Modelling Illegal Drug Participation*

Sarah Brown
Department of Economics
University of Sheffield,
United Kingdom

Mark N. Harris
School of Economics and Finance
Curtin University, Australia

Preety Srivastava
School of Economics, Finance and Marketing
Royal Melbourne Institute of Technology, Australia

Xiaohui Zhang
Department of Economics
University of Exeter, United Kingdom

September 2016

Abstract

We contribute to the small, but important, literature exploring the incidence and implications of mis-reporting 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 mis-report or not. We explore mis-reporting in the context of the consumption of a system of drugs and specify a *multivariate inflated probit model*. Compared to observed participation rates of 12.2, 3.2 and 1.3% (marijuana, speed and cocaine, respectively) 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 mis-report their participation is a staggering 65% for a hard drug like cocaine, and still some 31% and 17%, for the softer drugs of marijuana and speed.

JEL Classification: C3, D1, I1

Keywords: Discrete data, illegal drug consumption, inflated responses, mis-reporting.

*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 (BCEC) for their generous funding. We also wish to thank Steve Pudney for helpful comments and suggestions. The usual caveats apply.

1 Introduction and Background

Over the past three decades, the increased availability of micro level data sets has enabled researchers to explore an extensive range of research themes at the individual and household level. Such micro level data is invariably collected using survey techniques with the result that the quality of the data 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 mis-report due to the sensitive nature of the questions. Individuals may have an incentive to under-report 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 (for example, Berg and Lien 2006, Pudney 2007).

Mis-reporting will result in inaccurate estimates of the prevalence of such behaviours, which may lead one to question the validity of empirical conclusions drawn from surveys. Moreover, any systematic mis-reporting will likely 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 mis-reporting in survey data.

Mis-classification, or mis-reporting, often leads to the presence of "excess" zeros, 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, Heilbron 1989, Lambert 1992, Greene 1994, Pohlmeier and Ulrich 1995, Mullahey 1997); and, more recently, to zero-inflated ordered probit (ZIOP) models (Harris and Zhao 2007). Typically, the issue that arises is that "zero" observations can result from two distinct processes and that ignoring this can lead to seriously mis-specified models.

In this paper, we explore the modelling of *sensitive* response variables: variables where there is an associated loss-function (perceived or actual) involved for the individual in terms of the responses he/she reports. Here, it is clear that the researcher must be aware of the potential for mis-reporting. For consumption of goods with associated reporting loss-functions, the approach suggested here allows for these zero observations to correspond

to not only non-participants, but importantly also to those participants who, fearing repercussions, erroneously report zero-consumption.

Our particular application lies in the important area of mis-reporting within the context of the consumption of illicit drugs. Given the considerable individual and social costs 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 related to the addictive nature of drugs as well as the relationship between the consumption of different types of drugs. 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 socioeconomic determinants of drug use (Duarte, Escario, and Molina 2005, Ramful and Zhao 2009) and panel surveys to estimate rational addiction models (Becker, Grossman, and Murphy 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), p.26, comments 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 mis-reporting (specifically with regard to drug consumption) in individual level survey data.

Our approach is similar to that of Hausman, Abrevaya, and Scott-Morton (1998) who use a logit model to estimate mis-classification probabilities. They consider a binary choice model with two types of mis-classification: 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, Abrevaya, and Scott-Morton (1998) to allow for covariates to influence the mis-reporting/mis-classification decision; this will be very important for policy-makers to help identify those individuals with greater propensities to do so. Secondly, we acknowledge that many "similar" response variables of interest (various illicit drugs our example) are likely to be consumed jointly (here due to their common addictive nature), so that we extend the simpler univariate approach to a multivariate one. Finally, we apply this

new model to the consumption of illegal drugs (in Australia) and thereby provide new evidence as to the likely extent of mis-reporting across these. We also provide evidence as to the true rates of participation across these drugs as compared to 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 The Econometric Framework

2.1 An Inflated Probit (IP) 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, due to 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 $r^* \leq 0$. 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 mis-report. Again this is related to a second unobserved variable m such that $m = 1$ for $m^* > 0$ and $m = 0$ for $m^* \leq 0$, where $m = 0$ represents a mis-reporter 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

$$y = r \times m. \tag{3}$$

As such the observed realisation 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 $\boldsymbol{\varepsilon}_i$ will be related across equations (a point 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 ones on the diagonal and ρ on the off diagonals) the relevant probabilities will have the form

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

where Φ_2 denotes the cumulative distribution function (*c.d.f.*) of the standardised bivariate normal distribution. The first term on the right-hand side of equation (4) for $\Pr\left(y = 0 \left| \mathbf{x} \right.\right)$ represents a genuine non-participant; the second term, a (participant) mis-reporter. The expression for $\Pr\left(y = 1 \left| \mathbf{x} \right.\right)$ thus represents a (participant) true reporter. So here the probability of a zero observation has been "inflated" as it is a combination of the probability of non-participation plus that from mis-reporting. This approach thus models "mis-reporting" explicitly and as a function of a set of explanatory variables unlike the model developed by Hausman, Abrevaya, and Scott-Morton (1998) where mis-reporting is accounted for using constant terms; or by Dustmann and Soest (2001) who decompose mis-classification errors in panel data into time-persistent and time-varying components and where the probability of mis-classification is independent of respondent characteristics. However, it is very unlikely that such mis-reporting rates will be constant and homogeneous across individuals. Moreover, ignoring this heterogeneity (if present) could well lead to biased estimates of quantities of interest (such as true participation rates). Due to the zero-inflation and the correlated disturbances, we term this a correlated inflated probit (IPC) model. A test of $\rho = 0$ is jointly a test for independence of the two error terms and also one of the IPC *versus* the nested inflated probit (IP) one.

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

choice)

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

2.2 Extending to a Multivariate Inflated Probit (MIP) System

Often social bads such as licit and illicit drugs are consumed in a consumption bundle (see, for example, Collins, Ellickson, and Bell 1998, 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, Ramful and Zhao 2009). Due 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 mis-reporting equations: that is, via the unobservables. As a consequence, vital cross-drug information is lost when the IPC model is estimated in a univariate framework for several drugs of interest. The multivariate approach essentially isolates the joint impacts of observable and unobservable personal characteristics on the participation and mis-reporting of all three drugs and estimates the strength of the intrinsic correlations, via the unobservables, across the three drugs which are commonly considered closely related economic goods.

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

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

and the propensity to mis-report will be:

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

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 will assume so, both in the empirical application and also below, to simplify notation (*i.e.*, the same covariate specification applies for all drugs). Note that 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 mis-reporting 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 IPC 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 mis-reporting equation relating to the first drug. This results in a range of joint probabilities of interest, such as the polar cases of

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

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 ones 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 *c.d.f.*.

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 (RHS) relate to participation and true-reporting in all of the three respective drugs. On the other hand, $Pr(y_1 = 0, y_2 = 0, y_3 = 0|\mathbf{x})$, or equation (10), has a more complex form. This probability corresponds to an *observed* zero in each of the three drugs. This can occur in eight distinct ways; the individual can be:

1. a true non-participant in each drug; line 1.

2. a mis-reporting 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 $\left(-\mathbf{x}'_{r_2}\boldsymbol{\beta}_{r_2}, -\mathbf{x}'_{r_3}\boldsymbol{\beta}_{r_3}\right)$; line 2.

$\vdots$

8. a mis-reporting participant in all drugs; line 8.

Note that Σ_j defines the relevant sub-matrices of Σ with appropriate signs in the correlations. For example, the relevant lower sub-matrix of Σ_4 in the second RHS 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 MIP model can be estimated by ML but as the probabilities entering this are functions of high dimensional multivariate normal distributions, these are simulated using the GHK 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 using numerical gradients, and standard errors of these obtained by the delta method.

3 An Application to Drug Consumption

Empirical studies play a crucial role in identifying the socioeconomic and demographic factors 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 mis- and under-report illicit drug use. The extent of such mis- and under-reporting is likely to be influenced by a variety of factors such as gender and ethnicity (see, for example, Mensch and Kandel 1988, Fendrich and Vaughn 1994).

Mis-reporting 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 under-reporting 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 to computer-assisted telephone interviews (CATI), Kraus and Augustin (2001) find that a lower number of respondents would admit alcohol consumption if questioned by telephone compared to self-reports from questionnaires. In a similar vein, Hoyt and Chaloupka (1994) and Fendrich and Vaughn (1994) find that lower reported drugs use is associated with telephone interviews. The increased use of computer assisted self-interviewing 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, Carey, and Tu 2006).

In addition, given the apparent complex interrelationships between the demand for different types of illicit drugs, it is apparent that the extent of mis-reporting may vary across different types of drugs, arguably being particularly serious in the case of harder drugs (such as heroin and cocaine). Pudney (2007) analyses the consequences of mis-reporting of illicit drugs use for statistical inference using UK panel data containing repeated questions on self-reported lifetime drug use. The findings indicate serious under-reporting of the use of marijuana and cocaine, which in turn leads to biases in statistical modelling. For example, for one of the datasets analysed, under-reporting rates for marijuana (cocaine) with bounds averaging from 23 to 60% (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 mis-reporting problems in the case of hard drugs (see, for example, MacDonald and Pudney 2003). For example, in an early contribution, Wish (1987) analyses a sample of men arrested in New York City in 1984. For cocaine, the interview data indicated a drug use rate of 43%

as compared to 82% elicited from urine specimens. More recently, Lu, Taylor, and Riley (2001) compare under-reporting of drugs by validating information obtained via interviews with urinalysis for a sample of adult arrestees. The findings indicate significant levels of under-reporting 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 generalised, is not apparent and is arguably limited given that such data are based on somewhat atypical circumstances and samples. The modelling strategy outlined in this study, in contrast, only requires 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 we use for the model are drawn from the Australian National Drug Strategy Household Survey (NDSHS), which is a nationally representative survey of the Australian civilian population aged over 14 (NDSHS 2010). The earlier waves of the NDSHS used face-to-face and drop-and-collect methods to collect data. The computer-assisted telephone interview (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, on the other hand, were administered using only drop and collect and CATI, while the more recent surveys have been conducted only using the CATI method. Note that our dataset 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?* Due 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. This data has been used in several previous studies (for example, Harris and Zhao 2007, Van Ours and Williams 2011, Williams and Bretteville-Jensen 2014).

In terms of explanatory variables, we require two sets: one to determine participation and the other mis-reporting. While 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 (for example, Gill and Michaels 1991, Saffer

and Chaloupka 1999, Cameron and Williams 2001), we include a wide range of personal and demographic characteristics (such as: gender; marital status; educational attainment; whether the individual resides in a state where possession of small amounts of marijuana is decriminalised; income and so on; see the Online Appendix and Table 1). Inclusion of year and state dummies in both equations allows for the fact that both participation and mis-reporting rates may follow different trends over time, whilst also allowing for any difference in drug prices and policies across states.

We include a range of identifying variables in the mis-reporting equation: variables that (ostensibly) affect the mis-reporting decision(s) but not the participation one(s). These identifying variables $\mathbf{x}_m$ (to capture mis-reporting) mostly relate to the conditions under which the survey was administered, which as mentioned above, may potentially influence the extent to which individuals mis-report but will arguably be independent from any participation propensities. Specifically, we control for: if anyone else was present when the respondent was completing the survey questionnaire; if anyone helped the respondent complete the survey questionnaire; and if the drop-and-collect survey mode was used. These variables conform with the factors that have been associated with mis-reporting or misclassification in prior studies (for example, Mensch and Kandel 1988, O'Muircheartaigh and Campanelli 1998, Lu, Taylor, and Riley 2001, Kraus and Augustin 2001, Berg and Lien 2006).

Finally, we also include as an instrument variable indicating a general lack of trust in the survey which we measure 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/her and the survey in general (for example, Corbin and Morse 2003). For each respondent it is possible to calculate the total number of compulsory (asked to everyone) questions 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 statistics relating to the variables used in our econometric analysis for the pooled cross-section data set. The variables are fully described in the Online 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 IPC ones. Although the differences in the estimated parameters from the more complex MIP model are somewhat 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 mis-reporting 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 (whilst the full set are 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, Bray, Zarkin, and Wendling 2001, Saffer and Chaloupka 1998), decriminalisation 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; while 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 mis-reporting equations, and noting that a positive coefficient indicates a lower probability of mis-reporting, we see that being male is associated with a higher probability of mis-reporting speed but with lower chances of mis-reporting marijuana

consumption. Age has no significant effect on the probability of mis-reporting marijuana but we find a quadratic effect of age on the probability of mis-reporting 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 mis-reporting marijuana and speed but education does not seem to affect reporting behaviour of cocaine. The introduction of decriminalisation is likely to be associated with increased awareness of the potential consequences 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 decriminalisation on the probability of mis-reporting marijuana or any of the other drugs. We find an increasing trend in mis-reporting 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 employed. With respect to the effects of the identifying set of variables in the mis-reporting equation, the presence of anyone else when the respondent was completing the questionnaire is associated with a higher probability of mis-reporting across all three drugs, consistent with Hoyt 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/face to face interview (relative to drop-and-collect), is associated with a higher probability of mis-reporting across all three drugs. Finally, if the respondent had a general lack of trust in the survey then they have a higher chance of mis-reporting drug use. In summary, since three out 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 one may be interested in predicting with the current model. For each drug in isolation, one may be interested in: the marginal probability of participation; the joint probability of participation and mis-reporting; or the probability of accurate reporting, conditional on participation. In Table 4 we present some summary probabilities 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, based on 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 mis-reporting 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 to the sample proportions. We find that for all three drugs the joint probabilities mimic the observed sample proportions very closely.

Conditional on an individual participating, there was a 65% chance of mis-reporting cocaine use compared with 31% for marijuana. That is, of the small percentage of cocaine users in the population (recorded at 1.3% and estimated at just under 5%) 65% claimed to not 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 associated with the former. For speed the estimated conditional probability of mis-reporting was 17%, significantly less than the other two drugs. This lower mis-reporting 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 mis-reporting 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 zeros, we also present in Table 4 the predicted probability of zero for each of the three drugs broken down into two respective components: non-participation and mis-reporting. 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 mis-reporting (11.2%). In view of the low rates of participation, the low mis-reporting 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 and above, they represent nearly 2,016,000, 1,004,000 and 629,000 cases of unreported cases of marijuana, speed and cocaine use, respectively. Such under-reporting can thus have extremely important implications for drug policies.

Such probabilities can be thought of as *prior* probabilities. That is, 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 mis-reporting, it is possible to estimate *posterior* probabilities, analogous 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 zeros come from non-participation and mis-reporting respectively, using all the information we have on the individual: this attempts to answer the question, *given that an individual recorded a zero, what is the probability that they are a true non-participant versus a mis-reporting participant (given their observed characteristics)?* The posterior probabilities for the two types of zeros are given as

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

and

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

From Table 4, we find that close to 87% of the reported zeros for marijuana are estimated to come from genuine non-participation (and therefore 13% from mis-reported participation). Note that as with the prior probabilities presented earlier, these posterior ones for mis-reporting might appear, superficially, rather low. However, it is important to remember that the probabilities for mis-reporting here are not marginal, but joint of participation *and* mis-reporting. Thus given participation probabilities are very low for all of these drugs (estimated at some 23, 8 and 5%, respectively, for marijuana, speed and cocaine, see second row of Table 4), it is not surprising that these joint probabilities are also small. Moreover, as with all of 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; the conditional probability of mis-reporting cocaine conditional on marijuana participation; and so on. Indeed, it is not immediately obvious how one would undertake such an exercise if these drug equations were estimated separately. We can also estimate partial effects on all of 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, conditional 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 mis-reporting effects serving to operate in the same direction. For instance, males are 5.3 percentage points (pp) less likely to abstain from all three drugs and they have a 0.6pp lower chance of accurately reporting such zero consumption. This results in an overall 5.9pp lower probability of recording zero consumption for males compared to 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 mis-reporting effects operating in the opposite direction thereby serving to moderate the participation effects.

Turning to education, degree holders have a 1.8pp lower chance of abstaining from all three drugs but a 2pp higher chance of accurately reporting such non-participation resulting in an overall 0.2pp lower probability of recording joint zero consumption across all three drugs relative to those with less than year 12 qualifications. However, the overall effect is statistically insignificant. In terms of the additional variables in the mis-reporting 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 a decline in abstention or in other words an increase in drug use over time. On the other hand, 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 zero for speed and cocaine, conditional 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 3pp 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, consistent with a gateway effect for males.

5.3 Robustness Checks, False Positives, and Validation Exercises

The instruments we use to identify the mis-reporting 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 mis-classification literature and their statistical significance in the estimated model lend further support to their inclusion in the mis-reporting 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 (presumably a result of the very low recorded participation rate). While we do not present the results from the various model specifications here due to space constraints, 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 while 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).

While most of the reporting bias is believed to be in the direction of under-reporting there is also some evidence from the literature on over-reporting. Such false positive rates

are generally lower than false negatives (see, for example, Visher and McFadden 1991, Harrison and Hughes 1997). However, as a "litmus test" to gauge the likely magnitudes of any false-positives we conduct a simple test reversing the 1's to 0's and re-estimating the model. For all three drugs the mis-reporting biases appear to be very small ranging from 0.81% for cocaine to 4.67% for marijuana. We also extended the basic framework to jointly allow for false positives and negatives. 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 mis-reporting with regard to false positives are very low, and therefore would not unduly affect the main results reported in the paper.

We also restricted the sample to individuals who have reported having ever used the drug (the NDSHS does not collect information on previous month's use): if the model is well-specified mis-reporting rates should be significantly lower as stigma rates will obviously be much reduced in this sub-sample. Indeed, we do find much less mis-reporting in these sub-samples (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 sub-sample of those who have ever used marijuana). Again, this validation exercise, gives us strong confidence in our main findings. Unfortunately the sub-samples of those who have ever used speed and cocaine are too small for robust analysis.

The final validation exercise 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 stigmatised given their general acceptability in the community. Thus, we would expect less reporting bias in the case of legal drugs. Based on IPC models, we still find strong effects of the identifying variables in the mis-reporting equations (results 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% 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 associated with the former. Table 7 also reports mis-reporting probabilities: conditional on an individual participating, we see that for alcohol and tobacco, there is a 10% and 26% chance of mis-reporting, compared to the much higher values 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 MIP Model

A key contribution of this paper is the use of a multivariate model which, as noted earlier, allows us to jointly estimate drug participation and mis-reporting 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 mis-reporting 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 mis-reporting is prevalent.

The MIP model is also preferred to the IPC model as it takes into account the likely cross-drug correlations. If estimated in isolation where correlations across participation and mis-reporting equations exist, but are ignored, estimated parameters and subsequent analysis are potentially biased and/or inefficiently estimated since they are based on mis-specified models and/or models where not all relevant information is being utilised. 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% while the IPC model would only estimate it at 0.005%. Clearly the MIP model that fully accounts for correlations within and across drugs exhibits better performance than the IPC.

Indeed, if it were the case that 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 (MC) 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 to that of a univariate one, *i.e.*, the IPC model

where the participation and mis-reporting equations are correlated for each drug but not across the drugs. Additionally we also consider the model suggested by Hausman, Abrevaya, and Scott-Morton (1998), HAS, which provides a good basis for comparison; here the participation equation is specified as in the IPC or MIP model but the mis-reporting probabilities are constants (and not a function of covariates). We thus consider the (relative) performance of the HAS model (for consistency, considered in a systems framework) to the IPC and MIP models.

For the data generating process (*dgp*), we consider three scenarios: 1) the true MIP model; 2) three independent IPC models; and 3) the HAS form. With the latter, the error terms in the participation and mis-reporting equations are allowed to be correlated. Therefore, the MIP is a generalised model for both the IPC and the HAS model, in terms of allowing decisions to be taken jointly and mis-reporting 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: 1) the true average probability over Q replications, $\bar{P}(\cdot)$; 2) the estimated average probability over Q replications, $\hat{\bar{P}}(\cdot)$; and 3) the root mean square error of the estimated probability, $RMSE_{P(\cdot)}$. To shed light on estimated parameters, we also report the averaged root mean square error over all estimated parameters, $AveRMSE_{para}$.

As expected, for the (true) MIP model all estimated probabilities are very close to the corresponding true ones and with very low RMSE's. In fact, almost all RMSE's from the MIP model are lower than those from the IPC and HAS models. Although the IPC model performs well in estimating marginal probabilities of a single drug consumption and probabilities of mis-reporting 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 mis-reporting (that is, the HAS model) appears to generally result in even greater discrepancies all with associated with high RMSE's. For details, refer to results presented in Table A2 in the

Online Appendix. Although it might not be strictly valid to generalise these findings universally, they do 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 mis-reporting/mis-classification propensities are a function of covariates and these are ignored in estimations. The results on averaged RMSE's over all estimated parameters reinforce the above results, *i.e.,* finite sample estimation bias from the MIP model is essentially zero, with that from the IPC model being significantly higher and that from the HAS model higher still.

When the true *dgp* is an independent IPC model for each of the three drugs, we see that again the estimated probabilities from the MIP model are very close to the true ones and with low RMSE's. We also find that the MIP does an excellent job in estimating the true non-zero correlations; and the average estimated correlation coefficients across drugs are all very close to zero (their true values) and with very low RMSE's. Moreover, the averaged RMSE over all parameters (β's and ρ's) is very small at 0.002. Detailed results can be found in Table A3 and Table A4 from the Online 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 HAS model (where an individual's decision to mis-report is not influenced by their characteristics but cross and within drug correlations exist), although being over-specified, the MIP model performs exceptionally well in terms of estimating the probabilities considered. For example, $P(m_{coc} = 0|r_{coc} = 1)$, whilst the true probability is 94.15%, this is estimated as (on average) 94.13% by the MIP model, with a 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.

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 jointly estimate participation and mis-reporting behaviours across a range of drugs. Some previous studies have relaxed such distributional assumptions, for example, Chen, Hu, and Lewbel (2009) and Feng and Hu (2013) but they have not considered mis-reporting/mis-classification on multiple events, *i.e.*, not allowing for correlations across events. In our application, we do 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 one, therefore we finally 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). We repeated this exercise assuming other forms of non-normality, such as other mixing distributions, multivariate $t-$distributions with very low degrees of freedom, and so on; 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 mis-reporting 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 mis-report their true situation, leading (here) to an excess amount of zero observations in the context of questions relating to activities such as illicit drug consumption: individuals are likely to deny their participation due to a variety of reasons, such as fear of being caught, stigma and moral concerns.

Overall, we find that mis-reporting 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 mis-report. Interestingly, our findings suggest that the extent of mis-reporting is influenced by how the survey was administered, 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 is collected serve to influence the accuracy of the information obtained. Our findings suggest that accounting for mis-reporting is important in the context of using survey data related to sensitive activities, especially where such data is used to inform public policy.

References

Becker, G. S., M. Grossman, and K. M. Murphy (1994): "An empirical analysis of cigarette addiction," Discussion paper, National Bureau of Economic Research, No. w3322.

Berg, N., and D. Lien (2006): "Same-sex sexual behaviour: US frequency estimates from survey data with Simultaneous Misreporting and Non-Response," *Applied Economics*, 38(7), 757–769.

Bhat, C. R. (2003): "Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences," *Transportation Research Part B: Methodological*, 37(9), 837–855.

Bowling, A. (2005): "Mode of questionnaire administration can have serious effects on data quality," *Journal of Public Health*, 27(3), 281–291.

Cameron, L., and J. Williams (2001): "Cannabis, alcohol and cigarettes: substitutes or compliments," *The Economic Record*, 77(236), 19–34.

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

Collins, R., P. Ellickson, and R. Bell (1998): "Simultaneous polydrug use among teens: prevalence and predictors," *Journal of Substance Abuse*, 10(3), 233–253.

Corbin, J., and J. Morse (2003): "The unstructured interactive interview: issues of reciprocity and risks when dealing with sensitive topics," *Qualitative Enquiry*, 9(3), 335–354.

Cragg, J. (1971): "Some statistical models for limited dependent variables with application to the demand for durable goods," *Econometrica*, 39(5), 829–44.

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

Duarte, R., J. Escario, and J. Molina (2005): "Participation and consumption of illegal drugs among adolescents," *International Advances in Economic Research*, 11(4), 399–415.

Dustmann, C., and A. Soest (2001): "Language fluency and earnings: estimation with misclassified language indicators," *Review of Economics and Statistics*, 83(4), 663–674.

Farrelly, M., J. Bray, G. Zarkin, and B. Wendling (2001): "The joint demand for cigarettes and marijuana: evidence from the National Household Surveys on Drug Abuse," *Journal of Health Economics*, 20(1), 51–68.

Fendrich, M., and C. Vaughn (1994): "Diminished lifetime substance use over time: an inquiry into differential Under-reporting," *Public Opinion Quarterly*, 58(1), 96–123.

Feng, S., and Y. Hu (2013): "Misclassification errors and the underestimation of the US unemployment rate," *American Economic Review*, 103(2), 1054–70.

Gill, A., and R. Michaels (1991): "The determinants of illegal drug use," *Contemporary Economic Policy*, 9(3), 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.

Greene, W. (2012): *Econometric Analysis 7e*. Prentice Hall, New Jersey, USA, seventh edn.

Harris, M., and X. Zhao (2007): "A zero-inflated ordered Probit model, with an application to modelling tobacco consumption," *Journal of Econometrics*, 141(2), 1073–1099.

Harrison, L., and A. Hughes (1997): "The validity of self-reported drug use: improving the accuracy of survey estimates," NIH Publication No. 97-4147, NIDA Research Monograph 167, Rockville, MD: National Institute on Drug Abuse.

Hausman, J., J. Abrevaya, and F. Scott-Morton (1998): "Misclassification of the dependent variable in a discrete-response setting," *Journal of Econometrics*, 87(2), 239–269.

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

Hoyt, G., and F. Chaloupka (1994): "Effect of survey conditions on self-reported substance use," *Contemporary Economic Policy*, 12(3), 109–121.

Ives, R., and P. Ghelani (2006): "Polydrug use (the use of drugs in combination): a brief review," *Drugs: Education, Prevention, and Policy*, 13(3), 225–232.

Keane, M. P. (1994): "A computationally practical simulation estimator for panel data," *Econometrica: Journal of the Econometric Society*, 62(1), 95–116.

Kraus, L., and R. Augustin (2001): "Measuring alcohol consumption and alcohol-related problems: comparison of responses from self-administered questionnaires and telephone interviews," *Addiction*, 96(3), 459–471.

Labeaga, J. M. (1999): "A double-hurdle rational addiction model with heterogeneity: estimating the demand for tobacco," *Journal of Econometrics*, 93(1), 49–72.

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

Lu, N., B. Taylor, and K. Riley (2001): "The validity of adult arrestee self-reports of crack cocaine use," *American Journal of Alcohol Abuse*, 27(3), 399–419.

MacDonald, Z., and S. Pudney (2000): "Analysing drug abuse with british crime survey data: modelling and questionnaire design issues," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(1), 95–117.

——— (2003): "The use of self-report and drugs tests in the measurement of illicit drug consumption," Discussion Papers in Economics, No. 03/3, Department of Economics, University of Leicester.

Mensch, B., and D. Kandel (1988): "Underreporting of substance use in a national longitudinal youth cohort," *Public Opinion Quarterly*, 52(1), 100–124.

Morrison-Beedy, D., M. Carey, and X. Tu (2006): "Accuracy of audio computer-assisted self-interviewing (ACASI) and self-administered questionnaires for the assessment of sexual behavior," *AIDS and Behavior*, 10(5), 541–552.

Mullahey, J. (1986): "Specification and testing of some modified count data models," *Journal of Econometrics*, 33(3), 341–365.

——— (1997): "Heterogeneity, excess zeros and the structure of count data models," *Journal of Applied Econometrics*, 12(3), 337–350.

NDSHS (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 P. Campanelli (1998): "The relative impact of interviewer effects and sample design effects on survey precision," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 161(1), 63–77.

Pohlmeier, W., and V. Ulrich (1995): "An econometric model of the two-part decision-making process in the demand for health care," *Journal of Human Resources*, 30(2), 339–361.

Pudney, S. (2007): "Rarely pure and never simple: extracting the trust from self-reported data on substance use," Cemmap Working Paper 11/07, Institute for Fiscal Studies and Institute for Social and Economic Research.

——— (2010): "Drugs policy: what should we do about cannabis?," *Economic Policy*, 25(61), 165–211.

Ramful, P., and X. Zhao (2009): "Participation in marijuana, cocaine and heroin consumption in Australia: a multivariate Probit approach," *Applied Economics*, 41(4), 481–496.

Saffer, H., and F. Chaloupka (1998): "Demographics differentials in the demand for alcohol and illicit drugs," in *Economic Analysis of Substance Use and Abuse: An Integration of Econometric and Behavioral*, pp. 187–211. Chaloupka F., Bickel, W., Saffer, H., Grossman, M. (Eds.), Economic Research, University of Chicago Press, Chicago.

——— (1999): "The demand for illicit drugs," *Economic Inquiry*, 37(3), 401–411.

Smith, M. (2003): "On dependency in double-hurdle models," *Statistical Papers*, 44(4), 581–595.

Train, K. (2000): "Halton sequences for mixed logit," Department of Economics, UCB.

Van Ours, J. C., and J. Williams (2011): "Cannabis use and mental health problems," *Journal of Applied Econometrics*, 26(7), 1137–1156.

Visher, C. A., and K. McFadden (1991): *A comparison of urinalysis technologies for drug testing in criminal justice*. DIANE Publishing.

Williams, J., and A. L. Bretteville-Jensen (2014): "Does liberalizing cannabis laws increase cannabis use?," *Journal of Health Economics*, 36, 20–32.

Wish, E. (1987): *Drug Use Forecasting: New York 1984 to 1986*. US Department of Justice, National Institute of Justice Washington, DC.

Zhao, X., and M. Harris (2004): "Demand for marijuana, alcohol and tobacco: participation, frequency and cross-equation correlations," *Economic Record*, 80(251), 394–410.

Table 1: Summary Statistics, Sample Size 56,579

Variable	Mean	Std Dev	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

Table 2: Marijuana, Speed and Cocaine Consumption: Estimated Coefficients[a]

	Marijuana		Speed		Cocaine	
	Participation	Mis-reporting	Participation	Mis-reporting	Participation	Mis-reporting
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)*

Table 2: Marijuana, Speed and Cocaine Consumption: Estimated Coefficients[a] (Cont'd)

	Marijuana		Speed		Cocaine	
	Participation	Mis-reporting	Participation	Mis-reporting	Participation	Mis-reporting
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)
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. *Significant at 10% level; **Significant at 5% level; ***Significant at 1% level.

[a] A positive coefficient for participation indicates an increase in participation probability while a negative coefficient for mis-reporting indicates an increase in mis-reporting probability.

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 10% level; **Significant at 5% level; ***Significant at 1% level.

Table 4: Sample and Predicted Probabilities

	Marijuana	Speed	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	0.1206	0.0281	0.0137
$[Pr(r = 1, m = 1\|\mathbf{x})]$	$(0.0013)^{***}$	$(0.0007)^{***}$	$(0.0007)^{***}$
Probability of Mis-reporting Conditional on Participation	0.3064	0.1702	0.6467
$[Pr(m = 0\|r = 1, \mathbf{x})]$	$(0.0683)^{***}$	$(0.0698)^{***}$	$(0.1236)^{***}$
Components of the zeros:			
Non-participation $[Pr(r = 0\|\mathbf{x})]$	0.7674	0.9162	0.9514
	$(0.0178)^{***}$	$(0.0112)^{***}$	$(0.0224)^{***}$
Mis-reporting $[Pr(r = 1, m = 0\|\mathbf{x})]$	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
	$(0.0204)^{***}$	$(0.0098)^{***}$	$(0.0204)^{***}$
Mis-reporting $[Pr(r = 1, m = 0\|\mathbf{x},y = 0)]$	0.1308	0.0491	0.0310
	$(0.0204)^{***}$	$(0.0098)^{***}$	(0.0204)

Standard errors are given in parentheses. *Significant at 10% level; **Significant at 5% level; ***Significant at 1% level.

Table 5: Partial Effects on Selected Joint and Conditional Probabilities[a]

	$Pr(y_{mar}=0, y_{spd}=0, y_{coc}=0\|\mathbf{x})$			$Pr(y_{spd}=0, y_{coc}=0\|r_{mar}=1, \mathbf{x})$		
	Participation	Mis-reporting	Overall	Participation	Mis-reporting	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)***

Table 5: Partial Effects on Selected Joint and Conditional Probabilities[a] (Cont'd)

	$Pr(y_{mar}=0, y_{spd}=0, y_{coc}=0\|\mathbf{x})$			$Pr(y_{spd}=0, y_{coc}=0\|r_{mar}=1, \mathbf{x})$		
	Participation	Mis-reporting	Overall	Participation	Mis-reporting	Overall
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)
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.*Significant at 10% level; **Significant at 5% level; ***Significant at 1% level.

[a] A positive marginal effect for participation represents an increase in participation probability while a negative marginal effect for mis-reporting represents an increase in mis-reporting probability.

Table 6: Comparison across Specifications

	Main Spec	Spec 1	Spec 2	Spec 3	Spec 4	Spec 5
				Marijuana		
- Sample Rate of Part'n	0.1215	0.1215	0.1215	0.1215	0.1215	0.1215
- Joint Probability of Part'n and Accurate Reporting $[Pr(r_{mar}=1, m_{mar}=1\|\mathbf{x})]$	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 Part'n	0.0316	0.0316	0.0316	0.0316	0.0316	0.0316
- Joint Probability of Part'n and Accurate Reporting $[Pr(r_{spd}=1, m_{spd}=1\|\mathbf{x})]$	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 Part'n	0.0127	0.0127	0.0127	0.0127	0.0127	0.0127
- Joint Probability of Part'n and Accurate Reporting $[Pr(r_{coc}=1, m_{coc}=1\|\mathbf{x})]$	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.*Significant at 10% level; **Significant at 5% level; ***Significant at 1% level. The five specifications are similar to the main one except for the following: spec 1- present only; spec 2 - help only; spec 3- survey type only; spec 4 - trust only; spec 5 - survey type and trust only

Table 7: Comparison across Legal and Illegal Drugs

	Alcohol	Tobacco	Marijuana	Speed	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 Mis-Reporting Conditional on Participation $[Pr(m=0\|r=1, \mathbf{x})]$	0.101	0.257	0.357	0.154	0.643

Probabilities are estimated from IPC models.

7 Online Appendix (not for publication)

7.1 Definition of variables

- **Stage**: standardised age (mean subtracted and scaled by sample standard deviation).
- **Stagesq**: age-squared, standardised (mean subtracted and scaled by sample standard deviation).
- **Male:** = 1 for male; and = 0 for female.
- **Married:** = 1 if married or *de facto*; and = 0 otherwise.
- **Preschool:** = 1 if the respondent has pre-school aged child/children, and = 0 otherwise.
- **Singpar:** 1 if respondent comes from a single parent household, and = 0 otherwise.
- **Capital**: = 1 if the respondent resides in a capital city, and = 0 otherwise.
- **ATSI:** = 1 if respondent is of Aboriginal or Torres Strait Islander origin, and = 0 otherwise.
- **Work:** = 1 if mainly employed; and = 0 otherwise.
- **Study:** = 1 if mainly study; and = 0 otherwise.
- **Unemp** = 1 if unemployed; and = 0 otherwise.
- **Other** = 1 if retired, home duty, or volunteer work; and = 0 otherwise. This variable is used as the base of comparison for work status dummies and is dropped in the estimation.
- **Degree**: = 1 if the highest qualification is a tertiary degree, and = 0 otherwise.
- **Diploma:** = 1 if the highest qualification is a non-tertiary diploma or trade certificate, and = 0 otherwise.
- **Yr12**: = 1 if the highest qualification is Year 12, and = 0 otherwise.

- **Less than Year 12**: = 1 if the highest qualification is below Year 12, and = 0 otherwise. This variable is used as the base of comparison for education dummies and is dropped in the estimation.

- **Lrpinc:** Logarithm of real personal annual income before tax measured in thousands of Australian dollars.

- **Decrim:** = 1 if respondent resides in a state where small possession is decriminalised and = 0 otherwise.

- **Migr10:** = 1 if migrated to Australia in the last 10 years, and = 0 otherwise.

- **Present**: = 1 if anyone else was present when the respondent was completing the survey questionnaire; and = 0 otherwise.

- **Help**: = 1 if anyone helped the respondent complete the survey questionnaire; and = 0 otherwise.

- **Survtype**: = 1 if the computer-assisted telephone interview (CATI) or face-to-face method was used to collect data; and = 0 if drop and collect method was used.

- **Trust**: percentage of compulsory questions left unanswered in the survey. This is based on questions that require a response from every single participant and other questions with conditional branching where a path is set to answer subsequent questions based on how a participant answers a given question; the trust variable is the non-response rate.

- **YR07**: = 1 in year 2007 and = 0 otherwise; **YR04**: = 1 in year 2004 and = 0 otherwise; **YR01**: = 1 in year 2001 and = 0 otherwise (used as the base of comparison for time effect and is dropped in the estimation).

- **VIC**:=1 if living in Victoria and = 0 otherwise; **QLD**:=1 if living in Queensland and = 0 otherwise; **WA**:=1 if living in Western Australia and = 0 otherwise; **SA**:=1 if living in South Australia and = 0 otherwise; **TAS**:=1 if living in Tasmania and = 0 otherwise; **ACT**:=1 if living in Australian Capital Territory and = 0 otherwise; **NT**:=1 if living in Northern Territory and = 0 otherwise; **NSW**:=1 if living in New South Wales and = 0 otherwise (used as the base of comparison for time effect and is dropped in the estimation).

7.2 Participation partial effects compared to Probit ones

Table A1: Selected Partial Effects on Participation - MIP versus Probit[a]

	Marijuana		Speed		Cocaine	
	MIP	Probit	MIP	Probit	MIP	Probit
MALE	0.079	0.050	0.011	0.011	0.007	0.005
	(0.009)***	(0.003)***	(0.003)***	(0.002)***	(0.007)	(0.001)***
STAGE	-0.195	-0.013	-0.066	-0.090	-0.040	-0.033
	(0.046)***	(0.011)	(0.023)***	(0.005)***	(0.074)	(0.004)***
STAGESQ	-0.042	-0.112	0.033	0.058	0.020	0.024
	(0.031)	(0.012)***	(0.019)*	(0.005)***	(0.071)	(0.004)***
MARRIED	-0.101	-0.069	-0.016	-0.024	-0.021	-0.012
	(0.011)***	(0.003)***	(0.003)***	(0.002)***	(0.022)	(0.001)***
PRESCHOOL	-0.008	-0.019	-0.010	-0.012	-0.009	-0.008
	(0.010)	(0.004)***	(0.003)***	(0.003)***	(0.010)	(0.002)***
SINGPAR	0.007	0.006	0.002	-0.001	-0.003	-0.007
	(0.009)	(0.005)	(0.002)	0.003	(0.006)	(0.002)***
CAPITAL	-0.013	0.002	0.004	0.009	0.011	0.010
	(0.006)***	(0.003)	(0.001)***	(0.002)***	(0.011)	(0.001)***
ATSI	0.001	0.021	0.001	-0.006	-0.021	-0.006
	(0.021)	(0.010)**	(0.004)	0.006	(0.027)	(0.005)
WORK	0.016	-0.030	-0.003	-0.020	0.001	-0.006
	(0.011)	(0.004)***	(0.002)	(0.003)***	(0.006)	(0.002)***
STUDY	0.100	-0.051	0.010	-0.034	0.019	-0.011
	(0.031)***	(0.006)***	(0.005)*	(0.004)***	(0.021)	(0.003)***
UNEMP	0.026	0.042	0.001	0.006	0.006	-0.002
	(0.015)*	(0.008)***	(0.003)	0.005	(0.010)	(0.004)
DEGREE	0.035	-0.012	-0.011	-0.019	0.000	0.000
	(0.012)***	(0.004)***	(0.003)***	(0.003)***	(0.005)	(0.002)
YR12	0.007	0.000	-0.006	-0.003	0.001	0.000
	(0.009)	(0.004)	(0.002)***	0.002	(0.005)	(0.002)
DIPLOMA	0.013	0.006	-0.004	-0.001	-0.004	0.000
	(0.007)*	(0.004)*	(0.002)***	0.002	(0.007)	(0.002)
LRPINC	-0.031	0.009	0.001	0.012	0.001	0.005
	(0.007)***	(0.002)***	(0.001)	(0.001)***	(0.002)	(0.001)***
DECRIM	-0.049	-0.013	0.001	0.000	-0.008	0.003
	(0.017)***	(0.008)	(0.003)	0.004	(0.012)	(0.003)
MIGR10	0.016	-0.030	0.003	-0.006	0.007	0.006
	(0.019)	(0.006)***	(0.004)	(0.004)*	(0.010)	(0.002)***

Standard errors are given in parentheses.*Significant at 10% level; **Significant at 5% level; ***Significant at 1% level.

[a] A positive marginal effect represents an increase in participation probability.

Table A2: Monte Carlo Results: MIP as true DGP

		MIP		IPC		HAS	
	$\bar{P}$	$\hat{\bar{P}}$	RMSE	$\hat{\bar{P}}$	RMSE	$\hat{\bar{P}}$	RMSE
$P(r_{mar}=1)$	0.6028	0.5993	0.0466	0.5986	0.0547	0.0000	0.6028
$P(r_{spd}=1)$	0.0667	0.0668	0.0060	0.0669	0.0071	0.6137	0.5508
$P(r_{coc}=1)$	0.0223	0.0224	0.0022	0.0283	0.0294	0.0229	0.0013
$P(m_{mar}=0\|r_{mar}=1)$	0.8541	0.8454	0.0422	0.8541	0.0165	1.000	0.1459
$P(m_{spd}=0\|r_{spd}=1)$	0.7638	0.7641	0.0363	0.7646	0.0402	0.9978	0.2342
$P(m_{coc}=0\|r_{coc}=1)$	0.9766	0.9763	0.0051	0.9742	0.0448	0.9896	0.0147
$P(r_{mar}=1,r_{spd}=1)$	0.0479	0.0479	0.0087	0.0415	0.0087	0.0000	0.0480
$P(r_{mar}=1,r_{coc}=1)$	0.0184	0.0184	0.0023	0.0217	0.0222	0.0000	0.0184
$P(r_{spd}=1,r_{coc}=1)$	0.0051	0.0051	0.0018	0.0026	0.0029	0.0217	0.0167
$P(r_{mar}=1,r_{spd}=1,r_{coc}=1)$	0.0048	0.0047	0.0016	0.0021	0.0030	0.0000	0.0048
$AveRMSE_{para}$			0.0020		0.5187		0.6462

Sample is restricted to one year of data, $N = 16,334$, and $Q = 500$.

7.3 Monte Carlo results

Table A3: Monte Carlo Results: IPC as true DGP

	$\bar{P}$	$\hat{\bar{P}}$	RMSE
$P(r_{mar}=1)$	0.6028	0.5967	0.0752
$P(r_{spd}=1)$	0.0668	0.0723	0.0310
$P(r_{coc}=1)$	0.0223	0.0226	0.0024
$P(m_{mar}=0\|r_{mar}=1)$	0.8539	0.8409	0.0503
$P(m_{spd}=0\|r_{spd}=1)$	0.7625	0.7571	0.0406
$P(m_{coc}=0\|r_{coc}=1)$	0.9758	0.9762	0.0058
$P(r_{mar}=1,r_{spd}=1)$	0.0416	0.0456	0.0229
$P(r_{mar}=1,r_{coc}=1)$	0.0169	0.0169	0.0025
$P(r_{spd}=1,r_{coc}=1)$	0.0024	0.0026	0.0013
$P(r_{mar}=1,r_{spd}=1,r_{coc}=1)$	0.0019	0.0020	0.0010
$AveRMSE_{para}$			0.0020

Sample is restricted to one year of data, $N = 16,334$, and $Q = 500$.

Table A4: True and Estimated Correlation Coefficients with IPC as true DGP

	True Correlation Coefficients							Estimated Correlation Coefficients				
	M_{mar}	R_{mar}	M_{spd}	R_{spd}	M_{coc}	R_{coc}		M_{mar}	R_{mar}	M_{spd}	R_{spd}	M
M_{mar}	-						$\overline{\rho}$	-				
							RMSE	-				
R_{mar}	0.036	-					$\overline{\rho}$	0.0363	-			
							RMSE	0.0005	-			
M_{spd}			-				$\overline{\rho}$	-3.93E-05	-1.16E-05	-		
							RMSE	4.41E-04	4.08E-04	-		
R_{spd}			0.133	-			$\overline{\rho}$	-2.20E-05	-7.25E-06	0.1329	-	
							RMSE	8.32E-04	6.35E-04	0.0011	-	
M_{coc}					-		$\overline{\rho}$	-2.02E-06	-2.42E-06	5.13E-06	4.79E-06	
							RMSE	1.82E-04	1.19E-04	6.24E-05	1.49E-04	
R_{coc}					0.311	-	$\overline{\rho}$	2.31E-06	2.19E-06	5.43E-06	-1.83E-06	0.3
							RMSE	1.92E-04	1.36E-04	1.08E-04	1.30E-04	0.0

Sample is restricted to one year of data, $N = 16,334$, and $Q = 500$.

Table A5: Monte Carlo Results: HAS as true DGP

	$\bar{P}$	$\hat{\bar{P}}$	RMSE
$P(r_{mar} = 1)$	0.6028	0.6038	0.0337
$P(r_{spd} = 1)$	0.0668	0.0683	0.0173
$P(r_{coc} = 1)$	0.0223	0.0228	0.0032
$P(m_{mar} = 0 \mid r_{mar} = 1)$	0.8393	0.8145	0.1077
$P(m_{spd} = 0 \mid r_{spd} = 1)$	0.8571	0.8554	0.0175
$P(m_{coc} = 0 \mid r_{coc} = 1)$	0.9415	0.9413	0.0056
$P(r_{mar} = 1, r_{spd} = 1)$	0.0480	0.0492	0.0130
$P(r_{mar} = 1, r_{coc} = 1)$	0.0184	0.0188	0.0029
$P(r_{spd} = 1, r_{coc} = 1)$	0.0051	0.0052	0.0014
$P(r_{mar} = 1, r_{spd} = 1, r_{coc} = 1)$	0.0048	0.0049	0.0013
$AveRMSE_{para}$			0.0016

Sample is restricted to one year of data, $N = 16,334$,and $Q = 500$.

Table A6: Monte Carlo Results: MIP with Mixture Errors as true DGP

	$\bar{P}$	$\hat{\bar{P}}$	RMSE
$P(r_{mar} = 1)$	0.6027	0.5147	0.1172
$P(r_{spd} = 1)$	0.0668	0.0573	0.0281
$P(r_{coc} = 1)$	0.0223	0.0226	0.0028
$P(m_{mar} = 0 \mid r_{mar} = 1)$	0.8539	0.9063	0.0640
$P(m_{spd} = 0 \mid r_{spd} = 1)$	0.7632	0.7819	0.0502
$P(m_{coc} = 0 \mid r_{coc} = 1)$	0.9765	0.9762	0.0084
$P(r_{mar} = 1, r_{spd} = 1)$	0.0480	0.0349	0.0218
$P(r_{mar} = 1, r_{coc} = 1)$	0.0184	0.0169	0.0031
$P(r_{spd} = 1, r_{coc} = 1)$	0.0051	0.0044	0.0021
$P(r_{mar} = 1, r_{spd} = 1, r_{coc} = 1)$	0.0048	0.0037	0.0020
$AveRMSE_{para}$			0.0025

Sample is restricted to one year of data, $N = 16,334$, and $Q = 500$.