bankwest | Curtin University

BANKWEST CURTIN ECONOMICS CENTRE

WORKING PAPER SERIES

13/5: ECONOMETRIC MODELLING OF SOCIAL BADS

William Greene, Mark Harris, Preety Srivastava and Xueyan Zhao

Bankwest Curtin Economics Centre Working Paper Series
ISSN: 2202-2791
ISBN: 978-1-925083-04-0

William Greene[1], Mark Harris[2], Preety Srivastava[3] and Xueyan Zhao[4]

[1] Stern Business School, New York University, New York, USA

[2] Bankwest Curtin Economics Centre, Curtin Business School, Curtin University, Australia

[3] School of Economics, Finance and Marketing, Royal Melbourne Institute of Technology, Australia

[4] Department of Econometrics and Business Statistics, Monash University, Australia

Econometric Modelling of Social Bads*

William Greene[1], Mark N. Harris[2]†, Preety Srivastava[3] and Xueyan Zhao[4]
[1]Stern Business School
New York University
New York, U.S.A.
[2]Curtin Business School
Curtin University, Australia
[3]School of Economics, Finance and Marketing
Royal Melbourne Institute of Technology, Australia
[4]Department of Econometrics and Business Statistics
Monash University, Australia

June 2013

Abstract

When modeling "social bads", such as illegal drug consumption, researchers are often faced with a dependent variable characterised by an "excessive" amount of zero observations. Building on the recent literature on hurdle and double-hurdle models, we propose a double-inflated modeling framework, where the zero observations are allowed to come from: non-participants; participant misreporters (who have larger loss functions associated with a truthful response); and infrequent consumers. Due to our empirical application, the model is derived for the case of an ordered discrete dependent variable. However, it is similarly possible to augment other such zero-inflated models (zero-inflated count models, and double-hurdle models for continuous variables, for example). The model is then applied to a consumer choice problem of cannabis consumption. As expected we find that misreporting has a significant (estimated) effect on the recorded incidence of marijuana. Specifically, we find that 14% of the zeros reported in the survey is estimated to come from individuals who have misreported their participation.

JEL Classification: C3, D1, I1

Keywords: Ordered outcomes, discrete data, cannabis consumption, zero-inflated responses.

---

*We kindly acknowledge funding from the Bankwest Economic Centre, Curtin University and the Australian Reseach Council. The usual caveats apply.

†Corresponding author: Curtin Business School, Bentley Campus, Building 101, School of Economics and Finance, Curtin University, Perth 6102, Australia; tel +61 89266 9692, fax: +61 8 9266 3026. *Email address:* mark.harris@curtin.edu.au

1 Introduction and Background

The issue of the presence of "excess" zeros in empirical economics 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, Jones 1989, Smith 2003); to the so-called zero-inflated (augmented) Poisson (ZIP) 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.

Consider modelling a "sensitive" response variable: that is, one where there is an associated loss-function (either perceived or actual) involved for the individual in terms of the responses she reports. Here it is clear that the researcher must be aware of the potential for misreporting. Consider, for example, modelling of responses to illicit drug consumption: there will be a strong incentive for individuals to misreport (presumably under-report) their true consumption levels for fears of legal (and/or moral) repercussions (see, for example, Pudney 2007). For instance, there is a significant amount of literature (see, for example, Worcester and Burns 1975) that suggests that many respondent give answers that they believe are "socially acceptable"and, in essence, try to please the interviewer. Indeed, this has been shown to be the case specifically with regard to drug taking (Swadi 1990).

In research in areas of discrete random variables that are inherently ordered, such misreporting has been approached by allowing the model's inherent boundary variables to vary by observed personal characteristics (Greene and Hensher 2010). However, here we suggest a more fundamental form of modelling the misreporting which is likely to be present in data which is perceived to embody a strong loss-function (social and/or legal) for the individual. Examples of such would be licit and illicit drug consumption, use of sex workers and so on. Here we suggest that the likely build-up of "zero" observations will correspond to both non-participants and participant (but infrequent) consumers. This follows the standard double-hurdle type arguments. However, for these "goods" with associated reporting loss-functions, a third source will be those participants who, fearing repercussions, report zero-consumption when in fact, this is not so.

This concept can be applied to the range of models mentioned above that exhibit a preponderance of zeros such as ZIP, ZIOP and double-hurdle. Here, in view of our application to illicit drug use recorded on an ordinal scale, we focus on a ZIOP model; although the techniques can be similarly applied to other statistical models. Explicitly, we propose a three tiered approach: the first equation determines participation or not; the second, *conditional on participation*, determines whether an individual misreports (and knowingly wrongly reports zero consumption) or reports truthfully; and finally, *for participants who report truthfully*, an ordered probit model applies, which also allows for infrequent (*i.e.*, zero) consumption of "truth-telling" participants. We term this generalisation of the ZIOP model, the double zero-inflated model (DZIOP). In addition to such "fundamental" misreporting, we can also allow for more general under- (or over-) reporting, by allowing the boundary parameters to vary by observed characteristics. Our particular application lies in misreporting in the context of marijuana consumption.

These differing types of "zero" will be driven by different systems of consumer behaviour. Moreover, a particular explanatory variable could have different effects on the three decisions. Take part-legalisation and illicit drug consumption.[1] Although it is unclear what the effects of decriminalisation will be in terms of genuine participation and (conditional) consumption levels; but clearly it is likely to reduce the chances of erroneously reporting zero-consumption for participants. Neither standard statistical models, nor their zero-inflated counterparts, would be able to disentangle these effects.

There have been a few attempts in the literature to model misreporting. For example, Hausman, Abrevaya, and Scott-Morton (1998) account for misreporting or misclassification using constant terms and Dustmann and Soest (2001) decomposes misclassification errors in panel data into a time-persistent and a time-varying component where the probability of classification is independent of respondent characteristics or any other factor. Some have used the generalised ordered probit model (or variants thereof) to model misclassification, where the cut points are function of covariates (Kristensen and

[1] In our sample, in some States in some time periods, cannabis was part-legalised or decriminalised (see Section 3.1).

Johansson 2008, Gannon 2009).

Clearly such misreporting, especially in the policy sensitive areas such as drug-taking, can potentially bias inferences in econometric analyses and lead to inappropriate policy decision-making. With an increasing use of survey data for policy analysis, it is therefore crucial to explore the incidence and implications of misreporting.

2 The Econometric Framework

2.1 A Double Zero-Inflated Ordered Probit Model (DZIOP)

Following the ZIOP model of Harris and Zhao (2007), we start by defining a discrete random variable y that is observable and assumes the discrete ordered values of $0, 1, ..., J$. A standard order probit (OP) approach would map a single latent variable to the observed outcome y via so-called boundary parameters, with the latent variable being related to a set of covariates (Greene and Hensher 2010). However, the ZIOP model involves two latent equations: a probit selection equation and an OP equation. As with the more traditional double-hurdle models (Jones 1989), individuals here have to overcome two hurdles before one observes non-zero consumption: whether to participate, and then, conditional on participation, how much to consume *which also includes zero consumption.*

However, it is our contention here, that, especially regarding the consumption of "social bads" (licit, and in particular, illicit drugs for example), participants will intentionally misreport their true consumption patterns. In particular, we hypothesis that a (probably significantly large) proportion of participants will under-report their true consumption levels (non-participants will record zero-consumption by definition). Moreover, for these participants, we make the assumption that as opposed to under-reporting their true usage, they simply state zero-consumption. That is, we contend that if a participant is concerned with legal, or otherwise, ramifications of admitting drug use, he/she will typically not prefer the option of misreporting less than they actually use, as compared to misreporting "none at all": the alternative assumption would be akin to someone feeling more comfortable to admitting breaking the law, but just by "a bit less".[2]

Finally, for participants who do not misreport, as with the ZIOP model, these are free to select any of the $j = 0, ..., J$ outcomes. In this way observed zero-consumption can arise from: 1) non-participants; 2) participants who misreport; and 3) participants who do not

[2] However, as shown below, it is relatively easy to relax this assumption if required.

misreport, but who are infrequent consumers. Thus, as compared to say, a standard OP approach, the zero observations are "double-inflated": once by non-participants and then by misreporters.

Explicitly, we suggest a three-tiered sequencing of decision making. First, the individual make a decision whether to participate or not; secondly, for participants, there is the decision to misreport or not; finally, for participants who do not misreport, the decision on how much to consume.

Following Harris and Zhao (2007) we 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 misreport. 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) misreporter and $m = 1$ a (participant) true-reporter. Again, we can write this as a linear latent form as

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

Conditional on *jointly* that $r = 1$ and $m = 1$, consumption levels under Regime 1 for true-reporters are represented by a discrete variable $\widetilde{y}$ ($\widetilde{y} = 0, 1, ..., J$) generated by an OP model via a third latent variable $\widetilde{y}^*$ such that

$$\widetilde{y}^* = \mathbf{x}_y'\boldsymbol{\beta}_y + \varepsilon_y, \tag{3}$$

with the standard mapping of

$$\widetilde{y} = \begin{cases} 0 & \text{if } \widetilde{y}^* \leq 0, \\ j & \text{if } \mu_{j-1} < \widetilde{y}^* \leq \mu_j, \ (j = 1, ..., J-1) \\ J & \text{if } \mu_{J-1} \leq \widetilde{y}^*, \end{cases} \tag{4}$$

where $\boldsymbol{\mu}$ are boundary parameters to be estimated (we assume throughout, for identification, that $\mu_0 = 0$). Of course, as with the ZIOP model, $\widetilde{y}$ is not directly observed, nor is either r or m. Here the observability criterion for observed y is

$$y = r \times m \times \widetilde{y}. \tag{5}$$

Here, an observed $y = 0$ outcome can arise from three distinct sources: $r = 0$ (the individual is a non-participant); $r = 1$ (the individual is a participant) and jointly that $m = 0$ (the individual is a miss-reporter); and finally, that jointly $r = 1$, $m = 1$ and $\widetilde{y} = 0$ (the individual is a zero consumption true-reporting participant). In the same way, to observe a positive y, we require jointly that the individual is a participant ($r = 1$) and a true reporter ($m = 1$) and that $\widetilde{y}^* > 0$.

For the time-being, assume that the stochastic terms $\boldsymbol{\varepsilon}$ $(\varepsilon_r, \varepsilon_m, \varepsilon_y)$ are independent and follow standard Gaussian distributions, the full probabilities for $y = 0$ are given by

$$\begin{aligned}\Pr(y = 0\,|\mathbf{x}) = \Pr(r = 0|\mathbf{x}) + &\\ \Pr(r = 1|\mathbf{x}) \Pr(m = 0|\mathbf{x}, r = 1) + &\\ \Pr(r = 1|\mathbf{x}) \Pr(m = 1|\mathbf{x}, r = 1) \Pr(\widetilde{y} = 0\,|\mathbf{x}, r, m = 1) &\end{aligned} \tag{6}$$

and for the remaining outcomes

$$\Pr(y = j\,|\mathbf{x}) = \Pr(r = 1|\mathbf{x}) \Pr(m = 1|\mathbf{x}, r = 1) \Pr(\widetilde{y} = j\,|\mathbf{x}, r, m = 1), \quad (j = 1, ..., J). \tag{7}$$

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 (6) and (7). 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

$$\begin{aligned}\Pr(y = 0\,|\mathbf{x}) &= [1 - \Phi(\mathbf{x}_r'\boldsymbol{\beta}_r)] + \\ &\quad \Phi(\mathbf{x}_r'\boldsymbol{\beta}_r)[1 - \Phi(\mathbf{x}_m'\boldsymbol{\beta}_m)] + \\ &\quad \Phi(\mathbf{x}_r'\boldsymbol{\beta}_r)\Phi(\mathbf{x}_m'\boldsymbol{\beta}_m)\Phi\left(-\mathbf{x}_y'\boldsymbol{\beta}_y\right)\end{aligned}$$

and

$$\begin{aligned}\Pr(y = j\,|\mathbf{x}) &= \Phi(\mathbf{x}_r'\boldsymbol{\beta}_r)\Phi(\mathbf{x}_m'\boldsymbol{\beta}_m)\left[\Phi\left(\mu_j - \mathbf{x}_y'\boldsymbol{\beta}_y\right) - \Phi\left(\mu_{j-1} - \mathbf{x}_y'\boldsymbol{\beta}_y\right)\right], \ (j = 1, ..., J-1) \\ \Pr(y = J\,|\mathbf{x}) &= \Phi(\mathbf{x}_r'\boldsymbol{\beta}_r)\Phi(\mathbf{x}_m'\boldsymbol{\beta}_m)\left[1 - \Phi\left(\mu_{J-1} - \mathbf{x}_y'\boldsymbol{\beta}_y\right)\right].\end{aligned} \tag{8}$$

So here the probability of a zero observation has been "doubly-inflated" as it is a combination of the probability of "zero consumption" from the OP process and the probability of "non-participation" from the split probit model plus that from misreporting.

Note that as per the ZIOP model, there may or may not be overlaps with the variables in the partitions in $\mathbf{x}_r, \mathbf{x}_m$ and $\mathbf{x}_y$, although undoubtedly identification will be aided by such.

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

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=0}^{J} h_{ij} \ln \left[\Pr\left(y_i = j \left| \mathbf{x}, \boldsymbol{\theta} \right. \right) \right], \tag{9}$$

where the indicator function h_{ij} is

$$h_{ij} = \begin{cases} 1 \text{ if individual } i \text{ chooses outcome } j \\ 0 \text{ otherwise.} \end{cases} \quad (i = 1, ..., N; \; j = 0, 1, ..., J). \tag{10}$$

Clearly to apply a similar set-up to count or continuous dependent variables, one would simply replace the OP densities above by the appropriate ones for the data at hand.[3] Note that, to allow for general misreporting and/or reporting heterogeneity, it would be possible to allow the $\boldsymbol{\mu}$ to be functions of covariates (Greene and Hensher 2010). However, economic identification here would require additional variables determining these boundary shifts, and in our example we have no plausible candidates for this.

2.2 Generalising the Model to Correlated Disturbances (DZIOPC)

As described above, the observed realisation of the random variable y can be viewed as the result of three separate latent equations 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 $(\varepsilon_r, \varepsilon_m, \varepsilon_y)$ follow a multivariate normal distribution with covariance matrix Ω_3, whilst maintaining the identifying assumption of unit variances. The full observability criteria are thus

$$y = rm\widetilde{y} = \begin{cases} 0 & \text{if } (r^* \leq 0) \text{ or } (r^* > 0 \text{ and } m^* \leq 0) \text{ or } (r^* > 0 \text{ and } m^* > 0 \text{ and } \widetilde{y}^* \leq 0) \\ j & \text{if } \left(r^* > 0 \text{ and } m^* > 0 \text{ and } \mu_{j-1} < \widetilde{y}^* \leq \mu_j \right), \; (j = 1, ..., J-1) \\ J & \text{if } \left(r^* > 0 \text{ and } m^* > 0 \text{ and } \mu_{J-1} < \widetilde{y}^* \right), \end{cases} \tag{11}$$

[3]Note the above model shows some similarities to that considered by Kasteridisa, Munkinb, and Yen (2010), although their focus was on smoking and cessation decisions.

which translate into the following expressions for the probabilities

$$\Pr(y) = \begin{cases} \Pr\left(y = 0\,|\mathbf{x}\right) = \left[1 - \Phi\left(\mathbf{x}_r'\boldsymbol{\beta}_r\right)\right] + \Phi_2\left(\mathbf{x}_r'\boldsymbol{\beta}_r, -\mathbf{x}_m'\boldsymbol{\beta}_m; \Omega_2\right) + \Phi_3\left(\mathbf{x}_r'\boldsymbol{\beta}_r, \mathbf{x}_m'\boldsymbol{\beta}_m, -\mathbf{x}_y'\boldsymbol{\beta}_y; \Omega_3\right) \\ \Pr\left(y = j, |\mathbf{x}\right) = \Phi_3\left(\mathbf{x}_r'\boldsymbol{\beta}_r, \mathbf{x}_m'\boldsymbol{\beta}_m, \mu_j - \mathbf{x}_y'\boldsymbol{\beta}_y; \Omega_3\right) - \Phi_3\left(\mathbf{x}_r'\boldsymbol{\beta}_r, \mathbf{x}_m'\boldsymbol{\beta}_m, \mu_{j-1} - \mathbf{x}_y'\boldsymbol{\beta}_y; \Omega_3\right) \\ \Pr\left(y = J\,|\mathbf{x}\right) = \Phi_3\left(\mathbf{x}_r'\boldsymbol{\beta}_r, \mathbf{x}_m'\boldsymbol{\beta}_m, \mathbf{x}_y'\boldsymbol{\beta}_y - \mu_{J-1}; \Omega_3\right), \end{cases} \tag{12}$$

where $\Phi_3(.)$ and $\Phi_2(.)$ respectively, denote the *c.d.f.* of the standardised trivariate and bivariate normal distribution, and where Ω_2 is the relevant partition of the full Ω_3 matrix.

ML estimation would again involve maxmisation of equation (9) replacing the probabilities of (8) with those of (12) and re-defining $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\mu}', \Omega_3)'$.[4] A test of $\Omega_3 = I_3$ is a joint test for independence of the three error terms and thus a test of the more general model given by Equation (12) against the null of a simpler nested model of Equation (8).

3 An Application to Marijuana Consumption

Marijuana use imposes a high social and economic cost on society and has been a major concern to policy-makers worldwide. It is the most commonly used drug after tobacco and alcohol, particularly in the young population. A large amount of public funds have flowed into promotional campaigns and rehabilitation programs in many countries across the world in order to treat and prevent marijuana-related harm. This has resulted in a growing importance of research in order to develop sound policies and strategies. The quality of the evidence from these scientific investigations is however an important concern. Since marijuana possession and market transactions constitute illegal activities in most jurisdictions, there is a strong incentive for marijuana users to conceal their behaviour, from fear of punishment. The concealment of use can also result from embarrassment or social disapproval (Swadi 1990). Such misreporting can have a significant impact on research findings. A major focus of this paper would be to examine the profile of those people who misreport their marijuana consumption.

3.1 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 non-institutionalized Australian civilian population aged over 14 providing information on

[4] We evaluate these multivariate probabilities using the GHK simulator.

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 misreporting 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 computer-assisted telephone interview (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. There have been seven NDSHSs conducted since 1985. In this paper, due to consistency with respect to the key variable of interest, we use data from the three most recent, at the time of writing, surveys (2001, 2004 and 2007). Definitions of all variables used in the study are given in the Appendix. A sample of 50,153 individuals is thus available for estimation. This data set has been used in several previous studies (see, for example, Cameron and Williams 2001, Zhao and Harris 2004, Harris and Zhao 2007).

In this data set, neither the monetary expenditures nor the physical quantities of marijuana consumed are reported. The information on individuals' consumption of marijuana is given via a discrete variable measuring the participation and intensity of consumption in the last 12 months. In particular, the information in the data concerning an individual's consumption of marijuana is collected through the question "*Have you used Marijuana/Cannabis in the last 12 months*" and "*In the last 12 months, how often did you use Marijuana/Cannabis?*", where the responses to the frequency of use take the form of one of the following choices: not at all ($y = 0$); using marijuana once or twice a year ($y = 1$); using marijuana monthly or every few months ($y = 2$); and using marijuana everyday or once a week ($y = 3$).

In terms of explanatory variables, we have three blocks: $\mathbf{x}_r$, to determine participation; $\mathbf{x}_m$, for misreporting; and $\mathbf{x}_y$, to determine consumption levels. While many of the variables overlap (as we have no *a priori* information as to where they should appear in the model and where not), to facilitate identification we ensure that all three have exclusion restrictions. The common variables in the three equations include a wide range of personal and demographic characteristics, namely: gender; marital status; individual's (standardised) age; a dummy variable for whether there are preschool children in the household; whether the individual comes from a single parent household; a dummy variable for whether individual resides in a capital city; a dummy variable for whether the individual migrated to Australia in the last 10 years; and a dummy variable for whether the respondent is of Aboriginal or Torres Strait Islander origin. 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. Illicit drugs are just market commodities, and users are just market participants. 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. Income may act as a social class proxy in the participation and misreporting decisions but amount of consumption is likely to be directly proportional to the level of income as it is with any normal good.

The criminal justice environment is an important determinant of drug participation and consumption. At the same time it also increases the incentive to misreport. For instance, the fear of punishment may be heightened if users perceive that supplying accurate information could lead to legal repercussions. Australia has long-standing laws with regard to marijuana *decriminalisation*. South Australia was the first jurisdiction to implement an expiation system for minor marijuana offences in 1987. Under this scheme, simple marijuana offences such as possessing, or cultivating small amounts for personal use are subject to minor penalties although the sanctions for commercial dealings are rather significant. Similar expiation systems have since been introduced in a few other Australian states and territories and yet others have been gradually easing their laws against marijuana consumption in recent years. We therefore include in all three equations a variable to represent the decriminalisation status of marijuana use across the various Australian states and territories.

Drug culture or peer drug use has been identified as an important risk factor for drug consumption (see, for example, Kenkel, Reed III, and Wang 2002, Pudney 2004, Delaney, Harmon, and Wall 2008). We therefore include a variable in $\mathbf{x}_r$ and $\mathbf{x}_y$ that indicates the proportion of the individual's friends and acquaintances that use marijuana. Given evidence on the gateway effect of alcohol to harder drugs such as marijuana (Pacula 1998b) and the association between body piercing and tattoo procedures with risk-taking behaviours (see, for example, Deschesnes, Finès, and Demers 2006, Heywood, Patrick, Smith, Simpson, Pitts, Richters, and Shelley 2012), we include in $\mathbf{x}_r$ and $\mathbf{x}_y$ whether individual started drinking alcohol at a young age, i.e. below the legal age of 18 years, and whether the individual has ever undergone a body piercing procedure or a tattoo procedure. We also control in the variables for participation and consumption, for any change in trend in consumption through the inclusion of year dummies.

As noted above, other than demographic and socioeconomic characteristics, we include in $\mathbf{x}_r$ some additional variables to instrument participation and to more strongly achieve identification. In particular, an individual's attitude towards drug laws is very likely to influence his or her decision to initiate the use of drugs but unlikely to affect the amount of drugs he or she consumes. We thus include a dummy variable in $\mathbf{x}_r$ which takes value 1 if the individual believes that a small quantity of marijuana for personal use should be a criminal offence, and 0 otherwise. Onset of drug use is much higher among adolescents and these young people are also more price sensitive than older age-groups (see, for example, Gill and Michaels 1991, Hoyt and Chaloupka 1994, Saffer and Chaloupka 1999, Cameron and Williams 2001). To examine price sensitivity in youth participation in marijuana we include an interaction term in $\mathbf{x}_r$ to represent the cross product of prices and individuals under the age of 18. This also acts as an instrument to identify participation.

Importantly, in terms of the identifying instruments for misreporting ($\mathbf{x}_m$), we include several variables. These mostly relate to the conditions under which the survey was administered, and therefore may potentially influence the extent to which individuals misreport, but not their participation or consumption levels. 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 which takes a value 0 if the drop-and-collect method was used and takes a value 1 if the CATI or face-to-face method was used. These variables conform with the factors that have been associated with misreporting or misclassification in prior studies (see, for example, Mensch and Kandel 1988, Hoyt and Chaloupka 1994, O'Muircheartaigh and Campanelli 1998, Lu, Taylor, and Riley 2001, Kraus and Augustin 2001, Berg and Lien 2006) although none of these studies have modelled misreporting explicitly. We also include as instrument a variable indicating a general lack of trust in the survey which we proxy by the percentage of compulsory questions left answered in the survey.

Finally, in terms of the decision of the levels of consumption conditional on participation, a standard consumer demand framework applies with special characteristics for addictive goods (see, for example, Becker and Murphy 1988). We thus include standard demand-schedule own and cross-drug prices in $\mathbf{x}_y$ which also serve as identifying instruments. Other than marijuana price, we thus control for the price of a range of related drugs such as amphetamines, cocaine, heroin, alcohol and tobacco given evidence that certain drugs act as either compliments or substitutes to marijuana (see, for example, Cameron and Williams 2001, Zhao and Harris 2004, Ramful and Zhao 2009). Price series for marijuana, cocaine, amphetamines and heroin for individual years and states/territories are obtained from the Illicit Drug Reporting System (IDRS) (NDARC 2009). 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), replaced by the Australian Crime Commission (ACC) in recent years. 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 and the respective price of amphetamines, cocaine and heroin is measured 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 2010). All price and income series are deflated using the all-items CPI for individuals' respective states of residence.

Table 1 presents some summary statistics on the observed marijuana consumption. On average around 89% of individuals identify themselves as current non-users. With the way the survey questions are asked, these self-identified non-users or the *build-up of zero observations* will include genuine non-users, recent quitters, infrequent users who are not *currently* consuming marijuana, as well as potential users who might use when, say, the price falls. More importantly, these self-identified non-users may include misreporters who, for example, out of embarrassment, social disapproval or fear of repercussions, may prefer to identify themselves as non-users. Given users have incentive to misreport consumption and for users who report truthfully, the choices of consumption intensities are ordered, this presents a good case for the DZIOP(C) model(s) in order to identify the different types of zero observations and their potentially different driving factors. At the same time there is also the possibility of over-reporting, particularly in the intensity of consumption, because of memory difficulties. However, there is evidence that over-reporting is rarely a problem when analysing self-reported drug use (see, for example, and references therein, Swadi 1990).

3.2 The Results

Table 2 reports the estimated coefficients of the correlated DZIOP model. In particular, we report three sets of results from the three equations: participation, truthful reporting and levels of consumption. Turning firstly to the results relating to participation, we find that, consistent with existing evidence, age, being married, having preschool children in the household, living in a capital city and being a new migrant decrease the probability of participation; on the other hand, being a male, having started drinking at a young age, having a tattoo or body piercing and being of Aboriginal or Torres Strait Islander background are associated with higher participation (see, for example, Saffer and Chaloupka 1999, Cameron and Williams 2001, Deschesnes, Finès, and Demers 2006, Ramful and Zhao 2009). Consistent with literature, marijuana use among peers and the decriminalisation laws also have a positive impact on participation (see, for example, Saffer and Chaloupka 1999, Cameron and Williams 2001, Farrelly, Bray, Zarkin, and Wendling 2001, Kenkel, Reed III, and Wang 2002, Pudney 2004, Delaney, Harmon, and Wall 2008). In terms of education, we find those with higher qualifications are more likely to report participation. With respect to labour market status, relative to being retired or performing home duties, being unemployed is significantly associated with higher probability of participation. Nevertheless, we do not find evidence of an income effect on participation. Overall our results therefore shows mixed evidence of the association of marijuana use with socioeconomic status.

As mentioned above, the identification of our model relies largely on the exclusion variables. The effects of all three instruments are statistically significant. Support for restrictive marijuana laws is negatively associated with participation. Price effects of marijuana and tobacco on young people (captured through an interaction of age with the prices) indicate that a tobacco price rise has a negative effect on youth participation while a positive association is reported in the case of a marijuana price rise. It is important to note that price of marijuana is strongly associated with quality (see, for example, Cameron and Williams 2001) and because we are unable to control for the price variation due to quality, a positive price effect could well be picking up the drug quality effect on participation.

Focusing on the misreporting equation, we find that age, being a male and living in a capital city are associated with higher probability of truthful reporting while those from a single-parent household and of aboriginal status are more likely to misreport. Interestingly, we find higher level of education to be also associated with a higher probability of misreporting. While we would expect decriminalisation to increase honest reporting in view of lower legal implications, it is nevertheless associated with higher probability of misreporting. In terms of the instruments in the misreporting equation, all four of them are statistically significant and negative. This suggests that the presence of anyone else when the respondent was completing the questionnaire or having sought help from someone when the survey was conducted increases the probability of misreporting. Similarly, the CATI and face-to-face methods of interview (relative to drop-and-collect) also increases the probability of misreporting. Finally, if the individual demonstrated a lack of trust in the survey by refusing to give a response in general, he or she also had a higher probability of misreporting.

With respect to the levels of consumption, we find that gender, marital status, aboriginal status and peer use have statistically significant positive effects. According to the rational addiction model by Becker and Murphy (1988), drug users are rational, forward looking utility maximizers who base consumption decisions on full knowledge of the consequences of addiction. Current consumption by a young adult raises the user's marginal utility of future use but also reduces the overall utility in the future given that the rational user takes account of the addictive properties of drugs and their implications for future health and wealth. We thus allow for this non-linear age-consumption relationship through a quadratic specification for age. Our results indeed shows evidence of an inverted U-shaped distribution of levels of consumption with age. In other words, at both ends of the age distribution individuals are associated with lower levels of consumption. Having young children in the household, being employed or a full-time student, having higher qualifications are all associated with lower levels of consumption. While we do not find much evidence of any significant impact of household income on participation or misreporting, we do observe a general decline in the levels of use in the highest income groups.

Turning towards our set of price instruments we find that the price effect of marijuana[5]

[5] Unlike in the participation equation which focuses on price effects on young people, here we estimate full price effects, that is, on all ages.

is positive and significant, in accordance to the results in the participation equation. As noted above, a positive price effect could thus be suggesting that level of use increases with the quality of drug. Levels of marijuana consumption is also responsive to heroin prices suggesting that the two drugs are economic substitutes. However, the price effects of cocaine, speed, tobacco and alcohol are all statistically insignificant. In summary, with at least two identifying variables exhibiting high levels of significance and along with similarly strong identifying instruments in the participation and misreporting equations, we are confident that our overall model is well identified and our results are trustworthy.

3.3 Marginal Effects

As with any probability model, marginal effects are generally more informative than coefficients. There are several sets of marginal effects that may be estimated here. For example, one may be interested in the marginal effects of an explanatory variable on probabilities such as the probability of participation, $Pr(r = 1)$, the probability of misreporting, $Pr(m = 0)$, the probabilities for the levels of consumption *conditional* on participation and true-reporting, $Pr(\widetilde{y} = j|r = 1)$, and the overall probabilities for different levels of consumption, $\Pr(y = j)$. In particular, here, we are interested in the probability of reporting zeros. The marginal effect on the overall probability of observing zero consumption, $\Pr(y = 0)$, is the sum of the effects on the probabilities of the three types of zeros; that is, the probability of non-participation, the probability of misreporting and the probability of zero-consumption arising from participants who are infrequent or potential consumers. Note that the explanatory variables of interest may appear in only one of $\mathbf{x}_r$, $\mathbf{x}_m$ or $\mathbf{x}_y$, or in all three. For comparison purposes, in Table 3, we also present results from a Generalised Ordered Probit (GOP) model, where here the boundary parameters are specified as a function of variables in $\mathbf{x}_m$ that do not appear in $\mathbf{x}_y$. Standard errors of the marginal effects for all models are obtained using the Delta method (Greene 2008).

We report the marginal effects on $\Pr(y = 0)$ (estimated at sample means) coming from these three sources in the correlated DZIOP model in Table 3. For a further comparison, we also compare these results with marginal effects estimated from a ZIOP model that allows zero observations to come from two distinct sources, *i.e.*, non-participation and infrequent consumption/misreporting; and, as noted, from a GOP model that does not explicitly model zero observations coming from different sources but allows for the boundary parameters inherent in the OP model to be a function of the *zero-generating* variables.[6] For the DZIOPC model, the overall marginal effects are decomposed in three parts: non-participation, $\Pr(r = 0)$; participation and misreporting, $\Pr(r = 1, m = 0)$; and participation, truthful reporting and zero consumption, $\Pr(r = 1, m = 1, \widetilde{y} = 0)$. In contrast, we have two components in the ZIOPC model: non-participation, $\Pr(r = 0)$; and participation and zero consumption, $\Pr(r = 1, \widetilde{y} = 0)$.

Interestingly, we observe some important differences across the estimates from the alternate models for some explanatory variables such as living in a capital city, household income and education. A key example is the effect of education. The ZIOPC model indicates that those with higher qualifications have a lower probability of non-participation but a higher probability of participation with infrequent consumption. With an additional misreporting dimension in the DZIOPC model, we find that those with higher qualifications also have a higher probability of misreporting. For instance, from the ZIOPC results, relative to those with less than year 12 qualifications, degree holders have a 1.9 percentage points (pp) lower probability of being a non-participant and a 1.3 pp higher probability of being a participant with zero consumption, resulting in an overall 0.6 pp lower probability of observing zero consumption. The DZIOPC results, in contrast, indicate that degree holders have a 2.3 pp lower probability of being a non-participant and 0.6 pp higher respective probability of being a misreporter and of being a participant with zero consumption. Overall, degree holders have a 1.1 pp net negative effect on the probability of observing zero consumption relative to those with less than year 12 qualifications. Finally, basing policy advice on the GOP model results, one would conclude that education has no impact on marijuana consumption.

The effect of decriminalisation also highlights the potentials of the DZIOPC model. For instance, the GOP indicates that decriminalisation increases consumption. From the ZIOPC model, we find that decriminalisation is associated with higher probabilities of participation and while from the DZIOPC model we find that an easing of the criminal justice system is actually associated with both a higher probability of participation and a higher probability of misreporting.

[6] Specifically, we allow the boundary or threshold parameters of the OP model, which are generally constants in a standard OP, to be a function of the instruments that we used for misreporting.

3.4 Predicted Probabilities

A key output from such a model, relates to summary predicted probabilities, especially with regard to the zeros. Thus, there are several predicted probabilities that will be of interest with the DZIOP class of models. For example, one may be interested in the marginal probability of participation, $\Pr(r = 1)$. In terms of misreporting, one may be interested in the marginal probability of misreporting, $\Pr(m = 0)$; or the joint probability of participation and misreporting, $\Pr(r = 1, m = 0)$; or the probability of truthful reporting, conditional on participation, $\Pr(m = 1|r = 1)$. Similarly, there is a range of probabilities one may be interested in predicting for levels of consumption. However, our main interest in this paper is on the misreporting dimension. Therefore, to gain insight on our sources of the observed zeros, we present in the first row in Table 4 the predicted probability of the zeros broken down into three respective components (using the equations presented above): non-participation, misreporting and zero consumption. We find that the overall predicted probability of 88.7% of zero consumption in the population is made up of the respective probability of, non-participation (84.8%), misreporting (2.9%), and infrequent consumption (1.1%).

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 for their characteristics. However, to provide further insights into the extent of misreporting, it is possible to estimate *posterior* probabilities, analogous to those considered in latent class models (Greene 2008), that are conditional on knowing what outcome the individual chose. Specifically, here this allows us to make a prediction on what percentage of these zeros come from non-participation, misreporting and zero consumption, 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 he/she is a true non-participant or a misreporting participant or an infrequent consumer (given their observed characteristics)?* The posterior probabilities for the three types of zeros are given as (Greene 2008),

$$\Pr\left(r = 0|\mathbf{x}, \mathbf{y} = \mathbf{0}\right) = \frac{f(r = 0|x)}{f(y = 0)} \tag{13}$$

$$= \frac{1 - \Phi\left(\mathbf{x}_r'\boldsymbol{\beta}_r\right)}{\left[1 - \Phi\left(\mathbf{x}_r'\boldsymbol{\beta}_r\right)\right] + \Phi_2\left(\mathbf{x}_r'\boldsymbol{\beta}_r, -\mathbf{x}_m'\boldsymbol{\beta}_m; \Omega_2\right) + \Phi_3\left(\mathbf{x}_r'\boldsymbol{\beta}_r, \mathbf{x}_m'\boldsymbol{\beta}_m, -\mathbf{x}_y'\boldsymbol{\beta}_y; \Omega_3\right)}$$

$$\Pr\left(r=1,m=0|\mathbf{x}\right)=\frac{f(r=1,m=0|x)}{f(y=0)} \tag{14}$$

$$=\frac{\Phi_2\left(\mathbf{x}_r'\boldsymbol{\beta}_r,-\mathbf{x}_m'\boldsymbol{\beta}_m;\Omega_2\right)}{\left[1-\Phi\left(\mathbf{x}_r'\boldsymbol{\beta}_r\right)\right]+\Phi_2\left(\mathbf{x}_r'\boldsymbol{\beta}_r,-\mathbf{x}_m'\boldsymbol{\beta}_m;\Omega_2\right)+\Phi_3\left(\mathbf{x}_r'\boldsymbol{\beta}_r,\mathbf{x}_m'\boldsymbol{\beta}_m,-\mathbf{x}_y'\boldsymbol{\beta}_y;\Omega_3\right)}$$

$$\Pr\left(r=1,m=1,\tilde{y}=0|\mathbf{x}\right)=\frac{f(r=1,m=1,\tilde{y}=0|x)}{f(y=0)} \tag{15}$$

$$=\frac{\Phi_3\left(\mathbf{x}_r'\boldsymbol{\beta}_r,\mathbf{x}_m'\boldsymbol{\beta}_m,-\mathbf{x}_y'\boldsymbol{\beta}_y;\Omega_3\right)}{\left[1-\Phi\left(\mathbf{x}_r'\boldsymbol{\beta}_r\right)\right]+\Phi_2\left(\mathbf{x}_r'\boldsymbol{\beta}_r,-\mathbf{x}_m'\boldsymbol{\beta}_m;\Omega_2\right)+\Phi_3\left(\mathbf{x}_r'\boldsymbol{\beta}_r,\mathbf{x}_m'\boldsymbol{\beta}_m,-\mathbf{x}_y'\boldsymbol{\beta}_y;\Omega_3\right)}$$

From Table 4, we find that about 80% of the zeros come from genuine non-participation, 14% from those who have misreported their participation and 6% from those who reported zero consumption (estimated individually and averaged across). Moreover, the small estimated standard errors on these quantities is an indication that they have been estimated relatively accurately. These are important finding suggesting that misreporting and reporting infrequent use of drugs as zero consumption in survey data may lead to considerable underestimation of drug use prevalence.

3.5 Conclusions

When modeling "social bads", such as illegal drug consumption, researchers are often faced with a dependent variable characterised by an "excessive" amount of zero observations. Such zero observations could result from individuals misreporting activities regarded as being socially undesirable, illegal or which are associated with perceived social stigma, as is the case with drug-taking. The accuracy of the information gathered from surveys is therefore crucially dependent on the respondents providing reliable and accurate responses which otherwise may lead to information being mis-classified in survey data, and which can mask the incidence of such behaviours. Thus, misreporting potentially leads to inaccurate estimates of the prevalence of such behaviours and ultimately may lead one to question the validity of any conclusions drawn and raises concerns regarding how useful such data actually is to the policy-maker.

Building on the recent literature on hurdle and double-hurdle models, we propose a double-inflated modeling framework, where the zero observations are allowed to come from: non-participants; participant misreporters (who have larger loss functions associated with a truthful response); and infrequent consumers. Due to our empirical application, the model is derived for the case of an ordered discrete dependent variable. However, it is similarly possible to augment other such zero-inflated models (zero-inflated count models, and double-hurdle models for continuous variables, for example). The model is then applied to a consumer choice problem of cannabis consumption.

Overall, our results suggest that misreporting has a significant effect on the incidence of marijuana use. Specifically, we find that 14% of the zeros reported in the survey is estimated to come from individuals who have misreported their participation in marijuana. Our modelling framework provides important insights on misreporting in surveys compared to standard modelling techniques. For instance, a Generalised Ordered Probit suggests that education has no effect on marijuana participation while our model estimates a significant negative effect of education on participation offset by positive effects on misreporting and infrequent consumption. Interestingly, our findings suggest that the extent of misreporting is influenced by how the survey was administered as well as factors such as the presence of other individuals when the survey was completed and their general trust in such surveys. In order to enhance accuracy of information gathered from surveys, it is therefore important to pay attention to the conditions under which the survey data is collected. Our findings suggest that accounting for misreporting is important in the context of using survey data related to sensitive activities, especially where such data is used to inform public policy.

References

ABS (2010): Consumer Price Index 14th Series: By Region, All Groups, Cat. No. 6401.0, Australian Bureau of Statistics.

ACC (2010): Australian Illicit Drug Report, Australian Crime Commission, Commonwealth of Australia, Canberra.

Becker, G., and K. Murphy (1988): "A Theory of Rational Addiction," Journal of Political Economy, 96(4), 675–700.

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.

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

Delaney, L., C. Harmon, and P. Wall (2008): "Behavioral Economics and Drinking Behavior: Preliminary Results from an Irish College Study," Economic Inquiry, 46(1), 29–36.

Deschesnes, M., P. Finès, and S. Demers (2006): "Are Tattooing and Body Piercing Indicators of Risk-Taking Behaviours Among High School Students?," Journal of Adolescence, 29(3), 379–393.

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.

Gannon, B. (2009): "The Influence of Economic Incentives on Reported Disability Status," Health Economics, 18(7), 743–759.

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. (2008): Econometric Analysis. Prentice Hall, New Jersey, USA, sixth edn.

Greene, W., and D. Hensher (2010): Modeling Ordered Choices. Cambridge University Press.

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.

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, San Francisco.

Heywood, W., K. Patrick, A. Smith, J. Simpson, M. Pitts, J. Richters, and J. Shelley (2012): "Who Gets Tattoos? Demographic and Behavioral Correlates of Ever Being Tattooed in a Representative Sample of Men and Women," Annals of Epidemiology, 22(1), 51–56.

Hoyt, G., and F. Chaloupka (1994): "Effect of Survey Conditions on Self-Reported Substance Use," Contemporary Economic Policy, 12(3), 109–121.

Jones, A. (1989): "A Double-Hurdle Model of Cigarette Consumption," Journal of Applied Econometrics, 4(1), 23–39.

Kasteridisa, P., M. Munkinb, and S. Yen (2010): "Demand for cigarettes: a mixed binary-ordered probit approach," Applied Economics, 42, 413–426.

Kenkel, D., R. Reed III, and P. Wang (2002): "Rational addiction, Peer Externalities and Long Run Effects of Public Policy," Discussion paper, National Bureau of Economic Research.

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.

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.

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," 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.

Pacula, R. (1998b): "Does Increasing the Beer Tax Reduce Marijuana Consumption?," Journal of Health Economics, 17, 557–85.

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. (2004): "Keeping Off the Grass? An Econometric Model of Cannabis Consumption in Britain," Journal of Applied Econometrics, 19(4), 435–453.

——— (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.

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.

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

Swadi, H. (1990): "Validating andimproving the validity of self-reports in adolescent substancemisuse surveys," Journal of Drug Issues, 20(3), 473–486.

Worcester, R., and T. Burns (1975): "A Statistical Examination of the Relative Precision of Verbal ScalesA Statistical Examination of the Relative Precision of Verbal Scales," Journal of Market Research Society, 17(3), 181–197.

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

- y: Levels of marijuana/cannabis consumption; $y = 0$ if not current user, $y = 1$ if using marijuana/cannabis once or twice a year, $y = 2$ if using marijuana/cannabis monthly or every few months, and $y = 3$ if using marijuana/cannabis everyday or once a week.

- **STAGE**: standardised age.

- **STAGESQ**: standarised age-squared .

- **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.

- **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.

- **TAFE:** $= 1$ 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.
- **LESSYR12**: = 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.
- **HINC**: Household income before tax, measured in thousands of Australian dollars where **HINC1**= 1 for $0-$9,999; **HINC2**= 1 for $10,000-$19,999; **HINC3**= 1 for $20,000-$29,999; **HINC4**= 1 for $30,000-$39,999; **HINC5**= 1 for $40,000-$59,999; **HINC6**= 1 for $60,000-$89,999; **HINC7**= 1 for $90,000-$99,999; **HINC8**= 1 for $100,000 and above. **HINC8** is used as the base of comparison for income and is dropped in the estimation.
- **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.
- **YNGDRINK**: = 1 if started drinking at the age of 12, and = 0 otherwise.
- **TATTOO:** = 1 if undergone any tattoo procedure, and = 0 otherwise.
- **PIERCING**: = 1 if undergone any body piercing procedure, and = 0 otherwise.
- **PEER**: = 1 if most or all of respondent's friends and acquaintances use marijuana/cannabis.
- **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.
- **YGPMAR:** Cross product of marijuana price with an indicator that the respondent is under 18.

- **YGPTOB:** Cross product of tobacco price with an indicator that the respondent is under 18.
- **CRIMSUP**: = 1 if respondent believes small quantity of marijuana for personal use use should be a criminal offence; 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) method 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.

Table 1: Descriptive Statistics

Variable	Mean	Std Dev	Min	Max	# Obs
Dependent variable					
$y = 0$	0.888	-	-	-	43472
$y = 1$	0.036	-	-	-	1740
$y = 2$	0.034	-	-	-	1680
$y = 3$	0.042	-	-	-	2078
Explanatory variables					
MALE	0.466	0.499	0	1	48970
STAGE	-0.016	0.916	-1.716	2.903	48970
STAGESQ	-0.045	0.917	-1.244	4.137	48970
MARRIED	0.625	0.484	0	1	48970
PRESCHOOL	0.132	0.338	0	1	48970
SINPAR	0.067	0.251	0	1	48970
CAPITAL	0.644	0.479	0	1	48970
ATSI	0.012	0.110	0	1	48970
WORK	0.606	0.489	0	1	48970
STUDY	0.061	0.240	0	1	48970
UNEMP	0.023	0.149	0	1	48970
OTHER	0.310	0.463	0	1	48970
DEGREE	0.275	0.447	0	1	48970
YR12	0.126	0.332	0	1	48970
DIPLOMA	0.347	0.476	0	1	48970
LESSYR12	0.252	0.434	0	1	48970
HINC1	0.057	0.232	0	1	48970
HINC2	0.118	0.322	0	1	48970
HINC3	0.074	0.262	0	1	48970
HINC4	0.139	0.346	0	1	48970
HINC5	0.165	0.371	0	1	48970
HINC6	0.193	0.395	0	1	48970
HINC7	0.077	0.267	0	1	48970
HINC8	0.177	0.382	0	1	48970
DECRIM	0.256	0.437	0	1	48970
MIGR10	0.047	0.211	0	1	48970
YNGDRINK	0.597	0.490	0	1	48970
YR04	0.355	0.479	0	1	48970
YR07	0.320	0.466	0	1	48970
YR10	0.325	0.468	0	1	48970
TATTOO	0.105	0.307	0	1	48970
BODYPIER	0.077	0.267	0	1	48970
PEER	0.042	0.202	0	1	48970
YGPMAR	0.155	0.890	0	5.474	48970
YGPTOB	0.165	0.942	0	5.646	48970
CRIMSUP	0.290	0.454	0	1	48970
PRESENT	0.298	0.458	0	1	48970
HELP	0.231	0.421	0	1	48970
SURVTYPE	0.179	0.383	0	1	48970
TRUST	0.035	0.051	0	0.610	48970
LRPMAR	5.238	0.155	4.809	5.474	48970
LRPCOC	5.182	0.224	4.818	5.824	48970
LRPHER	5.524	0.335	4.831	6.348	48970
LRPSPD	4.660	0.476	3.514	5.346	48970
LRPTOB	5.560	0.051	5.450	5.646	48970
LRPALC	4.712	0.036	4.630	4.766	48970

Table 2: DZIOP Estimates

	Participation		Truthful Reporting		Levels of Consumption	
CONSTANT	-1.707	(0.094)**	1.888	(0.284)**	-10.450	(6.144)*
STAGE	-0.707	(0.036)**	0.746	(0.137)**	1.326	(0.178)**
STAGESQ					-1.360	(0.210)**
MALE	0.152	(0.039)**	0.437	(0.096)**	0.283	(0.052)**
MARRIED	-0.360	(0.036)**	-0.001	(0.106)	0.121	(0.047)**
PRESCHOOL	-0.099	(0.049)**	-0.129	(0.125)	-0.114	(0.057)**
SINGPAR	0.084	(0.059)	-0.288	(0.127)**	-0.073	(0.065)
CAPITAL	-0.069	(0.035)**	0.329	(0.089)**	-0.006	(0.045)
ATSI	0.362	(0.121)**	-0.554	(0.172)**	0.263	(0.125)**
WORK	0.065	(0.047)	0.042	(0.133)	-0.241	(0.062)**
STUDY	-0.124	(0.078)	0.393	(0.190)**	-0.303	(0.094)**
UNEMP	0.206	(0.078)**	0.248	(0.216)	0.037	(0.096)
DEGREE	0.254	(0.054)**	-0.650	(0.142)**	-0.458	(0.072)**
DIPLOMA	0.143	(0.041)**	-0.315	(0.120)**	-0.181	(0.053)**
YR12	0.107	(0.050)**	-0.230	(0.140)*	-0.183	(0.060)**
HINC1	-0.031	(0.127)	0.057	(0.378)	0.347	(0.173)**
HINC2	0.166	(0.094)*	0.260	(0.297)	0.128	(0.131)
HINC3	0.058	(0.069)	0.198	(0.175)	0.321	(0.082)**
HINC4	0.070	(0.067)	0.155	(0.172)	0.263	(0.084)**
HINC5	0.005	(0.053)	0.167	(0.142)	0.316	(0.064)**
HINC6	-0.003	(0.048)	0.022	(0.126)	0.227	(0.058)**
HINC7	-0.032	(0.045)	0.198	(0.129)	0.098	(0.055)*
DECRIM	0.133	(0.036)**	-0.300	(0.090)**	0.058	(0.053)
MIGR10	-0.292	(0.074)**	1.781	(2.043)	0.044	(0.110)
YNGDRINK	0.619	(0.032)**			-0.118	(0.074)
YR04	-0.013	(0.028)			0.050	(0.066)
YR07	-0.104	(0.030)**			0.041	(0.101)
TATTOO	0.308	(0.031)**			0.017	(0.046)
PIERCING	0.427	(0.038)**			-0.039	(0.053)
PEER	1.550	(0.061)**			0.499	(0.089)**
YGPMAR	0.543	(0.295)*				
YGPTOB	-0.563	(0.279)**				
CRIMSUP	-1.250	(0.049)**				
PRESENT			-0.237	(0.079)**		
HELP			-0.215	(0.097)**		
SURVTYPE			-0.215	(0.113)*		
TRUST			-1.471	(0.722)**		
LRPMAR					0.396	(0.134)**
LRPCOC					-0.031	(0.097)
LRPHER					0.201	(0.100)**
LRPSPD					-0.090	(0.058)
LRPTOB					1.025	(0.649)
LRPALC					0.838	(0.593)
μ_1					1.006	(0.150)**
μ_2					1.770	(0.182)**
ρ_{12}	-0.342	(0.154)**				
ρ_{13}	-0.435	(0.076)**				
ρ_{23}	-0.010	(0.196)				

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

Table 3: Marginal Effects on Selected Models

	GOP	ZIOPC			DZIOPC			
		Non Participation	Zero Consumption	Overall	Non Participation	Misreporting	Zero Consumption	Overall
	$\Pr(y=0)$	$\Pr(r=0)$	$\Pr(r=1,\widetilde{y}=0)$	$\Pr(y=0)$	$\Pr(r=0)$	$\Pr(r=1,m=0)$	$\Pr(r=1,m=1,\widetilde{y}=0)$	$\Pr(y=0)$
CONSTANT	0.472 (0.302)	0.111 (0.007)**	0.253 (0.141)*	0.364 (0.141)**	0.154 (0.008)**	-0.017 (0.009)*	0.135 (0.090)	0.271 (0.090)**
STAGE	-0.101 (0.007)**	0.048 (0.003)**	-0.042 (0.005)**	0.006 (0.006)	0.064 (0.005)**	-0.007 (0.005)	-0.017 (0.006)**	0.040 (0.007)**
STAGESQ	0.171 (0.007)**	0.000 (0.000)	0.002 (0.001)	0.002 (0.001)	0.000 (0.000)	0.000 (0.000)	0.018 (0.006)**	0.018 (0.006)**
MALE	-0.031 (0.002)**	-0.012 (0.002)**	-0.011 (0.001)**	-0.022 (0.002)**	-0.014 (0.003)**	-0.004 (0.002)*	-0.004 (0.001)**	-0.021 (0.002)**
MARRIED	0.034 (0.002)**	0.030 (0.003)**	0.000 (0.001)	0.030 (0.002)**	0.032 (0.003)**	0.000 (0.001)	-0.002 (0.001)*	0.031 (0.003)**
PRESCHOOL	0.014 (0.002)**	0.007 (0.003)**	0.005 (0.002)**	0.012 (0.003)**	0.009 (0.004)**	0.001 (0.002)	0.001 (0.001)	0.012 (0.003)**
SINGPAR	0.007 (0.003)**	0.000 (0.004)	0.002 (0.002)	0.003 (0.003)	-0.008 (0.005)	0.003 (0.002)	0.001 (0.002)	-0.004 (0.004)
CAPITAL	-0.003 (0.002)*	0.001 (0.002)	-0.001 (0.001)	0.000 (0.002)	0.006 (0.003)*	-0.003 (0.002)*	0.000 (0.001)	0.003 (0.002)
ATSI	-0.008 (0.006)	0.004 (0.006)	-0.008 (0.003)**	-0.004 (0.005)	-0.033 (0.011)**	0.005 (0.005)	-0.003 (0.003)	-0.031 (0.011)**
WORK	0.002 (0.003)	-0.008 (0.003)**	0.005 (0.002)**	-0.003 (0.003)	-0.006 (0.004)	0.000 (0.001)	0.003 (0.001)**	-0.003 (0.003)
STUDY	0.009 (0.004)**	-0.011 (0.006)*	0.008 (0.002)**	-0.003 (0.005)	0.011 (0.007)	-0.004 (0.003)	0.004 (0.002)**	0.012 (0.006)**
UNEMP	-0.021 (0.005)**	-0.016 (0.006)**	-0.003 (0.003)	-0.020 (0.005)**	-0.019 (0.007)**	-0.002 (0.004)	0.000 (0.003)	-0.021 (0.006)**
DEGREE	-0.002 (0.003)	-0.019 (0.004)**	0.013 (0.002)**	-0.006 (0.003)**	-0.023 (0.005)**	0.006 (0.003)*	0.006 (0.002)**	-0.011 (0.003)**
DIPLOMA	-0.002 (0.002)	-0.008 (0.003)**	0.005 (0.001)**	-0.004 (0.002)	-0.013 (0.004)**	0.003 (0.002)	0.002 (0.001)**	-0.008 (0.003)**
YR12	-0.005 (0.003)*	-0.008 (0.004)**	0.005 (0.002)**	-0.003 (0.003)	-0.010 (0.004)**	0.002 (0.002)	0.002 (0.001)**	-0.005 (0.004)
HINC1	-0.008 (0.007)	0.007 (0.009)	-0.009 (0.004)**	-0.001 (0.008)	0.003 (0.012)	-0.001 (0.007)	-0.004 (0.004)	-0.002 (0.010)
HINC2	-0.024 (0.006)**	-0.013 (0.008)*	-0.005 (0.004)	-0.017 (0.006)**	-0.015 (0.009)*	-0.002 (0.004)	-0.002 (0.003)	-0.019 (0.008)**
HINC3	-0.019 (0.004)**	0.000 (0.005)	-0.009 (0.002)**	-0.009 (0.004)**	-0.005 (0.006)	-0.002 (0.002)	-0.004 (0.002)**	-0.011 (0.005)**
HINC4	-0.016 (0.004)**	0.000 (0.005)	-0.008 (0.002)**	-0.008 (0.004)**	-0.006 (0.006)	-0.001 (0.002)	-0.003 (0.002)**	-0.011 (0.005)**
HINC5	-0.012 (0.003)**	0.004 (0.004)	-0.009 (0.002)**	-0.005 (0.003)	0.000 (0.005)	-0.002 (0.002)	-0.004 (0.002)**	-0.006 (0.004)
HINC6	-0.006 (0.003)**	0.005 (0.003)	-0.005 (0.002)**	-0.001 (0.003)	0.000 (0.004)	0.000 (0.001)	-0.003 (0.001)**	-0.003 (0.003)
HINC7	-0.003 (0.002)	0.002 (0.003)	-0.003 (0.001)*	-0.001 (0.003)	0.003 (0.004)	-0.002 (0.002)	-0.001 (0.001)	0.000 (0.003)
DECRIM	-0.006 (0.002)**	-0.004 (0.002)*	-0.001 (0.001)	-0.005 (0.002)**	-0.012 (0.003)**	0.003 (0.002)*	-0.001 (0.001)	-0.010 (0.002)**
MIGR10	0.011 (0.004)**	0.010 (0.005)**	-0.002 (0.002)	0.009 (0.004)**	0.026 (0.007)**	-0.016 (0.017)	-0.001 (0.002)	0.009 (0.014)
YNGDRINK	-0.061 (0.002)**	-0.041 (0.003)**	-0.007 (0.002)**	-0.048 (0.002)**	-0.056 (0.003)**	0.000 (0.000)	0.002 (0.001)	-0.054 (0.003)**
YR04	-0.005 (0.003)	0.000 (0.003)	-0.001 (0.002)	-0.001 (0.002)	0.001 (0.002)	0.000 (0.000)	-0.001 (0.001)	0.001 (0.002)
YR07	0.002 (0.005)	0.003 (0.003)	0.002 (0.002)	0.005 (0.003)*	0.009 (0.003)**	0.000 (0.000)	-0.001 (0.001)	0.009 (0.003)**
TATTOO	-0.028 (0.002)**	-0.021 (0.003)**	-0.003 (0.001)**	-0.024 (0.002)**	-0.028 (0.003)**	0.000 (0.000)	0.000 (0.001)	-0.028 (0.003)**
PIERCING	-0.034 (0.003)**	-0.028 (0.004)**	-0.004 (0.001)**	-0.032 (0.003)**	-0.038 (0.004)**	0.000 (0.000)	0.000 (0.001)	-0.038 (0.003)**
PEER	-0.134 (0.004)**	-0.091 (0.005)**	-0.024 (0.003)**	-0.115 (0.005)**	-0.140 (0.007)**	0.000 (0.000)	-0.006 (0.002)**	-0.146 (0.007)**
YGPMAR		-0.036 (0.029)	0.000 (0.000)	-0.036 (0.029)	-0.049 (0.027)*	0.000 (0.000)	0.000 (0.000)	-0.049 (0.027)*
YNGPTOB		0.038 (0.027)	0.000 (0.000)	0.038 (0.027)	0.051 (0.025)**	0.000 (0.000)	0.000 (0.000)	0.051 (0.025)**
CRIMSUP		0.102 (0.003)**	0.000 (0.000)	0.102 (0.003)**	0.113 (0.004)**	0.000 (0.000)	0.000 (0.000)	0.113 (0.004)**
PRESENT	0.000 (0.000)	0.000 (0.000)	0.048 (0.006)**	0.048 (0.006)**	0.000 (0.000)	0.002 (0.001)*	0.000 (0.000)	0.002 (0.001)*
HELP	0.000 (0.000)	0.000 (0.000)	-0.007 (0.003)**	-0.007 (0.003)**	0.000 (0.000)	0.002 (0.001)	0.000 (0.000)	0.002 (0.001)
SURVTYPE	0.000 (0.000)	0.000 (0.000)	-0.001 (0.002)	-0.001 (0.002)	0.000 (0.000)	0.002 (0.001)	0.000 (0.000)	0.002 (0.001)
TRUST	0.000 (0.000)	0.000 (0.000)	-0.005 (0.002)**	-0.005 (0.002)**	0.000 (0.000)	0.013 (0.009)	0.000 (0.000)	0.013 (0.009)
LRPMAR	-0.009 (0.007)	0.000 (0.000)	-0.017 (0.015)	-0.017 (0.015)	0.000 (0.000)	0.000 (0.000)	-0.005 (0.002)**	-0.005 (0.002)**
LRPCOC	-0.007 (0.005)	0.000 (0.000)	-0.020 (0.014)	-0.020 (0.014)	0.000 (0.000)	0.000 (0.000)	0.000 (0.001)	0.000 (0.001)
LRPHER	-0.010 (0.005)**	0.000 (0.000)	0.003 (0.001)**	0.003 (0.001)**	0.000 (0.000)	0.000 (0.000)	-0.003 (0.002)*	-0.003 (0.002)*
LRPSPD	0.003 (0.003)	0.000 (0.000)	0.002 (0.001)	0.002 (0.001)	0.000 (0.000)	0.000 (0.000)	0.001 (0.001)	0.001 (0.001)
LRPTOB	0.014 (0.032)	0.000 (0.000)	0.002 (0.002)	0.002 (0.002)	0.000 (0.000)	0.000 (0.000)	-0.013 (0.009)	-0.013 (0.009)
LRPALC	-0.045 (0.030)	0.000 (0.000)	0.018 (0.009)**	0.018 (0.009)**	0.000 (0.000)	0.000 (0.000)	-0.011 (0.008)	-0.011 (0.008)

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

Table 4: Predicted Probabilities

	Non-Participation	Misreporting	Zero Consumption	Full
Marginal Probability	0.848	0.029	0.011	0.887
of zero consumption	(0.008)**	(0.007)**	(0.006)*	(0.001)**
Posterior Probability	0.805	0.138	0.058	1
of zero consumption	(0.027)**	(0.025)**	(0.024)**	

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

The Bankwest Curtin Economics Centre is an independent economic and social research organisation located within the Curtin Business School at Curtin University. The Centre was established in 2012 through the generous support from Bankwest (a division of the Commonwealth Bank of Australia), with a core mission to undertake high quality, objective research on the key economic and social issues of relevance to Western Australia.

The Centre's research and engagement activities are designed to influence economic and social policy debates in state and Federal Parliament, regional and national media, and the wider Australian community. Through high quality, evidence-based research and analysis, our research outcomes inform policy makers and commentators of the economic challenges to achieving sustainable and equitable growth and prosperity both in Western Australia and nationally.

The Centre capitalises on Curtin University's reputation for excellence in economic modelling, forecasting, public policy research, trade and industrial economics and spatial sciences. Centre researchers have specific expertise in economic forecasting, quantitative modelling, microdata analysis and economic and social policy evaluation.

A suite of tailored and national economic models and methods are maintained within the Centre to facilitate advanced economic policy analysis: these include macroeconomic and time series models, micro(simulation) models, computable general equilibrium (CGE) models, spatial modelling methods, economic index analysis, and behavioural modelling methods.

CONTACT

Bankwest Curtin Economics Centre
Curtin University
Kent Street Bentley WA 6102
GPO Box U1987 Perth WA 6845

Tel: +61 8 9266 2873

bankwesteconomicscentre@curtin.edu.au
business.curtin.edu.au/bcec