

# The causal effect of substance use on adolescent sexual activity

Alex Acworth  
Sydney University

Nicolas de Roos  
Sydney University

Hajime Katayama\*  
Sydney University

March 27, 2007

## Abstract

Using the National Longitudinal Survey of Youth 1997, we estimate the causal effects of alcohol and marijuana use on youth sexual behaviour. Three primary aspects distinguish this study from those in the literature. First, we exploit panel data to control for unobserved heterogeneity through a difference in difference estimator. Second, we control for observed heterogeneity by adopting propensity score matching methods, and adopting a more complete set of control variables than has been considered to date. Third, we separately examine the effect of initiating substance use and ceasing substance use, and thus differentiate the likely effect of policy on current substance users and non-users. The results indicate striking differences across gender. For males, initiating alcohol or marijuana use increases the likelihood of engaging in sexual intercourse and uncontracepted sexual intercourse, while ceasing use decreases the likelihood only for alcohol. For females, there is no causal link in most cases. The soundness of our estimation method is confirmed by Rashad and Kaestner's (2004) test; our method identifies no causal relationship between smoking cigarettes and sexual activities.

**JEL Classification:** J13

**Keywords:** Teenage sex; Drugs; Alcohol use; Risky behaviour

## 1 Introduction

Because teenage sex often leads to undesirable consequences such as unwanted pregnancy and sexually transmitted diseases (STDs), increasing attention has been drawn to the sexual behaviour of adolescents and youth. The potential for alcohol and drug use to influence adolescent sexual practices is a well established hypothesis. Many studies have documented a strong, positive correlation between adolescent sexual practices and

---

\*Corresponding author. School of Economics and Political Science, Merewether Building H04, University of Sydney, NSW 2006, Australia, e-mail: h.katayama@econ.usyd.edu.au, phone: +61 (2) 9036 9171, fax: +61 (2) 9351 4341.

substance use (for example, Elliot and Morse (1989); Lowry et al. (1994); Harvey and Springer (1995)). However, the reported association may not reflect causality.

There are two important econometric issues to overcome to extract the causal impact of substance use. First, the analysis needs to address the potential endogeneity of substance use arising from unobserved heterogeneity (that is, selection on unobservables). It is likely that unobservable individual characteristics, such as an individual's rate of time preference or penchant for risk taking, affect both substance use and sexual activity. Without controlling for correlation between substance use and unobservables, the resulting estimates will be biased. Second, the relationship between substance use and sexual activity is likely a non-linear one. It is possible that substance use alters the slope of the relationship between sexual activity and observable determining factors. Moreover, observable factors may be associated with sexual activity in a complicated manner. For example, the interaction terms of observables as well as the quadratic terms of observables may be important in explaining the propensity for sexual activity. If observables are not adequately controlled for, the estimates will suffer from selection on observables.

Recently, two main methods have been employed in the literature to tackle the first issue. First, instrumental variables techniques have been adopted to deal with the potential endogeneity of substance use. For example, Rees et al. (2001) and Sen (2002) apply a bivariate probit model, and rely on exclusion restrictions to identify the impact of substance use. Second, by removing (time-invariant) unobserved heterogeneity, fixed effects estimators can control for individual characteristics. Grossman (2004) employs a fixed-effects model as well as the method proposed by Altonji (2005), where identification is achieved under the assumption that the amount of selection on unobservables is equal to the amount of selection on observables.

While the fixed-effects model is straightforward in a panel setting, applying a bivariate probit model is practically difficult as it requires instrumental variables that affect substance use but not sexual activity. In addition, the literature has not addressed the second issue, nonlinear selection on observables. Indeed, a recent study by Rashad and Kaestner (2004) raises some concerns about the estimation strategies employed in the literature. Rashad and Kaestner propose an informal test based on the idea that applicable estimation strategies should not mistakenly find a causal effect of smoking on sexual behaviour. They show that the estimation strategies used by Rees et al. (2001) and Sen (2002) would in fact provide evidence of a causal effect of smoking.

In this paper, we examine the causal impact of alcohol and marijuana use on sexual behaviour among youth, using data from the National Longitudinal Survey of Youth (NLSY97). Our estimation strategy is designed to control for both unobservable and observable heterogeneity, and contains several novel aspects relative to the literature. Our panel setting permits the use of a fixed effects estimator, allowing us to control for unobserved heterogeneity. We supplement our fixed effects estimator with two strategies to carefully control for observable heterogeneity. First, we include a more complete set of control variables than has been considered by the literature. Second, we incorporate propensity score matching techniques to control for observable heterogeneity that may enter in a non-trivial manner. Our final methodological innovation is to distinguish between positive and negative treatments, allowing us to obtain separate estimates of the impact on sexual activity of commencing and ceasing substance use. This distinction is

potentially of great relevance to policy makers in regions where a significant fraction of youths are already substance users.

Our preferred estimator is the Difference in Difference Propensity Score Matching (DIDPSM) estimator (Heckman et al. (1997, 1998)). Propensity score matching (Rosenbaum and Rubin (1983)) is used to deal with possible selection bias due to systematic differences between individuals who use substances and those who do not. The inclusion of the difference in difference component allows us to control for unobserved individual specific determinants of substance use. To our knowledge the DIDPSM estimator we consider is the first to pass the informal test of Rashad and Kaestner (2004).

With our preferred estimator, we find a more tenuous causal link between substance use and sexual activity than has been reported in the literature. For males, there is some evidence of a causal effect of initiating alcohol or marijuana use on sexual activity. For females, the evidence is weaker and dependent on the matching specification in our DIDPSM estimator. We find an appreciable difference between the effect of ceasing and initiating substance use. For males, we find evidence of a causal effect on sexual activity of ceasing alcohol consumption, but not ceasing marijuana use. For females we find no causal links for ceasing substance use.

The importance of carefully controlling for both observable and unobservable heterogeneity is highlighted by the difference in results of the estimators we consider. A linear probability model finds strong spurious causal links between smoking and sexual activity. Our other estimators mitigate this link by controlling either for unobserved heterogeneity (our difference in difference estimator) or observed heterogeneity (our propensity score matching estimator). However, only our preferred estimator satisfactorily passes the informal test proposed by Rashad and Kaestner (2004).

The rest of the paper is organized as follows. Below, we briefly consider the related literature. In section 3, we outline some of the methodological issues and discuss our estimation strategy. We briefly describe our data in section 4. In section 5, we present our results. We first consider the causal effect of substance use on sexual activity, before examining the performance of our estimators in the informal test proposed by Rashad and Kaestner (2004). Finally, concluding remarks are offered in section 6.

## 2 Literature Review

The positive correlation between adolescent substance use and sexual activity is well documented.<sup>1</sup> According to a study released by the National Center on Addiction and Substance Abuse (CASA)<sup>2</sup>, teens who consume alcohol are seven times more likely than nondrinkers to have sexual intercourse, while those who use drugs are five times more likely to engage in intercourse. Similar evidence is provided in Harvey and Springer (1995) with respect to drinking. Unprotected sexual intercourse amongst youth is found to be associated with alcohol use (Strunin and Hingson (1992); Graves and Leigh (1995); Fergusson and Lynskey (1996)) and both with alcohol use and drug use (Hingson et al.

---

<sup>1</sup>See, for example, Elliott and Morse (1989) and Lowry et al. (1994). Leigh and Stall (1993) and Donovan and McEwen (1995) provide a comprehensive review of the literature.

<sup>2</sup>“Dangerous Liaisons: Substance Abuse and Sex”, Columbia University.

(1990); Cooper et al. (1994); Morrison et al. (1998)). According to Graves and Leigh (1995), heavy drinking and marijuana use amongst youth are linked with having multiple sex partners. Rosenbaum and Kandel (1990) find that youths who use substances are more likely to initiate sexual intercourse at an early age.

While the literature has been comprehensive in establishing an association between substance use and risky sexual behaviour, the causal nature of this relationship remains unknown. Authors such as Laumann et al. (1994) and Sen (2002) propose that alcohol and substance use enhance sexual arousal, lower inhibitions and hamper judgment and thus potentially have a direct causal effect on adolescent sexual practices. By contrast, Rees et al. (2001), Grossman et al. (2004), and Grossman and Markowitz (2005) argue that alcohol appears to have no causal influence in determining a young adult's sexual practices. Instead these authors conclude that the association between teenage sexual practices and substance use could easily reflect difficult-to measure unobservable factors, such as an individual's rate of time preference or propensity for risk taking, that could potentially influence both substance use and sexual behaviour. As Rashad and Kaestner (2004) note, it is difficult to establish causality since an adolescent's sexual behaviour and substance use are likely to depend on a common set of unobservable personal and social factors.

A number of recent studies have used individual level data to test whether a direct causal relationship exists. Rees et al. (2001) and Sen (2002) address the issue by using two related econometric techniques that can mitigate the effect of unobserved omitted variable bias: instrumental variables and the bivariate probit model. According to Rees et al. (2001), a causal relationship may not exist between substance use and sexual behaviour. Without controlling for endogeneity, the effects of alcohol and marijuana use on sexual behaviour are significant. However, when accounting for endogeneity, the effects become small and even statistically insignificant. Sen (2002), who examines alcohol use only, finds that in contrast to Rees et al. (2001), alcohol use has a direct positive effect on the likelihood of sexual intercourse and uncontraceptive sexual intercourse amongst adolescents. Interestingly, Sen (2002) concludes that heavy drinking (5 or more drinks in a sitting) is not a causal determinant of sexual intercourse. Grossman et al. (2004) approach the issue using a fixed effects model, which can control for unobserved time invariant, person specific effects. When they use data from the NLSY97, alcohol and marijuana use are found to affect the likelihood of sexual intercourse; however, most respective estimates are insignificant when they use data drawn from the National Longitudinal Survey of Adolescent Health.

Given the similarity in estimation techniques, the divergent findings by the three papers may seem to be surprising. However, considering the well-known difficulties associated with practical application of instrumental variables and bivariate probit models, the conflicting results are not unanticipated. As Rashad and Kaestner (2004) note, the efficiency of these procedures depends crucially on the correlation between the instruments and the endogenous variable and the validity of the exclusion restrictions - the availability of variables that significantly affect substance use, but not sexual behaviour. According to Rashad and Kaestner (2004), the application of instrumental variables and the bivariate probit model by Rees et al. (2001) and Sen (2002) might be subject to identification issues. Rashad and Kaestner (2004) note that the instruments may be weakly correlated

or potentially uncorrelated with substance use and that the exclusion restrictions may not be valid. In addition, the estimates of the correlation between the errors in the treatment equation and the equation of interest are often large and statistically insignificant. It is then pointed out that either their identification strategies are somewhat weak or the original concern for simultaneity bias is tenuous.

Rashad and Kaestner (2004) informally test the validity of the Rees et al. (2001) and Sen (2002) estimation strategies by examining the relationship between smoking and sexual activity. Smoking and sexual activity have been found to be strongly and positively correlated (Leigh and Stall (1993)), but presumably they do not have a causal relationship. [WE SHOULD EXPLAIN THIS A LITTLE - why is there intuitively no causal relationship?] Rashad and Kaestner (2004) thus argue that if the Rees et al. (2001) and Sen (2002) approaches are valid, they should identify no causal relationship between smoking and sexual behaviour. Their strategies do indeed identify a causal relationship between smoking and sexual behaviour and thus the soundness of their approaches is challenged.

### 3 Methodology

Consider the following model of sexual behaviour and substance use:

$$Y_{it} = I(f(x_{it}, D_{it}, \gamma_i, \varepsilon_{it}) > 0) \quad (1)$$

$$D_{it} = I(g(x_{it}, z_{it}, \nu_{it}) > 0) \quad (2)$$

where  $Y_{it}$  is an indicator function for sexual behaviour for individual  $i$  at time  $t$ , and  $D_{it}$  is an indicator of substance use. Sexual behaviour depends on substance use, a vector of observable control variables,  $x_{it}$ , unobservable characteristics of the individual,  $\gamma_i$ , and an idiosyncratic shock,  $\varepsilon_{it}$ . Substance use itself depends on observable control variables, and an idiosyncratic shock,  $\nu_{it}$ . A set of controls,  $z_{it}$ , determine the decision to engage in substance use, but not the decision to engage in sexual behaviour.

There are two potential econometric challenges to inferring the causal effect of substance use on sexual behaviour. First, let us suppose that the relationship embodied in  $f(\cdot)$  is a nonlinear one. It is possible that substance use alters not only the constant term but also the slope of the relationship between sexual activity and its fundamental determining factors. It is also possible that observable determining factors affect sexual activity in a nontrivial manner. For example, sexual activity may interact with observable variables and the quadratic or cubic terms of observable variables. If there is any correlation between the observable controls,  $x$ , and substance use, a linear specification will then only partially control for selection on observables. Second, any correlation between the unobservable fixed effect,  $\gamma$ , and substance use leads to selection on unobservables. Both of these issues must be dealt with to obtain consistent estimates of the relationship between substance use and sexual behaviour.

Our approach is to adopt a difference in differences propensity score matching estimator (DIDPSM) (Heckman et al. (1997,1998)). Propensity score matching (PSM) is well suited to control for observable heterogeneity even of a non-linear nature and hence allows us to deal with the first issue. The PSM technique was originally proposed by Rosenbaum

and Rubin (1983) as a method for causal inference in observational studies. During the past decade, PSM has been extensively studied and applied in the program evaluation literature (for example, Heckman et al. (1997,1998); Dehejia and Wahba (1999); Smith and Todd (2005)). To deal with the second issue, unobserved heterogeneity, we exploit the panel nature of our data by supplementing the PSM approach with a difference in differences (DID) estimator. Importantly, unlike past studies, our approach does not rely on instrumental variables. Before laying out the details, let us briefly contrast this with the empirical approach adopted by the bulk of the literature.

Past studies (for example, Rees et al. (2001); Sen (2002)) do not address the first econometric issue. That is,  $f(\cdot)$  and  $g(\cdot)$  are modelled as linear separable parameteric functions. To deal with the second challenge, past studies typically use instrumental variables. Equation (2) indicates that there are a set of control variables,  $z$ , that influence substance use, but are excluded from the sexual behaviour relationship. This lays the foundations for the instrumental variables estimation strategy adopted by much of the literature. The strategy hinges on the availability of appropriate instruments. In practice, however, it is very hard to find such instruments. For the elements of  $z$  to be valid instruments, they must satisfy two main properties. First, the instruments must be truly excludable from equation (1). In practice, this has proven to be a difficult task. For example, an instrument used by Rees et al. (2001) is whether the state of residence required schools to offer alcohol and drug prevention education. This policy is likely to reflect the social attitudes of the state; states with a permissive attitude towards substance use will not implement such a policy. It is plausible that the same factors that influence attitudes toward substance use will affect attitudes towards sexual behavior. Thus, this variable might not be excludable from equation (1). Rashad and Kaestner (2004) apply a similar argument to an instrument used by Sen (2002), the year in which the state of residence raised the minimum drinking age to 21.

The second requirement is that the instruments must be correlated with the endogenous variable, substance use. It is questionable whether instruments used in past studies are indeed correlated with substance use. Most of the instruments used are collected at a state level. Examples of those instruments include per gallon beer tax, per pack cigarette tax, per capita spending on police protection, the number of arrests per violent crime and per capita alcohol consumption by adults (see Rees et al. (2001) and Sen (2002)). Past studies typically report the significance of those variables in equation (2), thereby justifying such instruments. Note, however, that past studies often do not account for state-cluster effects when computing the standard errors. This suggests that past studies might have underestimated the standard errors and consequently overestimated the significance of their instruments. These issues are addressed by Rashad and Kaestner (2004).

To facilitate discussion of our approach, consider an alternative representation of the relationship between sexual behaviour and substance use. There are two periods,  $t = 0, 1$ . We are interested in eliciting the effect of “treatment” or substance use on sexual activity. Let  $Y_{it}^T$  and  $Y_{it}^C$  denote indicators of the sexual behaviour of individual  $i$  at time  $t$ , for a treated group and a comparison group, respectively. The composition of the treated groups varies across specifications, and we will discuss this shortly.  $Y_{it}^T$  and  $Y_{it}^C$  are

specified as follows:

$$Y_{it}^C = g_{it}^C + \gamma_i + \theta_t^C + \mu_{it}^C \quad (3)$$

$$Y_{it}^T = g_{it}^T + \gamma_i + \theta_t^T + \mu_{it}^T \quad (4)$$

where  $g_{i0}^T$  ( $g_{i0}^C$ ) is the mean outcome for individual  $i$  in the treatment (comparison) group at time 0 and  $g_{i1}^T$  ( $g_{i1}^C$ ) is the mean outcome for individual  $i$  at time 1 with (without) treatment. The error term is decomposed into individual specific ( $\gamma$ ), time specific ( $\theta$ ), and stochastic ( $\mu$ ) components.

Note that, relative to our earlier specification (equations (1) and (2)), this specification is considerably more flexible.  $g_{it}^T$  and  $g_{it}^C$  can be arbitrary functions of the observables,  $x$ , rather than taking on a specific functional form (typically linear), as in equation (1). Further, in a typical specification of (1), substance use,  $D$ , is additively separable and enters linearly. In our specification,  $D$  enters flexibly. In addition, no distributional assumptions are made for the disturbances in the model above.

We first examine our PSM estimator. PSM works in a cross-sectional context and thus the model becomes

$$\begin{aligned} Y_i^C &= g_i^C + \mu_i^C \\ Y_i^T &= g_i^T + \mu_i^T. \end{aligned}$$

In this context, the treated group consists of all those individuals in the cross-section who engaged in substance use, while the comparison group is the complementary set. The idea of PSM is that two individuals with “similar” observable characteristics will have a similar propensity for substance use. If one individual does in fact engage in substance use, while the other does not, a comparison of these individuals reveals the effect of substance use without contamination by selection on observables. The effect of treatment for individual  $i$  is the difference between the two potential outcomes,

$$\tau_i = Y_i^T - Y_i^C.$$

We focus on the related concept, the average treatment effect on the treated (ATET),

$$\tau_{ATET} = E(\tau_i | D_i = 1) = E(Y_i^T | D_i = 1) - E(Y_i^C | D_i = 1),$$

which measures the expected impact of the treatment on individuals who were actually treated. ATET may be of more relevance to policy makers given that a certain fraction of youths never engage in substance use.

Notice that  $E(Y_i^T | D_i = 1)$  can be estimated by the sample average of  $Y$  for those individuals with treatment. However,  $E(Y_i^C | D_i = 1)$  is not observable because it is a measure of the expected outcomes for individuals who did not receive treatment had they been treated. Hence, we seek to match similar individuals.

Matching is valid under two main conditions. First, we require conditional mean independence. Specifically, conditional on the covariates  $x$ , the outcome  $Y^C$  is mean independent of the treatment  $D$ :

$$\text{Assumption 1: } E(Y_i^C | x_i, D_i) = E(Y_i^C | x_i).$$

This assumption implies that selection is based exclusively on observables.<sup>3</sup> Under this assumption, the missing counterfactual means can be constructed from the non-treated comparison group. This amounts to finding matched observations from the comparison group for each observation in the treatment group and comparing the means of the matched groups.

Our second requirement is that for all  $x_i$  there is a positive probability of not being treated, i.e.

$$\text{Assumption 2: } \Pr(D_i = 1|x_i) < 1.$$

This condition ensures the possibility of an untreated individual analogue for each individual with treatment.<sup>4</sup>

Note that the conditional independence assumption requires conditioning on all relevant covariates,  $x$ . Hence, matching can become very difficult if  $x$  is of high dimension. To overcome this problem, Rosenbaum and Rubin (1983) suggest using the propensity score, which is simply the conditional probability of assignment to a particular treatment given a vector of observed covariates:

$$P(x_i) = \Pr(D_i = 1|x_i)$$

Rosenbaum and Rubin (1983) show that conditional on the propensity score,  $P(x_i)$ , the potential outcomes are independent of treatment assignment. Hence, the PSM estimator is simply the mean difference in outcomes over the common support, appropriately weighted by the propensity score distribution of participants.

Formally, we consider the following estimator:

$$\hat{\tau}_{PSM} = \sum_{i \in T} \left( Y_i - \sum_{j \in C} W_{ij} Y_j \right) w_i, \quad (5)$$

where  $T$  is the set of treated individuals, and  $C$  is the control set. The weight  $W_{ij}$  is attached to each matched observation  $j$  for individual  $i$ .  $w_i$  is a reweighting that reconstructs the outcome distribution for the treated sample. Taken together,  $W$  and  $w$  describe the matching method.

We use four different matching methods: Nearest Neighbor, Radius, Stratification, and Kernel matching methods. Nearest Neighbour matching assigns a weight,  $(W_{ij})$ , of one to the closest control unit and zero to all others. If the propensity scores in the treatment and comparison groups are not evenly distributed, one should match with replacement; this prevents bad matches, thereby improving the quality of matching. However, this may increase the variance as fewer observations from the comparison group are used. Radius matching avoids bad matches by specifying a pre-defined tolerance. All control units with estimated propensity scores falling within a pre-defined radius from the treated unit are

---

<sup>3</sup>Strictly,  $D$  could depend on unobservables, but only in a restricted manner. See Wooldridge (2002) for details.

<sup>4</sup>In addition to Assumptions 1 and 2, we also require that the support of  $x$  is equal in both the control and treatment group. If there are areas in which there is no overlap between the control and treatment group, then matching has to be performed within the area of common support only. If the area of no overlap is significant and the treatment effect is heterogeneous, then the resulting estimates are potentially biased.

matched. Given the nature of Nearest Neighbour and Radius matching, there is no need to reconstruct the outcome distribution for the treated sample;  $w_i$  is set equal to one.

Stratification and Kernel matching use all of the observations from the comparison group. Stratification matching partitions the common support of propensity scores into intervals in such a way that within each interval treated and control units have, on average, the same propensity score. The ATET is calculated by averaging each of the blocks with weights given by the distribution of the propensity score. Kernel matching uses weighted averages of individuals in the comparison group to construct the counterfactual outcome. It achieves a small variance by using the maximum amount of information; the drawback is that there are potentially many bad matches which may lead to bias. Results from Kernel matching are robust to the choice of kernel (see DiNardo and Tobias (2001)). However, the choice of bandwidth can significantly affect the results (see Silverman and Silverman, 1986; Pagan and Ullah, 1999). The bandwidth choice involves a tradeoff between small variance and an unbiased estimate of the true density.

As we condition only on the propensity score, we need to examine if the matching procedure balances the distribution of all relevant covariates in the comparison and treatment groups. We use the approach developed by Dehejia and Wahba (1999, 2002) to examine the matching quality of the PSM estimator.<sup>5</sup> Specifically, we split the sample into  $k$  equally spaced intervals of the propensity score, ensuring that in each interval the average propensity score for the treatment group does not differ from that for the comparison group. Then we use t-tests to examine whether the distribution of the covariates ( $x$ ) do not differ across intervals.

The conditional mean independence assumption (Assumption 1), is a strong assumption. It indicates that unobservable heterogeneity does not play an important role. Given our panel data, we can relax this presumption with our DIDPSM estimator. If matching is combined with DID, one can control for unobserved heterogeneity ( $\gamma$ ) as well as the time specific ( $\theta$ ) component that may be arbitrarily correlated with the treatment  $D$  (Heckman et al. (1998); Blundell and Costa Dias (2000)).

Formally, our DIDPSM is described as follows:

$$\hat{\tau}_{DIDPSM} = \sum_{i \in T} \left( (Y_{i1} - Y_{i0}) - \sum_{j \in C} W_{ij} (Y_{j1} - Y_{j0}) \right) w_i, \quad (6)$$

For this estimator, the analogue of our conditional mean independence requirement is as follows:

$$\text{Assumption 1}^1: E(Y_{i1}^C - Y_{i0}^C | x_i, D_i) = E(Y_{i1}^C - Y_{i0}^C | x_i)$$

Equivalently,

$$E [(g_{i1}^C - g_{i0}^C) + (\theta_1^C - \theta_0^C) | x_i, D_i] = E [(g_{i1}^C - g_{i0}^C) + (\theta_1^C - \theta_0^C) | x_i]$$

Notice that the conditional mean independence assumption is expressed in terms of the before–after evolution instead of levels. For this assumption to hold, it is sufficient for

---

<sup>5</sup>There are several methods for testing the balancing property. For a high level summary see Caliendo and Kopeining (2005).

both  $(g_{i1}^C - g_{i0}^C)$  and  $(\theta_1^C - \theta_0^C)$  to be conditionally mean independent of the treatment  $D_i$  (Blundell and Costa Dias (2000)).

To implement our DIDPSM estimator, we consider two specifications of the treated group. First, we restrict our sample to those individuals who did not engage in substance use in period 1. Our treated group then comprises individuals who engaged in substance use in period 2. Notice that with this specification, we are examining the effect on sexual activity of an individual’s decision to take up substance use. In our second specification, we restrict our sample to those individuals who engaged in substance use in period 1. Our treated group consists of individuals who did not engage in substance use in period 2. Now, we can examine the effect of an individual’s decision to cease substance use. From a policy maker’s perspective, the two measures could have quite different implications. We return to this issue in Section 5.4.

Our empirical methodology then consists of a dual strategy of estimating equation (5) under Assumptions 1 and 2, and estimating equation (6) under Assumptions 1’ and 2. The latter we consider under the two alternative treatment specifications described above.

## 4 Data Description

We use data from the National Longitudinal Study of Youth, 1997 (NLSY97). The survey contains a sample of 8,984 respondents, aged 12 to 16 as of December 31, 1996. The sample is designed to be representative of youths in the United States. Respondents have been surveyed annually since 1997. The survey was conducted within the respondent’s household using a computer-assisted personal interviewing system (CAPI). To ensure consistency across respondents, interviewers were automatically guided through the survey by computer software. The program prevented interviewers from entering invalid answers, alerted them to implausible answers, and reduced the probability of inconsistent data both during the interview and over time.

An interesting feature of this survey is that it includes data on sexual activity and substance use. With respect to sexual activity, respondents aged 14 and over were asked how many times they had had sexual intercourse in the 12 months preceding the survey. Those who responded positively to having sexual intercourse were further asked about the regularity of contraception use. In regard to substance use, respondents were asked how often they had consumed alcohol, marijuana and cigarettes in the past 30 days. Importantly, these sensitive questions were completed in a self-administered section of the survey so as to minimize the potential influence of parents or the interviewer.

For this study, we use Round Three (1999), in which all respondents are aged between 14 and 20, and the sample of adolescents aged 14 years and over as of December 31, 1997 from Round Two, to create a two-period panel data set. To measure sexual activity, we follow Sen (2002) and construct two binary indicators; one represents participation in sexual intercourse and the other the simultaneous decision to engage in uncontracepted sexual intercourse. Similarly, two dummy variables are created to measure alcohol use; whether the respondent reported drinking any alcohol (“drinker”) and whether the respondent had consumed more than 5 drinks in a stretch (“heavy drinker”), over the past

30 days. By using the variable, “heavy drinker”, we attempt to identify the effect of heavy drinking per se.

To analyse the effect of marijuana use, we create a dummy variable indicating whether or not the respondent has smoked marijuana in the past 30 days. Finally, we construct a binary indicator of whether or not the respondent has smoked cigarettes in the past 30 days. As in Rashad and Kaestner (2004), we use this variable to examine whether our identification strategy successfully identifies no causal relationship between smoking cigarettes and sexual activity.

The design of the survey is such that the variables on sexual activity and substance use are constructed with different time durations (12 months for sexual activity and 30 days for substance use. Consequently, we assume that substance use in the preceding 30 days is a reasonable indication of substance use throughout the year. Faced with the same issue, Sen (2002) makes a similar assumption on alcohol use.

Table 1 presents the sample means of the substance use and the sexual behavior measures. Sexual activity and substance use appears widespread amongst youths and amongst males in particular. In 1998, approximately 37% of males and 35% of females practiced sexual intercourse and of these adolescents, 30% of males and 32% of females also participated in uncontracepted sexual intercourse. Alcohol use was the most prevalent amongst adolescents; in 1998, 37% and 22% of males and 35% and 16% of females were drinkers and heavy drinkers, respectively. A smaller, but still significant, proportion of adolescents used marijuana. The proportion of adolescents who are sexually active and substance users increases over the period, while the proportion of adolescents who practice uncontracepted sexual intercourse appears to remain constant.

Table 1: Sample means of substance abuse and sexual activity

|   | 1998  |        | 1999  |        |
|---|-------|--------|-------|--------|
|   | Male  | Female | Male  | Female |
| Intercourse in past 12 months                             | 0.371 | 0.348  | 0.411 | 0.406  |
| Intercourse w/o contraception in past 12 months           | 0.112 | 0.110  | 0.103 | 0.111  |
| Consumed any alcohol in the past 30 days                  | 0.373 | 0.351  | 0.421 | 0.399  |
| Consumed $\geq 5$ drinks in a stretch in the past 30 days | 0.219 | 0.159  | 0.264 | 0.185  |
| Smoked marijuana in the past 30 days                      | 0.163 | 0.113  | 0.175 | 0.135  |
| Smoked cigarettes in the past 30 days                     | 0.294 | 0.274  | 0.305 | 0.284  |
| Sample size   | 3407  | 3286   | 4170  | 4039   |

Notes: The sample sizes reported above represent the available observations once respondents under 14 or non-interviewed respondents have been removed. We remove missing observations on an estimation basis. The number of missing observations is small, and a breakdown by year for relevant variables is available on request.

Table 2 partitions substance users by sexual activity. Importantly, substance users represent a considerably larger proportion of sexually active respondents than non-sexually active respondents. For instance, in 1998, of those who had sexual intercourse, 54% of males and 52% of females were drinkers. Amongst non-sexually active respondents, however, only 28% of males and 26% of females drank alcohol. Further, males who practiced

uncontracepted sexual intercourse, as opposed to those who did not, were almost twice as likely to be drinkers. The largest disparities occurred between adolescents who used marijuana. For example, in 1999, the proportions of males and females who used marijuana were three and four times higher, respectively, for adolescents who had had sexual intercourse as opposed to those who had not had sexual intercourse. Males that engaged in sexual intercourse or risky sexual intercourse were more likely to be substance users than females who engaged in the same practices.

Table 2: Proportion of respondents with substance use conditional on sexual activity

|                                  | Drinkers |         | Heavy drinkers |         | Marijuana users |         |
|----------------------------------|----------|---------|----------------|---------|-----------------|---------|
|                                  | Males    | Females | Males          | Females | Males           | Females |
| 1998                             |          |         |                |         |                 |         |
| Intercourse                      | 0.538    | 0.521   | 0.366          | 0.291   | 0.296           | 0.219   |
| No Intercourse                   | 0.278    | 0.259   | 0.134          | 0.09    | 0.086           | 0.056   |
| Intercourse w/o contraception    | 0.609    | 0.560   | 0.451          | 0.361   | 0.333           | 0.279   |
| No Intercourse w/o contraception | 0.344    | 0.325   | 0.181          | 0.134   | 0.141           | 0.093   |
| 1999                             |          |         |                |         |                 |         |
| Intercourse                      | 0.576    | 0.551   | 0.413          | 0.292   | 0.321           | 0.243   |
| No Intercourse                   | 0.307    | 0.294   | 0.157          | 0.111   | 0.091           | 0.062   |
| Intercourse w/o contraception    | 0.676    | 0.561   | 0.536          | 0.327   | 0.407           | 0.305   |
| No Intercourse w/o contraception | 0.391    | 0.378   | 0.232          | 0.167   | 0.149           | 0.114   |

Notes: This table describes the number of substance users as a proportion of respondents in each of the sexual activity categories. Each category refers to observations in the past 12 months.

Respondents in the NLSY97 also provided detailed information on a range of socio-economic, personal and family characteristics. Using this rich information, we construct a set of control variables that are expected to affect both substance use and sexual activity. These variables include age, religious affiliation, dwelling, race, ethnicity, whether the respondent attended a public high school, whether the respondent lives in the south, whether the respondent is an urbanite, parental education, parental family status, and regional unemployment rates. We also construct two dummy variables for whether a parental figure is considered “strict” and for whether the respondent has experienced “hard times” throughout his/her childhood. The former controls for the relationship between the respondent and his/her parents and the latter for the respondent’s life experience. Furthermore, we use indices of substance use and delinquency to control for the respondent’s past involvement in risky behaviour. The substance use index assigns a weight of one if the respondent answers positively to having ever drunk alcohol, smoked cigarettes or used marijuana. The delinquency index is created from a set of 10 questions regarding the youth’s history<sup>6</sup>. To ensure that all control variables are exogenous and pre-treatment, we lag the controls one year relative to the outcome and treatment variables. Variable definitions are provided in appendix A.

Descriptive statistics of the control variables in 1998 are presented in Tables 3 and 4 for males and females, respectively. The control variables are partitioned into two groups;

<sup>6</sup>For more detail on these indices, see NLSY97 user guide appendix 9.

the treatment group (respondents who reported substance use) and the comparison group (respondents who did not use substances). For each control variable, we report the means of both the treatment and comparison groups. Systematic difference between substance users and non-users are evident, suggesting that selection may be an important issue. For example, within the treatment group, respondents are more likely to be white, have an urban residence, have a history of delinquency, and have used substances in the past. On the other hand, respondents in the comparison group are more likely to have a Baptist religious affiliation, be young, black, live in the south, attend a public school and have strict parental figures. Respondents who live with their biological parents are more likely to be in the comparison group for marijuana use.

Table 3: Descriptive statistics for control variables (males)

|                           | Drinkers  |         | Heavy drinkers |         | Marijuana users |         |
|---------------------------|-----------|---------|----------------|---------|-----------------|---------|
|                           | Treatment | Control | Treatment      | Control | Treatment       | Control |
| Baptist                   | 0.18      | 0.24    | 0.16           | 0.23    | 0.20            | 0.21    |
| Roman Catholic            | 0.33      | 0.26    | 0.35           | 0.27    | 0.29            | 0.29    |
| Parents Strict            | 0.31      | 0.43    | 0.27           | 0.41    | 0.31            | 0.39    |
| Age (1997)                | 16.27     | 15.64   | 16.47          | 15.71   | 16.24           | 15.84   |
| House                     | 0.82      | 0.78    | 0.84           | 0.78    | 0.81            | 0.79    |
| Apartment                 | 0.11      | 0.14    | 0.09           | 0.14    | 0.13            | 0.13    |
| Hard times                | 0.04      | 0.06    | 0.03           | 0.06    | 0.06            | 0.05    |
| Lives in South            | 0.33      | 0.41    | 0.31           | 0.40    | 0.30            | 0.39    |
| Has Urban residence       | 0.72      | 0.72    | 0.72           | 0.72    | 0.77            | 0.71    |
| Lives with mum and dad    | 0.51      | 0.49    | 0.51           | 0.50    | 0.43            | 0.51    |
| Lives w/o mum or dad      | 0.07      | 0.08    | 0.06           | 0.08    | 0.06            | 0.07    |
| Attended public school    | 0.87      | 0.88    | 0.86           | 0.88    | 0.87            | 0.88    |
| Parents highest education | 13.67     | 13.29   | 13.54          | 13.43   | 13.67           | 13.41   |
| White                     | 0.67      | 0.53    | 0.71           | 0.55    | 0.61            | 0.58    |
| Black                     | 0.17      | 0.32    | 0.13           | 0.30    | 0.23            | 0.26    |
| Hispanic                  | 0.22      | 0.20    | 0.22           | 0.21    | 0.20            | 0.21    |
| Mixed race                | 0.01      | 0.01    | 0.01           | 0.01    | 0.01            | 0.01    |
| Delinquency               | 1.39      | 0.69    | 1.61           | 0.75    | 1.99            | 0.76    |
| Substance Use             | 1.65      | 0.65    | 1.89           | 0.78    | 2.06            | 0.86    |
| Unemployment rate         |           |         |                |         |                 |         |
| Less than 3%              | 0.22      | 0.23    | 0.22           | 0.23    | 0.23            | 0.22    |
| In between 3% and 6%      | 0.57      | 0.57    | 0.57           | 0.57    | 0.56            | 0.57    |
| In between 6% and 9%      | 0.17      | 0.16    | 0.16           | 0.16    | 0.15            | 0.17    |
| In between 9% and 12%     | 0.02      | 0.01    | 0.02           | 0.01    | 0.02            | 0.02    |
| In between 12% and 15%    | 0.02      | 0.01    | 0.02           | 0.01    | 0.02            | 0.01    |
| Greater than 15%          | 0.01      | 0.01    | 0.01           | 0.01    | 0.01            | 0.01    |

Notes: We partition the control variables between the treatment and control groups of the endogenous substance use variables. The endogenous variables are taken from Round Three (1999). To ensure that the control variables are exogenous and pre-treatment we use the data from Round Two (1998).

Table 4: Descriptive statistics for control variables (females)

|                           | Drinkers  |         | Heavy drinkers |         | Marijuana users |         |
|---------------------------|-----------|---------|----------------|---------|-----------------|---------|
|                           | Treatment | Control | Treatment      | Control | Treatment       | Control |
| Baptist                   | 0.16      | 0.28    | 0.15           | 0.25    | 0.17            | 0.24    |
| Roman Catholic            | 0.30      | 0.25    | 0.33           | 0.26    | 0.28            | 0.27    |
| Parents Strict            | 0.31      | 0.42    | 0.27           | 0.40    | 0.30            | 0.39    |
| Age (1997)                | 16.22     | 15.77   | 16.34          | 15.86   | 16.18           | 15.91   |
| House                     | 0.81      | 0.77    | 0.81           | 0.78    | 0.81            | 0.78    |
| Apartment                 | 0.12      | 0.15    | 0.12           | 0.14    | 0.13            | 0.14    |
| Hard times                | 0.04      | 0.05    | 0.04           | 0.05    | 0.04            | 0.05    |
| Lives in South            | 0.33      | 0.42    | 0.31           | 0.40    | 0.33            | 0.39    |
| Has Urban residence       | 0.74      | 0.72    | 0.73           | 0.73    | 0.75            | 0.73    |
| Lives with mum and dad    | 0.48      | 0.45    | 0.49           | 0.46    | 0.37            | 0.48    |
| Lives w/o mum or dad      | 0.09      | 0.09    | 0.09           | 0.09    | 0.12            | 0.09    |
| Attended public school    | 0.85      | 0.88    | 0.84           | 0.87    | 0.83            | 0.87    |
| Parents highest education | 13.56     | 12.98   | 13.69          | 13.10   | 13.61           | 13.14   |
| White                     | 0.68      | 0.51    | 0.73           | 0.54    | 0.69            | 0.56    |
| Black                     | 0.17      | 0.33    | 0.12           | 0.30    | 0.18            | 0.28    |
| Hispanic                  | 0.20      | 0.22    | 0.19           | 0.22    | 0.16            | 0.22    |
| Mixed race                | 0.01      | 0.01    | 0.01           | 0.01    | 0.01            | 0.01    |
| Delinquency               | 0.81      | 0.36    | 1.00           | 0.44    | 1.36            | 0.41    |
| Substance Use             | 1.65      | 0.65    | 1.93           | 0.85    | 2.22            | 0.86    |
| Unemployment rate         |           |         |                |         |                 |         |
| Less than 3%              | 0.25      | 0.23    | 0.24           | 0.24    | 0.26            | 0.24    |
| In between 3 and 6%       | 0.54      | 0.55    | 0.55           | 0.55    | 0.56            | 0.54    |
| In between 6% and 9%      | 0.16      | 0.17    | 0.17           | 0.16    | 0.14            | 0.17    |
| In between 9% and 12%     | 0.02      | 0.01    | 0.01           | 0.02    | 0.01            | 0.02    |
| In between 12% and 15%    | 0.02      | 0.02    | 0.01           | 0.02    | 0.02            | 0.02    |
| Greater than 15%          | 0.00      | 0.01    | 0.00           | 0.01    | 0.00            | 0.01    |

Notes: We partition the control variables between the treatment and control groups of the endogenous substance use variables. The endogenous variables are taken from Round Three (1999). To ensure that the control variables are exogenous and pre-treatment we use the data from Round Two (1998).

## 5 Estimation Results

Following past studies, we analyse male and females separately. We are interested in how, if at all, substance use has shaped adolescents' sexual practices across the two-year panel period. In what follows, we compare the ATET estimates from four different models: two simple linear probability models (LPM) and (LPM\*), a propensity score matching (PSM) model, a difference in differences model (DID) and a difference in differences propensity score matching (DIDPSM) model. LPM\* differs from LPM by the inclusion of two important observable variables, the indices for substance use and delinquency. The LPM is used to benchmark the control variables usually used in the literature, while the

LPM\* demonstrates the importance of controlling for youths' observable history. The PSM model, which balances the confounding effect of observable factors, is used to detect whether or not observables play a confounding role in adolescent substance use and sexual practices. The DID model controls for potential unobserved individual heterogeneity which may be correlated with both substance use and sexual behaviour. Finally, the DIDPSM model is a combination of the PSM and DID models.

For the LPM, LPM\* and PSM model, we take the dependent variables from the Third and the control variables from the Second Rounds of the NLSY97 respectively. For the PSM model, our treatment and comparison groups are those respondents who did and did not use substances in 1999, respectively. For the DID and DIDPSM models, our comparison group consists of those respondents who did not use substances in either 1998 or 1999 and our treatment group consists of those respondents who did not use substances in 1998 but were substance users in 1999. As a dependent variable, we take the difference between 1999 and 1998 sexual intercourse indicators and we use 1998 control variables. In section 5.4 we also consider the reverse treatment for our DID and DIDPSM models. We will lay out the details when we come to this specification.

## 5.1 Propensity Score Models

Table 5 reports the probit estimates of the propensity scores for the DIDPSM model.<sup>7</sup> In each case, the balancing property is satisfied at significance levels better than 0.1%. At times, insignificant variables were removed to satisfy the balancing property. All the models are significant at the 1% level; however, the covariates did not individually perform well in determining females' marijuana use or heavy drinking by males.

Perhaps the most striking result is that, after controlling for observed heterogeneity, the indices of risky behaviour (the substance use index and delinquency index) continue to be important determinants of substance use, indicating that past risky behaviour has a real effect on substance use in the future. In particular, adolescents who have used substances in the past are substantially more likely to begin using them again. Delinquency amongst adolescent females encourages the uptake of marijuana use while females who live with their biological parents tend to avoid marijuana use. We also find that females who are Baptist are statistically less likely to become both drinkers and heavy drinkers over the period, while females with an urban residence tend to become drinkers and females whose parents did not complete year 10 high school tend to become heavy drinkers. For males, being white or older increases the likelihood that they will become drinkers while a history of delinquency appears to encourage males to become heavy drinkers and start using marijuana. Males who reside in urban areas are more likely to begin using marijuana over the period while males who live in the south tend not to use marijuana.

We examined the distribution of the propensity scores for the treatment and comparison groups.<sup>8</sup> Matching estimators would provide an inconsistent estimate of the ATET if there are large disparities in the mode and empirical support of the treatment and comparison group. The area of empirical support between the comparison and treatment

---

<sup>7</sup>Propensity scores for the PSM model are similar in sign and available from the authors on request.

<sup>8</sup>Histograms are available on request for both treatment and comparison groups, by gender, for both the PSM and DIDPSM.

Table 5: DIDPSM model - Positive treatment

|                                    | Drink   |         | Drink heavily |         | Marijuana |         |
|------------------------------------|---------|---------|---------------|---------|-----------|---------|
|                                    | Males   | Females | Males         | Females | Males     | Females |
| Unemployment < 3%                  | -0.049  | -0.069  | -0.089        | -0.108  |           | -0.042  |
|                                    | (-0.65) | (-0.88) | (-1.18)       | (-1.33) |           | (-0.48) |
| Unemployment 9% - 12%              | 0.38    | 0.075   | 0.303         | -0.007  |           | -0.643  |
|                                    | (1.5)   | (0.31)  | (1.39)        | (-0.03) |           | (-0.48) |
| Unemployment > 15%                 | -1.11   | -1.014  | -0.476        | -0.334  |           | 0.115   |
|                                    | (-1.83) | (-1.76) | (-1.01)       | (-0.75) |           | (0.26)  |
| Urban residence                    | 0.021   | 0.202   | 0.072         | 0.122   | 0.159     | 0.144   |
|                                    | (0.3)   | (2.54)  | (0.98)        | (1.49)  | (2.03)    | (1.63)  |
| Lives in the South                 | -0.025  | -0.027  | -0.077        | 0.032   | -0.239    | 0.097   |
|                                    | (-0.35) | (-0.38) | (-1.08)       | (0.43)  | (-3.16)   | (1.19)  |
| Baptist religious affiliation      | -0.081  | -0.192  | -0.047        | -0.207  |           | -0.173  |
|                                    | (-0.99) | (-2.36) | (-0.54)       | (-2.27) |           | (-1.78) |
| Lives with mum and dad             | 0.05    | 0.051   |               | 0.149   | -0.128    | -0.244  |
|                                    | (0.76)  | (0.75)  |               | (2.09)  | (-1.84)   | (-3.16) |
| Strict parental figure             | -0.067  | -0.103  | -0.084        | -0.111  |           | -0.114  |
|                                    | (-0.94) | (-1.4)  | (-1.19)       | (-1.42) |           | (-1.37) |
| Parents dropped out of high school |         | 0.239   | -0.146        | 0.48    | 0.046     | 0.25    |
|                                    |         | (1.64)  | (-1.04)       | (2.56)  | (0.29)    | (1.21)  |
| Attended public school             |         | 0.086   |               | -0.033  | 0.108     | -0.029  |
|                                    |         | (0.95)  |               | (-0.36) | (1.08)    | (0.29)  |
| Been through hard times            | -0.126  | -0.027  | -0.189        | -0.181  | 0.107     | -0.319  |
|                                    | (-0.89) | (-0.19) | (-1.23)       | (-1.07) | (0.72)    | (-1.65) |
| White                              | 0.199   | 0.15    | 0.164         | 0.074   | -0.206    | -0.071  |
|                                    | (2.92)  | (2.16)  | (1.47)        | (0.98)  | (-2.76)   | (-0.85) |
| Hispanic                           | 0.046   | 0.005   | 0.15          | 0.074   | -0.138    | -0.172  |
|                                    | (0.57)  | (0.06)  | (1.33)        | (0.83)  | (-1.5)    | (-1.68) |
| Age                                | 1.801   | -0.894  | 1.428         | -1.303  | 2.265     | 0.058   |
|                                    | (2.27)  | (-1.1)  | (1.78)        | (-1.59) | (2.58)    | (0.06)  |
| Age squared                        | -0.052  | 0.026   | -0.04         | 0.04    | -0.07     | -0.003  |
|                                    | (-2.15) | (1.06)  | (-1.64)       | (1.62)  | (-2.63)   | (-0.09) |
| Substance use index                | 0.381   | 0.422   | 0.366         | 0.374   | 0.425     | 0.451   |
|                                    | (10.11) | (11.33) | (11.39)       | (10.89) | (11.59)   | (11.72) |
| Delinquency index                  | 0.016   | -0.002  | 0.04          | 0.014   | 0.051     | 0.072   |
|                                    | (0.68)  | (-0.05) | (2.01)        | (0.44)  | (2.32)    | (2.15)  |
| Constant                           | -16.444 | 6.123   | -13.944       | 8.343   | -19.901   | -2.229  |
|                                    | (-2.53) | (0.92)  | (-2.11)       | (1.23)  | (-2.76)   | (-0.3)  |

Notes: The dependent variable is an indicator of sexual activity. See text for further details. T-statistics are reported in brackets.

groups is relatively similar in both the PSM model and the DIDPSM model. For the PSM model, the distribution of the comparison group is skewed to the left; however, the

treatment distribution compensates for this by being relatively flat. The distributions of the control and treatment groups in the DIDPSM model are quite symmetric and thus conducive to a high quality of matching.

## 5.2 Estimates of the ATET

Tables 6 and 7 provide the LPM, LPM\*, PSM and DIDPSM estimates of the ATET for substance use on sexual intercourse and uncontracepted sexual intercourse, for males and females, respectively. The estimated coefficients in the LPM uniformly indicate a significant and positive association between reported substance use and adolescent sexual practices. Marijuana use has the strongest association with adolescent sexual practices; for males (females), it increases the likelihood of being sexually active by 30% (34%) as well as that of engaging in uncontracepted activity by 16% (16%).

Table 6: ATET estimated coefficients (males)

|   | DID             | LPM              | LPM*            | PSM              |                  |                  | DIDPSM positive treatment |                 |                 |
|---|-----------------|------------------|-----------------|------------------|------------------|------------------|---------------------------|-----------------|-----------------|
|   |                 |                  |                 | Attrnd           | Attr             | Atts             | Attrnd                    | Attr            | Atts            |
|   |                 |                  |                 |                  | (0.0001)         | (0.06)           |                           | (0.0001)        | (0.06)          |
| <b>Sexual intercourse</b>                   |                 |                  |                 |                  |                  |                  |                           |                 |                 |
| Consumed alcohol at all                     | 0.104<br>(4.54) | 0.235<br>(16.27) | 0.136<br>(8.73) | 0.15<br>(6.15)   | 0.22<br>(6.58)   | 0.12<br>(5.77)   | 0.102<br>(3.12)           | 0.12<br>(2.87)  | 0.089<br>(3.77) |
| Consumed five or more drinks                | 0.137<br>(5.72) | 0.279<br>(17.18) | 0.167<br>(9.60) | 0.136<br>(5.439) | 0.243<br>(7.552) | 0.153<br>(8.908) | 0.125<br>(3.24)           | 0.109<br>(2.36) | 0.115<br>(4.33) |
| Used Marijuana                              | 0.126<br>(4.55) | 0.299<br>(16.41) | 0.167<br>(8.55) | 0.186<br>(5.87)  | 0.282<br>(8.85)  | 0.167<br>(8.70)  | 0.072<br>(1.60)           | 0.171<br>(3.24) | 0.122<br>(3.80) |
| <b>Sexual intercourse w/o contraception</b> |                 |                  |                 |                  |                  |                  |                           |                 |                 |
| Consumed alcohol at all                     | 0.06<br>(3.51)  | 0.104<br>(10.60) | 0.066<br>(6.21) | 0.083<br>(5.57)  | 0.082<br>(5.17)  | 0.069<br>(4.17)  | 0.063<br>(2.25)           | 0.063<br>(2.22) | 0.054<br>(2.88) |
| Consumed five or more drinks                | 0.065<br>(3.49) | 0.139<br>(12.77) | 0.097<br>(8.11) | 0.11<br>(5.04)   | 0.111<br>(5.19)  | 0.98<br>(5.61)   | 0.061<br>(2.22)           | 0.04<br>(1.10)  | 0.05<br>(2.24)  |
| Used Marijuana                              | 0.106<br>(4.83) | 0.156<br>(12.76) | 0.106<br>(7.93) | 0.11<br>(4.40)   | 0.153<br>(7.45)  | 0.111<br>(6.18)  | 0.117<br>(2.66)           | 0.113<br>(2.62) | 0.11<br>(3.92)  |

Note: T-statistics are reported in brackets. Matching is performed over the area of common support. Attrnd denotes Nearest Neighbour matching. Attr denotes Radius matching, where the size of the radius is reported in parenthesis. Atts denotes Stratification matching. Attk denotes Kernel matching where the bandwidth is specified in parenthesis. For the LPM, we control for age, squared age, baptist affiliation, parental upbringing, housing, past hard times, living in the South, urban residence, public schooling, ethnicity, unemployment conditions, and missing observations. For the LPM\* we control for the same covariates as the LPM and in addition delinquency and past substance use. Definitions are provided in Appendix A.

Table 7: ATET estimated coefficients (females)

|   | DID             | LPM              | LPM*            | PSM             |                 | DIDPSM positive treatment |                 |                 |                   |                 |                 |
|---|-----------------|------------------|-----------------|-----------------|-----------------|---------------------------|-----------------|-----------------|-------------------|-----------------|-----------------|
|   |                 |                  | Attrnd          | Attr            | Atts            | Attrnd                    | Attr            | Attk            |                   |                 |                 |
|   |                 |                  | (0.0001)        | (0.0001)        | (0.06)          | (0.0001)                  | (0.0001)        | (0.06)          |                   |                 |                 |
| <b>Sexual intercourse</b>                   |                 |                  |                 |                 |                 |                           |                 |                 |                   |                 |                 |
| Consumed alcohol at all                     | 0.095<br>(4.23) | 0.217<br>(14.52) | 0.094<br>(5.95) | 0.082<br>(2.44) | 0.162<br>(5.37) | 0.072<br>(4.15)           | 0.093<br>(4.86) | 0.103<br>(2.75) | 0.068<br>(1.52)   | 0.091<br>(3.41) | 0.094<br>(3.25) |
| Consumed five or more drinks                | 0.052<br>(2.02) | 0.242<br>(12.93) | 0.107<br>(5.60) | 0.123<br>(4.30) | 0.228<br>(5.19) | 0.124<br>(5.45)           | 0.136<br>(6.22) | 0.034<br>(0.83) | 0.051<br>(1.51)   | 0.042<br>(1.50) | 0.042<br>(1.44) |
| Used marijuana                              | 0.06<br>(2.02)  | 0.336<br>(16.01) | 0.166<br>(7.56) | 0.165<br>(4.61) | 0.309<br>(6.65) | 0.172<br>(6.30)           | 0.19<br>(6.42)  | 0.057<br>(1.12) | 0.097<br>(1.86)   | 0.068<br>(1.95) | 0.07<br>(2.00)  |
| <b>Sexual intercourse w/o contraception</b> |                 |                  |                 |                 |                 |                           |                 |                 |                   |                 |                 |
| Consumed alcohol at all                     | 0.021<br>(1.16) | 0.076<br>(7.27)  | 0.016<br>(1.43) | 0.014<br>(0.68) | 0.035<br>(1.96) | 0.008<br>(0.54)           | 0.017<br>(1.13) | 0.031<br>(0.91) | -0.032<br>(-1.05) | 0.033<br>(1.24) | 0.034<br>(1.57) |
| Consumed five or more drinks                | 0.04<br>(1.83)  | 0.104<br>(7.89)  | 0.039<br>(2.90) | 0.053<br>(2.15) | 0.091<br>(3.49) | 0.043<br>(2.17)           | 0.047<br>(2.45) | 0.036<br>(0.96) | 0.002<br>(0.51)   | 0.039<br>(1.37) | 0.036<br>(1.11) |
| Used marijuana                              | 0.034<br>(1.36) | 0.159<br>(10.98) | 0.08<br>(5.11)  | 0.055<br>(2.22) | 0.117<br>(3.29) | 0.078<br>(3.57)           | 0.087<br>(3.58) | 0.034<br>(0.71) | 0.024<br>(0.66)   | 0.04<br>(1.30)  | 0.036<br>(1.24) |

Note: T-statistics are reported in brackets. Matching is performed over the area of common support. Attrnd denotes Nearest Neighbour matching. Attr denotes Radius matching, where the size of the radius is reported in parenthesis. Atts denotes Stratification matching. Attk denotes Kernel matching where the bandwidth is specified in parenthesis. For the LPM, we control for age, squared age, baptist affiliation, parental upbringing, housing, past hard times, living in the South, urban residence, public schooling, ethnicity, unemployment conditions, and missing observations. For the LPM\* we control for the same covariates as the LPM and in addition delinquency and past substance use. Definitions are provided in Appendix A.

For both males and females, heavy drinking has a stronger relationship with sexual practices than drinking per se. For instance, males (females) who drink heavily are 28% (24%) more likely to have sexual intercourse and 14% (10%) more likely to engage in uncontracepted activity, while drinking increases the likelihood that males (females) are sexually active by 24% (22%) and participate in uncontracepted sexual intercourse by 10% (8%).

The results from the LPM\* model suggest that, although most recent research has focused on the role of unobservable characteristics, important observable variables have also been omitted. After including indices which control for youths' previous involvement in risky behaviour, both the magnitudes and statistical significance of the coefficients were at least halved for the majority of the estimates for females. Dramatic differences are also evident in the results for males. For instance, in contrast to the results discussed above, the LPM\* model suggests that males (females) who drink and drink heavily increase the likelihood of sexual activity by only 14% (9%) and 17% (11%) respectively. The most substantial difference is found for the association between drinking and unprotected sexual intercourse for females; the estimate of ATET has dropped from 8% to 2% and has become insignificant.

Our results for the PSM, illustrated in columns 4-7 of Tables 6 and 7, are qualitatively similar to the LPM\*. The point estimates for the PSM are similar to the LPM\*, while the T-statistics for the PSM are a little lower (with the possible exception of the Radius matching specification in Column 5). Taken together, the results of the LPM\* and PSM suggest that it is important to adequately control for observable heterogeneity.

Columns 1 and 8-11 of Tables 6 and 7 present the results from the DID and DIDPSM models, respectively. The DIDPSM estimates do not differ substantially in size from the DID estimates. This is not surprising, given that we have observed the resemblance between the LPM\* estimates and the PSM estimates. However, in terms of significance, the DIDPSM estimates do differ from the DID estimates. In most cases, the standard errors of the DIDPSM estimates are larger than those of the DID estimates. This is expected due to differences between the two estimators. When imputing a missing potential outcome, the DIDPSM ensures that the support condition does not fail and uses the "matched observations". On the other hand, the DID estimator uses the conditional regression function to impute missing outcomes and consequently, the support condition is not necessarily met. This means that effectively the DID uses more observations than the DIDPSM by using potentially "unmatched" observations. Since we would like to ensure that the treated individual and non-treated individuals are actually comparable, the following discussion focuses exclusively on the DIDPSM estimates.

For males, the results from the DIDPSM indicate that even after controlling for unobservables, alcohol and marijuana use continue to have significant and positive effects on sexual practices. Males who became drinkers (heavy drinkers) over the period are 9-12% (11-13%) more likely to engage in sexual activity. Both drinking and heavy drinking increase the likelihood of unprotected sexual intercourse by 5-6%. Furthermore, marijuana use increases the likelihood of sexual intercourse (unprotected sexual intercourse) by 12-17% (11-12%). In each case, the estimates from at least three of the four matching algorithms are statistically significant at levels better than 1%.

For females, in contrast, we observe a very limited association between substance use

and sexual practices. While drinking increases the likelihood of sexual activity by 9-10%, neither heavy drinking nor marijuana use is statistically associated with sexual practices at levels of significance better than 5%. The results therefore suggest that much of the previously reported association between substance use and female sexual behaviour can be explained by positive correlation between substance use and unobservable relevant factors.

In this way, the effects of substance use substantially differ between males and females. This could be partly because the risks of sexual activity are greater for females (for example, the complications of an unwanted pregnancy). Even if females are relaxed and aroused by substance use, they may still deliberate the full consequences of their actions. As a policy implication, the prevention of commencing substance use has the potential to reduce sexual practices for males, while it appears to have limited potential for females.

### 5.3 Sex and Cigarettes: An Informal Test

We now conduct an informal test for the soundness of our estimation strategy by examining the causal effect of smoking on sex behaviour along the lines of Rashad and Kaestner (2004). An appropriate estimation strategy must control adequately for observables as well as for unobservables, thereby identifying no causal relationship between smoking and sexual behaviour.

We use the same set of controls and indicators of sexual activity as described in section 4. Table 8 provides summary statistics and definitions for the endogenous smoking variable and Table 9 presents the estimated relationship between smoking and sexual behaviour for the linear probability model, the PSM model, the DID model and the DIDPSM model. Estimates of the propensity scores are similar to those of other substances and are available from the authors on request.

Table 8: Smoking variable used in the specification test

|                                   | 1998  |         | 1999  |         |
|-----------------------------------|-------|---------|-------|---------|
|                                   | Males | Females | Males | Females |
| Smoked in past month <sup>a</sup> | 0.294 | 0.274   | 0.305 | 0.284   |
| Sample Size                       | 3399  | 3274    | 4162  | 4024    |

Notes: <sup>a</sup>Equal to 1 if adolescent reports smoking in the past 30 days; 0 otherwise.

Table 9: ATET estimated effect of smoking behaviour

|   | DID             | LPM              | LPM*            | PSM              |                  |                 | DIDPSM positive treatment |                 |                 |                 |
|---|-----------------|------------------|-----------------|------------------|------------------|-----------------|---------------------------|-----------------|-----------------|-----------------|
|   |                 |                  |                 | Attrnd           | Attr             | Attk            | Attrnd                    | Attr            | Attk            |                 |
|   |                 |                  |                 |                  | (0.0001)         | (0.06)          |                           | (0.0001)        | (0.06)          |                 |
| <b>Sexual intercourse</b>                   |                 |                  |                 |                  |                  |                 |                           |                 |                 |                 |
| Males                                       | 0.087<br>(3.38) | 0.258<br>(16.98) | 0.139<br>(8.14) | 0.169<br>(6.91)  | 0.217<br>(5.24)  | 0.138<br>(6.98) | 0.085<br>(1.95)           | 0.07<br>(1.54)  | 0.074<br>(3.21) | 0.086<br>(2.69) |
| Females                                     | 0.06<br>(2.24)  | 0.272<br>(16.93) | 0.113<br>(6.25) | 0.089<br>(2.73)  | 0.252<br>(9.027) | 0.113<br>(4.73) | 0.055<br>(1.38)           | 0.069<br>(1.33) | 0.048<br>(1.42) | 0.054<br>(1.83) |
| <b>Sexual intercourse w/o contraception</b> |                 |                  |                 |                  |                  |                 |                           |                 |                 |                 |
| Males                                       | 0.031<br>(154)  | 0.096<br>(9.28)  | 0.041<br>(3.54) | 0.059<br>(2.88)  | 0.067<br>(3.17)  | 0.055<br>(3.45) | 0.023<br>(0.66)           | 0.029<br>(0.95) | 0.036<br>(1.36) | 0.034<br>(1.42) |
| Females                                     | 0.045<br>(2.04) | 0.106<br>(9.38)  | 0.02<br>(1.89)  | 0.024<br>(1.109) | 0.096<br>(5.18)  | 0.026<br>(1.28) | 0.082<br>(1.73)           | 0.058<br>(1.53) | 0.041<br>(1.59) | 0.037<br>(1.36) |

Note: T-statistics are reported in brackets. Matching is performed over the area of common support. Attrnd denotes Nearest Neighbour matching. Attr denotes Radius matching, where the size of the radius is reported in parenthesis. Attk denotes Stratification matching. Attk denotes Kernel matching where the bandwidth is specified in parenthesis. For the LPM, we control for age, squared age, baptist affiliation, parental upbringing, housing, past hard times, living in the South, urban residence, public schooling, ethnicity, unemployment conditions, and missing observations. For the LPM\* we control for the same covariates as the LPM and in addition delinquency and past substance use. Definitions are provided in Appendix A.

The LPM reveals smoking to be a strong determinant of sexual activity. The ATET of smoking in the past month is positively associated with both sexual intercourse and unprotected sexual intercourse. In the LPM, the magnitudes of the coefficients are large; for instance, a male (female) who smokes is 26% (27%) and 10 (11%) more likely to have sexual intercourse and unprotected sexual intercourse, respectively. Under the LPM\* model, the magnitude of the coefficients becomes much smaller, underlining the importance of controlling for past substance use and delinquency.

Although the PSM has reduced the significance of the coefficients, it still fails to remove the association between smoking and sexual activity. This suggests that unobserved factors are an important determinant of both smoking cigarettes and sexual activity. The DIDPSM, which controls for both observable and unobservable relevant factors, has successfully removed the spurious association between smoking and sexual behaviour. For females the results universally report no statistical association between smoking and sexual practices. We also find smoking and males' uncontracepted sexual activity to have no statistical association. However, the DIDPSM estimates report a weak relationship between males who smoke and males who engage in sexual intercourse. The estimated relationship is not robust across different matching algorithms; of the four matching algorithms, only two are significant at better than 5% levels. Taken together, these results show that our estimation strategy reasonably removes the spurious correlation between smoking and adolescent sexual practices.

## 5.4 The Effect of Ceasing Substance Use on Sexual Behaviour

It is potentially important for policy makers to know whether reducing substance use amongst adolescents will have a direct effect on their sexual activity. This is a very specific policy question which is not necessarily answered by simply establishing a causal relationship between sexual activity and substance use. To target this policy issue specifically, we apply the DIDPSM model to determine whether or not a policy which curtails the number of adolescents using substances, would have an effect on adolescent sexual practices. To achieve this, we use for our comparison group those respondents who used substances in both 1998 and 1999; and for our treatment group, those respondents who used substances in 1998 but stopped using substances in 1999. We use the same controls and dependent variable as discussed in the previous DIDPSM analysis. We refer to this exercise as a *negative* treatment, as it measures the effect of specifically ceasing (negating) substance use. We refer to our previous specification as a *positive* treatment, as it effectively measures the effect of commencing substance use. Since we calculate the ATET, we are the first to estimate the direct effect of a policy which curtails substance use amongst substance users.

Table 10 provides the probit estimates of the propensity score for the DIDPSM. Note that this model differs from previous models with regard to the sign of the coefficients. A positive coefficient implies adolescents are more likely to stop using substances. The propensity scores perform slightly worse than in the PSM model, considerably better than the DIDPSM model with regard to drinking and heavy drinking, and slightly worse than the DISPSM for marijuana. Interestingly, it appears that different controls are important for explaining why adolescents start and stop using substances. For example, males who

live in the south, have parents who dropped out of high school before Year 10 or have been through hard times are considerably more likely to stop drinking alcohol over the past year. However, these variables are not significant determinants of whether a male started to drink over the same period. Further, males who attended public schools are significantly less likely to give up both drinking and heavy drinking over the period but no more likely to start drinking or drinking heavily. Similarly, females who live with their biological parents, have poorly educated parents or have a history of delinquency are considerably more likely to continue drinking but are no more likely to start drinking over the period studied.

## 5.5 Estimates of the ATET: Negative Treatment

The DIDPSM results using our negative treatment are provided in Tables 11 and 12 for males and females, respectively. After controlling for both observable and unobservable relevant factors we find that quitting substance use has a differential impact on males and females. The DIDPSM estimates unambiguously report an insignificant statistical association between females' substance use and sexual behaviour. This implies policy setters will have very limited success in attempts to curtail the sexual practices of females via policies which reduce the number of adolescent female substance users.

For males, the story is quite different. The results show that reducing alcohol consumption amongst adolescent males appear to have a direct negative effect on their sexual practices. For instance, an adolescent male who quits alcohol over the period is 8-11% percent less likely to have sexual intercourse and 7-9% less likely to have risky sexual intercourse; males who stopped drinking heavily reduced the likelihood of sexual intercourse and risky sexual intercourse by 10-12% and 12-14%, respectively. The statistical significance of this relationship is reasonably robust across matching algorithms; in each case, four of the five matching algorithms are statistically significant at a 5% level or better. This result is interesting from a policy perspective. Educational programs which encourage responsible use of alcohol will significantly reduce the incidences of both sexual intercourse and sexual intercourse without contraception. Indeed, policies that target adolescent male drinkers have the potential to reduce the number of males engaging in sexual practices.

From our results, it is unclear whether marijuana use has a statistical relationship with adolescent male sexual practices. For instance, two of the five matching algorithms suggest that, at better than a 5% level of significance, males who stop smoking marijuana are 11-12% less likely to have sexual intercourse and 11% less likely to have uncontracepted sexual intercourse. The other three are not significant at the 10% level. This result is anticipated as we expect adolescent males to respond differently to the effect of marijuana use; for some individuals it is likely to increase sexual desire and hamper judgment while others may become anti-social and withdrawn.

Again we follow the advice of Rashad and Kaestner (2004) and, in a similar fashion described in section 5.3, apply the informal sex and cigarettes test of the validity of our results. Our comparison group comprises individuals who reported smoking in both 1998 and 1999 and our treatment group consists of individuals who were smokers in 1998 but quit smoking in 1999. We use 1998 controls and take the difference in sexual activity over

Table 10: DIDPSM model - Negative treatment

|                                    | Drink   |         | Drink heavily |         | Marijuana |         |
|------------------------------------|---------|---------|---------------|---------|-----------|---------|
|                                    | Males   | Females | Males         | Females | Males     | Females |
| Unemployment < 3%                  | 0.045   | -0.33   | -0.366        | -0.031  | -0.074    | 0.235   |
|                                    | (0.44)  | (-0.32) | (-2.6)        | (-0.21) | (-0.52)   | (1.46)  |
| Unemployment 9% - 12%              | 0.068   | -0.414  | 0.204         | -0.298  | -0.349    | -0.31   |
|                                    | (0.26)  | (-1.03) | (0.57)        | (-0.52) | (-0.8)    | (-0.44) |
| Unemployment > 15%                 | 0.121   | 0.636   | 0.068         | 0.52    | -0.348    |         |
|                                    | (0.28)  | (1.3)   | (0.13)        | (0.66)  | (-0.49)   |         |
| Urban residence                    | -0.091  | -216    | -0.037        | 0.083   | -0.087    | 0.102   |
|                                    | (-0.92) | (-2.19) | (-0.29)       | (0.57)  | (-0.58)   | (0.58)  |
| Lives in the South                 | 0.285   | 0.065   | 0.159         | -0.022  | 0.219     | 0.018   |
|                                    | (3.04)  | (0.65)  | (1.31)        | (-0.15) | (1.57)    | (0.1)   |
| Baptist religious affiliation      | -0.053  | 0.339   | 0.081         | 0.216   | 0.054     | -0.167  |
|                                    | (-0.44) | (2.82)  | (0.52)        | (1.14)  | (0.32)    | (-0.75) |
| Lives with mum and dad             | -0.082  | -0.282  | -0.191        | -0.255  | -0.113    | -0.166  |
|                                    | (-0.94) | (-3.13) | (-1.71)       | (-1.97) | (1.05)    | (-1.07) |
| Strict parental figure             | 0.051   | -0.017  | 0.125         | -0.219  | -0.001    | 0.186   |
|                                    | (0.51)  | (0.17)  | (0.93)        | (-1.43) | (-0.01)   | (1.07)  |
| Parents dropped out of high school | 0.66    | -0.548  | 0.475         | -0.251  | 0.118     | -0.461  |
|                                    | (3.02)  | (-2.74) | (1.75)        | (-0.93) | (0.41)    | (-1.05) |
| Attended public school             | -0.286  | 0.055   | -0.343        | 0.123   | -0.126    | 0.136   |
|                                    | (-2.5)  | (0.45)  | (-2.31)       | (0.7)   | (-0.8)    | (0.7)   |
| Been through hard times            | 0.401   | 0.122   | 0.624         | 0.062   | 0.083     | 0.104   |
|                                    | (2.14)  | (0.59)  | (2.86)        | (0.22)  | (0.3)     | (0.32)  |
| White                              | -0.438  | -0.254  | -0.416        | -0.366  | -0.193    | -0.346  |
|                                    | (-4.49) | (-2.52) | (-3.23)       | (-2.44) | (-1.39)   | (-1.91) |
| Hispanic                           | 0.003   | -0.01   | 0.059         | 0.126   | 0.22      | -0.177  |
|                                    | (0.03)  | (-0.08) | (0.41)        | (0.75)  | (1.4)     | (-0.81) |
| Age                                | -0.149  | 1.884   | -0.372        | 0.063   | 1.167     | -0.288  |
|                                    | (-0.14) | (1.75)  | (-0.27)       | (0.04)  | (-1.07)   | (-0.14) |
| Age squared                        | 0       | -0.061  | 0.007         | -0.004  | -0.093    | 0.005   |
|                                    | (0)     | (-1.88) | (0.17)        | (-0.1)  | (-3.62)   | (0.08)  |
| Substance use index                | -0.169  | -0.23   | -0.23         | -0.195  | -0.116    | -0.338  |
|                                    | (-3.06) | (-4.13) | (-3.07)       | (-2.3)  | (-1.07)   | (-2.1)  |
| Delinquency index                  | -0.022  | -0.083  | -0.021        | -0.077  | -0.093    | -0.07   |
|                                    | (0.99)  | (-2.33) | (-0.78)       | (-1.93) | (-3.62)   | (-1.72) |
| Constant                           | 2.097   | 13.355  | 4.469         | 1.033   | -8.935    | 4.64    |
|                                    | (0.23)  | (-1.5)  | (0.38)        | (0.08)  | (-0.65)   | (0.27)  |

Notes: The dependent variable is an indicator of sexual activity. See text for further details. T-statistics are reported in brackets.

the two years as our dependent variable. The probit model for the propensity scores of smoking performs similarly to those of other substances; the estimates are available on

Table 11: ATET estimates for the DIDPSM model (males)

|   | DID               | Attnd             | DID PSM (quit) treatment |                   |                   |                   |
|---|-------------------|-------------------|--------------------------|-------------------|-------------------|-------------------|
|   |                   |                   | Attr<br>(0.01)           | Atts              | Attk<br>(0.06)    | Attk<br>(0.01)    |
| <b>Sexual intercourse</b>                   |                   |                   |                          |                   |                   |                   |
| Consumed alcohol at all                     | -0.097<br>(-2.86) | -0.08<br>(-1.74)  | -0.101<br>(-2.90)        | -0.099<br>(-2.66) | -0.096<br>(-2.78) | -0.11<br>(-2.85)  |
| Consumed five or more drinks                | -0.101<br>(-2.31) | -0.114<br>(-1.68) | -0.122<br>(-2.55)        | -0.104<br>(-2.18) | -0.105<br>(-2.26) | -0.126<br>(-2.38) |
| Used Marijuana                              | -0.053<br>(-1.17) | -0.016<br>(-0.31) | -0.059<br>(-1.20)        | -0.039<br>(-0.83) | -0.039<br>(-0.78) | -0.04<br>(-0.86)  |
| <b>Sexual intercourse w/o contraception</b> |                   |                   |                          |                   |                   |                   |
| Consumed alcohol at all                     | -0.076<br>(-2.33) | -0.072<br>(-1.75) | -0.071<br>(-2.28)        | -0.086<br>(-2.37) | -0.08<br>(-2.51)  | -0.08<br>(-2.79)  |
| Consumed five or more drinks                | -0.134<br>(-3.03) | -0.045<br>(-0.69) | -0.139<br>(-2.71)        | -0.116<br>(-2.35) | -0.123<br>(-2.76) | -0.117<br>(-2.22) |
| Used Marijuana                              | -0.069<br>(-1.48) | -0.064<br>(-0.85) | -0.073<br>(-1.42)        | -0.126<br>(-2.45) | -0.115<br>(-2.24) | -0.076<br>(-1.46) |

Note: T-statistics are reported in brackets. Matching is performed over the area of common support. Attnd denotes Nearest Neighbour matching. Attr denotes Radius matching, where the size of the radius is reported in parenthesis. Atts denotes the Stratification matching. Attk denotes the Kernel matching where the bandwidth is specified in parenthesis. For DIDPSM, in this sample, we increased the radius of the matching estimator from 0.0001 to 0.01. At a radius of 0.0001 there were not sufficient controls to match to the treatment observation, all ATET estimates had t-statistics close to 0. The estimates using a radius of 0.0001 are available on request.

request. The DIDPSM estimates of the ATET are presented in Table 13. By balancing observable relevant variables across treatment and control groups and controlling for unobserved heterogeneity, we have successfully eliminated the spurious correlation between adolescents' sexual activities and smoking. This is universally true for all matching estimates for both males and females. As such we are confident that the ATET estimates provided in this section represent the causal effect of ceasing substance use amongst adolescents.

## 6 Concluding Remarks

While the correlation between adolescent substance use and sexual activity is well established, teasing out the causal relationship is a more challenging task. To deal with these challenges, we have adopted a methodology with three primary novel aspects relative to the literature. First, we exploit a panel data setting to control for unobserved heterogeneity through a difference in difference estimator. Second, we carefully control for observed heterogeneity by both exploiting a more complete set of control variables than has previously been adopted and by combining propensity score matching methods

Table 12: ATET estimates for the DIDPSM model (females)

|   | DID     | Attnd   | DID PSM (quit) treatment |         |                |                |
|---|---------|---------|--------------------------|---------|----------------|----------------|
|   |         |         | Attr<br>(0.01)           | Atts    | Attk<br>(0.06) | Attk<br>(0.01) |
| <b>Sexual intercourse</b>                   |         |         |                          |         |                |                |
| Consumed alcohol at all                     | 0.017   | 0.06    | -0.004                   | 0.025   | 0.026          | 0.027          |
|   | -0.54   | -1.37   | (-0.13)                  | -0.72   | -0.97          | -0.79          |
| Consumed five or more drinks                | -0.023  | -0.045  | -0.047                   | -0.028  | -0.027         | -0.044         |
|   | (-0.54) | (-0.71) | (-0.04)                  | (-0.51) | (-0.53)        | (-0.87)        |
| Used Marijuana                              | -0.032  | -0.019  | -0.031                   | -0.028  | -0.02          | -0.069         |
|   | (-0.60) | (-0.25) | (-0.47)                  | (-0.51) | (-0.34)        | (-1.04)        |
| <b>Sexual intercourse w/o contraception</b> |         |         |                          |         |                |                |
| Consumed alcohol at all                     | -0.053  | -0.059  | -0.024                   | -0.055  | -0.051         | -0.062         |
|   | (-1.64) | (-1.45) | (-0.85)                  | (-1.66) | (-1.81)        | (-1.85)        |
| Consumed more than five drinks              | -0.034  | -0.036  | -0.023                   | -0.032  | -0.036         | -0.011         |
|   | (-0.62) | (-0.59) | (-0.43)                  | (-0.62) | (-0.71)        | (-0.19)        |
| Used Marijuana                              | -0.094  | -0.082  | -0.087                   | -0.047  | -0.068         | -0.093         |
|   | (-1.44) | (-0.81) | (-1.31)                  | (-0.58) | (-0.88)        | (-1.39)        |

Note: T-statistics are reported in brackets. Matching is performed over the area of common support. Attnd denotes Nearest Neighbour matching. Attr denotes Radius matching, where the size of the radius is reported in parenthesis. Atts denotes the Stratification matching. Attk denotes the Kernel matching where the bandwidth is specified in parenthesis. For DIDPSM, in this sample, we increased the radius of the matching estimator from 0.0001 to 0.01. At a radius of 0.0001 there was not sufficient controls to match to the treatment observation, all ATET estimates had t-statistics close to 0. The estimates using a radius of 0.0001 are available from the authors on request.

with our difference in difference estimator. Third, we make a distinction between positive and negative treatments, allowing us to differentiate the likely effect of policy on current substance users and non-users.

Our study demonstrates the importance of carefully controlling for both unobserved and observed heterogeneity. To our knowledge, the DIDPSM estimator that we adopt is the first in the literature to pass the informal test proposed by Rashad and Kaestner (2004). Adopting a simple linear probability model, we found a strong spurious causal link between smoking and sexual activity. Incorporating our more complete set of control variables (in particular, the past substance use index) ameliorates this problem by more completely controlling for observable elements of the adolescent’s background. The results for our PSM model are in the same vein; by more completely controlling for observable characteristics, the spurious causal link is dampened. Our results using the DID estimator demonstrate that controlling for unobserved heterogeneity is important; in many instances the spurious causal link is removed. However, it is only when we adopt our DIDPSM estimator that the causal association is satisfactorily removed and we can pass the test proposed by Rashad and Kaestner (2004).

After controlling for observed and unobserved heterogeneity with our DIDPSM esti-

Table 13: ATET estimated effect of smoking - negative treatment

|   | DID PSM negative treatment |                   |                   |                   |                   |                   |                   |
|---|----------------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
|   | DID                        | Attnd             | Attr<br>(0.0001)  | Attr<br>(0.01)    | Atts              | Attk<br>(0.06)    | Attk<br>(0.01)    |
| <b>Sexual intercourse</b>                   |                            |                   |                   |                   |                   |                   |                   |
| Males                                       | -0.06<br>(-1.52)           | -0.073<br>(-1.26) | -0.039<br>(-0.34) | -0.068<br>(-1.82) | -0.065<br>(-1.52) | -0.063<br>(-1.15) | -0.067<br>(-1.44) |
| Females                                     | 0.029<br>(0.74)            | -0.011<br>(-0.18) | 0.019<br>(0.17)   | 0.041<br>(1.06)   | 0.012<br>(0.27)   | 0.028<br>(0.73)   | 0.015<br>(0.34)   |
| <b>Sexual intercourse w/o contraception</b> |                            |                   |                   |                   |                   |                   |                   |
| Males                                       | -0.047<br>(-1.21)          | -0.027<br>(-0.48) | -0.036<br>(-0.92) | -0.042<br>(-1.19) | -0.051<br>(-1.35) | -0.047<br>(-1.22) | -0.041<br>(-1.08) |
| Females                                     | 0.0183<br>(0.44)           | 0.072<br>(1.22)   | 0.1<br>(0.91)     | 0.054<br>(1.62)   | 0.019<br>(0.36)   | 0.024<br>(0.66)   | 0.014<br>(0.38)   |

Note: T-statistics are reported in brackets. Matching is performed over the area of common support. Attnd denotes Nearest Neighbour matching. Attr denotes Radius matching, where the size of the radius is reported in parenthesis. Atts denotes the Stratification matching. Attk denotes the Kernel matching where the bandwidth is specified in parenthesis. For DIDPSM, in this sample, we increased the radius of the matching estimator from 0.0001 to 0.01. At a radius of 0.0001 there was not sufficient controls to match to the treatment observation, all ATET estimates had t-statistics close to 0. The estimates using a radius of 0.0001 are available from the authors on request.

mator, we find that the causal link between adolescent substance use and sexual activity is more tenuous than reported in the literature. Our positive treatment measures the effect of initiating substance use on sexual activity. For males, we find some evidence of a causal link for both alcohol and marijuana on sexual activity using this treatment. For females, the evidence is mixed; a causal link is evident for a subset of our matching specifications in our DIDPSM estimator. Our negative treatment considers the effect of ceasing substance use on sexual activity. For males, we find a causal effect on sexual activity for alcohol but not marijuana. For females, we identify no causal links.

## References

- [1] Altonji, J. G., Elder, T. E., Taber, C. R., 2005. Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools. *Journal of Political Economy* 113(1), 151-184.
- [2] Blundell, R., Costa Dias, M., 2000. Evaluation methods for non-experimental data. *Fiscal Studies* 21(4), 427-468.
- [3] Caliendo, M., Kopeinig, S., 2005. Some practical guidance for the implementation of propensity score matching. IZA Discussion Paper No. 1588.

- [4] Cooper, M. L., Peirce, R. S., Huselid, R. F., 1994. Substance use and sexual risk taking among black adolescents and white adolescents. *Health Psychology* 13(3), 251-262.
- [5] Dehejia, R. H., Wahba, S., 1999. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94(448), 1053-1062.
- [6] Dehejia R. H., Wahba, S., 2002. Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics* 84(1), 151-161.
- [7] DiNardo, J., Tobias, J. L., 2001. Nonparametric density and regression estimation. *Journal of Economic Perspectives* 15(4), 11-28.
- [8] Donovan, C., McEwan, R., 1995. A review of the literature examining the relationship between alcohol use and HIV-related sexual risk-taking in young people. *Addiction* 90, 319-328.
- [9] Elliott, D. S., Morse, B. J., 1989. Delinquency and drug use as risk factors in teenage sexual activity. *Youth and Society* 21(1), 32-60.
- [10] Fergusson, D. M., Lynskey, M. T., 1996. Alcohol misuse and adolescent sexual behaviors and risk taking. *Pediatrics* 98, 91-96.
- [11] Graves, K. L., Leigh, B. C., 1995. The relationship of substance use to sexual activity among young adults in the United States. *Family Planning Perspectives* 27(1), 18-33.
- [12] Grossman, M. and Markowitz, S., 2005, I did what last night?!!! Adolescent risky sexual behaviors and substance use. *Eastern Economic Journal* 31(3), 383-405.
- [13] Grossman, M., Kaestner, R., Markowitz, S., 2004. Get high and get stupid: the effect of alcohol and marijuana use on teen sexual behavior. *Review of Economics of the Household* 2(4), 413-441.
- [14] Harvey, S. M., Spigner, C., 1995. Factors associated with sexual behavior among adolescents: a multivariate analysis. *Adolescence* 30(118), 253-264.
- [15] Heckman, J. J., Ichimura, H., Todd, P. E., 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *The Review of Economic Studies* 64(4), 605-654.
- [16] Heckman, J. J., Ichimura, H., Todd, P. E., 1998. Matching as an econometric evaluation estimator. *The Review of Economic Studies* 65(2), 261-294.
- [17] Hingson, R. W., Strunin, L., Berlin, B. M., Heeren, T., 1990. Beliefs about AIDS, use of alcohol and drugs, and unprotected sex among Massachusetts adolescents. *American Journal of Public Health* 80(3), 295-299.

- [18] Laumann, E. O., Gagnon, J. H., Michael, R. T., Michaels, S., 1994. The social organization of sexuality: sexual practices in the United States. University of Chicago Press, Chicago.
- [19] Leigh, B. C., Stall, R., 1993. Substance use and risky sexual behavior for exposure to HIV: issues in methodology, interpretation, and prevention. *The American Psychologist* 48(10), 1035-1046.
- [20] Lowry, R., Holtzman, D., Truman, B. I., Kann, L., Collins, J. L., Kolbe, L. J., 1994. Substance use and HIV-related sexual behaviors among US high school students: are they related? *American Journal of Public Health* 84(7), 1116-1120.
- [21] Morrison, T. C., Diclemente, R. J., Wingood, G. M., Collins, C., 1998. Frequency of alcohol use and its association with STD/HIV-related risk practices, attitudes and knowledge among an AfricanAmerican community-recruited sample. *International Journal of STD & AIDS*, 9(10), 608-612.
- [22] Pagan, A. R., Ullah, A., 1999. *Nonparametric Econometrics*. Cambridge University Press, Cambridge.
- [23] Rashad, I., Kaestner, R., 2004. Teenage sex, drugs and alcohol use: problems identifying the cause of risky behaviors. *Journal of Health Economics* 23(3), 493-503.
- [24] Rees, D. I., Argys, L. M., Averett, S. L., 2001. New evidence on the relationship between substance use and adolescent sexual behavior. *Journal of Health Economics* 20(5), 835-845.
- [25] Rosenbaum, E., Kandel, D. B., 1990. Early onset of adolescent sexual behavior and drug involvement. *Journal of Marriage and the Family* 52(3), 783-798.
- [26] Rosenbaum, P. R., Rubin, D. B., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41-55.
- [27] Sen, B., 2002. Does alcohol-use increase the risk of sexual intercourse among adolescents? Evidence from the NLSY97. *Journal of Health Economics* 21(6), 1085-1093.
- [28] Silverman, B. W., Silverman, S., 1986. *Density Estimation for Statistics and Data Analysis*. CRC Press, Florida.
- [29] Smith, J. A., Todd, P. E., 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(2), 305-353.
- [30] Strunin, L., Hingson, R. W., 1992. Alcohol, drugs, and adolescent sexual behavior. *International Journal of Addictions* 27, 129-146.
- [31] Wooldridge, J. M., 2002. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, Massachusetts.

## Appendix A: Definitions of observable variables

| Observable                              | Definition   |
|---|--|
| Age                                     | The age of the respondent as of December 31, 1996                |
| Parents' highest educ                   | The number of years of education of respondent's parents         |
| Substance use                           | Index of the respondent's previous substance use <sup>a</sup>    |
| Delinquency                             | Index of the respondent's history of delinquent behaviour        |
| <i>Binary variables<sup>b</sup></i>     |  |
| Baptist                                 | The respondent is of Baptist religious affiliation               |
| Roman Catholic                          | The respondent is of Roman Catholic religious affiliation        |
| Parents strict                          | The respondent reports having a strict parental figure           |
| House                                   | The respondent's dwelling is a house                             |
| Apartment                               | The respondent's dwelling is an apartment                        |
| Hard times                              | The respondent has experienced "hard times" (eg homeless living) |
| Lives in South                          | The respondent lives in the south                                |
| Urban residence                         | The respondent lives in an urban area                            |
| Lives with parents                      | The respondent lives with both of their biological parents       |
| Public school                           | The respondent attended a public high school                     |
| White                                   | The respondent is of White ethnicity                             |
| Black                                   | The respondent is of Black ethnicity                             |
| Hispanic                                | The respondent is of Hispanic ethnicity                          |
| Mixed race                              | The respondent is of mixed race ethnicity                        |
| Parental education                      | The respondent's parents did not complete Year 10 high school    |
| <i>Unemployment rate<sup>c</sup></i>    |  |
| Less than 3%                            | The respondent lives in an area with unemployment below 3%       |
| Between 3% and 6%                       | The respondent lives in an area with unemployment 3 - 6%         |
| Between 6% and 9%                       | The respondent lives in an area with unemployment 6 - 9%         |
| Between 9% and 12%                      | The respondent lives in an area with unemployment 9 - 12%        |
| Between 12% and 15%                     | The respondent lives in an area with unemployment 12 - 15%       |
| Greater than 15%                        | The respondent lives in an area with unemployment above 15%      |
| <i>Structural variables<sup>d</sup></i> |  |
| Urban missing                           | The respondent lives in an urban area                            |
| Stict missing                           | The respondent reports having a strict parental figure           |
| Parent missing                          | The respondent lives with both of their biological parents       |
| South missing                           | The respondent lives in the south                                |

Notes: <sup>a</sup> The index weights a value of 1 to each positive response to the question "have you ever used alcohol (marijuana) (cigarettes)".

<sup>b</sup> Binary variables take the value 1 if the listed condition is satisfied, and 0 otherwise.

<sup>c</sup> The unemployment rate is in the range indicated (binary variable).

<sup>d</sup> These variables were created where there was missing data on the specified control variable. As they have no economic interpretation, estimates of the propensity scores for these variables were not reported in the main body of the paper. Estimates are available from the authors on request.