ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2018)

Modelling illegal drug participation

Sarah Brown,

University of Sheffield, UK

Mark N. Harris,

Curtin University, Australia

Preety Srivastava

Royal Melbourne Institute of Technology, Australia

and Xiaohui Zhang

University of Exeter, UK

[Received November 2014. Final revision September 2016]

Summary. We contribute to the small, but important, literature exploring the incidence and implications of misreporting in survey data. Specifically, when modelling 'social bads', such as illegal drug consumption, researchers are often faced with exceptionally low reported participation rates. We propose a modelling framework where firstly an individual decides whether to participate or not and, secondly, for participants there is a subsequent decision to misreport or not. We explore misreporting in the context of the consumption of a system of drugs and specify a *multivariate inflated probit model*. Compared with observed participation rates of 12.2%, 3.2% and 1.3% (for use of marijuana, speed and cocaine respectively) the true participation rates are estimated to be almost double for marijuana (23%), and more than double for speed (8%) and cocaine (5%). The estimated chances that a user would misreport their participation is a staggering 65% for a hard drug like cocaine, and still about 31% and 17%, for the softer drugs of marijuana and speed.

Keywords: Discrete data; Illegal drug consumption; Inflated responses; Misreporting

1. Introduction and background

Over the past three decades, the increased availability of microlevel data sets has enabled researchers to explore an extensive range of research themes at the individual and household level. Such microlevel data are invariably collected by using survey techniques with the result that the quality of the data that are gathered hinges critically on the respondents providing reliable and accurate information. It is apparent, however, that the subject matter of some surveys may be such that respondents have an incentive to misreport because of the sensitive nature of the questions. Individuals may have an incentive to underreport activities which are regarded as socially undesirable or which are associated with perceived social stigma or legal consequences, such as smoking, alcohol, illicit drug consumption and sexual behaviours (e.g. Berg and Lien (2006) and Pudney (2007)).

Address for correspondence: Mark N. Harris, School of Economics and Finance, Curtin University, Perth, WA 6102, Australia.
E-mail: mark.harris@curtin.edu.au

Misreporting will result in inaccurate estimates of the prevalence of such behaviours, which may lead us to question the validity of empirical conclusions that are drawn from surveys. Moreover, any systematic misreporting will probably lead to biased inferences in econometric analyses and erroneous policy advice. Despite these extremely important implications there is a shortage of relevant research exploring the incidence and likely effects of such misreporting in survey data.

Misclassification, or misreporting, often leads to the presence of 'excess' 0s, which has long been of interest to the applied researcher. To address such concerns, hurdle and double-hurdle models have been developed and have found favour in areas ranging from a continuous dependent variable with a non-zero probability mass at (typically, but not exclusively) zero levels (Cragg, 1971; Smith, 2003), to the so-called zero-inflated (augmented) Poisson count data models (Mullahey, 1986, 1997; Heilbron, 1989; Lambert, 1992; Greene, 1994; Pohlmeier and Ulrich, 1995) and, more recently, to zero-inflated ordered probit models (Harris and Zhao, 2007). Typically, the issue that arises is that '0'-observations can result from two distinct processes and that ignoring this can lead to seriously misspecified models.

In this paper, we explore the modelling of *sensitive* response variables: variables where an associated loss function (perceived or actual) is involved for the individual in terms of the responses that he or she reports. Here, it is clear that the researcher must be aware of the potential for misreporting. For consumption of goods with associated reporting loss functions, the approach that is suggested here allows for these 0-observations to correspond not only to non-participants, but importantly also to those participants who erroneously report zero consumption.

Our particular application lies in the important area of misreporting within the context of the consumption of illicit drugs. Given the considerable individual and social costs that are associated with the consumption of illegal drugs (including increased crime, health issues and difficulties at school or work) it is not surprising that an extensive body of research exists exploring issues that are related to the addictive nature of drugs as well as the relationship between the consumption of different types of drug. However, as argued by MacDonald and Pudney (2000) and Pudney (2010), there is no consensus regarding policy advice relating to drug abuse and, furthermore, analysis of survey data relating to drug use could potentially contribute to the policy debate in this area. The use of cross-sectional surveys to model socio-economic determinants of drug use (Duarte *et al*., 2005; Ramful and Zhao, 2009) and panel surveys to estimate rational addiction models (Becker *et al*., 1994; Labeaga, 1999) and demand elasticities are therefore important tools of present-day policy making.

It is apparent that the shortcomings of this type of data should therefore be well understood in order to make appropriate policy decisions. Indeed, in the context of survey response rates and response accuracy, Pudney (2010), page 26, commented that

> 'these problems cannot be overcome completely and their impact on research findings is not yet well understood'.

Hence, we aim to contribute to the relatively small, but clearly important, literature exploring the incidence and extent of misreporting (specifically with regard to drug consumption) in individual level survey data.

Our approach is similar to that of Hausman *et al*. (1998) who used a logit model to estimate misclassification probabilities. They considered a binary choice model with two types of misclassification: the probability that the true 0 is recorded as a 1, and the probability that the true 1 is recorded as a 0. Our specific contributions to the literature are threefold. Firstly, we extend the general approach of Hausman *et al*. (1998) to allow for covariates to influence the

misreporting, or misclassification, decision; this will be very important for policy makers to help to identify those individuals with greater propensities to do so. Secondly, we acknowledge that many 'similar' response variables of interest (various illicit drugs in our example) are likely to be consumed jointly (here because of their common addictive nature), so we extend the simpler univariate approach to a multivariate approach. Finally, we apply this new model to the consumption of illegal drugs (in Australia) and thereby provide new evidence about the likely extent of misreporting across these. We also provide evidence about the true rates of participation across these drugs compared with a simple inspection of observed participation rates.

The rest of the paper is as follows. Section 2 describes the econometric setting; the empirical application to a system of drug participation equations is described in Section 3. The data and empirical results (including a series of robustness checks, validation exercises and Monte Carlo experiments) are detailed in Sections 4 and 5 respectively. Finally, Section 6 concludes.

2. Econometric framework

2.1. An inflated probit model

We start by defining a discrete random variable y that is observable and assumes the binary outcomes of 0 and 1. A standard probit approach would map a single latent variable to the observed outcome $y=1$ via an index function, essentially modelling participation rates. In the context of illegal drug use, we hypothesize that a (potentially significantly large) proportion of participants will actually report themselves as being non-participants, because of both moral and legal concerns about participation. Specifically, let r^* denote a binary variable indicating the split between regimes 0 (with $r=0$ for non-participants) and 1 ($r=1$ for participants). Although unobservable, r is related to a latent variable r^* via the mapping: $r=1$ for $r^*>0$ and $r=0$ for $r^*\leqslant 0$. Thus r^* represents the propensity for participation and is related to a set of explanatory variables $\mathbf{x}_r$ with unknown weights $\boldsymbol{\beta}_r$, and a standard normally distributed error term ε_r, such that

$$r^* = \mathbf{x}_r'\boldsymbol{\beta}_r + \varepsilon_r. \tag{1}$$

For participants ($r=1$), a second latent variable m^* represents the propensity to misreport. Again this is related to a second unobserved variable m such that $m=1$ for $m^*>0$ and $m=0$ for $m^*\leqslant 0$, where $m=0$ represents a misreporter and $m=1$ a true reporter. Again, we can write this (linear) latent form as

$$m^* = \mathbf{x}_m'\boldsymbol{\beta}_m + \varepsilon_m. \tag{2}$$

Of course, neither r nor m is observed; the observability criterion for observed y is

$$y = rm. \tag{3}$$

As such the observed realization of the random variable y can be viewed as the result of two independent latent equations, equations (1) and (2). However, these equations correspond to the same individual so it is likely that the vector of stochastic terms ε_i will be related across equations (which is a point that has been ignored in the previous literature). Allowing $(\varepsilon_r, \varepsilon_m)$ to follow a bivariate normal distribution with covariance matrix Ω (a 2×2 symmetric matrix with 1s on the diagonal and ρ on the off-diagonals) the relevant probabilities will have the form

$$\Pr(y) = \begin{cases} \Pr(y=0|\mathbf{x}) = 1-\Phi(\mathbf{x}_r'\boldsymbol{\beta}_r) + \Phi_2(\mathbf{x}_r'\boldsymbol{\beta}_r, -\mathbf{x}_m'\boldsymbol{\beta}_m; \Omega), \\ \Pr(y=1|\mathbf{x}) = \Phi_2(\mathbf{x}_r'\boldsymbol{\beta}_r, \mathbf{x}_m'\boldsymbol{\beta}_m; \Omega) \end{cases} \tag{4}$$

where Φ_2 denotes the cumulative distribution function of the standardized bivariate normal

distribution. The first term on the right-hand side of equation (4) for $\Pr(y=0|\mathbf{x})$ represents a genuine non-participant; the second term, a (participant) misreporter. The expression for $\Pr(y=1|\mathbf{x})$ thus represents a (participant) true reporter. So here the probability of a 0-observation has been 'inflated' as it is a combination of the probability of non-participation plus that from misreporting. This approach thus models misreporting explicitly and as a function of a set of explanatory variables unlike the model that was developed by Hausman *et al.* (1998) where misreporting is accounted for by using constant terms, or by Dustmann and Soest (2001) who decomposed misclassification errors in panel data into time persistent and time varying components and where the probability of misclassification is independent of respondent characteristics. However, it is very unlikely that such misreporting rates will be constant and homogeneous across individuals. Moreover, ignoring this heterogeneity (if present) could lead to biased estimates of quantities of interest (such as true participation rates). Owing to the 0-inflation and the correlated disturbances, we term this a correlated inflated probit (IP) model. A test of $\rho=0$ is jointly a test for independence of the two error terms and also one of the correlated IP *versus* the nested IP model.

Given the assumed form for the probabilities and an independently and identically distributed sample of size N from the population on $(y_i, \mathbf{x})$, $i=1,\ldots,N$, the parameters of the full model $\boldsymbol{\theta}=(\boldsymbol{\beta}_r', \boldsymbol{\beta}_m', \rho)'=\boldsymbol{\beta}'$ can be consistently and efficiently estimated by using maximum likelihood techniques; the log-likelihood function is (where h_{ij} is the usual indicator function for the observed choice)

$$l(\boldsymbol{\theta})=\sum_{i=1}^{N}\sum_{j}h_{ij}\ln\{\Pr(y_i=j|\mathbf{x},\boldsymbol{\theta})\}. \tag{5}$$

2.2. Extending to a multivariate inflated probit system

Often social bads such as licit and illicit drugs are consumed in a consumption bundle (see, for example, Collins *et al.* (1998) and Ives and Ghelani (2006)), given that they are habit forming. Instead of modelling the consumption of such social bads in isolation, the above set-up can be extended to a multivariate framework where participation decisions are considered to be taken jointly (see, for example, Zhao and Harris (2004) and Ramful and Zhao (2009)). Owing to unobservable characteristics (such as individual tastes, addictive traits and risk-taking attitudes) the decision to consume multiple drugs is very likely to be related through the error terms of the participation and misreporting equations, i.e. via the unobservables. As a consequence, vital cross-drug information is lost when the correlated IP model is estimated in a univariate framework for several drugs of interest. As detailed below, we consider a system of three drugs. The multivariate approach essentially isolates the joint effects of observable and unobservable personal characteristics on the participation and misreporting of all three drugs and estimates the strength of the intrinsic correlations, via the unobservables, across the three drugs which are commonly considered to be closely related economic goods.

For a set of k $(k=1,\ldots,K)$ multivariate correlated IP models, the propensity for participation will be

$$r_k^*=\mathbf{x}_{rk}'\boldsymbol{\beta}_{rk}+\varepsilon_{rk} \qquad (k=1,\ldots,K), \tag{6}$$

and the propensity to misreport will be

$$m_k^*=\mathbf{x}_{mk}'\boldsymbol{\beta}_{mk}+\varepsilon_{mk} \qquad (k=1,\ldots,K). \tag{7}$$

There is no necessary restriction that $\mathbf{x}_{rk}=\mathbf{x}_{rh}$ or that $\mathbf{x}_{mk}=\mathbf{x}_{mh}, \forall\, k\neq h,$ but we shall assume so, both in the empirical application and also below, to simplify the notation (i.e. the same covariate

specification applies to all drugs). Economic and mathematical identification here, though, will require that $\mathbf{x}_{rk} \neq \mathbf{x}_{mk}, \forall\, k$: for each drug equation we require exclusion restrictions with regard to the participation and misreporting equations (but these are not necessary *across* the different drug equations). The most general specification is to assume that the ε_{rk}s and the ε_{mk}s are freely correlated both within and across equations. This results in $s = 2K$ latent equations where the error terms jointly follow a multivariate normal distribution of order s with covariance matrix Σ given by

$$\Sigma_s = \begin{pmatrix} 1 & & & & & \\ \rho_{r_1m_1} & 1 & & & & \\ \rho_{r_1r_2} & \rho_{m_1r_2} & 1 & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ \rho_{r_1r_K} & \rho_{m_1r_K} & \rho_{r_2r_K} & \cdots & 1 & \\ \rho_{r_1m_K} & \rho_{m_1m_K} & \rho_{r_2m_K} & \cdots & \rho_{r_Km_K} & 1 \end{pmatrix}, \tag{8}$$

where $\rho_{r_1m_1} = \rho_{m_1r_1}$ and so on (Σ_s is symmetric).

Consider a system of correlated IP models for three illicit drugs. Since $K = 3$, we have six latent equations where, for example, $\rho_{r_1m_1}$ relates to the correlation between ε_{r_1} and ε_{m_1}, the respective error terms from the participation equation and the misreporting equation relating to the first drug. This results in a range of joint probabilities of interest, such as the polar cases of

$$\Pr(y_1 = 1, y_2 = 1, y_3 = 1|\mathbf{x}) = \Phi_6(\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}, \mathbf{x}'_{m_1}\boldsymbol{\beta}_{m_1}, \mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, \mathbf{x}'_{m_2}\boldsymbol{\beta}_{m_2}, \mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}, \mathbf{x}'_{m_3}\boldsymbol{\beta}_{m_3}; \Sigma_6) \tag{9}$$

and

$$\begin{aligned} \Pr(y_1 = 0, y_2 = 0, y_3 = 0|\mathbf{x}) = {} & \Phi_3(-\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}, -\mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, -\mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}; \Sigma_3) + \Phi_4(\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}, -\mathbf{x}'_{m_1}\boldsymbol{\beta}_{m_1}, \\ & -\mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, -\mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}; \Sigma_4) + \Phi_4(-\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}, \mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, -\mathbf{x}'_{m_2}\boldsymbol{\beta}_{m_2}, \\ & -\mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}; \Sigma_4) + \Phi_4(-\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}, -\mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, \mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}, -\mathbf{x}'_{m_3}\boldsymbol{\beta}_{m_3}; \Sigma_4) \\ & + \Phi_5(\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}, -\mathbf{x}'_{m_1}\boldsymbol{\beta}_{m_1}, \mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, -\mathbf{x}'_{m_2}\boldsymbol{\beta}_{m_2}, -\mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}; \Sigma_5) \\ & + \Phi_5(\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}, -\mathbf{x}'_{m_1}\boldsymbol{\beta}_{m_1}, -\mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, \mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}, -\mathbf{x}'_{m_3}\boldsymbol{\beta}_{m_3}; \Sigma_5) \\ & + \Phi_5(-\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}, \mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, -\mathbf{x}'_{m_2}\boldsymbol{\beta}_{m_2}, \mathbf{x}'_{r3}\boldsymbol{\beta}_{r_3}, -\mathbf{x}'_{m_3}\boldsymbol{\beta}_{m_3}; \Sigma_5) \\ & + \Phi_6(\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}, -\mathbf{x}'_{m_1}\boldsymbol{\beta}_{m_1}, \mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, -\mathbf{x}'_{m_2}\boldsymbol{\beta}_{m_2}, \mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}, -\mathbf{x}'_{m_3}\boldsymbol{\beta}_{m_3}; \Sigma_6), \end{aligned} \tag{10}$$

and also intermediate cases such as

$$\begin{aligned} \Pr(y_1 = 1, y_2 = 0, y_3 = 0|\mathbf{x}) = {} & \Phi_4(\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}, \mathbf{x}'_{m_1}\boldsymbol{\beta}_{m_1}, -\mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, -\mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}; \Sigma_4) \\ & + \Phi_5(\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}, \mathbf{x}'_{m_1}\boldsymbol{\beta}_{m_1}, \mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, -\mathbf{x}'_{m_2}\boldsymbol{\beta}_{m_2}, -\mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}; \Sigma_5) \\ & + \Phi_5(\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}, \mathbf{x}'_{m_1}\boldsymbol{\beta}_{m_1}, -\mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, \mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}, -\mathbf{x}'_{m_3}\boldsymbol{\beta}_{m_3}; \Sigma_5) \\ & + \Phi_6(\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}, \mathbf{x}'_{m_1}\boldsymbol{\beta}_{m_1}, \mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, -\mathbf{x}'_{m_2}\boldsymbol{\beta}_{m_2}, \mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}, -\mathbf{x}'_{m_3}\boldsymbol{\beta}_{m_3}; \Sigma_6), \end{aligned} \tag{11}$$

where Φ_n denotes the n-dimensional multivariate normal cumulative distribution function.

It is intuitive to take a closer look at these probabilistic expressions. Take, for example, equation (9), which corresponds to the probability of *observing* participation in all three drugs. Here all the six elements in parentheses on the right-hand side relate to participation and true reporting in all of the three respective drugs. However, $\Pr(y_1 = 0, y_2 = 0, y_3 = 0|\mathbf{x})$, or equation (10), has a more complex form. This probability corresponds to an *observed* 0 in each of the three drugs. This can occur in eight distinct ways; the individual can be

(a) a true non-participant in each drug, first term,
(b) a misreporting participant in drug 1, with the relevant (upper) integration limits being $\mathbf{x}'_{r_1}\boldsymbol{\beta}_{r_1}$ and $-\mathbf{x}'_{m_1}\boldsymbol{\beta}_{m_1}$, but a true non-participant in drugs 2 and 3 ($-\mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}$, $-\mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}$), second term,

$$\vdots$$

(h) a misreporting participant in all drugs, eighth term.

Σ_j defines the relevant submatrices of Σ with appropriate signs in the correlations. For example, the relevant lower submatrix of Σ_4 in the second right-hand side term of equation (10) is defined as

$$\Sigma_4 = \begin{pmatrix} 1 & & & \\ -\rho_{r_1m_1} & 1 & & \\ -\rho_{r_1r_2} & \rho_{m_1r_2} & 1 & \\ -\rho_{r_1r_3} & \rho_{m_1r_3} & \rho_{r_2r_3} & 1 \end{pmatrix}.$$

This multivariate IP (MIP) model can be estimated by maximum likelihood but, as the probabilities entering this are functions of high dimensional multivariate normal distributions, these are simulated by using the Geweke–Hajivassiliou–Keane algorithm (see, for example, Keane (1994)) and Halton sequences (Train, 2000; Bhat, 2003) of length 500. In addition, since the joint and conditional probabilities are highly non-linear functions of $\mathbf{x}$, partial effects are calculated by using numerical gradients, and standard errors of these obtained by the delta method.

3. Application to drug consumption

Empirical studies play a crucial role in identifying the socio-economic and demographic factors that are associated with the consumption of illicit drugs, providing invaluable information to facilitate well-targeted public health policies. One strand of the existing literature in this area focuses on exploring the determinants of the decision to take illegal drugs. However, one of the key issues in the empirical literature on drug addiction and the demand for illicit drugs relates to the accuracy of self-reported data and the incentive to misreport and underreport illicit drug use. The extent of such misreporting and underreporting is likely to be influenced by a variety of factors such as gender and ethnicity (see, for example, Mensch and Kandel (1988) and Fendrich and Vaughn (1994)).

Misreporting of drugs use may also be influenced by how the survey is conducted. In particular, the drop-and-collect and/or mail-back methods have been associated with lower underreporting of sensitive information (Bowling, 2005). Presumably, this is due to the greater anonymity, more privacy and confidentiality of the method. For instance, comparing the mail survey method with computer-assisted telephone interviews (CATIs), Kraus and Augustin (2001) found that a lower number of respondents would admit alcohol consumption if questioned by telephone compared with self-reports from questionnaires. In a similar vein, Hoyt and Chaloupka (1994) and Fendrich and Vaughn (1994) found that lower reported drugs use is associated with telephone interviews. The increased use of CATIs in the gathering of information has arguably improved the accuracy of such data although it is not clear to what extent the accuracy has been improved (Morrison-Beedy *et al.*, 2006).

In addition, given the apparent complex interrelationships between the demand for different types of illicit drugs, it is apparent that the extent of misreporting may vary across different types of drugs, arguably being particularly serious in the case of harder drugs (such as heroin and cocaine). Pudney (2007) analysed the consequences of misreporting of illicit drugs use for statistical inference by using UK panel data containing repeated questions on self-reported

lifetime drug use. The findings indicate serious underreporting of the use of marijuana and cocaine, which in turn leads to biases in statistical modelling. For example, for one of the data sets that were analysed, underreporting rates for marijuana and cocaine with bounds averaging respectively from 23% to 60% and from 31% to 95% for all individuals were observed.

Such findings are supported by the evidence from surveys which check self-reported data via drug tests (usually for prisoners or arrestees), which indicate serious misreporting problems in the case of hard drugs (see, for example, MacDonald and Pudney (2003)). For example, in an early contribution, Wish (1987) analysed a sample of men arrested in New York City in 1984. For cocaine, the interview data indicated a drug use rate of 43% compared with 82% elicited from urine specimens. More recently, Lu *et al*. (2001) compared underreporting of drugs by validating information obtained via interviews with urinalysis for a sample of adult arrestees. The findings indicate significant levels of underreporting for all drugs with accurate reporting declining from 64% in the case of marijuana to 46% in the case of opium.

However, the extent to which findings from such studies where cross-validation is possible can be generalized is not apparent and is arguably limited given that such data are based on somewhat atypical circumstances and samples. The modelling strategy that is outlined in this study, in contrast, requires only a single source of (cross-section) survey data without recourse to validation from other sources, such as drug tests or historical information on lifetime drug consumption.

4. The data

The data that we use for the model are drawn from the Australian National Drug Strategy Household Survey, which is a nationally representative survey of the Australian civilian population aged over 14 years (National Drug Strategy Household Survey, 2010). The earlier waves of the survey used face-to-face and drop-and-collect methods to collect data. The CATI method of data collection was introduced in the 2001 survey and all three methods were employed to collect data. The 2004 and 2007 surveys, in contrast, were administered by using only drop and collect and CATIs, whereas the more recent surveys have been conducted using only the CATI method. Note that our data set consists of independent cross-sectional surveys over time. The key question is 'have you used marijuana or cannabis (cocaine, speed) in the last 12 months?'. Owing to consistency with respect to the key variables of interest and the change in the collection method in more recent years, we use data from the 2001, 2004 and 2007 surveys in this study. A sample of 56 579 individuals is thus available for estimation. These data have been used in several previous studies (e.g. Harris and Zhao (2007), Van Ours and Williams (2011) and Williams and Bretteville-Jensen (2014)).

In terms of explanatory variables, we require two sets: one to determine participation and the other misreporting. Although many of these variables overlap, to facilitate identification we ensure that $\mathbf{x}_m$ have exclusion restrictions. In terms of common variables, in line with several past studies on drug consumption (e.g. Gill and Michaels (1991), Saffer and Chaloupka (1999) and Cameron and Williams (2001)), we include a wide range of personal and demographic characteristics (such as gender, marital status, educational attainment, whether the individual lives in a state where possession of small amounts of marijuana is decriminalized and income; see the on-line appendix and Table 1). Inclusion of year and state dummies in both equations allows for the fact that both participation and misreporting rates may follow different trends over time, while also allowing for any difference in drug prices and policies across states.

We include a range of identifying variables in the misreporting equation: variables that (ostensibly) affect the misreporting decision(s) but not the participation decision(s). These identifying variables in $\mathbf{x}_m$ (to capture misreporting) mostly relate to the conditions under which the

survey was administered, which, as discussed above, may potentially influence the extent to which individuals misreport but will arguably be independent from any participation propensities. Specifically, we control for whether anyone else was present when the respondent was completing the survey questionnaire, whether anyone helped the respondent to complete the survey questionnaire and whether the drop-and-collect survey mode was used. These variables conform with the factors that have been associated with misreporting or misclassification in prior studies (e.g. Mensch and Kandel (1988), O'Muircheartaigh and Campanelli (1998), Lu *et al*. (2001), Kraus and Augustin (2001) and Berg and Lien (2006)).

Finally, we also include as an instrument a variable indicating a general lack of trust in the survey which we measure by using the percentage of unanswered questions. This is based on the significant amount of literature suggesting that the longer a respondent spends with the interviewer the more trusting they are of both him or her and the survey in general (e.g. Corbin and Morse (2003)). For each respondent it is possible to calculate the total number of compulsory (asked to everyone) questions that are left unanswered (as a percentage); this is clearly both a strong proxy for the length of time spent completing the survey, and as such an indirect proxy for trust, and also (arguably) a direct measure of trust. Table 1 presents summary

Table 1. Summary statistics, sample size 56579

Variable	*Mean*	*Standard deviation*	*Minimum*	*Maximum*
Y_{mar}	0.1215	0.3267	0	1
Y_{spd}	0.0316	0.175	0	1
Y_{coc}	0.0127	0.1118	0	1
MALE	0.4662	0.4989	0	1
STAGE	−0.0238	0.9352	−1.7157	2.9028
STAGESQ	−0.0460	0.9349	−1.2437	4.1373
MARRIED	0.5931	0.4913	0	1
PRESCHOOL	0.1232	0.3287	0	1
SINGPAR	0.0704	0.2558	0	1
CAPITAL	0.6437	0.4789	0	1
ATSI	0.0133	0.1144	0	1
WORK	0.6239	0.4844	0	1
STUDY	0.0619	0.2409	0	1
UNEMP	0.0225	0.1482	0	1
DEGREE	0.2626	0.4400	0	1
YR12	0.1295	0.3358	0	1
DIPLOMA	0.3488	0.4766	0	1
LRPINC	9.7776	0.9324	6.6400	11.2708
DECRIM	0.2534	0.4350	0	1
MIGR10	0.0435	0.2040	0	1
YR04	0.3593	0.4798	0	1
YR07	0.2887	0.4532	0	1
VIC	0.2053	0.4039	0	1
QLD	0.1794	0.3837	0	1
WA	0.1107	0.3137	0	1
SA	0.0840	0.2775	0	1
TAS	0.0474	0.2126	0	1
ACT	0.0538	0.2257	0	1
NT	0.0478	0.2134	0	1
PRESENT	0.2916	0.4545	0	1
HELP	0.2144	0.4104	0	1
SURVTYPE	0.1671	0.3730	0	1
TRUST	0.0397	0.0615	0	0.6688

statistics relating to the variables that were used in our econometric analysis for the pooled cross-section data set. The variables are fully described in the on-line appendix.

5. Results

5.1. Estimated parameters

In estimating the joint MIP model (with 15 correlation coefficients) we note that the estimated coefficients and their statistical significance do not change dramatically from the univariate correlated IP results. Although the differences in the estimated parameters from the more complex MIP model are quite negligible, the main advantage of estimating the system model is that it allows us to estimate a whole range of joint and conditional probabilities of interest (see below). In Table 2 we present the MIP estimated coefficients for the participation and misreporting equations and, in Table 3, the correlation coefficients. In general, the system model does a very good job, in terms of statistical significance, of modelling such difficult data with such low observed (recorded) participation rates. Before briefly describing some main findings, Table 3 shows that there are several significant correlations both across and within drugs (and the full set is jointly significant), suggesting the existence of complex interrelationships in the participation and reporting decisions here, and that both observed and unobserved heterogeneity play a significant role.

With regard to participation, we find that gender and marital status are significantly associated with all three drugs. Age has a statistically significant effect on the participation probabilities of marijuana and speed. In line with previous literature (see, for example, Farrelly *et al*. (2001) and Saffer and Chaloupka (1998)), decriminalization is negatively associated with participating in marijuana consumption. Income is negatively associated with marijuana consumption, which may be a reflection of social class. The relationship between education and illicit drug consumption appears to be somewhat complex with the effect varying significantly across the three drugs: education has no significant effect on cocaine use; in the case of marijuana, higher levels of education are associated with a higher probability of participation, whereas, for speed, the more highly educated is the individual, the lower is their probability of participating. Such findings may reflect the different social norms, recreational activities and/or preferences across educational groups.

Turning to the misreporting equations, and noting that a positive coefficient indicates a lower probability of misreporting, we see that being male is associated with a higher probability of misreporting speed but with lower chances of misreporting marijuana consumption. Age has no significant effect on the probability of misreporting marijuana but we find a quadratic effect of age on the probability of misreporting speed and cocaine. Interestingly, income is positively associated with accurately reporting all three drugs. As expected, the more educated individuals have a higher probability of misreporting marijuana and speed but education does not seem to affect reporting behaviour of cocaine. The introduction of decriminalization is likely to be associated with increased awareness of the potential consequences that are associated with consuming illicit drugs, through increased debate as well as campaigns (such as the Australian National Campaign Against Drug Abuse). Surprisingly, we do not find evidence of any effect of decriminalization on the probability of misreporting marijuana or any of the other drugs. We find an increasing trend in misreporting behaviour across the years, reflecting a changing trend in opinions with regard to drug use.

The identification of our model hinges, to a large extent, on the exclusion strategy that is employed. With respect to the effects of the identifying set of variables in the misreporting equation, the presence of anyone else when the respondent was completing the questionnaire is associated with a higher probability of misreporting across all three drugs, consistent with Hoyt

Table 2. Marijuana, speed and cocaine consumption: estimated coefficients†

Variable	*Results for marijuana*		*Results for speed*		*Results for cocaine*	
	Participation	*Misreporting*	*Participation*	*Misreporting*	*Participation*	*Misreporting*
CONSTANT	0.243	−0.763	−2.347	0.435	−2.339	−6.020
	(0.286)	(0.384)§	(0.336)‡	(1.446)	(1.066)§	(1.574)‡
MALE	0.409	0.148	0.381	−0.206	0.192	0.054
	(0.029)‡	(0.049)‡	(0.045)‡	(0.118)§§	(0.071)‡	(0.153)
STAGE	−1.010	0.445	−2.211	3.880	−1.131	5.039
	(0.183)‡	(0.311)	(0.387)‡	(0.925)‡	(1.045)	(1.483)‡
STAGESQ	−0.218	−0.022	1.099	−3.197	0.566	−7.383
	(0.167)	(0.540)	(0.430)§	(1.579)§	(1.478)	(1.987)‡
MARRIED	−0.523	−0.053	−0.521	0.410	−0.594	0.545
	(0.034)‡	(0.082)	(0.048)‡	(0.210)§§	(0.111)‡	(0.381)
PRESCHOOL	−0.041	−0.256	−0.320	0.251	−0.247	0.152
	(0.053)	(0.075)‡	(0.060)‡	(0.198)	(0.105)§	(0.347)
SINGPAR	0.035	0.055	0.078	−0.050	−0.098	−0.106
	(0.048)	(0.061)	(0.062)	(0.105)	(0.156)	(0.274)
CAPITAL	−0.067	0.109	0.135	0.123	0.312	0.211
	(0.031)§	(0.044)§	(0.042)‡	(0.085)	(0.102)‡	(0.227)
ATSI	0.004	0.196	0.025	−0.367	−0.582	1.200
	(0.106)	(0.143)	(0.145)	(0.227)	(0.326)§§	(0.910)
WORK	0.083	−0.339	−0.111	−0.270	0.034	0.014
	(0.052)	(0.095)‡	(0.069)	(0.180)	(0.164)	(0.359)
STUDY	0.518	−0.432	0.319	−0.644	0.541	−0.275
	(0.132)‡	(0.110)‡	(0.146)§	(0.205)‡	(0.252)§	(0.441)
UNEMP	0.135	0.309	0.038	0.259	0.172	−0.083
	(0.078)§§	(0.143)§	(0.103)	(0.241)	(0.248)	(0.458)
DEGREE	0.181	−0.444	−0.367	−0.344	−0.011	0.044
	(0.052)‡	(0.076)‡	(0.063)‡	(0.148)§	(0.136)	(0.261)
YR12	0.034	−0.158	−0.202	−0.091	0.020	0.034
	(0.046)	(0.059)‡	(0.065)‡	(0.103)	(0.145)	(0.249)
DIPLOMA	0.066	−0.089	−0.149	0.006	−0.103	0.359
	(0.036)§§	(0.059)	(0.050)‡	(0.103)	(0.135)	(0.266)
LRPINC	−0.160	0.195	0.029	0.159	0.033	0.427
	(0.028)‡	(0.032)‡	(0.035)	(0.057)‡	(0.077)	(0.116)‡
DECRIM	−0.253	0.165	0.026	−0.048	−0.223	0.508
	(0.082)‡	(0.110)	(0.105)	(0.185)	(0.213)	(0.423)
MIGR10	0.082	−0.415	0.103	−0.897	0.205	−0.419
	(0.097)	(0.089)‡	(0.124)	(0.187)‡	(0.149)	(0.246)§§
YR04	0.117	−0.213	0.121	−0.207	0.179	−0.623
	(0.035)‡	(0.053)‡	(0.046)‡	(0.094)§	(0.103)§§	(0.246)§
YR07	0.238	−0.539	0.210	−0.917	0.450	−0.638
	(0.053)‡	(0.067)‡	(0.073)‡	(0.142)‡	(0.108)‡	(0.306)§
VIC	−0.143	0.117	−0.101	−0.177	−0.306	0.207
	(0.041)‡	(0.059)§	(0.058)§§	(0.107)§§	(0.098)‡	(0.232)
QLD	−0.136	0.120	−0.107	−0.175	−0.529	0.382
	(0.043)‡	(0.060)§	(0.059)§§	(0.111)	(0.133)‡	(0.343)
WA	0.260	0.062	0.140	0.199	−0.113	−0.279
	(0.066)‡	(0.085)	(0.085)	(0.147)	(0.156)	(0.290)
SA	0.294	−0.066	0.024	0.125	−0.153	−0.581
	(0.098)‡	(0.135)	(0.127)	(0.226)	(0.266)	(0.499)
TAS	−0.022	0.121	−0.239	−0.367	−0.621	−0.187
	(0.065)	(0.098)	(0.111)§	(0.208)§§	(0.307)§	(0.661)
ACT	0.093	−0.022	−0.166	0.139	−0.261	−0.346
	(0.102)	(0.143)	(0.146)	(0.271)	(0.256)	(0.535)

(*continued*)

Table 2 (*continued*)

Variable	*Results for marijuana*		*Results for speed*		*Results for cocaine*	
	Participation	*Misreporting*	*Participation*	*Misreporting*	*Participation*	*Misreporting*
NT	0.715	−0.200	0.295	−0.701	−0.389	−0.211
	(0.109)‡	(0.156)	(0.139)§	(0.250)‡	(0.268)	(0.596)
PRESENT		−0.192		−0.404		−0.364
		(0.039)‡		(0.088)‡		(0.155)§
HELP		−0.051		0.024		0.118
		(0.048)		(0.099)		(0.172)
SURVTYPE		−0.212		−0.288		−0.664
		(0.058)‡		(0.111)‡		(0.256)‡
TRUST		−1.435		−1.753		−3.518
		(0.347)‡		(0.761)§		(1.348)‡

†Standard errors are given in parentheses. A positive coefficient for participation indicates an increase in the probability of participation whereas a negative coefficient for misreporting indicates an increase in the probability of misreporting.
‡Significant at the 1% level.
§Significant at the 5% level.
§§Significant at the 10% level.

Table 3. Correlation coefficients†

	M_{mar}	R_{mar}	M_{spd}	R_{spd}	M_{coc}	R_{coc}
M_{mar}	—					
R_{mar}	0.069	—				
	(0.135)					
M_{spd}	0.504	0.199	—			
	(0.066)‡	(0.235)				
R_{spd}	0.205	0.601	0.078	—		
	(0.094)§	(0.040)‡	(0.372)			
M_{coc}	0.300	0.299	0.025	0.028	—	
	(0.107)‡	(0.265)	(0.122)	(0.128)		
R_{coc}	0.101	0.498	0.037	0.031	0.080	—
	(0.096)	(0.076)‡	(0.118)	(0.043)	(0.426)	

†Standard errors are given in parentheses.
‡Significant at the 1% level.
§Significant at the 5% level.

and Chaloupka (1994). Seeking help from someone to complete the questionnaire does not appear to have a significant effect on reporting participation in any of the drugs. Clearly, survey type, i.e. the CATI method or face-to-face interview (relative to drop and collect), is associated with a higher probability of misreporting across all three drugs. Finally, if the respondent had a general lack of trust in the survey then they have a higher chance of misreporting drug use. In summary, since three of the four identifying variables exhibit high levels of significance and in the expected direction, we are confident in our identification strategy and, consequently, our results overall.

5.2. Predicted probabilities and partial effects

There are numerous probabilities that one may be interested in in predicting with the current

Table 4. Sample and predicted probabilities†

	Results for marijuana	*Results for speed*	*Results for cocaine*
Sample rate of participation	0.1215	0.0316	0.0127
Marginal probability of participation, $\Pr(r=1\|\mathbf{x})$	0.2326	0.0838	0.0486
	(0.0178)‡	(0.0112)‡	(0.0224)‡
Joint probability of participation and accurate reporting, $\Pr(r=1,m=1\|\mathbf{x})$	0.1206	0.0281	0.0137
	(0.0013)‡	(0.0007)‡	(0.0007)‡
Probability of misreporting conditional on participation $\Pr(m=0\|r=1,\mathbf{x})$	0.3064	0.1702	0.6467
	(0.0683)‡	(0.0698)‡	(0.1236)‡
Components of the 0s			
Non-participation, $\Pr(r=0\|\mathbf{x})$	0.7674	0.9162	0.9514
	(0.0178)‡	(0.0112)‡	(0.0224)‡
Misreporting, $\Pr(r=1,m=0\|\mathbf{x})$	0.1120	0.0558	0.0349
	(0.0178)‡	(0.0111)‡	(0.0224)
Total	0.8794	0.9719	0.9863
	(0.0013)‡	(0.0007)‡	(0.0007)‡
Posterior probabilities			
Non-participation, $\Pr(r=0\|\mathbf{x},y=0)$	0.8692	0.9509	0.9690
	(0.0204)‡	(0.0098)‡	(0.0204)‡
Misreporting, $\Pr(r=1,m=0\|\mathbf{x},y=0)$	0.1308	0.0491	0.0310
	(0.0204)‡	(0.0098)‡	(0.0204)

†Standard errors are given in parentheses.
‡Significant at the 1% level.

model. For each drug in isolation, one may be interested in the marginal probability of participation, the joint probability of participation and misreporting or the probability of accurate reporting, conditionally on participation. In Table 4 we present some summary probabilities that are associated with each of the drugs (evaluated individually and then averaged over the sample). As expected, across all three drugs, the predicted marginal probabilities of participation are higher than the sample rates of participation as indicated by the survey responses. Specifically, on the basis of the survey responses, one would estimate participation rates in marijuana, speed and cocaine respectively to be 12.2%, 3.2% and 1.3%. However, we estimate, once misreporting has been taken into account, that these are significantly higher at 23.3%, 8.4% and 4.9% respectively. Given the small standard errors of these, they also appear to be quite precisely estimated. The joint probability of participation and accurate reporting (alternatively, the *recorded* probability of participation, $\Pr(y=1|\mathbf{x})$) allows us to assess the performance of our model as they are directly comparable with the sample proportions. We find that for all three drugs the joint probabilities mimic the observed sample proportions very closely.

Conditionally on an individual participating, there was a 65% chance of misreporting cocaine use compared with 31% for marijuana, i.e. of the small percentage of cocaine users in the population (recorded at 1.3% and estimated at just under 5%) 65% claimed not to be. This may appear high, but it is in line with previous studies (Pudney, 2007). This difference between cocaine and marijuana may reflect the greater risk that is associated with the former. For speed the estimated conditional probability of misreporting was 17%, which is significantly less than for the other two drugs. This lower misreporting rate may be related to the younger age and lower education of speed consumers, a demographic for which speed consumption may be considered more 'socially acceptable'. Overall, these findings suggest that misreporting in survey

data may lead to considerable underestimation of participation rates in the case of consumption of illicit drugs, especially with regard to both marijuana and cocaine in the current study.

To gain more insights into the source of the observed 0s, we also present in Table 4 the predicted probability of 0s for each of the three drugs broken down into two respective components: non-participation and misreporting. For example, the overall predicted probability of 87.9% of zero consumption in the case of marijuana is made up of the respective probability of non-participation (76.7%) and misreporting (11.2%). In view of the low rates of participation, the low misreporting components here (of 11.2%, 5.6% and 3.5%) may appear to be quite small. However, when translated to the Australian population aged 14 years and above, they represent nearly 2016000, 1004000 and 629000 cases of unreported cases of marijuana, speed and cocaine use respectively. Such underreporting can thus have extremely important implications for drug policies.

Such probabilities can be thought of as *prior* probabilities, i.e. they apply to a randomly selected individual from the population, about whom we know nothing except their characteristics. However, to provide further insights into the extent of misreporting, it is possible to estimate *posterior* probabilities, analogously to those considered in latent class models (Greene, 2012), that are conditional on knowing what outcome the individual chose. Specifically, this allows us to make a prediction on what percentage of these 0s comes from non-participation and misreporting respectively, using all the information that we have on the individual: this attempts to answer the question 'given that an individual recorded a 0, what is the probability that they are a true non-participant *versus* a misreporting participant (given their observed characteristics)?'. The posterior probabilities for the two types of 0s are given as

$$\begin{aligned}\Pr(r=0|\mathbf{x}, y=0) &= \frac{f(r=0|\mathbf{x})}{f(y=0)} \\ &= \frac{1-\Phi(\mathbf{x}_r'\boldsymbol{\beta}_r)}{1-\Phi(\mathbf{x}_r'\boldsymbol{\beta}_r)+\Phi_2(\mathbf{x}_r'\boldsymbol{\beta}_r, -\mathbf{x}_m'\boldsymbol{\beta}_m, -\rho_{rm})}\end{aligned} \tag{12}$$

and

$$\begin{aligned}\Pr(r=1, m=0|\mathbf{x}, y=0) &= \frac{f(r=1, m=0|\mathbf{x})}{f(y=0)} \\ &= \frac{\Phi_2(\mathbf{x}_r'\boldsymbol{\beta}_r, -\mathbf{x}_m'\boldsymbol{\beta}_m, -\rho_{rm})}{1-\Phi(\mathbf{x}_r'\boldsymbol{\beta}_r)+\Phi_2(\mathbf{x}_r'\boldsymbol{\beta}_r, -\mathbf{x}_m'\boldsymbol{\beta}_m, -\rho_{rm})}.\end{aligned} \tag{13}$$

From Table 4, we find that close to 87% of the reported 0s for marijuana are estimated to come from genuine non-participation (and therefore 13% from misreported participation). Note that, as with the prior probabilities that were presented earlier, these posterior probabilities for misreporting might appear, superficially, rather low. However, it is important to remember that the probabilities for misreporting here are not marginal, but joint of participation *and* misreporting. Thus, given that participation probabilities are very low for all these drugs (estimated at about 23%, 8% and 5% respectively for marijuana, speed and cocaine; see the second row of Table 4), it is not surprising that these joint probabilities are also small. Moreover, as with all the predicted probabilities, estimated standard errors are generally (relatively) very small, giving us greater confidence in their magnitudes.

Considering the full system of demand equations, as in the current approach, one may also be interested in any of numerous cross-drug probabilities such as the joint probability of participating in marijuana, speed and cocaine and the conditional probability of misreporting cocaine conditionally on marijuana participation. Indeed, it is not immediately obvious how one would undertake such an exercise if these drug equations were estimated separately. We can also

Table 5. Partial effects on selected joint and conditional probabilities†

Variable	$Pr(y_{mar}=0, y_{spd}=0, y_{coc}=0 \mid \mathbf{x})$			$Pr(y_{spd}=0, y_{coc}=0 \mid r_{mar}=1, \mathbf{x})$		
	Participation	*Misreporting*	*Overall*	*Participation*	*Misreporting*	*Overall*
MALE	−0.053	−0.006	−0.059	−0.030	0.047	0.017
	(0.009)‡	(0.003)§§	(0.007)‡	(0.015)§	(0.021)§	(0.021)
STAGE	0.143	−0.028	0.115	0.238	0.070	0.308
	(0.036)‡	(0.022)	(0.045)§	(0.074)‡	(0.112)	(0.107)‡
STAGESQ	0.016	0.010	0.026	−0.151	0.086	−0.065
	(0.021)	(0.024)	(0.033)	(0.068)§	(0.189)	(0.174)
MARRIED	0.067	0.002	0.069	0.048	−0.024	0.023
	(0.013)‡	(0.004)	(0.010)‡	(0.020)§	(0.027)	(0.026)
PRESCHOOL	0.008	0.011	0.019	0.040	−0.084	−0.044
	(0.007)	(0.005)§	(0.005)‡	(0.010)‡	(0.027)‡	(0.030)
SINGPAR	−0.005	−0.002	−0.008	−0.007	0.019	0.012
	(0.007)	(0.003)	(0.005)	(0.008)	(0.020)	(0.020)
CAPITAL	0.007	−0.005	0.002	−0.024	0.032	0.008
	(0.004)§§	(0.002)§	(0.003)	(0.008)‡	(0.015)§	(0.017)
ATSI	−0.002	−0.009	−0.012	0.005	0.051	0.056
	(0.016)	(0.012)	(0.014)	(0.025)	(0.058)	(0.065)
WORK	−0.009	0.015	0.007	0.017	−0.106	−0.089
	(0.006)	(0.005)‡	(0.005)	(0.009)§§	(0.034)‡	(0.035)§
STUDY	−0.064	0.020	−0.044	−0.023	−0.131	−0.153
	(0.017)‡	(0.007)‡	(0.017)‡	(0.024)	(0.040)‡	(0.038)‡
UNEMP	−0.016	−0.014	−0.030	−0.001	0.097	0.097
	(0.013)	(0.009)	(0.009)‡	(0.020)	(0.050)§§	(0.052)§§
DEGREE	−0.018	0.020	0.002	0.053	−0.140	−0.087
	(0.007)‡	(0.006)‡	(0.006)	(0.013)‡	(0.032)‡	(0.038)§
YR12	−0.002	0.007	0.005	0.026	−0.050	−0.024
	(0.006)	(0.003)§	(0.005)	(0.009)‡	(0.019)‡	(0.021)
DIPLOMA	−0.007	0.004	−0.003	0.022	−0.032	−0.009
	(0.005)	(0.003)	(0.004)	(0.008)‡	(0.019)§§	(0.020)
LRPINC	0.019	−0.009	0.010	−0.011	0.056	0.045
	(0.004)‡	(0.002)‡	(0.004)§	(0.007)	(0.013)‡	(0.015)‡
DECRIM	0.030	−0.008	0.022	−0.012	0.047	0.035
	(0.011)‡	(0.005)	(0.010)§	(0.016)	(0.035)	(0.036)
MIGR10	−0.011	0.020	0.009	−0.012	−0.123	−0.134
	(0.012)	(0.008)§	(0.010)	(0.017)	(0.034)‡	(0.036)‡
YR04	−0.015	0.010	−0.005	−0.012	−0.060	−0.071
	(0.005)‡	(0.003)‡	(0.005)	(0.007)	(0.019)‡	(0.019)‡
YR07	−0.030	0.025	−0.004	−0.021	−0.160	−0.180
	(0.007)‡	(0.006)‡	(0.009)	(0.013)	(0.035)‡	(0.031)‡
VIC	0.017	−0.005	0.012	0.010	0.035	0.045
	(0.006)‡	(0.003)§§	(0.005)§	(0.009)	(0.020)§§	(0.019)§
QLD	0.016	−0.005	0.011	0.014	0.035	0.049
	(0.006)§	(0.003)§§	(0.006)§§	(0.009)	(0.020)§§	(0.020)§
WA	−0.033	−0.003	−0.036	−0.004	0.022	0.018
	(0.009)‡	(0.004)	(0.008)‡	(0.014)	(0.028)	(0.028)
SA	−0.036	0.003	−0.033	0.013	−0.015	−0.003
	(0.013)‡	(0.007)	(0.011)‡	(0.019)	(0.043)	(0.043)
TAS	0.004	−0.005	−0.001	0.036	0.042	0.078
	(0.009)	(0.006)	(0.008)	(0.016)§	(0.034)	(0.035)§
ACT	−0.010	0.001	−0.009	0.028	−0.004	0.024
	(0.013)	(0.006)	(0.011)	(0.018)	(0.046)	(0.047)
NT	−0.090	0.010	−0.080	0.002	−0.058	−0.056
	(0.017)‡	(0.009)	(0.016)‡	(0.031)	(0.048)	(0.046)

(*continued*)

Table 5 (*continued*)

Variable	$Pr(y_{mar}=0, y_{spd}=0, y_{coc}=0\|\mathbf{x})$			$Pr(y_{spd}=0, y_{coc}=0\|r_{mar}=1, \mathbf{x})$		
	Participation	*Misreporting*	*Overall*	*Participation*	*Misreporting*	*Overall*
PRESENT	0.000	0.009	0.009	0.000	−0.055	−0.055
	(0.000)	(0.003)‡	(0.003)‡	(0.000)	(0.014)‡	(0.014)‡
HELP	0.000	0.002	0.002	0.000	−0.017	−0.017
	(0.000)	(0.002)	(0.002)	(0.000)	(0.015)	(0.015)
SURVTYPE	0.000	0.010	0.010	0.000	−0.059	−0.059
	(0.000)	(0.004)‡	(0.004)‡	(0.000)	(0.019)‡	(0.019)‡
TRUST	0.000	0.069	0.069	0.000	−0.409	−0.409
	(0.000)	(0.022)‡	(0.022)‡	(0.000)	(0.119)‡	(0.119)‡

†Standard errors are given in parentheses. A positive marginal effect for participation represents an increase in the probability of participation whereas a negative marginal effect for misreporting represents an increase in the probability of misreporting.
‡Significant at the 1% level.
§Significant at the 5% level.
§§Significant at the 10% level.

estimate partial effects on all these different marginal, joint and conditional probabilities. For brevity, we present partial effects for a joint and a conditional probability, which we discuss briefly. Full results are available from the authors on request. In particular, Table 5 presents partial effects on the probabilities of the two cases (estimated at sample means): the *recorded* probability of zero consumption of all three drugs ($\Pr(y_{mar}=0, y_{spd}=0, y_{coc}=0|\mathbf{x})$) and the probability of reporting zero consumption of speed and cocaine, conditionally on *predicted* participation in marijuana, i.e. $\Pr(y_{spd}=0, y_{coc}=0|r_{mar}=1, \mathbf{x})$.

Consider first the zero reported consumption of all three drugs ($\Pr(y_{mar}=0, y_{spd}=0, y_{coc}=0|\mathbf{x})$). It appears that being male is inversely associated with this probability, with the non-participation and misreporting effects serving to operate in the same direction. For instance, males are 5.3 percentage points (PPs) less likely to abstain from all three drugs and they have a 0.6-PP lower chance of accurately reporting such zero consumption. This results in an overall 5.9-PP lower probability of recording zero consumption for males compared with females. Some of the effects of main occupation and education are interesting with negative effects on the probability of reporting non-participation across all three drugs with the misreporting effects operating in the opposite direction, thereby serving to moderate the participation effects.

Turning to education, degree holders have a 1.8-PP lower chance of abstaining from all three drugs but a 2-PP higher chance of accurately reporting such non-participation, resulting in an overall 0.2-PP lower probability of recording joint zero consumption across all three drugs relatively to those with less than year 12 qualifications. However, the overall effect is statistically insignificant. In terms of the additional variables in the misreporting equation, positive statistically significant partial effects are apparent for three of the four survey-related variables, again highlighting the important role of survey conditions in the collection of accurate (or otherwise) information.

The negative year effects indicate an increase in drug use over time. In contrast, we observe a rise in accurate reporting of such non-participation across the years. We also observe some significant state effects on the probability of zero consumption.

Next we look at the joint probability of observing a 0 for speed and cocaine, conditionally on being a marijuana user ($\Pr(y_{spd}=0, y_{coc}=0|r_{mar}=1, \mathbf{x})$). Bringing an analogy with the gateway

effect where there is a progression from soft drugs to hard drugs, this probability allows us to examine zero reporting (or non-participation) in the case of the harder drugs, cocaine and speed, in a subpopulation of marijuana users. We find a significant association of factors such as gender, presence of young children, employment and education with the non-participation of speed and cocaine in the subpopulation of marijuana participants. For example males have a 3-PP lower probability of non-participation in speed and cocaine than females, if they are already marijuana users. Put differently, males are more likely to be hard-core drug users if they are already marijuana users, which is consistent with a gateway effect for males.

5.3. Robustness checks, false positives results and validation exercises

The instruments that we use to identify the misreporting equation are all survey related, which makes them unlikely to be related to drug participation, providing a strong case for identification. The importance of these factors in the misclassification literature and their statistical significance in the estimated model lend further support to their inclusion in the misreporting equation. Explicitly testing the validity of instruments in non-linear models is a difficult task (see, for example, Davidson and MacKinnon (1993)) and there may also be concerns that some of the instruments such as *present* and *help* are correlated with unobserved characteristics and are therefore potentially endogenous. In light of this we therefore perform a series of robustness checks to test whether our results change significantly with the inclusion and exclusion of the respective identifying variables. Comparing across the resulting marginal effects, we find that the results are generally robust to the various specifications for marijuana and speed although we observe some differences for cocaine (which are presumably a result of the very low recorded participation rate). Although for brevity we do not present the results from the various model specifications, they are available from the authors on request. Instead, in Table 6 we provide a comparison of the various specifications on the basis of the joint probability of participation and accurate reporting which we contrast with the sample rate of drug participation. Consistent with the results relating to the marginal effects, for marijuana and speed the joint probability of participation and accurate reporting mimics the sample rate of participation quite well whereas we see some differences for cocaine. Thus, in short, our findings do not appear to be heavily reliant on the particular choice of identifying variable(s).

Although most of the reporting bias is believed to be in the direction of underreporting there is also some evidence from the literature on overreporting. Such false positive rates are generally lower than false negative ones (see, for example, Visher and McFadden (1991) and Harrison and Hughes (1997)). However, as a 'litmus test' to gauge the likely magnitudes of any false positive results we conduct a simple test reversing the 1s to 0s and re-estimating the model. For all three drugs the misreporting effects appear to be very small, ranging from 0.81% for cocaine to 4.67% for marijuana. We also extended the basic framework to allow jointly for false positive and false negative results. The estimated rates for the former were found to be even lower (at 0.04%, 0.02% and 0.06% respectively). Both of these exercises suggest that the levels of misreporting with regard to false positive results are very low and therefore would not unduly affect the main results that are reported in the paper.

We also restricted the sample to individuals who have reported having ever used the drug (the National Drug Stategy Household Survey does not collect information on previous month's use): if the model is well specified misreporting rates should be significantly lower as stigma rates will obviously be much reduced in this subsample. Indeed, we do find much less misreporting in these subsamples (for example, the percentage difference bias between the observed proportions of marijuana users, 0.122, and the predicted rate of users, 0.259, drops significantly from 113% for the full sample to 55% for the subsample of those who have ever used marijuana). Again, this

Table 6. Comparison across specifications†

	Results for main specification	*Results for specification 1*	*Results for specification 2*	*Results for specification 3*	*Results for specification 4*	*Results for specification 5*
Marijuana						
Sample rate of participation	0.1215	0.1215	0.1215	0.1215	0.1215	0.1215
Joint probability of participation and accurate reporting, $\Pr(r_{\text{mar}}=1, m_{\text{mar}}=1\|\mathbf{x})$	0.1206 (0.0013)‡	0.1079 (0.0077)‡	0.1411 (0.0096)‡	0.1431 (0.0062)‡	0.1442 (0.0053)‡	0.1506 (0.0097)‡
Speed						
Sample rate of participation	0.0316	0.0316	0.0316	0.0316	0.0316	0.0316
Joint probability of participation and accurate reporting, $\Pr(r_{\text{spd}}=1, m_{\text{spd}}=1\|\mathbf{x})$	0.0281 (0.0007)‡	0.0386 (0.0011)‡	0.0373 (0.0010)‡	0.0383 (0.0008)‡	0.0360 (0.0010)‡	0.0352 (0.0019)‡
Cocaine						
Sample rate of participation	0.0127	0.0127	0.0127	0.0127	0.0127	0.0127
Joint probability of participation and accurate reporting, $\Pr(r_{\text{coc}}=1, m_{\text{coc}}=1\|\mathbf{x})$	0.0137 (0.0007)‡	0.0028 (0.0010)‡	0.0060 (0.0009)‡	0.0031 (0.0010)‡	0.0032 (0.0007)‡	0.0097 (0.0006)‡

†Standard errors are given in parentheses. The five specifications are similar to the main specification except for the following details: specification 1, present only; specification 2, help only; specification 3, survey type only; specification 4, trust only; specification 5, survey type and trust only.
‡Significant at the 1% level.

validation exercise gives us strong confidence in our main findings. Unfortunately the subsamples of those who have ever used speed and cocaine are too small for robust analysis.

The final validation exercise that we conduct involves estimating the model on legal drugs (alcohol and tobacco), which, unlike illegal drugs, do not pose any risk of legal prosecution and are much less stigmatized given their general acceptability in the community. Thus, we would expect less reporting bias in the case of legal drugs. On the basis of correlated IP models, we still find strong effects of the identifying variables in the misreporting equations (the results are available on request) and in Table 7 we present the recorded and predicted probabilities of participation and the implied percentage biases. Clearly the reporting biases are much smaller for alcohol (11%) and tobacco (63%) relative to the illegal drugs (where biases are 113%, 172%

Table 7. Comparison across legal and illegal drugs†

	Results for alcohol	*Results for tobacco*	*Results for marijuana*	*Results for speed*	*Results for cocaine*
Sample rate of participation	0.851	0.212	0.122	0.032	0.013
Predicted rate of participation, $\Pr(\|r=1, \mathbf{x})$	0.947	0.345	0.259	0.086	0.045
% bias	11%	63%	113%	172%	257%
Probability of misreporting conditional on participation, $\Pr(m=0\|r=1, \mathbf{x})$	0.101	0.257	0.357	0.154	0.643

†Probabilities are estimated from correlated IP models.

and 257%, for marijuana, speed and cocaine respectively). Moreover, *a priori* we would expect higher bias for tobacco relative to alcohol, due to the stronger adverse stigma that is associated with the former. Table 7 also reports misreporting probabilities, conditionally on an individual participating. Here we see that, for alcohol and tobacco, there is a 10% and 26% chance of misreporting, compared with the much higher values that were found for the illegal drugs. So, once more, this validation exercise gives us strong confidence in the model, the identifying variables and the empirical findings.

5.4. Finite sample performance of the multivariate inverse probit model

A key contribution of this paper is the use of a multivariate model which, as noted earlier, allows us to estimate jointly drug participation and misreporting decisions across a range of drugs. A parsimonious model of drug consumption such as a simple probit model that does not take into account any misreporting will yield not only erroneous prediction of drug participation rates but also biased parameter estimates. To highlight such differences, in Table A1 in the online appendix we compare partial effects (on the marginal probability of drug participation) of some selected covariates from the MIP model and simple probit models. Clearly, we see some contrasting effects from the two models, with differences in magnitude and statistical significance. In extreme cases such as tertiary degree and income, we have opposite effects of covariates on drug participation. Thus, a simple probit model is likely to provide biased parameter estimates and partial effects if misreporting is prevalent.

The MIP model is also preferred to the correlated IP model as it takes into account the likely cross-drug correlations. If estimated in isolation where correlations across participation and misreporting equations exist, but are ignored, estimated parameters and subsequent analysis are potentially biased and/or inefficiently estimated since they are based on misspecified models and/or models where not all relevant information is being utilized. For instance, from our data, the *observed* sample proportion of individuals jointly consuming marijuana, speed and cocaine is 0.64%. Using the MIP model we would predict this joint probability to be 0.614% whereas the correlated IP model would estimate it at only 0.005%. Clearly the MIP model that fully accounts for correlations within and across drugs exhibits better performance than the correlated IP model.

Indeed, if simple univariate estimations of these models yielded essentially the same results as the much more complex MIP approach, then researchers would surely prefer the former (although some quantities of potential interest would be lost, or not as easily obtained). To ascertain the relative performance of a range of models that could have potentially been considered here we conduct some Monte Carlo simulations. To make these findings more relevant to the application at hand we simulate on the observed data and estimated parameters. Explicitly, we compare the multivariate model performance with that of a univariate model, i.e. the correlated IP model where the participation and misreporting equations are correlated for each drug but not across the drugs. Additionally we also consider the model that was suggested by Hausman *et al*. (1998), which provides a good basis for comparison; here the participation equation is specified as in the correlated IP or MIP model but the misreporting probabilities are constants (and not a function of covariates). We thus compare the (relative) performance of the model of Hausman *et al*. (1998) (for consistency, considered in a systems framework) with the correlated IP and MIP models.

For the data-generating process (DGP) we consider three scenarios:

(a) the true MIP model,
(b) three independent IPC models and
(c) the Hausman *et al*. (1998) form.

With the last, the error terms in the participation and misreporting equations are allowed to be correlated. Therefore, the MIP is a generalized model for both the correlated IP and the Hausman *et al.* (1998) model, in terms of allowing decisions to be taken jointly and misreporting decisions to be influenced by covariates respectively. Finally, we consider an additional set of experiments to determine whether the model is sensitive to the underlying assumptions of normality.

5.4.1. Monte Carlo evidence

As noted, with such a highly specified model, a comparison of all estimated coefficients would not be particularly illuminating. Instead, in comparing across the various approaches we examine a range of estimated summary probabilities as we envisage that these would be of primary interest to policy makers. In each scenario, for a particular probability we present

(a) the true average probability over Q replications, $\bar{P}(\cdot)$,
(b) the estimated average probability over Q replications, $\hat{\bar{P}}(\cdot)$, and
(c) the root-mean-square error of the estimated probability, $\text{RMSE}_{P(\cdot)}$.

To shed light on estimated parameters, we also report the averaged root-mean-square error over all estimated parameters, $\text{AveRMSE}_{\text{para}}$.

As expected, for the (true) MIP model all estimated probabilities are very close to the corresponding true probabilities and with very low RMSEs. In fact, almost all RMSEs from the MIP model are lower than those from the correlated IP and Hausman *et al.* (1998) models. Although the correlated IP model performs well in estimating marginal probabilities of a single drug consumption and probabilities of misreporting conditional on actual consumption in a single drug, the estimated joint probabilities of consuming more than one type of drug appear to be quite out. Ignoring the influence of individual characteristics on misreporting (i.e. the model of Hausman *et al.* (1998)) appears generally to result in even greater discrepancies all associated with high RMSEs. For details, refer to the results presented in Table A2 in the on-line appendix. Although it might not be strictly valid to generalize these findings universally, they suggest that, if cross-drug correlations do exist, finite sample biased quantities of interest might result if ignored. And even more severe biases can arise if misreporting or misclassification propensities are a function of covariates and these are ignored in estimations. The results on averaged RMSEs over all estimated parameters reinforce the above results, i.e. the finite sample estimation bias from the MIP model is essentially 0, with that from the correlated IP model being significantly higher and that from the model of Hausman *et al.* (1998) higher still.

When the true DGP is an independent correlated IP model for each of the three drugs, we see that again the estimated probabilities from the MIP model are very close to the true probabilities and with low RMSEs. We also find that the MIP does an excellent job of estimating the true non-zero correlations; and the average estimated correlation coefficients across drugs are all very close to 0 (their true values) and with very low RMSEs. Moreover, the averaged RMSE over all parameters is very small at 0.002. Detailed results can be found in Table A3 and Table A4 in the on-line appendix. Thus, even if correlations do not exist across drugs but within, the MIP is still a 'safe option' in correctly estimating all quantities of interest. When the true DGP is the model of Hausman *et al.* (1998) (where an individual's decision to misreport is not influenced by their characteristics but cross- and within-drug correlations exist), although being overspecified, the MIP model performs exceptionally well in terms of estimating the probabilities that were considered. For example, for $\Pr(m_{\text{coc}} = 0 | r_{\text{coc}} = 1)$, whereas the true probability is 94.15%, this is estimated as (on average) 94.13% by the MIP model, with an RMSE of 0.0056. And once

more, the averaged RMSE of all estimated parameters is extremely small (at 0.0016). All relevant results are listed in Table A5 in the on-line appendix.

The assumption that the error terms in the multiple-equation system independently and identically follow a multivariate normal distribution could be considered relatively restrictive, but this specification does allow us to estimate jointly participation and misreporting behaviours across a range of drugs. Some previous studies have relaxed such distributional assumptions (e.g. Chen *et al*. (2009) and Feng and Hu (2013)) but they have not considered misreporting or misclassification on multiple events, i.e. not allowing for correlations across events. In our application, we indeed find significant cross-drug correlations, which, if ignored in estimation, can have adverse effects on the results (as demonstrated above). However, the assumption of normality can be viewed as an identifying assumption; therefore finally we conduct some experiments to ascertain how important this identifying assumption is. With the MIP model as the true DGP, we now allow the multivariate error terms to have non-normal distributions: following a mixture distribution of $0.95N(0, \Sigma_6) + 0.05N(0, I_6)$. The results demonstrate that the MIP model estimations are again robust to this scenario (Table A6 in the on-line appendix). We repeated this exercise assuming other forms of non-normality, such as other mixing distributions and multivariate t-distributions with very low degrees of freedom, and, again, the results were essentially robust to such violations. These findings, overall, give us strong confidence in our overall findings and approach.

6. Conclusions

In this paper we have explored the potential implications of misreporting in survey data in the context of reporting consumption of three illicit drugs (marijuana, cocaine and speed). The widespread use of data collected from individual and household level surveys by researchers and policy makers is clearly reliant on respondents supplying accurate and reliable information. Indeed, estimated participation rates of illegal drugs are invariably inferred from such sample-based data. It is apparent, however, that in the context of gathering sensitive information individuals may misreport their true situation, leading (here) to an excess amount of 0-observations in the context of questions relating to activities such as illicit drug consumption: individuals are likely to deny their participation for a variety of reasons, such as fear of being caught, stigma and moral concerns.

Overall, we find that misreporting has a significant effect on observed participation rates such that, across all three drugs, the predicted marginal probabilities of participation are substantially higher than indicated by the raw data. This is caused by some quite high propensities to misreport. Interestingly, our findings suggest that the extent of misreporting is influenced by how the survey was administered and how much trust participants placed on the survey, as well as factors such as the presence of other individuals when the survey was completed. We conclude that the conditions under which survey data are collected influence the accuracy of the information that is obtained. Our findings suggest that accounting for misreporting is important in the context of using survey data related to sensitive activities, especially where such data are used to inform public policy.

Acknowledgements

We are very grateful to the Associate Editor and two referees for valuable comments. We are also grateful to the Australian Research Council and the Bankwest Curtin Economics Centre

for their generous funding. We also thank Steve Pudney for helpful comments and suggestions. The usual *caveats* apply.

References

Becker, G. S., Grossman, M. and Murphy, K. M. (1994) An empirical analysis of cigarette addiction. *Discussion Paper w3322*. National Bureau of Economic Research, Cambridge.

Berg, N. and Lien, D. (2006) Same-sex sexual behaviour: US frequency estimates from survey data with simultaneous misreporting and non-response. *Appl. Econ.*, **38**, 757–769.

Bhat, C. R. (2003) Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transprtn Res.* B, **37**, 837–855.

Bowling, A. (2005) Mode of questionnaire administration can have serious effects on data quality. *J. Publ. Hlth*, **27**, 281–291.

Cameron, L. and Williams, J. (2001) Cannabis, alcohol and cigarettes: substitutes or compliments. *Econ. Rec.*, **77**, 19–34.

Chen, X., Hu, Y. and Lewbel, A. (2009) Nonparametric identification and estimation of nonclassical errors-in-variables models without additional information. *Statist. Sin.*, **34**, 949–968.

Collins, R., Ellickson, P. and Bell, R. (1998) Simultaneous polydrug use among teens: prevalence and predictors. *J. Subst. Abuse*, **10**, 233–253.

Corbin, J. and Morse, J. (2003) The unstructured interactive interview: issues of reciprocity and risks when dealing with sensitive topics. *Qual. Enq.*, **9**, 335–354.

Cragg, J. (1971) Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, **39**, 829–844.

Davidson, R. and MacKinnon, J. G. (1993) *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.

Duarte, R., Escario, J. and Molina, J. (2005) Participation and consumption of illegal drugs among adolescents. *Int. Adv. Econ. Res.*, **11**, 399–415.

Dustmann, C. and Soest, A. (2001) Language fluency and earnings: estimation with misclassified language indicators. *Rev. Econ. Statist.*, **83**, 663–674.

Farrelly, M., Bray, J., Zarkin, G. and Wendling, B. (2001) The joint demand for cigarettes and marijuana: evidence from the National Household Surveys on Drug Abuse. *J. Hlth Econ.*, **20**, 51–68.

Fendrich, M. and Vaughn, C. (1994) Diminished lifetime substance use over time: an inquiry into differential under-reporting. *Publ. Opin. Q.*, **58**, 96–123.

Feng, S. and Hu, Y. (2013) Misclassification errors and the underestimation of the US unemployment rate. *Am. Econ. Rev.*, **103**, 1054–1070.

Gill, A. and Michaels, R. (1991) The determinants of illegal drug use. *Contemp. Econ. Poly*, **9**, 93–105.

Greene, W. (1994) Accounting for excess zeros and sample selection in Poisson and Negative Binomial regression models. *Working Paper EC-94-10*. Stern School of Business, New York University, New York.

Greene, W. (2012) *Econometric Analysis*, 7th edn. Englewood Cliffs: Prentice Hall.

Harris, M. and Zhao, X. (2007) A zero-inflated ordered Probit model, with an application to modelling tobacco consumption. *J. Econmetr.*, **141**, 1073–1099.

Harrison, L. and Hughes, A. (1997) *The Validity of Self-reported Drug Use: Improving the Accuracy of Survey Estimates*. Rockville: National Institute on Drug Abuse.

Hausman, J., Abrevaya, J. and Scott-Morton, F. (1998) Misclassification of the dependent variable in a discrete-response setting. *J. Econmetr.*, **87**, 239–269.

Heilbron, D. (1989) Generalized linear models for altered zero probabilities and overdispersion in count data. *Technical Report*. Department of Epidemiology and Biostatistics, University of California, San Francisco.

Hoyt, G. and Chaloupka, F. (1994) Effect of survey conditions on self-reported substance use. *Contemp. Econ. Poly*, **12**, no. 3, 109–121.

Ives, R. and Ghelani, P. (2006) Polydrug use (the use of drugs in combination): a brief review. *Drugs Educ. Prevn Poly*, **13**, 225–232.

Keane, M. P. (1994) A computationally practical simulation estimator for panel data. *Econometrica*, **62**, 95–116.

Kraus, L. and Augustin, R. (2001) Measuring alcohol consumption and alcohol-related problems: comparison of responses from self-administered questionnaires and telephone interviews. *Addiction*, **96**, 459–471.

Labeaga, J. M. (1999) A double-hurdle rational addiction model with heterogeneity: estimating the demand for tobacco. *J. Econmetr.*, **93**, 49–72.

Lambert, D. (1992) Zero inflated Poisson regression with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Lu, N., Taylor, B. and Riley, K. (2001) The validity of adult arrestee self-reports of crack cocaine use. *Am. J. Alc. Abuse*, **27**, 399–419.

MacDonald, Z. and Pudney, S. (2000) Analysing drug abuse with British Crime Survey data: modelling and questionnaire design issues. *Appl. Statist.*, **49**, 95–117.
MacDonald, Z. and Pudney, S. (2003) The use of self-report and drugs tests in the measurement of illicit drug consumption. *Discussion Paper in Economics 03/3*. Department of Economics, University of Leicester, Leicester.
Mensch, B. and Kandel, D. (1988) Underreporting of substance use in a national longitudinal youth cohort. *Publ. Opin. Q.*, **52**, 100–124.
Morrison-Beedy, D., Carey, M. and Tu, X. (2006) Accuracy of audio computer-assisted self-interviewing (ACASI) and self-administered questionnaires for the assessment of sexual behavior. *AIDS Behav.*, **10**, 541–552.
Mullahey, J. (1986) Specification and testing of some modified count data models. *J. Econmetr.*, **33**, 341–365.
Mullahey, J. (1997) Heterogeneity, excess zeros and the structure of count data models. *J. Appl. Econmetr.*, **12**, 337–350.
National Drug Strategy Household Surveys (2010) Computer files for the unit record data from the National Drug Strategy Household Surveys. Social Science Data Archives, Australian National University, Canberra.
O'Muircheartaigh, C. and Campanelli, P. (1998) The relative impact of interviewer effects and sample design effects on survey precision. *J. R. Statist. Soc.* A, **161**, 63–77.
Pohlmeier, W. and Ulrich, V. (1995) An econometric model of the two-part decision-making process in the demand for health care. *J. Hum. Resour.*, **30**, 339–361.
Pudney, S. (2007) Rarely pure and never simple: extracting the trust from self-reported data on substance use. *Working Paper 11/07*. Institute for Fiscal Studies and Institute for Social and Economic Research, London.
Pudney, S. (2010) Drugs policy: what should we do about cannabis? *Econ. Poly*, **25**, 165–211.
Ramful, P. and Zhao, X. (2009) Participation in marijuana, cocaine and heroin consumption in Australia: a multivariate Probit approach. *Appl. Econ.*, **41**, 481–496.
Saffer, H. and Chaloupka, F. (1998) Demographics differentials in the demand for alcohol and illicit drugs. In *Economic Analysis of Substance Use and Abuse: an Integration of Econometrics and Behavioral Economic Research* (eds F. Chaloupka, W. Bickel, H. Saffer and M. Grossman), pp. 187–211. Chicago: University of Chicago Press.
Saffer, H. and Chaloupka, F. (1999) The demand for illicit drugs. *Econ. Inq.*, **37**, 401–411.
Smith, M. (2003) On dependency in double-hurdle models. *Statist. Pap.*, **44**, 581–595.
Train, K. (2000) Halton sequences for mixed logit. Department of Economics, University of California at Berkeley, Berkeley.
Van Ours, J. C. and Williams, J. (2011) Cannabis use and mental health problems. *J. Appl. Econmetr.*, **26**, 1137–1156.
Visher, C. A. and McFadden, K. (1991) *A Comparison of Urinalysis Technologies for Drug Testing in Criminal Justice*. Darby: Diane.
Williams, J. and Bretteville-Jensen, A. L. (2014) Does liberalizing cannabis laws increase cannabis use? *J. Hlth Econ.*, **36**, 20–32.
Wish, E. (1987) *Drug Use Forecasting: New York 1984 to 1986*. Washington DC: National Institute of Justice.
Zhao, X. and Harris, M. (2004) Demand for marijuana, alcohol and tobacco: participation, frequency and cross-equation correlations. *Econ. Rec.*, **80**, 394–410.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Online Appendix'.