WORKING PAPERS

The Role of Education in the Production of Health:
An Empirical Analysis of Smoking Behavior

Steven Tenn
Douglas A. Herman
Brett Wendling

WORKING PAPER NO. 292

June 2008

FTC Bureau of Economics working papers are preliminary materials circulated to stimulate discussion and critical comment. The analyses and conclusions set forth are those of the authors and do not necessarily reflect the views of other members of the Bureau of Economics, other Commission staff, or the Commission itself. Upon request, single copies of the paper will be provided. References in publications to FTC Bureau of Economics working papers by FTC economists (other than acknowledgment by a writer that he has access to such unpublished materials) should be cleared with the author to protect the tentative character of these papers.

BUREAU OF ECONOMICS
FEDERAL TRADE COMMISSION
WASHINGTON, DC 20580

THE ROLE OF EDUCATION IN THE PRODUCTION OF HEALTH: AN EMPIRICAL ANALYSIS OF SMOKING BEHAVIOR

Steven Tenn*
Douglas A. Herman
Brett Wendling

Federal Trade Commission

June 24, 2008

Abstract

We estimate the effect of education and student status on the propensity to smoke. Our estimation strategy accounts for the endogeneity of education by "differencing out" the impact of unobserved characteristics correlated with educational attainment. This is accomplished by exploiting education differences between similarly selected groups that are one year apart in their life cycle. The results indicate that an additional year of education does not have a causal effect on smoking. Unobserved factors correlated with educational attainment entirely explain their cross-sectional relationship. We do find, however, that being a student reduces the likelihood of smoking. This may be a peer effect, which prior research shows has a significant impact on smoking decisions.

Keywords: education, health, smoking
JEL classification codes: I12, I20

* Corresponding author. E-mail: stenn@ftc.gov. Mailing address: 600 Pennsylvania Ave NW, Washington, DC 20580. Telephone: (202) 326-3243. We thank Dan Hanner, Daniel Hosken, and David Schmidt for providing very helpful comments. The views expressed in this paper are those of the authors and do not necessarily represent the views of the Federal Trade Commission or any individual Commissioner.

THE ROLE OF EDUCATION IN THE PRODUCTION OF HEALTH: AN EMPIRICAL ANALYSIS OF SMOKING BEHAVIOR

I. Introduction

Education is correlated with a wide range of health measures (Grossman 2006). The better educated are less likely to smoke, abuse alcohol, be obese, or work in a hazardous profession. They also tend to produce healthier offspring, live longer, and are more likely to exercise. Despite the strong correlation between education and health, the causal mechanism underlying these relationships has not yet been determined. Several potential explanations have emerged from the literature. Education may teach individuals to convert health inputs into health outcomes more efficiently (Grossman 1972), or the better educated may employ a more efficient mix of health inputs (Kenkel 1991, Rosenzweig 1995, de Walque 2007a). A competing hypothesis is that education does not play a causal role in explaining health behaviors. Rather, unobserved characteristics that make individuals invest in education may also increase their investment in health. This can create a correlation between education and health even in the absence of any direct effect (Farrell and Fuchs 1982).

This paper adds to the growing health-education literature by exploring the impact of educational attainment on smoking behavior. We analyze smoking for two reasons. First, the relationship between smoking and health outcomes is well documented by medical science. Smoking is causally associated with cancer, cardiovascular diseases, respiratory diseases, and other serious medical conditions (U.S. Department of Health and Human Services 2004). In fact, smoking is the leading preventable cause of death in the United States (Mokdad et al. 2004). Chaloupka and Warner (2000) estimate that the costs associated with smoking exceed $100 billion annually in the United States alone. The second reason we focus on smoking is that there is a strong cross-sectional relationship between education and smoking, where the better educated are much less likely to smoke (U.S. Department of Health and Human Services 1989).

To the extent that this correlation is due to a causal effect of education, this relationship may be an important determinant of why the better educated are healthier.

The Surgeon General's reports on smoking and health highlight the need for prevention policies that target the less educated. The success of such programs depends, in part, on what factors cause smoking differences by education level. If education increases awareness of the negative health effects of smoking, information programs targeted at the less educated may reduce their smoking. An alternative possibility is that the less educated understand smoking has an adverse effect on health, but have unobserved attributes that affect their cost-benefit analysis, e.g., they may care less about the future. Information programs on the health consequences of smoking would be ineffective in this situation, although prevention programs that educate on the importance of the future might deter smoking (Becker and Mulligan 1997). Our analysis of whether education has a causal effect on smoking is instructive on the likely efficacy of these prevention programs.

Recent empirical research exploring the impact of education on smoking behavior employs "policy" instruments to identify causal effects. A variety of instrumental variables have been developed: the Vietnam draft (de Walque 2007b, Grimard and Parent 2007), high school graduation requirements (Kenkel et al. 2006), and college openings (Currie and Moretti 2003). While these studies all conclude that education significantly reduces the propensity to smoke, they arrive at this result using the same basic methodology. Specifically, they employ instruments that vary only by birth cohort and gender, or by birth cohort and geographic location. To avoid collinearity problems, interactions between these variables must either be excluded from the model or controlled for in a restrictive manner (see Section II).

This paper contributes to the literature by developing an alternative estimation strategy that identifies the impact of education under less restrictive conditions. This allows us to test whether similar findings can be obtained after relaxing the exclusion restrictions imposed in previous studies. We address the potential endogeneity of education by using a control group framework that matches individuals who differ by being one year apart in their life cycle. Those

who will acquire a given level of education in the following year are compared to individuals who are one year older and currently have that particular level of education. The key identification assumption is that these two groups have similar unobserved characteristics. This assumption is reasonable since they are born only one year apart, and make identical education decisions at the same point in their lives. The empirical methodology "differences out" the impact of unobserved characteristics correlated with education, thereby isolating its causal effect.

The effect of education on smoking behavior is estimated using data from the Tobacco Supplement of the Current Population Survey. Unlike recent studies which rely on instrumental variables, we find that an additional year of education has little impact on the propensity to smoke. Our estimates range from -0.7 to -0.2 percentage points, depending on how smoking is measured (ever, current, or everyday smoker).[1] Despite being precisely estimated, these effects are not statistically significant, nor are they large in magnitude. In contrast to education, our results indicate that being a student reduces the probability of smoking by 3.0 to 5.2 percentage points, depending on the smoking measure used. We hypothesize that this may be due to peer effects, which prior research shows have a significant impact on smoking decisions (Norton et al. 1998, Gaviria and Raphael 2001, Powell et al. 2005).

Our analysis has several distinguishing features. First, we report estimates of education's effect on smoking that both control for selection bias and are applicable to a significant fraction of the population. This overcomes a limitation of studies that use policy instruments, which generate estimates that apply only to those individuals whose education choice is affected by their instrument. Second, we estimate the effect of education for a more recent generation than has previously been considered. This allows us to assess whether education has an impact now that the harmful health effects of smoking are widely known. Lastly, the methodology employed allows us to identify the effect of education in an unrestrictive manner. In particular, we do not

[1] Results are taken from specification (i) of Table 3. Other specifications yield similar findings.

impose exclusion restrictions on interactions between age, generation, time, and geography that have been employed in prior research that uses instrumental variables.

The layout of the paper is as follows. Section II reviews the literature. Section III provides an example that illustrates how we identify the causal effect of education. The empirical methodology is detailed in Section IV. Section V describes the data. Results are presented in Section VI, followed by a discussion in Section VII of why our findings differ from prior research. Section VIII concludes.

II. Literature Review[2]

Three theories relating education to health have emerged from the literature. The theory of *productive efficiency* contemplates that the production function converting health inputs into health outputs depends on an individual's stock of human capital (Grossman 1972), a major component of which is education. Those with greater human capital are able to convert health inputs into positive health outcomes more efficiently. Alternatively, the theory of *allocative efficiency* is modeled on the premise that the better educated choose a more productive set of health inputs (Kenkel 1991, Rosenzweig 1995, de Walque 2007a). The acquisition of human capital may reveal information about the health production function that allows those with greater education to select a more efficient mix of inputs. For example, education may increase awareness of the negative health effects of smoking, facilitating a more informed cost-benefit analysis.

A third class of models explores the influence of time preference. Those with a lower discount rate are more likely to make long-run investments in education and health. An example of the latter is choosing not to smoke, since the negative health impacts from smoking do not typically materialize until many years later. Time preference is viewed as a "third factor" which

[2] An expansive literature considers the relationship between education and health. We focus on studies of education's effect on smoking. See Grossman (2006) for a review of the wider literature.

can create a positive correlation between education and health even in the absence of any causal relationship between the two variables (Farrell and Fuchs 1982). Causation is possible, however, if time preference is endogenously determined. Becker and Mulligan (1997) consider a model where education influences an individual's discount rate. If education increases one's future orientation, an individual is more likely to invest in health after acquiring additional education.

Recent contributions to the literature use instrumental variables to estimate the total effect of education across all three causal mechanisms described above. These studies conclude that education has a significant effect on a variety of health measures, including both inputs and outcomes. Examples include self-rated health (Adams 2002), mortality (Lleras-Muney 2005), and smoking (de Walque 2007b, Grimard and Parent 2007, Kenkel et al. 2006, Currie and Moretti 2003).

To be valid, an instrumental variable must be correlated with education but uncorrelated with unobserved characteristics that affect an individual's health. Early studies use background characteristics, such as parental income and education, as an instrument for education (Berger and Leigh 1989, Sander 1995, Leigh and Dhir 1997). This method has been criticized in the recent literature due to concerns about endogeneity. Parental education may be correlated with the child's education, but may also be correlated with unobserved determinants of the child's health. Researchers have tried to overcome this issue by exploiting policy changes that impact an individual's educational attainment. The benefit of relying on policy changes is that there is less reason to suspect that they affect an individual's health (other than through education).

Four studies use policy instruments to estimate the effect of education on smoking behavior.[3] Grimard and Parent (2007) and de Walque (2007b) use the Vietnam draft as an instrument for education. Grimard and Parent's instrument is an indicator variable for males

[3] Since smoking patterns vary significantly by country, we focus on studies in the United States. See also Arendt (2005), who exploits a reform in the Danish education system to estimate the effect of education on several health outcomes. He finds that education reduces the propensity to smoke. Since his instrument varies only by birth cohort, the identification issues in this paper are similar to those discussed below for Grimard and Parent (2007) and de Walque (2007b).

born between 1945 and 1950, while de Walque's instrument is a more complicated measure of induction risk into the Vietnam draft. Both find that an additional year of education significantly reduces the likelihood of smoking. The instruments in these studies vary only by gender and birth cohort. This necessitates an exclusion restriction on how they control for interactions between these variables. Grimard and Parent impose the functional form restriction that gender differences across birth cohorts vary only through a fourth order polynomial, while de Walque accounts for gender differences across birth cohorts via a linear trend.[4] These functional form assumptions may be restrictive since the first Surgeon General's Report on Smoking and Health came out in 1964, during the formative years of those affected by the Vietnam draft. As Grimard and Parent recognize, their analysis is appropriate only if this report had the same impact on men and women despite significant gender differences in smoking.

Other studies exploit local institutional changes that affect the supply of education. The instruments employed by Kenkel et al. (2006) measure the cost and difficulty of graduating high school, while Currie and Moretti (2003) use the availability of a local college as an instrument.[5] Both conclude that education significantly reduces the propensity to smoke. The instruments in these studies vary only by birth cohort and geographic area (state or county).[6] The identification assumption in these analyses is that variation in the supply of education is uncorrelated with unobserved factors that affect smoking behavior. A potential concern is that the policy variation exploited by these studies results from changes in unobserved characteristics that alter the cost or

[4] Grimard and Parent include a fourth order polynomial in age, rather than birth cohort. Age and birth year are very highly correlated since they use data from a narrow range of surveys, 1995-1999.

[5] While most studies use policy instruments that measure the "supply" of regulation, Kenkel et al. also rely (in part) on demand-side measures. For example, one of their instruments is the fraction of individuals who take the GED. Demand-side sources of variation such as this have been criticized as being potentially endogenous since they may be correlated with unobserved characteristics that influence health.

[6] These studies require an exclusion restriction on how they control for interactions between birth cohort and geography. If they had taken the flexible approach of including fixed effects for every combination of birth cohort and geographic area, these fixed effects would be perfectly collinear with the instrument employed. Much like Grimard and Parent and de Walque, Kenkel et al. and Currie and Moretti identify the effect of education only after imposing functional form and exclusion restrictions.

benefit of a given policy (Peltzman 1976, Becker 1983). For example, variation in high school graduation requirements over time may be a response to ability changes among the students. Graduation requirements may be correlated with the error term if ability is an omitted variable that affects the efficiency of health production.[7]

Although the four studies detailed above all conclude that education reduces the propensity to smoke, their identification strategy requires exclusion or functional form restrictions. Prior research shows that violations of these restrictions can lead to bias in instrumental variable applications (Meyer 1995, Bound et al. 1995, Bound and Jaeger 1996, Heckman 1996, Angrist and Krueger 2001) and other empirical analyses (Borjas 1985, Cawley et al. 1998). Recent trends suggest that these restrictions are a particular concern when analyzing smoking behavior (U.S. Department of Health and Human Services 1989, 1998, 2001). Gender differences in the propensity to smoke have declined over time. Recent generations are less likely to smoke, especially those whose formative years were subsequent to the first Surgeon General's Report on Smoking and Health in 1964. Regional smoking patterns also vary by time and generation. These trends highlight the need for robustly controlling for interactions between age, generation, time, and geography. Unfortunately, this is not possible in studies that employ instrumental variables that vary only along these dimensions.

While substantial progress has been made using policy instruments, these studies rely on an identification strategy that requires strong exclusion and functional form restrictions. Our analysis serves as a check on whether similar results can be obtained after relaxing these assumptions. The broader contribution of this paper is the introduction of an alternative method for testing theoretical models of how education affects health.

[7] Recognizing this issue, Kenkel et al. include an ability measure in the model specification. A potential concern is that ability is measured imperfectly, in which case the error term may include ability attributes that influence health.

III. Identification

We present a simple example that motivates the empirical methodology developed in Section IV. A control group framework is used that compares individuals who will acquire a given level of education in the following year to those who are one year older and currently have that particular level of education. The key identifying assumption is that these two groups have similar unobserved characteristics, which allows us to "difference out" the impact of the unobservables.

Consider the following stylized example, where for simplicity we assume the data is composed of six types ("groups") of individuals.

	Current Year			**Next Year**		
	Age	**Education**	**Student**	**Age**	**Education**	**Student**
Group 1	17	10	0	18	unknown	unknown
Group 2	17	11	1	18	unknown	unknown
Group 3	17	11	0	18	unknown	unknown
Group 4	16	10	0	17	10	0
Group 5	16	10	1	17	11	1
Group 6	16	10	1	17	11	0

We observe each individual *i*'s current age a_{it}, education e_{it}, student status s_{it}, and smoking decision $y_{it} \in \{0,1\}$. For those in groups 4 through 6 we also know their education and student status in the following year. This additional information is available due to the panel structure of the Current Population Survey (CPS), which as discussed in Section V is the data source used. Since the Tobacco Supplement of the CPS is given only occasionally, we do not observe their smoking decision in the following year. This example is constructed so that groups 1 and 4 are one year apart in their life cycle. That is, group 1's age, education, and student status in the current year is identical to group 4's age, education, and student status in the following year. The same relationship holds for groups 2 and 5 and groups 3 and 6.

A central concern in the literature that estimates the effect of education on health is that unobserved characteristics may be correlated with both variables. For example, an individual's time preference might affect whether he smokes. Data limitations typically prevent this variable from being included in the model specification. This is problematic since time preference is likely correlated with an individual's educational attainment decision. The variable controlling for an individual's education captures the effect of omitted correlated factors, leading to biased estimates. Suppose each group *k* has unobserved characteristics that have influence δ_k on their propensity to smoke. One might specify the following linear probability model where unobserved characteristics are controlled for through a set of group fixed effects.[8]

$$(3.1) \qquad \Pr(y_{it} = 1) = \sum_{k=1}^{6} \delta_k 1_{group_i = k} + \alpha_a a_{it} + \alpha_e e_{it} + \alpha_s s_{it}$$

The problem with specifying the model in this way is that parameters α_a, α_e, and α_s are not identified since a_{it}, e_{it}, and s_{it} are perfectly collinear with the set of group fixed effects. An additional assumption is required to identify the impact of education when unobserved characteristics are robustly controlled for in this manner. We assume that individuals with a given age, education, and student status in the current year have identical unobservable characteristics as those with the same age, education, and student status in the following year. As discussed in Section IV, this is a reasonable assumption since the two groups are born only one year apart, and make identical education decisions at the same point in their lives.

Recall that groups 1 and 4 are one year apart in their life cycle, and likewise for groups 2 and 5 and groups 3 and 6. Therefore, this assumption imposes three parameter restrictions: $\delta_1 = \delta_4$, $\delta_2 = \delta_5$, and $\delta_3 = \delta_6$. All of the parameters in equation (3.1) are identified once these restrictions are imposed. Ordinary least squares estimation of the regression model is equivalent

[8] See Section IV for discussion of why we employ a linear probability model.

to solving the following system of equations, where $\bar{y}^k$ denotes the average smoking rate of group *k*.

(3.2) $\bar{y}^1 = \delta_1 + 17\alpha_a + 10\alpha_e$

(3.3) $\bar{y}^2 = \delta_2 + 17\alpha_a + 11\alpha_e + \alpha_s$

(3.4) $\bar{y}^3 = \delta_3 + 17\alpha_a + 11\alpha_e$

(3.5) $\bar{y}^4 = \delta_1 + 16\alpha_a + 10\alpha_e$

(3.6) $\bar{y}^5 = \delta_2 + 16\alpha_a + 10\alpha_e + \alpha_s$

(3.7) $\bar{y}^6 = \delta_3 + 16\alpha_a + 10\alpha_e + \alpha_s$

Subtracting equation (3.5) from equation (3.2) yields $\hat{\alpha}_a = \bar{y}^1 - \bar{y}^4$. Since groups 1 and 4 are similarly selected, variation in smoking between them is due to their one year age difference. Subtracting equation (3.6) from equation (3.3) gives $\hat{\alpha}_a + \hat{\alpha}_e = \bar{y}^2 - \bar{y}^5$. Groups 2 and 5 are similarly selected, but differ by one year of age and one year of education. The smoking difference between the two groups is the combined impact of these two variables. Substituting for $\hat{\alpha}_a$ yields the following "difference in difference" estimator: $\hat{\alpha}_e = (\bar{y}^2 - \bar{y}^5) - (\bar{y}^1 - \bar{y}^4)$. Finally, the effect of being a student is obtained by comparing groups 3 and 6. Subtracting equation (3.7) from equation (3.4) gives $\hat{\alpha}_a + \hat{\alpha}_e - \hat{\alpha}_s = \bar{y}^3 - \bar{y}^6$, which simplifies to $\hat{\alpha}_s = (\bar{y}^2 - \bar{y}^5) - (\bar{y}^3 - \bar{y}^6)$ after substituting for $\hat{\alpha}_a$ and $\hat{\alpha}_e$. Groups 3 and 6 differ by age, education, and student status, while groups 2 and 5 differ only by age and education. The impact of student status is estimated by subtracting the smoking difference between groups 3 and 6 from the smoking difference between groups 2 and 5.

To summarize, the control group methodology identifies the effect of education from differences between similarly selected groups of individuals that are one year apart in their life cycle. This simple example provides the intuition for how the empirical methodology detailed in Section IV allows us to identify the effect of education on smoking behavior.

IV. Methodology

The empirical framework detailed in this section allows us to differentiate between whether education affects smoking through the causal channels detailed in Section II and Farrell and Fuch's (1982) competing hypothesis that the relationship between education and smoking is due to unobserved factors that affect both variables. In developing the model, minimal structure is imposed. For example, fixed effects are employed in place of potentially restrictive functional form assumptions. In addition, the use of a control group methodology allows us to identify the causal effect of education without the exclusion restrictions employed in prior research. As in the example given in Section III, we compare those who will acquire a given level of education in the following year to those who are one year older and currently have that particular level of education. The key identifying assumption is that these two groups have similar unobserved characteristics. This assumption is reasonable since they are born only one year apart, and make identical education decisions at the same point in their lives. The empirical methodology "differences out" the impact of the unobserved characteristics correlated with education, thereby isolating its causal effect.[9]

Let $y_{it} \in \{0,1\}$ denote individual *i*'s smoking behavior in time *t*, where each time period is one year apart. We model the decision to smoke as a function of age a_{it}, education e_{it}, student status s_{it}, and all other characteristics δ_{it} that affect smoking.

$$(4.1) \quad \Pr(y_{it} = 1) = \alpha_a a_{it} + \alpha_e e_{it} + \alpha_s s_{it} + \delta_{it}$$

The impact of all factors other than age, education, and student status is represented by δ_{it}. This variable controls for the effect of unobserved characteristics such as an individual's time preference, as well as observed characteristics such as sex, race, and state of residence. If all components of δ_{it} were included in the model specification, one could use equation (4.1) to

[9] This framework can be applied to other settings, such as Tenn (2007) who analyzes the effect of education on voter turnout.

obtain consistent estimates of the effect of education. A problem arises when only a subset of the characteristics contained in δ_{it} is included in the model. As is well known, the omission of unobserved correlated factors can result in biased estimates. First, we demonstrate how the control group methodology allows the causal effect of education to be identified even when δ_{it} is completely unobserved. Later we consider the situation where some, but not all, of the characteristics contained in δ_{it} are observable.

The effect of age, education, and student status is separated from all other characteristics δ_{it} so that the model can control for differences between the "treatment" and "control" groups. Specifically, we must account for the age and education the group one year further along in their life cycle has already acquired, but which their younger counterparts will not obtain until the following year. The model specification differentiates between student status and an individual's educational attainment. Student status captures environmental influences such as peer effects (Norton et al. 1998, Gaviria and Raphael 2001, Powell et al. 2005), whereas educational attainment may enhance the efficiency of health production.

In specifying equation (4.1) we rely on a linear probability model due to the difficulty of implementing a control group methodology in nonlinear discrete choice models such as the logit or probit.[10] As detailed below, the only restriction imposed on δ_{it} is that its expected value is the same across the treatment and control groups. In a nonlinear model additional assumptions would be required since the effect of education would depend on the entire distribution of δ_{it}, which is unobserved. The linear probability model allows us to avoid such restrictions. Note that despite being linear, equation (4.1) is still quite flexible since there are no restrictions on what characteristics δ_{it} might contain.

[10] The linear probability model has been widely employed in the health-education literature. In particular, use of a linear probability model enhances comparability to the latest research on education's effect on smoking since recent papers on the topic all rely upon that framework (de Walque 2007b, Grimard and Parent 2007, Kenkel et al. 2006, Currie and Moretti 2003). The motivation for using a linear probability model in these analyzes differs, however. They use linear models due to the ease of applying instrumental variables.

A longitudinal dataset is required to implement the control group methodology detailed below. As discussed in Section V, we rely on the Current Population Survey. The CPS panel is designed such that half of those surveyed in time *t* can potentially be matched to the previous year's survey (time *t*-1), and the other half to the following year's survey (time *t*+1). For those matched to the previous year's survey we observe $g_{i,t-1}$ and g_{it}, where $g_{it} = \{a_{it}, e_{it}, s_{it}\}$ denotes the triplet containing individuals *i*'s age, education, and student status in time *t*. For those matched to the following year's survey we observe g_{it} and $g_{i,t+1}$. The Tobacco Supplement is not given every year. Therefore, we do not observe smoking behavior in time *t*-1 or *t*+1 for individuals who participate in the Tobacco Supplement in time *t*. This is not an issue, however, since the control group methodology does not require that we observe future or past smoking behavior.

Due to panel attrition, some individuals who take the CPS survey in time *t* cannot be matched to the survey given in time *t*-1 or *t*+1. Since only g_{it} is observed for such individuals, we cannot apply the control group methodology to this group. We therefore exclude them from the analysis. This exclusion does not lead to bias so long as one recognizes that the estimates correspond to the effect of education among those who do not drop out of the panel, rather than the effect of education across all individuals.[11]

We take the expectation of equation (4.1) conditional on each respondent's age, education, and student status in the two years they participate in the panel, where m_{it} denotes an indicator variable for whether individual *i* can be matched from the survey given in time *t* to the survey given in time *t*+1.

$$(4.2) \quad \Pr(y_{it} = 1 \mid g_{i,t-1}, g_{it}, m_{i,t-1} = 1) = \alpha_a a_{it} + \alpha_e e_{it} + \alpha_s s_{it} + E(\delta_{it} \mid g_{i,t-1}, g_{it}, m_{i,t-1} = 1)$$

$$(4.3) \quad \Pr(y_{it} = 1 \mid g_{it}, g_{i,t+1}, m_{it} = 1) = \alpha_a a_{it} + \alpha_e e_{it} + \alpha_s s_{it} + E(\delta_{it} \mid g_{it}, g_{i,t+1}, m_{it} = 1)$$

[11] As discussed in Section II, previous studies estimate the impact of education for those affected by a particular policy instrument (e.g., those who entered college to avoid the Vietnam draft). Since the policy instruments in these analyses affect only a small fraction of the population, our estimation method comes closer to measuring the effect of education on the educated (i.e., "treatment on the treated").

The following assumption allows us to identify the causal effect of education for those who remain in the CPS panel for two years.

$$(4.4) \quad E(\delta_{it} \mid g_{it} = g, g_{i,t+1} = g', m_{it} = 1) = E(\delta_{it} \mid g_{i,t-1} = g, g_{it} = g', m_{i,t-1} = 1), \forall g, g'$$

Equation (4.4) states that individuals with a given age and education in the current and following year have identical unobserved characteristics (in expectation) as their counterparts with the same age and education in the previous and current year.

We believe this is a valid assumption for several reasons. First, the two groups are born only one year apart. We are unaware of research showing significant generational differences across cohorts born a single year apart. Instead, generational differences are typically thought to arise across much longer time frames. For example, Currie and Moretti (2003) control for generational effects using decade of birth dummy variables. Implicitly, they assume generational differences within each decade are sufficiently minor that they can be ignored. We make the much weaker assumption that generational differences across adjacent birth years are insignificant.

Second, the two groups make the same educational choices at the same point in their lives. It is reasonable to presume that both groups are affected by the same selection process regarding educational attainment. Without specifying a particular process for how individuals select into a given education choice, we simply assume that selection regarding educational attainment leads those who differ by being one year ahead in their life cycle to have the same unobserved characteristics as their younger counterparts one year behind.

Lastly, not only does equation (4.4) compare individuals who make the same education decision at the same point in their lives, they share the additional characteristic of not dropping out of the panel in the second year of the survey. Without specifying a particular selection process regarding sample attrition, we rely on the assumption that attrition affects the two groups similarly since both are comprised entirely of individuals who do not drop out of the panel in the second year they are given the survey.

Assumption (4.4) allows us to control for selection bias in a very flexible manner. By analyzing differences in smoking rates between groups with the same expected unobservables (per equation 4.4), unobserved characteristics are differenced out without making any additional assumptions regarding their distribution across individuals. This is significantly less restrictive than models that specify a particular form of selection (e.g., Heckman 1979).

Let g_i and g_i' respectively denote the triplet containing individual *i*'s age, education, and student status in the first and second year he participates in the panel. For individuals who can be matched to the previous year, $g_i = g_{i,t-1}$ and $g_i' = g_{it}$. For the remaining individuals who can be matched to the following year, $g_i = g_{it}$ and $g_i' = g_{i,t+1}$. Using this notation, equations (4.2) and (4.3) are combined into a single equation since they are a function of the same variables after identification assumption (4.4) is imposed. To simplify notation, we drop whether an individual can be matched across survey years from the set of conditioning variables; implicitly, all expectations are taken across the set of individuals who participate in the survey in both years.

$$(4.5) \quad \Pr(y_{it} = 1 \mid a_{it}, e_{it}, s_{it}, g_i, g_i') = \alpha_a a_{it} + \alpha_e e_{it} + \alpha_s s_{it} + E(\delta_{it} \mid g_{i,t-1} = g_i, g_{it} = g_i')$$

To clarify how the causal effect of education is identified, define $\delta^*(g, g', t) = \alpha_a a(g') + \alpha_e e(g') + \alpha_s s(g') + E(\delta_{it} \mid g_{i,t-1} = g, g_{it} = g')$, where $a(g')$ is the age element of g', and $e(g')$ and $s(g')$ are analogously defined. Equation (4.5) can then be written as follows.

$$(4.6) \quad \Pr(y_{it} = 1 \mid a_{it}, e_{it}, s_{it}, g_i, g_i') = \alpha_a (a_{it} - a(g_i')) + \alpha_e (e_{it} - e(g_i')) + \alpha_s (s_{it} - s(g_i')) + \delta^*(g_i, g_i', t)$$

For individuals who can be matched to the previous year, $g_i' = g_{it}$, so the first three terms of equation (4.6) equal zero. The last term $\delta^*(g, g', t)$, is estimated via a set of fixed effects for each combination of g, g', and *t*. These fixed effects recover average smoking behavior, separately for each survey year, conditional on an individual's current and previous age, education, and student status.

The remaining parameters $\{\alpha_a, \alpha_e, \alpha_s\}$, which reflect the effect of age, education and student status, are identified from changes in these variables between survey years for individuals who can be matched to the following year's survey. Under identification assumption (4.4), their propensity to smoke is equal to the average smoking rate of their counterparts one year ahead in their life cycle, adjusted for the effect of the age and education they will experience in the upcoming year (which their older counterparts have already experienced, but which they will not until one year later).

This completes the model specification. The model is estimated via weighted least squares using the sample weights provided by the CPS. Weights are employed since we take expectations over the distribution of unobserved characteristics in the population when deriving the model. Estimating the model without weights would be inconsistent with the derivation since the CPS over-samples certain population groups, such as those living in small states. This is primarily a theoretical concern, however, since similar results are obtained when sample weights are not used. To account for both heteroskedasticity and correlated errors across individuals, robust standard errors are reported that cluster by state of residence.

Generational Differences

For the remainder of the section we discuss potential reasons why identification assumption (4.4) might be violated, and how we modify the empirical framework to avoid bias. First, while generational differences across adjacent birth cohorts are unlikely to be significant, the assumption of no generational differences may be overly restrictive. Further, although sample attrition is likely to incur a comparable selection effect on cohorts one year apart in their life cycle, attrition might not have the exact same selection effect on the two groups.

To account for such differences, we relax assumption (4.4) by allowing the expected difference in unobservables between cohorts one year apart in their life cycle to be an arbitrary function $d(a,t)$ of age and time.

(4.7) $$E(\delta_{it} \mid g_{it} = g, g_{i,t+1} = g', m_{it} = 1) = E(\delta_{it} \mid g_{i,t-1} = g, g_{it} = g', m_{i,t-1} = 1) + d(a(g),t), \forall g, g'$$

This assumption is substantially weaker than equation (4.4), which is equivalent to assuming $d(a,t) = 0, \forall a,t$. Rather than imposing a functional form assumption, we estimate $d(a,t)$ via a set of fixed effects for every combination of age and time. Therefore, equation (4.7) flexibly accommodates any aggregate differences in unobserved characteristics between cohorts one year apart in their life cycle that vary by age, time, or birth year (the latter being true since birth year is determined by age and time).

Replacing equation (4.4) with equation (4.7) leads to the following modification of equation (4.6), where we define $\alpha_a(a,t) = \alpha_a - d(a,t)$.

(4.8) $$\Pr(y_{it} = 1 \mid a_{it}, e_{it}, s_{it}, g_i, g_i') = \alpha_a(a_{it},t)(a_{it} - a(g_i')) + \alpha_e(e_{it} - e(g_i')) + \alpha_s(s_{it} - s(g_i')) + \delta^*(g_i, g_i', t)$$

In equation (4.6) the marginal effect of age is the same for all individuals; in equation (4.8) it varies by age and time. The model simplifies in this manner since the term $a_{it} - a(g_i')$ is equivalent to a dummy variable for those matched to the following year's survey. It equals zero for individuals matched to the previous year, and equals -1 for those matched to the following year (since they are one year younger than their counterparts one year ahead in their life cycle).[12]

The set of coefficients $\alpha_a(a,t)$ is estimated via an interaction between $a_{it} - a(g_i')$ and a set of fixed effects for every combination of age and time. This flexibility allows the model to accommodate potential differences between cohorts one year apart in their life cycle. Unfortunately, it is not possible to identify the effect of age since $\alpha_a(a,t)$ captures the combined effect of age and unobserved characteristics $d(a,t)$. As our objective is to estimate the marginal effect of education, rather than age, this limitation is relatively minor.

[12] Since the CPS survey given in the second year of the panel may be administered on a different day of the month, the time elapsed between surveys ranges between 11 and 13 months. Therefore, the difference in reported age between survey years takes values between 0 and 2 years. To avoid this problem caused by measuring age in whole years, the change in age between surveys is taken to be one year for all individuals.

Additional Control Variables

In equation (4.1) the probability of being a smoker is a function of age, education, and student status, with the effect of all other variables captured by δ_{it}. By matching individuals one year apart in their life cycle, the control group methodology described above differences out the effect of δ_{it}. However, one might include additional control variables in the model specification to account for differences that potentially violate identification assumption (4.7).[13] For example, suppose that living with a parent makes it harder to conceal smoking, causing such individuals to be less likely to smoke. Since younger individuals are more likely to live with a parent, this characteristic can lead to differential smoking rates between cohorts one year apart in their life cycle. Whether this violates assumption (4.7) depends on whether the likelihood of living with a parent varies by education (if younger individuals are more likely to do so, independent of their education, then this effect would be absorbed into the set of age coefficients).

Although this example suggests that including additional controls can be beneficial, doing so can also be problematic. To analyze this issue let $\delta_{it} = X_{it}\beta + \delta_{it}^{u}$, where X_{it} is a set of observed characteristics and δ_{it}^{u} contains the effect of all remaining unobservables. If the model is derived taking expectations conditional on X_{it} (and the other variables employed earlier), one arrives at an equation analogous to (4.8), where
$b(g,g',t,X) = E(\delta_{it}^{u} \mid g_{i,t-1} = g, g_{it} = g', X_{it} = X) - E(\delta_{it}^{u} \mid g_{i,t-1} = g, g_{it} = g')$, and
$\delta^{u*}(g,g',t)$ is defined similarly to $\delta^{*}(g,g',t)$ after replacing δ_{it} with δ_{it}^{u}.[14]

$$(4.9)\quad \Pr(y_{it} = 1 \mid a_{it}, e_{it}, s_{it}, g_i, g_i', X_{it}) = \alpha_a(a_{it},t)(a_{it} - a(g_i')) + \alpha_e(e_{it} - e(g_i')) + \alpha_s(s_{it} - s(g_i')) + \delta^{u*}(g_i, g_i', t) + X_{it}\beta + b(g_i, g_i', t, X_{it})$$

[13] A second reason for including additional control variables is that doing so (weakly) increases the explanatory power of the model, potentially leading to more precise parameter estimates.

[14] Equation (4.7) must be modified so that the expected value of unobserved characteristics δ_{it}^{u} is the same across the treatment and control groups conditional on age, education, student status, and observed characteristics X_{it}.

Equation (4.9) contains two additional terms that are omitted from equation (4.8). The first, $X_{it}\beta$, controls for the impact of observed characteristics. The second term $b(g,g',t,X)$ is unobserved, and captures the difference in the expected value of the unobservables depending on whether one conditions on X_{it}. If this term is correlated (uncorrelated) with the control variables, omitting it from the model specification will (will not) lead to biased estimates of the effect of education on smoking.

Since it is not clear whether one should control for additional characteristics, we estimate the model both including and excluding a set of observed characteristics X_{it}. Doing so allows us to assess the robustness of the empirical methodology. As discussed in Section VI, the results are not sensitive to whether additional control variables are included in the model specification.

V. Data

The data used in the analysis are drawn from the Tobacco Supplement of the Current Population Survey (CPS). The CPS is a nationally representative household survey that is primarily used as a source of labor market and demographic information. While the basic monthly survey does not contain questions regarding smoking, the Tobacco Supplement has occasionally been given since 1992 and reports smoking behavior for each survey respondent. We use this data source since it best fulfills the requirements for implementing the control group framework detailed in Section IV. Namely, it is a longitudinal dataset with sufficient sample size to provide precise estimates of education's effect on smoking even though our identification strategy relies on a limited source of variation: educational differences between groups one year apart in their life cycle. The CPS Tobacco Supplement has previously been used by Grimard and Parent (2007) to analyze the effect of education on smoking behavior.

Since the CPS is more commonly used as a cross-sectional dataset, its panel structure may be unfamiliar. CPS respondents are surveyed for four consecutive months, removed from the sample for the next eight months, and then resurveyed for four more months before retiring from the panel. At any given time, half the respondents are in their first sequence of surveys and

can potentially be matched to the survey given one year later. The remaining individuals in their second sequence of surveys can potentially be matched to the survey given one year earlier.

A shortcoming of the CPS is high attrition from the panel. The primary reason is that the CPS does not follow individuals who move between surveys. Instead, the CPS interviews whoever currently lives at a given residence. Most recently, Neumark and Kawaguchi (2004) undertake a detailed analysis of attrition bias in the CPS. They conclude that the longitudinal advantages of the CPS outweigh any bias arising from panel attrition. However, Peracchi and Welch (1995) do uncover small biases when studying labor force transitions across matched CPS respondents. Although important economic insight can be revealed using matched data, both analyses conclude that attrition bias cannot be ignored and must be considered in the context of each particular study. For this reason, the methodology developed in Section IV explicitly accounts for the possibility that those who drop out of the panel have unobservable differences from those who remain.

A second difficulty with using the CPS relates to how it measures education. Although the CPS reported years of schooling until 1991, subsequent surveys collect information on the highest degree obtained. Education is reported in categories spanning multiple years of education, rather than the exact number of completed years of schooling. For example, an individual with 13 years of school in the first survey year and 14 years of school in the second might report "some college" for both periods, even though he obtained an additional year of education between the two years.

This data limitation poses a potential difficulty for our analysis, since single year differences in educational attainment is the source of variation relied upon. We overcome this problem by taking advantage of the fact that an individual's educational attainment e_{it} is his accumulated years of being a student, i.e., $e_{it} = \sum_{\tau \geq 1} s_{i,t-\tau}$. Even though change in education

$e_{i,t+1} - e_{it}$ is not directly reported by the CPS, it can be calculated as an individual's student status s_{it} in the earlier time period.[15]

The key assumption when measuring change in education in this manner is that those who are currently a student remain so for the rest of the year. The validity of this assumption is evaluated in two ways. First, we use the CPS to calculate the fraction of people in school over the course of the calendar year. We find very little variation in student status between September and April, which comprises the period when schools are traditionally in session (enrollment slightly declines in May, when schools with early calendars end the year, with a much larger drop between June and August that coincides with when most schools are on summer vacation). This pattern is consistent with the assumption that individuals who start the school year remain students for the rest of the academic calendar.

A second method of validating our measure of change in education is to compare it to the usual definition for CPS surveys given prior to 1992, which report years of schooling rather than highest degree obtained. Tenn (2007) finds that the two methods provide very similar results. This gives us confidence that student status in the earlier year is an accurate measure of an individual's change in education from one year to the next.

One complication in measuring the change in education between surveys in this manner is that the Tobacco Supplement is sometimes given in the middle of the academic calendar. Under the assumption that individuals who start the academic year complete that year of school, change in education can be measured as follows where θ_i denotes the number of months between September, the approximate start of the academic year, and when the Tobacco Supplement is given (which varies by year).

$$(5.1) \quad e_{i,t+1} - e_{it} = (1 - \frac{\theta_i}{9}) s_{it} + \frac{\theta_i}{9} s_{i,t+1}$$

[15] Since educational attainment is one's accumulated years of being a student, it is useful to consider how the effects of these two variables are separately identified. In the data, the majority of variation in student status comes from individuals leaving school. In contrast, most educational changes come in the middle of an individual's academic life and are therefore not associated with a change in student status.

The first term in equation (5.1) corresponds to the fraction of the previous year's academic calendar completed between time t and t+1, while the second term corresponds to the fraction of the current academic calendar completed as of when the survey was given.

If $s_{it} = s_{i,t+1}$, an individual's change in education between survey years equals zero if he is not a student in either year, and equals one if he is a student in both years. An individual's change in education is a fraction of a year only for those individuals who change student status between survey years. We recognize the potential for an individual's change in education to be mismeasured for this latter group. 16% of individuals in our dataset change student status between survey years. As a robustness check, in some specifications we restrict the data to the remaining 84% of individuals who do not change student status. Similar estimates for the effect of education are obtained, suggesting that measurement error does not have a major impact on our findings (see Section VI).[16]

A third limitation of the CPS is that it reports student status only for ages 16 to 24. As detailed in Section IV, it is important to differentiate between school enrollment, which may affect smoking via peer effects (and other environmental factors), and educational attainment, which may impact the efficiency of health production. Lacking this critical piece of information all other age groups are excluded from the analysis. This restriction is fairly minor since most people take up smoking, if ever, between these years. Data from the CPS Tobacco Supplement is used to construct Figure 1, which reports smoking rates by age.[17] Few people start to smoke prior to age 16, and there is little change in the percentage of smokers after age 24.[18]

While the Tobacco Supplement has been administered 17 times since its inception in 1992, many of the surveys cannot be used. Specifically, six surveys are given in May or June.

[16] The effect of student status is not identified after restricting the data, since doing so eliminates all variation in school enrollment between survey years.

[17] Following the literature, a "smoker" is defined as an individual who has smoked at least 100 cigarettes in his lifetime.

[18] See also, U.S. Department of Health and Human Services (1989).

Since many individuals are on summer break during that time, one cannot use equation (5.1) to calculate an individual's change in education for respondents in these surveys. Two additional surveys (September 1995 and January 1996) are excluded due to a change in sample design that prevents the matching of individuals across survey years. The September 1992 survey is also excluded since the CPS changed the way it measured education between 1991 and 1992, making matching to the previous year's survey problematic. After excluding these surveys, seven surveys given between 1998 and 2003 remain, as well as an earlier survey given in January 1993. To maximize the comparability of the data sample, we exclude the 1993 survey since it lies outside the narrow time frame covered by the remaining surveys. This avoids potential biases due to pooling data across distant years, during which time the model parameters may vary.

We match individuals across surveys using the following fixed characteristics: state of residence, gender, and household/individual identifiers (household id, household number, individual line number, and month in sample). As Madrian and Lefgren (2000) point out, data inaccuracies can result in the match of two distinct individuals rather than the same individual in two different periods. Based on their recommendations, matches are rejected if the difference in age between potential matches is not between zero and two years, if the education level reported in the follow-up survey is less than that reported in the first survey, or if different races are reported across surveys.[19] Approximately 5% of potential matches are invalidated due to these reasons.

The final dataset is constructed of individuals aged 16 to 24, residing in the United States,[20] from the Tobacco Supplements given in September 1998, January 1999, January 2000, November 2001, February 2002, February 2003, and November 2003. Across all seven surveys, this dataset comprises 41,882 individuals, or approximately six thousand observations per

[19] Starting in 2003, respondents can report multiple races. A match between an individual reporting a single race in 2002, but multiple races in 2003, is considered valid. This has little impact on our analysis, since 0.5% of the data sample reports multiple races.

[20] This includes all 50 states and the District of Columbia.

survey.[21] Table 1 reports summary statistics for each variable employed in the analysis. 19% of our sample has ever been a smoker, while 15% are current smokers and 11% smoke everyday. As expected given their average age of 19.5 years, the sample is primarily comprised of individuals who have (at least) started high school but have not graduated college, with 63% enrolled in school. In addition to age, education, and student status, in some specifications we control for additional characteristics that potentially explain smoking behavior: gender, race/ethnicity, marital status, native born, veteran status, living in a metropolitan statistical area, and whether the respondent currently lives with a parent. This is similar to the set of controls employed in previous studies of education's effect on smoking.

VI. Results

We begin by estimating the cross-sectional relationship between education and smoking using a model that ignores the endogeneity of education. A linear probability model is employed that controls for a variety of observable characteristics that potentially explain smoking behavior: gender, race/ethnicity, marital status, native born, veteran status, living in a metropolitan statistical area, and whether the respondent currently lives with a parent. To flexibly account for age, generation, and time, we include a set of fixed effects for every combination of age and survey year.[22] In addition, the model includes a set of fixed effects for state of residence that

[21] We arrive at the final data sample as follows. Across all seven surveys, 81,008 individuals aged 16 to 24 participated in the CPS Tobacco Supplement. Of these, 33,085 could not be matched to either the current or previous year. An additional 6,041 observations have missing student information due to not being aged 16 to 24 in both survey years. This leaves a final data sample of 41,882 individuals. The inclusion rate in our analysis is similar to other panel studies, such as Kenkel et al. (2006) who employ the National Longitudinal Study of Youth (NLSY). Importantly, our methodology explicitly corrects for selection bias that could arise due to sample attrition, which other studies of education's effect on smoking have not done.

[22] The Tobacco Supplement was given in February and November 2003. Throughout the analysis we treat these two surveys as being from different "years" to maximize the flexibility of the model specification.

controls for geographic variation in factors such as cigarette taxes and attitudes towards smoking.[23]

Table 2 presents estimates of the effect of education on smoking from this model, which does not control for selection bias. As expected, the educated are less likely to smoke. This is the case for all three smoking measures (ever, current, or everyday smoker). The difference between those with a high school and college degree is quite large, 17 to 21 percentage points depending on the smoking measure.[24] A potential concern with using a data sample of those aged 16 to 24 is that education's impact might not appear until later in life. The results shown in Table 2 suggest this is not the case; they are similar to estimates of education's effect from prior studies that do not control for selection bias, but which use data samples of older individuals who have largely completed their schooling.[25] This comparison illustrates that, when selection bias is ignored, the effect of education for our data sample is similar to the effect of education for older individuals. This is not surprising given that the decision to become a smoker is primarily made between the ages of 16 and 24 (see Figure 1).

The key question is whether the cross-sectional relationship between education and smoking is causal, or is instead due to the effect of unobserved characteristics correlated with both variables. Estimates obtained using the control group methodology detailed in Section IV answer this question. The results presented in specification (i) of Table 3 correspond to equation (4.8), which controls for age, education, student status, and the set of fixed effects that accounts

[23] In this and subsequent analysis, we checked whether the results are sensitive to letting the state fixed effects vary by year. Doing so would accommodate, for example, variation in cigarette taxes over time. Similar results were obtained.

[24] Since very few 16 to 24 year olds have an advanced degree, we estimate a single effect for all individuals with at least a college degree.

[25] For example, de Walque (2007b) finds that those with a college degree are 17 percentage points less likely to be a current smoker. Similarly, Grimard and Parent (2007) find that individuals with a college degree are 21 percentage points less likely to smoke everyday. Note that these estimates correspond to specifications that do not control for the endogeneity of education (specifically, they are not the instrumental variables estimates presented later in their analyses). As such, they are directly comparable to the results presented in Table 2.

for unobserved factors potentially correlated with education. Across all three measures of smoking (ever, current, or everyday smoker), education has little effect on smoking. An additional year of education reduces the probability of smoking by 0.2 to 0.7 percentage points, depending on the dependent variable employed (the standard errors range from 0.9 to 1.2 percentage points). These estimates are neither statistically significant nor economically large. This result contrasts with previous research that identifies the effect of education via instrumental variables, which finds that education has a large impact even after controlling for selection bias (see Section II).

The second result emerging from our analysis is that being a student leads to a moderate reduction in the probability of smoking. Across the three dependent variables, the point estimates range from -5.2 to -3.0 percentage points (the standard errors range from 1.0 to 1.2 percentage points). These estimates are statistically significant at any conventional level. While our analysis does not offer insight into the causal mechanism behind this result, one possibility is that we are measuring peer effects. Previous studies conclude that the smoking decisions of one's peers are an important determinant of smoking behavior (Norton et al. 1998, Gaviria and Raphael 2001, Powell et al. 2005). Since the smoking rate among students is quite low, it is plausible that being a student, and interacting with other students, lowers the likelihood of smoking.

Specification (ii) of Table 3 expands the model to include the control variables employed earlier (gender, race/ethnicity, marital status, native born, veteran status, living in a metropolitan statistical area, whether the respondent currently lives with a parent, and fixed effects for state of residence).[26] If groups one year apart in their life cycle have similar characteristics, as we assume, it is unnecessary to include these additional control variables in the model specification since their impact is differenced out when making comparisons between the two groups. The sensitivity of our results to the inclusion of additional variables allows us to assess the validity of

[26] The parameter estimates for these additional control variables are similar to those presented in Table 2.

this identification assumption. We find the results do not depend on whether additional control variables are included in the model. Education has little effect in either specification, while being a student reduces the likelihood of smoking.

Specification (iii) and (iv) include interactions that let the effect of education and student status differ for those in high school and college. Doing so accommodates potential differences in the health curriculum across educational settings. Those in high school often take health classes that inform on the consequences of smoking, whereas a college curriculum typically does not require such class work. We find this difference between high school and college has little impact on smoking behavior. The effects of high school and college education are not statistically different from zero, or each other, at any conventional level of significance. Being a high school or college student reduces the propensity to smoke by a similar magnitude. This is noteworthy given that different margins of variation identify these two effects. The effect of being a high school student is primarily identified from those *leaving* high school. In contrast, the effect of being a college student is primarily identified from individuals *starting* college.[27]

Sensitivity Analysis

Table 4 presents results from additional regressions that allow us to assess the impact of measurement error. Two measurement issues are considered. First, as detailed in Section V, for those individuals who take the Tobacco Supplement in the middle of the academic calendar we must estimate how much education they obtained between survey years. This is not an issue for those who have the same student status in both years. Such individuals are likely to have been either in school, or out of school, for the entire period. However, the remaining individuals who

[27] In further analysis we explore whether the effect of education might differ by race or gender. The model is estimated separately for males, females, whites, non-whites, white males, and white females. In all specifications the effect of education is small, and is not statistically significant at any conventional level. As before, we find student status reduces the probability of smoking. In some specifications this effect is not statistically significant. This is primarily due to larger standard errors when the sample is split by race and/or gender.

changed student status between surveys are potentially problematic, since they completed only a fraction of a year of school (which is estimated via equation 5.1).

To test whether measurement error in calculating each individual's change in education leads to attenuation bias in the estimated effect of education, we restrict the data sample to those individuals who do not change student status between survey years. As observed in line (a) of Table 4, restricting the dataset in this manner has little impact on the parameter estimates. This suggests that measurement error in calculating each individual's change in education between survey years is not a significant problem.

A second potential source of measurement error relates to how individuals are matched across CPS surveys. As described in Section V, data inaccuracies can result in the match of two distinct individuals rather than the same individual in two different periods. To eliminate "bad matches" we follow Madrian and Lefgren (2000) and remove individuals with implausible changes in certain characteristics (gender, age, race, and education). In particular, we required that education be weakly increasing across survey years. Individuals who increased education by more than one category were not excluded since this can occur for valid reasons (e.g., an individual may have skipped a grade in school). Nonetheless, we recognize the possibility that such changes could be due to measurement error in reported education. We therefore test whether our results are sensitive to excluding those who increase their education by more than one category in a single year. Line (b) of Table 4 reports the results of this sensitivity analysis, which are consistent with our previous findings. This suggests that measurement error in educational attainment is not a problem.[28]

A potential criticism of our analysis relates to the "incidental parameter" problem (Lancaster 2000). Bias can occur in certain settings where the number of model parameters increases as the sample size grows. For example, in analyses of panel datasets where fixed effects for each individual are employed, the number of model parameters increases as

[28] Line (c) imposes the restrictions from both line (a) and (b) of the table. Similar results are obtained.

individuals are added to the dataset. This does not occur in our analysis. Even though the model employs a large number of fixed effects to control for selection bias in education (one for every combination of year, age, education, and student status), the number of fixed effects is not an increasing function of the sample size. As such, arbitrarily precise estimates of these effects can be obtained as the number of individuals in the dataset becomes arbitrarily large. Nonetheless, to demonstrate that the large number of fixed effects included in the model is not an issue, we re-estimate the model after restricting the fixed effects that control for selection bias to be equal across survey years.[29] Since our analysis employs data from a narrow range of years, 1998-2003, this pooling assumption is plausible since selection bias regarding education choice is unlikely to have significantly changed over such a short period of time. Restricting the fixed effects to be identical across survey years greatly reduces the number of model parameters.[30]

The results from this restricted model are presented in Table 5. For baseline specification (i), a year of education reduces the likelihood of smoking by 0.2 to 0.6 percent points, depending on the measure of smoking behavior employed (ever, current, or everyday smoker). The standard errors range from 0.9 to 1.2 percentage points. None of these effects are large or statistically significant. Further, consistent with our earlier results being a student negatively impacts smoking behavior. Depending on the dependent variable, the effect of student status ranges from -4.5 to -2.7 percentage points (the standard errors range from 0.9 to 1.0 percentage points). These effects are statistically significant at any conventional level, indicating an impact from peer effects or some other environmental factor. Similar results are obtained in specification (ii) that controls for additional observable characteristics. These results show that

[29] To control for year effects in the restricted model, the specification includes a set of dummy variables for each survey.

[30] The number of fixed effects declines from 2,763 in the original model to 802 in the restricted model. The pooled model contains a larger number of fixed effects than one might expect because a small number of people have an unusual combination of age, education, and student status. Similar results are obtained after restricting the data sample to the 100 groups with the largest number of people, which represent 89% of the total observations.

our reliance on a large number of fixed effects does not explain why education has little impact on smoking.

VII. Discussion

Our results indicate that the strong cross-sectional relationship between education and smoking is due to unobserved factors correlated with both variables, rather than from a causal effect of education. To assess the plausibility of this finding, we examine whether an effect from education can be observed in the raw data. High school graduates are split into two groups depending on whether they have started college. We aggregate the data in this manner since a sizable fraction of high school graduates do not continue on to college. Far fewer people end their academic career at lower levels of education. Using the CPS Tobacco Supplements, the average smoking rate for each group is calculated separately by age.[31] This is done for ages 21 to 24. We do not compute smoking rates for older individuals because the CPS does not report student status beyond age 24, so we cannot be sure whether an individual has started college. We exclude those younger than 21 since the fraction of the population who has started college increases until that age. For every age between 21 and 24, however, 34% of the high school graduates in our sample have not started college. This stability is important to our analysis since it indicates that within-group variation in smoking rates by age is not affected by composition changes.

Figure 2 reports the fraction of each group that has ever been a smoker, separately by age.[32] While the college educated group is less likely to smoke, this is not evidence that education has a causal effect since unobserved factors may be correlated with education. Instead, we emphasize that the two age profiles are approximately parallel. While both groups

[31] In our earlier analysis, we restricted the data to individuals who do not drop out of the panel. Since we do not exploit the panel structure of the CPS in this section, this restriction is relaxed.

[32] Similar results are obtained using current smoking status or whether the respondent smokes everyday.

are more likely to smoke as they get older, the difference between them in the propensity to smoke is remarkably stable. Between the ages of 21 and 24, we calculate that individuals in the college group acquire nearly two years of schooling (on average), while the high school group obtains no additional education. If education has a causal effect on smoking, the smoking rate difference between the two groups should increase as the college group becomes better educated. Figure 2 does not reveal any evidence of this. Instead, the smoking gap between the two groups shrinks slightly for the older age groups. The results of our empirical analysis explain why this occurs: as those in the college group complete their education and leave school, they become more likely to smoke. Figure 2 is consistent with our conclusions that accumulated education does not affect smoking, but being a student reduces the propensity to smoke.

Two prior studies similarly conclude that unobserved factors are important determinants of smoking behavior. Farrell and Fuchs (1982) and DeCicca et al. (2002) use panel datasets to show that *future* education predicts *current* smoking behavior. Their results suggest that a significant portion of the relationship between education and smoking is not causal, and is instead due to the omission of unobserved factors correlated with both variables. The effect of education should therefore be much smaller after controlling for endogeneity. Recent studies have instead found that the effect of education (weakly) increases when instrumental variables are used to control for selection bias. One explanation is that the instrumental variable studies estimate a "local average treatment effect" for a non-representative sample of the population (Imbens and Angrist 1994). de Walque (2007b) and Grimard and Parent (2007) estimate the effect of education for those who would not have gone to college but for the Vietnam draft. Kenkel et al.'s (2006) estimates correspond to the group of individuals just on the margin of obtaining a high school (or GED) degree. Currie and Moretti (2003) estimate the effect of education for those who continue their education only if a college is located in the same county.

The literature recognizes that the effect of education for these groups is unlikely to represent the effect for the general population. For example, Grimard and Parent (2007) acknowledge that their instrumental variable results are too large to be credible estimates for the

general population (see pgs. 912-916). If applied to the general population their estimates imply that virtually all those with some college education would have become smokers had they not started college. Furthermore, their estimates imply that none of those with a high school education would have become smokers had they obtained further schooling. Neither of these two counterfactuals is plausible. While studies that employ policy instruments are useful for evaluating potential policy reforms that would affect a similar group of individuals (Card 2001), their results are likely not informative of the effect of education for the population at large.

In contrast, our results are representative of the effect of education during the primary years when individuals make their decision to become a smoker.[33] The results indicate that the average treatment effect is close to zero, casting doubt on the applicability of the causal theories detailed in Section II. Of course, our analysis does not exclude the possibility that education might have a causal effect for a subset of the population. As such, our results are not necessarily inconsistent with the findings of prior studies.

It is important to note that we estimate the effect of education for a recent generation, those born between 1974 and 1986. In contrast, Grimard and Parent (2007) identify the effect of education from males born between 1945 and 1950. de Walque (2007b) uses a different measure of induction risk that includes males born between 1937 and 1956. Kenkel et al. (2006) uses a data sample of those born between 1957 and 1964. The data sample employed by Currie and Morretti (2003) consists of women born between 1925 and 1975.

Information regarding the negative health effects of smoking did not become widespread until the 1950's and 1960's, culminating in the issuance of the first Surgeon General's Report on Smoking and Health in 1964 (Grossman 2006). For earlier generations it seems more likely that education played a meaningful role in spreading information about the consequences of smoking, particularly for the less educated. Knowledge of the health effects of smoking is widespread by the period of our data, 1998-2003, potentially limiting the informative value of education. We

[33] As noted earlier, our results apply only to those individuals who do not drop out of the CPS panel.

analyze the effect of education for those aged 16 to 24, which corresponds to when individuals are in high school or college. Higher education may have little impact on smoking if the current generation became aware of its negative consequences at an earlier age. It is possible that education has a larger effect on younger individuals in elementary and middle school who are less informed.

VIII. Conclusion

We explore whether an additional year of education affects an individual's propensity to smoke. In contrast to previous research, we do not find that education has a significant impact on smoking. The control group methodology that we employ compares those who will acquire a given level of education in the following year to those who are one year older and currently have that particular level of education. This framework allows us to difference out the impact of unobserved characteristics, isolating the causal effect of education. The results indicate that unobserved characteristics correlated with education entirely explain the cross-sectional relationship between education and smoking behavior.

We do find, however, that being a student reduces the propensity to smoke. The impact of being a high school or college student is similar in magnitude. Since students are less likely to smoke than non-students, we hypothesize that this finding is due to peer effects. This interpretation of our results corroborates prior research that concludes peer effects have a significant impact on smoking (Norton et al. 1998, Gaviria and Raphael 2001, Powell et al. 2005).

This paper evaluates the smoking behavior of those aged 16 to 24. The majority of those who ever decide to smoke do so during the age range of our analysis. While we find little evidence that high school and college education affect smoking behavior, future research is needed to explore whether education has a significant impact earlier in life (such as in elementary and middle school). Similarly, although education is unlikely to have a major impact

on the decision to become a smoker for older individuals (since very few start smoking after age 24), whether education affects the decision to quit smoking is an area for further analysis.[34]

Research on the causal effect of education on health is still relatively small. Although progress has been made using instrumental variables, our analysis highlights the need for exploring whether similar results can be obtained using alternative estimation strategies that measure the effect of education for different segments of the population and for different margins of educational attainment.

References

Adams, Scott J. 2002. "Educational Attainment and Health: Evidence from a Sample of Older Adults", *Education Economics* 10(1): 97-109.

Angrist, Joshua D. and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives* 15(4): 69-85.

Arendt, Jacob N. 2005. "Does Education Cause Better Health? A Panel Data Analysis Using School Reforms for Identification," *Economics of Education Review* 24(2):149-60.

Becker, Gary S. 1983. "A Theory of Competition among Pressure Groups for Political Influence," *Quarterly Journal of Economics* 98(3):371-400.

Becker, Gary S. and Casey B. Mulligan. 1997. "The Endogenous Determination of Time Preference," *Quarterly Journal of Economics* 112(3):729-58.

Berger, Mark C. and J. Paul Leigh. 1989. "Schooling, Self-Selection, and Health," *Journal of Human Resources* 24(3):433-55.

Borjas, George J. 1985. "Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants," *Journal of Labor Economics* 3(4):463-89.

Bound, John and David A. Jaeger. 1996. "On the Validity of Season of Birth as an Instrument in Wage Equations: A Comment on Angrist & Krueger's 'Does Compulsory School Attendance Affect Schooling and Earnings,'" NBER Working Paper 5835.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association* 90(430):443-50.

Card, David. 2001. "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica* 69(5):1127-60.

[34] Previous research that looks at smoking cessation obtains mixed results (Walque 2007b, Grimard and Parent 2007).

Cawley, John, James Heckman, and Edward Vytlacil. 1998. "Cognitive Ability and the Rising Return to Education," NBER Working Paper 6388.

Chaloupka, Frank J. and Kenneth E. Warner. 2000. "The Economics of Smoking," in *Handbook of Health Economics*, vol. 1B, Joseph P. Newhouse and Anthony J. Culyer, eds. Amsterdam: North-Holland.

Currie, Janet and Enrico Moretti. 2003. "Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings," *Quarterly Journal of Economics* 118(4):1495-1532.

de Walque, Damien. 2007a. "How Does the Impact of an HIV/AIDS Information Campaign Vary with Educational Attainment? Evidence from Rural Uganda." *Journal of Development Economics* 84(2):686-714.

de Walque, Damien. 2007b. "Does Education Affect Smoking Behaviors? Evidence using the Vietnam Draft as an Instrument for College Education," *Journal of Health Economics* 26(5):877-95.

DeCicca, Philip, Donald Kenkel, and Alan Mathios. 2002. "Putting Out the Fires: Will Higher Taxes Reduce the Onset of Youth Smoking?," *Journal of Political Economy* 110(1):144-69.

Farrell, Phillip and Victor R. Fuchs. 1982. "Schooling and Health: The Cigarette Connection," *Journal of Health Economics* 1(3): 217-30.

Gaviria, Alejandro and Steven Raphael. 2001. "School-based Peer Effects and Juvenile Behavior," *Review of Economics and Statistics* 83(2):257-68.

Grimard, Franque and Daniel Parent. 2007. "Education and Smoking: Were Vietnam Draft Avoiders Also More Likely to Avoid Smoking?," *Journal of Health Economics* 26(5):896-926.

Grossman, Michael. 1972. "On the Concept of Health Capital and the Demand for Health," *Journal of Political Economy* 80(2):223-55.

Grossman, Michael. 2006, "Education and Nonmarket Outcomes," in *Handbook of the Economics of Education*, vol. 2, Eric Hanushek and Finis Welch, eds. Amsterdam: North-Holland.

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

Heckman, James J. 1996. "Identification of Causal Effects Using Instrumental Variables: Comment," *Journal of the American Statistical Association* 91(434):459-62.

Imbens, Guido W. and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62(2):467-75.

Kenkel, Donald S. 1991. "Health Behavior, Health Knowledge, and Schooling," *Journal of Political Economy* 99(2):287-305.

Kenkel, Donald, Dean Lillard, and Alan Mathios. 2006. "The Roles of High School Completion and GED Receipt in Smoking and Obesity," *Journal of Labor Economics* 24(3):635-660.

Lancaster, Tony. 2000. "The Incidental Parameter Problem since 1948," *Journal of Econometrics* 95(2): 391-413.

Leigh, J. Paul and Rachna Dhir. 1997. "Schooling and Frailty Among Seniors", *Economics of Education Review* 16(1):45-57.

Lleras-Muney, Adriana. 2005. "The Relationship between Education and Adult Mortality in the United States," *Review of Economic Studies* 72(1):189-221.

Meyer, Bruce D. 1995. "Natural and Quasi-Experiments in Economics," *Journal of Business and Economic Statistics* 13(2):151-61.

Madrian, Brigitte C. and Lars J. Lefgren. 2000. "An Approach to Longitudinally Matching Current Population Survey (CPS) Respondents," *Journal of Economic and Social Measurement* 26(1):31-62.

Mokdad, Ali H., James S. Marks, Donna F. Stroup, and Julie L. Gerberding. 2004. "Actual Causes of Death in the United States," *JAMA* 291(10):1238-45.

Neumark, David and Daiji Kawaguchi. 2004. "Attrition Bias in Labor Economics Research using Matched CPS Files," *Journal of Economic and Social Measurement* 29(4):445–72.

Norton, Edward C., Richard C. Lindrooth, and Susan T. Ennett. 1998. "Controlling for the Endogeneity of Peer Substance Use on Adolescent Alcohol and Tobacco Use," *Health Economics* 7(5):439-53.

Peltzman, Sam. 1976. "Toward a More General Theory of Regulation," *Journal of Law and Economics* 19(2):211-40.

Peracchi, Franco and Finis Welch. 1995. "How Representative are Matched Cross-Sections? Evidence from the Current Population Survey," *Journal of Econometrics* 68(1):153-79.

Powell, Lisa M., John A. Tauras, and Hana Ross. 2005. "The Importance of Peer Effects, Cigarette Prices and Tobacco Control Policies for Youth Smoking Behavior," *Journal of Health Economics* 24(5):950-68.

Rosenzweig, Mark R. 1995. "Why are there Returns to Schooling?" *American Economic Review* 85(2):153-58.

Sander, William. 1995. "Schooling and Quitting Smoking," *Review of Economics and Statistics* 77(1):191-99.

Tenn, Steven. 2007. "The Effect of Education on Voter Turnout," *Political Analysis* 15(4):446-64.

U.S. Department of Health and Human Services. 1989. *Reducing the Health Consequences of Smoking: 25 Years of Progress. A Report of the Surgeon General.* Atlanta: U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. DHHS Publication No. (CDC) 89-8411.

U.S. Department of Health and Human Services. 1998. *Tobacco Use Among U.S. Racial/Ethnic Minority Groups—African Americans, American Indians and Alaska Natives, Asian Americans and Pacific Islanders, and Hispanics: A Report of the Surgeon General*. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

U.S. Department of Health and Human Services. 2001. *Women and Smoking. A Report of the Surgeon General*. Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, Office of the Surgeon General.

U.S. Department of Health and Human Services. 2004. *The health consequences of smoking: a report of the Surgeon General*. Washington, DC. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

Table 1: Summary Statistics

Variable	Mean	Std Dev
Smoker, Ever	18.9%	39.1%
Smoker, Currently	14.6%	35.3%
Smoker, Everyday	10.9%	31.2%
Age (in years)	19.5	2.4
Education, <=8th grade	2.0%	14.1%
Education, 9th grade	5.1%	22.0%
Education, 10th grade	12.8%	33.5%
Education, 11th grade	16.8%	37.4%
Education, 12th grade	26.9%	44.3%
Education, Some College	32.3%	46.7%
Education, College Degree	4.0%	19.7%
Student	63.5%	48.1%
Female	49.0%	50.0%
White	66.3%	47.3%
Black	14.0%	34.7%
Hispanic	14.4%	35.1%
Multiple Races	0.5%	6.9%
Other Races	4.9%	21.6%
Married	7.9%	27.0%
Born in the U.S.	89.6%	30.6%
Veteran	0.4%	6.5%
Live in an MSA	81.4%	38.9%
Live with a Parent	78.7%	41.0%
Number of Observations	41,882	

Notes: Current Population Survey Tobacco Supplements, 1998-2003. The "multiple races" category is available only in 2003. Age is measured in whole years. All other variables are binary. The number of observations for "current smoker" and "everyday smoker" is 41,803 due to missing data for some individuals.

Table 2: Effect of Education on Smoking Status when Education is Treated as an Exogenous Variable

	Ever Smoke (N=41,882)			Currently Smoke (N=41,803)			Smoke Everyday (N=41,803)		
	Est	SE		Est	SE		Est	SE	
Education, <=8th grade	23.5%	4.0%	*	22.9%	4.2%	*	19.4%	3.6%	*
Education, 9th grade	28.6%	2.8%	*	27.3%	2.9%	*	24.2%	2.7%	*
Education, 10th grade	27.6%	2.4%	*	26.2%	2.5%	*	23.3%	2.3%	*
Education, 11th grade	24.8%	1.8%	*	23.7%	1.8%	*	20.8%	1.9%	*
Education, 12th grade	20.5%	1.4%	*	19.3%	1.5%	*	16.6%	1.4%	*
Education, Some College	15.4%	1.4%	*	13.8%	1.2%	*	11.5%	1.0%	*
Student	-14.0%	0.9%	*	-12.6%	0.8%	*	-11.3%	0.8%	*
Female	-2.5%	0.6%	*	-2.4%	0.6%	*	-1.1%	0.5%	*
Black	-14.4%	0.8%	*	-10.6%	0.8%	*	-8.7%	0.8%	*
Hispanic	-12.4%	0.5%	*	-10.3%	0.5%	*	-9.3%	0.8%	*
Multiple Races	2.0%	2.4%		3.0%	2.4%		0.3%	1.9%	
Other Races	-4.2%	1.0%	*	-2.3%	0.9%	*	-2.5%	0.6%	*
Married	-7.3%	1.8%	*	-8.3%	1.3%	*	-5.5%	1.2%	*
Born in the U.S.	8.3%	1.4%	*	6.4%	1.3%	*	5.5%	1.1%	*
Veteran	2.6%	4.4%		2.6%	4.5%		-0.5%	3.9%	
Live in an MSA	1.3%	0.7%		0.5%	0.6%		0.8%	0.6%	
Live with a Parent	-11.0%	1.0%	*	-7.6%	0.8%	*	-5.7%	0.8%	*

Notes: The model also includes age × year and state of residence fixed effects. Statistical significance corresponds to *=5%. Robust standard errors are reported that cluster by state of residence. The omitted education category corresponds to those with a college degree. The omitted race category corresponds to whites. The sample size varies depending on the smoking measure employed due to missing observations.

Table 3: Effect of Education on Smoking Status

A. Ever Smoke (N=41,882)

	(i)		(ii)		(iii)		(iv)	
	Est	SE	Est	SE	Est	SE	Est	SE
Education	-0.5%	1.2%	-0.4%	1.3%				
Student	-5.2%	1.2% *	-4.9%	1.2% *				
Education, High School					-1.5%	2.1%	-1.6%	2.2%
Education, College					-0.4%	1.3%	-0.3%	1.2%
Student, High School					-5.5%	1.2% *	-5.1%	1.2% *
Student, College					-5.1%	1.3% *	-4.9%	1.3% *
Additional controls?		N		Y		N		Y

B. Currently Smoke (N=41,803)

	(i)		(ii)		(iii)		(iv)	
	Est	SE	Est	SE	Est	SE	Est	SE
Education	-0.7%	1.1%	-0.7%	1.1%				
Student	-3.9%	1.1% *	-3.7%	1.1% *				
Education, High School					-0.6%	2.2%	-0.8%	2.2%
Education, College					-0.8%	1.1%	-0.7%	1.1%
Student, High School					-4.8%	1.0% *	-4.4%	1.1% *
Student, College					-3.6%	1.2% *	-3.5%	1.2% *
Additional controls?		N		Y		N		Y

C. Smoke Everyday (N=41,803)

	(i)		(ii)		(iii)		(iv)	
	Est	SE	Est	SE	Est	SE	Est	SE
Education	-0.2%	0.9%	-0.2%	0.9%				
Student	-3.0%	1.0% *	-2.8%	1.0% *				
Education, High School					-1.1%	1.5%	-1.3%	1.5%
Education, College					-0.1%	0.9%	-0.1%	0.9%
Student, High School					-3.4%	1.1% *	-3.1%	1.2% *
Student, College					-3.0%	1.0% *	-2.8%	1.0% *
Additional controls?		N		Y		N		Y

Notes: The model controls for age, education, student status, and a set of fixed effects that accounts for selection bias in education choice (see Section IV). Specification (ii) and (iv) contain additional controls for gender, race/ethnicity, marital status, native born, veteran status, living in a metropolitan statistical area, living with a parent, and a set of fixed effects for state of residence. Statistical significance corresponds to *=5%. Robust standard errors are reported that cluster by state of residence. The sample size varies depending on the smoking measure employed due to missing observations.

Table 4: Effect of Education on Smoking Status, Measurement Error Sensitivity Analysis

A. Ever Smoke

	Education (i) Est	(i) SE	Education (ii) Est	(ii) SE
Baseline model without exclusions (N=41,882)	-0.5%	1.2%	-0.4%	1.3%
Exclude observations with:				
(a) Change in student status between survey years (N=35,029)	-0.1%	1.3%	0.0%	1.3%
(b) Education in previous year not in adjacent education level (N=39,188)	-0.5%	1.3%	-0.4%	1.2%
(c) Either (a) or (b) (N=32,661)	-0.3%	1.3%	-0.2%	1.3%
Additional controls?	N		Y	

B. Currently Smoke

	Education (i) Est	(i) SE	Education (ii) Est	(ii) SE
Baseline model without exclusions (N=41,803)	-0.7%	1.1%	-0.7%	1.1%
Exclude observations with:				
(a) Change in student status between survey years (N=34,967)	0.0%	1.2%	0.1%	1.2%
(b) Education in previous year not in adjacent education level (N=39,116)	-0.6%	1.1%	-0.6%	1.1%
(c) Either (a) or (b) (N=32,604)	-0.1%	1.2%	-0.1%	1.2%
Additional controls?	N		Y	

C. Smoke Everyday

	Education (i) Est	(i) SE	Education (ii) Est	(ii) SE
Baseline model without exclusions (N=41,803)	-0.2%	0.9%	-0.2%	0.9%
Exclude observations with:				
(a) Change in student status between survey years (N=34,967)	0.5%	1.0%	0.5%	1.0%
(b) Education in previous year not in adjacent education level (N=39,116)	-0.2%	0.9%	-0.1%	0.9%
(c) Either (a) or (b) (N=32,604)	0.4%	1.0%	0.3%	1.0%
Additional controls?	N		Y	

Notes: The model controls for age, education, student status, and a set of fixed effects that accounts for selection bias in education choice (see Section IV). The baseline model corresponds to the results reported earlier in Table 3. Specification (ii) contains additional controls for gender, race/ethnicity, marital status, native born, veteran status, living in a metropolitan statistical area, living with a parent, and a set of fixed effects for state of residence. Statistical significance corresponds to *=5%. Robust standard errors are reported that cluster by state of residence. The sample size varies depending on the smoking measure employed due to missing observations.

Table 5: Effect of Education on Smoking Status, Fixed Effects for Unobserved Characteristics Pooled across Survey Years

A. Ever Smoke (N=41,882)

	(i)			(ii)		
	Est	SE		Est	SE	
Education	-0.2%	1.2%		-0.2%	1.2%	
Student	-4.5%	0.9%	*	-4.4%	0.9%	*
Additional controls?		N			Y	

B. Currently Smoke (N=41,803)

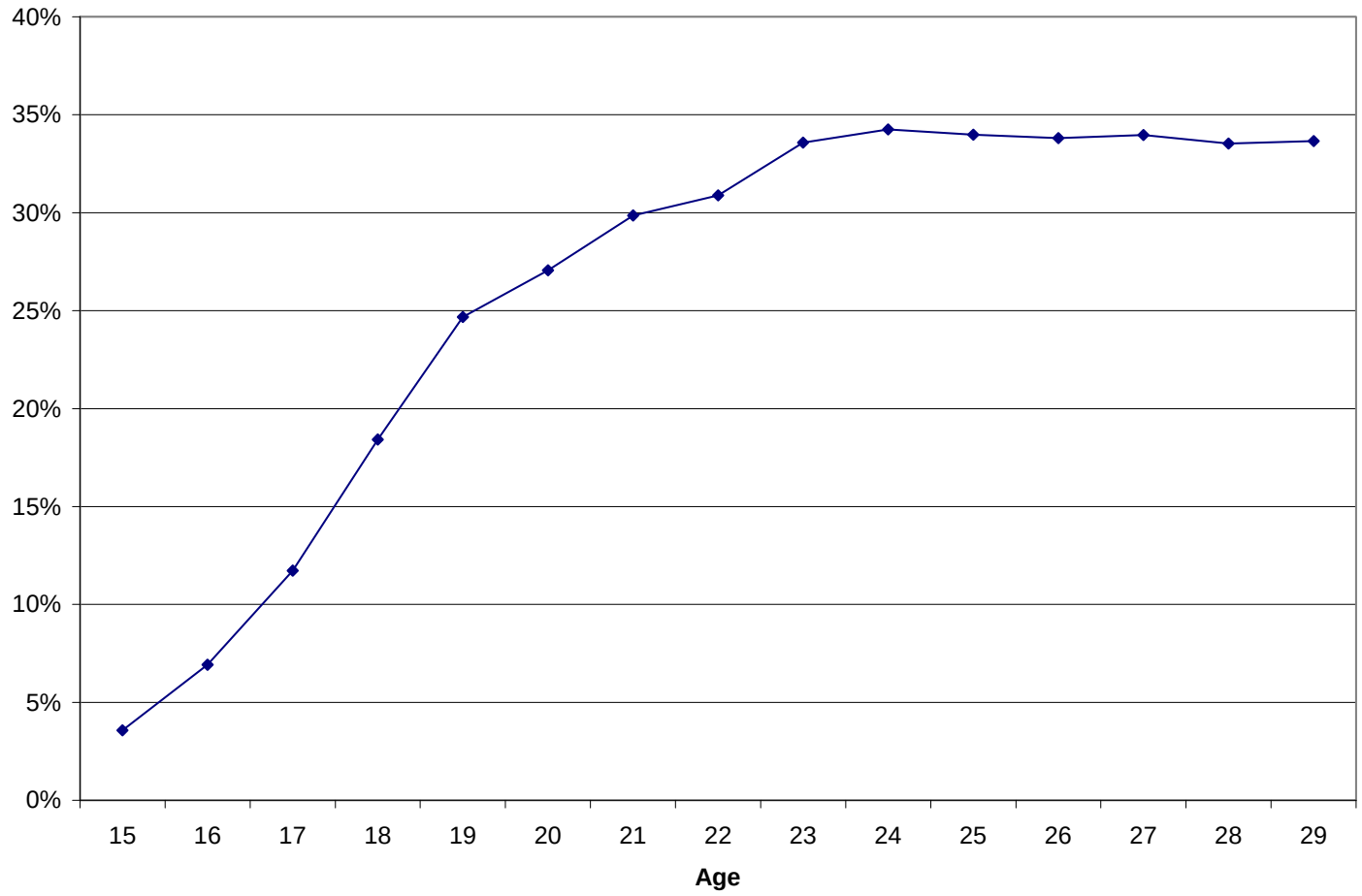
	(i)			(ii)		
	Est	SE		Est	SE	
Education	-0.6%	1.1%		-0.7%	1.1%	
Student	-3.1%	1.0%	*	-3.1%	1.0%	*
Additional controls?		N			Y	

C. Smoke Everyday (N=41,803)

	(i)			(ii)		
	Est	SE		Est	SE	
Education	-0.1%	0.9%		-0.1%	0.9%	
Student	-2.7%	0.9%	*	-2.6%	0.9%	*
Additional controls?		N			Y	

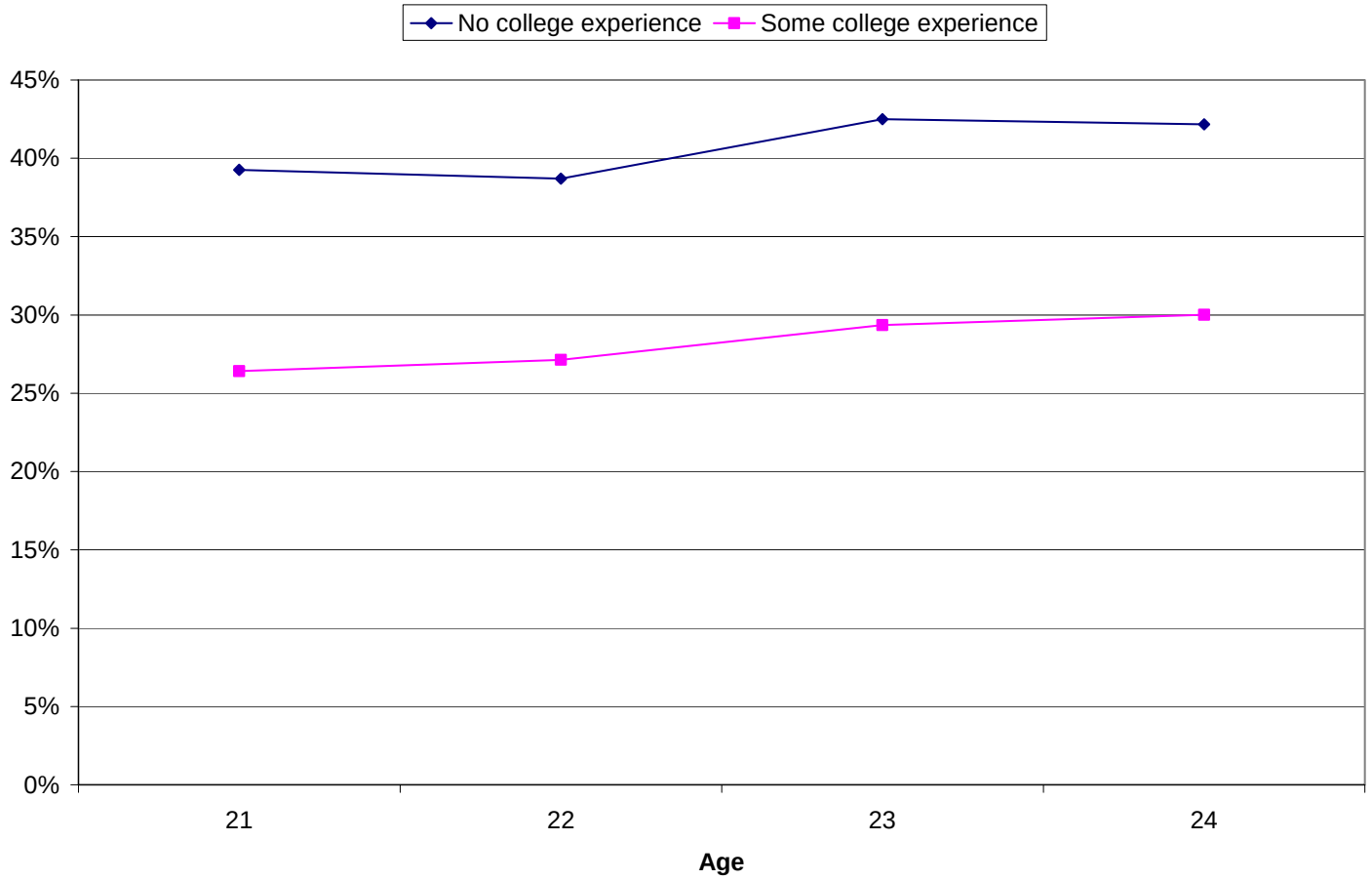
Notes: The model controls for age, education, student status, a set of fixed effects that accounts for selection bias in education choice (see Section IV), and a set of dummy variables for survey year. Specification (ii) contains additional controls for gender, race/ethnicity, marital status, native born, veteran status, living in a metropolitan statistical area, living with a parent, and a set of fixed effects for state of residence. Statistical significance corresponds to *=5%. Robust standard errors are reported that cluster by state of residence. The sample size varies depending on the smoking measure employed due to missing observations.

Figure 1: Percentage of Ever Smokers by Age



Notes: The Current Population Survey does not report smoking behavior prior to age 15.

Figure 2: Percentage of Ever Smokers among High School Graduates by Age and College Experience



Notes: For every age between 21 and 24, 34% of high school graduates have not started college.