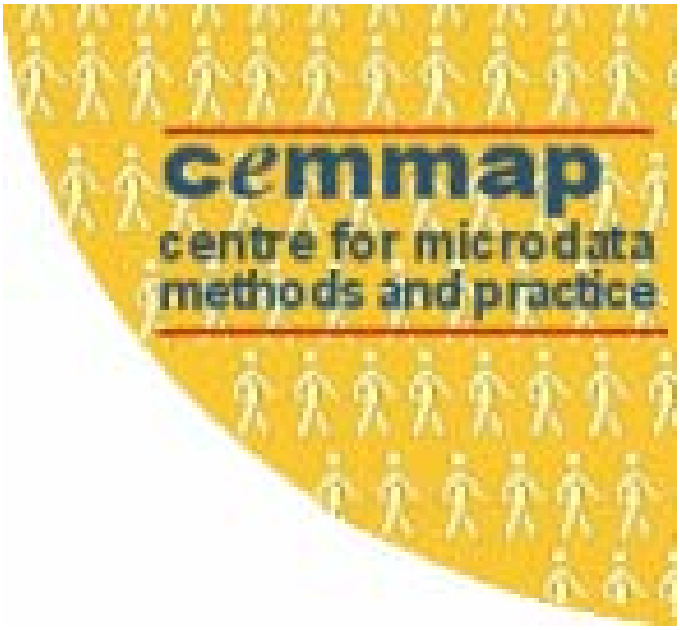
Rarely pure and never simple: Extracting the truth from self-reported data on substance use

cemmap working paper, No. CWP11/07

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Pudney, Stephen (2007) : Rarely pure and never simple: Extracting the truth from self-reported data on substance use, cemmap working paper, No. CWP11/07, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2007.1107

This Version is available at:
https://hdl.handle.net/10419/79364



RARELY PURE AND NEVER SIMPLE: EXTRACTING THE TRUTH FROM SELF-REPORTED DATA ON SUBSTANCE USE

Stephen Pudney

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP11/07

Rarely pure and never simple:
Extracting the truth from self-reported data on substance use

Stephen Pudney
Institute for Social and Economic Research
University of Essex

September 2006

ACKNOWLEDGEMENTS: Oscar Wilde's contribution gratefully acknowledged. I am also extremely grateful to Gabriella Conti for drawing my attention to the OCJS and BCS70 recanting problem and to her, Annette Jäckle, Heather Laurie and Peter Lynn for helpful discussions. This work was supported by the Economic and Social Research Council through the ULSC and MiSoC Research Centres (award nos. H562255004 and RES518285001). I bear sole responsibility for the conclusions and any errors.

ABSTRACT: We consider the misreporting of illicit drug use and juvenile smoking in self-report surveys and its consequences for statistical inference. Panel data containing repeated self-reports of 'lifetime' prevalence give unambiguous evidence of misreporting as 'recanting' of earlier reports of drug use. The identification of true initiation and reporting processes from such data is problematic in short panels, whilst more secure identification is possible in panels with at least five waves. Nevertheless, evidence from three UK datasets clearly indicates serious underreporting of cannabis, cocaine and tobacco use by young people, with consequent large biases in statistical modelling.

KEYWORDS: illicit drugs, smoking, misreporting, measurement error, nonparametric identification, OCJS, BCS70, BHPS

JEL CLASSIFICATION: C41, D12, I19

ADDRESS FOR CORRESPONDENCE: Stephen Pudney, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK; e-mail: spudney@essex.ac.uk.

1 Introduction

There is a huge research literature on "substance use"[1]- the consumption of tobacco, alcohol and illicit drugs - motivated primarily by concerns about their health impacts. Much of this impressive research edifice is based on self-report surveys which invite respondents to give details of their histories of drug use; see Pacula (1997), Kenkel *et. al.* (2001), Pudney (2003, 2004) and Van Ours (2003, 2007) for a representative sample of recent literature. An obvious objection to this research methodology is the strong possibility that survey respondents misreport, and particularly under-report, their drug use. This worry about data quality persists in spite of the use of computer-assisted audio interviewing methods (A-CASI), designed to ensure confidentiality of the interview process (Lessler and O'Reilly,1997). In some special cases, external bio-assay checks have suggested high rates of under-reporting (Magura and Kang, 1997; Lu *et. al.*, 2001) but these studies are based on highly selected samples gathered in difficult circumstances such as police custody. Others have examined the internal consistency of responses to groups of questions in a single cross-section survey questionnaire (Biemer and Wiesen, 2002) but most questionnaires do not allow this possibility and differences in the phrasing of questions make interpretation uncertain. The small amount of work done previously on contradictory responses to identical questions in a re-interview sequence ("recanting") reveals a modest rate of definite misreporting for illegal drugs in the US (Johnston and O'Malley, 1997) but that work does not attempt to estimate the full extent of misreporting and therefore tends to underestimate the scale of the problem.

Although our analysis is related to the statistical literature on measurement error (see Carroll *et. al.*, 1995), much of that literature assumes that misreporting is the outcome of an essentially unbiased statistical error process, rather than a pattern of behaviour reflecting the strong incentives that may exist (mainly) to under-report. Particularly in the case of sensitive issues like illicit drug use and under-age smoking, it is important to allow for the possibility of asymmetric reporting error. We do indeed find strong evidence of biased reporting behaviour.

Our aim is to examine evidence on the rate of misreporting and its impact on measured prevalence, using observations from recent UK panel and

[1]For simplicity of terminology, we use the word "drug" rather than "substance" henceforth.

cohort surveys. We analyse responses to lifetime prevalence questions of the form "have you ever...?". Self-contradictions in the observed sequences of responses give unambiguous evidence of misreports in some cases, which greatly assists with identification of the error rate and underlying initiation process. Section 2 of the paper outlines the three surveys we use here and the extensive evidence they give for the existence of misreporting. Panel length is an important issue and it is surprisingly rare for panel studies to contain long runs of repeated lifetime prevalence questions. Two of our three surveys have only two waves containing information on lifetime drug use. Section 3 examines the important case of two-wave panels and shows that identification is generally only available in the form of bounds on the parameters of interest. One of our objectives is to determine the width of these bounds and thus the degree of uncertainty associated with this kind of self-report data. The main part of this analysis, focusing on illegal drugs, rests on the critical assumption that there are no false positive reports. Evidence from responses to dummy questions about a fictional drug ("semeron") suggests that false positive responses are indeed very rare (Hamlyn *et. al.*, 2003). However, we also compare results obtained under this assumption with those obtainable under the more conventional (and, we argue, less plausible) assumption of two-sided, serially independent, random misclassification, which is an extension to the panel data context of a models previously used in the econometric literature (Bollinger, 1996; Lewbel, 2000).

Section 4 of the paper extends the analysis to longer panels and the specific case of under-age smoking as reported by young people in a panel for which up to five waves are available. Here, a two-sided misclassification model can be estimated, allowing also for some persistence in reporting behaviour. We find that under-reporting is again the dominant form of measurement error.

2 Evidence of under-reporting in UK surveys

There are few sources of re-interview data containing repeated questions on drug use. In this paper we use three UK surveys: the 2003-4 Offending, Crime and Justice Survey (OCJS), which is a conventional 2-wave annual panel; the 1970 British Cohort Survey (BCS70) which repeats the same drug use question in two waves widely separated in time; and the youth sample of the British Household Panel Survey, (BHPS) which contains up to five

interviews but covers smoking rather than illegal drug use.

2.1 The OCJS: prevalence of cannabis and cocaine

The OCJS was designed as a self-report survey of household-resident individuals, giving a wide range of information on respondents' drug use behaviour and other forms of illicit activity.[2] Fieldwork took place in January-July 2003 and respondents aged 10-25 were re-interviewed a year later. The initial survey had a 74% response rate and there was a 18.1% attrition rate in the panel subsample (of which 7.1% were refusals). Interviewing on sensitive issues was conducted using A-CASI. The OCJS contains a longitudinal element covering the 10-25 age group and we use this re-interview sample, containing up to 3,363 cases. We concentrate on two survey questions, asked initially in 2003 and then repeated in 2004:

> *"Have you ever taken cannabis (also known as marijuana, grass, hash, ganja, blow, draw, skunk,weed, spliff)?"*
>
> *"Have you ever taken cocaine (also known as charlie, C)?"*

The possible responses are: 1 = Yes, 2 = No, 3 = Don't Know, 4 = Don't want to answer. The two non-response categories together amount to only 0.61% and 0.44% of the sample for cannabis and cocaine respectively. These cases have been dropped from the analysis, as have the very small number of individuals who claim some previous use of the fictitious drug "semeron".[3]

Table 1 shows the transition matrix summarising the observed changes between the two years of interview. Conflicts are immediately apparent: 3.3% of the 2003 sample say they have previously used cannabis but then contradict that answer in 2004; for the less prevalent cocaine, this conflict occurs in 0.9% of the sample. Expressing these as transition rates, 13.0% of

[2]Release details: Home Office (Research, Development and Statistics Directorate, Offending Surveys and Research), National Centre for Social Research and BMRB (Social Research): Offending, Crime and Justice Survey, 2003 and 2004 [computer files]. Colchester, Essex: UK Data Archive [distributor], SN: 5248 (October 2005) and 5374 (July 2006).

[3]It is tempting to interpret "don't know" and "don't want to say" as disguised "No" and "Yes" respectively. Under this interpretation, 2004 measured prevalence rises from 28.7% to 29.1% for cannabis and 6.2% to 6.4% for cocaine. The subsequent analysis is not changed substantially by redefining prevalence rates in this way.

2003 self-declared cannabis users contradicted themselves in 2004, and 18.3% of cocaine users.

Table 1 Transition rates for self-reported lifetime prevalence; OCJS, BCS70 and BHPS samples

		2004 OCJS cannabis		
		Yes	No	n
2003 OCJS	Yes	0.870	0.130	830
cannabis	No	0.102	0.898	2437
n		970	2297	3267
		2004 OCJS cocaine		
		Yes	No	n
2003 OCJS	Yes	0.817	0.183	164
cocaine	No	0.024	0.976	3122
n		209	3077	3286
		2000 BCS70 cannabis		
		Yes	No	n
1986 BCS70	Yes	0.930	0.070	357
cannabis	No	0.469	0.531	4,982
n		2,670	2,669	5,339
		BHPS: smoked before current interview		
		Yes	No	n
BHPS: smoked before	Yes	0.926	0.074	2,240
previous interview	No	0.202	0.798	5,267
n		3,139	4,368	7,507

Note: all samples exclude individuals who claim any use of semeron; very small numbers of "don't know" responses were also excluded.

2.2 BCS70: cannabis prevalence

The BCS70 has followed through time a cohort of people born during the first week in April 1970.[4] The sweeps at age 16 and 30 both contained questions

[4]Release details: Butler, N. and Bynner, J.M., 1970 British Cohort Study : 16-year follow-up, 1986; and Joint Centre for Longitudinal Research, BCS70 follow-up 1999-2000

on drug prevalence. We restrict attention to cannabis because of the small number of positive responses at age 16 for other drugs. In both sweeps, the survey instrument was a postal questionnaire. At age 16, a randomised list anonymisation device was used in the following question.

> *"The next question [...] asks whether or not you have tried a number of substances some of which would under some circumstances be against the law. These are mixed in with a number of sporting activities and we have scrambled these by putting them into two lists - list A and list B. Please look at the box on this page to see whether you are to use list A or B when answering [...]. Please memorise whether it is list A or list B you are to use then erase the letter A or B with ink. Then proceed to use the list indicated for answering [...] Remember that nobody except you and us will know which list you are using*

The list of 16 activities following this question involves 7 types of illicit drug plus the fictional substance semeron and 8 sporting activities. Cannabis is the seventh item in the list.

At age 30, the following more conventional question was used:

> *As you know, many people have experimented with drugs at sometime. Have you ever tried cannabis, also known as blow, draw, puff, grass, skunk, weed, black, hash or red seal?*

The transitions between states nominated at ages 16 and 30 are summarised in Table 1; the BCS70 recanting rate of 7% is considerably lower than the 13% rate found in the OCJS panel.

This large difference may tell us something about the factors underlying self-contradictory reporting. The drug use questions are retrospective and recall error is a potential problem with such questions. A common finding is that recall error (in the form of non-recall of actual events) is more serious for more remote events, with a roughly linear decay profile (Lynn *et. al.*, 2005). This is completely inconsistent with the lower BCS70 recanting rate and suggests that recall error is not the primary source of the under-reporting problem. Differences in the BCS70 questions at age 16 and 30 may be relevant: for example, the age 16 question makes explicit reference to illegality,

[computer files], 2nd edition. Colchester, Essex: UK Data Archive [distributor], January 2003, SN: 3535 and 4396.

while the age 30 question hints at social acceptability. Attrition may also be an issue: it seems likely that subjects who are reliable participants also tend to be reliable reporters of their behaviour. However, the between-wave attrition rate for the OCJS was 18%, whereas only 3% of the BCS70 sample was lost between the age 16 and age 30 interviews. The relative maturity of the BCS70 means that panel conditioning may also be a contributory factor. However, the most persuasive explanation for the lower recanting rate in the BCS70 relates to the characteristics of the respondents and their incentives to misreport. At re-interview, they were aged 30 and, for those at risk of recanting, were answering questions relating to their behaviour of at least 14 years earlier, whereas OCJS respondents were, on average, much younger and responding to questions about relatively recent behaviour. It seems likely that people tend to be less sensitive about the remote past because it is less relevant to current self-image. Moreover, older people are likely, on average, to be more self-confident and consequently less concerned about the risk and consequences of disclosure.

2.3 BHPS: children's initiation into smoking

The BHPS is a nationally-representative annual household panel survey that began in the UK in 1991. Since 1994, children aged 11-15 have been included in the interviewing process, by means of a self-completion questionnaire covering a wide range of issues. Child respondents are asked to complete the questionnaire in privacy as far as possible, while face-to-face interviews are in progress with other household members. No information is available on the circumstances in which the questionnaire was completed. However, a significant change in the interviewing method occurred during our sample period. Before 2001, an audio questionnaire was used, with paper-based self-completion. Since 2001, the whole process has been paper-based, following reports that respondents found the audio questionnaire too time-consuming. There was no change in the questions themselves.

In Britain, it is not illegal for children to smoke but it is illegal to sell tobacco to anyone under the age of 16 or to purchase tobacco on behalf of a child. Thus smoking has much the same illicit character in this sample as illegal drug use has in the OCJS and BCS70 samples. The BHPS youth questionnaire contains a series of questions on smoking, the first of which is:

Have you ever tried a cigarette, even if it was only a single puff?

The sample transition rates are summarised in Table 1, which is constructed from the full sample 1994-2003; the BHPS recanting rate is 7% overall. The empirical recanting rate declines monotonically with age, falling from 16.7% at age 12 to 4.4% at age 15. The recanting rate is lower for females (5.2%) than males (9.7%), but the overall reported incidence is higher for females (37.1% compared with 34.6%, averaged over the 11-15 age range).

3 Two-wave panels

The length of a panel is critical to the indentifiability of the population processes of initiation and reporting. We begin with the case of a two-wave reinterview survey like the OCJS or BCS70.

3.1 Identification with under-reporting

The two time periods are indexed by $t = 0, 1$ and we define a binary variable Y_t equal to 1 if the respondent reports having used the drug prior to time t and 0 otherwise. The corresponding true drug status is D_t, which may differ from Y_t. We initially make the assumption that misreporting only takes the form of denial of drug use and thus $Y_t \leq D_t$. The analysis is conditional on observed covariates, which are of three kinds: X contains variables influencing drug-taking behaviour but not reporting behaviour; Z is a set of variables influencing both drug use and reporting; W contains variables influencing the propensity to misreport but not drug use itself. The function $\Pi_{jk}(X, Z, W)$ gives the probability of $Y_0 = j$ and $Y_1 = k$ $(j, k = 0, 1)$, conditional on X, Z, W. The analogous probabilities for true drug use are $P_{jk}(X, Z) = \Pr(D_0 = j, D_1 = k|X, Z)$, where $P_{10} = 0$, since 'past drug user' status is irreversible. For someone who has used drugs prior to time 0, define the probabilities $\Omega_{rs}(Z, W) = \Pr(Y_0 = r, Y_1 = s|D_0 = D_1 = 1, Z, W)$. For someone initiated into drug use between times 0 and 1, define $\omega_r(Z, W) = \Pr(Y_1 = r|D_0 = 0, D_1 = 1, Z, W)$. The probabilities of the possible observable outcomes are then:

$$\Pi_{01}(X, Z, W) = P_{11}(X, Z)\Omega_{01}(Z, W) + P_{01}(X, Z)\omega_1(Z, W) \quad (1)$$

$$\Pi_{10}(X, Z, W) = P_{11}(X, Z)\Omega_{10}(Z, W) \quad (2)$$

$$\Pi_{11}(X, Z, W) = P_{11}(X, Z)\Omega_{11}(Z, W) \quad (3)$$

The analogous equation for Π_{00} is redundant since the Π_{jk} sum identically to unity. Identification is subject to the following constraints:

$$\begin{aligned} P_{00}(X,Z) + P_{11}(X,Z) + P_{01}(X,Z) &= 1 && (4) \\ \sum_{j=0}^{1}\sum_{k=0}^{1} \Omega_{jk}(Z,W) &= 1 && (5) \\ \Omega_{rs}(Z,W) &\geq 0 && (6) \\ P_{jk}(X,Z) &\geq 0 && (7) \\ \omega_1(Z,W) &\geq 0 && (8) \end{aligned}$$

Local identification at a point (X, Z, W) is clearly problematic, since there are only five equality conditions to determine the eight unknowns P_{00}, P_{01}, $P_{11}, \Omega_{00}, \Omega_{01}, \Omega_{10}, \Omega_{11}, \omega_1$. In general, this means that only interval rather than point identification is available. Identification can be made sharper by means of exclusion restrictions (so that X and W are non-null) and by adding further *a priori* restrictions.

A possibility for the latter is a *homogeneity* assumption: independence of the misreporting distribution in period 1 and the history of drug use, so that $\Pr(Y_1 = 0|D_0 = D_1 = 1|Z, W) = \Pr(Y_1 = 0|D_0 = 0, D_1 = 1|Z, W)$ and thus:

$$\omega_1 = \Omega_{01} + \Omega_{11} \tag{9}$$

Another plausible restriction is *exchangeability*:

$$\Omega_{01} = \Omega_{10} \tag{10}$$

which is equivalent to a random effects structure for the misreporting process, with a persistent individual-specific component and an idiosyncratic time-varying component.

Another potential identifying restriction is serial independence, implying a rank 1 covariance structure $\Omega_{rs} = \Omega_r^0 \Omega_s^1$. Together with (9) or (10), this would give exact identification of all parameters. However, serial independence is not in the spirit of the under-reporting model. If there is a systematic tendency among some individuals to conceal past drug use, it is highly likely that this tendency will persist over time. Thus we do not assume serial independence in the case of under-reporting. We examine the polar opposite

case of unsystematic, serially-independent random misclassification in section 5 below.

Neither the homogeneity and exchangeability assumptions, nor exclusion restrictions are generally sufficient to give exact identification. The following identification result, proved in Appendix 1, gives necessary and sufficient conditions.

Proposition 1 *In the structure (1)-(10), the functions $P_{jk}(X,Z)$, $\Omega_{rs}(Z,W)$ and $\omega_1(Z,W)$ are locally identified at the point $Z=z$ if and only if there exist points $w \in S_{W|Z=z}$ and $x \in S_{X|Z=z}$ such that $\Omega_{00}(Z,w) = P_{00}(x,Z) = 0$, where $S_{W|Z=z}$ and $S_{X|Z=z}$ are the support sets for W and X conditional on $Z=z$.*

This result implies that, even if there are variables X and W excluded respectively from the drug use and reporting behaviour models and we continue to make the homogeneity and exchangeability assumptions, a further strong condition is required: among people of every type z, there must be some whose observable characteristics make them certain to report accurately in at least one of the two periods, and others who are certain to be drug users by the time of the period 1 interview. It is hard to be confident about either of these two conditions, so we must work with interval identification. We consider three specific parameters of particular interest: the misreporting rate, defined equivalently as $\omega_0 = 1-\omega_1 = \Omega_{01}+\Omega_{00} = \Omega_{00}+\Omega_{10}$; the initial prevalence rate, P_{11}; and the initiation or hazard rate, $h = P_{01}/(P_{01}+P_{00})$. Proposition 2, which is proved in the appendix, establishes bounds on these parameters in the case where we impose homogeneity and exchangeability.

Proposition 2 *In the structure (1)-(10), ω_0, P_{11} and h satisfy the following inequalities:*

$$\max_{X \in S_{X|Z,W}} \left(\frac{\Pi_{10}}{2\Pi_{10}+\Pi_{11}} \right) \leq \omega_0(Z,W) \leq \min_{X \in S_{X|Z,W}} (1-\Pi_{01}-\Pi_{11})$$

$$\max_{W \in S_{W|X,Z}} (2\Pi_{10}+\Pi_{11}) \leq P_{11}(X,Z) \leq \min_{W \in S_{W|X,Z}} \left(\frac{\Pi_{10}+\Pi_{11}}{\Pi_{01}+\Pi_{11}} \right)$$

$$\max_{W \in S_{W|X,Z}} \left(\frac{(\Pi_{10}+\Pi_{11})(2\Pi_{10}+\Pi_{11})}{(\Pi_{01}+\Pi_{11})(1-2\Pi_{10}-\Pi_{11})} \right) \leq h(X,Z) \leq 1$$

where $S_{X|Z,W}$ and $S_{W|X,Z}$ are the conditional support sets of X and W. These are the tightest possible bounds in the absence of further a priori information.

In proposition 2, homogeneity essentially serves to fix the value of ω_1 in relation to the Ω_{jk} and is fairly innocuous. However, exchangeability has the implication that $\Pi_{01} \geq \Pi_{10}$, which might be violated empirically in part of the support of (X, Z, W). Removing the exchangeability restriction gives the following wider bounds:

Proposition 3 *In the structure (1)-(9), the tightest possible bounds on* ω_0, P_{11} *and* h *are:*

$$\max_{X \in S_{X|Z,W}} \left(\frac{\Pi_{10}}{1 - \Pi_{00}} \right) \leq \omega_0(Z, W) \leq \min_{X \in S_{X|Z,W}} (1 - \Pi_{01} - \Pi_{11})$$

$$\max_{W \in S_{W|X,Z}} \max \left(\Pi_{10} + \Pi_{11}, \frac{\Pi_{10}}{1 - \Pi_{01} - \Pi_{11}} \right) \leq P_{11}(X, Z) \leq 1$$

$$0 \leq h(X, Z) \leq 1$$

These striking results are alarming at first sight. If we are interested in the process of initiation into drug use, the existence of under-reporting completely destroys the possibility of drawing any inferences about the hazard rate h, in the absence of further prior information. Similarly, the data can say little about the population prevalence of drug use, since only a lower bound on prevalence is available. Under the assumptions of proposition 3, this lower bound on prevalence is at least as great as $\Pi_{10} + \Pi_{11}$, the proportion of people admitting to drug use in period 0. The adjustment procedure used by Johnston and O'Malley (1997) uses this quantity and thus tends to under-adjust for misreporting.

3.2 Unconditional estimates

We begin with unconditional estimation, so that the vector (X, Z, W) is empty. The estimated bounds are given in Tables 3 and 4 for OCJS cannabis and cocaine and BCS70 cannabis. The bounds are wide. For example, in Table 3 the 2004 OCJS misreporting rate by established cannabis users lies in the estimated interval (0.115, 0.703), even when we impose homogeneity and exchangeability. However, the upper end of this interval has implications that can be ruled out *a priori*. The misreporting rate can only achieve its upper bound, 0.703, if over 85% of people aged 10-25 were cannabis

users prior to interview in 2003 and nobody at all was drug free in 2004 ($\widehat{P}_{00} = 0; \widehat{P}_{01} = 0.144; \widehat{P}_{11} = 0.856$). The lower end of the interval is more reasonable: a misreporting rate of 0.115 carries the implication that two-thirds of the 10-25 population remains cannabis-free in 2004, fewer than 5% first try cannabis between the 2003 and 2004 interviews and the remainder, under a third, had tried cannabis before 2003 ($\widehat{P}_{00} = 0.664; \widehat{P}_{01} = 0.048; \widehat{P}_{11} = 0.288$). A weak 'reality check' prior constraint could clearly narrow the range of uncertainty considerably. The bounds for cocaine are wider than for cannabis, with an interval for the 2004 misreporting rate of (0.154, 0.936) when both homogeneity and exchangeability are imposed. Again, the estimates are plausible near the lower bound (where $\widehat{P}_{00} = 0.925; \widehat{P}_{01} = 0.016; \widehat{P}_{11} = 0.059$) but not at the upper bound, where ($\widehat{P}_{00} = 0; \widehat{P}_{01} = 0.215$ and $\widehat{P}_{11} = 0.785$).

Table 3 Bounds analysis for OCJS cannabis and cocaine ($n = 3,267$ and $3,286$; standard errors in parentheses)

	Additional restrictions			
	CANNABIS			
Bound	None	Homogeneity	Exchangeability	Both
	Misreporting rate in 2004			
Lower	0.100	0.100	0.115	0.115
	(0.009)	(0.009)	(0.009)	(0.009)
Upper	0.761	0.703	0.735	0.703
	(0.008)	(0.008)	(0.008)	(0.008)
	True 'ever used' prevalence in 2003			
Lower	0.254	0.254	0.287	0.287
	(0.008)	(0.008)	(0.009)	(0.009)
Upper	1	1	0.957	0.856
	(-)	(-)	(0.006)	(0.018)
	True initiation rate 2003-4			
Lower	0	0	0.060	0.068
	(-)	(-)	(0.019)	(0.021)
Upper	1	1	1	1
	(-)	(-)	(-)	(-)
	COCAINE			
	Misreporting rate in 2004			
Lower	0.125	0.125	0.154	0.154
	(0.021)	(0.021)	(0.022)	(0.022)
Upper	0.958	0.936	0.949	0.936
	(0.004)	(0.004)	(0.004)	(0.004)
	True 'ever used' prevalence in 2003			
Lower	0.050	0.050	0.059	0.059
	(0.004)	(0.004)	(0.005)	(0.005)
Upper	1	1	0.986	0.785
	(-)	(-)	(0.003)	(0.043)
	True initiation rate 2003-4			
Lower	0	0	0.015	0.017
	(-)	(-)	(0.007)	(0.008)
Upper	1	1	1	1
	(-)	(-)	(-)	(-)

Table 4 Bounds analysis for BCS70 cannabis
($n = 5,339$; standard errors in parentheses)

	Additional restrictions			
Bound	None	Homogeneity	Exchangeability	Both
	Misreporting rate at age 30			
Lower	0.009	0.009	0.066	0.066
	(0.002)	(0.002)	(0.012)	(0.012)
Upper	0.889	0.500	0.882	0.500
	(0.006)	(0.007)	(0.006)	(0.007)
	True 'ever used' prevalence at age 16			
Lower	0.067	0.067	0.072	0.072
	(0.003)	(0.003)	(0.004)	(0.004)
Upper	1	1	0.567	0.370
	(-)	(-)	(0.007)	(0.006)
	True initiation rate between ages 16 and 30			
Lower	0	0	0.467	0.499
	(-)	(-)	(0.015)	(0.020)
Upper	1	1	1	1
	(-)	(-)	(-)	(-)

Point identification could be achieved if one further constraint were introduced. However, there is no obvious theoretical argument to generate such a constraint. Instead, we examine how the estimated implied misreporting and initiation rates vary as we alter the parameter P_{00} over a pre-selected range of values. Figure 1 shows the resulting loci for OCJS cannabis over the reasonable range 45-70%[5] for the drug-free rate, P_{00}. This indicates a misreporting rate somewhere between 12% and 34%. The misreporting rate only falls below 15% in the small region $P_{00} \in [0.652, 0.663]$. This is strong evidence of serious misreporting, with a substantial effect on measured prevalence. For OCJS cocaine (Figure 2), the implied misreporting rate varies from a very high level of over 70% to 15% as we vary the drug-free rate P_{00} from 80% to 93%.

[5] There is no solution to the identifying equations for P_{00} above 0.66 (OCJS cannabis), 0.92 (OCJS cocaine) or 0.46 (BCS cannabis).

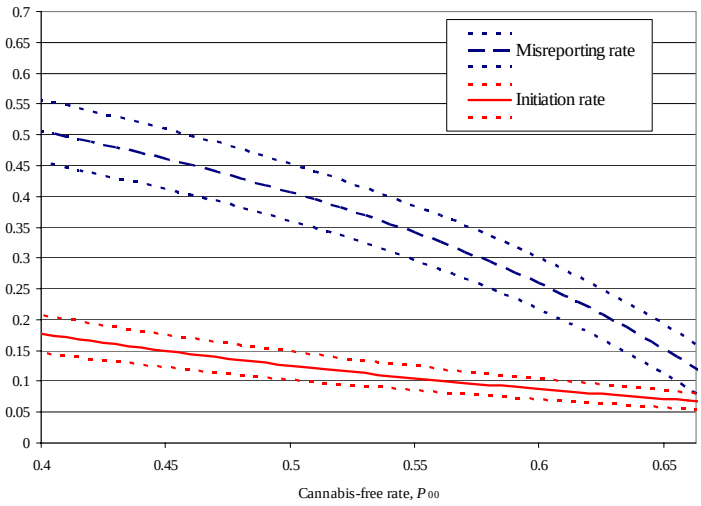


Figure 1 Estimates of the 2004 misreporting and initiation rates for alternative assumed values of the cannabis-free rate, P_{00}. (Dotted lines are 90% confidence intervals)

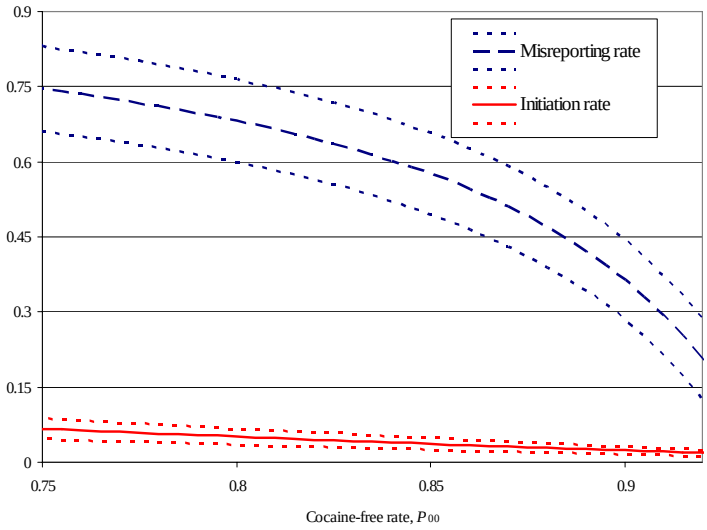


Figure 2 Estimates of the 2004 misreporting and initiation rates for alternative assumed values of the cocaine-free rate, P_{00}. (Dotted lines are 90% confidence intervals)

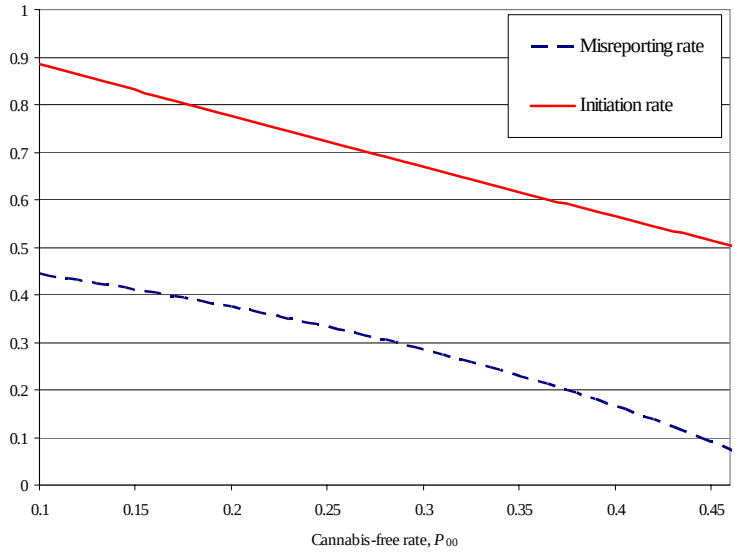


Figure 3 Estimates of the age 30 misreporting rate and the 16-30 initiation rate under alternative assumed values of the cannabis-free rate, P_{00} (confidence intervals too narrow to plot)

3.3 Conditional analysis

The first step in conditional analysis is to arrive at a good empirical approximation to the distribution $\Pi_{00}(X, Z, W)...\Pi_{11}(X, Z, W)$. We use a simple bivariate probit structure, using a set of covariates capturing some of the obvious potential influences on drug use. The covariates are defined and summarised in appendix Tables A1 and A2. RESET tests are used to check the specification of these models. The coefficients are given in appendix Tables A3 and A4;

It is difficult to find any variables that can confidently be assumed *a priori* to influence drug use but not reporting behaviour. Our strategy is to assume the misreporting probability to be uniform, except with respect to specific characteristics that we might expect to have some influence on reporting behaviour. These are: age, residence in the parental home, religious affiliation and ethnicity, which are assigned to Z. Other covariates associated with drug use (including employment status, locality and gender) are assumed independent of reporting behaviour. These variables are assigned

to X. We use the bounds set out in proposition 3, which does not impose the exchangeability assumption, and avoids the empirically invalid condition that $\Pi_{01}(X, Z, W) \geq \Pi_{10}(X, Z, W)$ everywhere.

There is strong evidence from other studies that interview conditions affect the nature of responses (Aquilino, 1997). In the OCJS analysis, we capture these contextual influences using two dummy variables for the presence of a parent during the interview and the respondent's need for interviewer help with the self-completion questionnaire. In the BCS70, drug use data come from a postal questionnaire and we have no information on the circumstances in which it was completed, so W is null.

The conditional bounds, based on the results of proposition 3, are summarised in Tables 5-7 for a set of hypothetical individuals. For the OCJS, we start from a baseline individual who is: male, aged 16 in 2003, living with parents in an owned house, not in work, with a self-reported religious affiliation and located in an area that is not perceived to have a particular drug problem. We contrast this individual with an economically-independent 21-year-old male, a baseline female, a baseline male but living in an area perceived to have a drug problem, and a male with a combination of risk factors (rented home, no religion, problem drug area). Table 5 gives the lower bound on 2003 prevalence for each of these individual types. The bound is relatively high for cannabis in certain groups, with prevalence estimated to be at least 62% for the highest risk group. For baseline females, the lower bound falls to below 26%. Bounds for the less prevalent drug cocaine are much lower but there is again a high degree of variation across individual types.

We also calculate a sample mean lower bound, calculated in the following way:

$$\widehat{\overline{P}}_{11}^{\min} = n^{-1} \sum_{i=1}^{n} \max_{W^0, W^1 \in S} \widehat{P}_{11}^{\min}(W^0, z_i^0, x_i^0, W^1, z_i^1, x_i^1) \tag{11}$$

where $W = (W^0, W^1), x_i = (x_i^0, x_i^1)$ and $z_i = (z_i^0, z_i^1)$ and superscript $t = 0, 1$ indicate the survey wave of observation. The set S is defined as the set of theoretically possible W-values, whether observed in the sample or not (the results are not altered materially if S is replaced by the empirical support set $\{W \mid \#(W, z_i, x_i) > 0\}$, where $\#(.)$ denotes the cell count).

Table 5 OCJS: estimated lower bounds on prevalence in 2003 (standard errors in parentheses)

Individual characteristics	**Cannabis**	**Cocaine**
Baseline 16-year-old	0.293	0.025
	(0.025)	(0.008)
Economically independent 21-year-old	0.562	0.158
	(0.039)	(0.042)
Female	0.258	0.012
	(0.024)	(0.005)
Drug-prevalent area	0.375	0.053
	(0.031)	(0.014)
High-risk individual	0.624	0.166
	(0.081)	(0.073)
Sample mean lower bound	0.258	0.050
Standard deviation of lower bound	0.225	0.065
Sample mean reported prevalence (2003)	0.252	0.049
$= n^{-1} \sum y_{0i}$	(0.008)	(0.004)

[1]Baseline: 16-year-old white male, with religious affiliation, in education, in parental-owned home, in median deprivation area with normal drug prevalence [2] As baseline, except 21-year-old , in work, living in own rented home [3] As baseline, except female [4] As baseline, but living in high-prevalence area [5] 16-year-old white male, previously in care, living in parental rented home, in high deprivation area with high drug prevalence

Table 6 gives OCJS bounds on the 2004 misreporting rate ω_0. The exclusion restrictions we have used are sufficient to reduce greatly the wide unconditional intervals in Table 3. We conclude that misreporting rates are at least 30-40% for some individual types. For cocaine, the range of uncertainty is greater and under-reporting may be much more serious than for cannabis. Note that it is possible for the empirical lower bound on ω_0 to exceed the empirical upper bound, due to parameter estimation error in the functions $\widehat{\Pi}_{rs}(X, Z, W)$ and we do observe this for cannabis in the case of economically independent 21-year-olds.

Table 6 OCJS: estimated bounds on misreporting rates
(standard errors in parentheses)

Individual characteristics	Cannabis	Cocaine
Baseline[1], completed questionnaire alone	[0.280, 0.366]	[0.168, 0.687]
without assistance	(0.105, 0.088)	(0.108, 0.110)
Baseline[1], parent present and need of	[0.354, 0.530]	[0.197, 0.822]
interviewer assistance	(0.218, 0.147)	(0.343, 0.103)
Economically independent[2] 21-year-old,	[0.264, 0.252]	[0.207, 0.446]
completed questionnaire without assistance	(0.075, 0.075)	(0.114, 0.120)
Economically independent[2] 21-year-old,	[0.310, 0.350]	[0.186, 0.589]
had need of interviewer assistance	(0.150, 0.134)	(0.280, 0.147)
No religion[3], completed questionnaire	[0.269, 0.334]	[0.230, 0.644]
alone without assistance	(0.098, 0.085)	(0.128, 0.115)
No religion[3], parent present and need	[0.344, 0.497]	[0.272, 0.789]
of interviewer assistance	(0.207, 0.147)	(0.422, 0.115)
Sample mean bounds	[0.230, 0.595]	[0.310, 0.950]
Standard deviations of bounds	[0.109, 0.250]	[0.376, 0.074]

[1]White 16-year-old in parental home, religious affiliation; [2] White 21-year-old in own home, religious affiliation; [3] White 16-year-old in parental home, no religious affiliation

Table 7 gives the results for the BCS70 case, where the two interview waves are separated by 14 years, our baseline individual is defined as a white male with religious affiliation and no history of local authority care or fostering who, at age 16, was not in work and lived in the (owned) parental home; at age 30 he was in work and living in an owned house. Departures from this baseline case explore the effects of differences in gender, religion, care and non-employment at 30.

The much lower incidence of recanting in the BCS70 data results in lower rates of under-reporting of cannabis use than we observe for the OCJS. The greater age of respondents at reinterview and the greater time separation between the two waves are plausible explanations for the apparently lower rate of under-reporting in the BCS70. The BCS70 analysis gives a much smaller lower bound on the 1986 prevalence rate, P_{11}, than we find for the OCJS in 2003. This is consistent with other evidence of strong growth in prevalence in the 1990s.

Table 7 BCS70: lower bound on prevalence and bounds on misreporting rates (standard errors in parentheses)

Individual characteristics (X, Z)	**Lower bound on prevalence rate**
Baseline[1]	0.052
	(0.006)
Female[2]	0.045
	(0.005)
In care[3]	0.074
	(0.045)
No religion[4]	0.077
	(0.013)
Out of work home renter at age 30	0.052
	(0.006)
Sample mean lower bound	0.067
Sample standard deviation of lower bound	0.017
Sample mean reported prevalence	0.067
	(0.003)
Individual characteristics (Z)	**Bounds on misreporting rate**
Baseline[5]	[0.065 , 0.233]
	(0.062 , 0.142)
No religion[6]	[0.045 , 0.144]
	(0.043 , 0.106)
Sample mean bounds	[0.042 , 0.323]
Standard deviation of bounds	[0.019 , 0.063]

[1]White male living in in parental home at 16 & own home at 30; religious affiliation; in education at 16 and in work at 30; never in care, owner-occupied housing at 16 & 30. [2] As baseline, but female. [3] As baseline, but has been in care; [4] As baseline, but no religious affiliation. [5]White, living in parental home at 16 & own home at 30; religious affiliation. [6]As baseline, but no religious affiliation.

3.4 Comparison with random misclassification

The analysis described above assumes a one-sided structure of misreporting that excludes false positive responses. This is different from the standard measurement error model. It is possible to extend the analysis to allow the possibility of false positive self-reports but, in a two-wave survey, without further strong assumptions identification is an insuperable problem. We resolve this by assuming serial independence in the misreporting process, so that reporting error is seen as unsystematic both over time and in terms of the direction of error. In this sense, the random misclassification assumption is the polar opposite of the one-sided, serially-dependent process assumed earlier. Define Ω_1 as the probability of misreporting for a drug user and Ω_0 as the corresponding probability for a non-user. Assume these to be constant over time, but not necessarily equal. This is essentially the same measurement error assumption, extended to the panel case, as used by Bollinger (1996) and Lewbel (2000) in their respective analyses of the regression model with mismeasured binary regressors and the binary choice model. The five unknown parameters satisfy the following four equalities:

$$\begin{aligned}
\Pi_{01} &= P_{00}\left(1-\Omega_0\right)^2 + P_{01}(1-\Omega_0)\Omega_1 + P_{11}\Omega_1^2 && (12)\\
\Pi_{10} &= P_{00}\left(1-\Omega_0\right)\Omega_0 + P_{01}(1-\Omega_0)\left(1-\Omega_1\right) + P_{11}\left(1-\Omega_1\right)\Omega_1 && (13)\\
\Pi_{11} &= P_{00}\left(1-\Omega_0\right)\Omega_0 + P_{01}\Omega_0\Omega_1 + P_{11}\left(1-\Omega_1\right)\Omega_1 && (14)\\
1 &= P_{00} + P_{01} + P_{11} && (15)
\end{aligned}$$

Sharp bounds on the two misreporting rates Ω_0 and Ω_1, the prevalence rate P_{11} and the initiation rate $h = P_{01}/(1-P_{11})$ are given by maximising or minimising the relevant quantity with respect to the parameters $P_{00}, P_{01}, P_{11}, \Omega_0$ and Ω_1, subject to equations (12)-(15) and inequalities constraining each parameter to the unit interval. These nonlinear programming problems are solved using an iterative constrained optimisation algorithm, with sample estimates $\widehat{\Pi}_{01}, \widehat{\Pi}_{10}$ and $\widehat{\Pi}_{11}$ substituted for the population probabilities. The results are given in Table 8.

Table 8 Estimated bounds on misreporting, prevalence and initiation rates under random misclassification assumptions (standard errors in parentheses)

Population characteristic	**OCJS**		**BCS70**
	cannabis	**cocaine**	**cannabis**
Misreporting rate for non-users (Ω_0)	[0 , 0.047] (- , 0.004)	[0 , 0.010] (- , 0.002)	[0 , 0.009] (- , 0.002)
Misreporting rate for users (Ω_1)	[0 , 0.130] (- , 0.012)	[0 , 0.182] (- , 0.030)	[0 , 0.070] (- , 0.014)
Prevalence rate (P_{11})	[0.217 , 0.292] (0.007 , 0.010)	[0.041 , 0.061] (0.003 , 0.005)	[0.058 , 0.072] (0.003 , 0.004)
Initiation rate ($P_{01}/(1-P_{11})$)	[0.057 , 0.070] (0.008 , 0.008)	[0.014 , 0.018] (0.003 , 0.004)	[0.464 , 0.502] (0.007 , 0.010)

The bounds on the prevalence and initiation rates are much tighter for this random misclassification model than for the under-reporting model used earlier (compare Tables 3 and 4). Given this good precision, there is little point in resorting to the stronger assumptions entailed by a conditional analysis. The main source of increased precision is the serial independence assumption used here, which would be less appropriate in a model based on systematic under-reporting. Nevertheless, the upper bound on the misreporting rate for users (Ω_1) is substantially larger than the rate for non-users (Ω_0), so this model is also consistent with a tendency towards under-reporting.

4 Longer panels: children's initiation into smoking

The BHPS youth sample gives up to five waves of data for any individual. From an initial period at time $t = 0$, the potentially mis-measured status Y_t evolves over a sequence of time periods $t = 1...T$. The underlying unobservable process for true status is $\{D_t\}$. We define period 0 to correspond to a sufficiently early age that it is safe to assume $D_0 = Y_0 = 0$ with probability 1. Note that the observation period during which data are collected may be any subset of periods $1...T$. However, we make a 'missing at random'

assumption, so that the selection of periods into the sample is independent of $\{D_t, Y_t\}$, conditional on the sequence of explanatory covariates $\{X_t\}$.

Since the transition from $D_{t-1} = 0$ to $D_t = 1$ is irreversible, the whole sequence $\{D_t\}$ can be represented by the transition date τ: a non-negative integer defined as the unique value of t for which $D_t > D_{t-1}$. We set $\tau = T+1$ in cases where the transition occurs after period T, so that τ is a discrete random variable with support $\{1...T+1\}$. This can be characterised in terms of the hazard rates $h_1...h_T$, giving:

$$\Pr(\tau|X_1...X_T) = h(X_\tau)^{I_1} \left[\prod_{s=1}^{\tau-1}(1-h(X_s))^{I_2}\right] \tag{16}$$

where $I_1 = 1$ if $\tau \leq T$ and 0 otherwise and $I_2 = 1$ if $\tau \geq 2$ and 0 otherwise.

The model allows for serial correlation in the reporting process. Specifically, we assume the following forms for the probabilities of false positive and false negative self-reports, conditional on X_t and the history, H_t, of the $\{D_t, Y_t\}$ process:

$$\Pr(Y_t = 1|D_t = 0, X_t, H_t) = p^+(\zeta^+_{t-1}, X_t) \tag{17}$$

$$\Pr(Y_t = 0|D_t = 1, X_t, H_t) = p^-(\zeta^-_{t-1}, X_t) \tag{18}$$

where $\zeta^+_t = (1-D_{t-1})Y_{t-1}$ and $\zeta^-_t = D_{t-1}(1-Y_{t-1})$ are indicators of false positive and false negative reports in the previous period. We do not explore other more persistent processes, since our observation period is short, covering only the age range 11-15.

If we could observe the whole sequence $\{(Y_t, X_t), t = 1...T\}$, the likelihood for a given individual would be:

$$L = \sum_{\tau=1}^{T+1}\left[h(X_\tau)^{I_1}\prod_{s=1}^{\tau-1}(1-h(X_s))^{I_2}\right]\Pr(Y_1...Y_T \,|\, \tau, X_1...X_T) \tag{19}$$

where the conditional probability of the reporting sequence is:

$$\begin{aligned}\Pr(Y_1...Y_T \mid \tau, X_1...X_T) = &\left[\prod_{t=1}^{\tau-1} p^+(0,X_t)^{Y_t}\left(1-p^+(0,X_t)\right)^{1-Y_t}\right]^{I_2} \\ &\times \left[p^-(0,X_\tau)^{1-Y_\tau}\left(1-p^-(0,X_t)\right)^{Y_t}\right]^{I_1} \\ &\times \left[\prod_{t=\tau+1}^{T} p^-(1-Y_{t-1},X_t)^{1-Y_t}\left(1-p^-(1-Y_{t-1},X_t)\right)^{Y_t}\right]^{I_1}\end{aligned} \tag{20}$$

Identification is a complicated issue in this serially-dependent, two-sided, multi-wave setting. For given $X_1...X_T$, the $2^T - 1$ independent observable probabilities of $\{Y_1...Y_T\}$ are polynomials in the $5T$ unknowns $h(X_t)$, $p^+(0, X_t)$, $p^+(1, X_t)$, $p^-(0, X_t)$ and $p^-(1, X_t)$. For the 5-wave BHPS panel, $2^T - 1 = 31 > 5T = 25$, so that the order condition, at least, is satisfied, even without exclusion restrictions.

There remains the difficulty that not all the Y_t are observable, either because of attrition or item non-response or because observation begins at a later age than a safe choice for the initial state. Since Y_t is binary, it is feasible to marginalise with respect to non-observed values by summation. Thus, if we observe no value of Y_t at dates $t_1...t_k$ the likelihood becomes:

$$L = \sum_{\tau=1}^{T+1} h(X_\tau)^{I_1} \left[\prod_{s=1}^{\tau-1}(1 - h(X_s))^{I_2}\right] \left[\sum_{Y_{t_1}} ... \sum_{Y_{t_k}} \Pr(Y_1...Y_T \mid \tau, X_1...X_T)\right] \tag{21}$$

In using this form, we are assuming that the covariates X_t are directly observed or can be constructed in all periods. This is automatically true for some variables like age or time but may require imputation in other cases.

In the implementation of this model, period 1 corresponds to age 8 so it is assumed that no-one smokes before their 8th birthday. The functions $h(.)$, $p^+(.)$ and $p^-(.)$ are specified as probits. The covariates in X represent the age and gender of the child, the year of observation,three dummy variables representing a 4-category socio-economic classification based on the higher occupational class of the two parents; and a binary variable indicating whether the child's mother had been observed as a smoker at any point in the observation period.[6] There is a single exclusion restriction, since the discontinuation of the taped questionnaire in 2001 (represented by a step-change dummy) affects only reporting behaviour and not true smoking. The dataset is summarised in Appendix Table A3. No difficulty was encountered in optimising the likelihood function and inverting the Hessian matrix at the optimum, so identification appears secure.

The estimated parameters are given in Tables 9 and 10, for two variants of the misreporting model. The simpler form of the model restricts the over-reporting probability p^- to be zero; the full specification uses probit forms for

[6] The father's smoking behaviour and more detailed social class variables were found to be insignificant.

each of p^+ and p^-. Table 9 compares the estimated hazard parameters from the two misreporting models with estimates obtained from a conventional discrete-time hazard model, where the transition date is taken to coincide with the first reported $0 \rightarrow 1$ transition in the time series $\{Y_t\}$. The two-sided misreporting model gives a much better sample fit than the one-sided model, which in turn is far superior to the simple hazard model. We find big differences in the hazard coefficients. In particular, the simple hazard model suggests that the initiation hazard rises strongly with age, that girls have a slightly higher hazard than boys and that there is a large positive demonstration effect of parental smoking. All of these findings are changed when we allow for misreporting: there is now no evidence of a rising hazard rate, a much larger and more significant gender difference and a much weaker impact of parental smoking. These differences are sufficient to demonstrate that it is not safe to neglect the possibility of misreporting bias in survey data.

Table 9 The impact of misreporting on BHPS hazard function parameter estimates ($n = 2622$; standard errors in parentheses)

	Misreporting model		**No reporting error**
Covariate	$p^+ = 0$	Variable p^+, p^-	$p^+ = p^- = 0$
Intercept	-1.307***	-1.425***	-1.665***
	(0.068)	(0.059)	(0.036)
Age	-0.020	0.490	3.233***
	(0.728)	(0.518)	(0.119)
Year	-0.014*	0.001	-0.038***
	(0.007)	(0.007)	(0.004)
Female	0.145***	0.312***	0.074**
	(0.043)	(0.054)	(0.031)
Managerial / professional parent	-0.141**	-0.113*	-0.122***
	(0.062)	(0.063)	(0.043)
Skilled white collar parent	0.000	0.008	-0.113**
	(0.065)	(0.066)	(0.045)
Skilled manual parent	0.067	-0.032	-0.028
	(0.061)	(0.064)	(0.039)
Mother smoker	0.181***	0.093*	0.285***
	(0.044)	(0.049)	(0.032)
Log-likelihood	-3411.0	-3366.8	-3909.5
Bayesian Information Criterion	6955.8	6938.3	7882.0
Akaike Information Criterion	6856.0	6785.6	7835.0

*, **, ***, *** = significantly different from zero at 10%, 5% and 1% levels

Table 10 gives the parameter estimates for the reporting probabilities and reports sample mean values of the misreporting probabilities. The two-sided model predicts a low probability of over-reporting (0.03 for a child who did not over-report in the previous year, or 0.17 for one who did). Both models imply a much higher under-reporting probability, averaging 0.14 or 0.52, depending on reporting behaviour in the previous year. Thus conventional measurement error models which view misreporting as an essentially unbiased process are clearly rejected here. The evidence in Table 10 suggests that the discontinuance of the taped questionnaire from 2001 onwards had the effect of increasing the rate of under-reporting. We also find that parental

smoking reduces the under-reporting probability, presumably because there is less within-family stigma associated with smoking. An important finding here is that parental influence has a greater impact on reporting behaviour than on smoking behaviour itself. Under-reporting is found to be more pronounced for girls than boys and for children in the higher socio-economic groups. Under-reporting declines strongly with age, but is highly persistent within individuals. The assumption of serially-independent misreporting would clearly be untenable here.

Table 10 Estimates of misreporting probabilities in the full misreporting model ($n = 2622$; standard errors in parentheses)

Covariate	2-sided model		1-sided model
	$p^+(X)$	$p^-(X)$	$p^-(X)$
Lagged misreporting	1.038***	1.576***	1.423***
	(0.204)	(0.140)	(0.113)
Intercept	-2.084***	-1.361***	-0.847***
	(0.155)	(0.158)	(0.090)
Age	2.284***	-6.072***	-4.050***
	(0.579)	(0.377)	(0.227)
Year	-0.035	-0.029	-0.017
	(0.026)	(0.020)	(0.012)
Female	-0.775***	0.484***	0.051
	(0.236)	(0.116)	(0.056)
Managerial / professional parent	-0.461**	0.225*	0.169**
	(0.228)	(0.121)	(0.080)
Skilled white collar parent	-0.024	0.346***	0.200**
	(0.180)	(0.126)	(0.086)
Skilled manual parent	0.222*	0.249**	0.208***
	(0.141)	(0.112)	(0.076)
Mother smoker	0.555***	-0.393***	-0.191***
	(0.135)	(0.093)	(0.056)
Interview mode change	-0.356*	0.473***	0.218**
	(0.216)	(0.141)	(0.089)
Mean predicted misreporting probabilities			
No error last year	0.035	0.138	0.178
	(0.006)	(0.012)	(0.011)
Misreporting last year	0.170	0.518	0.615
	(0.052)	(0.044)	(0.035)

*, **, ***, *** = significantly different from zero at 10%, 5% and 1% levels

5 Conclusions

The use of panel surveys, yielding sequential observations on lifetime prevalence, can give unambiguous evidence of the existence of misreporting, allowing us to identify important aspects of the process of initiation into drug use.

In very short panels, we may be limited to interval identification in the form of bounds on population parameters, rather than exact identification. We have evaluated these bounds for data on illicit drug prevalence using two UK panel surveys, each with two waves. A longer panel of data on smoking by children allows more precise identification of the dual processes of initiation and misreporting.

There are five broad conclusions. First, in all three surveys, the observed sequences of responses to questions of the form "have you ever...?" yield unambiguous evidence of a substantial degree of misreporting, in the form of 'recanting' of earlier positive self-reports. There is an indication that reporting error is most serious for relatively recent events, particularly for younger people.

Second, our analysis provides compelling evidence of asymmetric reporting error. Under-reporting of sensitive events is much more common than over-reporting, so the conventional measurement error model of essentially unbiased reporting appears untenable in this context, where survey subjects have a strong incentive to under-report.

Third, we find difficulty in drawing inferences about the precise nature of misreporting in very short panels, due to identification difficulties which can only be resolved completely by using implausible strong assumptions, such as serial independence of reporting behaviour. The bounds obtainable under weaker assumptions are nevertheless informative and we have estimated conditional bounds for 2003-4 OCJS data on cannabis and cocaine and for age 16 and 30 data from the BCS70 cohort. For the former, an analysis with an illustrative (but plausible) set of exclusion restrictions gives under-reporting rates within bounds averaging 23-60% over all sampled individuals for cannabis and 31-95% for cocaine. In the BCS70 case, where respondents are older and the time interval between interviews is much longer, cannabis under-reporting rates are much lower, within the average range 4-32%. There is evidence of substantial variation in the probability of under-reporting across individual types. Reporting error of this pattern and magnitude is clearly important for the purposes of statistical analysis of self-report data. For the true prevalence rate, only a lower bound is available and, in our analyses of cannabis in the 1986 BCS70 sweep and the 2003 OCJS sweep, this lower bound averages 7% and 25% respectively - a finding consistent with the strong growth in prevalence since the 1980s suggested by external sources such as the trends in drug-related deaths or drug seizures.

Fourth, in the case of a longer 5-wave panel, we have been able to es-

timate a full two-sided misreporting model for juvenile smoking, without a serial independence assumption for reporting behaviour. There is evidence of strong persistence in misreporting. We find large differences in the parameter estimates of a smoking initiation hazard model after allowance is made for reporting error, which suggests that measurement error biases are serious for models of this kind.

A final conclusion relates to survey questionnaire design. Our analysis deals with data on elapsed lifetime prevalence, which is important because contradictions in the sequence of responses give unambiguous evidence of error. Our analysis does not apply directly to self-reported behaviour in a limited period such as the last month or year; however it is implausible that reporting error is less serious for such variables. Panel designs often avoid simple repetition of questions in successive waves. For instance, the OCJS contains questions at wave 1 asking whether the respondent has ever committed certain types of crime; the following wave then contains questions about the previous 12 months rather than the whole past. Another growing practice is dependent interviewing, also used in some parts of the OCJS, where responses from earlier waves are fed back to respondents to aid them in making the current response. Although these practices have some advantages, they make it impossible to carry out the kind of analysis used here and may deprive us of a valuable source of evidence on measurement error.

References

[1] Aquilino, W. S. (1997). Privacy effects on self-reported drug use: interactions with survey mode and respondent characteristics, in L. Harrison and A. Hughes (eds.) *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*, 383-415. Rockville: National Institute on Drug Abuse, NIDA Research Monograph 167.

[2] Biemer, P. and Wiesen, C. (2002). Latent Class Analysis of Embedded Repeated Measurements: An Application to the National Household Survey on Drug Abuse, *Journal of the Royal Statistical Society, Series A* **165**, 97–119.

[3] Bollinger, C. R. (1996). Bounding mean regressions when a binary regressor is mismeasured, *Journal of Econometrics* **73**, 387-399.

[4] Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

[5] Hamlyn, B., Maxwell, C., Hales, J. and Tait, C. (2003). *2003 Crime and Justice Survey (England and Wales). Technical Report.* London: Home Office.

[6] Johnston, L. D. and O'Malley, P. M. (1997). The recanting of earlier reported drug use by young adults, in L. Harrison and A. Hughes (eds.) *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*, 59-80. Rockville: National Institute on Drug Abuse, NIDA Research Monograph 167.

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

[8] Lessler, J. T. and O'Reilly, J. M. (1997). Mode of interview and reporting of sensitive issues: design and implementation of audio computer-assisted self-interviewing, in L. Harrison and A. Hughes (eds.) *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*, 366-382. Rockville: National Institute on Drug Abuse, NIDA Research Monograph 167.

[9] Lewbel, A. (2000). Identification of the binary choice model with misclassification, *Econometric Theory* **16**, 603-609.

[10] Lynn, P., Buck, N. H., Burton, J., Jäckle, A. and Laurie, H. (2005). A review of methodological research pertinent to longitudinal survey design and data collection. University of Essex: ISER Working Paper no.2005-29.

[11] Lu, N T., Taylor, B. G. and Riley, K. J. (2001). The validity of adult arrestee self-reports of crack cocaine use, *American Journal of Alcohol Abuse* **27**, 399-419.

[12] Magura, S. and Kang, S.-Y. (1997). The validity of self-reported cocaine use in two high-risk populations, in L. Harrison and A. Hughes (eds.) *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*, 227-246. Rockville: National Institute on Drug Abuse, NIDA Research Monograph 167.

[13] Pacula, R. (1997). The modelling of the gateway effect, *Health Economics* **6**, 521-524.

[14] Pudney, S. E. (2003). The road to ruin? Sequences of initiation to drugs and crime in Britain, *Economic Journal* **113**, C182–198.

[15] Pudney, S. E. (2004). Keeping off the grass? An econometric model of cannabis consumption in Britain, *Journal of Applied Econometrics* **19**, 435-453.

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

[17] Van Ours. J. C. (2007). Cannabis, cocaine and jobs, *Journal of Applied Econometrics*, forthcoming.

Appendix 1: Proofs of propositions

A1.1 Conditions for exact identification

Consider first the sufficiency of the condition given in proposition 1. Equations (2), (3), (9) and (10) are satisfied only by functions of the form:

$$\begin{aligned}
P_{11}^{+}(X,Z) &= \lambda(Z)\overline{P}_{11}(X,Z) && (22)\\
\Omega_{10}^{+}(Z,W) &= \overline{\Omega}_{10}(Z,W)/\lambda(Z) && (23)\\
\Omega_{11}^{+}(Z,W) &= \overline{\Omega}_{11}(Z,W)/\lambda(Z) && (24)\\
\Omega_{01}^{+}(Z,W) &= \overline{\Omega}_{01}(Z,W)/\lambda(Z) && (25)\\
\omega_{1}^{+}(Z,W) &= \left[\overline{\Omega}_{01}(Z,W)+\overline{\Omega}_{11}(Z,W)\right]/\lambda(Z) && (26)\\
P_{01}^{+}(X,Z) &= \lambda(Z)\overline{P}_{01}(X,Z) && (27)
\end{aligned}$$

where $\lambda(Z)$ is an arbitrary positive function. To establish this, note that $\Pi_{10}(X,Z,W) = \overline{P}_{11}(X,Z)\overline{\Omega}_{10}(Z,W) = P_{11}^{+}(X,Z)\Omega_{10}^{+}(Z,W)$, implying that $P_{11}^{+}/\overline{P}_{11} = \overline{\Omega}_{10}/\Omega_{10}^{+}$. Since the left-hand side is independent of W and the right hand side independent of X, the common ratio must be a function, λ, of Z alone. This gives equations (22) and (23) and the implied forms for $\Omega_{11}^{+}, \Omega_{01}^{+}$, ω_{1}^{+} and P_{01}^{+} follow immediately. Equations (4) and (5) then give:

$$P_{00}^{+}(X,Z) = 1-\lambda(Z)+\lambda(Z)\overline{P}_{00}(X,Z) \tag{28}$$

$$\Omega_{00}^{+}(Z,W) = \frac{\lambda(Z)-1+\overline{\Omega}_{00}(Z,W)}{\lambda(Z)} \tag{29}$$

The choice of $\lambda(Z)$ is limited by the non-negativity restrictions on P_{00}^{+} and Ω_{00}^{+}. These imply:

$$\max_{W} 1-\overline{\Omega}_{00}(Z,W) \leq \lambda(Z) \leq \min_{X} \frac{1}{1-\overline{P}_{00}(X,Z)} \tag{30}$$

Thus, a sufficient condition for local exact identification is the existence of points w and x in the support sets of $W|Z=z$ and $X|Z=z$ where the limits of the interval (30) are unity, requiring $\overline{\Omega}_{00}(z,w)=0$ and $\overline{P}_{00}(x,z)=0$.

For necessity, we need to show that the bounds (30) cannot be tightened by using the other inequality constraints besides the non-negativity of $\overline{\Omega}_{00}$ and $\overline{P}_{00}$. To see this, note that the only inequality constraints not used in the derivation of (30) were the non-negativity conditions on $P_{01}^{+}, P_{11}^{+}, \Omega_{01}^{+}, \Omega_{10}^{+}, \Omega_{11}^{+}$ and ω_{1}^{+}, requiring only $\lambda(Z) \geq 0$, which is already implicit in (30). Thus the

existence of points x and w such that $\overline{\Omega}_{00}(z,w) = \overline{P}_{00}(x,z) = 0$ is both necessary and sufficient for exact identification.

A1.2: Bounds under homogeneity and exchangeability

Leave implicit the dependence of the unknown probabilities on X, Z, W. Equations (1), (9) and (10) imply $\Pi_{01} = \Pi_{10} + (P_{01}/P_{11})(\Pi_{10} + \Pi_{11})$, or:

$$P_{11} = \left(\frac{\Pi_{10} + \Pi_{11}}{\Pi_{01} - \Pi_{10}}\right) P_{01} \tag{31}$$

Thus solutions for P_{01} and P_{11} lie on the ray through the origin defined by (31). The non-negativity of P_{00} implies $1 - P_{01} - P_{11} \geq 0$ or $P_{11} \leq 1 - P_{01}$. Exchangeability and the non-negativity of Ω_{00} imply $1 - 2\Omega_{10} - \Omega_{11} \geq 0$ and thus $P_{11} \geq 2P_{11}\Omega_{10} + P_{11}\Omega_{11}$ or, using (2) and (3), $P_{11} \geq 2\Pi_{10} + \Pi_{11}$. These two conditions give upper and lower limits on P_{11} respectively, so the bounds on P_{11} are given by the intersection of the equations $P_{11} = 1 - P_{01}$ and $P_{11} = 2\Pi_{10} + \Pi_{11}$ with the ray (31). The intersection points are at $P_{11} = (\Pi_{10} + \Pi_{11})/(\Pi_{01} + \Pi_{11})$ and $2\Pi_{10} + \Pi_{11}$ respectively, as stated in proposition 2. It can be established that these bounds are sharp by solving for the remaining parameters at these two P_{11}-points and noting that all inequalities (6)-(8) are satisfied at each point.

The misreporting rate $1 - \omega_1$ can be written $(\Pi_{01} - \Pi_{10})/P_{01}$, using (1) and (10); (31) then gives:

$$1 - \omega_1 = 1 - \left(\frac{\Pi_{10} + \Pi_{11}}{P_{11}}\right) \tag{32}$$

This is increasing in P_{11}, so the lower and upper bounds on $1 - \omega_1$ are given by substituting the lower and upper bounds for P_{11} in (32).

The hazard or initiation rate h is defined as $P_{01}/(1 - P_{11})$. Using (31); this can be written as:

$$h = \left(\frac{\Pi_{01} - \Pi_{10}}{\Pi_{10} + \Pi_{11}}\right) \frac{P_{11}}{1 - P_{11}} \tag{33}$$

which is increasing in P_{11}; substitution of the lower and upper bounds for P_{11} then gives the bounds on h appearing in proposition 2.

Since P_{11} and h are known *a priori* to depend only on X, Z, and ω_0 to depend only on Z, W, we can take the largest (smallest) lower (upper) bound

over all W for the former and the largest (smallest) lower (upper) bound over all X for the latter.

A1.3: Bounds under homogeneity only

Solve equations (1), (2), (3) and (9), for given fixed values of P_{11} and ω_1:

$$\begin{aligned}
\Omega_{11} &= \Pi_{11}/P_{11} && (34)\\
\Omega_{10} &= \Pi_{10}/P_{11} && (35)\\
\Omega_{01} &= \omega_1 - \Pi_{11}/P_{11} && (36)\\
P_{01} &= (\Pi_{01} + \Pi_{11})/\omega_1 - P_{11} && (37)
\end{aligned}$$

Ω_{11} and Ω_{10} are non-negative provided P_{11} is; the non-negativity conditions corresponding to (36) and (37) are:

$$\begin{aligned}
\omega_1 &\geq \Pi_{11}/P_{11} && (38)\\
\omega_1 &\leq (\Pi_{01} + \Pi_{11})/P_{11} && (39)
\end{aligned}$$

Non-negativity of P_{00} and Ω_{00} require respectively:

$$\begin{aligned}
\omega_1 &\geq \Pi_{01} + \Pi_{11} && (40)\\
\omega_1 &\leq 1 - (\Pi_{10}/P_{11}) && (41)
\end{aligned}$$

The inequalities (38)-(41) are a complete characterisation of the admissible region for ω_1 and P_{11}. The four functions bounding this region are all monotonic in P_{11}, so the extremal points will be located at intersections of these functions. (38) and (39) do not intersect, so there are five intersection points to consider:

$$\begin{aligned}
\omega_1^{(A)} &= \Pi_{01} + \Pi_{11}; \quad P_{11}^{(A)} = \Pi_{11}/(\Pi_{01} + \Pi_{11})\\
\omega_1^{(B)} &= \Pi_{11}/(\Pi_{10} + \Pi_{11}); \quad P_{11}^{(B)} = \Pi_{10} + \Pi_{11}\\
\omega_1^{(C)} &= \Pi_{01} + \Pi_{11}; \quad P_{11}^{(C)} = 1\\
\omega_1^{(D)} &= (\Pi_{01} + \Pi_{11})/(\Pi_{01} + \Pi_{10} + \Pi_{11}); \quad P_{11}^{(D)} = \Pi_{01} + \Pi_{10} + \Pi_{11}\\
\omega_1^{(E)} &= \Pi_{01} + \Pi_{11}; \quad P_{11}^{(E)} = \Pi_{10}/(1 - \Pi_{01} - \Pi_{11})
\end{aligned}$$

Checking (38)-(41) shows points C and D to be feasible for all configurations of Π_{jk}; points A and B are feasible iff $(\Pi_{01} + \Pi_{11})(\Pi_{10} + \Pi_{11}) \leq \Pi_{11}$; and point E is feasible iff $(\Pi_{01} + \Pi_{11})(\Pi_{10} + \Pi_{11}) \geq \Pi_{11}$. The upper bound on

P_{11} is therefore always unity. For the other bounds, there are two cases to consider.

If $(\Pi_{01}+\Pi_{11})(\Pi_{10}+\Pi_{11}) \leq \Pi_{11}$:

$$\begin{aligned}
\omega_1^{\min} &= \min\{\omega_1^A, \omega_1^B, \omega_1^C, \omega_1^D\} = \min\{\Pi_{01}+\Pi_{11}, \Pi_{11}/(\Pi_{10}+\Pi_{11})\} \\
&= (\Pi_{10}+\Pi_{11})^{-1} \min\{(\Pi_{10}+\Pi_{11})(\Pi_{01}+\Pi_{11}), \Pi_{11}\} = \Pi_{01}+\Pi_{11}
\end{aligned}$$

$$\begin{aligned}
\omega_1^{\max} &= \max\{\omega_1^A, \omega_1^B, \omega_1^C, \omega_1^D\} \\
&= \max\{\Pi_{11}/(\Pi_{10}+\Pi_{11}), (\Pi_{01}+\Pi_{11})/(\Pi_{01}+\Pi_{10}+\Pi_{11})\} \\
&= (\Pi_{01}+\Pi_{11})/(\Pi_{01}+\Pi_{10}+\Pi_{11})
\end{aligned}$$

$$\begin{aligned}
P_{11}^{\min} &= \min\{P_{11}^A, P_{11}^B, P_{11}^C, P_{11}^D\} = \min\{\Pi_{11}/(\Pi_{01}+\Pi_{11}), \Pi_{10}+\Pi_{11}\} \\
&= (\Pi_{01}+\Pi_{11})^{-1} \min\{\Pi_{11}, (\Pi_{01}+\Pi_{11})(\Pi_{10}+\Pi_{11})\} = \Pi_{10}+\Pi_{11}
\end{aligned}$$

If $(\Pi_{01}+\Pi_{11})(\Pi_{10}+\Pi_{11}) \geq \Pi_{11}$:

$$\omega_1^{\min} = \min\{\omega_1^C, \omega_1^D, \omega_1^E\} = \Pi_{01}+\Pi_{11}$$

$$\omega_1^{\max} = \max\{\omega_1^C, \omega_1^D, \omega_1^E\} = (\Pi_{01}+\Pi_{11})/(\Pi_{01}+\Pi_{10}+\Pi_{11})$$

$$\begin{aligned}
P_{11}^{\min} &= \min\{P_{11}^C, P_{11}^D, P_{11}^E\} = \min\{\Pi_{01}+\Pi_{10}+\Pi_{11}, \Pi_{10}/(1-\Pi_{01}-\Pi_{11})\} \\
&= (1-\Pi_{01}-\Pi_{11})^{-1} \min\{(1-\Pi_{01}-\Pi_{11})(\Pi_{01}+\Pi_{10}+\Pi_{11}), \Pi_{10}\} \\
&= (1-\Pi_{01}-\Pi_{11})^{-1} \min\{(\Pi_{10}+\Pi_{00})(1-\Pi_{00}), \Pi_{10}\} \\
&= (1-\Pi_{01}-\Pi_{11})^{-1} \min\{\Pi_{10}+\Pi_{00}(1-\Pi_{00}-\Pi_{10}), \Pi_{10}\} \\
&= \Pi_{10}/(1-\Pi_{01}-\Pi_{11})
\end{aligned}$$

Note that $\Pi_{10}+\Pi_{11} \gtrless \Pi_{10}/(1-\Pi_{01}-\Pi_{11})$ according as $(\Pi_{01}+\Pi_{11})(\Pi_{10}+\Pi_{11}) \lessgtr \Pi_{11}$. Consequently, $P_{11}^{\min} = \max\{\Pi_{10}+\Pi_{11}, \Pi_{10}/(1-\Pi_{01}-\Pi_{11})\}$. For the initiation rate, note that removing the exchangeability restriction cannot reduce the set of feasible values for h, so the upper bound remains 1. For the lower bound, note that the point $P_{01}=0, P_{11}=\theta, \Omega_{01}=\Pi_{01}/\theta, \Omega_{10}=\Pi_{10}/\theta, \Omega_{11}=\Pi_{11}/\theta$ is feasible for any choice of $\theta > 1-\Pi_{00}$; all such points imply $h=0$, so this is the lower bound.

Appendix 2: Data definitions and summaries

Table A1 Definitions and sample characteristics of covariates: 2003/4 OCJS ($n = 3,090$)

Covariate	Sample mean	
	2003	**2004**
Parent present: parent present during interview	0.377	0.249
Interviewer help: interviewer helped with A-CASI self-completion	0.015	0.009
Age: (age in years-16)/10	-0.0099	0.0913
In work: respondent employed or self-employed	0.324	0.349
Low deprivation: in top 3 deciles of ONS index of multiple deprivation	0.275	0.272
High deprivation: in bottom 3 deciles of ONS index of multiple deprivation	0.296	0.293
Parental home: respondent lives with parents	0.867	0.845
Owner-occupied: home is owner-occupied	0.679	0.694
Non-religious: respondent gives no religious affiliation	0.371	0.451
Non-white ethnicity: any self-assessed ethnicity except "white"	0.084	
Care: has been in local authority care or a foster home	0.014	
Female: respondent is female	0.503	
Drug area: neighbourhood with many drug dealers/users	0.203	0.232

Table A2 Definitions and sample characteristics of covariates: 1986 and 2000 waves BCS70 ($n = 5,339$)

Covariate	**Sample mean**	
	age 16	**age 30**
In work: respondent employed or self-employed	0.177	0.837
Parental home: respondent lives with parents	0.900	0.114
Owner-occupied: home is owner-occupied	0.464	0.668
Non-religious: respondent gives no religious affiliation	0.094	0.249
Non-white ethnicity: any self-assessed ethnicity except "white"	0.029	
Care: has been in local authority care or a foster home	0.006	
Female: respondent is female	0.555	

Table A3 Definitions and sample characteristics of covariates: 1994-2004 BHPS ($n = 2622$)

Covariate	Sample mean
Age: (age in years as at December of interview year - 12)/10	0.098
Year: year of interviewer - 2000	-0.311
Female: dummy = 1 if female, 0 if male	0.490
Managerial/professional parent: either parent has managerial/professional occupation	0.240
Skilled white collar parent: parent with higher occupational status is skilled white collar	0.184
Skilled manual parent: parent with higher occupational status is skilled manual	0.252
Mother smoker: mother self-reports smoking at any wave of child panel	0.366
Interview mode change: dummy = 1 for interviews in 2001-4, 0 otherwise	0.331

Table A4 Bivariate probit estimates of the conditional distribution of self-declared lifetime prevalence in 2003 and 2004 OCJS

Covariate	Cannabis		Cocaine	
	2003	**2004**	**2003**	**2004**
Covariates in W				
Parent present	-0.094*	-0.143***	-0.055	-0.136
Interviewer help needed	-0.331**	-0.312	-0.386*	-0.337*
Covariates in Z				
(Age-16)/10	1.989****	1.837****	2.119****	3.733****
((Age-16)/10)^2	-3.938****	-4.720****	-1.751*	-8.986****
((Age-16)/10)^3	5.126***	6.920****	3.123	10.712*
((Age-16)/10)^4	-2.924	-3.514**	-3.201	-4.445
Parental home	-0.016	-0.018	0.118	-0.135*
Non-religious	0.111***	0.078**	0.221****	0.096*
Non-white ethnicity	-0.360****	-0.319****	-0.459***	-0.383***
Intercept	-0.578****	-0.505****	-2.135****	-1.837****
Covariates in X				
In work	0.169***	0.067	0.052	0.106*
Least deprived area	0.182****	0.035	0.327****	0.177**
Most deprived area	0.066	-0.079	0.159*	-0.046
Owner-occupied home	-0.046	-0.016	-0.108	-0.056
Been in care	0.518****	0.093	0.455*	0.462**
Female	-0.107**	-0.045	-0.318****	-0.248****
Problem drug area	0.209****	0.381****	0.331****	0.290****
ρ	0.853****		0.913****	
RESET $\chi^2(2)$-statistic	1.38 $(P = 0.503)$		3.01 $(P = 0.222)$	
$\chi^2(4)$ for W-variables	10.29 $(P = 0.036)$		4.78 $(P = 0.311)$	

*, **, ***, **** = significantly different from zero at 20%, 10%, 5% and 1% levels

Table A5 Bivariate probit estimates of the conditional distribution of self-declared lifetime prevalence in the age 16 and 30 BCS70 sweeps

Covariate	**Age 16**	**Age 30**
Covariates in Z		
Parental home	-0.155***	-0.323****
Non-religious	0.202***	0.332****
Non-white ethnicity	-0.060	-0.058
Intercept	-1.273****	
Covariates in X		
In work	0.055	0.102***
Owner-occupied home	-0.200****	-0.285****
Been in care	0.178	0.307
Female	-0.067*	-0.464****
ρ	0.608****	
RESET $\chi^2(2)$-statistic	1.70 $(P = 0.429)$	

*, **, ***, **** = significantly different from zero at 20%, 10%, 5% and 1% levels