Modelling Illegal Drug Participation*

Sarah Brown*, Mark N. Harris**,†and Preety Srivastava**
*Department of Economics
University of Sheffield
Sheffield
United Kingdom
**Department of Econometrics and Business Statistics
Monash University
Australia

February 2011

Abstract

We contribute to the small literature exploring the incidence and implications of mis-reporting in survey data. Specifically, when modeling "economic bads", such as illegal drug consumption, researchers are often faced with exceptionally low reported participation rates. Building on the recent literature on hurdle and double-hurdle models, we propose a zero-inflated modeling framework, where, firstly, an individual decides whether to participate or not and, secondly, for participants, there is a decision to mis-report or not. Furthermore, given that we explore mis-reporting in the context of the consumption of three illicit drugs, we specify a *multivariate inflated probit model*. We find that mis-reporting has a significant effect on drug participation rates. Across all three drugs, the predicted marginal probabilities of participation are substantially higher that the sample rate of participation as indicated by the survey responses.

JEL Classification: C3, D1, I1

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

*We wish to thank Steve Pudney for helpful comments and suggestions. The usual caveats apply.
†Corresponding author: Department of Econometrics and Business Statistics, Building 11, Wellington Road, Clayton Campus, Monash University, Victoria 3800, Australia; tel +61 39905 9414; fax +61 39905 5474. *Email address:* mark.harris@monash.edu

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 has generally been gathered using survey techniques with the accuracy of the data gathered from such surveys, therefore, being dependent 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 the true situation due to the sensitive nature of the questions. For example, individuals may have an incentive to underreport activities which are regarded as socially undesirable or which are associated with perceived social stigma such as smoking, alcohol or illicit drugs consumption or sexual behaviour (see, for example, Berg and Lien 2006). In a similar vein, self-reported versus the true incidence of cheating behaviour has attracted some interest in the economics literature. For example, Caudill et al. (2005) study undergraduate cheating behaviour by applying a logit model applicable in the case of a misclassified dependent variable, as developed by Hausman, Abrevaya, and Scott-Morton (1998), who show that ignoring such misclassification can result in biased and inconsistent estimates. Their findings indicate that the incidence of cheating estimated at 70% is considerably higher than the self-reported incidence of 51%. Thus, it is apparent that mis-reporting is potentially prevalent in a wide range of areas of economics. Moreover, such mis-reporting leads to inaccurate estimates of the prevalence of such behaviours, which may lead one to question the validity of empirical conclusions drawn from such surveys, with mis-reporting leading to biased inference in econometric analysis as well as potentially to inappropriate decision-making by policy-makers. Despite this, however, there is a shortage of research exploring the incidence and implications of such mis-reporting in survey data.

Mis-reporting can lead to the presence of "excess" zeros in empirical economics, 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 arise from two distinct processes and that ignoring this can lead to seriously mis-specified models. For example, Harris and Zhao (2007) considered ordered amounts of tobacco consumption, and argue that the presence of zeros arises from both non-participants and infrequent consumers. They showed that ignoring this inherent decomposition can lead to quite significant biases.

In this paper, we explore the modelling of "sensitive" response variables: that is, variables where there is an associated loss-function (either 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. Indeed, in research in areas of discrete random variables that are inherently ordered, such miss-reporting has been approached by allowing the model's inherent boundary variables to vary by observed personal characteristics (see, for example, Maassen van den Brink and Groot 1999, Kristensen and Johansson 2008). However, here we suggest a more fundamental form of modelling the mis-reporting which is likely to be present in data which is perceived to embody a strong loss-function (social and/or legal) for the individual. We suggest that the likely build-up of "zero" observations will correspond to both non-participants and participants. However, for these "goods" with associated reporting loss-functions, an additional source will be those participants who, fearing repercussions, report zero-consumption when in fact, this is not so. To be specific, we suggest a two-tiered sequencing of decision making. First, the individual makes a decision whether to participate or not; secondly, for participants, there is the decision to mis-report or not. The second stage allows for our contention, that, especially regarding participation of "economic bads" (licit, and in particular, illicit drugs for example), participants may intentionally mis-report their true consumption patterns. So within our econometric framework the probability of a zero observation is "inflated" as it is a combination of the probability of "non-participation" from the split probit model plus that from mis-reporting. In particular, we hypothesize that a potentially significantly large proportion of participants may actually report themselves as being non-participants, due to both moral and legal concerns about participation.

Our particular application lies in mis-reporting in 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 prescriptions relating to drugs abuse and, furthermore, analysis of survey data relating to drug use could potentially contribute to the policy debate in this area. It is apparent, therefore, that it is important to understand the shortcomings of such data, such as the possibility of under-reporting, which may mask the true extent of the problem, 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 literature exploring the implications of mis-reporting in individual level survey data.

2 The Econometric Framework

2.1 An Inflated Probit Model (IP)

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 this single latent variable to the observed outcome $y = 1$ via an index function being strictly non-negative, and one would then model participation rates. However, it is our contention here, that, especially regarding participation of "economic bads" (licit, and in particular, illicit drugs for example), participants may intentionally mis-report their true consumption patterns. In particular, 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.

Explicitly, we suggest a two-tiered sequencing of decision making. First, the individual makes a decision whether to participate or not; secondly, for participants, there is the decision to mis-report or not. Let r^* denote a binary variable indicating the split between Regime 0 (with $r = 0$ for non-participants) and Regime 1 (with $r = 1$ for participants). Although unobservable, r is related to a latent variable r^* via the mapping: $r = 1$ for $r^* > 0$ and $r = 0$ for $r^* \leq 0$. 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 (participant) mis-reporter and $m = 1$ a (participant) true-reporter. Again, we can write this as a linear latent form as

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

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

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

Under the assumption of independence, and that the stochastic terms $\boldsymbol{\varepsilon}$ $(\varepsilon_r, \varepsilon_m)$ are independent and follow standard Gaussian distributions, the full probabilities for $y = 0$ are given by

$$\Pr(y = 0 \,|\mathbf{x}) = \Pr(r = 0|\mathbf{x}) + \Pr(r = 1|\mathbf{x}) \Pr(m = 0|\mathbf{x}, r = 1) \tag{4}$$

and for $y = 1$ are

$$\Pr(y = 1 \,|\mathbf{x}) = \Pr(r = 1|\mathbf{x}) \Pr(m = 1|\mathbf{x}, r = 1) \tag{5}$$

These expressions can be stated simply in terms of joint probabilities by writing conditional probabilities as joint over marginals. The marginals in the denominator of these then cancel with the same when are entered into equations (4) and (5). Moreover, by independence these joint probabilities are simply products of the marginals such that, under the usual assumption of normality, they are given, respectively, by

$$\Pr(y = 0 \,|\mathbf{x}) = [1 - \Phi(\mathbf{x}_r' \boldsymbol{\beta}_r)] + \Phi(\mathbf{x}_r' \boldsymbol{\beta}_r) [1 - \Phi(\mathbf{x}_m' \boldsymbol{\beta}_m)]$$

and

$$\Pr(y = 1 \,|\mathbf{x}) = \Phi(\mathbf{x}_r' \boldsymbol{\beta}_r) \Phi(\mathbf{x}_m' \boldsymbol{\beta}_m) \tag{6}$$

So here the probability of a zero observation has been "inflated" as it is a combination of the probability of "non-participation" from the split probit model plus that from mis-reporting.

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 $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\mu}')'$ can be consistently and efficiently estimated using maximum likelihood techniques; the log-likelihood function is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=0}^{J} h_{ij} \ln \left[\Pr\left(y_i = j \left| \mathbf{x}, \boldsymbol{\theta} \right. \right) \right], \tag{7}$$

where the indicator function h_{ij} is

$$h_{ij} = \begin{cases} 1 \text{ if individual } i \text{ chooses outcome } j \\ 0 \text{ otherwise.} \end{cases} \quad (i = 1, ..., N;\; j = 0, 1, ..., J). \tag{8}$$

2.2 Generalising the Model to Correlated Disturbances (IPC)

As described above, the observed realisation of the random variable y can be viewed as the result of two separate latent equations, equations (1) and (2), with uncorrelated error terms. 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. So, we can now extend the model to have $(\varepsilon_r, \varepsilon_m)$ follow a bivariate normal distribution with covariance matrix Ω, whilst maintaining the identifying assumption of unit variances. Thus Ω will have the form

$$\Omega = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{9}$$

and the relevant probabilities will have the form

$$\Pr(y) = \begin{cases} \Pr\left(y = 0 \left| \mathbf{x} \right.\right) = \left[1 - \Phi\left(\mathbf{x}_r' \boldsymbol{\beta}_r\right)\right] + \Phi_2\left(\mathbf{x}_r' \boldsymbol{\beta}_r, -\mathbf{x}_m' \boldsymbol{\beta}_m; \Omega\right) \\ \Pr\left(y = 1 \left| \mathbf{x} \right.\right) = \Phi_2\left(\mathbf{x}_r' \boldsymbol{\beta}_r, \mathbf{x}_m' \boldsymbol{\beta}_m; \Omega\right) \end{cases} \tag{10}$$

where $\Phi_2(.)$ denotes the *c.d.f.* of the standardised bivariate normal distribution. ML estimation would again involve maximisation of Equation (7) replacing the probabilities of (6) with those of (10) and re-defining $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = (\boldsymbol{\beta}', \rho)'$. A test of $\rho = 0$ is a joint test for independence of the two error terms and thus a test of the more general model given by Equation (10) against the null of a simpler nested model of Equation (6).

3 Extending to a Multivariate System

Often "economic bads" such as licit and illicit drugs are consumed in a consumption bundle given that they are habit-forming. Instead of modelling the consumption of such "economic bads" in isolation, it can be extended to a multivariate framework where participation decisions are considered to be taken jointly by the same individual (see, for example, Harris and Zhao 2004, Ramful and Zhao 2009). The IP approach described in Section 2.1 ignores the potential cross-product correlations across multiple commodities for the same individual. Due to unobservable characteristics such as individual tastes, addictive traits and risk-taking attitudes, an individual's decision to consume multiple drugs can be potentially related through the error terms of the participation equations, that is, via the unobservables. As a consequence, vital cross-drug information is lost when the IP model is estimated in a univariate framework.

For a set of k multivariate Inflated Probit models, the propensity for participation will be:

$$r_k^* = \mathbf{x}_{rk}'\boldsymbol{\beta}_{rk} + \varepsilon_{rk}, \tag{11}$$

and the propensity to mis-report will be:

$$m_k^* = \mathbf{x}_{mk}'\boldsymbol{\beta}_{mk} + \varepsilon_{mk}.;;(k = 1, ..., K) \tag{12}$$

There is no restriction that $\mathbf{x}_{rk} = \mathbf{x}_{rh}$ or $\mathbf{x}_{mk} = \mathbf{x}_{mh}, \forall k \neq h$, but we will assume so both in the empirical application and also below to simplify notation. 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 $2K$ latent equations where the error terms jointly follow a multivariate normal distribution of order $s = 2K$ with covariance matrix Σ given by

$$\Sigma_s = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1s} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2s} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{s1} & \rho_{s2} & \rho_{s3} & \cdots & 1 \end{pmatrix}$$

Consider a system of Inflated models for three illicit drugs. Since $K = 3$, we have six latent equations with a variance covariance matrix defined as

$$\Sigma_6 = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & \rho_{15} & \rho_{16} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} & \rho_{25} & \rho_{26} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} & \rho_{35} & \rho_{36} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 & \rho_{45} & \rho_{46} \\ \rho_{51} & \rho_{52} & \rho_{53} & \rho_{54} & 1 & \rho_{56} \\ \rho_{61} & \rho_{62} & \rho_{63} & \rho_{64} & \rho_{65} & 1 \end{pmatrix}$$

where, for example, ρ_{12} relates to the correlation between ε_{r1} and ε_{m1} i.e. the respective error terms from the participation equation and the mis-reporting equation relating to the first drug; ρ_{13} is the correlation between ε_{r1} and ε_{r2}, i.e. the respective error terms from the participation equation for the first drug and the participation equation for the second drug; and ρ_{14} , the correlation between ε_{r1} and ε_{m2}, i.e. the respective error terms from the participation equation for the first drug and the mis-reporting equation for the second drug, etc. This results into a range of joint probabilities of interest such as the polar case of

$$\begin{aligned} Pr(y_1 = 1, y_2 = 1, y_3 = 1|\mathbf{x}_{r1}, \mathbf{x}_{r2}, \mathbf{x}_{r3}) = \\ \Phi_6(\mathbf{x}'_{r1}\boldsymbol{\beta}_{r1}, \mathbf{x}'_{m1}\boldsymbol{\beta}_{m1}, \mathbf{x}'_{r2}\boldsymbol{\beta}_{r2}, \mathbf{x}'_{m2}\boldsymbol{\beta}_{m2}, \mathbf{x}'_{r3}\boldsymbol{\beta}_{r3}, \mathbf{x}'_{m3}\boldsymbol{\beta}_{m3}; \Sigma_3) \end{aligned} \tag{13}$$

and

$$\begin{aligned} & Pr(y_1 = 0, y_2 = 0, y_3 = 0|\mathbf{x}) = \\ & \Phi_3(-\mathbf{x}'_{r1}\boldsymbol{\beta}_{r1}, -\mathbf{x}'_{r2}\boldsymbol{\beta}_{r2}, -\mathbf{x}'_{r3}\boldsymbol{\beta}_{r3}; \Sigma_3) \\ & + \Phi_4(\mathbf{x}'_{r1}\boldsymbol{\beta}_{r1}, -\mathbf{x}'_{m1}\boldsymbol{\beta}_{m1}, -\mathbf{x}'_{r2}\boldsymbol{\beta}_{r2}, -\mathbf{x}'_{r3}\boldsymbol{\beta}_{r3}; \Sigma_4) \\ & + \Phi_4(-\mathbf{x}'_{r1}\boldsymbol{\beta}_{r1}, -\mathbf{x}'_{r2}\boldsymbol{\beta}_{r2}, -\mathbf{x}'_{m2}\boldsymbol{\beta}_{m2}, -\mathbf{x}'_{r3}\boldsymbol{\beta}_{r3}; \Sigma_4) \\ & + \Phi_4(-\mathbf{x}'_{r1}\boldsymbol{\beta}_{r1}, -\mathbf{x}'_{r2}\boldsymbol{\beta}_{r2}, -\mathbf{x}'_{r3}\boldsymbol{\beta}_{r3}, -\mathbf{x}'_{m3}\boldsymbol{\beta}_{m3}; \Sigma_4) \\ & + \Phi_5(\mathbf{x}'_{r1}\boldsymbol{\beta}_{r1}, -\mathbf{x}'_{m1}\boldsymbol{\beta}_{m1}, \mathbf{x}'_{r2}\boldsymbol{\beta}_{r2}, -\mathbf{x}'_{m2}\boldsymbol{\beta}_{m2}, -\mathbf{x}'_{r3}\boldsymbol{\beta}_{r3}; \Sigma_5) \\ & + \Phi_5(\mathbf{x}'_{r1}\boldsymbol{\beta}_{r1}, -\mathbf{x}'_{m1}\boldsymbol{\beta}_{m1}, -\mathbf{x}'_{r2}\boldsymbol{\beta}_{r2}, \mathbf{x}'_{r3}\boldsymbol{\beta}_{r3}, -\mathbf{x}'_{m3}\boldsymbol{\beta}_{m3}; \Sigma_5) \\ & + \Phi_5(-\mathbf{x}'_{r1}\boldsymbol{\beta}_{r1}, -\mathbf{x}'_{r2}\boldsymbol{\beta}_{r2}, -\mathbf{x}'_{m2}\boldsymbol{\beta}_{m2}, \mathbf{x}'_{r3}\boldsymbol{\beta}_{r3}, -\mathbf{x}'_{m3}\boldsymbol{\beta}_{m3}; \Sigma_5) \\ & + \Phi_6(\mathbf{x}'_{r1}\boldsymbol{\beta}_{r1}, -\mathbf{x}'_{m1}\boldsymbol{\beta}_{m1}, -\mathbf{x}'_{r2}\boldsymbol{\beta}_{r2}, -\mathbf{x}'_{m2}\boldsymbol{\beta}_{m2}, \mathbf{x}'_{r3}\boldsymbol{\beta}_{r3}, -\mathbf{x}'_{m3}\boldsymbol{\beta}_{m3}; \Sigma_6) \end{aligned} \tag{14}$$

and also intermediate ones such as

$$
\begin{aligned}
Pr(y_1=1,y_2=0,y_3=0|\mathbf{x}) \quad = & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (15)\\
& \Phi_4(\mathbf{x}_{r1}'\boldsymbol{\beta}_{r1},\mathbf{x}_{m1}'\boldsymbol{\beta}_{m1},-\mathbf{x}_{r2}'\boldsymbol{\beta}_{r2},-\mathbf{x}_{r3}'\boldsymbol{\beta}_{r3};\Sigma_4)\\
& +\Phi_5(\mathbf{x}_{r1}'\boldsymbol{\beta}_{r1},\mathbf{x}_{m1}'\boldsymbol{\beta}_{m1},\mathbf{x}_{r2}'\boldsymbol{\beta}_{r2},-\mathbf{x}_{m2}'\boldsymbol{\beta}_{m2},-\mathbf{x}_{r3}'\boldsymbol{\beta}_{r3};\Sigma_5)\\
& +\Phi_5(\mathbf{x}_{r1}'\boldsymbol{\beta}_{r1},\mathbf{x}_{m1}'\boldsymbol{\beta}_{m1},-\mathbf{x}_{r2}'\boldsymbol{\beta}_{r2},\mathbf{x}_{r3}'\boldsymbol{\beta}_{r3},-\mathbf{x}_{m3}'\boldsymbol{\beta}_{m3};\Sigma_5)\\
& +\Phi_6(\mathbf{x}_{r1}'\boldsymbol{\beta}_{r1},\mathbf{x}_{m1}'\boldsymbol{\beta}_{m1},\mathbf{x}_{r2}'\boldsymbol{\beta}_{r2},-\mathbf{x}_{m2}'\boldsymbol{\beta}_{m2},\mathbf{x}_{r3}'\boldsymbol{\beta}_{r3},-\mathbf{x}_{m3}'\boldsymbol{\beta}_{m3};\Sigma_6)
\end{aligned}
$$

where the six elements in parentheses on the RHS of equation (13) relate to participation and true-reporting in the three respective drugs. $Pr(y_1=0,y_2=0,y_3=0|\mathbf{x})$ has a more complex form with the three elements in the first RHS term of equation (14) representing non-participation in all three drugs. The first two elements in the second term relate to participation ($\mathbf{x}_{r1}'\boldsymbol{\beta}_{r1}$) but mis-reporting ($-\mathbf{x}_{m1}'\boldsymbol{\beta}_{m1}$) of the first drug, and the remaining two elements relating to non-participation in the other two drugs. The last term of equation (14) with dimension six relates to participation but mis-reporting of the three respective drugs. On the other hand, the six elements in the last term on the RHS of equation (15) represent participation ($\mathbf{x}_{r1}'\boldsymbol{\beta}_{r1}$) and true-reporting ($\mathbf{x}_{m1}'\boldsymbol{\beta}_{m1}$) of the first drug, and participation but mis-reporting of the second and third drugs. Σ_j defines the relevant submatrices of Σ with appropriate signs in the correlations. For example, Σ_3 in the second RHS term of $Pr(y_1=0,y_2=0,y_3=0|\mathbf{x}_{r1},\mathbf{x}_{r2},\mathbf{x}_{r3},\mathbf{x}_{m1},\mathbf{x}_{m2},\mathbf{x}_{m3})$ is defined as

$$
\Sigma_4=\begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{15}\\ \rho_{21} & 1 & \rho_{23} & \rho_{25}\\ \rho_{31} & \rho_{32} & 1 & \rho_{35}\\ \rho_{51} & \rho_{52} & \rho_{53} & 1 \end{pmatrix}
$$

Given an $i.i.d.$ sample of N individuals making an observed choice of $(0,1)$ across all three drugs with associated probability P_i^* of the form defined above, the log-likelihood function will simply be

$$
\log L=\sum_{i=1}^{N}\log P_i^* \tag{16}
$$

The multivariate Inflated Probit model can then be estimated by maximising this log-likelihood function. Because the probabilities that enter the likelihood are functions of high dimensional multivariate normal distributions, these are simulated using the GHK algorithm. We follow the recent literature and use Halton sequences to generate the uniform variates required to evaluate the GHK probability simulator. In addition, since the joint and conditional probabilities are highly non-linear functions of $\mathbf{x}$, analytical solutions of marginal effects are difficult to obtain. Thus, the marginal effects are calculated using numerical gradients. As is standard in the literature, the standard errors of the marginal effects are then estimated using the delta method (using the estimated Hessian) which provides an approximation to the asymptotic distributions of the marginal effects.

4 An Application to Drug Consumption

As mentioned in Section 1, 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, which is not surprising given the considerable individual and social costs associated with the consumption of illegal drugs. Consequently, large amounts of public funds are spent worldwide on educational programs and promotional campaigns to reduce the consumption of drugs. Empirical studies play a crucial role in helping to identify 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. For example, in one of the early contributions Sickles and Taubman (1991) use data drawn from the US National Longitudinal Survey of Youth (NLSY) to explore the question as to who uses illegal drugs. The findings, which are based on a logistic model, suggest that socio-economic variables such as religious preference have a statistically significant influence on drug use. Similarly, Gill and Michaels (1991) analyse the US NLSY in order to explore the determinants of an individual's decision to use illegal drugs using a probit model. Their findings suggest that personal attributes rather than economic factors, including marital status, ethnicity and family background, play the dominant role in such decisions. The authors however acknowledge that due to the absence of information on prices, their findings should be regarded as "tentative". Saffer and Chaloupka (1999) analyse the Household Surveys of Drug Abuse using a variety of probit specifications to explore the effects of drug prices on participation for a range of illegal drugs. Their findings suggest that drug participation responds to prices and there is evidence of complementarity across some drug types. More recently, Duarte, Escario, and Molina (2005) find that illegal drug use among Spanish adolescents is determined by economic factors such as income as well as socio-demographic characteristics such as personal habits, family environment and receiving information relating to the adverse effects of drug use.

A key issue 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-report illicit drug use such as under-reporting, the extent of which may be influenced by a variety of factors. In terms of differences across socio-economic groups, Mensch and Kandel (1988) find that females and ethnic minorities have a tendency to underreport drug consumption. Similarly, Fendrich and Vaughn (1994) find that ethnicity has an important influence on under-reporting of substance abuse.

Mis-reporting of drugs use may also be influenced by how the survey is conducted. For instance, responding directly to an interviewer may lead to under-reporting (see, for example, OŠMuircheartaigh and Campanelli 1998), which accords with the findings of Mensch and Kandel (1988) who find that reported drugs use is higher with self-completed questionnaires. Whilst Hoyt and Chaloupka (1994) find that lower reported drugs use is associated with telephone interviews. The increased use of computer assisted self-interviewing in the gathering of such data has arguably improved the accuracy of such data. Although, it is not clear to what extent the accuracy has been improved. 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 "hard" 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 information of self-reported lifetime drugs use, i.e. repeated questions relating to whether individuals have ever taken particular drugs. The findings indicate serious under-reporting of the use of cannabis and cocaine, which in turn leads to bias in statistical modelling. For example, for one of the data sets analysed, which focuses on a relatively young age group, under-reporting rates for cannabis (cocaine) with bounds averaging 23 to 60% (31 to 95%) for all individuals were found. 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) analysed a sample of men recently arrested in Manhattan Central Booking in 1984. For cocaine, the interview data indicated a drug use rate of 43% as compared to a drug use rate of 82% elicited from urine specimens. More recently, Lu, Taylor, and Riley (2001) compare under-reporting of crack cocaine use with that of other 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.

It should be acknowledged, however, that the extent to which such findings from studies, where such cross validation is possible, can be generalised, however, is not apparent and is arguably limited given that such samples are based on somewhat atypical circumstances. The modelling strategy outlined above, in contrast, relies on a single source of cross-section survey data without recourse to validation from other sources such as drug tests or historical information on lifetime drugs consumption.

4.1 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 noninstitutionalized Australian civilian population aged over 14 providing information on drug use patterns, attitudes and behaviour (NDSHS 2010). A multi-stage, stratified area sample design ensured a random sample of households in each geographical stratum. As mentioned above, there has been some discussion in the existing literature regarding the potential for mis-reporting to be influenced by how the survey is conducted. The earlier waves of the NDSHS used face-to-face and drop and collect methods to collect data. The CATI method of data collection was introduced in the 2001 survey. In that particular survey, 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. Although CATI has become an increasingly popular method of data collection in the last decade with the widespread use of mobile and computer, as mentioned earlier, it is unclear to what extent it improves the accuracy of the data collected.

In this data set, neither the monetary expenditures nor the physical quantities of the illicit drugs consumed are reported. The information on individuals' drugs consumption is given via a discrete variable measuring whether they have consumed the drug in question over the last 12 months. There have been seven surveys conducted through the NDSHS since 1985. In this paper, due to consistency with respect to the key variables of interest, we use data from the three most recent surveys (2001, 2004 and 2007). After removal of missing values, a sample of 50,153 individuals is used for estimation. This data set has been used in several previous studies (Cameron and Williams 2001, Williams 2003, Zhao and Harris 2004). We focus on three illicit drugs, namely, marijuana, speed (amphetamines) and cocaine since information on state level prices is available for these three drugs. Information on the price of heroin is also available - however the extremely low prevalence rate of heroin consumption at 0.2%, precludes us from modelling the consumption of this drug. As mentioned above, the absence of such price data has been problematic in some of the existing studies in this area (see, for example, Gill and Michaels 1991). Hence, the inclusion of such information is an important feature of our empirical analysis. The prices of the three illicit drugs are obtained at state level from the Illicit Drug Reporting System (IDRS). The IDRS collects such data predominantly from interviewing injecting drug users and key informants who have regular contact with illicit drug users but which may potentially exhibit coverage error (NDARC 2009). In occasional cases where a price report is missing, it is constructed using information from the Australian Bureau of Criminal Intelligence (ABCI), recently replaced by the Australian Crime Commission (ACC). The ABCI/ACC is an alternative source for drug prices, which collects information on drugs through covert police units and police informants (ACC 2010). The advantage of using price data from the IDRS is that they are provided with unified measures and fewer missing observations. To be specific, the price of marijuana is measured in dollars per ounce, the price of amphetamines is measured in dollars per gram and the price of cocaine is measured in dollars per gram.

In terms of explanatory variables, we control for a range of personal and demographic characteristics, namely: gender; marital status; a quadratic in the individual's standardised age; a dummy variable for whether the individual migrated to Australia in the last 10 years; a dummy variable for whether the respondent is of Aboriginal or Torres Strait Islander origin; whether the individual has ever undergone a tattoo procedure; whether the individual has ever undergone a body piercing procedure; a dummy variable for whether there are preschool children in the household; and, finally, we control for whether the individual comes from a single parent household. We also control for educational attainment distinguishing between four categories of highest educational attainment: a tertiary degree; a non-tertiary diploma or trade certificate; year 12 education; and less than year

12 education, which is the omitted category. In terms of regional controls, we include dummy variables for whether the individual resides in a capital city and for whether the individual resides in a state where small possession of illicit drugs is decriminalised. In terms of the individual's economic situation, we control for the natural logarithm of real personal annual income before tax measured in Australian Dollars and the individual's main labour market status, i.e. employed, studying, unemployed and other activities such as retired, on a pension or performing home duties, which form the omitted category. As mentioned above, the inclusion of prices of the three illicit drugs is an important feature of our data set hence we control for the natural logarithm of the real price of marijuana measured in dollars per ounce, the natural logarithm of the real price of amphetamines measured in dollars per gram and the natural logarithm of real price of cocaine measured in dollars per gram. We also control for the prices of a range of complimentary/ substitute drugs, as well as own price, namely the natural logarithm of the real price index of alcohol, the natural logarithm of the real price index of tobacco and, finally, the natural logarithm of the real price of heroin in dollars per gram. The data on alcohol and tobacco prices are obtained in the form of indices from the Australian Bureau of Statistics (ABS 2003). The additional variables used to capture "mis-reporting" relate to the conditions under which the survey was administered, which as mentioned above, may potentially influence the extent to which individuals mis-report. 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 the survey type, i.e. if the drop-and-collect method was used, which takes a value of 0, or if the computer-assisted telephone interview (CATI) method or face-to-face method is used, which takes the value of 1; if the decision not to use drugs was influenced by fear of being caught by police/being convicted by court/going to prison; if the respondent had chosen not to ever try drugs because of pressure from family/friends; if the respondent had chosen never to take drugs because he/she did not want family/friends/teacher/employer to find out; and if the respondent had not taken drugs because of a lack of availability.

Table 1 presents summary statistics relating to the variables used in our econometric analysis for the pooled cross-section data set. It is apparent that out of the three illicit drugs, the consumption of marijuana is the most prevalent at 12%, followed by speed at 3% and cocaine at 1%. In terms of personal characteristics, 47% of the sample are male and 60% of the sample are married, with 63% of the sample being employed and the most populated highest equational attainment category is diploma level education at 35%. In terms of the variables employed to capture the possibility of lying, 43% of the sample were interviewed with another individual present, 22% of the sample indicated that they were given help to complete the questionnaire, 18% were interviewed using the CATI or face-to-face method, 8% would not try drugs because of fear of legal consequences, 6% were under pressure from family/friends not to take drugs, with 4% not trying drugs because they did not want family/friends/teacher/employer to find out and 3% because of lack of availability.

5 Results

In Table 2 we present the estimated coefficients for the participation and mis-reporting equations for the three drugs. Turning firstly to the results relating to participation, being male has a strong statistically significant positive effect on the probability of participation in the case of all three drugs. Age, on the other hand, appears to have a statistically insignificant effect on the probability of reporting consumption of marijuana and cocaine, yet a negative effect on the probability of reporting consumption of speed. Being married has a strong inverse effect on the probability of participation across all three drugs, whereas having preschool children is inversely associated with reporting consumption of speed and cocaine, yet insignificantly related to reporting consumption of marijuana. In terms of other personal characteristics, it is interesting to note that the effects of having a tattoo or body piercing are positive and statistically significant in the case of all three drugs.

In accordance with intuition, whether the individual resides in a state where small possession of illicit drugs is decriminalised is positively associated with reporting consumption of marijuana and statistically insignificantly associated with reporting consumption of the other two drugs. With respect to labour market status, relative to being retired or performing home duties, currently studying is positively associated with the probability of reporting consumption of marijuana yet insignificantly related to the probability of reporting consumption of speed or cocaine. Being unemployed, on the other hand, is positively associated with reporting positive consumption of cocaine yet statistically insignificant in the case of marijuana and speed.

The results relating to the price effects suggest complex interrelationships between the demand for different types of illicit drugs. Whilst the price of marijuana does not appear to influence the probability of reporting consumption of any of the three drugs, the price of speed is inversely associated with the probability of reporting consumption of speed and marijuana yet statistically insignificant in the case of cocaine. These findings suggest a negative own price effect in the case of speed and a negative cross price effect in terms of the effect of the price of speed on the consumption of marijuana which accords with a complementary relationship between these two drugs. The price of cocaine, on the other hand, appears to be insignificantly related to reporting consumption of marijuana and speed with a positive own price effect, albeit at the borderline of statistical significance. In terms of the effects of the prices of other complementary or substitute drugs, the price of tobacco does not appear to influence the probability of reporting positive consumption for any of the three drugs, whereas, in contrast, the price of alcohol is strongly positively associated with reporting consumption in the case of all three drugs suggestive of a substitute relationship between alcohol and the consumption of the three drugs. In terms of the price of heroin, positive effects on the consumption of speed and marijuana are apparent suggesting substitution effects whilst a negative effect exists in the case of cocaine in accordance with a complementary relationship between heroin and cocaine, which ties in with the findings of Jofre-Bonet and Petry (2008) , who explore polydrug use patterns for heroin and cocaine addicts in the US based on experiments measuring drug elasticities following changes in heroin and cocaine prices in the context of an Almost Ideal Demand System, and find that for heroin addicts, cocaine complements heroin consumption whilst cocaine addicts complement cocaine with heroin.

Turning to the mis-reporting equation and focusing on the statistically significant effects, the effect of being male only attains statistical significance in the case of marijuana, where it is positively associated with the probability of mis-reporting. The effect of age is statistically significant for both speed and cocaine where the results suggest a quadratic effect in both cases. Having body piercing is positively associated with mis-reporting across all three drugs, whereas having a tattoo only has a positive influence on mis-reporting in the case of marijuana. In terms of labour market status, a negative effect from being in work on the probability of mis-reporting in the case of marijuana is apparent, whereas studying is inversely associated with mis-reporting in the case of marijuana and speed and, finally, being unemployed is positively associated with mis-reporting in the case of marijuana. Interestingly, income is positively associated with mis-reporting across all three drugs. This tends to indicate that individuals with higher socioeconomic status are less likely to report their consumption of illegal drugs with honesty.

With respect to the effects of the additional set of variables in the mis-reporting equation, it is apparent that the presence of anyone else when the respondent was completing the questionnaire is inversely associated with the probability of mis-reporting across all three drugs, whereas if anyone helped the respondent complete the questionnaire only has a negative influence on mis-reporting in the case of marijuana. Survey type, the CATI method or face to face interview, is inversely associated with mis-reporting across all three drugs. It is apparent that the effect of the variable capturing whether the decision not to use drugs was influenced by fear of being caught has a strong inverse association with mis-reporting across all three drugs. This is also the case for the variable capturing pressure from family and friends. Finally, the effect of whether the respondent had not taken drugs due to a lack of availability is positively associated with mis-reporting in the case of speed and cocaine.

In terms of the correlation coefficients presented in Table 3, strong statistically significant correlation between the mis-reporting equations are found for: marijuana and speed; marijuana and cocaine; and cocaine and speed. Similarly, in terms participation, statistically significant correlation coefficients are found for: speed and marijuana; cocaine and marijuana; and cocaine and speed. Interestingly, a strong correlation coefficient is found between the cocaine participation equation and the marijuana mis-reporting equation. Statistically significant correlation coefficients are also found between: the speed participation equation and the marijuana mis-reporting equation; and, finally, the cocaine mis-reporting equation. The estimated correlation coefficients therefore suggest the existence of complex interrelationships between participation in the consumption of the three illicit drugs as well as in the propensity to mis-report participation across the three drugs.

In terms of the marginal effects presented in Table 4, for brevity, we present marginal effects for two joint probabilities: the probability of reporting zero consumption of all three drugs and the probability of reporting positive consumption of all three drugs. Focusing on these two extreme cases, it is apparent that being male, having a tattoo and body piercing are all inversely associated with the probability of reporting zero consumption of all three drugs, with the participation and mis-reporting effects serving to operate in the same direction. The effect of income is interesting with a positive effect on the probability of reporting non participation across all three drugs with the mis-reporting effect operating in the opposite direction thereby serving to moderate the participation effect. In the case of the marginal effects related to studying, the reverse pattern is apparent with studying being inversely associated with the probability of reporting zero participation of the three drugs whereas the mis-reporting effect serves to increase the overall effect. The only price effects that attain statistical significance are: the inverse effects from the prices of heroin and alcohol and a positive effect from the price of speed. In terms of the additional variables in the mis-reporting equation, positive statistically significant marginal effects are apparent for the presence of anyone else when the respondent was completing the questionnaire, survey type, fear of being caught and pressure from family and friends. In terms of the marginal effects related to the probability of reporting consumption of all three drugs, the marginal effects are very small and statistically insignificant throughout, which is in accordance with our expectations, since the probability of participating in all three drugs is very small.

Table 5 summarises the key finding from our analysis, which suggests that the impact of mis-reporting has a significant effect on the drug participation rates. Specifically, across all three drugs, the predicted marginal probabilities of participation are substantially higher that the sample rates of participation as indicated by the survey responses. Furthermore, it apparent that the predicted probabilities of mis-reporting conditional on participation are substantial across the three drugs, particularly in the case of marijuana and cocaine. Thus, our findings suggest that mis-reporting in survey data may lead to considerable underestimation of participation rates in the case of consumption of illicit drugs.

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, namely 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. It is apparent however that in the context of gathering sensitive information individuals may mis-report the true situation, leading to an excess amount of zero observations in the context of questions relating to activities such as illicit drug consumption, where individuals may deny their participation in such activities due to a variety of reasons such as fear of being caught. The zero-inflated modelling framework proposed in this paper is based on a two stage decision-making process whereby an individual firstly decides whether to participate in the activity in question and then decides whether or not to mis-report their behaviour. Furthermore, given that we apply this framework to survey data relating to the consumption of three illicit drugs, it is apparent that such drugs may be consumed jointly hence we expand our modelling approach to a multivariate framework whereby the participation decisions are modelled jointly. The estimated correlation coefficients across the participation and mis-reporting equations suggest the existence of complex interrelationships in illicit drug behaviour. We find that mis-reporting has a significant effect on the drug participation rates such that, across all three drugs, the predicted marginal probabilities of participation are substantially higher than that in the sample rate of participation as indicated by the raw survey data. This is found to be particularly pronounced in the case of the two 'harder' drugs, i.e. speed and cocaine. Interestingly, our findings suggest that the extent of mis-reporting is influenced by how the survey was administered as well factors such as the presence of other individuals when the survey was completed. Such findings suggest 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 since policy prescriptions in some cases may be based on inaccurate information.

References

ABS (2003): Consumer Price Index 14th Series: By Region by Group, Sub-Group and Expenditure Class - Alcohol and Tobacco, Cat. No. 6455.0.40.001, Australian Bureau of Statistics.

ACC (2010): Australian Illicit Drug Report, various reports.

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.

Cameron, L., and J. Williams (2001): "Cannabis, Alcohol and Cigarettes: Substitutes or Compliments," *The Economic Record*, 77(236), 19–34.

Caudill, S., et al. (2005): "Analysing Misleading Discrete Responses: A Logit Model Based on Misclassified Data*," *Oxford Bulletin of Economics and Statistics*, 67(1), 105–113.

Cragg, J. (1971): "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods.," *Econometrica*, 39, 829–44.

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.

Fendrich, M., and C. Vaughn (1994): "Diminished lifetime substance use over time: An inquiry into differential underreporting," *Public Opinion Quarterly*, 58(1), 96.

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, Stern School of Business, New York University.

Harris, M., and X. Zhao (2004): "Modelling Tobacco Consumption with a Zero-Inflated Ordered Probit Model," Working Paper 14/04, Monash University, Department of Econometrics and Business Statistics, Monash University, Australia.

——— (2007): "A Zero-Inflated Ordered Probit Model, with an Application to Modelling Tobacco Consumption," *Journal of Econometrics*, 141(2), 1073–1099.

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," Discussion paper, University of California, University of California, San Francisco.

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

Jofre-Bonet, M., and N. Petry (2008): "Trading apples for oranges?:: Results of an experiment on the effects of Heroin and Cocaine price changes on addicts' polydrug use," *Journal of Economic Behavior & Organization*, 66(2), 281–311.

Kristensen, N., and E. Johansson (2008): "New evidence on cross-country differences in job satisfaction using anchoring vignettes," *Labour Economics*, 15(1), 96–117.

Lambert, D. (1992): "Zero Inflated Poisson Regression with an Application to Defects in Manufacturing," *Technometrics*, 34, 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, 399–419.

Maassen van den Brink, H., and W. Groot (1999): "Job satisfaction and preference drift.," *Economics Letters*, 63(3), 363–367.

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*.

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

Mullahey, J. (1986): "Specification and Testing of Some Modified Count Data Models," *Journal of Econometrics*, 33, 341–365.

——— (1997): "Heterogeneity, Excess Zeros and the Structure of Count Data Models," *Journal of Applied Econometrics*, 12, 337–350.

NDARC (2009): "Australian Drug Trends 2007: Findings from the Illicit Drug Reporting System (IDRS)," Various Reports, National Drug and Alcohol Research Centre, University of New South Wales: Sydney.

NDSHS (2010): "Computer Files for the Unit Record Data from the National Drug Strategy Household Surveys," .

OŠMuircheartaigh, 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, 339–361.

Pudney, S. (2007): "Rarely Pure and Never Simple: Extracting the Trust from Self-Reported Data on Substance Use," *Cemmap Working paper*, (CWP11/07).

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

Sickles, R., and P. Taubman (1991): "Who uses illegal drugs?," *The American Economic Review*, 81(2), 248–251.

Smith, M. (2003): "On Dependency in Double-Hurdle Models," *Statistical Papers*, 44, 581–595.

Williams, J. (2003): "The Effects of Price and Policy on Marijuana Use: What Can Be Learned from the Australian Experience?," *Health Economics*, 13(2), 123–137.

Wish, E. (1987): *Drug use forecasting: New York 1984 to 1986*. US Dept. of Justice, National Institute of Justice.

Zhao, X., and M. Harris (2004): "Demand for Marijuana, Alcohol and Tobacco: Participation, Frequency and Cross-Equation Correlations," *Economic Record*, 80(251), 394–410.

Appendix: Definition of Variables

- **Stage**: standarised age.
- **Stagesq**: standarised age-squared .
- **Male:** = 1 for male; and = 0 for female.
- **Married:** = 1 if married or de facto; and = 0 otherwise.
- **Preskool:** = 1 if the respondent has pre-school aged child/children, and = 0 otherwise.
- **Sinpar:** 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.
- **Unemp** = 1 if unemployed; and = 0 otherwise.
- **Study:** = 1 if mainly study; 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.
- **Tattoo:** = 1 if undergone any tattoo procedure, and = 0 otherwise.
- **Bodypier:** = 1 if undergone any bodypiercing procedure, and = 0 otherwise.
- **Lrpmar:** Logarithm of real price for marijuana measured in dollars per ounce.
- **Lrpcoc:** Logarithm of real price of cocaine measured in dollars per gram.
- **Lrpspd:** Logarithm of real price of speed measured in dollars per gram.
- **Lrpher:** Logarithm of real price of heroin measured in dollars per gram.
- **Lrptob:** Logarithm of real price index for tobacco.
- **Lrpalc:** Logarithm of real price index for alcoholic drinks.
- **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) method or face-to-face method was used to collect data; and = 0 if drop and collect method was used.
- **Fear**: = 1 if the respondent's decision not to use drugs was influenced by fear of being caught by police/being convicted by court/going to prison; and = 0 otherwise.
- **Press**: = 1 if the respondent had chosen not to ever try drugs because of pressure from family/friends; and = 0 otherwise.

- **Found**: = 1 if the respondent had chosen never to take drugs because he/she did not want family/friends/teacher/employer to find out; and = 0 otherwise.
- **Avail**: = 1 if the respondent had not taken drugs because of a lack of availability; and = 0 otherwise.

Table 1: Summary Statistics

Variable	Mean	Std Dev	Minimum	Maximum	Cases
Ymar	0.118	0.323	0	1	50345
Yspd	0.031	0.173	0	1	50345
Ycoc	0.013	0.114	0	1	50345
MALE	0.473	0.499	0	1	50345
STAGE	-0.037	0.931	-1.716	2.903	50345
STAGESQ	-0.060	0.928	-1.244	4.137	50345
MARRIED	0.595	0.491	0	1	50345
PRESKOOL	0.123	0.329	0	1	50345
SINPAR	0.070	0.255	0	1	50345
CAPITAL	0.647	0.478	0	1	50345
ATSI	0.013	0.113	0	1	50345
WORK	0.633	0.482	0	1	50345
STUDY	0.063	0.244	0	1	50345
UNEMP	0.022	0.147	0	1	50345
DEGREE	0.271	0.445	0	1	50345
YR12	0.131	0.337	0	1	50345
DIPLOMA	0.349	0.477	0	1	50345
LRPINC	9.796	0.934	6.640	11.271	50345
DECRIM	0.256	0.437	0	1	50345
MIGR10	0.044	0.206	0	1	50345
TATTOO	0.108	0.311	0	1	50345
BODYPIER	0.083	0.276	0	1	50345
LRPMAR	5.239	0.156	4.809	5.474	50345
LRPCOC	5.182	0.224	4.818	5.824	50345
LRPHER	5.527	0.335	4.831	6.348	50345
LRPSPD	4.660	0.477	3.514	5.346	50345
LRPTOB	5.559	0.051	5.451	5.646	50345
LRPALC	4.712	0.036	4.630	4.766	50345
PRESENT	0.431	0.495	0	1	50153
HELP	0.229	0.420	0	1	50345
SURVTYPE	0.180	0.384	0	1	50345
FEAR	0.088	0.283	0	1	50345
PRESS	0.056	0.230	0	1	50345
FOUND	0.044	0.205	0	1	50345
AVAIL	0.029	0.168	0	1	50345

Table 2: Marijuana, Speed and Cocaine Consumption: Estimated Coefficients

	Marijuana		Speed		Cocaine	
	Participation	Mis-Reporting	Participation	Mis-Reporting	Participation	Mis-Reporting
CON	-4.105	-1.219	-9.444	-0.540	-16.740	-7.378
	(3.714)	(0.380)**	(5.466)*	(1.014)	(7.974)**	(2.151)**
MALE	0.340	0.314	0.397	-0.035	0.288	0.047
	(0.048)**	(0.057)**	(0.049)**	(0.109)	(0.070)**	(0.191)
STAGE	-0.232	0.402	-1.724	3.217	-0.016	3.764
	(0.201)	(0.286)	(0.416)**	(0.694)**	(0.822)	(1.594)**
STAGESQ	-0.899	-0.267	0.692	-2.888	-0.718	-5.811
	(0.210)**	(0.477)	(0.486)	(0.917)**	(1.205)	(2.348)**
MARRIED	-0.491	-0.026	-0.475	0.262	-0.519	0.388
	(0.051)**	(0.087)	(0.056)**	(0.182)	(0.092)**	(0.354)
PRESKOOL	0.031	-0.311	-0.370	0.224	-0.331	0.688
	(0.084)	(0.086)**	(0.075)**	(0.204)	(0.108)**	(0.429)
SINPAR	-0.034	0.055	0.044	-0.154	-0.334	0.441
	(0.071)	(0.075)	(0.072)	(0.125)	(0.138)**	(0.378)
CAPITAL	-0.125	0.194	0.066	0.254	0.374	0.025
	(0.047)**	(0.051)**	(0.050)	(0.095)**	(0.093)**	(0.305)
ATSI	-0.206	0.631	-0.179	0.144	-0.594	1.777
	(0.129)	(0.203)**	(0.155)	(0.309)	(0.325)*	(1.847)
WORK	0.098	-0.296	-0.094	-0.199	0.080	0.022
	(0.076)	(0.101)**	(0.088)	(0.218)	(0.170)	(0.442)
STUDY	0.436	-0.364	0.231	-0.444	0.310	0.030
	(0.180)**	(0.123)**	(0.144)	(0.238)*	(0.219)	(0.489)
UNEMP	0.155	0.298	-0.013	0.317	0.689	-0.844
	(0.116)	(0.159)*	(0.128)	(0.296)	(0.314)**	(0.563)
DEGREE	0.270	-0.397	-0.244	-0.135	0.082	0.087
	(0.073)**	(0.082)**	(0.073)**	(0.156)	(0.128)	(0.292)
YR12	0.089	-0.134	-0.113	-0.008	0.033	0.185
	(0.066)	(0.069)*	(0.074)	(0.119)	(0.134)	(0.284)
DIPLOMA	0.093	-0.149	-0.129	0.055	-0.066	0.370
	(0.055)*	(0.067)**	(0.060)**	(0.119)	(0.127)	(0.290)
LRPINC	-0.150	0.161	0.053	0.158	-0.011	0.601
	(0.035)**	(0.036)**	(0.040)	(0.068)**	(0.083)	(0.144)**
DECRIM	0.089	-0.031	0.085	-0.073	-0.057	-0.301
	(0.048)*	(0.053)	(0.053)	(0.097)	(0.091)	(0.229)
MIGR10	0.077	-0.392	0.138	-0.804	0.192	-0.258
	(0.134)	(0.103)**	(0.150)	(0.227)**	(0.127)	(0.317)
TATTOO	0.372	0.228	0.502	-0.052	0.317	-0.033
	(0.058)**	(0.071)**	(0.051)**	(0.131)	(0.079)**	(0.236)
BODYPIER	0.229	0.525	0.410	0.434	0.382	0.416
	(0.073)**	(0.063)**	(0.058)**	(0.109)**	(0.085)**	(0.256)
LRPMAR	0.091	-	0.072	-	0.169	-
	(0.096)		(0.134)		(0.209)	
LRPCOC	0.082	-	-0.036	-	0.192	-
	(0.056)		(0.072)		(0.106)*	
LRPHER	0.308	-	0.262	-	-0.217	-
	(0.067)**		(0.085)**		(0.116)*	
LRPSPD	-0.096	-	-0.107	-	0.046	-
	(0.036)**		(0.046)**		(0.077)	
LRPTOB	-0.493	-	-0.757	-	-0.196	-
	(0.355)		(0.515)		(0.753)	
LRPALC	1.066	-	2.091	-	3.059	-
	(0.424)**		(0.594)**		(3.059)**	
PRESENT	-	-0.145	-	-0.384	-	-0.372
		(0.035)**		(0.090)**		(0.175)
HELP	-	-0.083	-	0.014	-	-0.123
		(0.047)*		(0.104)		(0.187)
SURVTYPE	-	-0.222	-	-0.384	-	-0.671
		(0.055)**		(0.120)**		(0.245)**
FEAR	-	-0.854	-	-1.207	-	-1.865
		(0.084)**		(0.207)**		(0.530)**
PRESS	-	-0.961	-	-1.009	-	-1.651
		(0.095)**		(0.268)**		(0.562)**
FOUND	-	-0.019	-	-0.096	-	0.020
		(0.087)		(0.257)		(0.438)
AVAIL	-	0.081	-	0.581	-	1.999
		(0.094)		(0.296)**		(0.841)**

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

Table 3: Correlation Coefficients

	1M	1R	2M	2R	3M	3R
1M	-					
1R	-0.089	-				
	(0.163)					
2M	0.554	0.081	-			
	(0.126)**	(0.188)				
2R	0.522	0.452	-0.074	-		
	(0.104)**	(0.048)**	(0.208)			
3M	0.536	0.163	0.521	0.449	-	
	(0.302)*	(0.224)	(0.195)**	(0.273)*		
3R	0.429	0.314	0.215	0.525	-0.093	-
	(0.133)**	(0.069)**	(0.146)	(0.054)**	(0.489)	

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

Table 4: Marginal Effects on Selected Probabilities

	$Pr(y_{mar}=0, y_{spd}=0, y_{coc}=0 \mid \mathbf{x}_{r1}, \mathbf{x}_{r2}, \mathbf{x}_{r3})$			$Pr(y_{mar}=1, y_{spd}=1, y_{coc}=1 \mid \mathbf{x}_{r1}, \mathbf{x}_{r2}, \mathbf{x}_{r3})$		
	Participation	Mis-reporting	Overall	Participation	Misreporting	Overall
CON	0.601	0.083	0.683	-0.010	-0.002	-0.012
	(0.500)	(0.037)**	(0.496)	(0.020)	(0.003)	(0.026)
MALE	-0.047	-0.020	-0.068	0.000	0.000	0.000
	(0.011)**	(0.007)**	(0.007)**	(0.001)	(0.000)	(0.001)
STAGE	0.045	-0.033	0.012	0.000	0.001	0.001
	(0.027)*	(0.023)	(0.029)	(0.001)	(0.001)	(0.001)
STAGESQ	0.109	0.024	0.134	0.000	-0.002	-0.002
	(0.037)**	(0.035)	(0.025)**	(0.001)	(0.002)	(0.002)
MARRIED	0.067	0.001	0.068	0.000	0.000	0.000
	(0.014)**	(0.006)	(0.010)**	(0.001)	(0.000)	(0.001)
PRESKOOL	-0.001	0.020	0.019	0.000	0.000	0.000
	(0.011)	(0.008)**	(0.007)**	(0.000)	(0.000)	(0.000)
SINPAR	0.004	-0.003	0.000	0.000	0.000	0.000
	(0.010)	(0.006)	(0.009)	(0.000)	(0.000)	(0.000)
CAPITAL	0.016	-0.013	0.003	0.000	0.000	0.000
	(0.007)**	(0.004)**	(0.005)	(0.000)	(0.000)	(0.000)
ATSI	0.028	-0.042	-0.014	0.000	0.001	0.000
	(0.035)	(0.023)*	(0.040)	(0.001)	(0.001)	(0.000)
WORK	-0.012	0.020	0.008	0.000	0.000	0.000
	(0.010)	(0.008)**	(0.006)	(0.000)	(0.000)	(0.000)
STUDY	-0.058	0.025	-0.034	0.000	0.000	0.000
	(0.025)**	(0.010)**	(0.020)*	(0.001)	(0.000)	(0.000)
UNEMP	-0.019	-0.020	-0.039	0.000	0.000	0.000
	(0.017)	(0.015)	(0.019)**	(0.001)	(0.000)	(0.000)
DEGREE	-0.033	0.026	-0.007	0.000	0.000	0.000
	(0.011)**	(0.007)**	(0.009)	(0.000)	(0.000)	(0.000)
YR12	-0.010	0.009	-0.002	0.000	0.000	0.000
	(0.009)	(0.005)*	(0.007)	(0.000)	(0.000)	(0.000)
DIPLOMA	-0.011	0.010	-0.001	0.000	0.000	0.000
	(0.007)	(0.005)*	(0.005)	(0.000)	(0.000)	(0.000)
LRPINC	0.019	-0.011	0.008	0.000	0.000	0.000
	(0.006)**	(0.003)**	(0.005)*	(0.000)	(0.000)	(0.000)
DECRIM	-0.012	0.002	-0.010	0.000	0.000	0.000
	(0.007)*	(0.004)	(0.005)**	(0.000)	(0.000)	(0.000)
MIGR10	-0.011	0.027	0.016	0.000	0.000	0.000
	(0.018)	(0.011)**	(0.016)	(0.000)	(0.000)	(0.000)
TATTOO	-0.052	-0.015	-0.067	0.000	0.000	0.000
	(0.012)**	(0.007)**	(0.010)**	(0.001)	(0.000)	(0.001)
BODYPIER	-0.033	-0.035	-0.068	0.000	0.000	0.000
	(0.013)**	(0.009)**	(0.007)**	(0.001)	(0.000)	(0.001)
LRPMAR	-0.012	-	-0.012	0.000	-	0.000
	(0.013)		(0.013)	(0.000)		(0.000)
LRPCOC	-0.010	-	-0.010	0.000	-	0.000
	(0.007)		(0.007)	(0.000)		(0.000)
LRPHER	-0.042	-	-0.042	0.000	-	0.000
	(0.010)**		(0.010)**	(0.000)		(0.000)
LRPSPD	0.013	-	0.013	0.000	-	0.000
	(0.005)**		(0.005)**	(0.000)		(0.000)
LRPTOB	0.070	-	0.070	0.000	-	0.000
	(0.048)		(0.048)	(0.001)		(0.001)
LRPALC	-0.154	-	-0.154	0.002	-	0.002
	(0.059)**		(0.059)**	(0.005)		(0.005)
OPRESENT	-	0.010	0.010	-	0.000	0.000
		(0.003)**	(0.003)**		(0.000)	(0.000)
HELP	-	0.005	0.005	-	0.000	0.000
		(0.003)	(0.003)		(0.000)	(0.000)
SURVTYPE	-	0.015	0.015	-	0.000	0.000
		(0.004)**	(0.004)**		(0.000)	(0.000)
FEAR	-	0.058	0.058	-	-0.001	-0.001
		(0.011)**	(0.011)**		(0.001)	(0.001)
PRESS	-	0.065	0.065	-	-0.001	-0.001
		(0.015)**	(0.015)**		(0.001)	(0.001)
FOUND	-	0.001	0.001	-	0.000	0.000
		(0.008)	(0.008)		(0.000)	(0.000)
AVAIL	-	-0.007	-0.007	-	0.001	0.001
		(0.010)	(0.010)		(0.001)	(0.001)

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

Table 5: Sample and Predicted Probabilities

	Marijuana	Speed	Cocaine
Sample Rate of Participation	0.1183	0.0308	0.0132
Predicted Marginal Probability of Participation	0.2739	0.0785	0.0335
Lower 95% level	0.0002	0.0000	0.0000
Upper 95% level	0.8834	0.5714	0.2196
Predicted Probability of Misreporting			
Conditional on Participation	0.5324	0.2856	0.6729
Lower 95% level	0.2265	0.0344	0.0495
Upper 95% level	0.9524	0.9606	1.0000