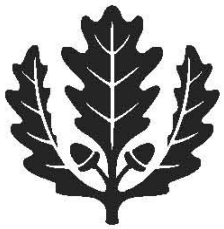

University of Connecticut

Department of Economics Working Paper Series

Estimating the Effects of Friendship Networks on Health Behaviors of Adolescents

Jason M. Fletcher
Yale University
Columbia University

Stephen L. Ross
University of Connecticut

Working Paper 2011-26

December 2011

341 Mansfield Road, Unit 1063
Storrs, CT 06269–1063
Phone: (860) 486–3022
Fax: (860) 486–4463
http://www.econ.uconn.edu/

This working paper is indexed on RePEc, http://repec.org/

Estimating the Effects of Friendship Networks on Health Behaviors of Adolescents[*]

Jason M. Fletcher
Yale University
Columbia University

Stephen L. Ross
University of Connecticut

This Draft: December 28, 2011

Abstract

This paper estimates the effects of friends' health behaviors, smoking and drinking, on own health behaviors for adolescents while controlling for the effects of correlated unobservables between those friends. Specifically, the effect of friends' health behaviors is identified by comparing similar individuals who have the same friendship opportunities because they attend the same school and make similar friendship choices, under the assumption that the friendship choice reveals information about an individual's unobservables. We combine this identification strategy with a cross-cohort, within school design so that the model is identified based on across grade differences in the clustering of health behaviors within specific friendship patterns. Finally, we use the estimated information on correlated unobservables to examine longitudinal data on the on-set of health behaviors, where the opportunity for reverse causality should be minimal. Our estimates for both behavior and on-set are very robust to bias from correlated unobservables.

Key Words: Peer Effects, Friendship Networks, Adolescent Health, Smoking, Drinking, Cohort Study

JEL Codes: D85, I19, I21, J13

[*] We received valuable comments from numerous seminar participants at Baylor University, Cornell University, Lafayette College, Lehigh University, Texas A&M, University of California-Santa Barbara, University of Texas-Austin, Yale University, Population Association of American Conference, the Annual Health Econometrics Workshop, NBER Summer Institute, Urban Economics Association, and the Second Annual Economics of Risky Behaviors (AMERB) conference. We thank Michael Anderson, Tao Chen, Ethan Cohen-Cole, Bill Evans, Don Kenkel, Brian Krauth, Anna Mueller, Bruce Sacerdote, Rusty Tchernis and Gautam Tripathi for specific comments that improved the paper.
Fletcher and Ross gratefully acknowledge support from the NICHD (1R21 HD066230-01A1). Fletcher thanks the Robert Wood Johnson Foundation Health & Society Scholars program for its financial support.
This research uses data from Add Health, a program project designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris, and funded by a grant P01-HD31921 from the National Institute of Child Health and Human Development, with cooperative funding from 17 other agencies. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Persons interested in obtaining data files from Add Health should contact Add Health, Carolina Population Center, 123 W. Franklin Street, Chapel Hill, NC 27516-2524 (addhealth@unc.edu).

Extended Abstract

Researchers typically examine peer effects by defining the peer group broadly (all classmates, schoolmates, neighbors) because of the lack of friendship information in many data sources as well as recently to enable the use of plausibly exogenous variation in peer group composition across cohorts in the same school. A classic concern when examining the effects of peers or friends who are chosen by the student is that those choices were made because the friends shared similar unobservables. This paper estimates the effects of friends' health behaviors on own health behaviors for adolescents while insulating these estimates from the effects of correlated unobservables. Specifically, the effect of friends' health behaviors is identified by comparing similar individuals who have the same friendship opportunities because they attend the same school and make similar friendship choices, under the assumption that the friendship choice reveals information about an individual's unobservables. We combine this identification strategy with a cross-cohort, within school design so that the model is identified based on across grade differences in the clustering of health behaviors within specific friendship patterns. This strategy also allows us to separate the effect of friends' behavior on own behavior from the effect of friends' observable attributes on behavior, a key aspect of the reflection problem. We use a partial equilibrium model of friendship formation in order to derive the conditions under which our identification strategy will provide consistent estimates, and a key assumption required for our strategy to be feasible is supported by the empirical patterns of across cohort variation that we observe in our data. Finally, we use the estimated information on correlated unobservables to examine longitudinal data on the on-set of health behaviors where the opportunity for reverse causality should be minimal. Our preferred estimates suggest that estimates in specifications that do not fully take into account correlated unobservables appear to be overstated by no more than 10% in terms of the cross-sectional incidence and no more than 20% in models of the on-set of smoking or drinking.

Introduction

Individuals in modern societies are socially connected in a multitude of ways. For example, the social networking website Facebook.com has increased its membership by 100 million users during 2009, and now there are over 800 million users worldwide. Individuals use their social networks to receive and send information as well as establish, update, and enforce social norms of behavior. Both information acquisition as well as the impacts of social norms within social networks could have large effects on the health behaviors of individuals, particularly adolescents, who are particularly responsive to peer pressure (Brown et al. 1997). This heightening of peer influence also takes place during the developmental stage when many of the most costly health outcomes and behaviors are initiated. Our analysis will use detailed information on individual's health related behaviors and friendship networks from the National Longitudinal Study of Adolescent Health (Add Health) to examine the role of social interactions in these behaviors.

Many studies of social interactions find evidence of clustering of outcomes or behaviors above and beyond the clustering that might have been expected based on individuals' observables, including studies of crime (Glaeser, Sacerdote and Scheinkman 1996), employment (Topa 1999, Bayer, Ross, and Topa 2008), welfare usage (Bertrand, Luttmer, and Mullainathan 2000), pre-natal care (Aizer and Currie 2004), and youth health behaviors (Weinberg 2008).[1] We also observe unexpectedly high levels of clustering on health behavior within grades of students at the same school in our data. Specifically, if we look within schools, very little variation remains across grades in student composition in terms of racial or socio-economic variables, but we observe substantial across grade variation in health behaviors for student populations that are nearly identical. The purpose of this paper is to examine whether the within friendship clustering of health behaviors that lies underneath the clustering in specific grades is consistent with the influence of friendship networks.

Specifically, the primary purpose of our analysis is to examine the impact of friends' health behaviors on a student's own behavior while controlling for the likelihood that these students are friends because they have similar unobservabes. Our controls for correlated unobservables are built on the idea that individuals who make the same friendship choices are likely to be more similar overall than might be indicated by their observables. Specifically, we

[1] See Ross (2011), Durlauf (2004) and Ioannides and Loury (2004) for relevant literature reviews.

examine a partial equilibrium model of friendship formation and use the model to illustrate the effect of controlling for fixed effects associated with clusters of observationally equivalent individuals who face the same friendship opportunity set and make the same friendship choices. We show that if individual students face a shock in terms of exposure to health behaviors, then as the number of friendship choices becomes large the unobservables of individuals in the same friendship choice cluster will converge to the same value and so a cluster fixed effect will act as a non-parametric control for unobservable attributes that influence friendship formation and might affect health behaviors. Significantly, this structure also allows us to separate the influence of friends' behaviors on individual behavior from the influence of the observable attributes of those friends, a key part of the reflection problem, because the within friendship cluster comparisons are made between individuals who have observationally equivalent sets of friends and so have similar contextual effects, at least based on observables.

Our identification strategy relies on several empirical features of adolescent friendship networks. First, a large literature suggests that individuals exhibit strong racial, gender, and age preferences when choosing their friends—likes choose likes (Mayer and Puller 2008, Weinberg 2008). Second, data from the Add Health suggests that most friendships occur within grades, which is important for our use of cross-cohort variation in our identification strategy. Finally, as discussed above, individual grades within schools are quite homogenous over racial and socio-economic composition. Specifically, we will estimate models of youth drinking and smoking in high school that control for the share of same sex-same school-same grade friends who exhibit this behavior and fixed effects based on clusters of individuals who have the same race, ethnicity, and maternal educational attainment (individual observables), same school (same friendship opportunity set over observables), and same number of friends overall and for each racial, maternal education, and other demographic subgroups (similar friendship choices). In our preferred specification, we will randomly choose one individual from each grade per cluster so that the model estimates are explicitly identified based on variation across cohorts within a school.

This approach is similar to earlier analyses by Dale and Krueger (2002) and Fu and Ross (2010) who use fixed effects for individuals who are equivalent on key attributes and then have the same outcome or make the same choice as a reduced form control in order to minimize bias from unobservables. However, our analysis has the advantage over these earlier studies because

the identification strategy contains a clear source of exogenous variation that can create within cluster differences in environment, namely differences in exposure to health behaviors associated with belonging to a particular cohort or grade of students. Further, our friendship formation model demonstrates the importance of having such a source of exogenous variation for identification.

This strategy can be illustrated by the following thought exercise: consider a 9^{th} grader and 10^{th} grader who attend the same high school. As we show in detail below, these students face very similar friendship opportunities with respect to racial, gender, and socioeconomic composition of their same-grade classmates, and yet there is substantial clustering of health behaviors into specific cohorts within schools. Thus, if we compare two students who choose similar "types" of friends based on race, maternal education, and other demographic characteristics, there will exist substantial differences in health behaviors between the across cohort friendship opportunities, and those differences in friends' health behaviors is arguably quasi-random. The key is that the age difference between the 9^{th} grader and the 10^{th} grader (who attend the same high school and have the same preferences for "types" of friends) has effectively randomized these two students into their actual friendship network.

We find evidence that this strategy produces somewhat smaller (no more than 10% smaller) "network effect" estimates than the more standard school fixed effect models, and after controlling for correlated unobservables we still find very robust evidence of network effects on the smoking and drinking behavior of adolescents. Further, we find that peer health behaviors are statistically insignificant and not strong predictors of predetermined student or family attributes. We also find no evidence that peer behavior affects student behavior for our placebo variable, chest pain, while specifications that do not use our controls do show "peer effects" in chest pain. In addition, we demonstrate that the effect of controlling for friendship choices on estimates of the influence of friends is quite similar across subsamples of students with different numbers of friends. We find no evidence of falling causal estimates of friends' behaviors as we increase the number of friends, suggesting that the large number of friendship demographics included in the construction of the fixed effect are sufficient to control for individual unobservables even with a small number of actual friends.

Finally, while our fixed effect estimator insulates us from bias from correlated unobservables, neither the fixed effects nor our counterfactuals address concerns about reverse

causation where for example individuals who smoke choose friends who smoke. In a second analysis, we show how our controls for correlated unobservables from the cross-sectional analysis can be combined with longitudinal data to examine the effect of friends' smoking or drinking behavior at baseline on the onset of smoking or drinking prior to follow-up interviews one year later. As in cross-sectional analyses, we find robust evidence that having friends who smoke or drink contributes to the onset of smoking or drinking behavior, and the inclusion of our controls for unobservable student attributes that affect friendship choice have little impact on those estimates.

Background Literature

A large body of research across multiple disciplines has shown very strong correlations in health behaviors for individuals who are socially connected. One reason there has been so much research and policy interest in exploring how networks affect health behaviors and outcomes is the potentially large set of health interventions and policies that could be proposed to leverage social influences on health behaviors. While the promise of using social networks to affect health is compelling, so too are the empirical issues inherent in detecting causal effects of social networks using observational data.

The estimation of the causal effects of social networks on health related behaviors is particularly challenging (Manski 1993). First, individuals who are friends share common unobservables because they chose each other as friends, because they self-select into the same social network, and because individuals in the same social network are simultaneously affected by their shared environment. Second, it is difficult to separate the influence of an individual's behavior and an individual's attributes in determining the health behaviors of his or her friends. Third, individuals may select their friends or peers based on the behavior of those peers, rather than adapting their behavior to the behavior of friends whom they have already selected. Unfortunately, failure to overcome these empirical difficulties casts considerable doubt on the current knowledge base linking the health behaviors among individuals in the same social network. Each of these biases can lead a researcher to incorrectly infer that social networks have a causal influence on behavior. Thus, policies intended to utilize social networks to enhance interventions to reduce unhealthy behaviors could be unable to affect change if social networks do not actually have causal effects. Providing evidence of the causal mechanisms and the likely

effects of policies is essential to be able to properly leverage social network effects on health behaviors.

There have been two directions that researchers have taken in estimating peer effects on health behaviors: [1] focus on broadly defined peer groups, such as all classmates in a school, in order to either (a) exploit cross-cohort population variation[2] in classmate composition (Bifulco et al. 2011, Fletcher 2010, in press, Trogdon et al. 2008, Lundburg 2006, Clark and Loheac 2007) and/or (b) use instrumental variable strategies (Powell et al. 2005, Gaviria and Raphael 2001[3]) or [2] focus on narrowly defined peer groups, such as nominated friends, where the issues with endogeneity are thornier and the estimates are likely less credible (Trogdon et al. 2008, Christakis and Fowler 2007, 2008, Renna et al. 2008, Halliday and Kwak 2008, 2009). In this paper, we seek to combine the more credible research designs from the first literature with the more credible peer group definitions of the second literature.

Since we focus on friendship networks as the definition of peer group in this paper, it is necessary to outline what other researchers have done previously and how our strategy adds to the literature in this area[4]. Renna et al. (2008), Trodgon et al. (2008) and Halliday and Kwak (2009) focus on estimating social contagion in obesity and control for endogeneity of friendship in part by using school fixed effects. However, since substantial friendship sorting occurs within schools, school fixed effects likely do not provide a full solution to the endogeneity of friend selection, unless students select friends randomly within schools. In fact, our estimates show positive "peer effects" in chest pain when we only control for school fixed effects, suggesting that school fixed effects may not be sufficient to control for endogeneity. In addition, Renna et al. (2008), Trodgon et al. (2008) and Halliday and Kwak (2008) use instruments for friends' outcomes, including friends' parents' outcomes or attributes.. It is unclear whether these

[2] See also the similar literature estimating peer effects in education outcomes (Angrist and Lang 2004, Friesen and Krauth 2011, Gould, Lavy, and Paserman 2009, Hoxby 2000, Lavy, Paserman, and Schlosser In Press, Lavy and Schlosser 2011, Hanushek et al. 2003)

[3] Instruments used in these analyses are often questionable, such as census poverty measures. Fletcher (2010) provides suggestive evidence that these instruments are invalid and proposes alternatives. Trodgon et al. (2008) and Fletcher (2010) use a combination of fixed effects and instruments.

[4] There have been recent examinations of the effects of social networks on obesity and smoking in the medical literature (Christakis and Fowler 2007, 2008), where "friends" are measured by the names respondents provide as potential contact sources for future survey waves. In order to control for endogeneity of friendships, Christakis and Fowler assume that including lags of the outcome for both the respondent and his/her friend is sufficient, and further they do not control for common environmental factors. Cohen-Cole and Fletcher (2008a) show that adding controls for environmental factors eliminates any detectable social network effects for obesity, and Cohen-Cole and Fletcher (2008b) show more generally that these parsimonious models will produce social network effects even in outcomes where none are expected to exist, such as for height.

instruments are adequate, though, as they are observable or correlated with observables at the time of friendship selection.

Calvó-Armengol et al. (2009) and Patachini and Zenou (2010) have extended the literature by using a network fixed effects approach in their examination of peer effect in education outcomes. Adolescents are assumed to choose among mutually exclusive networks of friends. Within these networks, their best friends (based on friendship nominations) are used as the peer exposure and their model of behavior controls for network fixed effects. The maintained assumption with this approach is that adolescents endogenously choose a friendship group, but within that group, actual "best friends" are random, an assumption that is verified for observables. However, in many cases, social networks are sufficiently dense that a large fraction of a school's students are contained in the largest network, raising questions about the practical differences between school and network fixed effects. Patachini and Zenou (2010) also use the outcomes of friends' friends (once removed in the network) as instruments.

All of these studies rely on information about the individual and their friends in order to identify the effect of friend's behavior. Whether identification is based on controlling for lagged outcomes, instrumenting for friends attributes or controlling for network fixed effects, all of these studies use variation across individuals who are in the same social environment and so reasonably may have contributed to that variation through their own choices. In the next section, we develop a simple model of friendship formation and demonstrate circumstances under which consistent estimates of the effect of friends' health behavior on own health behavior can be uncovered, and show that identification in the presense of correlated unobservables requires an exogenous shock in exposure to potential friends who exhibit certain behaviors. Following the literature on peer effects, we propose that across cohort variation within schools can provide this exogenous variation in exposure to health behaviors and demonstrate empirically that health behaviors vary substantially more across cohorts than student attributes, like race or parental education, evidence consistent with our identification strategy.

Identification Strategy

In this paper, we seek to estimate the effect of friends' health behaviors while overcoming several of the key empirical obstacles that we outlined above. The primary focus of our analysis is to address selection into friendships and other social relationships based on a students unobservables, but our fixed effect strategy also addresses reflection, at least on

observables, in that students in the same cluster have friends with the same observables. In addition, in our follow-up analysis, we examine the on-set of health behaviors where reverse causality is unlikely to be a concern.

The intuition behind our approach is that we seek to form comparison groups based on information in the data that describes the friendship *options* of students as well as the students' *choices* of friends (given these options) following the premise that individuals who make similar decisions or have similar outcomes when facing the same set of options likely are very similar on both observable and unobservable attributes. The beginning of this section illustrates this intuition, the next two subsections derive formal results.

We begin with a slight modification to the relatively straightforward linear-in-means model of social interactions (Case and Katz 1991, Manski, 1993; Moffit, 2001; Brock and Durlauf, 2001) by restricting social interactions to arise from a subset of individuals "friends" within a social environment (or school s) and dividing the unobservable into two components: an unobservable that also affects friendship choice ε_i and an orthogonal unobservable error that does not enter the friendship choice model μ_i.[5] Specifically, we consider the following empirical model:

$$H_{is} = \beta_s + \left(\frac{1}{n_i}\sum_{j\in\Omega_{is}} H_{js}\right)\beta_1 + \left(\frac{1}{n_i}\sum_{j\in\Omega_{is}} X_j\right)\beta_2 + X_i\beta_3 + \varepsilon_{is} + \mu_{is} \qquad (1)$$

where H_{is} indicates a particular health behavior, such as tobacco consumption, of individual *i* in a broad social environment or school *s*, X_i contains the individual's observable attributes, n_i is the number of friends of person *i* or more generally the number of friendship choices, Ω_{is} defines the set of individual *i*'s friends in *s*, H_{js} and X_{js} indicate the health behavior and observable attributes of individuals within Ω_{is}, and β_s is a school fixed effect.

[5] An alternative specification might involve a single unobservables each for determining health behavior and friendship outcomes. The specification is equation (1) is equivalent to such a model with the imposition of one restriction. We start with a model where the composite unobservables in equation (1) and a friendship formation model, $\tilde{\mu}_{is}$ and ε_{is}, are correlated, and then we can define μ_{is} as $\tilde{\mu}_{is} - E[\tilde{\mu}_{is} \mid \varepsilon_{is}]$ where we assume that the $E[\tilde{\mu}_{is} \mid \varepsilon_{is}] = \alpha_0 + \alpha_1\varepsilon_{is}$ so that the composite error $\tilde{\mu}_{is}$ depends upon the uncorrelated disturbances μ_{is} and ε_{is} and α_1 is simply initialized to one in the health behavior model and generality is maintained by allowing ε_{is} to enter the friendship formation model in a general manner.

As Manski (1993) demonstrates, even without the correlations in social networks that are caused by sorting into and within networks based on unobservables, e.g. ε_{is} orthogonal to $\left(\frac{1}{n_i}\sum_{j\in\Omega_{is}} X_j\right)$, this model is intrinsically unidentified.[6] This occurs because group member characteristics that might explain the health of group members *j* and so act as instruments for health behavior cannot be excluded from the second stage regression for the health behaviors of *i* because these attributes may just as reasonably directly influence *i*'s behaviors (the reflection problem).[7] [8]

Our identification strategy is to sort students into clusters *c* based on comparing similar students who faced similar friendship options and made similar friendship choices. This sorting is based on both observable (to the researcher) and unobservable characteristics. Following the standard selection argument: if two individuals make similar choices and differ on observables, then they are expected to differ on unobservables, as well (Heckman, 1976). Similarly, if two individuals are the same on observables and make similar choices, they are expected to be quite similar on unobservables (Altonji, Elder, and Taber 2005). Therefore, as argued by Dale and Krueger (2002) and Fu and Ross (2010), the inclusion of fixed effects for such clusters should assure that we are comparing students who are similar on both observables and unobservables, which weakens the correlation between peers' behaviors and a student's unobservable characteristics. Further, since all students in a cluster should have similar observable characteristics, the inclusion of the fixed effect also captures the observables associated with the students' peers while allowing the effect of behavioral differences within a cluster to identify the effect of friends' behavior on individual behavior. This feature of the approach solves the

[6] By this we mean that there is insufficient information in the regression to estimate uniquely the parameters of interest (β_1 in particular).

[7] For example, if one observes clustering of criminal behavior among friends whose parents have less education, even after controlling for all possible individual and environmental factors that might explain such clustering available in the data, we still cannot conclusively determine whether the clustering is caused because having friends whose parents have less education contributes to criminal behavior or individuals whose parents have less education are more likely to engage in criminal behavior and such criminal behavior influences the behavior of the individual's friends. See Brock and Durlauf (2001, 2006) for recent methodological progress on this problem.

[8] As noted by Sacerdote (2001) and Bayer and Ross (2008), when social network effects are determined in part by unobservable characteristics, even random assignment cannot solve this identification problem. While random assignment breaks the correlation between the health behavior *i*'s peers and *i*'s unobservable characteristics, the coefficient estimate on the behavior of peers is a composite of both the direct effect of peer's behaviors and the effect of peers' unobservable characteristics.

empirical problem outlined above and isolates the causal effect of student behaviors on the behavior of their friends from the effect of observable friends' attributes.

Specifically, define a cluster of individuals *c* in the same school who are observationally equivalent on X_i and choose observationally equivalent friends based on X_j. This structure implies that the individual and friendship group observables are the same within a cluster so that the contribution of the variables that determine clusters to individual's health behavior are constant within cluster or

$$\beta_s + \left(\frac{1}{n_i} \sum_{j \in \Omega_{is}} X_j \right) \beta_2 + X_i \beta_3 = \beta_s + \left(\frac{1}{n_i} \sum_{j \in \Omega_{ks}} X_j \right) \beta_2 + X_k \beta_3 \qquad (2)$$

for all $i, k \in c$. Further, we assume that the models that define selection over friendships on health behaviors and on observable attributes depend monotonically on the same observable vector of attributes X_i and the same single index unobservable ε_{is}. This assumption is central to our identification strategy. Without monotonicity, multiple values of the unobservable might be consistent with the observed friendship choices for observationally equivalent individuals. With monotonicity, at least on the unobservable, individuals who face the same friendship options based on the available social network (*s*) and make the same choices should have similar values on their unobservable because if they differed substantially on the unobservable they would likely have made different friendship choices.

Specifically, we can define ρ_c as a cluster fixed effect where based on the discussion in the preceeding paragraph

$$\rho_c \approx \beta_s + \left(\frac{1}{n_i} \sum_{j \in \Omega_{is}} X_j \right) \beta_2 + X_i \beta_3 + \varepsilon_{is} \approx \beta_s + \left(\frac{1}{n_i} \sum_{j \in \Omega_{ks}} X_j \right) \beta_2 + X_k \beta_3 + \varepsilon_{ks} \qquad (3)$$

Further, based on the construction of μ as an idiosyncratic disturbance, $E[\mu_{is} \mid \rho_{c_i}] = 0$ and substituting equation (2) into equation (1) yields

$$H_{ics} \approx \left(\frac{1}{n_i} \sum_{j \in \Omega_{is}} H_j \right) \beta_1 + \rho_c + (\mu_{is} - \bar{\mu}_c) \qquad (4)$$

where $(\mu_{is} - \bar{\mu}_c)$ represents the deviation of the right hand side expression in equation (3) from the average of this expression for all individuals in cluster *c*, $\bar{\mu}_c$.

Next, in order to understand the circumstances when our fixed effect estimator will yield consistent estimates of the effects of friendship networks, we develop a partial equilibrium model of friendship formation and use the friendship model to examine the properties of the specification in equation (4)

Partial Equilibrium Model of Friendship Formation

We begin this subsection by repeating equation (1)

$$H_{is} = \beta_s + \tilde{H}_{is}\beta_1 + \tilde{X}_{is}\beta_2 + X_i\beta_3 + \varepsilon_{is} + \mu_{is} \quad (5)$$

where we define $\tilde{H}_{is}$ and $\tilde{X}_{is}$ as $\frac{1}{n_i}\sum_{j\in\Omega_{is}} H_{js}$ and $\frac{1}{n_i}\sum_{j\in\Omega_{is}} X_j$, respectively, restricting H_{is} to only take on the values of 1 (healthy) or 0 (unhealthy) and X_i to only take on the values 1 (good) or 0 (bad) where the good type is defined agnostically as the type that is more likely to exhibit healthy behavior, and without loss of generality assume that β_2 and β_3 are non-negative.[9] Further, we assume that μ_{is} is an idiosyncratic error so that

<u>*Assumption 1*</u>: $E[\mu_{is} \mid \tilde{H}_{is}, \tilde{X}_{is}, X_i] = 0$

Now we define the likelihood of observing a specific health behavior H_{is} and type X_i for a selected friend by the following general set of functions

$$Pr_{is}\left[X_j = x, H_j = h | X_i, \varepsilon_{is}, \pi_{is}, j \in \mathrm{K}_i\right] = f_{sxh}(X_i, \varepsilon_{is}, \pi_{is}) \quad (6)$$

where π_{is} is an additional unobservable that does not enter equation (5), but influences friendship formation over health behaviors. The function f_{sxh} is defined over the four combinations of the outcomes for X and H and can vary across schools s since the social environment varies across schools. The four probabilities must sum to one for a given school for any value of the functions' arguments because they are probabilities.

We assume that the probabilities of having a friend who is of good type and who exhibits healthy behavior are not directly influenced by own health behavior (Assumption 2), are monotonic in the individual's unobservable attributes that influence health behavior (Assumption 3), and that additional unobservable attributes exist that have a monotonic influence on friendship formation concerning health behavior, but have no influence on either own health behavior or friendship formation over other friendship attributes (Assumption 4). While the

[9] See Brock and Durlauf (2001, 2006) for an alternative identification approach for the reflection problem that applies when behavior is discrete.

unobservables might be correlated with X_i, some variance must remain in the friend's health behabvior unobservable term that does not enter own health behavior after conditioning on X_i. These assumptions can be summarized as follows

Assumption 2: $\frac{\partial f_{s11}}{\partial H} = 0, \frac{\partial f_{s10}}{\partial H} = 0, \frac{\partial f_{s01}}{\partial H} = 0, \frac{\partial f_{s00}}{\partial H} = 0.$

Assumption 3: $\frac{\partial f_{s11}}{\partial \varepsilon} + \frac{\partial f_{s10}}{\partial \varepsilon} > 0$ *and* $\frac{\partial f_{s11}}{\partial \varepsilon} + \frac{\partial f_{s01}}{\partial \varepsilon} > 0.$[10]

Assumption 4: $\frac{\partial f_{s11}}{\partial \pi} = -\frac{\partial f_{s10}}{\partial \pi} > 0, \frac{\partial f_{s01}}{\partial \pi} = -\frac{\partial f_{s01}}{\partial \pi} > 0$, *and* $Var[\pi_{is} \mid X_i] \neq 0$

While Assumption 3 will be maintained throughout, we will examine the implications of relaxing Assumption 2 in the next subsection by allowing own health behavior to influence friendship formation over friends' health behavior. Assumption 4 is designed to capture the across cohort variation described in our identification strategy. Our maintained assumption is that membership in a cohort is based on age and so exogenous conditional on school, and so is not directly associated with own health behavior, except of course through the well-known age-gradient in unhealthy behaviors such as smoking and drinking. Further, cohort membership creates a shock to the health behavior composition of potential friends while leaving the exogenous attributes of potential friends relatively unchanged.

Now, we define a cluster *c* as all students in a school who are of the same type, have the same number of friends, and make the same friendship choices over "friendship type".

Definition 1: *A cluster c in school s is defined so that* $X_{is} = X_{ks}$, $n_i = n_k$ *and* $\frac{1}{n_i}\sum_{j\in\Omega_{is}} X_{js} = \frac{1}{n_k}\sum_{j\in\Omega_{ks}} X_{js}$ *for all i and k in cluster c and their exist no individuals l outside of cluster c where* $X_{is} = X_{ls}$, $n_i = n_l$ *and* $\frac{1}{n_i}\sum_{j\in\Omega_{is}} X_{js} = \frac{1}{n_l}\sum_{j\in\Omega_{ls}} X_{js}$.

Our first important result is that the bias in our estimate of β_1 in equation (4) limits to zero as the number of friends, or more generally the number of friendship choices, becomes large. In our empirical work, the number of choices made is substantially larger than the number

[10] The assumption of a positive relationship between good type and the individual's friendship formation propensity y_{is} is made without loss of generality because one can reverse the relationship by designating healthy behavior as unhealthy. However, once this assumption is made, the sign of the relationship between y_{is} and having friends who exhibit healthy behavior is meaningful. If this relationship is positive, then one's type has the same effect on health behavior composition of friendships as it has on composition of friends over type, and this assumption cannot be undone by reversal because the definition of what individual type means is nailed down by β_3 and the coefficient of one on ε_{is} in equation (5)

of friends because students choose over many demographic attributes including race, ethnicity, and maternal education. Further, later in the paper, we demonstrate that our results are robust to focusing on the subsample of students with the largest number of friends.[11]

Proposition 1: Under Assumptions 1 through 4 plus Definition 1, the bias arising from estimating the cluster fixed effects model in equation (4) limits to zero as n_i becomes large for all i in the sample.

Proof: See Appendix

A key derivation in Proposition 1 is that the bias in the cluster fixed effects estimate of β_1 can be written as

$$\phi_1^c = \frac{Cov[\varepsilon_{is} - \bar{\varepsilon}_c, (\tilde{H}_{is} - \bar{H}_c)]}{Var[(\tilde{H}_{is} - \bar{H}_c)]} \tag{7}$$

where $\bar{\varepsilon}_c$ and $\bar{H}_c$ are the cluster specific mean of ε_{is} and $\tilde{H}_{is}$, respectively. The bias limits to zero because, as the number of choices becomes large, two individuals *i* and *k* can only belong to the same cluster if $\varepsilon_{is} = \varepsilon_{ks}$. However, the within cluster variation in ε cannot limit to zero while within-cluster variance of $\tilde{H}_{is}$ remains unless $\tilde{H}_{is}$ contains variation associated with π. Therefore, the non-zero variance assumption in Assumption 4 is crucial to the consistency of our estimator.

Second, even when the number of friends is small, we can show that the inclusion of cluster fixed effects reduces the bias in estimates of the effect of friend's health behavior on own health behavior with the imposition of a couple of additional assumptions. First, we create a linear projection of $\tilde{H}_{is}$

$$\tilde{H}_{is} = \lambda_s + \tilde{X}_{is}\lambda_1 + X_i\lambda_2 + V_{is} \tag{8}$$

such that $V_{is} = V(\tilde{X}_{is}, X_i, \varepsilon_{is}, \pi_{is})$. We assume that the conditional expectation of V_{is} is zero and that the conditional variance of V_{is} is less than or equal to the variance of V_{is}.

Assumption 5: $E[V_{is} \mid \tilde{X}_{is}, X_i] = 0$ and $Var[V_{is} \mid \delta_c] \leq Var[V_{is}]$.

[11] A second implicit assumption is that the number of observations or students increases more quickly than the number of friendship choices. This assumption is required in order to assure consistency in the fixed effect estimates. Later, in our empirical work, we demonstrate that key findings on behavior on-set are robust in subsamples that focus on individuals in friendship choice clusters that contain a larger number of students.

The first part of Assumption 5 implies that

$$E[\tilde{H}_{is} \mid \tilde{\tilde{X}}_{is}, X_i] = \lambda_s + \tilde{\tilde{X}}_{is}\lambda_1 + X_i\lambda_2 \quad (9)$$

This restriction is essentially a law of large numbers style assumption where we assume that the average of this residual is zero over repeated realizations of $\tilde{H}_{is}$ and $\tilde{\tilde{X}}_{is}$ for a given X_i. This assumption would be standard if $\tilde{\tilde{X}}_{is}$ did not depend upon ε_{is}. The second half of Assumption 5 is something that can be theoretically violated in principle, but in practice we expect that variances will decline after conditioning on additional information, and we also directly verify this assumption for estimates of V_{is} using our data.

Proposition 2: Under Assumptions 1 through 5 plus Definition 1, the bias arising from estimating the cluster fixed effects model in equation (4) has the same sign and is smaller than the bias that arises for the OLS model described in equation (5).

Proof: See Appendix

In the appendix, we also consider a situation where the shock to friendship formation over health behavior π_{is} also affects friendship formation over the exogenous attributes, which might be the case if we were considering two individual in the same grade who had unobservable differences (associated with themselves rather than the cohort) that lead them to choose friends with different health behaviors and so likely also caused them to choose friends with different attributes. We show that Proposition 1 may not hold when we relax Assumption 4 along this margin, and in fact the sign of the bias may be reversed relative to OLS when we control for friendship cluster FE's that contain variation that does not satisfy Assumption 4, such as within grade variation in friends' health behaviors.

Simultaneity of Health Behavior and Friendship Sorting Model

In this section, we extend the friendship formation function so that friendship formation over health behavior depends upon one's own health behaviors creating true simultaneity between one's own health choices and the selection of friends based on their health choices. Specifically, we relax Assumption 1 so that own health behavior influences the likelihood of having friends who exhibit a health behavior, but do not allow own health behavior to affect friendship formation over the observable attributes. So

$$Pr_{is}\left[X_j = x, H_j = h | H_i, X_i, \varepsilon_{is}, \pi_{is}, j \in \mathrm{K}_i\right] = f_{sxh}(H_i, X_i, \varepsilon_{is}, \pi_{is}) \quad (10)$$

with

Assumption 6: $\frac{\partial f_{s11}}{\partial H} = -\frac{\partial f_{s10}}{\partial H} > 0$ *and* $\frac{\partial f_{s01}}{\partial H} = -\frac{\partial f_{s01}}{\partial H} > 0$.

Therefore, the idiosyncratic error μ_{is} does not have a conditional expectation of zero because it influences the health behavior of friends $\tilde{H}_j$ through one's own health behavior, and the bias in the coefficient on friend's health behaviors contains a second term φ_1.

Proposition 3: Under Assumptions 1 and 3 through 6 plus Definition 1 the cluster fixed effects model estimate of the effect of friends' health behavior limits to a reduced form estimate that is the sum of β_1 and φ_1 as n_i becomes large for all i in the sample.

Proof: See Appendix.

Proposition 3 illustrates why our fixed effects estimator only addresses bias from correlated unobservables, but not endogeneity where smoking causes someone to select friends who smoke. If β_1 is zero, it is still possible for φ_1 to be non-zero, because high values of μ_{is} will increase the likelihood of smoking and as a result lead to a larger numbers of smoking friends other things equal, V_{is}. We cannot address this limitation theoretically, but later in the paper we examine an empirical model on the onset of health behaviors where reverse causality should not be a serious concern. At that time, we will present the methods for conducting the required two stage estimates.

Friendship Data

In order to accomplish our research goals, we use the only available national dataset containing rich friendship network information as well as health behaviors, the National Longitudinal Study of Adolescent Health (Add Health). The Add Health is a school-based, longitudinal study of the health-related behaviors of adolescents and their outcomes in young adulthood.[12] For this paper, we focus on the In-School data collection, which utilized a self-administered instrument to more than 90,000 students in grades 7 through 12 in a 45- to 60-

[12] In short, the study contains an in-school questionnaire administered to a nationally representative sample of students in grades 7 through 12 in 1994-95 and three in-home surveys that focus on a subsample of students in 1995 (Wave 1), and approximately one year (Wave 2) and then six years later (Wave 3). The fourth wave of the survey was collected in 2008/9. The study began by using a clustered sampling design to ensure that the 80 high schools and 52 middle schools selected were representative of US schools with respect to region of country, urbanicity, size, type, and ethnicity. Eligible high schools included an 11th grade and enrolled more than 30 students. More than 70 percent of the originally sampled high schools participated. Each school that declined to participate was replaced by a school within the stratum.

minute class period between September 1994 and April 1995. The questionnaire focused on topics including socio-demographic characteristics, family background, health status, risk behaviors, and friendship nominations. In particular, each student respondent was asked to identify up to 10 friends (5 males, 5 females) from the school's roster. Based on these nominations, social networks within each school can be constructed and characterized, linking the health behaviors of socially connected individuals.

Of the nearly 90,000 students in the schools originally surveyed, several reductions in the sample size were made in order to construct the analysis sample. First, nearly 4,500 students did not have individual identification numbers assigned. Nearly 12,000 students did not nominate any friends and 5,000 individuals nominated friends who were not able to be linked with other respondents due to nominations based on incomplete information ("nicknames" rather than names, or the nominated friend did not appear on the Add Heath school roster, etc.) These issues reduced the sample to approximately 66,000 respondents. Appendix Table 1A presents an analysis of the correlates associated with individuals being dropped from the sample for these reason discussed above, as well as additional sources of selection arising from the empirical specification discussed below.[13]

In this paper, our main focus is on individuals with same-sex/same-grade level friends, which reduces the sample to approximately 58,000 students.[14] One reason to focus on same-sex friends is that romantic relationships may be nominated as "friends". In addition, most previous studies of friendship networks also limit the network definition to same-sex friends. We limit our analysis to same-grade friends in order to use cross-cohort (grade) variation in friendship opportunities and choices, as we describe below. In order to retain sample size, we impute missing covariates, such as maternal education, and control for missingness, but we do not impute missing outcomes.

[13]. Briefly, race, gender, family structure, and missingness on other variables predicts sample selection in to the original 66,000 observations to some extent, however health behaviors are not robust important predictors. In regards to same-sex/same-grade friendship nominations, the likelihood of making such nominations increases by grade and is smaller for more advantaged students. We find that the proportion of smokers in the grade (potential friends) is not related to these nomination patterns, however, individuals with drinking grademates are slightly more likely to nominate same-grade/same-gender friends (a 10 point increase in grademates drinking is associated with a 1 percentage point increase in the probability).

[14] Of the 66,000 students, 4,300 do not nominate any same grade friends and 4,100 do not nominate any same-grade/same-gender friends (that is, they nominate same grade friends but no same-grade/same gender friends).

Table 1 presents descriptive statistics of the analysis sample and shows that approximately 34% of the sample reports smoking and 54% of the sample reports drinking alcohol. The average adolescent nominates 2.4 same-sex friends. In Table 2 we present the distribution of friends' health behaviors in the data. Friendship networks include considerable variation, including individuals who have no smoking/drinking friends through individuals who have all smoking/drinking friends.

Evidence of Variation in Friendship Options

As we demonstrate above, identification of the effect of friend's health behavior requires a shock in exposure to potential friends with specific health behaviors. In our empirical analysis, we control for fixed effects associated with similar students who make the similar friendship choices on student attributes, but because they belong to different cohorts of the same school draw groups of friends who exhibit differing health behavior. That is, the dataset contains multiple cohorts within each surveyed high school, which allows us to combine our friendship type fixed effects with the use of cross-cohort, within-school variation and in doing so are able to compare students who face similar friendship options (are in the same school) and make similar friendship choices. This extension relies heavily on the assumption that individuals who attend the same school, but different grades, have essentially the same "types" of friendship options.

To what extent do students in the same school face similar friendship options? Using the Add Health data, we show below in Table 3 that controlling for school and grade effects can predict over 95% of the variation in racial composition of potential friends (classmates) in the data. Likewise, controlling for school and grade predicts 93% of the variation in peers' maternal education level and 96% of the variation in classmate nativity. These findings suggest that students in different grades but who attend the same school have very similar friendship options based on race and family background of peers.

In addition, there is substantially more variation across cohort, within schools in unhealthy behaviors. Using the same regression analysis, our data show that we only predict 77% of peer smoking rates and 81% of peer drinking rates. Thus, these results suggest that there is substantial variation in exposure to health behaviors of potential friends (classmates) even within school, while at the same time the friendship options based on race, maternal education, and nativity is nearly identical for students across grades within the same school. We use these features of our data to make comparisons within schools of students who face similar

environments in terms of friendship opportunities and make similar friendship choices over attributes, but have different friendship outcomes over health behavior and unhealthy behavior outcomes.

Empirical Specification

Our friendship clusters are based on students in the same school choosing sets of friends with very similar demographic attributes. As there is evidence that adolescents have strong preferences to befriend classmates based on age, gender, and race (Mayer and Puller 2008; Weinberg 2008), we create our "individual type-friendship type clusters" by focusing primarily on those attributes. Given a limited sample, there is clearly a trade-off between how restrictive we make our definitions of observationally similar individuals and of same friendship types. We begin by placing the most weight on obtaining very specific "friendship-type" clusters. The reason behind this focus in that most of our demographic variables are binary and so after controlling for individual-type on those variables very little information is left that can be used in our specification tests in order to examine whether peer attributes can explain predetermined student attributes. For example, we examine whether peer attributes can explain student race or ethnicity in a model that only controls for within school friendship types. However, we also examine model specifications that include the student's race (white, black, Hispanic, and Asian) and whether their mother is a college graduate in the creation of individual type-friendship type clusters, and then for years of maternal education we can test whether peer within cluster variation can explain a student's own maternal education.

The friendship clusters are based on the following exogenous characteristics of chosen friends, including (1) race (black vs. Hispanic vs. white vs. Asian vs. other) (2) maternal education (no college vs. some college vs. college graduate) (3) family structure (living with mother vs. not living with mother) and (4) nativity (native vs. foreign born). Specifically, the number of friends chosen from each characteristic is used in the cluster. Importantly, our clusters are quite flexibly created, such that an individual who chooses five black friends is in a different cluster than an individual who chooses four black friends.[15] In yet another refinement of our

[15] As an example, friendship cluster 15 could be created based on nominating four friends such that: friend A is white, has a college educated mother, lives with his mother, and is native born; friend B is white, has a mother with some college, lives with his mother ,and is native born; friend C is white, has a college educated mother, lives with his mother, and is foreign born; friend D is black, has a college educated mother, lives with mother, and is native

cluster approach, in some analyses we also include grade levels-pairs within the clusters, so that 7^{th} and 8^{th} graders are compared to each other (and $9^{th}/10^{th}$ and $11^{th}/12^{th}$) in order to move closer to the thought experiment described in the introduction.

In our final model, as discussed above, we restrict our comparisons to students in different grades who are observationally equivalent on *X* and chose the same friendship set on the *X*'s. Friends' health behaviors are based on own-grade friendships, and so these estimates are based entirely on comparisons across cohorts. Specifically, one student's friends' health behavior could not vary from another student's in the comparison group because one student selected a given student and another selected away from the same student. In order to accomplish this, we randomly choose only one student in each grade from each friendship type cluster so that the estimated effect of peer behavior cannot be identified off of within grade variation. In these estimates, the substantial differences in health behavior across cohorts provide the shock to the health behavior of potential same-grade friends that identifies the effect of friends on health behavior. In practice, we present the average of the parameter estimates resulting from several random draws of one individual in each cluster per cohort.

Finally, the rich structure of friendship type clusters, as outlined above, will create singleton clusters of students—those students who have unique or "unusual" friendship preferences. These singleton clusters will, implicitly, not contribute to the identification of the network effects estimates, as there will be no within-cluster variation to exploit. Tables 2A and 3A examines the significance of excluding the variation associated with these observations from our estimates of the effects of friends health behaviors. While we find some evidence that attrition on this dimension varies with observable attributes, the estimated relationship between smoking and drinking status and placement in a single cluster is fairly small (Table 2A). In addition, we repeat the substantive analyses presented below for subsamples excluding observations associated with singleton clusters and their exclusion has no effect on the pattern of estimates observed (Second row of Table 3A).

Evidence of Friendship Selection

We can partially test the validity of our approach by examining whether students seem to be sorting into specific friendship patterns within our friendship clusters. Specifically, we test

---

born. Cluster 16 could be identical except the individual nominated four white friends instead of three white friends and one black friend; Cluster 17 could be identical to cluster 15 except all the nominated friends are native born.

whether a student's own observable attributes correlate with the attributes of their friends within student clusters. Following the logic of Altonji, Elder, and Tabor (2005), if individuals do not sort on observables into friendships within clusters, it is very unlikely that they have sorted based on unobservable characteristics. For example, if we find no evidence of additional correlation between an individual's own parental education and the parental education of their friends after conditioning on the average level of correlation for all students in this cluster, which might include broader educational categories, then it is unlikely that students are sorting based on unobservable characteristics like the parents' involvement with the students' education or the parents' educational and academic expectations since those unobservable characteristics are likely correlated with parental education. Similar diagnostic tests have been used elsewhere (Bayer, Ross and Topa 2008; Bifulco, Fletcher and Ross 2011).

In Table 4A, we present evidence from these diagnostic tests. Each set of rows examines the correlation between a different "outcome" (individual-level characteristic) and friend's characteristics. Columns add controls from left to right. The first column and row shows the correlation between whether an individual is of Hispanic ethnicity (vs non-Hispanic) and the average of his or her friends' maternal education levels (-0.03). Column 2 controls for school fixed effects and reduces the coefficient by 1/3, but the estimated effect is still sizable and statistically significant. Column 3 controls for school by cluster fixed effects and reduces the coefficient to $1/10^{th}$ the size of the baseline regression, and Column 4 yields similar estimates after adding grade-pairs to the clusters so that $7^{th}/8^{th}$, $9^{th}/10^{th}$, and $11^{th}/12^{th}$ graders are compared. Column 6 adds individual characteristics to the cluster definition, including race and whether the student's mother graduated from college, and Column 7 estimates the Column 6 model selecting one observation per cohort per cluster and weighting clusters back up to their original size for comparability to Column 6, though the model is not identified for these two columns for this outcome (student race). Similar results arise for whether the individual is white in Row 2.[16]

In Row 3, we examine the correlation between own-maternal education and the average maternal education of friends. Here, the correlation is quite high—0.33—in the baseline specifications. Again, the inclusion of school fixed effects leads to only a moderate reduction in the coefficient estimate. However, when we add school X cluster fixed effects in column 3, the

[16] The estimated effects in OLS for explaining whether an individual is black is small relative to the standard error in our cluster fixed effect estimates and so a counterfactual based on whether the student is black is non-informative.

coefficient estimate is reduced by more than two-thirds, but is still statistically significant. Finally, we include individual characteristics in Column 5 in the clusters definitions, and the correlation between own and friends' maternal education falls to 0.01 and is not statistically significant. The one observation per cohort sample results in Column 6 indicate a slight increase in the magnitude of the estimates as compared to Column 6, but the effects are still statistically insignificant and substantially smaller than the estimates in the school fixed effects model.

In a second set of balancing tests (Table 4B), we examine the correlations between individual characteristics and friends' health behaviors in order to further assess our ability to control for observables and unobservables in our estimation strategy. In the first row, we show that maternal education is highly associated with friends' drinking behaviors. However, when we control for clustering, the coefficient is reduced by over 90% and is no longer statistically significant. In row 2, we find similar evidence from the correlation between maternal education and friends' smoking behaviors. In row 3, we find that individuals with highly educated mothers are more likely to have friends with caring mothers. However, as we add cluster fixed effects in the final column, this correlation is reduced over 80% and is no longer statistically significant. This result is a strong test of the adequacy of our clusters, as maternal caring might be a typically unobserved characteristic that researchers would worry is not completely captured in our clusters.[17] In two of the three cases, the effect size increases when we shift to the one per cohort sample, but as before the estimated effects are still insignificant and small relative to the school fixed effect estimates. These findings are suggestive evidence that our cluster controls are significantly reducing endogeneity bias associated with students choosing their friends both overall and when compared to school fixed effect models.

Finally, in Table 4C, we examine a placebo variable, chest pain, where we would not expect chest pain in one student to have a causal effect of chest pain for a second. While friend's chest pain correlates with student chest pain even after including student fixed effects, this effect completely disappears once we have controlled for the type of friendship choices made by students.

[17] Of course, we will control for maternal caring in our results, so any residual correlations in unobservables between the respondent and his friends will be net of these controls and the cluster fixed effects

Results

Table 5 presents estimates for adolescent smoking where same-sex/same-grade friends are used to define the friendship network. In Column 1, the baseline results suggest that increasing the share of friends who smoke by 10 percentage points would increase own-smoking by nearly 3.9 percentage points. In Column 2, we follow some of the previous literature and control for high school fixed effects; however this only reduces the coefficient from 0.388 to 0.368 for friends' smoking. In Column 3 we do not use school fixed effects, but instead use our friendship cluster fixed effects. As discussed above, we create cluster fixed effects based on several aspects of the respondent's friendship nomination patterns, including (a) number of nominations (b) race of nominated friends (white vs. black vs Hispanic vs. Asian vs. other race), maternal education of nominated friends (college graduate vs. non college graduate), whether friend is native born, and whether friend lives with his/her mother. With the inclusion of cluster fixed effects, the coefficient estimate mirrors that of the school fixed effects results (column 1 vs. column 3) declining from 0.39 to 0.37 and little reduction in the estimates is observed. However, when we control for school by cluster fixed effects in column 4 and so control for same friendship-type choices given the same friendship opportunity set, we observe a substantially larger decline in the estimated to 0.31. The last three columns limit comparisons to adjacent grades (7/8, 9/10, 11/12), incorporate same observables into the cluster definitions and restrict the sample to one observation per cohort in turn.

All of the estimates based on within friendship cluster comparisons fall between 0.295 and 0.315. The lowest estimates are associated with models that contain variation within grade, which is consistent with the possibility that within grade-within cluster variation could bias estimates in the opposite direction from the bias in OLS or school FE models. Focusing on our preferred estimates using the one observation per cluster per cohort, we see less than a 10% reduction relative to our school fixed effect estimates with our inclusion of individual-friendship type fixed effects. Significantly, all reductions due to cluster fixed effects including those based on estimates that include variation within cluster and cohort are less than 20%, which is quite small relative to the declines in estimates across the same model specifications for our balancing tests where the declines are typically on the order of 75 to 90 percent. As discussed above, as we control for richer cluster definitions, the sample size used to identify the coefficients is reduced due to "singleton clusters". As mentioned earlier in Appendix Table 3A, we show that the

change in composition is not the explanation for our results by estimating the baseline results in Table 5 using the non-singleton sample across columns.

Table 6 examines drinking behaviors. Baseline results in column 1 suggest that a 10 percentage point increase in friends' drinking is associated with a 3.3 percentage point increase in own-drinking. Like the results for smoking, school fixed effects (added in column 2) reduce this association by a modest amount to 3.0. Using the same cluster definition as in smoking, the results using friendship-cluster fixed effects (but not school fixed effects) in column 3 the coefficient is reduced slightly, suggesting that increasing friends' drinking by 10 points will increase own drinking by 3.2 percentage points. As before, when we control for school by cluster fixed effects in column 4, our estimated effect falls to about 2.5 percentage points. In our preferred specification, the one per cohort sample, the estimated effect is 2.8 percentage points, and our best estimate of causal effects is only 6 percent below the school fixed effect estimates.

The bottom line of these findings is that our estimates suggest only minimal bias from correlated unobservables in school fixed effect estimates of friends' smoking or drinking on a student's own behavior. After controlling for variables like gender, race, maternal education, gender, family structure and nativity, factors that are observed in many educational samples, friendship formation within a grade appears to be relative random at least in terms of unobservables that have a systematic influence on health behaviors. This finding has relevance for the growing empirical literature on networks (Calvo-Armengol et al. 2009, Trogdon et al. 2008). Most studies in this area are identified by network fixed effects assuming that individual links within each network are formed randomly. At least after controlling for the role of readily observable demographics in friendship formation, our research is supportive of this assumption for smoking and drinking. However, the assumption of exogenous friendship formation may not be valid for other behaviors or outcomes, especially in light of our spurious results on chest pain for the school fixed effects model.

Two Stage Models and an Application to Longitudinal Data

If our estimates are to capture the causal effects of friends' behaviors, an additional assumption is required that a student's own smoking behavior does not directly cause the student to form friendships with students who smoke. We can neither directly test whether this assumption holds or credibly argue that the assumption is reasonable. One reasonable option is to look at models of the on-set of health related behaviors with a longitudinal sample where reverse

causality is less of a concern. However, quite frequently, longitudinal datasets are substantially smaller than cross-sectional samples, and the Add Health survey is no exception with the initial in-school survey attempting to interview the full population of each sampled school, while only a fraction of these students are followed over time.

However, the analysis described above generates information on ε_{is} associated with all individuals who are in the same cluster *c*. Therefore, we can combine this information with the longitudinal data for examining the onset of health behaviors, such as drinking and smoking, which by construction could not have caused the earlier friendship choices. The classic threat to identification in studies of the on-set of health behaviors is that the same set of unobservables that caused the student to select their friends also lead to later smoking, but in our case the information generated on each cluster can provide a control for those correlated unobservables.

Specifically, we estimate the fixed effect associated with each school, student type and friendship pattern cluster, and that fixed effect provides an estimate of the unobservable for each individual within the cluster. The fixed effect estimate for each cluster is an average across all cohorts including the cohort from which a student is drawn, which in small samples can lead to an overcorrection from the fixed effects and an understatement of the true estimate. The standard solution to this problem is to calculate the control, in this case the fixed effect, omitting the information that creates this correlation, in this case the information from the same cohort. Guryan, Kroft and Notowidigdo (2009), however, show that the estimated coefficient on this corrected control may be negatively biased because an individual's own information is negatively correlated with the average of the relevant population (that omits the individual) from which that individual draws peers. In our context of a school fixed effects model, their recommended control for mitigating this bias is simply the average fixed effect estimate across all individuals in the school except for the cohort-cluster to which the subject individual belongs. The resulting estimating equation would be

$$H_{ics} \approx \left(\frac{1}{n_i} \sum_{j \in \Omega_{is}} H_j \right) \beta_1 + \hat{\rho}_{c-i} + \hat{\rho}_{s-i} + (\mu_{is} - \overline{\mu}_c) + (\rho_c - \hat{\rho}_{c-i} - \hat{\rho}_{s-i}) \qquad (11)$$

where $\hat{\rho}_{c-i}$ and $\hat{\rho}_{s-i}$ are the estimated cluster fixed effect omitting the information on the fixed effect from a student's own cohort and the mean (weighted by the number of students) of the cluster fixed effect estimates over all clusters at a school again omitting the student's cohort contribution to their cluster fixed effect estimate.

In order to obtain the most precise estimates for our preferred specification, we collapse the data at the level of the student type, friendship choices, cohort and school in order to obtain estimated based only on cross-cohort information using the entire sample. The parameter estimates and the estimated student type-friendship choice-school FE's are weighted based on the number of students in each student type-friendship choice-cohort-school cluster. We also focus in this analysis on comparisons across all grades in a school, again in order to increase the amount of information used to calculate the FE's. Note that the fixed effects are only identified for friendship clusters in schools containing at least one student in at least two cohorts.

Table 7 presents models from the cross-sectional sample. The first two columns present both the school FE and the student type by friendship choice by school FE model estimates for the individual subsample where one individual is selected randomly per cluster. These cluster FE estimates are very similar to the cluster FE estimates in Tables 5 and 6, which controlled for adjacent grade pairs, with the smoking estimates being identical between the two models and the drinking estimate being less that 5 percent larger than the estimate that includes grade pairs in the cluster definition. The next two columns present the student weighted estimates using the collapsed data with school and school by cluster FE's. The cluster FE's estimates for these two samples are very similar and continue to indicate very little bias from correlated errors relative to the school FE estimates for smoking or drinking.

In the last two columns, we present the two stage estimates using the individual sample and the predicted cluster FE's. The first of the last two columns includes our prediction of the cluster FE that includes information from an individual's own cohort. Naturally, the estimate on the fixed effect is large and highly significant because it contains information on the individual's own smoking behavior. The inclusion of this variable reduces the effect of friend's smoking or drinking, representing a classic overcorrection that occurs with fixed effect estimates based on small samples. In the last column, we present the estimates for the fixed effect model that includes the predicted cluster fixed effect that excludes information from the individual's own cohort plus the Guryan et al. (2009) control at the school level. The Guryan et al. control is negative and significant as expected. The estimate on the predicted fixed effect, however, is very near zero and statistically insignificant. As in the single stage fixed effect estimates, there is no evidence of substantial bias from correlated unobservables for friendship formation within

schools, and the estimated effect of friends smoking or drinking is quite near to the school FE estimates.

Table 8 presents effect estimates for the onset of smoking or drinking as a function of the behavior of a student's friends at baseline. The first two columns present the OLS and school fixed effect estimates for full sample of all students surveyed at followup who did not smoke or drink, respectively, at baseline. The third column presents the school FE estimates for a sample restricted to those for which a student type by friendship choice by school FE can be estimated (students present in the cluster in that school for at least two cohorts). The fourth column adds the predicted cluster fixed effect using information from the student's own cohort and the fifth column controls for the predicted fixed effect omitting that information and for the Guryan control. The inclusion of school fixed effects erodes the OLS estimates for the friends' effect on the onset of drinking, but has little impact on the effect of friends' smoking on smoking onset. More significantly, the coefficient on the prediction of the cluster FE is effectively zero in the last two columns, and the effect of friends' behavior is relatively unchanged as compared to the school FE presented in column 3. Note that the estimate on the predicted cluster FE are zero even when the own cohort information is included because the smokers and drinkers from the first wave who helped drive these cluster predictions have been deleted from the sample in order to study onset. Similarly, the Guryan control is relatively uninformative in these models.

The central threat to identification in our models of the on-set of smoking and drinking is that the fixed effects may contain considerable noise due to the smaller number of observations used to estimate each fixed effect. Table 9 presents results based on the size of the school-student type-friendship pattern clusters used to calculate the cluster FE. The average numbers of observations associated with the fixed effect estimates are 2.5 for the subsample of small clusters and 17 for the subsample of big clusters. Obviously, these subsamples are not randomly determined because, for example, clusters involving individuals with mostly own race friends will tend to have a substantially larger number of students than clusters of students who have many friends across racial lines. Regardless, if student friendship formation patterns capture information on unobservables that influence smoking or drinking, then we should find larger, more precisely estimated parameters on the predicted fixed effects when those fixed effects are more accurately measured because they are based on more observations. However, all our results are robust across the subsamples. The estimates on the predicted fixed effects are uniformly

small and statistically insignificant, and all our results are robust to the inclusion of the predicted fixed effects. The coefficient on smoking does erode somewhat for the big cluster subsample when the additional controls are included, but we have verified that those changes are driven entirely by the inclusion of the Guryan et al. control, rather than the predicted fixed effect.

Robustness to Number of Friends

One natural concern with our identification strategy is that students only report up to a maximum of five same gender friends and many report only 1 or 2 same gender friends. While students make friendship choices over a wide variety of student attributes that are observable, one still might be concerned that our positive estimates of friendship effects are driven by students who have a very small number of friends and that in those cases the number of choices being made is insufficient to eliminate the bias from student unobservables that influence friendship choice. In Table 10, we divide the sample by the number of friends where the first panel presents estimates for smoking and the second panel presents estimates for drinking. The general pattern of results remains the same with the inclusion of school by friendship cluster fixed effects leading to reductions in estimates by less than 10 percent relative to the school fixed effects for smoking and small increases, less than 5 percent, for drinking. In terms of the magnitudes, the changes in the estimated effect size due to the inclusion of cluster FE's is actually smaller for the subsample of students with 4-5 friends for smoking and drinking in both percentage terms and absolute changes.

In Table 11, we present estimates for the effect of friends' behavior on the on-set of smoking (first panel) and drinking (second panel) for the 1-3 and 4-5 friends subsamples. Since the restricted subsample for 4-5 friends is quite small (between 200 and 300 students), we estimate a joint model for the entire sample interacting dummies for number of friends with the share of friends smoking or drinking, rather than estimate separate models for each subsample.[18] In the first two columns, we present the school FE estimates for the individual longitudinal sample. The relationship between friends' behavior and own behavior is robust in both subsamples for both drinking and smoking, and as in cross-sectional estimates of behavioral effects in Table 10 the relationship for on-set is also stronger for the subsample with 4-5 friends.

[18] It is not realistic to estimate 60-70 school fixed effects in samples with only a couple of hundred observations. The pattern and significance arising from models using split samples are qualitatively similar to the estimates in columns 3 and 4, but the magnitudes of the parameter estimates are quite unstable.

Columns 3 and 4 then present the estimates for the restricted sample. Three of the four significant findings persist in the restricted sample, but the estimate on friend's smoking for the 4-5 friend subsample is very near zero. It is important to note that this zero estimate is for a very small subsample (less than 300 observations), and that the zero estimate does not provide evidence in either support of or against the validity of our identification strategy. Regardless, the inclusion of the predicted cluster FE has very little influence on any of the estimates, reducing the effect in the three positive and significant estimates by less than three percent and changing the zero estimate for friends' smoking for 4-5 friends by only 0.002. Further, we can calculate the predicted cluster FE that does not omit own cohort for the entire longitudinal sample. When we estimate the model including those predicted FE's, the estimated effects of friend's smoking are 0.050 and 0.068 for 1-3 and 4-5 friends, respectively, and those reductions are nearly identical in magnitude across the two subsamples at about 6 percent.

Conclusions

While researchers typically examine peer effects by defining the peer group broadly, this paper focuses attention on actual friends and implements a new research design to study the effects of friend's health behaviors on own health behaviors for adolescents. The main idea is to combine a cross-cohort, within school design with controls for friendship options and friendship choices through the use of "school-student type-friendship pattern" fixed effects. We show that in the Add Health data used in this paper, there is evidence that our design is successful in narrowing down relevant comparison groups by controlling for the friendship choices and friendship options of adolescents. Our initial estimates also suggest that all results are robust to the restriction of sample to one student per cluster per cohort, which assures that the model is only identified based on comparisons of students across cohorts in the same school.

Further, we use a model of friendship formation to investigate the circumstances under which our identification strategy will provide consistent estimates. We find that our approach can be applied under quite general circumstances. For example, our model allows for a very general non-linear process of friendship selection and allows for correlation between observable attributes and unobservables that affect friendship formation. In addition, we show how to apply the information gained from our analysis to smaller longitudinal samples in order to control for correlated unobservables in models of behavior on-set where the simultaneity between own health behavior and friendship choice is unlikely to be able to explain the estimated effects of

friends' behaviors. The key assumptions required to apply this identifications strategy are that unobservable determinants of health behavior have a monotonic affect on the patterns of friendship formation and that individuals experience some type of shock in exposure to health behavior of potential friends that does not directly enter own health behavior. This shock assures that some variation remains in friends' health behavior even after eliminating variation across individuals in friendship outcomes. In our application, this "treatment" is the variation across cohorts in the exposure to friends' health behavior. Our empirical analysis is very supportive of this assumption in that we find very small variation in the demographic attributes of students across cohorts in the same school, but substantially larger variation in health behavior.

Friends' drinking and smoking appears to have a substantial impact on a student's own smoking and drinking and on the onset of smoking and drinking, and controls for correlated unobservables does little to erode the estimated effect. However, counterfactual analyses of race, ethnicity, maternal education and chest pain find that no evidence of an effect of friends' attributes remains after applying our identification strategy. There continues to be little evidence of bias in the school FE models when we focus on a subsample of students that have the larger numbers of friends or on subsamples of clusters where we observed the largest number of students, which represent circumstances where our analysis should have the greatest potential for identifying bias.

Literature Cited

Aizer, Anna, and Janet Currie. 2004. ‘‘Networks or Neighborhoods? Correlations in the Use of Publicly Funded Maternity Care in California.’’ *Journal of Public Economics* 88(12):2573–85

Altonji, J G., Elder, TE, CR Tabor. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools.” *Journal of Political Economy*, 2005 113: 151-184.

Angrist, Joshua and Kevin Lang. 2004. “Does School Integration Generate Peer Effects? Evidence from Boston’s Metco Program.” *American Economic Review*, 94(5): 1613-1634.

Bayer. P and S. Ross. Identifying individual and group effects in the presence of sorting: A neighborhood effects application. *National Bureau of Economic Research* Working Paper #12211, 2008

Bayer, P., Ross, S. and G. Topa. Place of work and place of residence: Informal hiring networks and labor market outcomes. *Journal of Political Economy*, 2008, 116 (6): 1150-1196

Ballester, C., Calvó-Armengol, A. and Y. Zenou. Who's who in networks. Wanted: the key player. E*conometrica*, 2006, 74, 1403-1417

Bearman, P., Moody J., and K. Stovel. Chains of affection: the structure of adolescent romantic and sexual networks. *American Journal of Sociology*, 2004, 110 (1): 44-91

Bertrand, Marianne, Erzo Luttmer, and Sendhil Mullainathan. 2000. ‘‘Network Effects and Welfare Cultures.’’ Quarterly Journal of Economics 115(3):1019–55.

Bifulco, R., Fletcher, J. and S. Ross. 2011. The effect of classmate characteristics on individual outcomes: evidence from the add health. *American Economic Journal: Economic Policy*

Blundell, Richard and Monica C. Dias. 2009. Alternative Approaches to Evaluation in Empirical Microeconomics. *Journal of Human Resources*, 44, 565-640.

Bramoulle Y. and B. Rogers. Diversity and popularity in social networks. Social Science Research Network, Working Paper, 2009.

Brock, William A. and Stephen N. Durlauf. 2006. Identification of Binary Choice Models with Social Interactions. *Journal of Econometrics*, forthcoming.

Brock, William A. and Steven N. Durlauf. 2001. Discrete choice with social interactions. *Review of Economic Studies*, 68, 235-260.

Brown, B. B., Dolcini, M. M. & Leventhal, A. (1997) Transformations in peer relationships at adolescence: implications for health-related behavior. In: Schulenberg, J.,Maggs, J. L. & Hurrelmann, K., eds. *Health Risks and Developmental Transitions During Adolescence*, pp. 161–189. New York: Cambridge University Press.

Calvó-Armengol, A., Patacchini, E. and Y. Zenou. Peer effects and social networks in education. *Review of Economic Studies*, 2009

Case, Anne and Lawrence Katz. 1991. The Company You Keep: The Effects of Family and Neighborhood on Disadvantaged Youths. NBER Working Paper 3705

Christakis N, Fowler J. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 2007, 357: 370-9.

Christakis N, Fowler J. The collective dynamics of smoking in a large social network. *New England Journal of Medicine*, 2008, 358: 249-58.

Clark, Andrew and Youenn Loheac. (2007). “‘It Wasn’t Me, It Was Them!’ Social Influence in Risky Behavior by Adolescents.” *Journal of Health Economics*

Cohen-Cole, E. and Fletcher, J.M. Detecting implausible social network effects in acne, height, and headaches: Longitudinal analysis. *British Medical Journal*, 2008a, 337: a2533.

Cohen-Cole, E. and Fletcher, J.M. Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic. *Journal of Health Economics* 2008b, 27 (5): 1382-1387.

Cutler, D. and Glaeser, E. Social interactions and smoking. *National Bureau of Economic Research* Working Paper #13477, 2007

Dale, S. and A. Krueger. Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *Quarterly Journal of Economics*, 2002, 117(4):1491-1527.

Durlauf, Steven. 2004. Neighborhood Effects. In *The Handbook of Regional and Urban Economics, Volume 4: Cities and Geography*, edited by V. Henderson and J.F. Thisse. Elsevier Science/North Holland

Fletcher, J.M. 2010. Social interactions and smoking: Evidence using multiple student cohorts, instrumental variables, and school fixed effects. *Health Economics*.

Fletcher, J.M. (in press). "Peer Influences on Alcohol Consumption." Journal of Population Economics.

Friesen, Jane and Brian Krauth. 2011. Ethnic Enclaves in the Classroom. *Labour Economics* 18 (5): 656-66.

Fu, Shihe and Stephen L. Ross. 2010. Wage Premia in Employment Clusters: How Important is Worker Heterogeneity? Working Paper

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

Glaeser, Edward L., Bruce Sacerdote, and José A. Scheinkman (1996), "Crime and Social Interactions," *Quarterly Journal of Economics*, vol. 111, 507-548.

Gould, Eric, Victor Lavy, and Daniele Paserman. 2009. Long Term Classroom Peer Effects: Evidence from Random Variation in Enrollment of Disadvantaged Immigrants. *Economic Journal*, 119, 1243-1269.

Guryan, Jonathan, Kory Kroft and Matthew J. Notowidigdo. 2009. Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments. *American Economic Journal: Applied Economics*, 1, 34-68.

Halliday, Timothy J. and Kwak, Sally. 2009. Weight Gain in Adolescents and Their Peers. *Economics and Human Biology*, 7(2), 181-190.

Halliday, Timothy J. & Kwak, Sally. 2008. What is a Peer? The Role of Network Definition in the Estimation of Endogenous Peer Effects. IZA Discussion Paper #3335.

Hanushek, E.A., Kain, J.F., Markman, J.M., Rivkin, S.G., 2003. Does peer ability affect student achievement? *Journal of Applied Econometrics* 18 (5), 527–544.

Ioannides, Yannis M. and Linda Datcher Loury. 2004. Job Information Networks, Neighborhood Effects, and Inequality. *Journal of Economic Literature*, 42, 1056-93.

Jackson, M.O. Social and Economic Networks. Princeton, NJ: Princeton University Press, 2008.

Lavy, Victor, Daniele Passerman and Analia Schlosser. In Press. Inside the Black Box of Ability Peer Effects: Evidence from Variation in Low Achievers in the Classroom. *Economic Journal*.

Lavy, Victor and Analia Schlosser. 2011. "Mechanisms and Impacts of Gender Peer Effects at School." *American Economic Journal: Applied Economics.*

Liu X, Lee LF, Kagel J. Dynamic discrete choice models with lagged social interactions: With an application to a signaling game experiment. Ohio State: Mimeo, 2006.

Lundborg, Peter. (2006). "Having the Wrong Friends? Peer Effects in Adolescent Substance Use." *Journal of Health Economics*, Vol 25: 214-233

Manski, Charles. (1993). "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies*. 60

Manski, Charles. (1995). Identification Problems in the Social Sciences. Harvard University Press: Cambridge, MA

Manski, Charles. (2000). "Economic Analysis of Social Interactions." *Journal of Economic Perspectives*. 14:3

Mayer. A and S. Puller. The old boy (and girl) network: Social network formation on university campuses. *Journal of Public Economics*, 2008, 92 (1-2): 329-347

Powell, Lisa, John Taurus, and Hana Ross. "The Importance of Peer Effects, Cigarette Prices, and Tobacco Control Policies for Youth Smoking Behavior." *Journal of Health Economics*, 2005, Vol 24, pp. 950-968.

Renna, Francesco, Irina Grafova, and Nidhi Thakur. (2008). "The Effects of Friends on Adolescent Body Weight." Economics & Human Biology Volume 6, Issue 3, Pages 377-387.

Ross, Stephen L. 2011. Social interactions within cities: Neighborhood environments and peer relationships. In *Handbook of Urban Economics and Planning* (Eds. N. Brooks, K. Donaghy, G. Knapp). Oxford University Press.Sacerdote. B. Peer effects with random assignment: Results for Dartmouth roommates.
Quarterly Journal of Economics, 2001, 116 92): 681-704.

Trogdon, Justin, James Nonnemaker, Joanne Pais. (2008). "Peer Effects in Adolescent Overweight." Journal of Health Economics, 27(5): 1388-1399

Topa, Giorgio (2001), "Social Interactions, Local Spillovers, and Unemployment", *Review of Economic Studies*, Vol. 68, pp. 261-295.

Weinberg, B. (2008). Social interactions with endogenous associations. Ohio State University Working Paper

Appendix One

Proof of Propositions and Generalization of the Shock to Friendship Formation

Proposition 1. Under Assumptions 1 through 4 plus Definition 1, the bias arising from estimating the cluster fixed effects model in equation (4) limits to zero as n_i becomes large for all i in the sample.

Proof: First, based on equations (5) and (6), the probability of a friend exhibiting healthy behavior depends upon the individual's own observable and unobservable attributes that also directly influence own health behavior, the resulting correlations will bias OLS estimates of β. In order to characterize the bias from OLS estimation of equation (1) or (5), we write the expectation of equation (5) as

$$E[H_{is} \mid \tilde{H}_{is}, \tilde{X}_{is}, X_i] = \beta_s + \tilde{H}_{is}\beta_1 + \tilde{X}_{is}\beta_2 + X_{it}\beta_3 + E[\varepsilon_{is} \mid \tilde{H}_{is}, \tilde{X}_{is}, X_i] \quad \text{(A1)}$$

and substitute the linear projection of ε_{is} on the conditioning variables, $\phi_s + \tilde{H}_{is}\phi_1 + \tilde{X}_{is}\phi_2 + X_i\phi_3$, into equation (5).[19] This yields

$$E[H_{is} \mid \tilde{H}_j, \tilde{X}_{is}, X_i] = (\beta_s + \phi_s) + \tilde{H}_{is}(\beta_1 + \phi_1) + \tilde{X}_{is}(\beta_2 + \phi_2) + X_{it}(\beta_3 + \phi_3) \quad \text{(A2)}$$

Based on this linear projection, we define the bias in the estimated coefficient on $\tilde{H}_j$ as

$$\phi_1 = \frac{Cov[\varepsilon_{is}, \tilde{H}_{is} - E[\tilde{H}_{is} \mid \tilde{X}_{is}, X_i]]}{Var[\tilde{H}_{is} - E[\tilde{H}_{is} \mid \tilde{X}_{is}, X_i]]} \quad \text{(A3)}[20]$$

In terms of the health behavior equation, a cluster fixed effect will take on the following value

$$\delta_c = \overline{H}_c\beta_1 + \tilde{X}_{is}\beta_2 + X_i\beta_3 + \bar{\varepsilon}_c + \bar{\mu}_c \quad \text{(A4)}$$

where $\overline{H}_c$, $\bar{\varepsilon}_c$ and $\bar{\mu}_c$ are the means of $\tilde{H}_{is}$, ε and μ within the cluster *c*.

---

[19] This linear decomposition is typically imposed when examining problems associated with errors-in-variables in a linear model, Even without imposing any linearity assumptions, one can interpret the estimates of β as the best linear predictor of *H* conditional on $\tilde{H}_{is}$, $\tilde{X}_{is}$, X_i, and ε_{is}, and ϕ_{i1} is the relative bias in those estimates if one is unable to condition on ε_{is}.

[20] This arises from the standard omitted variables formula for a regressor that is orthogonal to all other regressors and othogonality is obtained using a conditioning argument where $y = \alpha_1 Z + \alpha_2 X + u$ can be rewritten as the following conditional regression $y = \alpha_1(Z - E[Z|X]) + \alpha_1 E[Z|X] + \alpha_2 X + u$.

After controlling for cluster fixed effects in equation (5), the health behavior model takes the following form:

$$H_{is} = (\tilde{H}_{is} - \overline{H}_c)\beta_1 + \delta_c + (\varepsilon_{is} - \bar{\varepsilon}_c) + (\mu_{is} - \bar{\mu}_c) \quad \text{(A5)}$$

The bias associated with the estimated coefficient on $(\tilde{H}_{is} - \overline{H}_c)$ in this model is

$$\phi_1^c = \frac{Cov[\varepsilon_{is} - \bar{\varepsilon}_c, (\tilde{H}_{is} - \overline{H}_c) - E[(\tilde{H}_{is} - \overline{H}_c) \mid \delta_c]]}{Var[(\tilde{H}_{is} - \overline{H}_c) - E[(\tilde{H}_{is} - \overline{H}_c) \mid \delta_c]]} = \frac{Cov[\varepsilon_{is} - \bar{\varepsilon}_c, (\tilde{H}_{is} - \overline{H}_c)]}{Var[(\tilde{H}_{is} - \overline{H}_c)]} \quad \text{(A6)}$$

Note that the expectation of the within cluster deviation in $\tilde{H}_{is}$ is zero because all observable information that influences the composition of friends on health behavior, i.e. observed attributes (X_i) or proxies for unobservable factors ($\tilde{X}_{is}$ for ε_{is}) are the same for all individuals in a cluster.

Now, the probability of a friend being of good type can be written as

$$Pr_{is}[X_j = 1] = f_{s11}(X_i, \varepsilon_{is}, \pi_{is}) + f_{s10}(X_i, \varepsilon_{is}, \pi_{is}) = f_s^X(X_i, \varepsilon_{is}) \quad \text{(A7)}$$

where the derivative of f_s^X is positive. As the number of friends becomes large,

$$\lim_{n_i \to \infty} \tilde{\mathrm{X}}_{\mathrm{is}} = f_s^X(X_i, \varepsilon_{is}) \quad \text{(A8)}$$

because as the number of draws goes to infinity the empirical frequency must equal the probability.

Since all individuals in cluster *c* have the same observable type X_{is} and the same fraction of good type friends, $\tilde{\mathrm{X}}_{\mathrm{is}}$, equation (A8) implies that

$$f_s^X(X_{is}, \varepsilon_{is}) = f_s^X(X_{is}, \varepsilon_{ks}) \quad \text{for all } i, k \in c \quad \text{(A9)}$$

when the number of friends is large.

However, equation (A9) can only hold if $\varepsilon_{is} = \varepsilon_{ks}$ for all *i* and *k* in the cluster, and so from equation (A6)

$$\lim_{n_i \to \infty} \phi_1^c = \frac{\lim_{n_i \to \infty} Cov[\varepsilon_{is} - \bar{\varepsilon}_c, (\tilde{H}_{is} - \overline{H}_c)]}{\lim_{n_i \to \infty} Var[(\tilde{H}_{is} - \overline{H}_c)]} = 0 \quad \text{(A10)}$$

because the within cluster variation in ε limits to zero while the within cluster variance of $\tilde{\mathrm{H}}_{\mathrm{is}}$ contains variation associated with π and so is strictly positive.#[21]

[21] As the number of friends becomes large, the cluster fixed effect model acts as a non-parametric control function estimator for absorbing unobservables that influence both friendship choice and health behavior. Specifically, using our notation, Blundell and Dias (2009) formally define a δ for equation (5) as $(\varepsilon_{is}, \mu_{is}) \perp (\tilde{H}_{is}, \tilde{X}_{is}, X_i) \mid \delta_c$,

Proposition 2: Under Assumptions 1 through 5 plus Definition 1, the bias arising from estimating the cluster fixed effects model in equation (11) has the same sign and is smaller than the bias that arises for the OLS model described in equation (5).

Proof: Using equation (8), the bias from the cohort fixed effect model in equation (A6) reduces to

$$\phi_1^c = \frac{Cov[\varepsilon_{is} - \bar{\varepsilon}_c, (\tilde{H}_{is} - \bar{H}_c)]}{Var[(\tilde{H}_{is} - \bar{H}_c)]} = \frac{Cov[\varepsilon_{is} - \bar{\varepsilon}_c, V_{is} - \bar{V}_c]}{Var[V_{is} - \bar{V}_c]} \tag{A11}$$

where $\bar{V}_c$ is the cohort mean of V_{is}.

The variance of the mean of a set of correlated variables is a well known expression

$$Var[\bar{V}_c] = \frac{1}{m_i} Var[V_{is}] - \frac{m_i - 1}{m_i} Cov[V_{is}, V_{ks} \mid i, k \in c] = \tag{A12}$$

where m_i is the number of individual in i's cluster. Similarly,

$$Cov[V_{is}, \bar{V}_c] = \frac{1}{m_i} Var[V_{is}] - \frac{m_i - 1}{m_i} Cov[V_{is}, V_{ks} \mid i, k \in c] \tag{A13}$$

so that the denominator of equation (A11) takes the form

$$Var[V_{is} - \bar{V}_c] = \left(1 - \frac{1}{m_i}\right)\left(Var[V_{is}] - Cov[V_{is}, V_{ks} \mid i, k \in c]\right) \tag{A14}$$

Turning to the numerator of equation (A11), the three relevant covariance terms are $Cov[\varepsilon_{is}, \bar{V}_c]$, $Cov[V_{is}, \bar{\varepsilon}_c]$ and $Cov[\bar{\varepsilon}_c, \bar{V}_c]$, which take the following form as illustrated for

$$Cov[\varepsilon_{is}, \bar{V}_c] = \frac{1}{m_i} Cov[\varepsilon_{is}, V_{is}] - \frac{m_i - 1}{m_i} Cov[\varepsilon_{is}, V_{ks} \mid i, k \in c] \tag{A15}$$

---

and conditional on δ OLS will yield consistent estimates of β. For large n_i, observations in the same cluster do not vary over ε, $\tilde{X}_{is}$ or X_i, and μ_{ist} is assumed to be an idiosyncratic disturbance.

Using all three covariance terms,

$$Cov[\varepsilon_{is} - \bar{\varepsilon}_c, V_{is} - \bar{V}_c] = \left(1 - \frac{1}{m_i}\right)\left(Cov[\varepsilon_{is}, V_{is}] - Cov[\varepsilon_{is}, V_{ks} \mid i, k \in c]\right) \quad \text{(A16)}$$

and Equation (A11) can be rewritten using equations (A14) and (A16) as

$$\phi_1^c = \frac{\left(Cov[\varepsilon_{is}, V_{is}] - Cov[\varepsilon_{is}, V_{ks} \mid i, k \in c]\right)}{\left(Var[V_{is}] - Cov[V_{is}, V_{ks} \mid i, k \in c]\right)} \quad \text{(A17)}$$

Next, using equations (17) and (18) the OLS bias in equation (9) reduces to

$$\phi_1 = \frac{Cov[\varepsilon_{is}, \tilde{H}_{is} - E[\tilde{H}_{is} \mid \tilde{X}_{is}, X_i]]}{Var[\tilde{H}_{is} - E[\tilde{H}_{is} \mid \tilde{X}_{is}, X_i]]} = \frac{Cov[\varepsilon_{is}, V_{is}]}{Var[V_{is}]} \quad \text{(A18)}$$

Note that the first terms in the numerator and denominator in equation (A17) are the same as the numerator and denominator in equation (A18). Equation (A17) will be smaller than equation (A18) if the relative or percentage reduction in the first numerator term caused by the second numerator term in equation (A17) is smaller than the equivalent reduction in the denominator or if

$$\frac{Cov[\varepsilon_{is}, V_{ks} \mid i, k \in c]}{Cov[\varepsilon_{is}, V_{is}]} > \frac{Cov[V_{is}, V_{ks} \mid i, k \in c]}{Var[V_{is}]} \quad \text{(A19)}^{22}$$

Without additional loss of generality, we can create a linear projection of V_{is} on ε_{is}

$$V_{is} = \xi_s + \varepsilon_{is}\xi_1 + U_{is} \quad \text{(A20)}$$

where $U_{is} = U(\tilde{X}_{is}, X_i, \varepsilon_{is}, \pi_{is})$ and $Cov[\varepsilon_{is}, U_{is}]$.

Further, $Cov[\varepsilon_{is}, U_{ks} \mid i, k \in c]$ and $Cov[U_{is}, U_{ks} \mid i, k \in c]$ both also equal zero because the all sources of a linear relationship between the $\tilde{H}'s$ within cohort has been eliminated. U_{ks} depends on π_{ks}, but any linear dependence with ε_{ks} and X_{is} has been eliminated from U through the linear projections and selection into clusters does not depend upon or correlate with π_{is} due to Assumption 3 and so does not contribute to the covariances.

[22] This condition holds regardless of the sign of the covariances. For example, if the covariances in the numerator of equation (A19) are both negative, they imply an increase in both the numerator and denominator and the bias is reduced if the numerator in equation (A18) increases by less. This requires that the right hand side of equation (A19) be larger magnitude, which is then smaller in value because the terms of negative.

Using equation (A20) and the above results, we can rewrite equation (A19) as

$$\frac{Cov[\varepsilon_{is},\varepsilon_{ks}\mid i,k\in c]}{Var[\varepsilon_{is}]}>\frac{Cov[\varepsilon_{is},\varepsilon_{ks}\mid i,k\in c]}{Var[\varepsilon_{is}]+\frac{1}{\xi_1^2}Var[U_{is}]} \tag{A21}$$

The variance of U_{is} is unambiguously positive because of the variation associated with π_{is} so this condition holds as long as $Cov[\varepsilon_{is},\varepsilon_{ks}\mid i,k\in c]$ is positive.

From equation (A1) and Assumption 1, we know that the probability of having good type friends f_s^X increases monotonically with ε_{is} and so the expected value of $\widetilde{X}_{is}$ must also increase monotonically with ε_{is}.[23] Therefore, we can express the fraction of good type friends as a monotonic function of ε_{is} and a stochastic variable of unknown form

$$\tilde{X}_{is}=g_s^X(X_i,\varepsilon_{is},\upsilon_{is}) \tag{A22}$$

Since the two individuals in the same cluster have the same fraction of good type friends $\widetilde{X}_{is}$ and are of the same type themselves X_i

$$g_s^X(X_i,\varepsilon_{is},\upsilon_{is})=g_s^X(X_i,\varepsilon_{ks},\upsilon_{ks}) \tag{A23}$$

where ν_{is} is an idiosyncratic error term so that $E[\varepsilon_{is},\nu_{is}]=0$.

The implicit function theorem and monotonicity assumption allows us to rewrite (A13) as

$$\varepsilon_{is}=g_\varepsilon^{-1}(X_i,\upsilon_{is},g_s^X(X_k,\varepsilon_{ks},\upsilon_{ks}))\equiv\widetilde{g}(\varepsilon_{ks},X_i,\upsilon_{is},\upsilon_{ks}) \tag{A24}$$

where g_ε^{-1} is the partial inverse of g_s^X with respect to the ε_{is} argument and is monotonically increasing in the third argument, g_s^X, for person *k*, and since ε_{ks} only enters the equation once and is inside of two monotonic functions $\tilde{g}$ can be defined as a monotonic function of ε_{ks}. The covariance can now be rewritten as

$$Cov[\varepsilon_{is},\varepsilon_{ks}\mid i,k\in c]=Cov[\tilde{g}(\varepsilon_{ks},X_i,\upsilon_{is},\upsilon_{ks}),\varepsilon_{ks}\mid i,k\in c]>0 \tag{A25}$$

which is unambiguously positive due to the monotonicity of $\tilde{g}$.

In order to sign the cohort fixed effects bias in equation (A17) relative to the OLS bias in equation (A18), we substitute equation (A20) in the numerators of the bias expressions. For OLS, the expression reduces to

$$sign(\phi_1)=sign\left(\frac{\xi_1 Var[\varepsilon_{is}]}{Var[V_{is}]}\right)=sign(\xi_1) \tag{A26}$$

[23] The following argument also holds for a monotonically decreasing function.

which takes the same sign as ξ_1. For the cohort fixed effects model,

$$sign(\phi_1^c) = sign\left(\frac{\xi_1\left(Var[\varepsilon_{is}] - Cov[\varepsilon_{is}, \varepsilon_{ks} \mid i,k \in c]\right)}{\left(Var[V_{is}] - Cov[V_{is}, V_{ks} \mid i,k \in c]\right)}\right) = sign(\xi_1) \quad \text{(A27)}$$

which will take the same sign as ξ_1 if both the terms in the numerator and denominator are unambiguously positive. The positive numerator and denominator hold due to Assumption 5 combined with the fact that a covariance of two related draws from a distribution cannot exceed the variance of this distribution. Specifically,

$$Var[V_{is}] \geq Var[V_{is} \mid \delta_c] \geq Cov[V_{is}, V_{ks} \mid i,k \in c] \quad \text{(A28)}$$

#

Generalizing the Shock to Friendship Composition

In this section, we relax Assumption (4) concerning the shock to friendship composition over health behavior so that this shock affects friendship composition over both health behavior and attributes. Assumption (4) is primarily supported by our across cohort identification strategy, and may be violated in models that are identified by within cohort variation in friendship choices. In that context, this extension is considered for two reasons: 1. To illustrate that Assumption (4) is crucial for our identification strategy and 2. To illustrate the potential bias in certain models that we estimate that include information from within cohort variation.

One possible alternative is to redefine the set of functions that describe the likelihood of observing a specific health behavior H_{is} and type X_i as

$$Pr_{is}\left[X_j = x, H_j = h | X_i, \varepsilon_{is}, \pi_{is}, j \in \mathrm{K}_i\right] = f_{sxh}(X_i, y_{is} = \varepsilon_{is} + \alpha\, \pi_{is}, \pi_{is}) \quad \text{(A29)}$$

and replace assumptions (3) and (4) with

Assumption 7: $\frac{\partial f_{s11}}{\partial y} + \frac{\partial f_{s10}}{\partial y} > 0, \frac{\partial f_{s11}}{\partial y} + \frac{\partial f_{s01}}{\partial y} > 0, \frac{\partial f_{s11}}{\partial \pi} = -\frac{\partial f_{s10}}{\partial \pi} > 0$ *and* $\frac{\partial f_{s01}}{\partial \pi} = -\frac{\partial f_{s01}}{\partial \pi} > 0$

which retains our monotonicity assumption in the effect of attributes on friendship, but now over a linear combination of ε_{is} and π_{is}. As the number of friends becomes large,

$$\lim_{n_i \to \infty} \widetilde{\mathrm{X}}_{\mathrm{is}} = f_s^X(X_i, \varepsilon_{is} + \alpha\pi_{is}) \quad \text{(A30)}$$

and

$$f_s^X(X_{is}, \varepsilon_{is} + \alpha\pi_{is}) = f_s^X(X_{is}, \varepsilon_{ks} + \alpha\pi_{ks}) \quad \text{for all } i,k \in c \quad \text{(A31)}$$

where c is defined based on constant X_{is} and $\widetilde{\mathrm{X}}_{\mathrm{is}}$ as in Definition 1. This implies that

$$\varepsilon_{is} + \alpha\pi_{is} = \varepsilon_{ks} + \alpha\pi_{ks} \quad \text{for all } i,k \in c \quad \text{(A32)}$$

Further, equation (A32) implies that

$$\varepsilon_{is} - \bar{\varepsilon}_c = \alpha(\pi_{is} - \bar{\pi}_c) \tag{A33}$$

where $\bar{\varepsilon}_c$ and $\bar{\pi}_c$ are the cohort means of ε_{is} and π_{is}.

Now as in Theorem 2, we expand V_{is} from equation (8) in terms of the relevant disturbances as

$$V_{is} = \varsigma_0 + \varepsilon_{is}\varsigma_1 + \pi_{is}\varsigma_2 + \tilde{U}_{is} \tag{A34}$$

And using equation (A23)

$$V_{is} - \bar{V}_c = (\varepsilon_{is} - \bar{\varepsilon}_c)\varsigma_1 + (\pi_{is} - \bar{\pi}_c)\varsigma_2 + (\tilde{U}_{is} - \bar{U}_c) = (\varepsilon_{is} - \bar{\varepsilon}_c)(\varsigma_1 - \varsigma_2/\alpha) \tag{A35}$$

where $\bar{U}_c$ is the cohort mean of $\tilde{U}_{is}$.

The bias from the cohort fixed effect model as shown in equation (A11) can be rewritten using equation (A35) as

$$\phi_1^c = \frac{Cov[\varepsilon_{is} - \bar{\varepsilon}_c, V_{is} - \bar{V}_c]}{Var[V_{is} - \bar{V}_c]} = \frac{(\varsigma_1 + \varsigma_2/\alpha)Var[\varepsilon_{is} - \bar{\varepsilon}_c]}{(\varsigma_1 + \varsigma_2/\alpha)Var[\varepsilon_{is} - \bar{\varepsilon}_c] + Var[\tilde{U}_{is} - \bar{U}_c]} \tag{A36}$$

The same substitution into the OLS bias expression from equation (A19) yields

$$\phi_1 = \frac{Cov[\varepsilon_{is}, V_{is}]}{Var[V_{is}]} = \frac{\varsigma_1 Var[\varepsilon_{is}]}{\varsigma_1 Var[\varepsilon_{is}] + Var[\tilde{U}_{is}]} \tag{A37}$$

because the unconditional covariance between ε_{is} and $\tilde{U}_{is}$ is zero.

In general, Proposition 1 will not hold for arbitrary values of the underlying parameters because the presence of π_{is} allows within cohort variation in ε_{is} to remain even as the number of friends becomes large. Further, the sign of the bias may differ from the OLS bias. If for example OLS estimates overstate the effect of friends' health behavior ($\varsigma_1 > 0$), the cluster fixed effect estimates under Assumption 5 may understate the effect. Specifically, if effects of π_{is} on friendship formation over attributes (α) differs in sign from the effects of π_{is} on friends' health behavior (ς_2), then $\varsigma_1 + \varsigma_2/\alpha$ is opposite sign of ς_1. This would arise if the direct effect of π_{is} on friendship formation on health behavior was opposite in sign and dominated the effect through *y*. Finally, based on Proposition 2, the sign of the OLS and cluster FE estimates are the same when π_{is} does not enter friendship formation over attributes and so our non-cohort cluster FE estimates that contain within cohort variation may produce estimates that lie below (relative to the OLS or school FE estimates) the true value of the parameter.

Proposition 3. Under Assumptions 1 and 3 through 6 plus Definition 1 the cluster fixed effects model estimate of the effect of friends' health behavior limits to a reduced form estimate that is the sum of β_1 and φ_1 as n_i becomes large for all i in the sample.

Proof: Given the assumptions, the linear projections of H_{is} and μ_{is} can be written as

$$E[H_{is} \mid \tilde{H}_{is}, \tilde{X}_{is}, X_i] = \beta_0 + \tilde{H}_{is}\beta_1 + \tilde{X}_{is}\beta_2 + X_{it}\beta_3 + E[\varepsilon_{is} + \mu_{is} \mid \tilde{H}_{is}, \tilde{X}_{is}, X_i] \quad \text{(A38)}$$

$$E[\mu_{is} \mid \tilde{H}_{is}, \tilde{X}_{is}, X_i] = \varphi_0 + \tilde{H}_{is}\varphi_1 \quad \text{(A39)}$$

$$E[H_{is} \mid \tilde{H}_j, \tilde{X}_{is}, X_i] = (\beta_0 + \phi_0 + \varphi_0) + \tilde{H}_{is}(\beta_1 + \phi_1 + \varphi_1) + \tilde{X}_{is}(\beta_2 + \phi_2) + X_{it}(\beta_3 + \phi_3) \quad \text{(A40)}$$

where using the expansion in equation (8) the new bias term may be expressed as

$$\varphi_1 = \frac{Cov[\mu_{is}, \tilde{H}_{is} - E[\tilde{H}_{is} \mid \tilde{X}_{is}, X_i]]}{Var[\tilde{H}_{is} - E[\tilde{H}_{is} \mid \tilde{X}_{is}, X_i]]} = \frac{Cov[\mu_{is}, V_{is}]}{Var[V_{is}]} \quad \text{(A41)}$$

Taking the expectation of the cluster fixed effects model in equation (A5) yields

$$E[H_{is} \mid \tilde{H}_{is} - \overline{H}_c, \delta_c] = (\tilde{H}_{is} - \overline{H}_c)\beta_1 + \delta_c + E[(\varepsilon_{is} - \bar{\varepsilon}_c) \mid \tilde{H}_{is} - \overline{H}_c, \delta_c] + E[(\mu_{is} - \bar{\mu}_c) \mid \tilde{H}_{is} - \overline{H}_c, \delta_c] \quad \text{(A42)}$$

The form of the bias in the estimated coefficient on $(\tilde{H}_{is} - \overline{H}_c)$ that is associated with the expectation over $(\varepsilon_{is} - \bar{\varepsilon}_c)$ has been previously defined in equation (A6). Again exploiting the expansion in equation (8), the bias associated with the expectation over $(\mu_{is} - \bar{\mu}_c)$ is

$$\varphi_1^c = \frac{Cov[\mu_{is} - \bar{\mu}_c, (\tilde{H}_{is} - \overline{H}_c) - E[(\tilde{H}_{is} - \overline{H}_c) \mid \delta_c]]}{Var[(\tilde{H}_{is} - \overline{H}_c) - E[(\tilde{H}_{is} - \overline{H}_c) \mid \delta_c]]} = \frac{Cov[\mu_{is} - \bar{\mu}_c, V_{is} - \overline{V}_c]}{Var[V_{is} - \overline{V}_c]} \quad \text{(A43)}$$

By equation (A40), the expectation of the estimate of the effect of friends' health behavior in the cluster fixed effects model is $(\beta_1 + \phi_1^c + \varphi_1^c)$, and Proposition 1 establishes that ϕ_1^c limits to zero with the number of friends.

Following the derivations in equations (A12) through (A17) except for μ instead of ε yields

$$\phi_1^c = \frac{(Cov[\mu_{is}, V_{is}] - Cov[\mu_{is}, V_{ks} \mid i, k \in c])}{(Var[V_{is}] - Cov[V_{is}, V_{ks} \mid i, k \in c])} \quad \text{(A44)}$$

However, membership in the cluster c only depends upon X_{is} and $\tilde{X}_{is}$ and so provides no information concerning the expectation of either μ_{is} or V_{is} since μ_{is} is orthogonal to these variables by assumption and V_{is} is orthogonal by construction. Therefore the covariance terms between *i* and *k* are zero,

$$\varphi_1^c = \frac{Cov[\mu_{is}, V_{is}]}{Var[V_{is}]} = \varphi_1 \tag{A45}$$

and

$$\lim_{n_i \to \infty} (\beta_1 + \phi_1^c + \varphi_1^c) = \lim_{n_i \to \infty} (\beta_1 + \phi_1^c + \varphi_1) = \beta_1 + \varphi_1 \#$$

Table 1
Descriptive Statistics
Add Health
Analysis Sample From In School Survey: Same Grade/Same Sex Friends
N~65,000

Variable	Mean	Std Dev	Min	Max
Smoke	0.35	0.48	0	1
Drink	0.54	0.50	0	1
Male	0.47	0.50	0	1
White	0.59	0.49	0	1
Hispanic	0.14	0.35	0	1
Black	0.18	0.38	0	1
Asian	0.06	0.23	0	1
Live with Mom	0.93	0.26	0	1
Maternal Years of Education	13.41	2.33	0	18
Maternal Caring Scale	4.78	0.61	1	5
Native Born	0.92	0.28	0	1
Grade = 7	0.14	0.35	0	1
Grade = 8	0.14	0.35	0	1
Grade = 9	0.21	0.41	0	1
Grade = 10	0.19	0.40	0	1
Grade = 11	0.17	0.37	0	1
Grade = 12	0.15	0.36	0	1
Missing	0.43	0.49	0	1
Number of Nominations	2.41	1.53	0	5
Proportion White	0.60	0.43	0	1
Proportion Black	0.17	0.35	0	1
Proportion Hispanic	0.13	0.29	0	1
Proportion Asian	0.06	0.19	0	1
Proportion Other Race	0.04	0.14	0	1
Proportion Mom Less High School	0.15	0.28	0	1
Proportion Mom Some College	0.18	0.28	0	1
Proportion of Mom College Grad	0.35	0.31	0	1
Proportion Native	0.92	0.22	0	1
Proportion Live with Mom	0.93	0.18	0	1

Table 2

Distribution of Health Behaviors in Friendship Networks

% Smoke	Freq.	Percent	Cum.	% Drink	Freq.	Percent	Cum.
Same Sex Friends							
0.00	22,994	42.51	42.51	0.00	12,509	23.18	23.18
0.10				0.10			
0.20	1,534	2.84	45.34	0.20	931	1.73	24.91
0.30	7,270	13.44	58.78	0.30	5,542	10.27	35.18
0.40	1,154	2.13	60.91	0.40	1,064	1.97	37.15
0.50	7,146	13.21	74.12	0.50	7,713	14.3	51.45
0.60	770	1.42	75.55	0.60	1,135	2.1	53.55
0.70	2651	4.9	80.45	0.70	3,774	6.99	60.55
0.80	1,748	3.23	83.68	0.80	3,440	6.38	66.92
0.90				0.90			
1.00	8,830	16.32	100	1.00	17,847	33.08	100
Total	54,097	100		Total	53,955	100	

Table 3

Variation in Friendship Options

Peer Variable	R-squared
% Maternal College Graduate	92.5%
% Black	97.2%
% Hispanic	97.4%
% White	
% Asian	93.8%
% Native Born	96.1%
Mean Maternal Caring Scale	55.1%
% Smoke Cigarettes	76.5%
% Drink Alcohol	80.9%

Notes: The results reported indicate the R-squared from a regression of a complete set of school-level and grade-level dummy variables on the grade-level measure of peer characteristics or peer health behaviors

N~65,000

Table 4A
Balancing Tests of Friendship Sorting

Outcome	Hispanic	Hispanic	Hispanic	Hispanic	Hispanic	Hispanic
Specification	OLS	OLS	OLS	OLS	OLS	OLS
Fixed Effects	None	School	School-Cluster	School-GradePair-Cluster	School-GradePair-Cluster-X	School-GP-Cluster-X One Per Cohort
Friends' Maternal Education	-0.032***	-0.022***	-0.003	-0.003		
	(0.006)	(0.005)	(0.002)	(0.003)		
Observations	65456	65456	65456	65456		
R-squared	0.027	0.306	0.696	0.725		
Outcome	White	White	White	White	White	White
Specification	OLS	OLS	OLS	OLS	OLS	OLS
Fixed Effects	None	School	School-Cluster	School-GradePair-Cluster	School-GradePair-Cluster-X	School-GP-Cluster-X One Per Cohort
Friends' Maternal Education	0.026***	0.019***	0.002	0.002		
	(0.007)	(0.004)	(0.002)	(0.003)		
Observations	65495	65456	65456	65456		
R-squared	0.003	0.009	0.570	0.752		
Outcome	Maternal Education	Maternal Education	Maternal Education	Maternal Education	Maternal Education	Maternal Education
Specification	OLS	OLS	OLS	OLS	OLS	OLS
Fixed Effects	None	School	School-Cluster	School-GradePair-Cluster	School-GradePair-Cluster-X	School-GP-Cluster-X One Per Cohort
Friends' Maternal Education	0.331***	0.197***	0.024	0.007	-0.010	0.018
	(0.026)	(0.021)	(0.020)	(0.014)	(0.017)	(0.072)
Observations	65456	65456	65456	65456	65456	49511
R-squared	0.061	0.123	0.530	0.586	0.869	0.975

Each set of rows and each column displays coefficients from separate regressions. All regressions control for grade-level fixed effects.

Table 4B

Balancing Tests of Friendship Sorting (Health Behaviors)

Outcome	Maternal Education	Maternal Education	Maternal Education	Maternal Education	Maternal Education	Maternal Education
Specification	OLS	OLS	OLS	OLS	OLS	OLS
Fixed Effects	None	School	School-Cluster	School-GradePair-Cluster	School-GradePair-Cluster-X	School-GP-Cluster-X One Per Cohort
Friends' Drinking	-0.239***	-0.280***	-0.179***	-0.064	-0.059	-0.004
	(0.057)	(0.033)	(0.046)	(0.048)	(0.066)	(0.203)
Observations	53895	53895	53895	53895	54027	43797
R-squared	0.002	0.112	0.603	0.665	0.915	0.980
Outcome	Maternal Education	Maternal Education	Maternal Education	Maternal Education	Maternal Education	Maternal Education
Specification	OLS	OLS	OLS	OLS	OLS	OLS
Fixed Effects	None	School	School-Cluster	School-GradePair-Cluster	School-GradePair-Cluster-X	School-GP-Cluster-X One Per Cohort
Friends' Smoking	-0.303***	-0.337***	-0.234***	-0.091*	-0.031	-0.050
	(0.074)	(0.048)	(0.059)	(0.046)	(0.072)	(0.217)
Observations	54027	54027	54027	54027	53564	43895
R-squared	0.003	0.113	0.602	0.665	0.939	0.981
Outcome	Maternal Education	Maternal Education	Maternal Education	Maternal Education	Maternal Education	Maternal Education
Specification	OLS	OLS	OLS	OLS	OLS	OLS
Fixed Effects	None	School	School-Cluster	School-GradePair-Cluster	School-GradePair-Cluster-X	School-GP-Cluster-X One Per Cohort
Friends' Maternal Caring	0.216***	0.157***	0.082**	0.079**	0.039	0.053
	(0.034)	(0.023)	(0.041)	(0.035)	(0.047)	(0.137)
Observations	51017	51017	50289	51017	50289	41009
R-squared	0.002	0.112	0.601	0.602	0.938	0.980

Each set of rows and each column displays coefficients from separate regressions. All regressions control for grade-level fixed effects.

Table 4C

Placebo Test of Friendship Effects (Chest Pain)

Outcome	Chest Pain	Chest Pain	Chest Pain	Chest Pain	Chest Pain	Chest Pain
Specification	OLS	OLS	OLS	OLS	OLS	OLS
Fixed Effects	None	School	School-Cluster	School-GradePair-Cluster	School-GradePair-Cluster-X	School-GP-Cluster-X One Per Cohort
Friends' Chest Pain	0.049***	0.044***	0.019	0.005	-0.020	-0.018
	(0.012)	(0.012)	(0.023)	(0.029)	(0.030)	(0.113)
Observations	48,843	48,843	48,843	48,843	48,843	36,596
R-squared	0.006	0.010	0.554	0.614	0.698	0.900

All regressions control for grade-level fixed effects.

Table 5

Friendship Network Effects on Smoking

Outcome	Smoke	Smoke	Smoke	Smoke	Smoke	Smoke	Smoke
Specification	OLS	OLS	OLS	OLS	OLS	OLS	OLS
Friends	Same Sex/ Same Grade	Same Sex/ Same Grade	Same Sex/ Same Grade	Same Sex/ Same Grade	Same Sex/ Same Grade	Same Sex/ Same Grade	Same Sex/ Same Grade
Fixed Effects	None	School	Cluster	School-Cluster	School-GradePair-Cluster	School-GradePair-Cluster-X	School-GP-Cluster-X One Per Cohort
% Smoke	0.385***	0.366***	0.368***	0.308***	0.297***	0.295***	0.333***
	(0.010)	(0.011)	(0.012)	(0.019)	(0.022)	(0.028)	(0.097)
Age	0.035***	0.034***	0.033***	0.035***	0.036***	0.039***	0.036**
	(0.004)	(0.004)	(0.005)	(0.008)	(0.010)	(0.015)	(0.015)
Male	-0.010**	-0.014***	-0.017***	-0.012	-0.010	-0.016	-0.002
	(0.005)	(0.005)	(0.006)	(0.010)	(0.013)	(0.020)	(0.018)
Hispanic	-0.024**	0.001	-0.002	-0.003	-0.010	-0.006	
	(0.011)	(0.009)	(0.011)	(0.022)	(0.028)	(0.040)	
Black	-0.093***	-0.108***	-0.093***	-0.088***	-0.092***	-0.054	
	(0.008)	(0.009)	(0.013)	(0.028)	(0.033)	(0.048)	
Asian	-0.081***	-0.080***	-0.056***	-0.089***	-0.098***	-0.072	
	(0.009)	(0.011)	(0.012)	(0.025)	(0.028)	(0.068)	
Live with Mom	-0.066***	-0.065***	-0.072***	-0.077***	-0.072***	-0.072	-0.073*
	(0.009)	(0.009)	(0.009)	(0.019)	(0.024)	(0.061)	(0.037)
Maternal Education	-0.004***	-0.005***	-0.003***	-0.006***	-0.006**	-0.003	-0.006*
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.006)	(0.004)
Maternal Caring Index	-0.071***	-0.071***	-0.070***	-0.070***	-0.074***	-0.073***	-0.082***
	(0.003)	(0.004)	(0.004)	(0.007)	(0.009)	(0.015)	(0.013)
Native Born	0.067***	0.059***	0.050***	0.054**	0.056**	0.043	0.047
	(0.014)	(0.010)	(0.014)	(0.024)	(0.023)	(0.045)	(0.034)
Observations	50249	50249	50249	50249	50249	50249	40427
R-squared	0.140	0.147	0.248	0.581	0.651	0.772	0.775

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Additional Controls: Grade dummies, Constant, Missing Indicator

Table 6

Friendship Network Effects on Drinking

Outcome	Drink	Drink	Drink	Drink	Drink	Drink	Drink
Specification	OLS	OLS	OLS	OLS	OLS	OLS	OLS
Friends	Same Sex/ Same Grade	Same Sex/ Same Grade	Same Sex/ Same Grade	Same Sex/ Same Grade	Same Sex/ Same Grade	Same Sex/ Same Grade	Same Sex/ Same Grade
Fixed Effects	None	School	Cluster	School-Cluster	School-GradePair-Cluster	School-GradePair-Cluster-X	School-GP-Cluster-X One Per Cohort
% Drink	0.329***	0.302***	0.316***	0.253***	0.239***	0.235***	0.297***
	(0.011)	(0.011)	(0.013)	(0.019)	(0.022)	(0.033)	(0.095)
Age	0.032***	0.031***	0.034***	0.032***	0.033***	0.033**	0.025
	(0.004)	(0.004)	(0.005)	(0.009)	(0.011)	(0.015)	(0.057)
Male	0.005	0.002	0.003	0.007	0.006	0.009	0.022
	(0.004)	(0.004)	(0.005)	(0.010)	(0.012)	(0.020)	(0.066)
Hispanic	0.018*	0.026***	0.033***	0.044*	0.044		
	(0.010)	(0.008)	(0.011)	(0.023)	(0.027)		
Black	-0.039***	-0.051***	-0.070***	-0.074***	-0.072**		
	(0.010)	(0.011)	(0.015)	(0.028)	(0.028)		
Asian	-0.101***	-0.118***	-0.078***	-0.081***	-0.078**		
	(0.011)	(0.012)	(0.014)	(0.026)	(0.032)		
Live with Mom	-0.066***	-0.065***	-0.071***	-0.069***	-0.075***	-0.050	-0.091
	(0.009)	(0.009)	(0.009)	(0.018)	(0.021)	(0.076)	(0.250)
Maternal Education	-0.004***	-0.006***	-0.004**	-0.006***	-0.006**	0.005	0.006
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.010)	(0.022)
Maternal Caring Index	-0.064***	-0.064***	-0.064***	-0.063***	-0.062***	-0.065***	-0.069
	(0.004)	(0.003)	(0.004)	(0.007)	(0.009)	(0.014)	(0.057)
Native Born	0.086***	0.084***	0.070***	0.079***	0.083***	0.077	0.118
	(0.014)	(0.013)	(0.012)	(0.022)	(0.028)	(0.052)	(0.209)
Observations	49656	49656	49656	49656	49656	49656	40570
R-squared	0.153	0.163	0.270	0.609	0.674	0.807	0.942

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Additional Controls: Grade dummies, Constant, Missing Indicator

Table 7
Two Stage Estimates of Friendship Effects on Smoking and Drinking

Outcome	Smoke	Smoke	Smoke	Smoke	Smoke	Smoke
Specification	School FE	School-Cluster-X One Per Cohort	School FE	School-Cluster-X	School FE, Cluster FE Est. w/ own Cohort	School FE, Cluster FE Est. w/out own Cohort
Sample	Individual	Individual	Collapsed	Collapsed	Individual	Individual
% Smoke	0.340***	0.333***	0.363***	0.336***	0.318***	0.328***
	(0.014)	(0.024)	(0.017)	(0.027)	(0.010)	(0.011)
Predicted FE w/ or w/out own cohort					1.037***	-0.002
					(0.003)	(0.009)
Guryan et. al. Control						-171.477***
						(16.991)
Observations	11,373	11,373	11,373	11,373	17,500	17,500
R-squared	0.147	0.539	0.192	0.553	0.377	0.392
Outcome	Drink	Drink	Drink	Drink	Drink	Drink
Specification	School FE	School-Cluster-X One Per Cohort	School FE	School-Cluster-X	School FE, Cluster FE Est. w/ own Cohort	School FE, Cluster FE Est. w/out own Cohort
Sample	Individual	Individual	Collapsed	Collapsed	Individual	Individual
% Drink	0.289***	0.312***	0.320***	0.321***	0.287***	0.288***
	(0.014)	(0.024)	(0.016)	(0.023)	(0.010)	(0.012)
Predicted FE w/ or w/out own cohort					1.037***	0.005
					(0.004)	(0.011)
Guryan et. al. Control						-179.230***
						(15.565)
Observations	11,340	11,340	11,340	11,340	17,435	17,435
R-squared	0.154	0.550	0.209	0.568	0.385	0.403

Table 8

Longitudinal Analysis of the onset of Smoking and Drinking

Outcome	Smoke	Smoke	Smoke	Smoke	Smoke
Specification	OLS	School FE	School FE	School FE, Cluster FE Est. w/ own Cohort	School FE, Cluster FE Est. w/out own Cohort
Sample	Full Sample	Full Sample	Restricted	Restricted	Restricted
% Smoke	0.063***	0.061***	0.053***	0.051***	0.045***
	(0.011)	(0.012)	(0.016)	(0.017)	(0.017)
Predicted FE w/ or w/out own cohort				-0.018	0.002
				(0.035)	(0.021)
Guryan et. al. Control					11.743**
					(5.392)
Observations	6,137	6,137	1,895	1,895	1,895
R-squared	0.019	0.041	0.093	0.093	0.095
Outcome	Drink	Drink	Drink	Drink	Drink
Specification	School FE	School FE	School FE	School FE, Cluster FE Est. w/ own Cohort	School FE, Cluster FE Est. w/out own Cohort
Sample	Full Sample	Full Sample	Restricted	Restricted	Restricted
% Drink	0.120***	0.112***	0.110***	0.107***	0.114***
	(0.015)	(0.016)	(0.027)	(0.028)	(0.028)
Predicted FE w/ or w/out own cohort				-0.044	-0.028
				(0.056)	(0.028)
Guryan et. al. Control					-10.836
					(11.298)
Observations	4,310	4,310	1,228	1,228	1,228
R-squared	0.034	0.074	0.119	0.120	0.121

Table 9
Longitudinal Analysis of the onset of Smoking and Drinking
Stratified by Cluster Size

Outcome	Smoke	Smoke	Smoke	Smoke
Specification	School FE	School FE	School FE, Cluster FE Est. w/out own Cohort	School FE, Cluster FE Est. w/out own Cohort
Sample	Restricted Small Cluster	Restricted Large Cluster	Restricted Small Cluster	Restricted Large Cluster
% Smoke	0.055*	0.052**	0.047	0.041**
	(0.029)	(0.021)	(0.031)	(0.021)
Predicted FE w/ or w/out own cohort			-0.025	0.054
			(0.029)	(0.036)
Guryan et. al. Control			9.439	18.294**
			(11.739)	(8.465)
Observations	833	1,062	833	1,062
R-squared	0.190	0.125	0.191	0.133
Outcome	Drink	Drink	Drink	Drink
Specification	School FE	School FE	School FE, Cluster w/out own Cohort	School FE, Cluster w/out own Cohort
Sample	Restricted Small Cluster	Restricted Large Cluster	Restricted Small Cluster	Restricted Large Cluster
% Drink	0.151***	0.079**	0.151***	0.082**
	(0.048)	(0.037)	(0.047)	(0.038)
Predicted FE w/ or w/out own cohort			-0.015	-0.034
			(0.044)	(0.070)
Guryan et. al. Control			-1.374	-24.185*
			(18.174)	(14.256)
Observations	571	657	571	657
R-squared	0.238	0.175	0.238	0.179

Table 10
Friendship Effects by Number of Friends

Outcome	Smoke	Smoke	Smoke	Smoke
Specification	School FE	School FE	School-Cluster-X FE	School-Cluster-X FE
Sample	Collapsed Sample 1-3 Friends	Collapsed Sample 4-5 Friends	Collapsed Sample 1-3 Friends	Collapsed Sample 4-5 Friends
% Smoke	0.326***	0.599***	0.307***	0.589***
	(0.017)	(0.031)	(0.027)	(0.058)
Observations	9,201	2,154	9,201	2,154
R-squared	0.187	0.284	0.546	0.604
Outcome	Drink	Drink	Drink	Drink
Specification	School FE	School FE	School-Cluster-X FE	School-Cluster-X FE
Sample	Collapsed Sample 1-3 Friends	Collapsed Sample 4-5 Friends	Collapsed Sample 1-3 Friends	Collapsed Sample 4-5 Friends
%Drink	0.290***	0.539***	0.297***	0.543***
	(0.016)	(0.033)	(0.025)	(0.063)
Observations	9,173	2,149	9,173	2,149
R-squared	0.201	0.313	0.559	0.621

Table 11

Friendship Effects on On-set by Number of Friends

Outcome	Smoke	Smoke	Smoke	Smoke	Smoke	Smoke
Specification	School FE	School FE	School FE	School FE	School FE- Cluster FE Est. w/out own Cohort	School- Cluster FE Est. w/out own Cohort
Sample	Individual 1-3 Friends	Individual 4-5 Friends	Restricted 1-3 Friends	Restricted 4-5 Friends	Restricted 1-3 Friends	Restricted 4-5 Friends
% Smoke	0.053***	0.073***	0.067***	0.001	0.066***	-0.001
	(0.015)	(0.017)	(0.020)	(0.029)	(0.021)	(0.029)
Observations	6,137		1,895		1,895	
R-squared	0.042		0.094		0.094	
Outcome	Drink	Drink	Drink	Drink	Drink	Drink
Specification	School FE	School FE	School FE	School FE	School FE- Cluster FE Est. w/out own Cohort	School- Cluster FE Est. w/out own Cohort
Sample	Individual 1-3 Friends	Individual 4-5 Friends	Restricted 1-3 Friends	Restricted 4-5 Friends	Restricted 1-3 Friends	Restricted 4-5 Friends
% Drink	0.088***	0.151***	0.083***	0.220***	0.081***	0.217***
	(0.016)	(0.029)	(0.028)	(0.071)	(0.029)	(0.073)
Observations	4,310		1,228		1,228	
R-squared	0.075		0.124		0.124	

Appendix Two

Tables on Sample Selection due to Research Design

Table 1A

Predictors of Dropped Sample

Outcome	Means	No ID	No Friend Nominations	No Found Nominations	Any Drop	No Same Grade	No Same Grade/ Gender Friends
Means		0.047	0.14	0.2	0.24	0.065	0.129
Fixed Effects		School	School	School	School	School	School
Age	15	0.005***	0.026***	0.016***	0.040***	0.032***	0.024***
		(0.001)	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)
Male	0.502	0.007***	0.072***	0.016***	0.083***	0.030***	0.007***
		(0.001)	(0.005)	(0.002)	(0.006)	(0.003)	(0.002)
Hispanic	0.155	0.004	0.014**	0.002	0.020***	-0.006	-0.001
		(0.003)	(0.006)	(0.004)	(0.007)	(0.006)	(0.004)
Black	0.19	0.007***	0.037***	0.027***	0.060***	0.017***	0.013***
		(0.002)	(0.006)	(0.005)	(0.007)	(0.005)	(0.004)
Asian	0.056	-0.003	0.002	0.001	0.004	-0.038***	-0.017***
		(0.004)	(0.009)	(0.009)	(0.013)	(0.008)	(0.005)
Native Born	0.9	-0.003	-0.023***	-0.027***	-0.042***	-0.007	-0.018***
		(0.003)	(0.005)	(0.006)	(0.007)	(0.008)	(0.006)
Live with Mom	0.92	0.037***	0.292***	0.059***	0.306***	0.152***	0.067***
		(0.011)	(0.019)	(0.017)	(0.023)	(0.008)	(0.006)
Mom Education	13.36	-0.000	-0.001*	-0.001	-0.002***	-0.001**	-0.001***
		(0.000)	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)
Mom Care	4.76	-0.002*	-0.009***	-0.005***	-0.013***	-0.006***	-0.006***
		(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Smoke	0.36	0.005**	0.007**	0.011***	0.017***	0.024***	0.013***
		(0.002)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)
Drink	0.55	0.000	-0.020***	-0.001	-0.017***	0.001	-0.004
		(0.001)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)
% Black	0.19	-0.010	-0.110	0.015	-0.097	-0.046	-0.043
		(0.031)	(0.084)	(0.051)	(0.107)	(0.036)	(0.035)
% Hispanic	0.16	0.036	0.089	0.067	0.162*	-0.065	-0.039
		(0.070)	(0.084)	(0.054)	(0.097)	(0.049)	(0.049)
% Mom Grad	0.32	0.040	-0.064	-0.032	-0.062	0.002	-0.035
		(0.035)	(0.052)	(0.032)	(0.065)	(0.042)	(0.036)
% Smoke	0.36	0.060	0.042	0.011	0.072	-0.092**	-0.061*
		(0.042)	(0.055)	(0.037)	(0.074)	(0.039)	(0.032)
% Drink	0.55	-0.000	-0.089	0.012	-0.047	0.103***	0.058
		(0.027)	(0.056)	(0.039)	(0.066)	(0.039)	(0.038)
Constant		-0.081**	-0.458***	-0.161***	-0.542***	-0.576***	-0.317***
		(0.032)	(0.049)	(0.040)	(0.060)	(0.039)	(0.039)
Observations		88995	84789	73363	88995	59470	59470
R-squared		0.173	0.188	0.210	0.245	0.246	0.122

Notes: Grade fixed effects controls and missing indicators are now shown. “No ID” is a binary variable indicating whether the respondent received an identification number in the survey. “No Friend Nominations” is a binary

variable indicating whether the respondent made zero friend nominations. “No Found Nominations” is a binary variable indicating whether the respondent nominated friends who were not able to be matched within sample (such as friends outside of school).

Table 2A
Predictors of "Unusual Type"—Single Cluster Membership

Outcome	Single Cluster	Single Cluster	Single Cluster	Single Cluster	Single Cluster
Cluster	Friend Xs	X, School	X,S, Own Xs	X,S Cohort	X,S,X, Cohort
Mean of dependent variable	0.06	0.32	0.37	0.44	0.53
Fixed Effects	School	School	School	School	School
Age	-0.007***	-0.033***	-0.035***	-0.036***	-0.042***
	(0.002)	(0.004)	(0.004)	(0.004)	(0.004)
Male	0.006**	-0.005	-0.014**	-0.015***	-0.017***
	(0.003)	(0.006)	(0.006)	(0.006)	(0.006)
Hispanic	0.020***	0.014	0.013	0.044***	0.078***
	(0.005)	(0.011)	(0.012)	(0.016)	(0.021)
Black	0.015**	0.029	0.030	0.041	0.043
	(0.007)	(0.021)	(0.023)	(0.025)	(0.026)
Asian	0.079***	0.091***	0.097***	0.132***	0.192***
	(0.012)	(0.018)	(0.018)	(0.025)	(0.027)
Native Born	-0.013**	0.000	0.006	0.005	0.006
	(0.006)	(0.010)	(0.009)	(0.010)	(0.011)
Live with Mom	-0.007*	-0.049***	-0.069***	-0.153***	-0.162***
	(0.004)	(0.008)	(0.009)	(0.011)	(0.012)
Mom Educatoin	-0.000	-0.000	-0.000	0.012***	0.013***
	(0.000)	(0.001)	(0.001)	(0.002)	(0.002)
Mom Care	0.000	0.000	0.002	0.003	0.003
	(0.001)	(0.003)	(0.003)	(0.003)	(0.003)
Smoke	-0.009***	-0.030***	-0.033***	-0.030***	-0.031***
	(0.002)	(0.005)	(0.005)	(0.005)	(0.005)
Drink	-0.003	-0.006	-0.007	-0.004	-0.001
	(0.003)	(0.005)	(0.004)	(0.005)	(0.005)
% Black	0.012	-0.023	-0.094	-0.090	-0.151*
	(0.028)	(0.055)	(0.072)	(0.064)	(0.085)
% Hispanic	0.035	0.097	0.049	0.077	-0.051
	(0.050)	(0.087)	(0.094)	(0.087)	(0.097)
% Mom College Grad	0.042	-0.081	-0.099	-0.038	-0.085
	(0.038)	(0.069)	(0.084)	(0.070)	(0.078)
% Smoke	0.024	0.037	-0.001	0.064	0.040
	(0.028)	(0.058)	(0.067)	(0.074)	(0.076)
% Drink	-0.058**	-0.125***	-0.109**	-0.144***	-0.099*
	(0.028)	(0.044)	(0.053)	(0.054)	(0.057)
Constant	0.207***	0.972***	1.084***	1.040***	1.242***
	(0.042)	(0.063)	(0.070)	(0.070)	(0.071)
Observations	59470	59470	59470	59470	59470
R-squared	0.068	0.087	0.078	0.096	0.092

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Grade fixed effects not shown.

Table 3A

Analysis of the Change in Composition of the Sample Due to Singleton Clusters

Outcome	Smoke	Smoke	Smoke	Smoke	Smoke	Smoke	Smoke
Group	None	School	Cluster	School/Cluster	School-Cluster	School-Cluster-GradePair	School-Cluster-GradePair-Xs
% Smoke	0.385***	0.366***	0.368***	0.348***	0.308***	0.297***	0.295***
	(0.010)	(0.011)	(0.012)	(0.012)	(0.019)	(0.022)	(0.028)
Observations	50959	50959	50959	50959	50959	50959	50959
R-squared	0.135	0.143	0.245	0.252	0.580	0.649	0.761
Outcome	Smoke	Smoke	Smoke	Smoke	Smoke	Smoke	Smoke
Group	None	School	Cluster	School/Cluster	School-Cluster	School-Cluster-GradePair	School-Cluster-GradePair-Xs
% Smoke	0.385***	0.366***	0.383***	0.363***	0.364***	0.358***	0.357***
	(0.010)	(0.012)	(0.011)	(0.011)	(0.011)	(0.012)	(0.011)
Observations	50249	50249	46842	46842	32032	28557	20120
R-squared	0.140	0.147	0.140	0.148	0.140	0.139	0.142
Outcome	Smoke	Smoke	Smoke	Smoke	Smoke	Smoke	Smoke
Group	None	School	Cluster	School/Cluster	School-Cluster	School-Cluster-GradePair	School-Cluster-GradePair-Xs
% Smoke	0.385***	0.357***	0.357***	0.323***	0.299***	0.291***	0.295***
	(0.010)	(0.011)	(0.011)	(0.012)	(0.016)	(0.018)	(0.018)
Observations	50249	20120	20120	20120	20120	20120	20120
R-squared	0.140	0.142	0.142	0.190	0.352	0.403	0.449

Outcome	Drink	Drink	Drink	Drink	Drink	Drink	Drink
Group	None	School	Cluster	School/Cluster	School-Cluster	School-Cluster-GradePair	School-Cluster-GradePair-Xs
% Drink	0.329***	0.302***	0.315***	0.286***	0.253***	0.240***	0.235***
	(0.011)	(0.011)	(0.013)	(0.013)	(0.019)	(0.022)	(0.029)
Observations	50019	50019	50019	50019	50019	50019	50019
R-squared	0.153	0.163	0.270	0.280	0.608	0.673	0.790
Outcome	Drink	Drink	Drink	Drink	Drink	Drink	Drink
Group	None	School	Cluster	School/Cluster	School-Cluster	School-Cluster-GradePair	School-Cluster-GradePair-Xs
% Drink	0.329***	0.302***	0.324***	0.297***	0.308***	0.304***	0.296***
	(0.011)	(0.011)	(0.012)	(0.012)	(0.013)	(0.013)	(0.013)
Observations	50019	50019	46616	46616	31878	28421	20021
R-squared	0.153	0.163	0.153	0.164	0.146	0.147	0.143
Outcome	Drink	Drink	Drink	Drink	Drink	Drink	Drink
Group	None	School	Cluster	School/Cluster	School-Cluster	School-Cluster-GradePair	School-Cluster-GradePair-Xs
% Drink	0.329***	0.296***	0.296***	0.257***	0.233***	0.225***	0.235***
	(0.011)	(0.013)	(0.013)	(0.014)	(0.018)	(0.018)	(0.019)
Observations	50019	20021	20021	20021	20021	20021	20021
R-squared	0.153	0.143	0.143	0.199	0.365	0.413	0.467

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Each column and row is from a separate regression. The first row of each set repeats the results from Table 5 (6). The second row reproduces the Column 1 results with the non-singleton samples. The third row presents results of each specification with the final column's sample