Marijuana on Main Street?
Estimating Demand in Markets with Limited Access

Liana Jacobi and Michelle Sovinsky*

December 15, 2015

Abstract

Marijuana is the most common illicit drug with vocal advocates for legalization. Among other things, legalization would increase access and remove the stigma of illegality. Our model disentangles the role of access from preferences and shows that selection into access is not random. We find that traditional demand estimates are biased resulting in incorrect policy conclusions. If marijuana were legalized those under 30 would see modest increases in use of 28%, while on average use would increase by 48% (to 19.4%).Tax policies are effective at curbing use, where Australia could raise a billion (and the US $12 billion).

JEL Classification: L15, K42, H2

Keywords: legalization, drugs, discrete choice models, limited information, selection correction

Marijuana is the most widely used illicit drug in the world (ONDCP, 2004). According to the United Nations Office of Drugs and Crimes (2012), there are 119 to 224 million users worldwide. While the nature of the market makes it difficult to determine total sales with certainty, estimates indicate sales in the United States are around $150 billion per year (Miron, 2005). Despite the attempts to regulate use, in nearly every country, the market for illicit drugs remains pervasive.

The marijuana market has the most vocal advocates for legalization of all illicit drugs. Within Europe: Germany, the Netherlands, Portugal, Spain and Switzerland all currently exhibit liberal attitudes of law enforcement towards marijuana possession. The United States has a more punitive system, but in late 2012 recreational use of marijuana became legal in Washington and Colorado and in 2014 Oregon and Alaska joined them. Uruguay became the first nation to legalize marijuana in December of 2013. In Australia, there have been

[0] *Liana Jacobi is at the University of Melbourne, Department of Economics, 111 Barry Street Carlton VIC 3053, Australia, ljacobi@unimelb.edu.au; Michelle Sovinsky is at the the University of Mannheim, Department of Economics, L3, 5-7 Mannheim, Germany and CEPR, michelle.sovinsky@gmail.com. Corresponding author is Sovinsky. We are grateful to Peter Arcidiacono, Gerard van den Berg, Eve Caroli, Sofronis Clerides, Alon Eizenberg, Rosa Ferrer, T. Head, Carlos Noton, Alison Ritter, Terry Sovinsky, and seminar participants at Adelaide, Bocconi, Duke, Rotman School of Business, Cyprus, Dauphine (Paris), Haifa, Hebrew, Mannheim, Monash, Sydney, University Technology Sydney, Virginia, Warwick, the Barcelona Summer Forum, the Cannabis Policy Conference (Melbourne), the Conference for Mental Health and Well-Being (Melbourne), CEPR (Athens) and the IIOC (Washington DC) meetings for helpful comments and suggestions.

many campaigns in the larger cities to legalize marijuana. Indeed, for the past 30 years there has been a debate regarding marijuana legalization in many countries.[1] Those in favor of legalization cite the harsh consequences a criminal record can have for young users who are otherwise law-abiding citizens, the costs of black-market violence, the exposure to harder drugs from dealer interactions, the high expenditures on enforcement, and the foregone sales tax revenues. Those opposed are concerned about the impact on health outcomes and that legalization could result in lower prices, hence generating higher use. This is of particular concern if use among young adults increases and marijuana usage serves as a "gateway" to subsequent consumption of other harder drugs (DeSimone, 1998; Van Ours, 2003; Bretteville-Jensen and Jacobi, 2011).

Much of the discussion surrounding marijuana drug policy is concerned with the following questions. First, by how much would the prevalence and intensity of use rise under legalization? Second, to what extent would at risk groups (such as youth) be impacted by legalization? Finally, could government policies (such as taxation) be effective in curbing use? In this paper, we provide a methodology for examining the consequences of legalizing illicit drugs, which helps lead us to answers to these questions.

During the last two decades there have been many empirical studies that assess the impact of decriminalization on marijuana use. These include Miron and Zwiebel (1995), Pacula, et. al. (2000), Clements and Zhao (2009), Pacula, et. al. (2010), Pudney (2010), Caulkins, et. al. (2011), Donohue, Ewing, and Peloquin (2011), and Williams, et. al. (2011).[2] However, decriminalization and legalization differ in significant ways. The first important way concerns limited accessibility. Given that illicit drugs are not as easy to find as legal products, one can argue that non-users have very little information about how to get marijuana, which is the first step to becoming a user. Under decriminalization it is still necessary to seek out suppliers in order to purchase the drug. If marijuana were legalized, purchasing it would be as difficult as purchasing cigarettes or alcohol. Second, while decriminalization removes criminal penalties, using the drug is still illegal. In fact, in the Australian National Drug Household Survey, a significant fraction of non-users report not using marijuana because it is illegal. Legalization would obviously remove this stigma

[1] Pacula, et. al. (2010) provides a literature review.

[2] There is also a large literature on drug policies and the effect of decriminalization or enforcement on crime; see, for example, Adda, et. al. (2011) and Sickles and Taubman (1991).

(and cost) associated with illegal behavior, which may result in use among some current non-users. The third way in which decriminalization and legalization differ concerns the impact on dealers. Decriminalization makes it less costly for potential users in that they face a fine for using the drug instead of the harsher cost of a criminal punishment. In contrast, selling the drug is still illegal and hence dealers, should they be arrested, incur the same penalties regardless of the decriminalization status of the state. In other words, decriminalization does not impact the costs (broadly defined to include the risk of criminal prosecution) faced by dealers, while legalization eliminates the risk of arrest leading to lower costs. For these reasons, models that focus on the impact of decriminalization will not provide us with answers to what will happen to use under legalization.

This paper provides the first approach to modeling and estimating the impact of legalization on use. To do so we explicitly consider the role played by accessibility, the impact of illegal actions on utility, as well as the impact on the supply side. We present a model of consumer behavior that includes the impact of illegal behavior on utility and the impact of limited accessibility (either knowing where to buy or being offered an illicit drug) on using marijuana. We apply the model to data from the Australian National Drug Strategy Household Survey.[3] These data are particularly suited for our purposes as they contain information both on use and also on access and enable us to identify the preference parameters on marijuana use. For example, we obtain estimates for price elasticities of demand (for an illicit good) taking into account selection into access. Modeling both of these effects is particularly important for drawing correct inferences about choices that individuals would make under a policy of legalization, where the accessibility issue would essentially disappear. We find that predictions based on a model that does not consider selection are biased due to ignoring the important role that selection based on observables and unobservables plays in the context of marijuana use and more generally in the use of illicit drugs. This is the first paper to estimate demand for an illicit drug that considers selection into access.

Our modelling framework also directly addresses an issue that is prevalent in studies of illicit markets: the fact that prices are not observed for each purchase. We construct an empirical price distribution by exploiting prevalences on the type of marijuana used (i.e., leaf, head), based on individual-level use data and market-level price data, to obtain implied

[3] Several studies use these data to examine issues related to marijuana, such as Damrongplasit, et. al. (2010) and Williams (2004).

prices faced by users and non-users. This allows us to estimate a model with individual prices while not observing these in the data.

We use the demand side estimates to conduct counterfactuals on how use would change under legalization, how effective government policies would be at curbing use, and what tax revenues could be raised under legalization. We also consider differences across age groups (including teenagers) and conduct counterfactuals of how much taxes would need to be imposed to return the probability of underage use to what it was before legalization (at the individual level). The counterfactual analysis is implemented under different post-legalization prices to allow for different supply side scenarios.

We find that selection into who has access to marijuana is not random, and the results suggest estimates of the demand curve will be biased unless selection is explicitly considered. Our results indicate that if marijuana were legalized in Australia and accessibility were not an issue the probability of use would increase by almost 50% to 19.4%. Obviously there would be an impact on prices due to the law change, and the results show taxes of 25% are effective to offset the increase in use due to the legal status change. The overall probability of use would be 40% higher than current levels (at 18.3%). Individuals under 30 would see a more modest increase in the probability of use of 28% on average, while the the average probability of underage use would increase by 34% (to 33.7% from 25.1%). Our results suggest legalization in Australia could raise as much as a billion in taxes. For a population the size of the US our results indicate that a policy of marijuana legalization would raise approximately $12 billion.

We also predict tax revenues for (a state the size of) Colorado that opened it's first retail outlets in 2014 and implemented a 25% tax which corresponds to one of the tax increases we conduct in our counterfactuals. This allows us to compare realized tax revenues in Colorado to the predicted tax revenues generated from our counterfactual model (for a population the size of Colorado). Our estimates indicate (a population the size of) Colorado would collect approximately $68.2 million annually in tax revenues. After we account for the estimated losses to the black market, we predict tax collections of around $61.5 million on average. Excluding medical marijuana and licensing fees, Colorado collected $56.1 million in taxes in 2014. Hence, our mid-range tax predictions are within 10% of the realized tax revenue that was generated in 2014 in Colorado.[4]

[4] Our results do not include taxes from medical marijuana sales or from licensing fees, but in Colorado's

Finally, we find that the average price per gram would have to be \$158, about four times the current level, in order to experience no rise in use among underage users in a post-legalized world. Increasing prices by four-fold is not feasible given that we would expect most users to resort to the black market. However, a tax increase of 25% is sufficient to realize 34% of the goal (where two-thirds of these individuals are female).

The previous literature on decriminalization already mentioned is not concerned with the impact of limited access on consumption decisions. In this sense, the approach presented in our paper is conceptually more closely related to the empirical IO literature that examines markets with limited consumer information. These include papers by Sovinsky Goeree (2008), Ching, Erdem and Keane (2009), Ching and Hayashi (2010), Clerides and Courty (2010), Kim, Albuquerque, and Bronnenberg (2010), and Eliaz and Spiegler (2011). Our estimation methodology corrects for sample selection in the tradition of Heckman (1979). In addition, there is a small but growing literature addressing sample selection in empirical IO including Veugelers and Cassiman (2005) and Eizenberg (2011). We employ the Bayesian estimation approach as it allows us to address the issue of unobserved prices in a novel way in addition to being well suited to deal with discrete response variables and the resulting complex likelihood structure. It also provides a natural framework for the counterfactual policy analyses.

The paper is structured as follows. Section 1 gives an overview of the data. Sections 2 and 3 outline the model and the estimation technique. We discuss our parameter estimates in Section 4. We present results of counterfactual policy experiments in Section 5. We examine the robustness of our results to alternative specifications in Section 6. Finally, we conclude and discuss directions for future work in Section 7.

1 Data

Cannabis comes in a variety of forms and potency levels. The herbal form consists of the dried flowering tops, leaves and stalks of the plant. The resinous form consists of the resin secreted from the plant and resin oil. In this paper we focus on the most commonly used forms of cannabis: the leaf of the plant, the flowering tops (or head) of the plant, and a high potency form selectively bred from certain species (sinsemilla, called skunk). The leaf, head,

case these made up a small portion of overall tax raised - including these fees Colorado raised \$67.5 milllion in taxes in 2014. See https://www.colorado.gov/pacific/revenue/colorado-marijuana-tax-data and skunk are collectively known as marijuana.[5]

We use data from two primary sources. The first are individual-level cross-section data from the Australian National Drug Strategy Household Survey (NDSHS). The NDSHS was designed to determine the extent of drug use among the non-institutionalized civilian Australian population aged 14 and older.[6] About 20,000 (different) individuals are surveyed every 2 or 3 years from all Australian states/territories. We use data from three waves: 2001, 2004, and 2007. These data are particularly useful as they not only contain demographic, market, and illicit drug use information, but they also contain a number of variables on accessibility to marijuana. These latter questions are crucial in order to estimate our model. The second primary source are market-level pricing data collected from drug seizures by the Australian Bureau of Criminal Intelligence. We discuss these data in more detail in the remainder of the section.

1.1 Marijuana Use

We present descriptive statistics from the NDSHS data in Table 1. We restrict the data to individuals aged between 16 and 60. The average age of a respondent in our sample is just under 40. Approximately 43% are male and 2% of the sample are of Aboriginal descent. About 60% of the sample live in a major city. We construct an indicator variable equal to one if individuals report their health status is good, very good, or excellent. About 56% of individuals report being in good or better health.[7] The majority of the sample have earned a trade degree or reached a higher level of education.

The second panel presents information about marijuana use. Nearly half of the population has tried marijuana at least once in their lifetime, where the average age of onset is 19. In every year the survey asks "Have you used marijuana in the last 12 months?" We use this question to construct our binary response variable on marijuana use for our analysis of the

[5] We do not consider hashish (the resin or resin oil of the plant) as these forms are much harder to obtain and have a much higher level of the psychoactive component.

[6] Households were selected in a multi-stage stratified area sample design in order to provide a random sample of households within each geographical stratum. For the 2001, 2004 and 2007 surveys the self-completion drop-and-collect method was used (in about 85% of the cases) and computer-assisted telephone interviews for the remaining cases. Respondents were requested to indicate their level of drug use. Responses were sealed so the interviewer did not know the answers. If collection was not possible a reply-paid pre-addressed envelope was provided.

[7] Our measure of health status is the self-reported answer to "Would you say your health is: 1=excellent; 2=very good; 3=good; 4=fair; 5=poor."

extensive margin. In 2001 just over 15% reported using marijuana in the past year, but this declined to around 11% by 2007. Although the rates of marijuana use are considerable, most people who use marijuana do not use on a daily basis. Those that report they use marijuana daily or habitually is around 3%. We also define an ordered variable of cannabis use for an analysis of the intensive margin employed in our estimation of tax revenues. Based on a question on frequency of use, we classify users as infrequent if they use only quarterly, biannually or annually (about 6% of the sample) and as frequent if they use more often (about 8% of the sample). We should note that hard core drug users are less likely to return the survey or to be available for a telephone survey. Hence, our study will reflect mostly recreational users.

	Year		
	2001	2004	2007
Demographics			
Male	43%	42%	42%
Age	38	39	40
Aboriginal Descent	2%	2%	2%
Live in City	62%	60%	59%
In Good, Very Good, or Excellent Health	57%	54%	58%
High School Education	16%	15%	14%
Trade Degree	36%	35%	37%
University Degree	22%	25%	28%
Marijuana Use			
Ever Used	43%	43%	44%
Used in Last 12 Months	15%	14%	11%
Use Infrequently (Quarterly, Biannually or Annually)	6%	6%	4%
Use Frequently (Monthly, Weekly or Daily)	9%	8%	7%
Report Use as a Habit	3%	3%	2%
Use Daily	3%	2%	2%
Illegality Reason Not to Try/Use	13%	21%	16%
Average Age First Used	19	19	19
Number of Observations	18655	19885	13657

Notes: 48 individuals reported "no opportunity to use" as a reason they had not used but were recorded as using in the last 12 months; 45 individuals reported that cannabis was not available to them but were recorded as using in the last 12 months. We dropped these 93 individuals.

Table 1: Annual Descriptive Statistics

In Australia the use of marijuana for any purpose is illegal. To assess the role the legal status of marijuana plays in the decision to use, we construct a variable that is intended to capture the (dis)utility associated with the illegal status of marijuana. It is an indicator variable with a value of one if either of the following questions are answered affirmatively:

"Did the fear of legal consequences influence your decision never to use marijuana/cannabis" and "Would you try cannabis/marijuana if it were legal." On average in the sample 17% of individuals who do not use marijuana say they do not use because it is illegal or the fear of legal consequences influenced their decision not to use. Almost 90% of these individuals report they have access to marijuana.[8]

1.2 Marijuana Access

The NDSHS survey also asks questions regarding how accessible marijuana is to the individual, which is particularly suited to the focus of this research. We construct three measures of accessibility (Access 1, Access 2, and Access 3) based on the answers to these questions, which are summarized in Table 2.[9] If the individual reports that they used or had been offered the drug in the past 12 months (about 24% of the sample) then they must have had access to the drug. Hence we set all our accessibility measures to one. Second, individuals report how difficult it would be to obtain marijuana. If they indicate it is "very easy" (about 28% of the sample) then we set all accessibility variables to one; or if the response is "difficult to obtain" (about 6%) or "impossible" (about 12%) then we set all access variables equal to zero. Third, non-users were asked why they didn't use the drug. If they answer it was "too difficult to get" or they had "no opportunity" (about 8% of the non-user sample) then we also set all accessibility variables to zero.[10] About 23% of individuals indicated it was "fairly easy" to obtain and 7% that it was "fairly difficult." The variation in our accessibility measures comes from the answers to these two questions, where our broadest definition (Access 1) assumes individuals who report it is "fairly difficult" to obtain have access; Access 2 assumes these individuals would not have access but those who say it is "fairly easy" to obtain would; and the strictest measure (Access 3) assumes access is restricted for both groups.

[8] We also considered an alternative disutility variable based on the question "If marijuana/cannabis were legal to use, would you ..." where we coded 0 if the answer is "Not use it - even if legal and available;" equal to 1 if the answer is "Try it; or Use it as often/more often than now" and -1 if respond "use it less often than now." However, such a variable is difficult to interpret and problematic due to the (small) set of subjects who would use it less often.

[9] While not all individuals answered all the questions, we have answers to at least one question for each user.

[10] There were 48 individuals which reported no opportunity to use as a reason they had not used but who were recorded as using in the last 12 months; there were an additional 45 individuals who reported that cannabis was not available to them but who used in the last 12 months. We drop these 93 observations.

As Table 2 illustrates, our broadest definition indicates about 59% of the sample has access to marijuana; our intermediate measure shows 53% have access on average; whereas our most restrictive definition indicates under 36% have access. The results in this paper are presented using the intermediate definition of access (Access 2) and robustness of the results to the access variable are presented in section 6. We should note that we are not relying on this variation to identify the correlation of the error terms, instead we conduct robustness checks of our estimates under different access definitions. The last row of Table 2 gives the mean conditional probability of use given access. It shows that the probability of use among those with access increases as we tighten the definition of access. The last column presents the probability of use for each access component and shows that better access translates into higher usage rates. These data are consistent with the idea that individuals who want to use marijuana try harder to gain access. Over our sample period access decreases overall by about 10 percentage points within each access measure. For example, access as measured by our intermediate measure decreases from 58% in 2001 to 46% in 2007.

	Value of Access Variable			Access	Probability
Questions on Access	Access 1	Access 2	Access 3	Percentage	of Use
Offered Marijuana	1	1	1	24%	57%
How difficult/easy to get cannabis if wanted some?					
Very Easy to Obtain	1	1	1	28%	31%
Fairly Easy to Obtain	1	1	0	23%	17%
Fairly Difficult to Obtain	1	0	0	7%	8%
Difficult to Obtain	0	0	0	6%	2%
Impossible	0	0	0	12%	1%
No Opportunity to Use	0	0	0	8%	0%
Mean Access	59%	53%	36%		
Probability of Use Given Access	23%	26%	37%		

Table 2: Access Variables Definition and Statistics

1.3 Descriptive Patterns

Table 3 provides descriptive statistics for access (Access 2) and use. Marijuana use and access vary with age and are the most prevalent among those in their twenties and thirties. Use declines to under 0.4% for those in their sixties. Males and younger people are more likely to have access and, conditional on having access, to use marijuana. Marijuana use varies across states, ranging from 12% in Victoria to 20% in the Northern Territory. If we compute the percentage of users among those with access (as opposed to the percentage of users among the entire population) the percent with access that report using marijuana is higher on average with a lower variance across states. This is consistent with non-random access. For example, 12% of residents of Victoria used in the past year, while 14% Tasmanians used last year. However, fewer individuals in Victoria had access to marijuana and, once this fact is accounted for, both states saw the same amount of use conditional on access (24%). That is, the distribution of use among those that have access is different than the unconditional distribution of use. This suggests that if we were only to rely on the distribution of use (instead of the conditional distribution) to identify how use would change after access were increased (due to legalization) we could have a different impression of the impact. Furthermore, this observed variation in access and use will allow us to separately identify the impact of explanatory variables that impact both use and access. Four states/territories have decriminalized the possession of small quantities of marijuana via the introduction of infringement schemes.[11] Both use and access are higher in states where marijuana use is decriminalized. In some specifications, we include whether the state is decriminalized and the maximum number of grams for which possession is a minor offense.[12]

As Table 3 shows the percent that report having access to marijuana varies across states. This may be related to the fact that growing conditions vary both across states (and years), which may impact how much marijuana is available to purchase.[13] For example, temperature is an important component impacting growing conditions. The mean average temperature

[11] The decrimilized states are South Australia (SA), Northern Territory (NT), Australia Capital Territory (ACT), and Western Australia (WA). Under an infringement scheme individuals which are found to have violated the law with a minor marijuana offence are fined but are not jailed. In other states and territories (Tasmania (TAS), Victoria (VIC), New South Wales (NSW), and Queensland (QLD)) possession of any amount of marijuana is a criminal offence, and individuals may be jailed for possession of any quantity. These jurisdictions have introduced "diversion schemes" where the police may issue a caution of diversion into treatment or education for a minor offence instead of jail time. The number of cautions issued before a criminal conviction varies by jurisdictions. The diversion schemes were introduced at different times: in 1998 in TAS and VIC; in 2000 in NSW, and 2001 in QLD. The state of WA gradually introduced the schemes between 2000 to 2003. Minor cannabis offences only refer to the possession of cannabis, not the possession of a plant. Trafficking and possessions of larger amounts of cannabis are serious offences that incur large monetary fines and long prison sentences.

[12] What constitutes a minor offense and the fine varies by state. These include possession of small amount of marijuana plant material (i.e., bulbs, leaves)(SA and NT), growing of one plant (SA) or two plants. The quantity considered a minor offence varies by cannabis type (plant versus resin) and ranges from 100 grams of plant material in SA to 25 grams in ACT.

[13] For more information regarding the growing seasons in Australia please see http://www.cannabisculture.com/articles/2674.html.

in a year is about 77 degrees Fahrenheit (25 Celsius) in Queensland but only 51 Fahrenheit (11 Celsius) in Tasmania.[14] Rainfall is another market condition that may impact growing seasons. Rainfall also varies across states, where Victoria sees more than four times as much average rainfall than does South Australia. Another potential issue to consider for accessibility is that most cannabis is grown in the outback hence the location of the individual (specifically if they live in a city) may impact how much is available to purchase.

Demographic Group or State	Percent Used in Last 12 Months	Percent Report Access	Percent With Access that Use	Average State Price*	Number of Observations
On Average	13%	53%	26%		52197
Male	17%	58%	30%		22146
Teenager	25%	71%	35%		3349
Age in Twenties	25%	73%	35%		9958
Age in Thirties	15%	58%	26%		13570
Age in Forties	10%	47%	20%		12408
Age Fifty or Over	4%	34%	11%		11744
New South Wales	12%	50%	24%	41.79	14194
Victoria	12%	48%	24%	33.51	10967
Queensland	13%	51%	26%	33.09	9395
Western Australia	18%	60%	30%	42.31	5820
South Australia	14%	57%	25%	41.05	4214
Tasmania	14%	57%	24%	26.08	2317
ACT	13%	51%	26%	28.38	2653
Northern Territory	20%	65%	31%	38.18	2637
Decriminalized	15%	57%	27%	38.90	12933
Not Decriminalized	13%	51%	25%	36.26	39264

Notes: These are based on access variable definition 2. *The Average State Price Data are not from the individual survey but are market level data from from Australian Bureau of Criminal Intelligence.

Table 3: Descriptive Statistics by Use and Access

Indeed, preliminary patterns in the data suggest that both temperature and whether an individual lives in a city are correlated with access. Specifically, a regression of access (using any access variable) with temperature, living in a city, other individual characteristics and state-fixed effects as regressors indicates that temperature and living in a city have statistically significant (negative) impacts on access. They are jointly significantly different than zero as well (with a χ^2 test-statistic of 441 and a p-value of 0.00). We provide more discussion of our use of these variables as exclusionary restrictions in Section 3.5.

[14] These data were obtained from http://www.bom.gov.au/climate/change/#tabs=Tracker&tracker=timeseries

1.4 Prices

Our market-level pricing data come from the Australian Bureau of Criminal Intelligence that publishes prices based on undercover buys in its Illicit Drug Data Reports.[15] Given that marijuana is an illicit drug there are a few data issues to resolve regarding the prices. First, we do not observe prices in all years due to different state procedures in filling in forms and the frequency of drug arrests of that certain marijuana form. To deal with missings across time we use linear interpolation when we observe the prices in other years. Second, the price per gram is the most frequently reported price, but in some quarters the only price available is the price per ounce.[16] We cannot simply divide the price per ounce by 28 to convert it to grams as quantity discounts are common (Clements 2006). However, assuming price changes occur at the same time with gram and ounce bags, when we observe both the gram and ounce prices we substitute the corresponding price per gram for the time period in which it is missing when the price per ounce is the same in the period where both are reported. Third, some prices are reported in ranges, in which case we use the mid-point of the reported price range. We deflate the prices using the Federal Reserve Bank of Australia Consumer Price Index for Alcohol and Tobacco where the prices are in real 1998 AU$. These data are reported on a quarterly or semi-annual basis. We construct an annual price per gram measure by averaging over the periods.

The major psychoactive chemical compound in marijuana is delta-9-tetrahydrocannabinol (or THC). The amount of THC absorbed by marijuana users differs according to the part of the plant that is used (e.g., leaf, head), the way the plant is cultivated (e.g., hydro), and the method used to imbibe the drug. On average marijuana contains about 5% THC, where the flowering tops contain the highest concentration followed by the leaves (Adams and Martin, 1996). Marijuana that is grown hydroponically (hydro), indoors under artificial light with

[15] This is the common source of pricing data used by researchers. We also considered using alternative pricing data reported in the Illicit Drug Reporting System National Reports. These are self-reported prices from a non-representative sample of injection users. Unfortunately they are less believable in that there is virtually no variation in nominal prices across years, states, and quality types: 88% of the observations are either 20 or 25 (with a mean of 23 and standard deviation of 3).

[16] A joint contains between 0.5 to 1.5 grams of plant material. We have data from 1998 to 2007 so when the data are missing for one year/state we use all years from 1998-2008 for linear interpolation. This is necessary for the price of hydro for - 2 states in 2001; 1 state in 2007; for the price of head for - 4 states in 2001; 2 states in 2004; and 1 state in 2007; and for the price of leaf for - 2 states in 2001; 3 states in 2004; and 4 states in 2007. The states with missing prices vary across years and type, and when interpolation is necessary we never have fewer than 6 points to use to interpolate the missing one point.

nutrient baths, typically has higher concentrations of THC relative to naturally grown leaf and head (Poulsen and Sutherland, 2000). The THC levels found in hydro are similar to those found in skunk.

	Year		
	2001	2004	2007
Median Market Prices by Gram			
Leaf	30	33	37
Head	30	34	37
Hydro	33	34	38
Individual Use by Type			
Leaf	46%	43%	39%
Head	80%	77%	70%
Hydro	23%	19%	40%

Notes: These are real prices in 1998$. The price data are market level data from the Australian Bureau of Criminial Intelligence.

Table 4: Prices and Use by Type

The NDSHS survey contains information about which form of marijuana the user uses. We combine these information with type-specific prices to simulate an individual price for person i. The details of how we construct simulated prices for users, as well as for non-users, are discussed in detail in section 3. Table 4 presents market prices and the individual percentage of use per type by year. The pricing data exhibit variation across markets within a type and across types within a market.[17] Given the higher amount of THC present in hydro and skunk (hereafter collectively referred to as hydro), hydro demands a higher price. The most common forms of marijuana used are leaf and head, but their use has been declining over time. In contrast, the last row shows that users have moved into using more hydro in the latter year. This is consistent with patterns seen in the rest of the world.[18] Given that the forms of marijuana vary in THC content, in the model we include a variable to capture

[17] Note that while overall we observe an increase in price, these reported prices represent averages over the different states/territories within each survey year. Prices vary substantially across states and territories and often exhibit different trends. This is also evident in the graph of the price of leaf per gram for the different Australian states and teritories provided in appendix A.1.1.

[18] According to the Australian Bureau of Criminal Intelligence (1996), the increase in hydroponic systems may be related to the fact that, unlike external plantations, hydroponic cultivation is not affected by the growing seasons of the region. It is common to use types in combination (i.e., a bag might contain leaf and head), hence the percentages do not sum to one.

the level of THC, which can be thought of as the "quality" of the marijuana product.[19]

2 Model

Our paper concerns the impact of legalization on marijuana use. Given that illicit drugs are not as easy to find as legal products, one can argue that non-users have less information about how to get marijuana, which is the first step to becoming a user. If marijuana were legalized, purchasing it would be as difficult as purchasing cigarettes or alcohol. Furthermore, legalization would remove the "breaking the law" hindrance, which may result in use among some current non-users.

An individual chooses whether or not to consume marijuana in market m which is defined as a state-year combination.[20] The indirect utility individual i obtains from using marijuana in market m depends on a number of factors including the price the individual pays (p_{im}), demographic characteristics (represented by the vector d_i), such as gender, age splines (young adult, college age, pensioner, etc.), education variables, whether they are of aboriginal descent, and health status.[21] One caveat is that we do not observe individual prices, p_{im}. However we know something about the distribution of the prices from the data, as discussed in section 1, which we use to construct an empirical price distribution, given by $\widehat{P}_m(p_{im})$, to generate p_{im}. We discuss the construction of the empirical price distribution in detail in section 3.1.

Market specific variables can also impact the benefit of consuming marijuana. These are represented by x_m and include the year in which the marijuana was purchased, average rainfall in the state-year, the proportion of high quality marijuana sold in the market, and state-fixed effects. We also include variables related to legality (represented by the vector L_{im}) that include the (dis)taste an individual has for engaging in illegal behavior, and the

[19] We could simulate the quality faced by an individual using a similar methodology as we use for simulating prices. However, the focus of our paper is on the impact of prices and access on use and, so, to reduce computational complexity, we choose not to simulation qualities together with simulating prices.

[20] Our model focuses on whether (or how much) to use of marijuana instead of what quality of marijuana to purchase. We wish to focus on whether individuals will use post legalization (and how much this will change) and modeling product characteristic choices will complicate matters as it is likely that the "products" available will change post legalization. The baseline model is at the extensive margin (the decision whether to use) in the past 12 months.

[21] Health status may be endogeneous to use. We run robustness checks without health status as a control variable, and the results do not change.

amount that can be grown for a minor offense.[22] Given that the age of an individual may influence their sensitivity to paying for marijuana or their view of doing something that is illegal ($L_{im}^{illegal}$), we also include an interaction of the age brackets (d_i^{age}) with price and with the (dis)utility of illegal behavior.[23] Specifically, the indirect utility is represented by

$$U_{imj} = \alpha_0 + p_{im}\alpha_1 + p_{im}d_i^{age\prime}\alpha_2 + d_i'\beta_1 + x_m'\beta_2 + L_{im}'\delta_1 + L_{im}^{illegal}d_i^{age\prime}\delta_2 + \epsilon_{imj}, \qquad p_{im} \sim \widehat{P}_m(p_{im}), \tag{1}$$

where $j = 1, \epsilon_{imj}$ is an idiosyncratic error term, and $\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2, \delta_1$ and δ_2 are (vectors of) parameters to be estimated.[24]

Individuals have utility from not using marijuana, which we model as

$$U_{im0} = \alpha_0 + \epsilon_{im0}, \tag{2}$$

where all non stochastic terms are normalized to zero, because we cannot identify relative utility levels.

One innovation of this paper is to model the role of accessibility in marijuana use. We allow for the possibility that whether an individual knows where to buy is a function of i's observed characteristics and market characteristics. The probability that person i has access to marijuana in market m is given by

$$\phi_{im} = \Pr(h_i'\gamma_1 + w_{im}'\gamma_2 + \eta_{im} > 0), \tag{3}$$

where h_i represents individual attributes and includes all variables in d_i except health status. The market-specific variables that influence access (w_{im}) include whether an individual lives in a city, average rainfall, average temperature, and state fixed effects. The η_{im} is

[22] In one specification we do not include state-fixed effects and instead include an indicator for whether marijuana use is decriminalized in the market.

[23] There may be individual characteristics that are not observed by the econometrician that impact the utility one obtains from using marijuana. We also estimated specifications with random coefficients on legality and prices. However, once we include demographic interactions there is not enough additional variation to identify the random coefficients.

[24] Our data are not longitudinal so we cannot control for (endogenous) lagged use. Therefore, one should consider our model as capturing use among recreational users and not accounting for the role possibly played by addiction. We think addiction is less of an issue for our data because, as discussed previously, our data capture mostly recreational use: only 3% report daily use or that use is a habit. However, we conduct robustness checks where we consider only non-habitual and non-daily users. These results are discussed in section 6.

an individual-market-specific error term and γ_1 and γ_2 are (vectors of) parameters to be estimated.

It is likely that access to marijuana and the use decision are correlated (due to selection). For example, some individuals may have high levels of utility for using marijuana, and therefore will search for where to purchase it. This can be captured by correlation in observables (such as demographics) and correlation in the error terms in the indirect utility and access equations.[25]

The probability that individual i chooses to use marijuana in market m depends upon the probability they know where to purchase marijuana (ϕ_{im}) and the probability they would use it given availability. Let

$$R_{im} \equiv \{U_{im1}(p_{im}, d_i, x_m, L_{im}, \epsilon_{im1}) \geq U_{im0}(x_m, \epsilon_{im0}), \phi^*_{im}(h_i, w_{im}, \eta_{im}) > 0\}$$

define the set of variables that results in consumption of marijuana given the parameters of the model, where $\phi^*_{im} = h_i'\gamma_1 + w_{im}\gamma_2 + \eta_{im}$. The probability i chooses to use marijuana in market m (the individual market share) is given by

$$\begin{aligned} S_{im} &= \int_{R_{im}} dF_{\epsilon,\eta,p}(\epsilon, \eta, p) && (4) \\ &= \int_{R_{im}} dF_{\epsilon,\eta}(\epsilon, \eta) d\widehat{P}_m(p_{im}), && (5) \end{aligned}$$

where $F(\cdot)$ denotes distribution functions, the latter equality follows from independence assumptions, and $\widehat{P}_m(p_{im})$ represents the market-specific empirical price distribution.

Notice that some variables are included only in the access equation (3) and do not impact utility directly. These excluded variables include the average temperature in the market and whether the individual lives in a city. Likewise some variables are modeled as impacting utility only. The variables excluded from the access equation include price, variables related to legality (L_{im}), the proportion of high quality of cannabis sold, and health status. We discuss the motivation for these exclusionary restrictions in detail in section 3.5.

The approach described so far informs us about the extensive margin, i.e., how people move from no marijuana use to marijuana use. We also want to get an estimate of the tax

[25] Our goal is to account for potential selection in use that could arise from individuals having access to marijuana. This could come from individuals searching for marijuana or being offered marijuana. We wish to get accurate estimates after controlling for selection not to understand search decisions. For this reason, as well as data limitations, we do not estimate a search model. See Galenianos, Pacula, and Persico (2009) for a theoretic search model applied to illicit markets.

revenue that would be raised under legalization, which requires information about per unit use. Ideally, we would have information about quantity used. Unfortunately these data are not available, but we have information on frequency of use and the average amount used per session (as discussed in section 1) that we use to construct a quantity variable associated with each frequency. We model use frequency for individual i in market m in terms of three frequencies: no use, infrequent use and frequent use with $j = 0, 1, 2$, respectively for the indirect utility in equation 1 (for $j = 1, 2$) and equation 2 (for $j = 0$). An infrequent user is one who uses once quarterly, biannually, or annually. A frequent user is one who uses monthly, weekly, or daily. Specifically, the frequency of use variable for those who use is given by

$$\begin{aligned} y_{im} &= 1 \text{ if } 0 < \widetilde{y}_{im} + \nu_{im} \leq \tau \\ \text{and } y_{im} &= 2 \text{ if } \tau < \widetilde{y}_{im} + \nu_{im}, \end{aligned}$$

where $\widetilde{y}_{im} = \widetilde{\alpha}_0 + p_{im}\widetilde{\alpha}_1 + p_{im}d_i^{age\prime}\widetilde{\alpha}_2 + d_i'\widetilde{\beta}_1 + x_m'\widetilde{\beta}_2 + L_{im}'\widetilde{\delta}_1 + L_{im}^{illegal}d_i^{age\prime}\widetilde{\delta}_2$, the variable τ is a cut-off parameter to be estimated, and ν_{im} is an idiosyncratic random shock. In the next section we discuss how we estimate the probability of use model and the frequency of use model. Details on how we compute tax revenues are provided in section 5.3.

Our approach differs from the rest of the literature in a couple of fundamental ways. First, we model accessibility directly. An implicit assumption in economic models that have been considered to date is that all individuals have access to marijuana. In our framework, this equivalent to $\phi_{im} = 1$ and that there is no correlation in observables and in the errors in the indirect utility and access equations. It further implies that observed consumption reflects a choice based on preferences only which can lead to biased parameters, even without the presence of correlation on unobservables. Second, we model the (dis)utility from engaging in illegal behavior directly. We are able to do both of these things because we have data on whether individuals have access to the drug and their feelings about engaging in illegal behavior. Modeling accessibility is particularly important for drawing correct inferences about choices that individuals would make under a policy of legalization, where the accessibility issue would essentially disappear. Third, we directly address an issue that is prevalent in studies of illicit markets: the fact that prices are not observed for each purchase. To do so we use individual-level data on the type of marijuana used (i.e., leaf, head, hydro) combined with market-level pricing data to obtain an implied price faced by users and non-users. This allows us to estimate a model with individual prices while not observing these in the data.

3 Econometric Specification

We propose and estimate two econometric models for marijuana access and utility based on the models specified in the previous section. The first considers the extensive margin of whether an individual uses or not. The second addresses the intensive margin of frequency of marijuana use. Prior to discussing these models we describe the method for dealing with unobserved individual prices. The last two subsections address the estimation strategy and identification exclusionary restrictions.

3.1 Unobserved Individual Prices

As mentioned previously we do not observe individual prices, p_{im}. In the model section, we introduced the general idea of an empirical price distribution $\widehat{P}_m(p_{im})$ to address this challenge. To construct this empirical distribution we exploit information on the average market-level marijuana prices per gram ($\overline{p}_{mt}$) for each of the three types $t = 1, 2, 3$ (leaf, head, hydro) and summarized in vector $\overline{p}_m = \{\overline{p}_{m,leaf}, \overline{p}_{m,head}, \overline{p}_{m,hydro}\} = \{\overline{p}_{mt} : t = 1, 2, 3\}$. These are based on the prices reported by the Australian Bureau of Criminal Intelligence as detailed in Section 1.4. Further we observe individual-level (binary) data from NDSHS on type used by an individual. Based on these responses for all individuals in a market, we construct market level probabilities of using a type in each market, $\overline{\pi}_m = \{\overline{\pi}_{m,leaf}, \overline{\pi}_{m,head}, \overline{\pi}_{m,hydro}\} = \{\overline{\pi}_{mt} : t = 1, 2, 3\}$.[26] Our aim is to exploit these observed quantities to construct an empirical distribution for the price per gram that an individual faces, $p_{im} \sim \widehat{P}_m(p_{im})$, taking into account the consumption of the three types and price differences across types. We specify distributions of prices and probabilities of use for each type by market, denoted $F_p(p_{imt})$ and $F_\pi(\pi_{imt})$, respectively as truncated normals, where

$$\begin{aligned} p_{imt} &\sim F_p(p_{imt}) \ , \ F_p(p_{imt}) = TN_{(0,\infty)}(\overline{p}_{mt}, \Omega^p_{mt}) \ \text{ for } \ t = 1, 2, 3 \\ \pi_{imt} &\sim F_\pi(\pi_{imt}) \ , \ F_\pi(\pi_{imt}), = TN_{(0,\infty)}(\overline{\pi}_{mt}, \Omega^\pi_{mt}) \quad s.t. \ \sum_t \pi_{imt} = 1. \end{aligned} \tag{6}$$

with the means set at the observed market averages and variances set using information across all markets. Assuming that the "average" price (p_{im}) a subject faces depends on the relative use of each type we then define this price, as a function of the random variables $\{\pi_{imt}, p_{imt} : t = 1, 2, 3\}$, as an average of the prices over the three different types weighted

[26] Note that, by construction, these market type use probabilities do not vary within a market.

by their respective use probabilities

$$p_{im}|\pi_{imt}, p_{imt} = \sum_{t=1}^{3}(\pi_{imt} * p_{imt}).$$

The price p_{im} reflects the average price faced by individual i in market m based on draws from the market and type specific distributions of price and the probability of use. The implied marginal empirical distribution of price for individuals in a market is given by

$$\widehat{P}_m(p_{im}) = \int \sum_{t=1}^{3}(\pi_{imt} * p_{imt})\ dF_p(p_{imt})\ dF_\pi(\pi_{imt}) \tag{7}$$

assuming independence in the distributions across types and across prices and usage as implied by expression 6.[27] This method of generating individual prices from an empirical distribution improves upon the typical approach in the literature that uses average market prices as those do not vary within a market neither by type used nor probability of use of each type, whereas we can generate a distribution of prices in each market. Importantly, this approach also allows us to obtain the implied price faced by users and non-users in a symmetric way and to properly address the econometric issue of unobserved individual prices in estimation by integration. Note that while the analytical form of the distribution is unknown, it can be easily approximated within our Bayesian estimation framework by a simple extension of the MCMC algorithm for the model estimation, essentially expanding the parameter space to include the vector of individual prices for access subjects to be estimated. We discuss further details of the simulation to solve the integral and construction of the empirical distribution in Appendix A.1 which also contains graphs of the implied empirical price distributions.

One other point regarding the prices concerns potential endogeneity issues. Specifically, as the prices are not the individually reported purchase price, it may be the case that price is correlated with the error term if it reflects unobserved quality that is not included as a regressor. As we note in section 1, prices are higher the higher is the potency, which can be thought of as measure of the quality of the marijuana. As we include a measure of the potency to control for quality this should ameliorate any endogeneity concerns. We also

[27] We assume that the distributions of prices and market usage are independent across types. Alternatively, we could allow for correlation in prices across types, across usage of types, and/or correlation in the joint distribution of prices and type. We tried to do this, however, we do not have enough individual level variation in the data to allow us to identify 3 more covariance parameters for each of the markets. We assume that prices are independent of ease of access which is a potential limitation mainly driven by data.

conduct robustness checks to further investigate issues related to price endogeneity. These details can be found in section 6.

3.2 Extensive Margin: Probability of Marijuana Use

Suppose we have a sample of $i = 1, .., n$ individuals. Let $a_{im} = 0, 1$ denote whether an individual has access to marijuana ($a_{im} = 1$) or not ($a_{im} = 0$), where access to marijuana will depend on a vector of covariates of individual attributes and market characteristics and a random shock η_{im}. Here we assume that an individual's indicator of having access to marijuana can be modeled in terms of a probit

$$a_{im} = I[\mu^a_{im} + \eta_{im} > 0] \text{ where } \eta_{im} \sim N(0, 1), \tag{8}$$

where $\mu^a_{im} \equiv h'_i\gamma_1 + w_{im}\gamma_2$ so that $\phi_{im} = \Pr(a_{im} = 1) = \Phi(\mu^a_{im})$. Further, we let $u_{im} = 0, 1$ denote whether individual i has a positive (indirect) utility from using marijuana relative to the outside good. For ease of exposition, we refer to u_{im} as net-utility.[28] We have

$$u_{im} = I[U_{im1} > U_{im0}] = I[\mu^u_{im} > \varepsilon_{im}], \tag{9}$$

where $\mu^u_{im} \equiv \alpha_0 + p_{im}\alpha_1 + p_{im}d^{age\prime}_i\alpha_2 + d'_i\beta_1 + x'_m\beta_2 + L'_{im}\delta_1 + L^{illegal}_{im}d^{age\prime}_i\delta_2$ and $\varepsilon_{im} \equiv \epsilon_{im0} - \epsilon_{im1}$, where ε_{im} is a mean zero stochastic term distributed i.i.d. normal across markets and individuals.

To account for the correlation between marijuana access and use decisions as a result of unobserved confounders we assume a joint normal distribution for the two error terms and let

$$\begin{pmatrix} \eta_{im} \\ \varepsilon_{im} \end{pmatrix} \sim N\left(0, \Xi = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \tag{10}$$

where the off-diagonal element ρ reflects the correlation between the two decisions and the diagonal elements are 1 due to the standard identification restriction for binary response variables.

In our setting with limited access, the net-utility from marijuana use is not observed for all individuals, but only reflected in the observed consumption decisions of those individuals

[28] Ching, Erdem and Keane (2009) contains a similar model, although they do not directly observe the outcome variable in the selection equation. They show that in a model with more than two alternatives in the second stage, it is possible to identify the parameters in the first stage (selection stage) provided that there are exclusion restrictions.

with access. Let indicator $c_{im} = 0, 1$ denote whether consumer i is observed using marijuana. Observed consumption can be expressed in terms of access and preferences (net-utility) based on our joint model as

$$\Pr(c_{im} = 1) = \Pr(a_{im} = 1) \Pr(u_{im} = 1 | a_{im} = 1)$$

$$\Pr(c_{im} = 0) = \Pr(a_{im} = 0) + \Pr(a_{im} = 1)(\Pr(u_{im} = 0 | a_{im} = 1),$$

where $\Pr(u_{im} = j | a_{im} = 1)$ for $j = 0, 1$ is the net-utility conditional on access. The first line states that marijuana consumption reflects access to marijuana and a positive net-utility from use, while the second line shows that zero consumption could arise from: (1) no access or (2) access and negative net-utility. In other words, the observed zero consumption is inflated with zeros reflecting access only. Observing access for each individual allows us to contribute those zeros correctly to the access model. Only for individuals with access the decision whether to use marijuana reflects the net-utility from use so that for those subjects $u_{im} = c_{im}$.

Thus, we observe three possible cases, $(a_{im} = 1, u_{im} = 1)$, $(a_{im} = 1, u_{im} = 0)$ and $(a_{im} = 0)$ and the likelihood contribution for the observed access and net-utility of individual i in market m can therefore be expressed as

$$\begin{aligned} \Pr(a_{im} = 0 | \boldsymbol{\theta}) &= \Pr(\mu^a_{im} + \eta_{im} \leq 0) & \text{if } a_{im} = 0 \\ \Pr(a_{im} = 1, u_{im} = 0 | \boldsymbol{\theta}) &= Pr(\mu^a_{im} + \eta_{im} > 0, \mu^u_{im} + \varepsilon_{im} \leq 0) & \text{if } a_{im} = 1, u_{im} = 0 \\ \Pr(a_{im} = 1, u_{im} = 1 | \boldsymbol{\theta}) &= Pr(\mu^a_{im} + \eta_{im} > 0, \mu^u_{im} + \varepsilon_{im} > 0) & \text{if } a_{im} = 1, u_{im} = 1 \end{aligned} ,$$

where $\boldsymbol{\theta}$ refers to the vector of all model parameters. In other words, given our normal error specifications we have a univariate probit for access for subjects with no access and a bivariate probit for access and net-use for subjects with access. Hence we can first rewrite the above expressions for the likelihood contribution of individual i, assuming that price is observed, more compactly as

$$\Phi(-\mu^a_{im})^{(1-a_{im})} \; \Phi_2(\mu^a_{im}, -\mu^u_{im}; \rho)^{(a_{im})(1-u_{im})} \; \Phi_2(\mu^a_{im}, \mu^u_{im}; \rho)^{(a_{im})(u_{im})}, \tag{11}$$

where $\Phi(\cdot)$ refers to the CDF of a standard normal distribution and $\Phi_2(\cdot)$ to the CDF of a standard bivariate normal distribution.

In contrast, the likelihood contribution in the standard reduced form model for marijuana used in the empirical literature is formulated without reference to access limitation, yielding a simple univariate binary choice model in terms of the observed consumption decision (c_{im}) for access and non-access subjects, rather than in terms of the net-utility (u_{im}) observed for access subjects. In the empirical analysis we *report* results on the traditional probit model in consumption $\Phi(-\mu^c_{im})^{(1-c_{im})}\Phi(\mu^c_{im})^{(1-c_{im})}$ as the benchmark model where the mean μ^c_{im} includes controls both for use and access. If access depends on observables that also affect use, estimates from the simple model will reflect effects from access and use. Presence of unobserved confounders that affect use and access will add further bias to the preference parameters already contaminated with access effects.

Let $\mathbf{a}_m = \{a_{1m}, ..., a_{n_m m}\}$ denote the vector of access variables for all n_m subjects in market m, $\mathbf{u}_m = \{u_{1m}, ..., u_{n_{1_m} m}\}$ the vector of net-utility variables for the n_{1_m} subjects in market m with access to marijuana and $\mathbf{W}_m = \{\mathbf{W}_{1m}, ..., \mathbf{W}_{n_m m}\}$ the matrix of all covariates excluding price. Grouping subjects in each market by marijuana access, we define the sets I_{m1} for all subjects with access and I_{m0} for all subjects with no access. Taking into account that the price is unobserved, the likelihood of observing the data $(\mathbf{a}_m, \mathbf{u}_m)$ for all subjects in market m can then be be expressed in two parts for the set of non-access and access subjects as

$$f(\mathbf{a}_m, \mathbf{u}_m|\boldsymbol{\theta}, \mathbf{W}) = \prod_{I_{m0}} \Pr(a_{im} = 0|W_{im}, \boldsymbol{\theta}) \prod_{I_{m1}} \int \Pr(a_{im} = 1, u_{im} = j|W_{im}, \boldsymbol{\theta}, p_{im})\, d\widehat{P}_m(p_{im}), \tag{12}$$

where p_{im} is the individual-specific price coming from the distribution defined in equation (7). The expression under the integral is the term $\Phi_2(\mu^a_{im}, -\widetilde{\mu}^u_{im}; \rho)^{(a_{im})(1-u_{im})} + \Phi_2(\mu^a_{im}, \widetilde{\mu}^u_{im}; \rho)^{(a_{im})(u_{im})}$ from (11) where the mean term $\widetilde{\mu}^u_{im}$ uses the price $p_{im} \sim \widehat{P}_m(p_{im})$. For all individuals in the sample the likelihood is simply a product over the likelihoods for all markets $m = 1, ..., M$, $f(\mathbf{a}, \mathbf{u}|\boldsymbol{\theta}, \mathbf{W}) = \prod_{m=1}^M f(\mathbf{a}_m, \mathbf{u}_m|\boldsymbol{\theta}, \mathbf{W})$, where $\mathbf{a} = \{a_1, ..., a_M\}$, $\mathbf{u} = \{u_1, ..., u_M\}$ and $\mathbf{W} = \{\mathbf{W}_1, ..., \mathbf{W}_M\}$ refer to the observed data for all sample subjects.

3.3 Intensive Margin: Frequency of Marijuana Use

We also consider an extended version of the above model to estimate a model of frequency of use. We use the ordered marijuana use response variable $y_{im} = j$, $j = 0, 1, 2$ for all individuals with access, where the three categories refer to "no use", "infrequent use" and "frequent use," respectively. Extending the model for use given in equation (9), we have

$$y_{im} = 0 \text{ if } (\mu^u_{im} + \nu_{im}) \leq 0,\ y_{im} = 1 \text{ if } 0 < (\mu^u_{im} + \nu_{im}) \leq \tau \text{ and } y_{im} = 2 \text{ if } \tau < (\mu^u_{im} + \nu_{im}),$$

where τ is a cut-off parameter to be estimated and ν_{im} refers to the random shock in the latent utility of marijuana use in the ordered probit model. The mean μ^u_{im} depends as before on a set of individual characteristics such as demographics and price, market specific variables, and legality related variables. As in the bivariate probit model above we allow for selection based on unobservables and unobservables and model the access and use decision jointly. We again assume a joint normal distribution of the error terms of the use and access model, $(\eta_{im}, \nu_{im}) \sim N(0, \Xi)$, to allow for the correlation in unobservables, with the access model specified as in equation (8). Under the ordered probit outcome for marijuana the likelihood contribution of individual i is

$$\begin{aligned}
\Pr(a_{im} = 1, y_{im} = 0|\boldsymbol{\theta}) &= Pr(\mu^a_{im} + \eta_{im} > 0, \mu^u_{im} + \nu_{im} \leq 0) && \text{if } a_{im} = 1, y_{im} = 0 \\
\Pr(a_{im} = 1, y_{im} = 1|\boldsymbol{\theta}) &= Pr(\mu^a_{im} + \eta_{im} > 0, 0 < \mu^u_{im} + \nu_{im} \leq \tau) && \text{if } a_{im} = 1, y_{im} = 1 \\
\Pr(a_{im} = 1, y_{im} = 2|\boldsymbol{\theta}) &= Pr(\mu^a_{im} + \eta_{im} > 0, \tau < \mu^u_{im} + \nu_{im}) && \text{if } a_{im} = 1, y_{im} = 2.
\end{aligned}$$

As before we have $\Pr(a_{im} = 0|\boldsymbol{\theta}) = \Pr(\mu^a_{im} + \eta_{im} \leq 0)$ for non-access subjects. Addressing the issue of the unobserved individual prices as described in section 1.4, the likelihood contribution of all subjects in market m is

$$f(\mathbf{a}_m, \mathbf{y}_m|\boldsymbol{\theta}, \mathbf{W}) = \prod_{I_{m0}} \Pr(a_{im} = 0|W_{im}, \boldsymbol{\theta}) \prod_{I_{m1}} \int \Pr(a_{im} = 1, y_{im} = j|W_{im}, \boldsymbol{\theta}, p_{im})\, d\widehat{P}_m(p_{im}),$$

where $\mathbf{y}_m = \{y_{1m}, ..., y_{n1_m m}\}$ is the vector of the ordered response variable on frequency of use for all subjects in market m.

3.4 Estimation Strategy

Based on these likelihood expressions we can identify the parameters for the access and the net-utility models and the correlation. We estimate both models via standard Bayesian Markov Chain Monte Carlo methods, building closely on Chib and Jacobi (2008) and Bretteville-Jensen and Jacobi (2011). The details are provided in Appendix A. Bayesian methods are increasingly used in empirical analysis including in empirical IO (see for example Jiang, Manchanda and Rossi, 2009).[29] The methods are well suited to deal with discrete response variables and the more complex likelihood structure arising from the joint modeling of marijuana use and access via the data augmentation approach. In addition, the Bayesian approach provides a natural framework to implement our counterfactual analysis of marijuana use under legalization. Specific to our context, the Bayesian approach enables us to address the issue of dealing with unobserved individual prices in a realistic and flexible approach described above which further complicates the form of the likelihood in the models by requiring the integration of the joint distribution of access and use for access subjects over the prices using the constructed empirical distribution of the weighted average price.

As described in detail in Appendix A, since the estimation of the model by Bayesian simulation methods exploits the conditional structure of the likelihood (and posterior distribution), we can apply standard simulation techniques to estimate the posterior distribution of the model parameters by simulating the prices from the empirical distribution at each iteration of our algorithm. Similarly, we can implement our predictive analysis using individual specific prices from market specific distributions where needed for counterfactual scenarios. In the remainder of the paper we report the means and standard deviations of the parameters and significance based on the posterior credibility intervals based on the draws from the posterior distributions obtained from the MCMC algorithm described in Appendix A. For the counterfactual use results based on the predictive analysis we report the means and standard deviations based on the draws from the predictive distributions of the probability of use.

3.5 Exclusionary Restrictions

To allow for identification of the parameters of our model due to data variation (rather than model non-linearities) we implement exclusionary restrictions. There are two variables that we argue impact access (via the effect on growing seasons or availability) but should not affect the utility of consuming conditional on access. These are the average mean temperature in the market and whether the consumer lives in a major city. Marijuana growing seasons are impacted by the temperature and hence the temperature in the state or time period should impact the supply available to purchase. In addition, marijuana is usually grown in sparsely

[29] We also estimated the baseline models using frequentist MLE methods and Bayesian methods using average market prices. We obtain the same results for the parameter point estimates (mean) and standard error (standard deviation) up to three decimal places of precision under both estimation approaches.

populated areas ("the outback") and hence it is easier to obtain outside of cities. The utility an individual obtains from using marijuana is a function of a variety of demographic characteristics, and we argue that whether an individual lives in a city does not *per se* influence the benefit from using conditional on other demographics. However, if there is something about living in a city or the temperature that impacts the utility obtained from using marijuana (such as a stressful environment) then our exclusionary restrictions would not be valid.

Probit Estimates
using subsample of Individuals who were given marijuana

Individual Attributes		Market and Policy Variables	
Male	0.342 **	Price	-0.002
	(0.015)		(0.001)
Age in Teens Spline	0.124 **	High Potency	-0.168
	(0.018)		(0.154)
Age in Twenties Spline	-0.036 **	Impact of Illegality	-0.561 **
	(0.004)		(0.025)
Age in Thirties Spline	-0.039 **	Grams Possession is not Minor Offense	0.024 **
	(0.003)		(0.011)
Age in Forties Spline	-0.028 **	Average Total Rainfall (in mm)	0.040 **
	(0.003)		(0.016)
Age over Forties Spline	-0.076 **		
	(0.005)	Exclusionary Restrictions	
Highest Education is High School	-0.057 **	Average Mean Temperature	-0.056
	(0.025)		(0.037)
Highest Education is Trade Degree	0.025	Live in City	-0.015
	(0.021)		(0.017)
Highest Education is University Degree	-0.069 **		
	(0.024)		
Of Aboriginal Descent	0.147 **		
	(0.051)		
In Good, Very Good, or Excellent Health	-0.264 **		
	(0.015)		
Number of Observations	13236		

Notes: Standard deviations in parentheses. * (**) indicates 90% (95%) Bayesian confidence interval does not contain zero. Includes state fixed effects and year fixed effects.

Table 5: Exclusionary Restrictions Validation

Fortunately, we are able to test the validity of the exclusionary restrictions by considering a subset of the individuals that were offered the drug. The presumption being that these individuals do not need to search, so this subsample should be relatively free of the selection problem. We estimated the use model with this subsample of individuals and included all explanatory variables to examine if temperature and living in a city are insignificant.[30] The results are presented in Table 5. As the table indicates both exclusionary restrictions are insignificant in the use equation for the subset of individuals that were offered the drug.

[30] We thank an anonymous referee for this suggestion.

Hence, these results suggest “living in a city” and “temperature” have some validity as exclusionary restrictions.[31]

While it is not necessary for identification, we have also included some variables in the usage equation but not the selection equation. These include price, high potency, health status, and legality variables. The main motivation for excluding price and potency from the selection equation is that consumers who do not have access are unlikely to know the price or potency of the marijuana they would obtain. Regarding legality variables, a reasonable concern is that the legality status and punishment for using marijuana could deter individuals from searching and hence should not be excluded as they may impact access. However, we note that the legality status variable is derived from questions regarding use such as “Did the fear of legal consequences influence your decision never to use marijuana” and “Would you try cannabis/marijuana if it were legal.” Therefore, the variable does not capture the effect of the current legality status which is what would likely impact search costs and hence access. Hence, we excluded legality impact from the selection equation. The variable “Grams possession is a minor offense” varies across states and may impact access if it translates into less search because the fines are too high, in which case it should not be excluded from the access equation. However, our selection equation includes state fixed effects which will capture differences in enforcement across states, and which will also pick up differences in fines across states. This motivated our choice to include penalties in the use equation. We also estimated the model with these excluded variables in the selection equation and this yielded no noticeable changes in any coefficient estimates on individual attributes and market and policy variables.

4 Results

In this section we discuss the role that access plays in marijuana use and the importance of correcting for selection into use. To do so we first compare the results from the selection model to estimates from the models that do not consider the role of selection explicitly. We then examine age-related differences in sensitivity to policy variables such as price and legality in the selection model. In all selection model specifications we use the intermediate

[31] Rainfall also influences growing conditions and hence may impact access to marijuana. However, as Table 5 shows, the text of the exclusionary restrictions indicates rainfall does impact utility on a selected sample that is free from selection. Therefore, we include it in both the use and access equations.

definition of the level of access (i.e., Access 2 in Table 2). We present robustness checks using the other definitions of access in Section 6.

Table 6 presents results from two probit models of marijuana use (specifications P1 and P2) and results from the baseline model corrected for selection (specification 1). The P1 specification includes a dummy variable indicating whether marijuana use is decriminalized in the market, while the others include state fixed effects. As we discussed earlier, previous literature has not accounted for restricted access and selection, therefore we refer to the results from the probit models as results from the standard approach. The simple probit models are based on the naive observed consumption variable c_{im} as defined in section 3.2 that includes the zeros from no-access subjects. Hence, even in the absence of selection on unobservables, the simple probit will yield biased estimates of the structural use parameters if access is a function of observables that also affect use.

Results from all specifications indicate that males and individuals in their teens and twenties are more likely to use marijuana relative to females and other age categories and that use is declining with age. They also indicate that aboriginal individuals are more likely to use, while those who report being in better health are less likely to use. In addition, individuals with only a high school education or those with a university degree are less likely to use relative to other education groups. For specifications P1 and P2, estimates of individual attribute parameters are similar. However, estimates vary with respect to market variables, which, for the probits, could be attributed to differences across states that are not controlled for in P1 (for example, variation in enforcement of marijuana laws). We focus on the differences between the standard P2 model (which includes state fixed effects) and the baseline selection model for the comparison.

Estimates from the baseline selection model reported in Table 6 illustrate that access is not randomly distributed across individuals, the same observables impact access and use conditional on access. Also results indicate that, conditional on age, individuals whose highest education is a trade degree are more likely to have access. In addition, there is selection on unobservables. This is reflected by the fact that the distribution of the correlation in unobservables (ρ) is positive. It is centered at 0.2 with the 95% Bayesian confidence interval excluding zero. These results indicate that it is important to correct for selection in marijuana use because accessibility is non-random across individuals.

	Standard Model Probit of Use		Selection Model Bivariate Probit with Selection	
	(P1)	(P2)	(1)	
			Use	Access
Individual Attributes				
Male	0.341 **	0.342 **	0.316 **	0.282 **
	(0.015)	(0.015)	(0.021)	(0.012)
Age in Teens Spline	0.122 **	0.124 **	0.111 **	0.138 **
	(0.017)	(0.018)	(0.020)	(0.016)
Age in Twenties Spline	-0.036 **	-0.036 **	-0.031 **	-0.038 **
	(0.003)	(0.004)	(0.004)	(0.003)
Age in Thirties Spline	-0.039 **	-0.039 **	-0.032 **	-0.040 **
	(0.003)	(0.003)	(0.004)	(0.003)
Age in Forties Spline	-0.028 **	-0.028 **	-0.025 **	-0.020 **
	(0.003)	(0.003)	(0.004)	(0.002)
Age over Forties Spline	-0.076 **	-0.076 **	-0.070 **	-0.053 **
	(0.005)	(0.005)	(0.007)	(0.003)
Highest Education is High School	-0.061 **	-0.058 **	-0.073 **	-0.003
	(0.025)	(0.025)	(0.028)	(0.020)
Highest Education is Trade Degree	0.025	0.025	-0.018	0.112 **
	(0.021)	(0.021)	(0.024)	(0.015)
Highest Education is University Degree	-0.076 **	-0.069 **	-0.069 **	-0.019
	(0.023)	(0.023)	(0.026)	(0.017)
Of Aboriginal Descent	0.151 **	0.147 **	0.130 **	0.181 **
	(0.050)	(0.050)	(0.058)	(0.046)
In Good, Very Good, or Excellent Health	-0.262 **	-0.264 **	-0.238 **	
	(0.015)	(0.015)	(0.017)	
Market and Policy Variables				
Price	0.001	-0.002	-0.004 **	
	(0.001)	(0.001)	(0.002)	
High Potency	-0.277 *	-0.17	-0.139	
	(0.153)	(0.156)	(0.176)	
Impact of Illegality	-0.562 **	-0.561 **	-0.481 **	
	(0.024)	(0.025)	(0.028)	
Grams Possession is not Minor Offense	0.0003	0.024 **	0.005 *	
	(0.0004)	(0.011)	(0.003)	
Average Total Rainfall (in mm)	0.026 **	0.041 **	0.056 **	-0.040 **
	(0.005)	(0.016)	(0.017)	(0.012)
Average Mean Temperature	0.014 **	-0.057		-0.267 **
	(0.002)	(0.036)		(0.018)
Live in City	0.013	-0.015		-0.188 **
	(0.016)	(0.017)		(0.013)
Decriminalized	0.127 **			
	(0.021)			
Correlation (ρ)			0.230 **	
			(0.116)	
State Fixed Effects Included	No	Yes	Yes	Yes

Notes: Standard deviations in parentheses. * (**) indicates 90% (95%) Bayesian confidence interval does not contain zero. All specifications include year fixed effects and a constant in access and use. Number of observations is 52,197.

Table 6: Estimates of Baseline Selection Model and Standard Probits

Selection results further indicate that accessibility is declining with increased rainfall and higher temperatures, both of which adversely impact the growing seasons. The results also show that individuals that live in a city are less likely to have access to marijuana. This is consistent with the reported growing patterns of marijuana in Australia, where it is usually grown in sparsely populated areas and hence it is easier to obtain outside of cities.

There are important differences between the probit models and the selection model regarding the impact of policy variables on use. Most notably, the P1 and P2 models suggest that individuals are less sensitive to price and more sensitive to legalization laws than the selection model indicates. This is of particular importance as these are market specific variables that the government can control through policy. The selection model indicates participation into using marijuana is more elastic with respect to price: a 10% increase in price would reduce marijuana smoking rates by 2.0% while the simple probit estimates indicate the probability of use is not significantly impacted by price. The magnitude of the participation elasticity (i.e, the probability of use with respect to price) from the selection model (-0.20) is consistent with estimates of cigarette participation elasticities from prior studies (which range from -0.25 to -0.50).[32] This similarity is not surprising as marijuana is combined with tobacco when consumed in Australia. Furthermore, as foreshadowed by the estimated coefficients, the simple probit estimates substantially overstate the elasticity of participation with respect to legality. The probit elasticities imply the impact of a change in legal status would have two times as large an impact on the probability of using marijuana (elasticity of -0.22) than that predicted by a model that corrects for selection (elasticity of -0.11).

Table 7 presents selected parameter results of selection models with age interactions. All specifications include the same control variables as those from Table 6. For ease of comparison we reproduce the relevant results for the baseline specification 1 from Table 6. Specification 2 shows results with price and age interactions. The results indicate that there is variation in price sensitivity across age groups. Individuals in their twenties and thirties are the most price sensitive age groups. This implies that increases in prices (via a tax, for example) will have less of an impact on teens but will otherwise influence use among young individuals.[33] Specification 3 results indicate that there is age variation in the disutility of participating in illegal activities. Teenagers and individuals in their twenties exhibit the most sensitivity to the legal status of marijuana of all age groups. The last specification contains interactions of age with price and legality. The results mirror those of the previous specifications. Overall, the findings indicate that variables associated with price and legality (two policy instruments)

---

[32] See the literature review in Chaloupka et al (2002) and Chaloupka and Warner (2000).

[33] If teens are more occasional users relative to those in their twenties and thirties, this would explain why they would be less influenced by price changes.

will both have an impact on individuals in their twenties and thirties relative to other age groups, but only the latter will influence teen use.

	Selection Models with Interactions of:							
	No interactions		Price and Age		Illegality and Age		Illegality, Price and Age	
Specification:	(1) from Table 6		(2)		(3)		(4)	
	Use	Access	Use	Access	Use	Access	Use	Access
Age Splines								
Age in Teens	0.111 **	0.138 **	0.129 **	0.139 **	0.116 **	0.138 **	0.130 **	0.138 **
	(0.020)	(0.016)	(0.025)	(0.016)	(0.021)	(0.016)	(0.025)	(0.016)
Age in Twenties	-0.031 **	-0.038 **	-0.023 **	-0.038 **	-0.033 **	-0.038 **	-0.025 **	-0.038 **
	(0.004)	(0.003)	(0.006)	(0.003)	(0.004)	(0.003)	(0.005)	(0.003)
Age in Thirties	-0.032 **	-0.040 **	-0.032 **	-0.040 **	-0.033 **	-0.040 **	-0.033 **	-0.040 **
	(0.004)	(0.003)	(0.006)	(0.002)	(0.004)	(0.002)	(0.005)	(0.003)
Age in Forties	-0.025 **	-0.020 **	-0.034 **	-0.020 **	-0.024 **	-0.020 **	-0.035 **	-0.020 **
	(0.004)	(0.002)	(0.006)	(0.002)	(0.004)	(0.002)	(0.006)	(0.002)
Age over Forties	-0.070 **	-0.053 **	-0.074 **	-0.053 **	-0.073 **	-0.053 **	-0.075 **	-0.053 **
	(0.007)	(0.003)	(0.009)	(0.003)	(0.007)	(0.003)	(0.009)	(0.003)
Price and Interactions:								
Price	-0.004 **				-0.003 **			
	(0.002)				(0.002)			
Age in Teens			-0.001				-0.001	
			(0.002)				(0.002)	
Age in Twenties			-0.003 *				-0.003 *	
			(0.002)				(0.002)	
Age in Thirties			-0.005 **				-0.005 **	
			(0.002)				(0.002)	
Age in Forties			-0.003				-0.003	
			(0.002)				(0.002)	
Age over Forties			-0.001				-0.002	
			(0.002)				(0.002)	
Illegality Interactions:								
Impact of Illegality	-0.481 **		-0.478 **					
	(0.028)		(0.030)					
Age in Teens					-0.498 **		-0.505 **	
					(0.072)		(0.073)	
Age in Twenties					-0.591 **		-0.590 **	
					(0.051)		(0.051)	
Age in Thirties					-0.393 **		-0.372 **	
					(0.052)		(0.052)	
Age in Forties					-0.484 **		-0.497 **	
					(0.068)		(0.068)	
Age over Forties					-0.330 **		-0.328 **	
					(0.090)		(0.093)	

Notes: Standard deviations in parentheses. * (**) indicates 90% (95%) Bayesian confidence interval does not contain zero.
Includes all controls in Table 6 including individual attributes, year and state fixed effects in use and access.

Table 7: Selected Parameter Estimates for Price, Age, and Illegality Interactions

The estimates from Tables 6 and 7 concern the extensive margin of marijuana use. We present the results from the selection model of frequency of use for the price-age interaction specification in Table 8. The results indicate that price increases would significantly decrease frequency of use and that the impact of price increases on frequency of use varies across age groups. The price elasticity of demand for the frequency of use model indicates a 10%

increase in price would reduce use frequency by on average 1.7%.[34] The results indicate that infrequent users would reduce their use frequency by 1.2% on average, while frequent users would reduce their use by 2.1%. While there alternative ways to measure intensity of use making direct comparisons across the literature difficult, our results are nonetheless consistent with estimates found in studies of cigarette price elasticity of demand.[35] Demographics and other market variable estimates exhibit similar patterns as their corresponding estimates from the extensive margin of use selection model. We use the frequency parameter estimates to compute the tax revenues raised under a variety of counterfactual scenarios discussed in section 5.3.

	Frequency Selection Model Ordered Probit with Selection					
	Use			Access		
	Mean	Std Dev		Mean	Std Dev	
Individual Attributes						
Male	0.326	(0.020)	**	0.282	(0.012)	**
Age in Teens Spline	0.100	(0.020)	**	0.138	(0.016)	**
Age in Twenties Spline	-0.023	(0.004)	**	-0.038	(0.003)	**
Age in Thirties Spline	-0.028	(0.004)	**	-0.040	(0.003)	**
Age in Forties Spline	-0.026	(0.004)	**	-0.020	(0.002)	**
Age over Forties Spline	-0.067	(0.007)	**	-0.053	(0.003)	**
Highest Education is High School	-0.124	(0.027)	**	-0.003	(0.020)	
Highest Education is Trade Degree	-0.065	(0.023)	**	0.113	(0.015)	**
Highest Education is University Degree	-0.144	(0.026)	**	-0.019	(0.017)	
Of Aboriginal Descent	0.157	(0.055)	**	0.181	(0.047)	**
In Good, Very Good, or Excellent Health	-0.259	(0.017)	**			
Market and Policy Variables						
Price Age in Teens Interaction	-0.002	(0.002)				
Price Age in Twenties Interaction	-0.003	(0.002)	*			
Price Age in Thirties Interaction	-0.004	(0.002)	**			
Price Age in Forties Interaction	-0.002	(0.002)				
Price Age over Forties Interaction	-0.004	(0.002)	**			
High Potency	-0.002	(0.170)				
Impact of Illegality	-0.489	(0.029)	**			
Grams Possession is not Minor Offense	0.005	(0.003)	*			
Average Total Rainfall (in mm)	0.054	(0.017)	**	-0.040	(0.012)	**
Average Mean Temperature				-0.269	(0.018)	**
Live in City				-0.187	(0.013)	**
Correlation (ρ)	0.118	(0.087)				
Cut-off (τ)	0.478	(0.012)	**			

Notes: Standard deviations in parentheses. * (**) indicates 90% (95%) Bayesian confidence interval does not contain zero. Includes year fixed effects, state fixed effects, and a constant in access and use. Number of observations is 52,184.

Table 8: Frequency of Use Estimates

[34] This is a weighted average of the elasticities for infrequent and frequent users.

[35] The International Agency for Research on Cancer provides an overview of the literature and reports price elasticities in the range of -0.2 and -0.6. See http://www.iarc.fr/en/publications/list/handbooks/

5 Policy Analysis

We use the results from the selection model to investigate the effect of legalization and to improve our understanding about individual decision making in that context. We aim to address the following policy concerns: (i) what role does access play in marijuana use; (ii) what role do other factors (such as demographic characteristics, illegality of the drug, prices, etc.) play in the decision to use the drug, (iii) can we use policy to restrict use among young adults, and (iv) how does legalization impact tax revenues.

5.1 Impact of Accessibility and Legalization on Use

We decompose the impact of legalization in three ways: the part of the increase in use due to increased accessibility, the part due to the removal of the stigma associated with breaking the law, and the part due to potential changes in prices (due to supply side cost changes or tax policies). If marijuana were legalized, then accessibility would not be a hurdle; in the model this implies $\phi_{im} = 1$. In addition, the disutility associated with illegal activity would be zero; in the model this implies $L_{im}^{illegal} = 0$. Furthermore, dealers would no longer face penalties for selling. To address this issue, we compute the counterfactuals under various assumptions about how price would change: (i) price would not change; (ii) price would increase by 25%; (iii) price would decline to the price of cigarettes; and (iv) price would decline to the marginal costs of production. Notice that since we do not model the supply side prices are taken as exogenous. In all scenarios, we change the environment and compute the predicted probability of use that would arise in the counterfactual world implied by the parameter estimates from the selection model, focusing on specifications 1 and 4.[36]

We discuss our choice of counterfactual prices in turn. The first scenario (no change in price) is not realistic, however it serves as a benchmark for other counterfactuals. Scenario (ii), a 25% price increase, is motivated by tax proposals made in the United States, where legalization laws were recently passed. Specifically, in 2013 Colorado and Washington legalized marijuana use for recreational purposes. Amendment I-502 requires state lawmakers to establish a system of state-licensed growers, processors and retail stores, where they propose to tax marijuana 25%. Scenario (iii) is more reasonable as marijuana is typically mixed with

[36] Notice that our model allows for selection on unobservables, as well as observables. When we estimate the model each person will have a vector of realizations of the unobserved term from each iteration of the MCMC algorithm (which is correlated with use according to $\widehat{\rho}$). We use those unobserved terms to correct for selection on unobservables when computing the counterfactuals.

tobacco in Australia. Finally, the last scenario, pricing at marginal cost, serves as a lower bound on the price of marijuana. We use marijuana marginal production cost estimates reported in Caulkins, et. al. (2011). These estimates are based on the costs for growing other herbs (e.g., the price of plants, growing fertilizer, labor, etc.).[37]

We first estimate the probabilities of use under the various counterfactual scenarios for access, legalization status and price for our baseline selection model using the Bayesian predictive approach. The prediction is based on the conditional probability of use implied by the selection model in order to take into account the role of selection on unobservables on use in addition to the effects of observables such as price, demographic, market and legality variables.[38] Legality and price variables are adjusted to predict use under legality and different policy scenarios. Under the Bayesian approach we (i) estimate the counterfactual use probabilities exploiting all information on the estimated parameters summarized in the posterior distribution, (ii) obtain the full distribution of the predicted use and (iii) implement the counterfactual analysis for subgroups that differ, for example, by reported access before legalization and age, in a coherent manner. Further details are provided in Appendix A.2. In the remainder of the section we report the point estimates in terms of the means and standard deviations of the counterfactual use probabilities for the overall population as well as for different subgroups.

Table 9 displays the counterfactual results which indicate that both access and illegality concerns play substantial roles in the decision to use marijuana. The first row replicates the data under the current legal environment. The second row shows how the probability of use would change if accessibility were not an issue in an environment where use was still illegal. That is we assume all other aspects of the counterfactual world stay the same other than access, so we recompute the probability of use assuming that $\phi_{im} = 1$ for all individuals. In this scenario, the probability of use among current non-users without access would increase

[37] Caulkins reports that, in the US, wholesale prices range from \$500 to \$1500 per pound. Due to electrical usage costs of growing hydro are higher, between \$2000–4500 per pound. If cannabis is grown outside production costs are estimated to be less than \$20 per pound. The costs in Australia are likely to be of the same magnitude as the costs of low-skilled labor and raw inputs in the US.

[38] The conditional probability of use is $\Phi(\mu^u_{im} + \rho\eta_{im})$ which is the standard normal c.d.f. $\Phi(\cdot)$ evaluated at the mean that consists of two terms. The first term (μ^u_{im}) is the mean of the marijuana use model and accounts for preferences. Under the full interaction specification this term is $p_{im}\alpha_1 + p_{im}d_i^{age\prime}\alpha_2 + d_i'\beta_1 + x_m'\beta_2 + L_{im}'\delta_1 + L_{im}^{illegal}d_i^{age\prime}\delta_2$ with age cohort specific effects of price and illegality. The second term corrects for selection on unobservables where the effect of the unobservables is captured by ρ. A detailed description is given in Appendix A.2 where we also describe how the unobservable term can be obtained based on reported access and the estimated access model.

to 10% resulting in an overall increase of 37% in the probability of use (from 13.1% to 18%). If marijuana were legalized (i.e., we set $L_{im}^{illegal} = 0$) and accessibility were not an issue, then use would increase by 48% to 19.4%. Obviously there would be an impact on prices due to the law change, and the results show taxes of 25% are effective to offset the increase in use due to the legal status change. Interestingly, while the overall probability of use would be 40% higher than current levels (at 18.3%), the policy of legalization with a 25% tax on current prices would not impact the behavior of users who currently have access (probability of use moves from 25% to 25.1%). This particularly highlights the significant role played by access and the importance of considering selection into access on the prevalence of use.

Environment			Predicted Probability of Use For Current Consumers: All		With No Access		With Access	
Accessible	Legal	Price	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
No Change	No	No Change	0.131		0.000		0.250	
Accessible	No	No Change	0.180	(0.129)	0.103	(0.080)	0.250	(0.127)
Accessible	Yes	No Change	0.194	(0.130)	0.116	(0.082)	0.264	(0.125)
		25% Increase	0.183	(0.126)	0.108	(0.078)	0.251	(0.122)
		Cigarette	0.238	(0.147)	0.149	(0.099)	0.319	(0.137)
		Cost	0.241	(0.148)	0.151	(0.100)	0.322	(0.137)

Notes: These use estimates from the baseline specification 1 in Table 6. The first row is a prediction for a person with the typical access characteristics. All 95% Bayesian Prediction Intervals exclude zero.

Table 9: Counterfactual Use Results

5.2 Legalization and Use among Young Adults

Finding ways to limit use of drugs among young adults is an important issue in the legalization debate. As our estimates from the various specifications of the selection model highlight, age plays an important role in access and use. In addition, the results from Table 7 show that the impact of prices and legality varies by age group. Therefore, we use the estimates from the model with interactions of age group with prices and legality (specification 4 in Table 7) to compute counterfactual use probabilities by age group. This allows us to conduct various age-specific counterfactuals which yield insight into the prevalence of use among youths in a legalized setting.

The counterfactual results indicate that if marijuana were freely accessible at the current prices then we would see an increase in the probability of use of 37% on average, but, as Table 10 shows, this has the least impact on individuals in their teens and twenties, where the probability of use increases by only 20% on average (from 25% to 30%). If, in addition to being accessible, it is no longer illegal to use marijuana the probability of use increases by 48% on average, but again this has the largest impact on individuals 30 and older where the use probability increases by 67% on average. However, prices will not remain constant - a tax of 25% over the current price would see the probability of use increase only by 40% over current use on average, while individuals under 30 would see a more modest increase in the probability of use of 28% on average.

Environment			Predicted Probability of Use For Individuals in Age Bracket:				
Accessible	Legal	Price	Teen	Twenties	Thirties	Forties	Fifty or Older
No Change	No	No Change	0.251	0.251	0.145	0.099	0.037
Accessible	No	No Change	0.304	0.300	0.195	0.149	0.067
			(0.148)	(0.142)	(0.111)	(0.095)	(0.059)
Accessible	Yes	No Change	0.337	0.323	0.207	0.161	0.072
			(0.134)	(0.131)	(0.108)	(0.093)	(0.060)
		25% Increase	0.333	0.312	0.192	0.154	0.07
			(0.135)	(0.130)	(0.104)	(0.090)	(0.059)
		Cigarette	0.356	0.371	0.272	0.190	0.081
			(0.139)	(0.139)	(0.125)	(0.103)	(0.066)
Smoked Cigarette in the Past Year			84.0%	70.8%	51.5%	43.1%	32.8%
Daily Cigarette Smoker			25.2%	34.5%	31.1%	29.4%	23.3%
Report Current Access to Marijuana			70.6%	72.3%	57.6%	47.5%	32.5%

Notes: This is a prediction of use for a person with the typical access characteristics using estimates from the state fixed effects specification with age interacted with prices and legality (spec 4 in Table 7). Standard deviations are in parenthesis; All 95% Bayesian Prediction intervals exclude zero.

Table 10: Counterfactual Use Results by Age Group

These results reflect the fact that the impact of accessibility on use probability differs considerably by age group. Those in their teens and twenties exhibit almost the same pre-legal levels of access to marijuana (70.6% and 72.3%, respectively) and thus react similarly to the removal of the access barrier. In contrast, for older age groups, with lower access under illegality, the impact of accessibility alone leads to a larger proportional increase in use.

In addition to variation across the mean in use, the shape of the probability of use distribution varies by age across all legalization scenarios. This is illustrated in Figure 1 that presents the distributions of the probability of use under the counterfactual of legalized marijuana with 25% higher price for the different age groups. The age distributions are centered at different means and also exhibit different shapes. Specifically, we see that the distribution of use among individuals below their teens and twenties (pink line) are more dispersed and symmetric compared to the distributions for individuals in their thirties (blue line), their forties (red line) and those fifty or older (green line). The three oldest age groups have distributions that are increasingly skewed to the right with most mass over small probabilities of use.

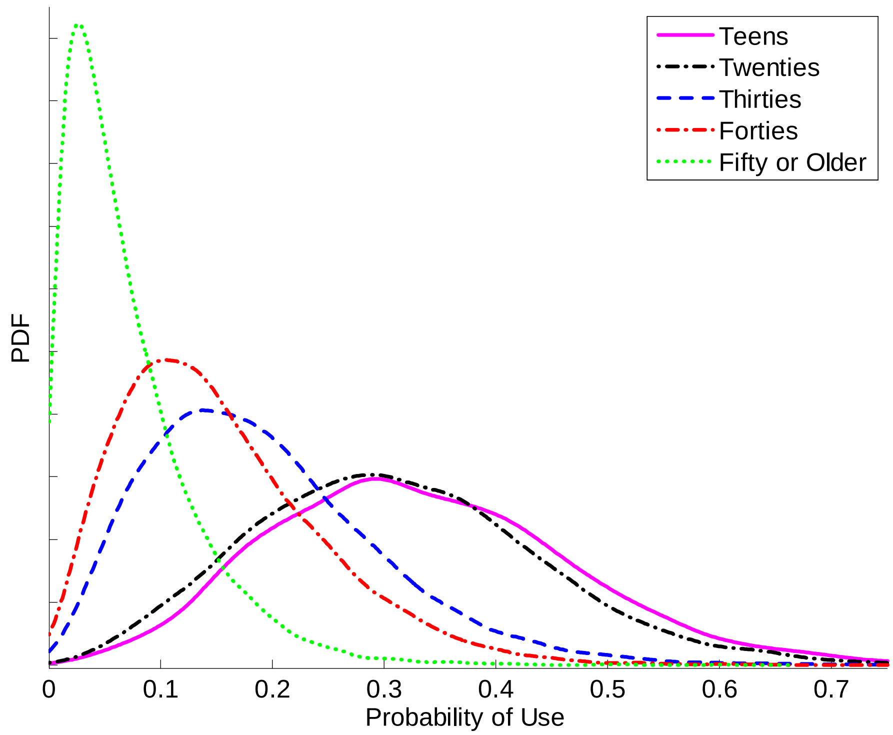

Figure 1: Predicted Use Probablity Distributions (Legalized; 25% Price Increase)

Even though our results show that teenagers and young adults have lower percentage increases in use probabilities, they still remain the age groups with the highest probabilities of use. This raises concerns for legalization opponents and questions for policy makers with regard to possible interventions. For example, one important policy tool is taxes, so it would be worthwhile to know to what extent taxes may be used to curb use, in particular among the most vulnerable group of teenagers. We explore this issue by conducting a further set of "price counterfactuals," that address the question how much taxes would need to be implemented to return the post-legalization predicted probability of underage use (of 33.7%) to the pre-legalization levels (of 25.1%). Since use would remain illegal for teenagers (as it is for alcohol), we consider an environment where access is not restricted but use is illegal.[39]

We find that the average price per gram would have to be $158, about four times the current level, in order for only 25.1% of teenagers on average to use in a post-legalized world

[39] We describe the procedure in detail in Appendix B.

(that is to experience none of the 8.9 percentage points increase in use among underage users). Increasing prices by four-fold is not feasible given that we would expect most users to resort to the black market. However, the current proposed tax increase of 25% is sufficient to realize 34% of the goal (where two-thirds of these individuals are female). In order to move the probability of post-legalization non-use 40% closer to pre-legalization levels, prices would have to almost double; and they would have to almost triple to move 50% closer to the pre-legalization level of 25.1% of use among the underage population. Hence, our results indicate that, in a post-legalization world we will see an increase in the probability of underage use of 34% on average.

5.3 Tax Revenues and the Black Market

We use the frequency of use estimates (presented in Table 8) to compute annual tax revenue under two taxation regimes $r = 1, 2$. The first regime uses the cigarette tax rate assuming the base price is marginal cost. The second regime is motivated by the proposal in the US to tax marijuana at 25% over current prices. Notice the first scenario involves a lower price than currently paid and the second scenario a higher price. Together these two tax scenarios should provide a reasonable idea of the bounds on tax revenue that could be generated from sales under legalization.

In order to compute the tax revenues we must: generate a new price under each tax regime, determine with which frequency an individual would consume given the new price, generate an average quantity consumed per session, link the average quantity consumed with the frequency of consumption to determine the total quantity consumed by an individual under the relevant tax regime, and determine the tax revenue generated by individual i. The total tax revenue under regime r is obtained by summing the individual tax revenue over individuals. We discuss each of these elements in turn.

To obtain the new prices faced by individuals under these tax regimes we follow the same strategy as outlined in Section 3.1 as closely as possible to obtain individual specific market price under taxation regime r (denoted p^r_{im}). Under tax regime 2, the only change is that the distribution of the market prices by type (given in equation 6) are no longer centered at the prices of marijuana by type (p_{mt}), but instead are centered at $1.25p_{mt}$. Finally, since there are not three types of cigarettes to mimic marijuana types, we adjust the raw cigarette prices using market level information on the proportion of high potency marijuana used. We then construct market-level empirical distributions for p^1_{im} in terms of a truncated normal distribution centred at the adjusted cigarette price.

Given the new prices, and taking into account selection on unobservables, we compute the probability that an individual's consumption falls into one of the three frequency categories. Specifically, let $\widehat{G}_{ikr}$ denote the probability that individual i's predicted frequency of use under tax scenario r falls in category $k = 0, 1, 2$, which depends on the utility of use (using the new prices p^r_{im}) and an additional term for the selection on unobservables: $\widehat{G}_{i0r}$ refers to the predicted probability of no use, $\widehat{G}_{i1r}$ the predicted probability of infrequent use and $\widehat{G}_{i2r}$ to the predicted probability of frequent use, with $\sum_k \widehat{G}_{ikr} = 1$.

To determine the average amount consumed per session we use data from the NDSHS on the average quantity consumed for those subjects in our sample who consume. Analogous to our approach for generating unobserved individual prices in Section 3.1, we use these data to construct an average amount consumed per session in each market q^{obs}_m. Then for each person (even individuals who did not consume pre-legalization) we draw an average quantity consumed (q_{im}) from a truncated normal that is centered at the observed market average quantity with a variance set using information on the variance in the data within a market. Specifically,

$$q_{im} \sim TN_{(0,\infty)}(q^{obs}_m, \Omega^q_m) \tag{13}$$

The total quantity consumed under frequency k, Q_{ik}, is given by

$$\mathrm{Q}_{ik} = \begin{cases} 0 & \text{when } k = 0 \\ [1,4] * q_{im} & \text{when } k = 1 \\ [12,365] * q_{im} & \text{when } k = 2 \end{cases}, \tag{14}$$

which is zero under no use and computed based on the number of sessions per year associated with the frequency and the average amount consumed per session (q_{im}) for infrequent ($k = 1$) and frequent ($k = 2$) use. An infrequent user is one who uses once quarterly, biannually, or annually. A frequent user is one who uses monthly, weekly, or daily. The intervals represent the lower and upper bounds on the units of consumption associated with the frequency definition (e.g., the interval takes on the lower value of 12 for a frequent (monthly) consumer who consumes 12 times per year). We compute tax revenues implied by the midpoint and the upper and lower bounds of these intervals.[40] Notice that the implicit assumption is that the average amount consumed per using session does not change. That is, we assume price changes influence quantity through use frequency (that is into which k category an individual falls) but not the average amount consumed per session. For example, perhaps a user smokes a marijuana joint during a party once per month. We assume that the post tax price may change the frequency with which the user smokes (to once every few months for example) but when he smokes he still consumes one joint.

We can then obtain the annual individual tax revenue for individual i that would be realized under the prices p_{im}^{r} according to

$$\text{Tax Revenue}_{ir} = \sum_{k} \widehat{G}_{ikr} \ \text{Q}_{ik} \ Tax_r,$$

where Tax_r is the tax paid under regime r. For regime 1 this is the difference between the (quality adjusted) cigarette price (p_{im}^{1}) and the (type-specific) marginal cost of marijuana production and for regime 2 this is the difference between p_{im}^{2} (drawn with 25 % higher prices on average) and the pre-legalization price, p_{im}. For further details regarding the above discussion please see Appendix A.3.[41]

Total tax revenue for tax regime r is obtained by summing Tax Revenue_{ir} over individuals. For Australia (and for US under the assumption that Americans are similar to Australians), our results indicate that a policy of marijuana legalization would raise a minimum amount between \$77 million to \$220 million (\$1 billion to \$3 billion) annually, depending on which taxation scheme is used and assuming individuals consume at the lower level of the frequency interval.[42] It is less likely all individuals consume at the upper end of the frequency interval (i.e., to do so would mean all monthly users are treated as daily users), but if individuals consume at the midpoint of the frequency intervals then tax revenues would increase to between \$320 million to \$915 million (\$4 billion to \$12 billion) annually depending on the

[40] Note that we compute three separate quantities consumed per person: one computed at lower bound of frequency (so in this example 1 and 12), one at mean of interval; and one at the upper bound of the interval (4 and 365).

[41] The frequency model predicts a certain probability a person lies in each of the 3 intervals. For example, a hypothetical person falls into category $k = 0$, 20% of the time; into $k = 1$, 70% of the time; and into $k = 2$, 10% of the time. So the quantity consumed for this person is 0, 20% of the time; $[1, 4] * q_{im}$, 70% of the time; and $[12, 365] * q_{im}$, 10% of the time. To compute the average tax per user we consider a person a user if the predicted frequency they fall into category $k = 0$ is lower than 50%.

[42] These calculations are based on population in 2014 as reported by http://worldpopulationreview.com/countries/.

taxing scheme. We should note that our findings are consistent with those from a 2005 report funded by the Marijuana Policy Project (Miron, 2005) which estimates legalization would raise tax revenues of $2.4 billion annually in the United States if it were taxed like most consumer goods and over $6 billion annually if it were taxed similarly to alcohol or tobacco.

According to the Framework Convention on Tobacco Control (2012), illicit trade in cigarettes accounts for approximately one-tenth of global sales. Likewise, it is reasonable to conjecture that some marijuana users will purchase from the black market, especially if tax rates are high. We compute an adjusted tax revenue that allows for some sales to be lost to the black market. If we assume that 10% of the sales will be lost to the black market, tax revenues for Australia (the US) would decline to between $70 million to $823 million ($872 million to $10 billion). We also computed how much tax revenue would be raised if all users who currently use (i.e., those who are currently willing to do something illegal) would buy on the black market instead of in the legal market. In this situation tax revenues would be between $61 million to $727 million ($763 million and $9 billion).

Colorado opened its first retail outlets in January 2014 and has a tax system similar to our tax regime 2. This gives us a nice experiment for the tax revenue predictions our model would generate for a state the size of Colorado. Our results based on the midpoint prediction is $68.2 million annually, which reduces to $61.5 million after losses to the black market are taken into account. Excluding medical marijuana and licensing fees, Colorado collected $56.1 million in taxes in 2014, which is close to the mid-range of our predictions.[43]

To summarize, in the worst case tax revenue scenario - all current users purchase on the black market - legalization in Australia (or the US) would still result in tax revenues of $61 million (over $700 million) annually. At the other extreme, the government would raise almost a billion ($12 billion) in taxes. Furthermore, governments would see cost reductions under legalization as they would not incur nearly as high of costs of enforcement.

6 Robustness Checks

We conducted a number of robustness checks of our results. First, given the importance of the role played by access in our results we reran our baseline and interaction specifications

[43] Our results do not include taxes from medical marijuana sales or from licensing fees, but in Colorado's case these made up a small portion of overall tax raised - including these fees Colorado raised $67.5 milllion in taxes in 2014. See https://www.colorado.gov/pacific/revenue/colorado-marijuana-tax-data

using two different definitions of the accessibility variable. The first access variable definition is more inclusive (Access 1 in Table 2) the second is more restrictive (Access 3 in Table 2). The results using either access definition are virtually identical to those in the main text of the paper - they indicate that there is selection on observables and the parameter estimates are almost identical for all variables. The only notable difference is that the aboriginal effect decreases and education university is no longer significant under Access 3.[44]

As we would expect, the correlation coefficient somewhat increases under the more stringent Access 3 variable (0.282) and decreases under the less stringent Access 1 variable (0.116), relative to our baseline Access 2 (0.230). Given that the more significant and somewhat larger positive correlation under Access 3 adds further evidence for the presence of unobservables that affect access and use, our key conclusions from Section 4 are supported by the robustness check.[45]

Even though we observe very little change in the model coefficient estimates under Access 1 and Access 3, the counterfactual use results will depend on the specific access variable since one of the key drivers of the increased use in the counterfactual scenarios is the removal of the access barrier. In particular for the tightly defined Access 3 variable, removing the accessibility barrier has a much larger effect as only 36% of subjects had access under illegality compared to 53% under Access 2 (and 59% under Access 1). Hence, under Access 3 the predicted probability of use increases more under legalization (to 0.25 rather than 0.18 under current prices). Under Access 1 we obtain almost identical results to those reported in Table 9. Changing the access variable definitions yields similar results on coefficients and significance for the interaction specifications reported in Table 7. The only notable change is that all education variables are significant in the access equation. For our counterfactual results in Table 10 based on specification 4 from Table 7, the relative patterns across the age groups remain the same under the alternative access specifications. Teenagers and those in the their twenties still behave in a similar manner, with slightly higher predicted probability of use for teens. Similarly, we see a slight increase in the probability of use for the other age groups compared to the Access 2 results.

[44] The parameter estimates are available on request.

[45] The correlation coefficient is also somewhat sensitive to the set of exclusion restrictions. Excluding either the city variable or both city and temperature (identification solely based on distributional assumptions), the ρ estimates becomes insignificant under the main access 2 variable. It remains positive and significant witout the inclusion of the city iv under the more restrictive access 3 variable in the baseline selection model.

We also conducted two model specification checks. In the first we consider the role of addiction or habitual use on current consumption. As discussed previously, our data are not longitudinal and hence we do not have information on use in previous periods so we cannot include a lagged (endogenous) use variable in the regression. However, we do observe in the data the frequency with which individuals use marijuana. Approximately 3% of the sample report using daily or that use is a habit. We reran the regressions excluding these individuals. The results are the same as those we obtain when we include this group, with one notable change: the variable "aboriginal" is no longer significant. These results suggest that our findings are not driven by the impact of habitual users. In the second specification check we consider that there may be individual characteristics that are not observed by the econometrician that impact the utility one obtains from marijuana use. We estimated specifications that include random coefficients on legality and prices. However, once we include demographic interactions there is not enough additional variation to identify the random coefficients.

Finally, we ran robustness checks of our results to the potential endogeneity of some covariates. The first concerns the endogeneity of health status where a potential concern is reverse causality - use influences health status. We reran our baseline specification without health status and there are no notable changes in the results. Second, as the prices are not individual reported purchase prices there may be some concern that price is correlated with the error term, and, therefore endogenous. We include a measure of the potency to control for quality to ameliorate this concern. As discussed in Section 1, prices are higher the higher is potency, which can be thought of as measure of the quality of the marijuana. In our setting for price endogeneity to be an issue it would be necessary for something unobserved (and hence in the error term) that impacts pricing decisions and that also matters to the consumer that is not related to quality. Furthermore, if we had access to individual price paid then there would have to be something that impacted marginal costs on an individual level that was endogenous to the demand side error term, which would make endogeneity less of a concern. Fortunately, in the 2007 wave of the data, respondents were asked to report the price of the most recent purchase (and the quantity purchased) and the quality of marijuana purchased. As these are individual prices reported by quality type they are less likely to be correlated with the error term. Unfortunately reported prices are only available in one wave so we cannot use them for the entire analysis. However, the estimates using reported prices for the 2007 wave of the data are not significantly different from those reported in Table 6, hence, we are not concerned that price endogeneity is an issue once quality of marijuana is accounted for.

7 Conclusions

We present a model of marijuana use that disentangles the impact of limited accessibility from consumption decisions based solely on preferences. We find that both play an important role and that individuals who have access to the illicit market are of specific demographics. Our results indicate that observables and unobservables from marijuana use and access are positively related and that the elasticities of legalization and price are all significantly different in the selection model relative to the standard approach. The selection model indicates demand is much more elastic with respect to price. Counterfactual results indicate that making marijuana legal and removing accessibility barriers would have a smaller relative impact on younger individuals but still a large impact in magnitude. The probability of use among underage youth would increase by 38% and the probability of use among individuals in their thirties and forties would more than double.

We found that prices would need to be four-fold higher than current levels in order to keep the frequency of post-legalization underage use the same as pre-legalization use even if underage users would still face the same restrictions as they face for alcohol use. Increasing prices by four-fold is not feasible given that we would expect most users to resort to the black market. However, the current proposed tax increase of 25% is sufficient to realize 34% of the goal (where two-thirds of these individuals are female).

For Australia, our results indicate that a policy of marijuana legalization would raise a minimum amount between $77 million to $915 million annually, depending on which taxation scheme is used. If all users who currently use (i.e., those who are currently willing to do something illegal) would buy on the black market tax revenues would be between $61 million to $727 million.

Our study provides insight on the potential impacts of legalizing marijuana use. Needless to say, there are many aspects to legalization that are not addressed in this study. One such important issue concerns the long-run implications of legalization. For example, social acceptance may change post-legalization which could have implications for use over the long-run. Another issue concerns product characteristics, it is reasonable that the "products" available will change post legalization. For example, perhaps the product line offerings will resemble those from medical marijuana stores - that is less smoke-based and more candy- or cookie-based. We don't provide a methodology that addresses what type of marijuana to consume but this is likely to play a role in the future as we may see an expansion in product line offerings. Finally, we do not address potential implications of use for health, labor, or criminal outcomes. These are all important topics for future research.

Finally, the nature of our model is structural, which allows us to generate predictions for a policy which has not yet been implemented in many places considering legalization. As such this represents a departure from current policies and, as was alluded to above, consumer perceptions may evolve as the market opens up. To the extent that consumers' preferences change to a great extent following this regulatory change the credibility of the structural approach could be strained as our counterfactual scenarios assume preferences are the same post legalization. This is another area where future research would be valuable.

References

Adams, B. and B. Martin. 1996. "Cannabis: Pharmacology and Toxicology in Animals and Humans." *Addiction,* 91, 1585-1614.

Adda J., B. McConnell and I. Rasul. 2011. "Crime and the Depenalization of Cannabis Possession: Evidence from a Policing Experiment," http://www.ucl.ac.uk/~uctpimr/research/depenalization.pdf.

Albert J. and S. Chib. 1993. "Bayesian Analysis via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts." *Journal of Business and Economic Statistics*, 11, 1-15.

Australian Crime Commission Illicit Drug Data Report. 2000. Report Number 2000-01, Australian Crime Commission (ACC) URL: http://www.crimecommission.gov.au/publications/illicit-drug-data-report.

Australian Crime Commission Illicit Drug Data Report. 2003. Report Number 2003-04, Australian Crime Commission (ACC) URL: http://www.crimecommission.gov.au/publications/illicit-drug-data-report.

Australian Crime Commission Illicit Drug Data Report. 2006. Report Number 2006-07, Australian Crime Commission (ACC) URL: http://www.crimecommission.gov.au/publications/illicit-drug-data-report.

Bretteville-Jensen, A. and L. Jacobi. 2011. "Climbing the Drug Staircase: A Bayesian Analysis of the Initiation of Hard Drug Use." *Journal of Applied Econometrics*, 26 (7), 1157-1186.

Caulkins, J., B. Kilmer, R. MacCoun, R. Pacula R., and P. Reuter. 2011. "Design Considerations for Legalizing Cannabis: Lessons Inspired by Analysis of California's Proposition 19," *Addiction.*

Chaloupka, F., K. Cummings, C. Morley, and J. Horan. 2002. "Tax, Price and Cigarette Smoking: Evidence from the Tobacco Documents and Implications for Tobacco Company Marketing Strategies." *Tobacco Control,* 11(Suppl I), i62–i72.

Chaloupka, F., and K. Warner. 2000. "The Economics of Smoking." In The Handbook of Health Economics Volume 1, Part 2 (eds. Newhouse J.P. and Cuyler A.J.), Amsterdam: Elsevier, 1539-627.

Chib S. and E. Greenberg. 1998. "Analysis of Multivariate Probit Models." *Biometrika*, 85, 347-361.

Chib S. and L. Jacobi. 2008. "Analysis of Treatment Response Data from Eligibility Designs. *Journal of Econometrics*, 144, 465-478.

Ching, A., Erdem, T. and M. Keane. 2009. "The Price Consideration Model of Brand Choice," *Journal of Applied Econometrics*, 24(3), 393-420.

Ching, A. and F. Hayashi. 2010. "Payment Card Rewards Programs and Consumer Payment Choice," *Journal of Banking & Finance,* 34(8),1773-1787.

Clements, K. 2006. "Pricing and Packaging: The Case of Marijuana." *Journal of Business,* 79(4), 2019-44.

Clements, K.and X. Zhao. 2009. Economics and Marijuana: Consumption, Pricing and Legalisation Volume 422, Cambridge and New York: Cambridge University Press.

Clerides, S. and P. Courty. 2010. "Sales, Quantity Surcharge, and Consumer Inattention," University of Cyprus, Department of Economnics Discussion Paper 2010-07.

Damrongplasit, K., C. Hsaio, and X. Zhao. 2010. "Decrimilization and Marijuana Smoking Prevalence: Evidence from Australia." *Journal of Business and Economic Statistics*, 28(3), 344-356.

DeSimone, J. 1998. "Is Marijuana a Gateway Drug?" *Eastern Economic Journal*, 24(2), 149-64.

Donohue, J., B. Ewing, D. Peloquin. 2011. "Rethinking America's Illegal Drug Policy," NBER Working Papers: 16776.

Eizenberg, A. 2011. "Upstream Innovation and Product Variety in the U.S. Home PC Market." SSRN Working Paper: 1760828.

Eliaz, K. and R. Spiegler. 2011. "Consideration Sets and Competitive Marketing." *Review of Economic Studies*, 78(1), 235-262.

Framework Convention on Tobacco Control. 2012. "Encouraging Consumption, Funding Organized Crime." http://global.tobaccofreekids.org/en/solutions/international_issues/illicit_trade_smuggling.

Galenianos, M., Pacula R., and N. Persico. 2009. "A Search-Theoretic Model of the Retail Market for Illicit Drugs," NBER Working Paper 14980.

Heckman, J. 1979. "Sample Selection Bias as a Specification Error," *Econometrica,* 47(1), 153–61.

Jiang R., P. Manchandab, and P. Rossi. 2009. "Bayesian Analysis of Random Coefficient Logit Models using Aggregate Data." *Journal of Econometrics,* 149, 136-148.

Kim, J., P. Albuquerque, and B. Bronnenberg. 2010. "Online Demand Under Limited Consumer Search," *Marketing Science,* 29(6), 1001-1023.

Miron, J. 2005. "The Budgetary Implications of Marijuana Prohibition." http://www.prohibitioncosts.org/mironreport/.

Miron, J. and J. Zwiebel. 1995. "The Economic Case Against Drug Prohibition." *Journal of Economic Perspectives,* 9, 175-92.

Office of National Drug Control Policy (ONDCP). 2004. The Economic Costs of Drug Abuse in the United States, 1992-2002. Publication 207303. Washington, D.C.: Executive Office of the President.

Pacula, R, M. Grossman, F. Chaloupka, J. O'Malley, L. Johnston, and M. Farrelly. 2000. "Marijuana and Youth," NBER Working Papers No. 7703.

Pacula, R., B. Kilmer, M. Grossman and J. Chaloupka. 2010. "Risks and Prices: The Role of User Sanctions in Marijuana Markets. " *B.E. Journal of Economic Analysis and Policy: Contributions to Economic Analysis and Policy* 10.

Poulsen, H. and G. Sutherland. 2000. "The Potency of Cannabis in New Zealand from 1976 to 1996." *Science and Justice*, 40, 171-176.

Pudney, S. 2010. "Drugs Policy: What Should we do About Cannabis?" *Economic Policy,* 25, 165-211.

Sickles R., and P. Taubman. 1991. "Who Uses Illegal Drugs?" *American Economic Review, Papers and Proceedings* 81(2), 248-251.

Sovinsky Goeree, M. 2008. "Limited Information and Advertising in the US Personal Computer Industry." *Econometrica,* 76(5), 1017-1074.

United Nations Office of Drugs and Crimes. 2012. "World Drug Report, 2012," http://www.unodc.org/unodc/en/data-and-analysis/WDR-2012.html.

Van Ours, J. 2003. "Is Cannabis a Stepping-Stone for Cocaine?" *Journal of Health Economics,* 22, 539-54.

Veugelers, R. and B. Cassiman. 2005. "R&D Cooperation Between Firms and Universities. Some Empirical Evidence from Belgian Manufacturing. *International Journal of Industrial Organization* 23(5–6), 355–379.

Williams, J. 2004. "The Effects of Price and Policy on Marijuana Use: What Can Be Learned from the Australian Experience?" *Health Economics*, 13, 123–137.

Williams, J., J. van Ours, and M. Grossman. 2011. "Why Do Some People Want to Legalize Cannabis Use?" NBER Working Paper No. 16795

For Online Publication:

A Estimation

A.1 Model Fitting for Probit Model with Selection

For the estimation of the Probit model for marijuana use with selection based on binary access via MCMC methods we introduce the latent continuous access and marijuana use variables $\{a^*_{im}\}$ and $\{u^*_{im}\}$ and use the common latent variable representation of the probit

$$a^*_{im} = \mu^a_{im} + \eta_{im} = \tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma} + \eta_{im}, \quad a_{im} = I[a^*_{im} > 0]$$

$$u^*_{im} = \mu^u_{im} + \varepsilon_{im} = \tilde{\mathbf{x}}'_{im}\boldsymbol{\beta} + \varepsilon_{im}, \quad u_{im} = I[u^*_{im} > 0] \qquad if \quad a_{im} = 1$$

where for each sample subject $\tilde{\mathbf{h}}_{im}$ refers to the combined covariate vector for the access model containing intercept, individual attributes, state fixed effects, market-specific variables influencing access, and $\tilde{\mathbf{x}}_{im}$ is the combined covariate vector for the net utility model that contains the price p_{im}, individual attributes, market specific variables, year fixed effects, and state fixed effects in addition to the intercept. We define the vector of model parameters as $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \rho)$. Under the assumption that $(\eta_{im}, \varepsilon_{im}) \sim N_2(0, \Xi)$, where Ξ is $2x2$ covariance matrix with 1 on the diagonal and ρ on the off-diagonal and following to the definition of the likelihood contribution given in equation 11, the likelihood of the model for all subjects in market m augmented with the latent access and net-use variables, $f(\mathbf{a}, \mathbf{u}, \{a^*_{im}\}, \{u^*_{im}\} | \boldsymbol{\theta}, \mathbf{W}, \{p_{im}\})$ can be expressed as

$$\prod_{i:a_{im}=0} \mathcal{N}(a^*_{im} | \tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma}, \mathbf{1}) \; I[a^*_{im} \leq 0]^{1-a_{im}}$$

$$\prod_{i:a_{im}=1} \mathcal{N}(u^*_{im} | \tilde{\mathbf{x}}'_{im}\boldsymbol{\beta} + \rho(a^*_{im} - \tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma}, 1-\rho^2) \times \{ I[u^*_{im} \leq 0]^{1-u_{im}} + I[u^*_{im} > 0]^{u_{im}} \}$$

$$\times \mathcal{N}(a^*_{im} | \tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma}, \mathbf{1}) I[a^*_{im} > 0]^{a_{im}}$$

where the inclusion of the latent data improves the tractability of the likelihood (Albert and Chib, 1993). The joint distribution of access and use for access subjects is now expressed in terms of the marginal-conditional decomposition. The indicator functions ensure that we choose the correct bivariate distribution with the distribution of latent use truncated according to the observed use.

For the Bayesian analysis we proceed with the common assumption of normal independent priors for the slope coefficients and correlation coefficient. The latter is restricted to the region $R = -1 < \rho < 1$ to ensure the positive definiteness of $\mathbf{\Xi}$. The joint prior is given by

$$\pi(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\beta}|\mathbf{b}_0, \mathbf{B}_0)\ \mathcal{N}(\boldsymbol{\gamma}|\mathbf{g}_0, \mathbf{G}_0)\ \mathcal{N}(\rho|r_0, R_0) \times R$$

The prior means are set at zero. In combination with large prior variances this implies relatively uninformative prior assumptions. It should be noted that in the context of our very large data set the influence of the prior is very small as the information from the data via the likelihood will dominate the inference about the model parameters summarized in the posterior distribution. The posterior distribution, with the parameter space augmented by the latent access and marijuana variables, $\pi(\boldsymbol{\theta}, \mathbf{a}^*, \mathbf{u}^*|\mathbf{a}, \mathbf{u})$, is proportional to the product of the likelihood and the prior. We employ a straight forward Metropolis within Gibbs simulation algorithm with five blocks to generate draws from the posterior distribution of the parameter vector, as well as the marginal distributions of each parameter. By augmenting the parameter space with the latent access and net-use variables, the priors on the regression coefficients are conditionally conjugate, thus allowing for normal updates of slope the coefficients. The latent variables are also normal updates. A Metropolis Hastings update is used for the correlation parameter as the structure of the covariance matrix and the likelihood do not allow a Gibbs update. The detailed steps of the algorithm are as follows:

First, we draw a^*_{im} from $\mathcal{N}(a^*_{im}|\tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma}, 1)\ I[a^*_{im} \leq 0]$ for $i \in I_0$ and from $\mathcal{N}(a^*_{im}|\tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma} + \rho(u^*_{im} - \widetilde{\mu}^u_{im}), 1 - \rho^2)\ I[a^*_{im} > 0]$ for those subjects with $i \in I$, where $i \in I_0$ refers to the subset of subjects with no access and $i \in I_1$ to those with access.

In the second step, we draw u^*_{im} for all subjects $i \in I_1$ from either $\mathcal{N}(u^*_{im}|\tilde{\mathbf{x}}'_{im}\boldsymbol{\beta} + \rho(a^*_{im} - \tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma}), 1-\rho^2)\ I[u^*_{im} \leq 0]$ if $u_{im} = 0$ or from $\mathcal{N}(u^*_{im}|\tilde{\mathbf{x}}'_{im}\boldsymbol{\beta} + \rho(a^*_{im} - \tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma}), 1-\rho^2)\ I[u^*_{im} > 0]$ if $u_{im} = 1$.

In the third step, we draw $\boldsymbol{\gamma}$ from $\mathcal{N}(\hat{\boldsymbol{\gamma}}, \hat{\mathbf{G}})$ with

$$\hat{\boldsymbol{\gamma}} = \hat{\mathbf{G}}[\mathbf{G}_0^{-1}\mathbf{g}_0 + \sum_{i \in I_0} \tilde{\mathbf{h}}_{im} a^*_{im} + \sum_{i \in I_1} \tilde{\mathbf{h}}_{im}(1-\rho^2)^{-1}(a^*_{im} - \rho(u^*_{im} - \tilde{\mathbf{x}}'_{im}\boldsymbol{\beta})]$$

$$\hat{\mathbf{G}} = [\mathbf{G}_0^{-1} + \sum_{i \in I_0} \tilde{\mathbf{h}}_{im}\tilde{\mathbf{h}}'_{im} + \sum_{i \in I_1} \tilde{\mathbf{h}}_{im}(1-\rho^2)^{-1}\tilde{\mathbf{h}}'_{im}]^{-1}.$$

In the fourth step we draw $\boldsymbol{\beta}$ based on the subjects in I_1 from $\mathcal{N}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{B}})$ where

$$\hat{\boldsymbol{\beta}} = \hat{\mathbf{B}}[\mathbf{B}_0^{-1}\mathbf{b}_0 + \sum_{i \in I_1} \tilde{\mathbf{x}}_{im}(1-\rho^2)^{-1}(u^*_{im} - \rho(a^*_{im} - \tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma}))]$$

$$\hat{\mathbf{B}} = [\mathbf{B}_0^{-1} + \sum_{i \in I_1} \tilde{\mathbf{x}}_{im}(1-\rho^2)^{-1}\tilde{\mathbf{x}}'_{im}]^{-1}.$$

In the last step we update ρ in Metropolis Hastings step based on the subjects in I_1, since the conditional posterior distribution of ρ is not tractable. Following Chib and Greenberg (1998) we generate proposal value ρ' from a tailored student-t density $t_\nu(\mu, V)$ where μ is the mode of

$$ln(\prod_{I \in I_1} \mathcal{N}(a^*_{im}, u^*_{im} | W_{im}\boldsymbol{\delta}, \Xi) \text{ , where } \mathbf{W}_{im} = \begin{pmatrix} \tilde{\mathbf{h}}'_{im} \\ \\ \tilde{\mathbf{x}}'_{im} \end{pmatrix}, \ \boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\gamma} \\ \\ \boldsymbol{\beta} \end{pmatrix} \text{ and } \Xi = \begin{pmatrix} 1 & \rho \\ \\ \rho & 1 \end{pmatrix}$$

and V is the inverse of the Hessian of the density evaluated at μ. The proposed value ρ' is accepted with probability

$$\alpha = \min\left(1 \ , \ \frac{\pi(\rho') \ \prod_{I \in I_1} \mathcal{N}(a^*_{im}, u^*_{im} | \mathbf{W}_{im}\boldsymbol{\delta}, \Xi') \ t_\nu(\rho|\mu, V)}{\pi(\rho) \ \prod_{I \in I_1} \mathcal{N}(a^*_{im}, u^*_{im} | \mathbf{W}_{im}\boldsymbol{\delta}, \Xi) \ t_\nu(\rho'|\mu, V)}\right).$$

We repeat the above steps for M iterations after an initial burn-in phase of M_0 iterations to allow for the convergence of the chain. We obtain a vector of M draws for each model parameter that reflects the (marginal) posterior distribution of each parameter. (Under the Bayesian approach the model parameters are random variables and all information about the parameters from the estimation is summarized in their respective posterior distributions.) In the main text we provide summaries of the posterior distributions in terms of the posterior means (coefficient estimate) and standard deviations or the 90% and 95% credibility intervals.

A.1.1 Simulated Prices

We address the issue of the unobserved individual prices as described in section 3.1 by adding an additional step to the above described algorithm to draw the individual price for each subject i in market m with access from the market specific empirical price distribution (7). At the beginning of each iteration g of the algorithm we generate an individual price $p^{(g)}_{im}$ from

$$\sum_{t=1}^{3} \pi^g_{imt} \times p^g_{imt} \ , \quad \sum_{t=1}^{3} \pi^g_{imt} = 1$$

where t takes the values $t = 1, 2, 3$. The probability of using type t (π^g_{imt}) and the corresponding price for type t (p^g_{imt}) are drawn from the constructed empirical distributions (see equation 6) based on the observed data. Here we use normal distributions truncated at zero below and centered at the observed values of the prices and usage for each type in each market, $TN_{(0,\infty)}(\overline{p}_{mt}, \Omega^p_{mt})$ and $TN_{(0,\infty)}(\overline{\boldsymbol{\pi}}_{mt}, \Omega^{\pi}_{mt})$, respectively. The variances are based on observed variation in the data. By generating the price by type and usage of type from these distributions we can exploit the information on prices by type and market and usage of types within a market among users in the data, while at the same time allowing for some variation of prices and usage among subjects in a market. Note that different choices of distributions are possible and independence of the distributions across types is not necessary but chosen here based on the data restrictions. The graphs below show the resulting empirical price distributions for a subset of markets.

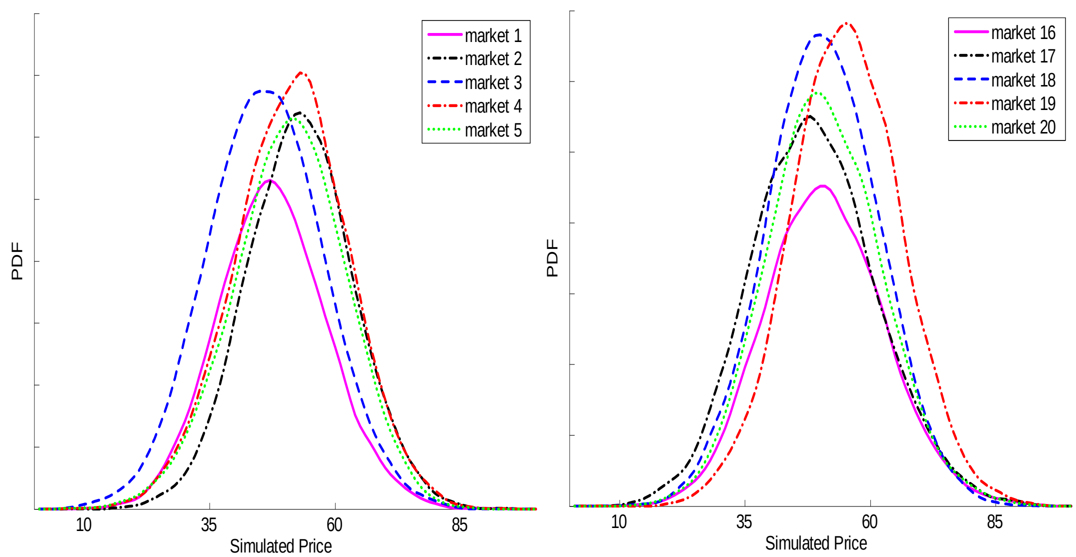

Figure A.1: Simulated Price Distributions

A.2 Marijuana Use Prediction Counterfactual

In Section 5 we report the probabilities of marijuana use for different counterfactual scenarios under various specifications of the selection model for the extensive margin of marijuana use. The reported probabilities are obtained using the standard Bayesian approach for prediction. The approach enables us to exploit all the information about the parameters summarized in the posterior distribution and to compute credibility intervals (Bayesian confidence intervals) for the probabilities of use and address the issue of unobserved individual prices via an empirical distribution generated as described above.

Let $n = 1$ refer to a random subject in market m from the sample, with demographic characteristics and market features in the use model and access model given by vectors $\tilde{\mathbf{x}}_{n+1,m}$, including price $p_{n+1,m}$, and $\tilde{\mathbf{h}}_{n+1,m}$, respectively. Under the selection model on marijuana use we can obtain the probability of marijuana use for the subject from the expression

$$\begin{aligned}\Pr(u_{n+1,m} &= 1|\mathbf{a},\mathbf{u}) = \int \Pr(u_{n+1,m} = 1|a_{n+1,m}, \tilde{\mathbf{x}}_{n+1,m}, \tilde{\mathbf{h}}_{n+1,m}, p_{im}) \;\; d\widehat{P}_m(p_{n+1,m}) \; dF_\pi(\boldsymbol{\theta}\;) \\ & dF_{data}(\tilde{\mathbf{x}}_{n+1,m}, \tilde{\mathbf{h}}_{n+1,m})\end{aligned}$$

where $\Pr(u_{n+1,m} = 1|a_{n+1,m}, \tilde{\mathbf{x}}_{n+1,m}, \tilde{\mathbf{h}}_{n+1,m}, p_{im}) = \Phi(m_{n+1,m}|\boldsymbol{\theta},\tilde{\mathbf{x}}_{n+1,m}, \tilde{\mathbf{h}}_{n+1,m}, p_{im})$ is the conditional probability of use (assuming access) with the conditional mean given by $m_{n+1,m} = \tilde{\mathbf{x}}'_{n+1,m}\;\boldsymbol{\beta}\;+\rho\eta_{n+1,m}$. The term $\rho\eta_{n+1,m}$ accounts for selection on unobservables, where the value of the unobservable term $\eta_{n+1,m}$ is found using information on the distribution of unobservables in the data by exploiting the data on the observed access of a subject within our model. Specifically, if $a_{n+1,m} = 0$, then it follows directly from the model of access that $\mu^a_{n+1,m} + \eta_{n+1,m} \leq 0$, or $\eta_{n+1,m}\; \leq -\mu^a_{n+1,m}$ where $\mu^a_{n+1,m} = \tilde{\mathbf{h}}'_{n+1,m}\boldsymbol{\gamma}$. As $\eta_{n+1,m}$ follows a standard normal distribution we generate $\eta_{n+1,m}$ from $TN_{(-\infty,-\mu^a_{n+1,m})}(0,1)$ for the subject. Similarly, for the case of $a_{n+1,m} = 1$ we have $\mu^a_{n+1,m} + \eta_{n+1,m} > 0$, so that we generate $\eta_{n+1,m}$ from $TN_{(-\mu^a_{n+1,m},\infty)}(0,1)$. Note that if we set $\rho\eta_{n+1,m} = 0$ we would implement the prediction based on the marginal model for marijuana use. While predictions are often based on the marginal model, this approach would lead us to considerably underpredict the benchmark case of use under pre-legalization relative to the observed use pre legalization, due to ignoring the important role that selection on unobservables plays in the context of marijuana use and more generally in the use of illicit drugs.

As indicated by the above integral expression, from the conditional probability we integrate out the prices based on the empirical price distribution $\widehat{P}_m(p_{im})$, the model parameters using the posterior distribution $F_\pi = \pi(\boldsymbol{\theta}\;|\mathbf{a},\mathbf{u})$ and the individual and market characteristics based on the empirical distribution of the sample. The above integral expression can be estimated in a straight forward manner using the draws from the posterior distribution from the MCMC algorithm discussed in the previous section. Essentially, at each iteration of the MCMC algorithm after the burn-in phase, we add an additional step where vectors $\tilde{\mathbf{x}}_{n+1,m}$

Scenario	Access	Legality	Price
1	no	no	pre-legality
2	yes	no	pre-legality
3	yes	yes	pre-legality
4	yes	yes	25% increase
5	yes	yes	cigarette price
6	yes	yes	price at marginal cost

and $\tilde{\mathbf{h}}_{n+1,m}$ are drawn from the data and $\Phi(m_{n+1,m})$ is computed using the current MCMC draws on the model parameters and prices. The resulting vector of probabilities describes the predictive distribution of the probability of use. In tables 9 and 10 we report the mean probabilities and standard deviations of the predictive distributions.

We implemented the predictions under the following counterfactual scenarios s with the first scenario being the status quo:

For the prediction under these different scenarios let $\widetilde{v}_{n+1,m}$ denote the vector of market and demographic characteristics without the price and the disutility variables, $p_{n+1,m}$ and $L^{illegal}_{n+1,m}$. Let $\boldsymbol{\beta}_v, \beta_p$ and β_l denote the corresponding parameter (vectors) . We can then write the conditional probability of use under scenario s for our baseline model specification without interactions as

$$\begin{aligned}\Pr(u_{n+1,m} &= 1|a_{n+1,m}, \tilde{\mathbf{x}}_{n+1,m}, \tilde{\mathbf{h}}_{n+1,m}, p_{n+1,m}) \\ &= \Phi^{\ s}(\widetilde{v}'_{n+1,m}\ \boldsymbol{\beta}_v + p^s_{n+1,m}\beta_p + L^{illegal,s}_{n+1,m}\beta_l + \rho\eta_{n+1,m})\end{aligned}$$

where $p^s_{n+1,m}$ is drawn from an empirical distribution using the approach described above, with the mean of the price distributions for each type adjusted according to the assumed price change for scenarios 4 to 6. The disutility variable $L^{illegal,s}_{n+1,m}$ is set to zero for all subjects in scenarios that assume legality ($s > 2$), and otherwise remains as unchanged. Note that by predicting use for any random subject in the sample we assume that marijuana is accessible for all subjects. To account for limited access in our benchmark scenario $s = 1$ we follow the described approach, but set $\Phi(m_{n+1,m}) = 0$ if $a_{n+1,m} = 0$. For the predicted probabilities of use by various demographic groups presented in table 10, we draw $\widetilde{v}_{n+1,m}$ and $\tilde{\mathbf{h}}_{n+1,m}$ from the corresponding subsamples of subjects with no access and subjects with access. For those with no access we set $\Phi(m_{n+1,m}) = 0$ under scenario 1. Finally, the counterfactual use results in Section 5 are generated using our baseline model specification as well as our model specification with price and legality interaction terms. For the latter the expression in the mean probability of use $(p^s_{n+1,m}\beta_p + L^{illegal,s}_{n+1,m}\beta_l)$ is replaced with the corresponding interaction terms with age brackets.

A.3 Model Fitting and Counterfactuals for Ordered Probit Model with Selection

We also estimate an ordered probit model with selection (intensive use margin model) for the discrete ordered marijuana frequency of use variable, $y_{im} = 0, 1, 2$, for the analysis of tax revenues in section 5.3:

$$y^*_{im} = \tilde{\mathbf{x}}'_{im}\boldsymbol{\beta} + \nu_{im}, \text{ where } y_{im} = 0 \text{ if } y^*_{im} \leq 0,\ y_{im} = 1 \text{ if } 0 < y^*_{im} \leq \tau \text{ and } y_{im} = 2 \text{ if } \tau < y^*_{im}$$

where τ refers to the cut-off point that has to be estimated. The first cut-off point has been set to zero for identification purposes. The model for access remains unchanged and as before we assume that access and marijuana use may both be affected by unobserved factors so that $(\eta_{im}, \nu_{im}) \sim N_2(0, \Xi)$ where Ξ is $2x2$ covariance matrix with 1 on the diagonal and ρ on the off-diagonal. The likelihood of the model, $f(\mathbf{a}, \mathbf{y}, \{a^*_{im}\}, \{y^*_{im}\}|\boldsymbol{\xi}, \mathbf{W}, \{p_{im}\})$ where $\boldsymbol{\xi} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \rho, \tau)$, can be again expressed in terms of the latent data to improve the tractability of the likelihood (Albert and Chib, 1993) as

$$\begin{aligned}
\prod_{i:a_{im}=0} \mathcal{N}(a^*_{im}|\tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma}, \mathbf{1})\ I[a^*_{im} &\leq 0]^{a_{im}} \\
\prod_{i:a_{im}=1} \mathcal{N}(a^*_{im}|\tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma}, \mathbf{1}) I[a^*_{im} &> 0]^{1-a_{im}}\ \{\mathcal{N}(y^*_{im}|\tilde{\mathbf{x}}'_{im}\boldsymbol{\beta} + \rho(a^*_{im} - \tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma}\ , 1-\rho^2) \\
\times(I[y_{im} &= 0]\ I[y^*_{im} \leq 0] + I[y_{im} = 1]\ I[0 < y^*_{im} \leq \tau] + \ I[y_{im} = 2]\ I[y^*_{im} > \tau])\}
\end{aligned}$$

We again assume independent normal priors for $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \rho)$ as in the probit model with selection (extensive use model). For the cut-off points it is sufficient to assume a priori that $\tau > 0$.

To simulate the posterior distribution $\pi(\boldsymbol{\xi}, \mathbf{a}^*, \mathbf{y}^*|\mathbf{a}, \mathbf{y})$ we employ a 6 step MCMC algorithm that is an extended and modified version of the 5 step algorithm for the Bivariate Probit with Selection discussed above. We add a 6th step to draw the cut-off point and also adjust the generation of the latent utility $\mathbf{y}^*$. For the latter, we draw y^*_{im} for all subjects $i \in I_1$ from the truncated normal distributions $\mathcal{TN}_{(a,b)}(y^*_{im}|\tilde{\mathbf{x}}'_{im}\boldsymbol{\beta} + \rho(a^*_{im} - \tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma}), 1-\rho^2)\ I[a < y^*_{im} \leq b]$, where $(a = -\infty, b = 0)$ for $k = 0$, $(a = 0, b = \tau)$ for $k = 1$ and $(a = \tau, b = +\infty)$ for $k = 2$. To update the cut-off point we employ a Metropolis Hastings algorithm as the conditional posterior distribution is of an unknown form. To improve the performance we update the cut-off point marginalized over the latent utilities $\{y^*_{im}\}$ and generate the proposal values from the tailored student-t density $q(\tau) = t_{10}(\mu, V)$, where here μ is the mode of the likelihood of the access subjects with with y_{im}=1 and y_{im} =2, $f(\mathbf{a} = \mathbf{1}, \{a^*_{im}\}, \{y_{im}{=}1\}, \{y_{im}{=}2\}|\boldsymbol{\gamma}, \boldsymbol{\beta}, \rho, \mathbf{W})$ and V is the inverse of the Hessian of the density evaluated at μ. We maximize the proportional conditional likelihood expression (omitted $\mathcal{N}(a^*_{im}|\tilde{\mathbf{h}}'_{im}\boldsymbol{\gamma}, \mathbf{1})I[a^*_{im} > 0]^{1-a_{im}}$ as it does not depend on cut-off points)

$$ln\left(\prod_{I_1:\ y_{im}=1}\left[\Phi(\frac{\tau - m_{im}}{\sigma}) - \Phi(\frac{-m_{im}}{\sigma})\right]\right) + ln\left(\prod_{I_1:\ y_{im}=2}\left[1 - \Phi(\frac{\tau - m_{im}}{\sigma})\right]\right)$$

where $m_{im} = \tilde{\mathbf{x}}'_{im}\boldsymbol{\beta} + \rho(a^*_{im} - \tilde{\mathbf{h}}'_{im}\gamma)$ and $\sigma = \sqrt{1-\rho^2}$. The maximization is subject to the constraint that $\tau > 0$.

The proposed value τ , with $\tau > 0$, is accepted with probability

$$\alpha = \min\left(1\ ,\ \frac{f(\{y_{im}{=}1\}, \{y_{im}{=}2\}|\{a^*_{im}\}, \mathbf{a} = \mathbf{1}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \rho, \tau', \mathbf{W})\ t_\nu(\tau|\mu, V)}{f(\{y_{im}{=}1\}, \{y_{im}{=}2\}|\{a^*_{im}\}, \mathbf{a} = \mathbf{1}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \rho, \tau, \mathbf{W})\ t_\nu(\tau'|\mu, V)}\right),$$

where again we use the conditional form of the likelihood of marijuana use, omitting the marginal likelihood of access as it does not depend on the cut-off point. As in the algorithm for the probit model we draw the prices for the access subjects from the corresponding empirical distribution at the beginning of each iteration of the algorithm. The estimates are presented in table 8.

For the estimation of the tax revenues in Section 5.3 we again employ the Bayesian predictive approach described in Appendix A.2. Instead of predicting the probability of use we predict the probability of use in each category k, $\widehat{G}_{ikr}$, for each subject i (in market m) in the sample under two different tax regimes $r = 1, 2$ from

$$\widehat{G}_{ikr} = \int \Pr(y_{im} = k|a_{i,m}, \tilde{\mathbf{x}}_{im}, \tilde{\mathbf{h}}_{im}, p^r_{im})\ \ d\widehat{P}_m(p^r_{im})\ dF_\pi(\boldsymbol{\xi}\ )\ \ dF_{data}(\tilde{\mathbf{x}}_{im}, \tilde{\mathbf{h}}_{im})$$

where as before we integrate over the price distribution, now also depending on the tax regime, the posterior distribution of posterior distribution of the relevant model parameters and the empirical distribution of the data (covariates). As in the extensive use model the prediction is based on the conditional probability of use, now for each category. For example, for $k = 0$ we have $\Pr(y_{im} = 0|\cdot) = \Phi(-\mu_{im})$ with $\mu_{im} = \tilde{\mathbf{x}}'_{im}\boldsymbol{\beta} + \rho\eta_{im}$ ($\tilde{\mathbf{x}}'_{im}\boldsymbol{\beta}$ includes age interactions). Under tax regime 1 with 25% tax on current prices, following equation 7 the price the individual faces under the tax regime is generated from

$$p^1_{im} \sim \sum_{t=1}^{3} \int (\pi_{imt} * p^1_{imt})\, dF_\pi(\pi_{imt})\, dF_p(p^1_{mt})$$

where we adjust equation 6 and now have $p^1_{imt} \sim TN_{(0,\infty)}(1.25 * p_{imt}, \Omega^p_{mt})$ with the mean set at the current average market price of each type plus a 25% tax. Under tax regime 2 the price is set at the marginal cost of production of each type, MC_t, so that the price distribution simplifies to

$$p^2_{im} \sim \sum_{t=1}^{3} \int (\pi_{imt} * MC_t)\, dF_\pi(\pi_{imt})$$

As before the prices reflect a weighted average over the prices of three different types based on the usage probabilities of each type and its distribution $F_\pi(\pi_{imt})$.

B Price Counterfactual Calculation

For the counterfactual price calculation under the ordered probit (intensive margin) with selection we find the counterfactual price that implies a predicted "post-legal" probability of no use among teenagers under legalization that is comparable to the probability of no use observed in the data before legalization. Given the model we implement the analysis at the market level, finding the counterfactual prices for all teenagers in market m to match the observed probability of no use in their market, call this S^{obs}_m. Let $S^{Post}_{im}(\widehat{\Theta}, data; p_{im})$ represent the probability of no use for teenager i in market m under a counterfactual of legalization, and p^{CF}_{im} the counterfactual price, where p^{CF}_{im} is chosen so that

$$\frac{\sum_{n_i=1}^{n_m} S^{Post}_{im}(\widehat{\Theta}, data; p_{im} = p^{CF}_{im})}{n_m} = S^{obs}_m$$

where n_m refers to the number of teenagers in market m and the estimated parameter set $\widehat{\Theta}$ (here means of the posterior distribution of the parameters). To find the counterfactual prices for each teenager in market m, we first predict post legalization probability of no use, as described in Appendix A.2, with price p_{im} coming from the corresponding (pre-legal) empirical distribution (7). We then find the counterfactual price, where p^{CF}_{im} is the price that equates

$$S_{im}^{Post}(\widehat{\Theta}, data; p_{im}^{CF}) = S_m^{obs}.$$

From our ordered probit model on marijuana use with selection it follows that the probability of no use is $S_{im}^{Post}(\widehat{\Theta}, data; p_{im}^{CF}) = \Phi(-(\mu_{im} = f(\widehat{\Theta}, data; p_{im}^{CF}))$, where μ_{im} is the conditional mean of marijuana use taking into account preferences and the selection of unobservables for teenager i so that under our model specification with price and age interactions

$$\Phi(-(\mu_{im} = p_{im}^{CF}\widehat{\beta}_{p,teen} + L_{im}^{illegal}\widehat{\beta}_l + \widetilde{v}_{im}' \widehat{\boldsymbol{\beta}}_v + \widehat{\rho}\eta_{im})) = S_m^{obs},$$

where $L_{im}^{illegal}$ denotes the disutility from illegality variable, $\widetilde{v}_{n+1,m}$ the vector of demographic characteristics and market characteristics without the price. $\widehat{\beta}_{p,teen}$, $\widehat{\beta}_l$ and $\widehat{\boldsymbol{\beta}}_v$ and refer to the estimated coefficients (posterior means) on price for teenagers, the effect of disutility and the effects of the elements in $\widetilde{v}_{n+1,m}$, respectively. As before $\rho\eta_{im}$ accounts for selection on unobservables with $\eta_{im} \sim TN_{(-\infty,-\widehat{\mu}_{im}^a,)}$ if $a_{im} = 0$ and $\eta_{im} \sim TN_{(-\widehat{\mu}_{im}^a,\infty)}$ if $a_{im} = 1$ using the observed access before legalization (see also Appendix A.2). The counterfactual price is obtained from

$$p_{im}^{CF} = \max\left\{\frac{-\Phi^{-1}(S_m^{obs}) - \widetilde{v}_{im}' \widehat{\boldsymbol{\beta}}_v - L_{im}^{illegal}\widehat{\beta}_l - \widehat{\rho}\eta_{im}}{\widehat{\beta}_{p,teen}}, 0\right\},$$

where Φ^{-1} is the inverse of the normal CDF and the maximum condition ensures non-negative prices. The latter is needed as some teenagers have $S_{im}^{Post}(p_{im}) >= S_m^{obs}$ and for some teenagers their probability of no use is far above the market average at current prices and we would obtain a negative price to lower it to the market average level. (An alternative approach is to set $p_{im}^{CF} = p_{im}$ for teenager with $S_{im}^{Post}(p_{im}) >= S_m^{obs}$ and only compute the counterfactual price as described above for teenagers who's probability of no use is below the market average. This approach yields similar results.) In the main paper we provide summaries of the estimated counterfactual prices for teenagers needed to keep the proportion of non-users at pre-legal levels and also by gender. Since use remains illegal for teenagers the disutility variable $L_{im}^{illegal}$ remains unchanged.