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

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

* 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

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*

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

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

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 $\quad_k$ 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]

$$(3.1) \qquad \Pr(y_{it} \quad 1) \quad \sum_{k \ 1}^{6} \ _k 1_{group_i \quad k} \quad _a a_{it} \quad _e e_{it} \quad _s s_{it}$$

The problem with specifying the model in this way is that parameters $\quad_a$, $\quad_e$, and $\quad_s$ 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: $\quad_1 \quad \quad_4$, $\quad_2 \quad \quad_5$, and $\quad_3 \quad \quad_6$. 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) $\bar{y}^1 \quad _1 \quad 17 \; _a \quad 10 \; _e$

(3.3) $\bar{y}^2 \quad _2 \quad 17 \; _a \quad 11 \; _e \quad _s$

(3.4) $\bar{y}^3 \quad _3 \quad 17 \; _a \quad 11 \; _e$

(3.5) $\bar{y}^4 \quad _1 \quad 16 \; _a \quad 10 \; _e$

(3.6) $\bar{y}^5 \quad _2 \quad 16 \; _a \quad 10 \; _e \quad _s$

(3.7) $\bar{y}^6 \quad _3 \quad 16 \; _a \quad 10 \; _e \quad _s$

Subtracting equation (3.5) from equation (3.2) yields $\hat{}_a \quad \bar{y}^1 \quad \bar{y}^4$. 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 $\hat{}_a \quad \hat{}_e \quad \bar{y}^2 \quad \bar{y}^5$. 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 $\hat{}_a$ yields the following "difference in difference" estimator: $\hat{}_e \quad (\bar{y}^2 \quad \bar{y}^5) \quad (\bar{y}^1 \quad \bar{y}^4)$. Finally, the effect of being a student is obtained by comparing groups 3 and 6. Subtracting equation (3.7) from equation (3.4) gives $\hat{}_a \quad \hat{}_e \quad \hat{}_s \quad \bar{y}^3 \quad \bar{y}^6$, which simplifies to $\hat{}_s \quad (\bar{y}^2 \quad \bar{y}^5) \quad (\bar{y}^3 \quad \bar{y}^6)$ after substituting for $\hat{}_a$ and $\hat{}_e$. 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.

obtain consistent estimates of the effect of education. A problem arises when only a subset of the characteristics contained in $_{it}$ 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 $_{it}$ is completely unobserved. Later we consider the situation where some, but not all, of the characteristics contained in $_{it}$ are observable.

The effect of age, education, and student status is separated from all other characteristics $_{it}$ 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 $_{it}$ 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 $_{it}$, 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 $_{it}$ 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 pap(a)5(ci(, Gav on)-4(edu)-4(catio)-4icuTJ 0.0038 Tm (it)Tjdtw 0o)1(d5bg53c(p)-5(a)5(p)cu8rdot003 Tw -4.up4ica0.00

The following assumption allows us to identify the causal effect of education for those who

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, $g_i \quad g_{i,t\ 1}$ and $g_i' \quad g_{it}$. For the remaining individuals who can be matched to the following year, $g_i \quad g_{it}$ and $g_i' \quad g_{i,t\ 1}$. 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) $\Pr(y_{it} \quad 1 \mid a_{it}, e_{it}, s_{it}, g_i, g_i') \quad {}_a a_{it} \quad {}_e e_{it} \quad {}_s s_{it} \quad E(\ {}_{it} \mid g_{i,t\ 1} \quad g_i, g_{it} \quad g_i')$

To clarify how the causal effect of education is identified, define ${}^*(g, g', t) \quad {}_a a(g') \quad {}_e e(g') \quad {}_s s(g') \quad E(\ {}_{it} \mid g_{i,t\ 1} \quad g, g_{it} \quad g')$, 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) $\Pr(y_{it} \quad 1 \mid a_{it}, e_{it}, s_{it}, g_i, g_i') \quad {}_a(a_{it} \quad a(g_i')) \quad {}_e(e_{it} \quad e(g_i')) \quad {}_s(s_{it} \quad s(g_i')) \quad {}^*(g_i, g_i', t)$

For individuals who can be maess ma

The remaining parameters $\{\ _a,\ _e,\ _s\}$, which reflect the effect of age, education and student status, are identified from changes in these variables between survey years for

(4.7) $E(\ _{it} \mid g_{it} \quad g, g_{i,t\ 1} \quad g', m_{it} \quad 1) \quad E(\ _{it} \mid g_{i,t\ 1} \quad g, g_{it} \quad g', m_{i,t\ 1} \quad 1) \quad d(a(g),t), \quad g, g'$

This assumption is substantially weaker than equation (4.4), which is equivalent to assuming $d(a,t) \quad 0, \quad a,t$. 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 $_{a}(a,t) \quad _{a} \quad d(a,t)$.

(4.8) $\Pr(y_{it} \quad 1 \mid a_{it}, e_{it}, s_{it}, g_i, g_i') \quad _{a}(a_{it}, t)(a_{it} \quad a(g_i')) \quad _{e}(e_{it} \quad e(g_i'))$
$\quad _{s}(s_{it} \quad s(g_i')) \quad ^{*}(g_i, g_i', t)$

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 $a_{it} \quad a(g_i')$ 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 $_{a}(a,t)$ is estimated via an interaction between $a_{it} \quad a(g_i')$ 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 $_{a}(a,t)$ 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

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 $_{it}$. By matching individuals one year apart in their life cycle, the control group methodology described above differences out the effect of $_{it}$. 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 abso

Equation (4.9) contains two additional terms that are omitted from equation (4.8). The first, X_{it} , 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 markmrstiifica W theh4.pr f Tinfa theiliaf. . Thly p onlemasedalysieholu5ely

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

$e_{i,t\ 1}$ 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

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 $s_{it} \quad s_{i,t\ 1}$, 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

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

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

Table 2 presents estimates of the effect of

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

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 year

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

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 pr

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

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.

on the decision to become a smoker for older indi

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

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		

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

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

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%
(%ea501(-)-7(0)5(42%)-650(0)-1(.9)5(%)]TJ0.3018 Tc 0.0013 Tw -34.991 educatlevel (N=39,188)	-0.5%	1.3%	-0.4%	1.2

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)

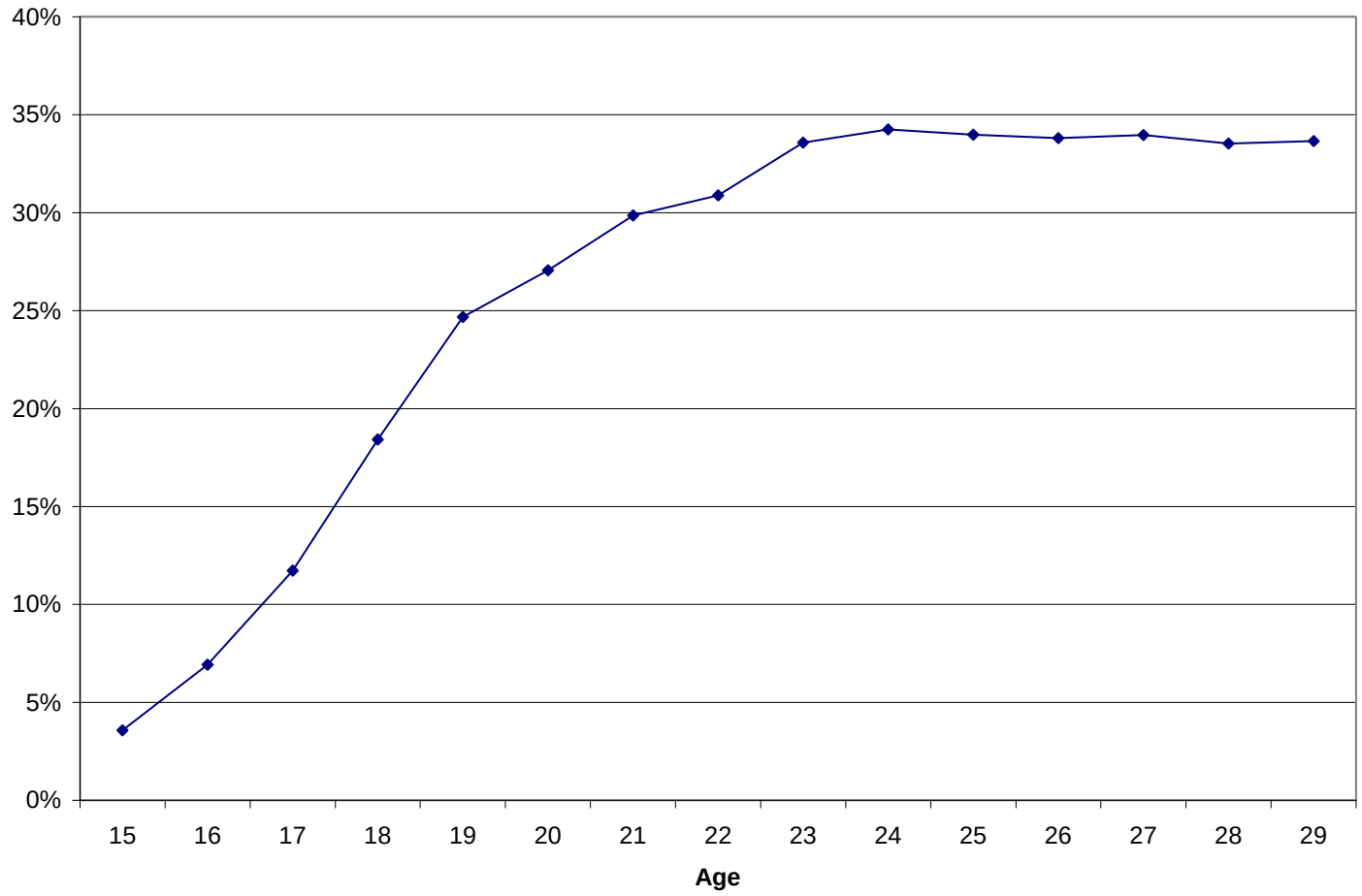
	(i)			(ii)		
	Est	SE		Est	SE	
Education	-0.6%	1.1%		-0.7%	1.1%	
Student	-3.1%	1.0%	*	-3.1%	1.0%	*
Additional controls?		N			Y	

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	

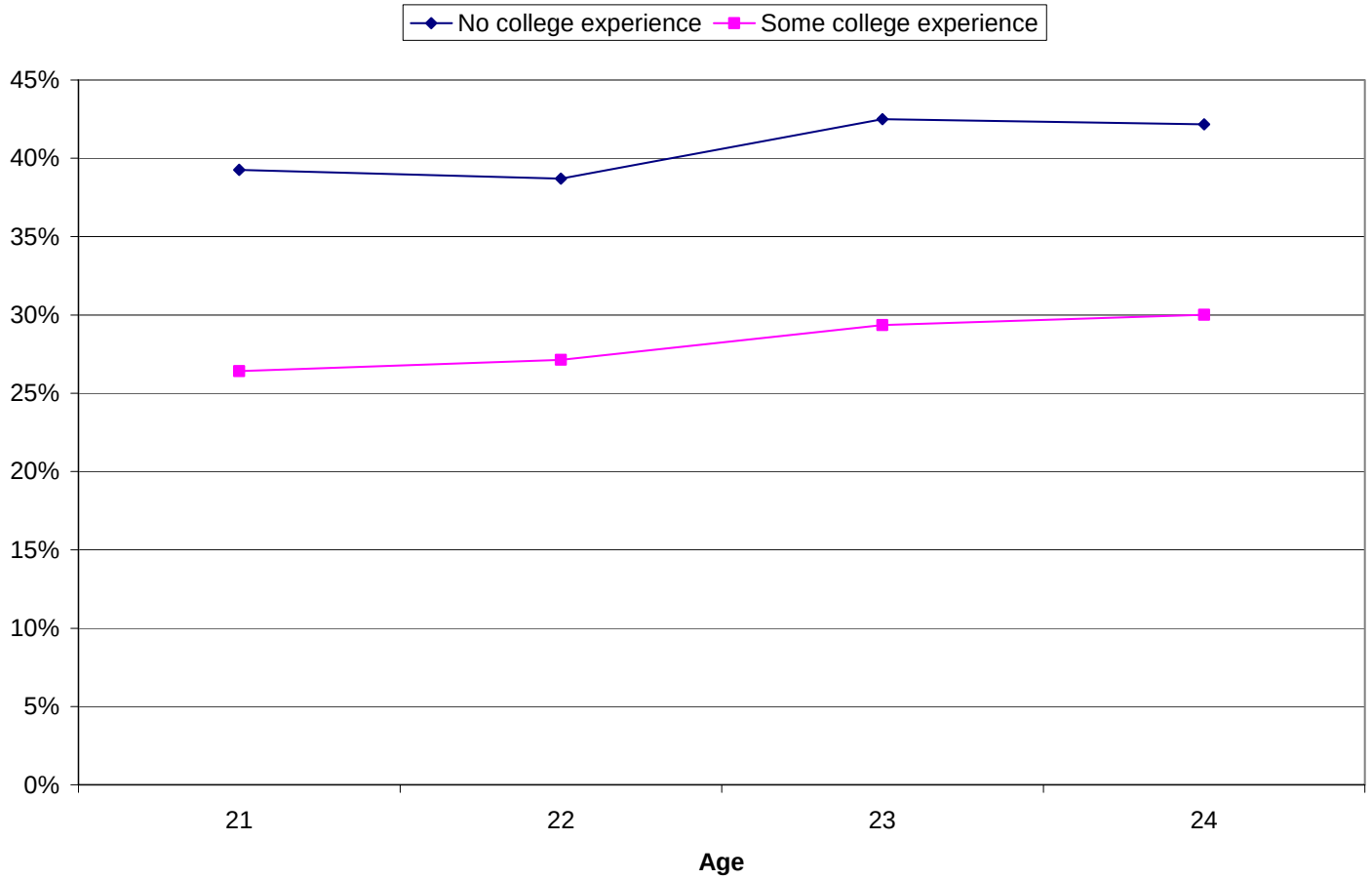
Notes: The model controls for age, education, student status, a set of fixed effects that accounts for selection bias in education choice (see Section IV), and a set of dummy variables for survey year. Specification (ii) contains additional controls for

Figure 1: Percentage of Ever Smokers by Age



Notes

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.