J Popul Econ (2004) 17:351–367
DOI 10.1007/s00148-003-0158-x

Journal of
Population
Economics

Peer effects on substance use among American teenagers

Daiji Kawaguchi

Institute of Social and Economic Research, Osaka University, 6-1 Mihogaoka, Ibaraki, Osaka 567-0047 Japan (Fax: +81-6-6878-2766; e-mail: kawaguch@iser.osaka-u.ac.jp)

Received: 3 June 2002/Accepted: 20 February 2003

Abstract. The widespread use of illicit substances by American teenagers has attracted the interest of both the general public and academic researchers. Among the various factors that people believe influence youth substance use, peer effects are identified as a critical determinant. Identifying peer effects, however, is known to be a difficult task. In an attempt to overcome known difficulties, I estimate peer effects on substance usage among American teenagers using perceived peer behavior in the *National Longitudinal Survey Youth 97*. The data indicate robust peer effects. Moreover, the results do not change substantially in school and household fixed effects estimations.

JEL classification: C13, I12, Z13

Key words: Subjective perception, reflection problem, youth behavior

1. Introduction

The widespread use of illicit substances by American teenagers attracts both public attention and research interest. The changing percentages of substance

This is a chapter of my Ph.D. dissertation submitted to Michigan State University. I am especially grateful to David Neumark for his constructive comments and continuous encouragement. I thank Jeff Biddle, Ali Berker, Frederic Durousseau, Gigi Foster, Daniel Hamermesh, John Strauss, Jeffrey Wooldridge, the participants in the labor economics workshop at Michigan State University and 2001 Joint Statistical Meeting, and an anonymous referee for helpful comments. Scott Adams and David Wetzell gave me detailed comments, as well as advice on English. Donna Maurer provided careful editorial assistance. All remaining errors are, of course, my own. Due to a confidentiality agreement with the BLS on the usage of the NLSY 97 GeoCode file, the part of the dataset that is constructed from the GeoCode file cannot be released. *Responsible editor*: Daniel S. Hamermesh

users during the 1990s are plotted in Fig. 1. Although the percentage of alcohol users dropped in the early 1990s, it still remained high. This figure shows the steady trend of cigarette users at a high level.[1] It is also notable that the percentage of marijuana users increased from 4.4% to 9.7% between 1990 and 1997.

These figures have sparked much public interest about the reasons why teenagers use substances and what policy makers can do to reduce this usage. Besides the price of substances, peer effects or peer pressure is identified as a critical determinant, since the use of substances is considered to be a highly social behavior.[2]

Reacting to this interest, economists and sociologists have tried to estimate the existence and strength of peer effects. Identifying peer effects is not easy, since an observed behavior shared by a teenager and his/her peer may result from unobserved factors that group members share, rather than from peer effects. In addition, identifying peer effects becomes complex when the average reference group's outcome is used as a measurement of peers' behavior. Determining whether a teenager's behavior affects his/her peers or vice versa is difficult. Manski (1993) articulated this as the "reflection problem." In addition, both current substance users and the backgrounds of group members may affect individual behaviors. Although both effects are called peer effects, each has different policy implications. Distinguishing between these two effects, however, is known to be difficult (Manski 1993).

This paper employs a critically different strategy for identifying peer effects; I identify peer effects by using teenagers' subjective perceptions. Manski (1993) wrote:

> Given that identification based on observed behavior alone is so tenuous, experimental and subjective data will have to play an important role in future efforts to learn about social effects.

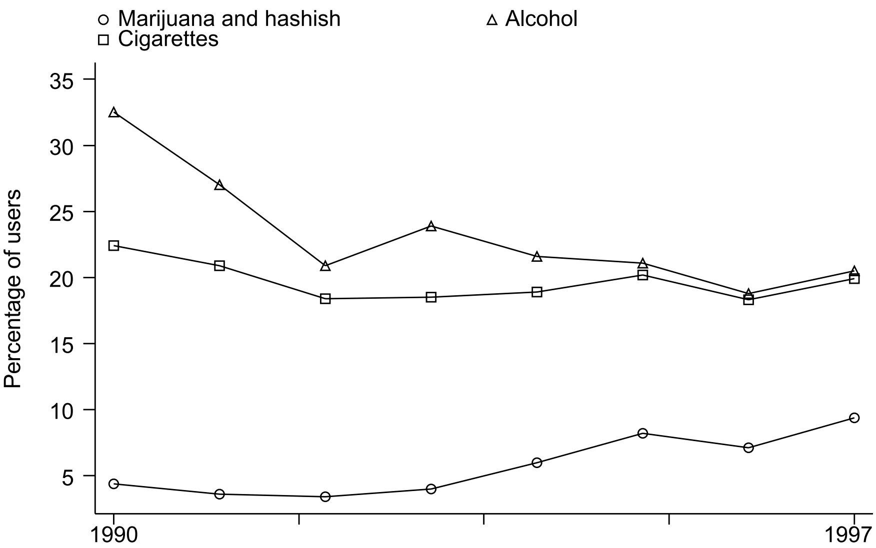

Fig. 1. Youth substance use, age 12–17, current users. (*Source*: National Household survey on Drug Abuse, annual.)

Using teenagers' subjective perceptions of peer behaviors, identifying peer effects is free from the problems that arise with using an average outcome as a peer variable. According to Manski (2000), this approach has not been taken seriously since he originally suggested it in 1993.[3]

In addition, I employ school and household fixed effects estimations to ensure the robustness of my results.

2. The reflection problem

Manski (1993) articulated several issues concerning the identification of peer effects using an average group outcome as a peer variable. A short sketch of the problem follows. Let y be an outcome of interest, x be an index of an individual's reference group, and z be attributes that affect the outcome directly. The outcome is characterized by

$$y = \alpha + \beta E(y|x) + E(z|x)\gamma + z\eta + u, \;\; E(u|x,z) = x\delta. \qquad (1)$$

If $\beta \neq 0$, then an individual behavior is affected by the mean of the group outcome, $E(y|x)$. This is the "endogenous effect." If $\gamma \neq 0$, an individual behavior is affected by the group mean of the exogenous variable (background of group members); this is the "contextual effect." If $\delta \neq 0$, the model exhibits the "correlated effect." People in the same group behave similarly because their shared group characteristics are correlated with unobservable factors, such as social institutions. The correlated effect can also arise from the formation of groups based on individuals unobserved characteristics. Individuals are likely to sort themselves into a group in which group members share similar unobserved characteristics. This phenomenon is often called "endogenous sorting."

Taking a conditional expectation on z and x, the model becomes

$$E(y|x,z) = \alpha + \beta E(y|x) + E(z|x)\gamma + z\eta + x\delta. \qquad (2)$$

To discuss the identification of parameters, we need to solve for the conditional expectation of y in terms of x and z. Using the iterated law of expectations,

$$E(y|x) = \alpha + \beta E(y|x) + E(z|x)\gamma + E(z|x)\eta + x\delta. \qquad (3)$$

Solving the expression for $E(y|x)$ and substituting it into (2), we obtain

$$E(y|x,z) = \alpha/(1-\beta) + E(z|x)[(\gamma + \beta\eta)/(1-\beta)] + x\delta/(1-\beta) + z\eta. \qquad (4)$$

The composite parameters are identified if $[1, E(z|x), x, z]$ are linearly independent. This linear independence assumption fails when one of the following occurs.

1. The individual attributes, z, are a function of a group index, x, since $E(z(x)|x) = z(x)$. Consequently $E(z|x)$ and z are linearly dependent. This case occurs when group members share the same exogenous variables.
2. z does not vary with x. Since $E(z|x)$ is constant, $E(z|x)$ and z are linearly dependent. This case occurs when average exogenous variables are identical across groups.
3. $E(z|x)$ is a linear function of x since $E(z|x) = \theta x$. This is a theoretical possibility rather than an actual one, since x is an index for a group.

Even if linear independence is assured, the endogenous effect (β) is not identified when the contextual effect is present (i.e., $\gamma \neq 0$). As Manski (1993) stressed, distinguishing the endogenous effect from the contextual effect is important because these two effects have critically different policy implications. Consider, for example, that a lecture on the health effects of smoking is provided to one class, but not to other classes in a particular school. If the lecture effectively makes this group of students quit smoking, the effect propagates to the students in the other classes through the endogenous effect. If the endogenous effect does not exist, the effect of the lecture is limited to the class where the lecture is given. While the endogenous effect implies this "social multiplier," the contextual effect and the correlated effect do not share this same implication.

Moreover, Manski (1993) warned that, in general, the sample correspondence of $E(y|x)$, denoted by $E_N(y|x)$, is not identical to $E(y|x)$. As a result, $\hat{\beta}$ can be calculated in the sample, even if β is not identified in the population. Thus, the successful calculation of $\hat{\beta}$ does not imply anything about the identification. After all, the the mixture of the endogenous effect and the contextual effect can be identified under the assumption of the linear independence of $[1, E(z|x), x, z]$. Distinguishing between the endogenous effect and the contextual effect is, however, impossible on principle.

The difficulty of the identification arises because the mean of the outcome is used as an explanatory variable. I avoid this complication in this study because the direct subjective perceptions of peers' behavior, instead of the peers' average behavior, are the key explanatory variable. The advantage of this approach is that once the linear independence assumption of the explanatory variables is assured, the endogenous effect and the contextual effect are *separately* identified. In the next section, the model with subjective perception is discussed.

3. The model with perceived peer behaviors

The model with perceived peer behaviors is specified as follows:

$$y = \alpha + \beta p + E(z|x)\gamma + z\eta + u, \tag{5}$$

where y is an outcome of interest, x consists of attributes characterizing an individual's reference group, and z is attributes that affect the outcome directly. The variable p is an individual's subjective perception of peer behaviors. Once $E(u|p, z, x) = 0$ and the linear independence of $[1, p, z, E(z|x)]$ are assured,[4] the parameters, α, β, γ, and η, are consistently estimated through OLS. Since the group average of observed behaviors is not used as an explanatory variable as in Manski (1993), no complication of identification arises.

The crucial assumption in this model is that perceived rather than actual peer behaviors determine individual behaviors. This assumption is consistent with the conformist model. Akerlof (1997) argued that individuals receive utility from behaving like an "average" person in a reference group. It is natural to assume that perceived peer behaviors produce the image of the "average" person. This "conformist" behavior may be reinforced through the formation of group norms. By forming group norms that require members to engage in similar behaviors, each individual group member enjoys higher utility. Becker (1996) considered a utility function (U) that has its own norm (N) and the norm of peers (N_P) as arguments and its cross derivative to be positive ($\partial^2 U/(\partial N \partial N_P) > 0$). Thus individuals can obtain higher utility by forming their own norms similar to the group norm. It is natural again to assume that each individual perceives group norms through perceived peer behaviors. Individuals behave according to their norms. To summarize, the assumption that perceived peer behaviors determine individual behaviors is a natural one if peer effects operate through "conformist" preferences or the enforcement of group norms.

At the same time, there are many other reasons why peer effects exist. For example, American children play baseball instead of cricket because other children know how to play baseball, and, consequently, it is easier to find playmates. Moreover, since playing an actual game is much more fun than playing catch, it may be an important goal for children to find many playmates easily. In this example, the utility obtained from playing baseball depends on the number of players with whom a child plays (up to 18 players). By the same token, teenagers may obtain higher utility from tobacco when they consume it with their friends. In these situations, actual peer behaviors, rather than perceived ones, affect an individual's behaviors. Although this possibility cannot be ruled out, it is simply assumed that an individual behavior is influenced only by the perceived peer behaviors.[5] Relaxing this assumption causes complications in identifying the endogenous effect as noted earlier. Thus, the interpretation of the endogenous effect in this paper should be restricted by this assumption.

4. Literature review

Many papers have investigated peer effects. In this section, the papers are reviewed in order by the identification strategy used. The identification of the endogenous effect is especially focused because of its unique policy implication associated with the "social multiplier." A thorough review of the existing literature reveals the strengths and weaknesses of each identification strategy. In particular, the limitation of an identification strategy that only uses observed behavior becomes clear.[6]

Many studies of peer effects assume the absence of the contextual effect to estimate the endogenous effect (Case and Katz 1991; Sacerdote 2001; Gaviria and Rapheael 2001). Replacing the population mean $E(y|x)$ with the sample average $E_N(y|x)$, the (2) is estimated assuming $\gamma = 0$. The assumption $\gamma = 0$ is not severe if the research interest lies in examining broadly defined peer effects. However, if the existence of the "social multiplier" is the main concern, the assumption is restrictive, since all of the observed peer effects are attributed to the endogenous effect by assumption. Thus, the identification based on $\gamma = 0$ is not appropriate for deriving policy implications because the impact of policy intervention depends on the existence of the "social multiplier." In addition, the absence of the contextual effect assures that $E(z|x)$ can be used as IV for $E(y|x)$ to deal with the possible endogeneity of peers' behavior (Evans et al. 1992; Gaviria and Raphael 2001). However, this identification strategy is valid only if the assignment of x is exogenous. Thus endogenous sorting into peer groups should be ruled out in this IV strategy.[7]

Sacerdote (2001) and Foster (2002) examined the peer effects among randomly assigned roommates in college dorms and this natural experimental situation effectively ruled out endogenous sorting. To deal with the endogenous sorting based on family unobserved characteristics, Aaronson (1998) used a sibling estimator to allow for a correlation between family unobserved characteristics and neighborhood quality. He estimated the effect of neighborhood quality on a teenager's dropping out of high school. Since the neighborhood quality cannot vary within the same family, he used a sibling sample from families that moved. Although these studies effectively overcome the issue of endogenous sorting, it is still impossible to distinguish the endogenous effect from the contextual effect.

Studies of Biddle (1991) and Norton et al. (1998) are examples of research that exploit the lag structure of the endogenous effect for the identification. Biddle (1991) regressed the current demand for personalized license plates on the previous year's demand using state-level aggregate data. Norton et al. (1998) regressed sixth to ninth graders' current use of substances on the average use of substances among students who attended the same primary school. Due to the lag structure, these estimators are immune from Manski (1993)'s criticism; however, one should realize that identifying the endogenous effect crucially depends on excluding the contemporaneous peer effect by assumption.

A series of studies used economic theoretical predictions as identification information. Those studies basically constructed a game theoretic model and estimated the best response function of agents or the relationships of individual behavior and the average group behavior implied by the equilibrium outcome. Neumark and Postlewaite (1998) estimated the importance of "relative income concern" as an explanation for the rapid increase of the labor force participation (LFP) rate among U.S. married women. Munshi (2000) analyzed technological diffusion during the Green Revolution in India. These studies are persuasive because they exploited the theoretical prediction as identification information and did much more than regress individual behavior on group behavior.

5. Estimation

The review of the existing research clearly indicates the limitation of studies that used only observed behavior to identify the endogenous effect. As has been discussed, economic theory can provide important information for identifying these effects. In the context of substance usage by teenagers, however, obtaining such a theoretical prediction is difficult because the teenagers' preferences, which are necessary to derive behavioral predictions theoretically, are largely unknown *a priori*. I thus use perceived peer behaviors, which are potentially error ridden, as additional information to identify the endogenous effect. Descriptions of the data, the econometric model, and the results of the estimations follow.

5.1. Data

The data set used in this study is the National Longitudinal Survey Youth 97 GeoCode file. The sample construction is summarized in Table 1. I used the set (10) ($N = 6356$) as a sample for the cross-sectional studies, and I applied a

Table 1. Sample construction

		N	Average incidence (non weighted)		
			Smoking	Drinking	Marijuana use
(1)	Entire sample	8984	0.162	0.185	0.086
(2)	All outcomes are available	8940	0.161	0.185	0.085
(3)	Demographic variables are available	8851	0.161	0.185	0.086
(4)	Relationships with parents are available	8833	0.161	0.185	0.086
(5)	All peer variables are available	8518	0.165	0.190	0.088
(6)	School characteristics are available	7498	0.166	0.192	0.091
(7)	Grade in school is available	7495	0.166	0.192	0.091
(8)	Parent's HGC available	7491	0.166	0.192	0.091
(9)	Variables from parent questionnaire are available	6615	0.168	0.193	0.092
(10)	Proxy variables are available* (Basic analysis sample)	6356	0.168	0.195	0.092
(11)	Within school duplication occurs** (School fixed effect analysis sample)	6312	0.167	0.196	0.092
(12)	Siblings data are available*** (Sibling fixed effect analysis sample)	2458	0.170	0.192	0.093

Note:
* Proxy variables for school quality (if experienced threat, if something stolen in school, feel safe in school) and neighborhood quality (if any gang in neighborhood).
** Since the school identification numbers are not available, quasi-school id's, which are composed of county dummy, school size, and student-teacher ratio, are used. The Bureau of Labor Statistics assigns the last two variables based on the school id number.
*** Siblings were determined by an identical household id.

quasi school fixed effects estimation and a sibling fixed effects estimation to the set (11) (N = 6312) and (12) (N = 2458) respectively. In Table 1, the sample means of outcomes are tabulated. From the tabulation, we can confirm that the sample selection does not drastically change the sample's properties in terms of outcomes.

The outcomes are constructed by using questions about substance use in the last 30 days. The respondent who smokes/drinks more than or equal to one cigarette/drink is defined as a smoker/drinker; similarly, the respondent who uses marijuana more than or equal to once per month is defined as a marijuana user.

In order to construct the peer variables, respondents were asked about their subjective perceptions of their peers' behavior with the following questions:

> "What percentage of kids (in your grade / in your grade when you were last in school)
> (smoke/smoked) cigarettes?
> (get/got) drunk at least once a month?
> (have / ever) used marijuana, inhalants, or other drugs?"

Respondents were allowed to answer the questions with one of the following five responses:

1. almost none (less than 10%)
2. about 25%

3. about half (50%)
4. about 75%
5. almost all (more than 90%).

From these categories, I constructed perceived peers' behaviors in which "almost none" was coded as 0 and "almost all" was coded as 1.

Descriptive statistics of individual substance uses and perceived peer substance uses are tabulated in Table 2. An interesting finding is that the respondents systematically overestimated peer behaviors and the degree of overestimation is not negligible. It is worth noting that this measurement error is not a problem if we assume that the variable that affects the respondents' behaviors is perceived peer behaviors rather than "objective" (for econometricians) peer behaviors.[8]

5.2. The model

This paper adopts a binary response model because the dependent variable in this study is the incidence of substance use. The essence of the discussion for linear models carries over, however. Using the latent variable framework, the substance use by a teenager is specified as

$$y^* = \alpha + \beta p + E(z|x)\gamma + z\eta + u, \quad y = 1(y^* \geq 0), \tag{6}$$

where y^* is the latent variable indicating the individual's tendency to smoke and y is a binary variable set to one if the individual is a substance user. The variable p is perceived peer behaviors. The vector z contains a set of student, family, school, and regional characteristics that may affect individual substance use. The variable x is an index of the group, which is "schools" in this model. The determination of perceived behavior is specified as

$$p = \theta_1 E(y|x) + z\theta_2 + E(z|x)\theta_3 + v. \tag{7}$$

I assume $u|x, z, v \sim Normal(0, 1)$. By including several measures of parental involvement in vector z, such as participation in PTA meetings, and many variables (85 total) that may affect a youth's substance use, the assumption $u|x, z, v \sim Normal(0, 1)$ is plausible.

There are still potential sources of omitted variable bias, however. For example, state anti-drug campaigns to reduce teenager substance use is a possible omitted variable that may affect both a respondent's and his or her peers' substance use (a "correlated effect," using Manski (1993)'s terminology). To reduce this possibility, I include many regional variables in z that characterize the county where the teenager lives (e.g., state cigarette tax rate, beer tax rate, county-level poverty rate, county-level unemployment rate, the county's demographic composition, and other characteristics). It is still fair to say, though, that the assumption $u|x, z, v \sim Normal(0, 1)$ can be violated. Thus, I will later relax this assumption and use a fixed effects estimation.

The Probit estimator is a consistent estimator under the assumption of $u|x, z, v \sim Normal(0, 1)$. This means that the error term of the behavioral equation (u) is not dependent with the error term in the equation determining perceptions of peer behaviors (v) after conditioning on (x, z). The

Table 2. Descriptive statistics of substance use and subjective measures of peers' behaviors by grade

Grade	Smoked last 30 days	Peers who smoke (Subjective)	Drunk last 30 days	Peers who get drunk (Subjective)	Used marijuana last 30 days	Peers who use illegal drug (Subjective)	Number of Observation
4	0	0	0	0	0	0	2
5	0.044 (0.043)	0.118 (0.055)	0.041 (0.033)	0.077 (0.039)	0	0.068 (0.038)	37
6	0.058 (0.013)	0.147 (0.013)	0.039 (0.011)	0.037 (0.007)	0.018 (0.007)	0.084 (0.010)	426
7	0.097 (0.009)	0.273 (0.009)	0.084 (0.009)	0.111 (0.007)	0.036 (0.006)	0.174 (0.008)	1294
8	0.153 (0.011)	0.375 (0.009)	0.146 (0.011)	0.202 (0.008)	0.067 (0.008)	0.257 (0.009)	1319
9	0.238 (0.013)	0.529 (0.008)	0.269 (0.013)	0.389 (0.009)	0.145 (0.011)	0.426 (0.010)	1416
10	0.260 (0.014)	0.553 (0.008)	0.329 (0.015)	0.481 (0.009)	0.142 (0.011)	0.476 (0.010)	1204
11	0.260 (0.014)	0.553 (0.008)	0.329 (0.015)	0.481 (0.009)	0.154 (0.016)	0.474 (0.010)	624
12	0.291 (0.021)	0.555 (0.011)	0.371 (0.022)	0.513 (0.012)	0.154 (0.016)	0.487 (0.060)	34
Total	0.190 (0.006)	0.427 (0.004)	0.214 (0.006)	0.303 (0.004)	0.099 (0.004)	0.333 (0.004)	6356

Note:
1. All statistics are calculated using sampling weight.
2. Standard errors of means are in parentheses.

identification of β hinges on this assumption. Intuitively, individual teenagers have some shock regarding what they think about peers' behavior and then they react to those perceptions. But the shock about these perceptions should not be correlated with the error term in the teenager's behavioral equation. This assumption arguably rules out the possibility that the perception of peers' behavior is contaminated by the teenager's own behavior. For example, if a teenager perceives that many of his/her peers smoke because he/she smokes, then the assumption is violated. Thus, it should be noted that this is a strong identifying assumption. Including a rich set of control variables, z, however, might reduce the risk of violating this assumption. Under this assumption, applying a Probit to (6) renders the estimator consistent up to scale. The cell average of z within a school is used instead of $E(z|x)$.[9]

5.3. The results

I report the results of the above model in Table 3.

For cigarette smoking, the marginal effect of 0.205 means that a 10 percentage point increase in the subjective perception of the peer smoking probability increases the probability of smoking by 2.05 percentage points. This estimate is statistically significant. For alcohol drinking, the estimated marginal effect is 0.263 and for marijuana usage, it is 0.141.

The results clearly show the existence and statistical significance of peer effects. When a teenager's perception of the percentage of his/her peers who use a substance increases by 10 percentage points, the probability that he/she will use the substance increases from 1.4 to 2.6 percentage points. Although the difference in identification strategy prohibits me from making a serious comparison of the estimates, the estimated peer effect is comparable to the estimated effect of Gaviria and Raphael (2001) for smoking and alcohol drinking (0.150 for smoking, 0.106 for alcohol drinking, and 0.254 for drug use). The absence of the contextual effect ($\gamma = 0$) is not rejected at a 5% significance level through the Wald test. In summary, the results of the estimation robustly show the existence of peer effects in the *causal* sense. Moreover, by using perceived peer behavior as the key independent variable, the results show that peer effects work through the endogenous effect. This implies the existence of the "social multiplier."

The causal interpretation depends on the assumption $u|x, z, v \sim Normal(0, 1)$. The multitude of variables in z (86 in total), however, makes this assumption realistic.

6. The quasi school fixed effects estimation

Although the previous section made the best possible effort to assure the assumption $u|x, z, v \sim Normal(0, 1)$ is correct, the school index x nonetheless may still contain some information that systematically predicts teenagers' substance use that the regional or school characteristics included in z fail to capture. In other words, there might be unobserved regional and school factors that encourage the teenagers' substance use. For example, suppose a cigarette shop is located just in front of High School A. Moreover, suppose

Table 3. Probit estimates of incidence of substance use

Dependent variable	(1) Incidence of cigarette smoking in the last 30 days	(2) Incidence of alcohol drinking in the last 30 days	(3) Incidence of marijuana use in the last 30 days
Method of estimation	Probit	Probit	Probit
Peer (fraction)			
Peer smoke	1.024 (0.078) [0.205]		
Peer drunk		1.128 (0.072) [0.263]	
Peer illegal drug			1.459 (0.087) [0.141]
z (other control variables)	Yes	Yes	Yes
E[z |x] (contextual effect)	Yes	Yes	Yes
Wald statistics for Contextual effect	86.17 (0.272)	81.78 (0.393)	100.04 (0.055)
Log likelihood	−2371.238	−2602.054	−1526.628
Sample size	6345	6327	6308

Note:
1. The vector z contains following variables:
z (independent variables): Dummies if respondent lives with mother, father, biological mother, biological father, foster mother, or foster father. Female dummy, age, black dummy, other minority dummy, Hispanic dummy, school grade dummies (grade 5 – grade 12), [Middle/junior high] school and high school dummies, catholic school dummy, private school dummy, student/teacher ratio category dummies (3 categories), school size category dummies (3 categories), Census regional dummies, urban dummy and proxy variables for unobserved school characteristics (if experience threat, if something stolen at school, feel safe in school), parents background variables (parent born in U.S., parent speaks a language other than English at home, parent was with both biological parents at age of 14), last year's household income category dummies (less than \$20,000, \$20,001 – \$40,000, \$40,001 – \$60,000, \$60,001 – \$80,000, more than \$80,001, and household income not available), household size, number of household members less than age 18 and less than 6, proxy variables for parent's involvement in education (often or sometimes participate in PTA activity, often or sometimes volunteer in school education), dummies for mother's education and father's education, dummy if any gang in neighborhood, county level variables (share of white population, black population, Indian population, Hispanic population, share of population under 5 years old, 5-17 years old, 18-20 years old, 21-24 years old, 25-34 years old, 35-44 years old, 45-54 years old, 55-64 years old, 65-74 years old and 75+ years old, share of males in population), state tax rates (cigarettes tax and beer tax). There are 86 variables in total, counting each dummy of the categorical variable as a variable.
2. 11, 29, and 48 observations were dropped from the basic analysis sample due to perfect prediction for the estimation of (1), (2), and (3) respectively.
3. Standard errors are in parentheses for the estimated probit coefficient. Marginal effects evaluated at the sample mean are in brackets. For Wald statistics, p-values are in parentheses.

this "unobservable" makes a student in the high school 50 percentage points more likely to smoke. Then the assumption $u|x, z, v \sim Normal(0, 1)$ is violated, because

$$u|x, z, v \sim \begin{cases} Normal(0.5, 1), & \text{if } x = A, \\ Normal(0, 1), & \text{Otherwise.} \end{cases} \tag{8}$$

This possibility is addressed through the school fixed effects estimation, which can be represented in the equation

$$y_{ij}^{*} = \beta p_{ij} + z_{ij}\gamma_1 + a_j + u_{ij}, \ y_{ij} = 1(y_{ij}^{*} \geq 0). \tag{9}$$

Here i is a subscript for an individual and j is a subscript for a school. The coefficient γ_2 is not identified because $E(z|x)_j$ is invariant within a school. The random variable $a(x)_j$ captures the school-specific unobservable that affects the teenager's substance usage, such as a cigarette shop at the school gate.

Under the assumptions $u|x, z, v, a \sim Normal(0, 1)$ and $a|x, z, v \sim N(0, \sigma_a^2)$, the random effects estimator is consistent up to scale. The assumption $a|x, z, v \sim Normal(0, \sigma_a^2)$ excludes the dependence between unobserved school heterogeneity and the explanatory variables. To allow for the dependence of unobserved school characteristics on explanatory variables, I assume a specific form of dependence of unobserved school heterogeneity and explanatory variables as follows:

$$a|x, z, v \sim Normal(\bar{p}\delta_1 + \bar{z}\delta_2, \sigma_a^2), \tag{10}$$

where $\bar{p}$ and $\bar{z}$ are the school average of the perceptions of peers' behavior and individual backgrounds respectively. The combination of (9) and (10) becomes the "fixed effects" Probit estimator, since the estimator is still consistent with the dependence of heterogeneity and explanatory variables. If $\delta_1 = 0$ and $\delta_2 = 0$, then the random effects Probit estimator is preferable because it is more efficient.

Since the respondents' school identification number (ID) is not available in my data set, I created a quasi school ID by matching the respondents' county ID, school size, and student/teacher ratio.[10]

The results of the random effects Probit estimation and the test statistics under the null $\delta_1 = 0$ and $\delta_2 = 0$ appear in Table 4. None of the tests reject the null of $\delta_1 = 0$ and $\delta_2 = 0$; thus I only report the result of the random effects estimation. Marginal peer effects for smoking, drinking, and marijuana usage are 0.205, 0.258, and 0.138 respectively. These numbers are similar to the Probit results. This is probably due to the fact that the vector z already contains enough information to capture the school characteristics.

7. The household fixed effects estimation using the sibling sample

To reinforce the previous results, I estimated the household fixed effects model using sibling samples. One concern in the previous research (Evans et al. 1992) was the endogeneity of peer quality due to omitted household characteristics. Peer quality can be endogenous, since parents who are willing to invest in their children may send their children to a school with good peers and parental care can also directly affect a child's behaviors. This endogeneity problem can be avoided by controlling household unobserved heterogeneity using the household fixed effects estimation.[11]

I estimate the following model:

$$y_{ij}^{*} = \beta p_{ij} + z_{ij}\gamma_1 + c_j + u_{ij}, \ y_{ij} = 1(y_{ij}^{*} \geq 0) \tag{11}$$

where i is a subscript for an individual, j is a subscript for a household, and c is a household's unobserved heterogeneity. The same econometric discussion

Table 4. Incidence of substance use using within school and household duplication data. (Pseudo school random & family fixed effects probit estimation.)

Dependent variable	(1) Incidence of cigarette smoking in last 30 days	(2) Incidence of alcohol drinking in last 30 days	(3) Incidence of marijuana use in last 30 days	(4) Incidence of cigarette smoking in last 30 days	(5) Incidence of alcohol drinking in last 30 days	(6) Incidence of marijuana use in last 30 days
Method of estimation unobserved effects		Random-effects probit school			Fixed-effects probit household	
Peer (fraction)						
Peer smoke	1.014 (0.077) [0.205]			0.723 (0.206) [0.077]		
Peer drunk		1.116 (0.072) [0.258]			1.165 (0.190) [0.169]	
Peer drug			1.425 (0.086) [0.138]			1.580 (0.252) [0.031]
z (Other control variables)	Yes	Yes	Yes	Yes	Yes	Yes
Hausman test	83.48 (0.315)	74.63 (0.587)	84.33 (0.292)	3.62 (<0.000)	3.03 (0.002)	3.90 (<0.000)
Log likelihood	−2394.503	−2637.821	−1562.674	−883.623	−952.772	−541.188
Sample size	6312	6312	6312	2458	2458	2458

Note:
1. School id's are not available in the data, so alternative school id's were created from county id, school size, and student-teacher ratio. The surveyor assigned the last two variables based on the confidential school id numbers.
2. The same control variables that appear in Table 3 were included. Some of the variables that do not vary within household were dropped in the fixed effects estimation.
3. Standard errors are in parentheses for the estimated Probit coefficient. Marginal effects evaluated at the sample mean are in brackets. For test statistics, p-values are in parentheses.
4. The fixed effects Probit estimation includes household average perceived peer behavior in the specification to allow for the dependence of individuals' perceived peer behavior and unobserved household characteristics.
5. Hausman test statistics are calculated under the null hypothesis that unobserved heterogeneity is not correlated with the group average of explanatory variable(s). The statistics are Wald statistics and the p-values are in parentheses. The statistics are distributed as an χ^2 distribution for the school unobserved effects models and as a t distribution for household unobserved effects models. Random effects specifications are favored for school unobserved effects models and fixed effects specifications are favored for family unobserved effects models.

from the previous section applies. The only the exception for the fixed effects estimation is that I assume

$$c|x, z, v \sim Normal(\bar{p}\delta_1, \sigma_c^2). \tag{12}$$

The specification only allows the dependence of c and $\bar{p}$, since siblings tend to share the characteristics in z, and allowing for the dependence of c and $\bar{z}$ causes serious multicolinearity between z and $\bar{z}$. This prohibits the convergence of parameters in the maximum likelihood estimation.

The results of the fixed effects estimation and the test statistics for $\delta_1 = 0$ appear in Table 4. All of the tests reject the null of $\delta_1 = 0$. Thus, the fixed effects estimator is preferred and only the fixed effects results are reported. Peer effects for smoking, drinking, and marijuana usage are 0.077, 0.169, and 0.031 respectively. The fixed effects coefficient estimates are smaller than those obtained from the pooled Probit estimation due to the positive correlation between household unobserved heterogeneity and the peer variable. Nevertheless, even after allowing for the correlation between household heterogeneity and the peer variable through the fixed effects model, the estimated peer effects are practically and statistically significant.

8. Peer effects within different demographic groups

Thus far, the existence of peer effects is a robust result. Next, it is interesting to investigate the strength of peer effects within different demographic groups. Knowing the groups in which the endogenous effects are strong, policy makers can effectively target policies that are likely to discourage youth substance use within those groups, since they can expect larger policy effects through the larger "social multiplier."[12] To estimate the difference in the endogenous effects across groups, I assume the following model in which the strength of peer effects may depend on the different demographic groups:

$$y^* = \alpha + \beta_1 p + p * z_1\beta_2 + z\gamma + E(z|x)\eta + u, \tag{13}$$

where z_1 is the part of z that defines demographic groups.[13] Since the exogeneity of p was not rejected in the school fixed effects estimation, I assume $u|x, z, v \sim Normal(0, 1)$. Under this assumption, the Probit estimator is a consistent estimator up to scale.

The results of the regressions appear in Table 5. Some of the estimated coefficients on the interaction terms are statistically significant and the effects are not negligible. For all three substances, there was no gender difference. Fewer peer effects were found among black teenagers. For smoking behavior, peer effects among black teenagers are about one- to two-thirds of that found among white teenagers. The smaller peer effects among black teenagers are statistically different from zero for all substances. Hispanics are also less vulnerable to peer pressure. An expectation is that minority teenagers might not obtain as much utility as non-minority teenagers from imitating each other. Teenagers with both biological parents are less likely to be affected by their peers in their smoking. As for drinking, and marijuana use, the coefficient on the interaction term for the peer variable and "both biological parents" is not statistically significant. The first result may imply that teenagers who do not have both biological parents present are more likely to

Table 5. Probit estimates of incidence of substance use

Dependent variable	(1) Incidence of cigarette smoking in the last 30 days	(2) Incidence of alcohol drinking in the last 30 days	(3) Incidence of marijuana use in the last 30 days
	Probit	Probit	Probit
Peer			
Peer's substance usage	1.327	1.305	1.588
(portion)	(0.137)	(0.130)	(0.144)
	[0.277]	[0.315]	[0.172]
Female × peer's usage	0.073	−0.053	0.067
	(0.139)	(0.124)	(0.151)
	[0.015]	[−0.013]	[0.007]
Black × peer's usage	−0.861	−0.466	−0.658
	(0.167)	(0.154)	(0.180)
	[−0.180]	[−0.113]	[−0.071]
Hispanic × peer's usage	−0.064	−0.105	−0.200
	(0.189)	(0.161)	(0.196)
	[−0.013]	[−0.025]	[−0.022]
Both biological parents	−0.160	0.044	0.147
× peer's usage	(0.122)	(0.117)	(0.135)
	[−0.034]	[0.011]	[0.016]
z (Other control variables)	Yes	Yes	Yes
Log likelihood	−2412.822	−2652.988	−1578.565
Sample size	6345	6327	6345

Note:
1. The same control variables as presented in Table 3 are included.
2. 11, 29, and 48 observations were dropped from the basic analysis sample due to perfect prediction for the estimation of (1), (2), and (3) respectively.
3. Standard errors are in parentheses for the estimated Probit coefficient. Marginal effects evaluated at the sample mean are in brackets. For test statistics, p-values are in parentheses.

depend on their peers to form their behavior. This result is consistent with Steinberg (1987)'s.

9. Conclusion

The estimation of peer effects on substance usage through perceived peer behaviors shows significant peer effects. When the perceived peer substance use increases by ten percentage points, the probability that a teenager will use substances increases by two to three percentage points. Moreover, the endogenous effect is found to be more important than the contextual effect when explaining the peer effects on youth substance use. This finding implies that current peer behaviors, rather than peer backgrounds, determine individual behaviors. Thus, if some exogenous shock reduces a group's substance use, this reduction affects other groups of youths through the endogenous effect. Hence, policy makers can expect a "social multiplier" effect in policies that discourage youth substance use.

In my model, the endogenous effect is identified when perceived peer behaviors are exogenous. To assure this exogeneity assumption, I used a rich set of controls consisting of parent, neighborhood, and school characteristics. Moreover, the robustness of the results was confirmed through the school and household fixed effects estimations. We also find that the strength of peer effects depends on the demographic group to which a teenager belongs. Peer effects are found to be large among white teenagers and teenagers.

Although this paper finds a robust peer effect, this study does not shed enough light on the mechanism of peer effects itself. Thus this study is still a reduced form study of peer effects. More direct study of the mechanism of peer effects is left for future research.

Endnotes

1 On the other hand, Gruber (2001) reports that the percentage of cigarette smokers increased by one-third between 1991 and 1997. (Based on a youth behavior risk survey for 9th through 12th graders, the number increased from 27% to 36% between 1991 and 1997.)

2 See Los-Angeles-Times (1999) for interviews with youth smokers on the reasons why they smoke. For the effect of alcohol price on alcohol consumption among youth, see Yamada et al. (1996) for example.

3 A seminal work in this field by Case and Katz (1991) asked youths living in low-income, inner-city Boston neighborhoods about their perceptions of their neighborhoods. They did not, however, use the information to explain the youths' behavior.

4 The assumption of linear independence here is more restrictive than the assumption needed in Manski (1993), since p is newly added to the list of variables that should be linearly independent.

5 Of course, the actual peer behaviors may determine perceived peer behaviors; however, here I assume that once perceived peer behaviors are included in the behavioral equation, the equation does not include actual peer behaviors.

6 Many of studies surveyed in this section used nonlinear models such as probit. However, this section discusses identification strategies in the context of linear models, since the discussions of identification fundamentally carry over. Moreover, the identification of parameters should not depend only on the nonlinearity assumption.

7 To measure the degree of endogenous sorting, Gaviria and Raphael (2001) divided their sample into two sub-samples: families that had moved in the previous two years and families that had not moved. They argued that if endogenous sorting is widespread, estimates of peer effects for families that had recently moved into a new neighborhood should be larger. Their findings were mixed. Different peer effects were found in the two sub-samples for marijuana usage but not for drinking, smoking, church attendance, or dropping out of high school.

8 The difference between the subjective measure and the objective measure of the peer behaviors implies that we cannot calculate the size of the social multiplier effect from the size of endogenous effect (β), since we need to know how individuals formed their perceptions. Only with perceived peer behaviors can we confirm the existence of the endogenous effect without assuming the absence of the contextual effect (i.e., $\gamma = 0$), as discussed before.

9 Although this first step estimation may change the asymptotic distribution of the Probit estimators, it is known that under the null of $H_0 : \gamma = 0$, the asymptotic distribution is not affected.

10 School size is classified into five categories: 1-299, 300-499, 500-749, 750-999, and 1000+ students. Student/teacher ratio is classified into four categories: less than 14, 14-17, 18-21, and more than 22. The Bureau of Labor Statistics assigns these two variables to each respondent based on its confidential information. Thus, when each school in a county differs in either school size or student/teacher ratio, I can identify all of the schools. If all of the schools in a county share both the same school size and student/teacher ratio, I just identify the county. In the worst-case scenario, it is assumed that $E(z|x)$ is constant within a county. This mis-matching becomes serious if the variation of $E(z|x)$ within a county is huge. However, the direction of bias in the estimator of γ caused by this measurement error is not clear *a priori*, since the measurement error is mean reverting (to the county mean) and not classical. At least the contextual effect that operates at the county level is captured, however.

[11] The sibling method, however, does not necessarily solve the endogeneity issue. Since the between households difference of outcomes and perceived peer behaviors are wiped away, the identification solely depends on the within household variation of outcomes and perceptions. If each sibling in a household has unobserved characteristics that determine both substance use and perceptions of peer behaviors, the sibling estimator is still biased. If this within household unobservable plays a more important role than the between households unobservable, the sibling estimator can be more biased. However, this discussion is unlikely to apply in the context of substance use. See Bound and Solon (1999) and Neumark (1999) for possible biases in the sibling estimator of the return to education.

[12] This discussion assumes, though, that sensitivity to the policy (a part of η) is the same across groups.

[13] I also estimated the endogenous effect using a sub-sample of demographic groups. The results obtained were qualitatively the same as the results obtained in this section.

References

Aaronson D (1998) Using Sibling Data to Estimate the Impact of Neighborhoods on Children's Educational Outcomes. *Journal of Human Resources* 33(4):915–946

Akerlof G (1997) Social Distance and Social Decisions. *Econometrica* 65(5): 1005–1027

Becker G (1996) *Accounting for Tastes*. Harvard University Press, Cambridge, Massachusetts

Biddle J (1991) A Bandwagon Effect in Personalized License Plates? *Economic Inquiry* 29(2):375–388

Bound J, Solon G (1999) Double Trouble: On the Value of Twin-Based Estimation of the Return to Schooling. *Economics of Education Review* 18:169–182

Case A, Katz L (1991) The Company you Keep: The Effects of Family and Neighborhood on Disadvantaged Youths. NBER Working Paper Series May

Evans W, Oates W, Schwab R (1992) Measuring Peer Group Effects: A Study of Teenager Behavior. *Journal of Political Economy* 100(5):966–991

Foster G (2002) Peerless Performers: The Absence of Robust Peer Effects at a Large Public University. University of Maryland

Gaviria A, Raphael S (2001) School Based Peer Effects and Juvenile Behavior. *Review of Economics and Statistics* 83(2):257–268

Gruber J (2001) Youth Smoking in the 1990's: Why Did it Rise and What are the Long-Run Implications? *American Economic Review* 91(2):85–90

Los-Angeles-Times (1999) Smoking Among Youths Continue to Rise. January 8

Manski C (1993)Identification of Endogenous Social Effects: The Reflection Problem. *Review of Economic Studies* 60(3):531–542

Manski C (2000) Economic Analysis of Social Interactions. *Journal of Economic Perspective* 68(3):115-36

Munshi K (2000) Social Leaning in a Heterogeneous Population: Technology Diffusion in the Indian Green Revolution. University of Pennsylvania

Neumark D (1999) Biases in Twin Estimates of the Return to Schooling. *Economics of Education Review* 18:143–148

Neumark D, Postlewaite A (1998) Relative Income Concerns and the Rise in Married Women's Employment. *Journal of Public Economics* 70:157–183

Norton E, Lindrooth R, Ennett S (1998) Controlling for the Endogeneity of Peer Substance Use on Adolescent Alcohol and Tobacco use. *Health Economics* 7:439–453

Sacerdote B (2001) Peer Effects with Random Assignment: Results for Dartmouth Roommates. *Quarterly Journal of Economics* 116:681–704

Steinberg L (1987) Single Parents, Step Parents, and the Susceptibility of Adolescents to Antisocial Peer Pressure. *Child Development* 58:269–275

Yamada T, Kendix M,Yamada T (1996) The Impact of Alcohol Consumption and Marijuana Use on High School Graduation. *Health Economics* 5:77–92