Published in final edited form as:

Health Econ. 2018 October ; 27(10): 1450–1483. doi:10.1002/hec.3780.

Estimating the effects of friends on health behaviors of adolescents

Jason M. Fletcher¹ and Stephen L. Ross²

¹La Follette School of Public Affairs, Departments of Sociology, Agricultural and Applied Economics, and Population Health Sciences, University of Wisconsin-Madison, Madison, WI, USA

²Department of Economics, University of Connecticut, Storrs, CT, USA

Abstract

This paper estimates the effects of friends' smoking and drinking on own behavior while controlling for correlated unobservables between friends. The effect of friends' behaviors is identified by comparing similar individuals who have similar friendship opportunities and make similar friendship choices, exploiting the idea that friendship choice reveals information about unobservables. We combine this identification strategy with an across-cohort within school design so that identification arises in our reduced form estimates from across-grade differences in the clustering of health behaviors. Finally, we use estimated information on correlated unobservables to examine longitudinal data on the onset of health behaviors, where the likelihood of reverse causality should be minimal. We find evidence that this strategy produces somewhat smaller (no more than 16% smaller) friendship effect estimates than the more standard school fixed effect models consistent with at most modest bias from correlated unobservables.

Keywords

adolescent health; cohort study; drinking; friendship effects; peer effects; smoking

1 | INTRODUCTION

Many studies of social interactions find evidence of clustering of outcomes beyond the clustering expected based on individuals' observables, including studies of crime (Glaeser, Sacerdote, & Scheinkman, 1996), employment (Bayer, Ross, & Topa, 2008), welfare usage

Correspondence Jason M. Fletcher, University of Wisconsin-Madison, Madison, WI, USA. jmfletcher@wisc.edu.

²⁴A 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 onset are robust in subsamples that focus on individuals in friendship choice clusters that contain a larger number of students.

²⁸This condition holds regardless of the sign of the covariances. For example, if the covariances in the numerator of Equation (B19) are both negative, they imply an increase in both the numerator and denominator and the bias is reduced if the numerator in Equation (B18) increases by less. This requires that the right hand side of Equation (B19) be larger magnitude, which is then smaller in value because the terms are negative.

³³All student demographics used in the creation of friendship clusters are included as regressors in all models unless they are collinear with the fixed effects used in that particular model.

(Bertrand, Luttmer, & Mullainathan, 2000), prenatal care (Aizer & Currie, 2004), and health behaviors (Weinberg, 2008).¹ However, without exploiting information on relationships, it is difficult to know whether these clusters are driven by social interactions or shared environments, such as information sources (Aizer & Currie, 2004) or classroom disruption (Aizer, 2009; Hoekstra & Carrell, 2010).

We observe high clustering on smoking and drinking within grades of students at the same school, but little variation across grades in composition over socioeconomic variables. The purpose of this paper is to examine whether the within friendship clustering of health behaviors underlying the grade clustering is consistent with spillovers across friends. The within-grade clustering changes the likely exposure of students to friends who exhibit those behaviors, and we exploit this variation to test whether friendship level exposure increases the student's own likelihood of exhibiting these behaviors. We provide substantial evidence consistent with at most minimal bias arising from students sorting into friendships on unobservable attributes.

A few studies have attempted to identify the causal effect of social relationships on health-related behaviors (Christakis & Fowler, 2008; Eisenberg, Golberstein, & Whitlock, 2014; Halliday & Kwak, 2012; Renna, Grafova, & Thakur, 2008; Trogdon, Nonnemaker, & Pais, 2008).² Although many studies exploit the quasirandom sorting of students into grades to study peer effects,³ studies on social relationships are far more problematic because students select their friends. Therefore, even if assignment to grade is quasirandom, the inclusion of school or network fixed effects may do little to address endogeneity in the within-grade variation provided by friendship links, and individual attributes that might act as instruments are invariably at risk of being correlated with individual unobservables.⁴

Several earlier studies exploit across cohort variation to identify the effect of friendship by examining aggregate friendship patterns. Babcock (2008) finds that being in a socially connected cohort raises the likelihood of high school graduation and college attendance, and Nathan (2008) finds that racial heterogeneity of friendships improves academic performance especially among Whites. However, as with previous studies that exploit aggregation (Cutler & Glaeser, 1997), the effects may be driven by the associated aggregate environment (Ross, 2011).

Our primary identification strategy is to exploit friendship choices made by students on other dimensions, such as friends' race or family background, to address bias from friends selecting each other on unobservables. The intuition behind the identification strategy is individuals who are similar on observables and make the same choices must be more similar on their unobservables in expectation than two random individuals selected from the population.⁵ For example, two White students who both have minority friends are likely

¹See Ross (2011) for a recent literature review.

²Several recent studies attempt to estimate the causal effects of friendship links on nonhealth outcomes (Bramoullé, Djebbari, & Fortin, 2009; Conti, Galeotti, Muller, & Pudney, 2013; Patacchini & Zenou, 2012).

³See Lavy, Passerman, and Schlosser (2012); Friesen and Krauth (2011); Bifulco, Fletcher, and Ross (2011); and Fletcher (2010, 2012).

⁴An exception is Eisenberg et al. (2014) who uses quasirandomly selected college roommates. See Fletcher and Marksteiner (2017) for an application of RCTs to examine health behavior spillovers between spouses.

much more similar on unobservables to each other than they are to other White students in the same school who do not have minority friends.

Specifically, this study examines the relationship between a student's own health behaviors, smoking and drinking, and friends' health behaviors making only comparisons between observationally similar students who (a) have very similar friendship opportunities over race/ethnicity and family background because they are in the same school, (b) make the same friendship choices over those attributes, but (c) are exposed to different frequencies of health behaviors because they are in different grades. Several previous studies have used similar identification strategies. Dale and Krueger (2002) examine returns from college quality comparing students who applied to similar quality schools. Fu and Ross (2013) examine returns from agglomeration comparing individuals who chose the same residential location, and Aliprantis and Richter (2012) compare participants in the Moving to Opportunity Program whose neighborhood choices imply similar preferences.

Similar to Dale and Krueger (2002), we (a) establish conditions under which our estimates are not biased by correlated unobservables between friends and (b) demonstrate that even when these conditions do not hold, the identification strategy reduces the bias from correlated unobservables. In our case, the estimates are consistent as the number of friendship choices of each student becomes large. Our model specification does not address bias from reverse causality where people sort into friendship based on health behaviors, which we address later by estimating a model of health onset. A key assumption behind our consistency results is that students must face some shock in the health behavior of friends,⁶ in this case, differences in health behavior clustering across cohorts.⁷

Using Wave 1 of the Add Health, we regress health behavior on average exposure to friends who exhibit the same health behavior, grade fixed effects, and fixed effects for clusters of students who attend the same school, make the same friendship choices in terms of the number of friends with specific demographic attributes, and are observationally equivalent on demographic attributes. Further, in our preferred estimates, we restrict our sample to one randomly selected individual from each grade in each cluster, identifying the model based solely on variation across cohorts. This specification satisfies a key criterion, at least within cluster, established by Angrist (2014) for peer effect studies: "... a clear separation between the subjects of a peer effects investigation and the peers who provide the mechanism for causal effects." In our sample, peers are restricted to same grade friends, and in the one student per cohort sample, a student's same grade friends or potential friends never appear in their cluster.

⁵This intuition is the mirror image of sample selection where individuals who differ on observables and make the same choice are expected to differ on unobservables.

⁶Similarly, Dale and Krueger (2002) require the existence of randomness in the student's choice that cannot be explained by student unobservables that affect future performance, but it is harder to interpret this source of randomness as a well-defined treatment.

⁷For example, smoking or drinking levels in a cohort may be affected by cohort specific environmental factors, such as life events of teachers, or the students in each cohort may have coordinated around different social structures. Without such variation, the effect of average friends' behaviors is unidentified asymptotically. As the number of friends increases, the within-cluster variance of the unobservables associated with friendship choice falls and the number of friends over which health behavior is averaged increases. Therefore, unless some outside source of variation in friends' health behaviors exists, students in the same cluster should be asymptotically exposed to the same average friends' behaviors given that composition, eliminating all within cluster variation.

This strategy has only small effects (reductions of no more than 16%) on the friendship effect estimates for smoking and drinking, relative to the standard school fixed effect estimates. Balancing tests show that friend health behaviors tend to be weak predictors of predetermined student attributes, even though those student attributes are predictive of a student's own smoking and drinking. Notably, our strategy reduces the estimates in the balancing tests by between 45% and 100%, with most reductions exceeding two thirds of the school fixed effect estimates. Therefore, our controls for correlated unobservables erode a substantial portion of the spurious correlation between friend's behavior and exogenous student attributes, and yet the estimated friendship effects on smoking and drinking are left mostly intact.

Next, we address a second major bias: Individuals may select each other as friends because they share the same health behaviors (reverse causation). We do so by estimating a model of the “onset” of health behaviors: whether a student began drinking or smoking between Waves 1 and 2 of the Add Health. Because Wave 2 behaviors cannot have a causal effect on Wave 1 friendships, the primary concern in models of onset is correlated unobservables, not reverse causality, and our fixed effect structure is designed to capture correlated unobservables. Because the Add Health followed only a subsample of students through Wave 2, we cannot estimate a model with a full set of school by friendship choice by student-type fixed effects. Therefore, we use the estimated fixed effects from our above analyses as a control function in the onset models. We find no evidence of bias from correlated unobservables in the estimated effect of friend's smoking or drinking in Wave 1 on the likelihood of starting to drink in Wave 2.

We also conduct a series of additional tests to examine the validity of our identification strategy. Our estimates are robust to including cohort level controls for peer demographic composition and peer smoking or drinking. Second, our results are also robust for individuals with above average numbers of friends where more friendship choice information is available. Third, our health onset results are robust to considering individuals in above average size school, friendship choice, and student-type clusters, so that cluster fixed effect estimates have less measurement error.

2 | IDENTIFICATION STRATEGY

Our primary focus selection into friendships is based on a student's unobservables. Our approach is to form comparison groups based on friendship “options” of students as well as the students' “choices” of friends (given these options). The intuition follows the premise that individuals who make similar decisions when facing the same set of options likely are very similar in their overall tendency to make such choices based on their observable and unobservable attributes, and so observationally different individuals with the same outcome should have different unobservables, and observationally similar individuals should have similar unobservables (Heckman, 1979).

We begin with a slight modification to the linear-in-means model of social interactions (Brock & Durlauf, 2001; Manski, 1993) by restricting interactions to arise from exposure to friends within a social environment (or school s) and dividing the unobservable contributor

to health behaviors into two components, one affecting friendship choice ε_i and one not affecting friendship choice μ_i .

Specifically, we consider the following empirical model:

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

where H_{is} indicates a health behavior of individual i in a social environment s , X_i contains the individual's observable attributes, n_i is the number of friends of person i , 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 the individuals j within Ω_{is} , and β_s is a school fixed effect. Although H_{is} is binary, we proceed with a linear model to facilitate comparison with traditional reduced form models of peer effects.⁸

Even without correlations in social relationships arising from sorting on unobservables, that is, ε_{is} is orthogonal to $\left(\frac{1}{n_i} \sum_{j \in \Omega_{is}} X_j \right)$, the linear in means model above is intrinsically unidentified (Manski, 1993).⁹ This occurs because group member j 's characteristics X_{js} that explain the health of this group member H_{js} cannot be excluded from the second stage regression for the health behavior of i because these attributes X_{js} may directly influence i 's behaviors (the reflection problem).

Our identification strategy sorts students into clusters c containing similar students who faced similar friendship options and made similar friendship choices. Specifically, define a cluster of individuals c in the same school s who are observationally equivalent on X_i and choose observationally equivalent friends X_j . This structure implies that individual and friendship group observables in Equation (1) are the same within a cluster so that contributions to health behavior are constant,

$$\beta_s + \left(\frac{1}{n_i} \sum_{j \in \Omega_{is}} X_{js} \right) \beta_2 + X_{is} \beta_3 = \beta_s + \left(\frac{1}{n_i} \sum_{j \in \Omega_{ks}} X_{js} \right) \beta_2 + X_{ks} \beta_3 \quad (2)$$

For all $i, k \in c$

Our identification strategy suggests that the individual students in cluster c should also have very similar unobservables. Specifically, we define ρ_{cs} as a cluster fixed effect where

⁸See, for example, Lavy, Passerman, and Schlosser (2008), Friesen and Krauth (2008), Bifulco et al. (2011).

⁹Estimates β_1 and β_2 cannot be separately identified. Brock and Durlauf (2001) show that this model is identified once the linearity restriction is relaxed, and Bramoullé et al. (2009) examine identification using lags along the friendship network structure (i.e., friends of friends).

$$\rho_{cs} \approx \beta_s + \left(\frac{1}{n_{ij}} \sum_{j \in \Omega_{is}} X_{js} \right) \beta_2 + X_{is} \beta_3 + \varepsilon_{is} \approx \beta_s + \left(\frac{1}{n_{ij}} \sum_{j \in \Omega_{ks}} X_{js} \right) \beta_2 + X_{ks} \beta_3 + \varepsilon_{ks}. \quad (3)$$

Next, based on the construction of μ as an idiosyncratic disturbance, $E[\mu_{is} | \rho_{c_i}] = 0$ and substituting Equation (3) into Equation (1) yields

$$H_{ics} \approx \left(\frac{1}{n_{ij}} \sum_{j \in \Omega_{is}} H_{js} \right) \beta_1 + \rho_{cs} + \mu_{ics}. \quad (4)$$

In Appendices A and B, we formally present the model and demonstrate that the estimated effect of friends' health behaviors on a students' own health behavior in Equation (4) converges to the true value as the number of friendships becomes large if (a) friends' health behaviors contain variation that cannot be entirely explained by friendship choices, (b) unobservables that affect health behavior have a monotonic relationship with friendship choices, and (c) own health behavior does not influence friendship choices. With an additional technical assumption, we also show that the inclusion of the cluster fixed effect reduces estimation bias.

In our preferred model, we restrict our comparisons to students in different grades. Specifically, each grade contributes one student per school-student type-friendship pattern cluster so that these estimates are based entirely on comparisons across cohorts. Our strategy is implemented by randomly choosing only one student in each grade from each cluster, and we present the average of the parameter estimates resulting from several random draws of one individual in each cluster per cohort.

Note that this identification strategy does not solve the problem of reflection on the friends' unobservables. We cannot distinguish between the spillover effects of friends' health behavior directly and the spillover effects of friends' unobservables such as values that might drive those behaviors. As noted by Sacerdote (2001) and Bayer and Ross (2008), even random assignment cannot solve this identification problem.

3 | FRIENDSHIP DATA

We use the only available national dataset containing friendship network information and 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. For this paper, we focus on the In-School data collection (Wave 1), which utilized a self-administered instrument to more than 90,000 students in Grades 7 through 12 in a 45- to 60-min class period between September 1994 and April 1995. The questionnaire focused on topics including sociodemographic characteristics, family background, health status, and risk behaviors. In addition, each

student respondent was asked to identify up to 10 close friends (five males, five females) from the school's roster, and friendship links within the sample are created based on two students nominating each other, reciprocal nominations.¹⁰ The analysis sample requires several restrictions to the Wave 1 sample, see Appendix C.

Table 1 presents descriptive statistics of the analysis sample. Approximately 34% of the sample reports smoking and 54% of the sample reports drinking alcohol. The average adolescent nominates 2.8 same-sex, close friends. Same grade nominations represent approximately 85% of all nominations, and the likelihood of making cross-grade nominations does not appear to be related to cohort composition. We impute missing covariates, such as maternal education, and control for missingness, but we do not impute missing outcomes.

3.1 | Evidence of variation in friendship options

As discussed 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 similar friendship choices on student attributes, but because they belong to different grades, draw friends who exhibit differing health behavior. This strategy 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 2 that controlling for school and grade effects can predict over 98% of the variation in racial composition of potential friends (classmates). Likewise, controlling for school and grade predicts 93% of the variation in peers' maternal education and 96% of the variation in classmate nativity. These findings suggest that students in different grades of the same school have very similar friendship options in terms of race/ethnicity and family background. However, using the same regression analysis, our data show that we only predict 77% of peer smoking rates and 81% of peer drinking rates. Thus, there is substantial variation in exposure to health behaviors of potential friends (classmates) across cohorts, whereas the friendship options based on race, maternal education, and nativity are nearly identical across grades.

3.2 | Empirical specification and evidence on identification

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 (Fletcher, Ross, & Zhang, 2013; Mayer & 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

¹⁰Requiring nominations to be reciprocal may eliminate some friendships, but at worst, this restriction adds measurement error attenuating estimates. Given minimal declines in effect estimates, measurement error is modest at most.

obtaining very specific friendship-type clusters because this is the dimension over which students are making choices.

The friendship clusters are based on the following exogenous characteristics of self-reported same gender friends, including (a) race/ethnicity (black vs. Hispanic vs. White vs. Asian vs. other), (b) maternal education (no college vs. some college vs. college graduate), (c) family structure (living with mother vs. not living with mother), and (d) nativity (native vs. foreign born). Specifically, the number of friends chosen for 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.¹¹ The friendship variables above cover most of the clearly exogenous attributes available in the Wave 1 sample.

We define observationally equivalent students over three key variables: the student's race/ethnicity (White, Black, Hispanic, and Asian), whether their mother is a college graduate, and gender. Our data contain over 3,800 clusters of students in the same school with identical friendship patterns and with two or more grades represented, and we have approximately 4,500 school by cluster by student-type fixed effects with two or more grades represented. Based on these empirically relevant groups, the typical school has on average 56 unique groups, and the standard deviation is 43 with a maximum of 200 groups in any school. The mean group size in the average school is eight students, and the standard deviation of this mean size across schools is three students with a maximum mean size of 27 students.¹²

The more parsimonious specification of observationally equivalent students has the added advantage of allowing for additional balancing tests over student observables. For family structure and nativity, we can examine whether within-cluster differences in friend's drinking or smoking can explain whether the student lives with their mother or whether the student's mother is an immigrant. For years of maternal education, we can test whether within-cluster variation in friends' behavior explains differences in students' own maternal education conditional on college completion. Finally, we present counterfactuals for paternal education and nativity. Any correlation between friends' health behaviors and the student's own demographic attributes must arise from students selecting into friendships, and no correlation is suggestive of little sorting over key student attributes (Bayer et al., 2008; Bifulco et al., 2011). Tables D1-D4 present our balancing tests, which support our assumption of quasirandom variation across-grades within schools.¹³

¹¹ 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 born.

¹² One might be concerned if a lot of friendship links were between students in the same fixed effect group. If so, we would be comparing the smoking of individuals who are all friends with each other and all smoking. However, only 0.7% of our friendship links involve students in the same FE group.

¹³ Table C6 demonstrates that the balancing test attributes explain individual's own health behaviors.

4 | RESULTS

The singleton clusters discussed above do not contribute to the estimates, and we restrict our sample to clusters that contain observations in at least two cohorts in the school so that we have the same effective sample for all models.¹⁴ Table 3 presents estimates of Equation (4) for this sample on adolescent smoking and drinking where same-sex/same-grade friends are used to define the friendship network. In column 1 (OLS),¹⁵ the baseline results for smoking (row 1) suggest that increasing the share of friends who smoke by 10% points would increase own smoking by nearly 3.5% points. In column 2, we follow some of the previous literature and control for school fixed effects; however, this only reduces the coefficient from 0.350 to 0.332 for friends' smoking. In column 3, we control for school by cluster fixed effects and so control for same friendship-type choices given the same friendship opportunity set, and we observe a substantially larger decline in the estimated to 0.306.¹⁶ The fourth column incorporates individual observables into the FE definition. The fifth column presents the results for our preferred specification restricting the sample to one observation per cohort.

Estimates based on within friendship cluster by school comparisons (columns 3–5) fall between 0.306 and 0.320. The lowest estimates are associated with models that contain variation within grade, which is consistent with findings in the appendix that within-grade–within-cluster variation biases estimates in the opposite direction from the bias in OLS. Similarly in row 2, school fixed effect estimates of friends' drinking are 0.271, and the estimates after controlling for friendship cluster range between 0.250 and 0.287. Focusing on our preferred estimates using one observation per FE per cohort, we observe 8.4% and 4.4% reductions relative to our school fixed effect estimates in the effects of friends' smoking and drinking, respectively. Significantly, all reductions due to cluster fixed effects are less than 16%, which is quite small relative to declines in estimates for our balancing tests, typically between 60% and 95%.¹⁷

Next, our results might be driven by broader peer effects or other effects associated with cohort composition, and we include controls for the demographic composition of each cohort in each school over all of our control variables. These results are shown in Table 4. The inclusion of cohort level peer controls has no impact on the estimated effects of friend's health behaviors. As a further robustness test, we also add controls for the share of peers in the cohort that smoke or drink for the models of smoking and drinking, respectively. Adding these controls reduces the magnitude of the estimated effects by only 1–2%.¹⁸

¹⁴Tables C3, C4, and C5 examine the compositional effects of dropping singleton clusters.

¹⁵The linear controls include all variables used to create the friendship choice clusters.

¹⁶It is interesting to note that the inclusion of friendship cluster fixed effects produces estimates that are very similar in magnitude to Column 2. Rather, it is controlling for friendship choices conditional on the friendship opportunity set (school) that leads to the significant reductions in the estimated effect.

¹⁷Our results are robust to splitting the sample into subsamples of 9th/10th graders and 11th/12th graders reducing the time gap between comparison individuals. The 9th/10th grade subsample also addresses concerns about sample attrition due to dropouts or moves. We also investigate models where smoking or drinking depends upon both friends' smoking and friends' drinking. The estimated effect magnitudes are essentially split between the two friends' behavior variables with the majority of the effect being associated with same behavior.

¹⁸We also estimate two-way fixed effect models including school by grade fixed effects in order to capture school unobservables at the cohort level and school by grade by student-type fixed effects to allow cohort unobservables to vary across students. These controls have minimal effects reducing our estimates by 2% to 6%.

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 such as 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 relatively 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 (Calvó-Armengol, Patacchini, & Zenou, 2009; Trogon et al., 2008) that are often identified by assuming that individual network links are formed randomly. After controlling for readily observable demographics, our research is supportive of this assumption for smoking and drinking, but this result may not generalize to other behaviors or outcomes. We report additional tests of robustness based on alternative cluster definitions, subsamples by cluster size and subsamples by number of friends in Appendix E.

4.1 | Two stage models and an application to longitudinal data

If our estimates are to capture the causal effects of friends' behaviors (rather than simply assess the bias arising from correlated unobservables), an additional assumption is required—a student's own smoking behavior does not directly cause the student to form friendships with students who smoke. We can neither directly test this assumption nor credibly argue that the assumption is reasonable. In fact, Eisenberg et al. (2014) show that the eventual closeness of randomly assigned roommates is related to the risky behavior of the roommates. One reasonable option is to look at models of the onset of health-related behaviors where reverse causality is less of a concern. However, longitudinal datasets are often substantially smaller than cross-sectional samples, and the Add Health survey is no exception, with only a fraction of the students being followed over time. Therefore, the estimation of the high dimension fixed effect model in Tables 4 and 5 is not feasible.

As an alternative, we will use the fixed effect estimate from the cross-sectional analysis, presented above, as a proxy for the e_{is} associated with all individuals who are in the same cluster c . We then estimate the effect of peers on the onset of drinking or smoking conditional on the estimated fixed effect. The classic threat to identification in studies of onset is that unobservables that caused the student to select their friends also lead to later smoking, but in our case, the information generated on each cluster provides a control for those correlated unobservables.

The fixed effect estimate for each cluster is an average across all cohorts m including the cohort n from which a student is drawn, which in small samples can lead to an overcorrection from the inclusion of the fixed effect estimates. The standard solution is to calculate the control, the fixed effect, omitting the information that creates this correlation, in this case, the information from the same cohort.

$$\hat{\rho}_{cs-i} = \sum_{k \in \{c, m, s\}, m \neq n} \left(H_{kcs} - \left(\frac{1}{n_k} \sum_{j \in \Omega_{ks}} H_j \right) \hat{\beta}_1 \right), \quad (5)$$

where the mean predicted residual for the cluster c is summed over all individuals k in school s who belong to the same cluster but do not belong to the same cohort n as student i .

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, their recommended control for mitigating this bias is simply the average fixed effect estimate across all individuals in the school except the cohort-cluster to which the individual belongs.

$$\hat{\rho}_{s-i} = \sum_{k \in \{d, m, s\}, \{d, m\} \neq \{c, n\}} \left(H_{kcs} - \left(\frac{1}{n_k} \sum_{j \in \Omega_{ks}} H_j \right) \hat{\beta}_1 \right)$$

where the mean predicted residual for the school s is summed over all individuals k in school s except those who belong to the same cluster c and cohort n as student i .

The resulting estimating equation¹⁹ is

$$H_{ics} \approx \left(\frac{1}{n_i} \sum_{j \in \Omega_{is}} H_j \right) \beta_1 + \hat{\rho}_{cs-i} + \hat{\rho}_{s-i} + (\mu_{is} - \bar{\mu}_c) + (\rho_c - \hat{\rho}_{c-i} - \hat{\rho}_{s-i}). \quad (6)$$

We demonstrate that this exercise is reasonable by replicating our Wave 1 results from Table 3. We restrict the sample to school by friendship pattern by student-type cells that contain students at least three cohorts/grades so that we have information from two cohorts on which to base $\hat{\rho}_{cs-i}$. Table 5 presents models from this exercise. The first two columns present the school and the student type by friendship choice by school fixed effect estimates for the individual subsample where one individual is selected randomly per cluster for this restricted sample. Both estimates are very similar to the estimates in columns 2 and 5 of Table 3.

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. 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 addition of the constructed controls moves the estimated effects of friends' smoking significantly closer to the cluster fixed effect estimates in column 2. 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.

¹⁹All results are robust to whether or not we include the Guryan style correction.

Next, we focus on two subsamples of students who completed the Wave 2 survey of the Add Health and either did not smoke or did not drink in Wave 1. For these subsamples, we construct two health behavior onset variables that equal one if they report smoking or drinking, respectively, in Wave 2 and zero otherwise. Table 6 presents effect estimates for the onset of smoking or drinking as a function of the Wave 1 behavior of a student's Wave 1 friends. The first two columns present the OLS and school fixed effect estimates for these subsamples. The third column presents the school FE estimates for a sample restricted to student type by friendship choice by school clusters with students in three cohorts. The OLS and school fixed effect estimates are positive and statistically significant, and the sample restriction has no impact on the estimated effects.

The fourth column of Table 6 presents estimates conditional on 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. 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 with the school FE presented in column 3 with the estimated smoking effect declining by 17% and the estimated drinking effect increasing by 3% of the OLS estimates. Note that the estimate on the predicted cluster FE when the own cohort information is included is not biased upwards, unlike in Table 5, because the smokers and drinkers from the first wave who helped drive these cluster predictions have been deleted from the sample.²⁰

5 | CONCLUSIONS

Although researchers typically examine peer effects by defining the peer group broadly, this paper focuses attention on actual friends and constructs and implements a novel research design to study the effects of friend's health behaviors on own health behaviors. The main idea is to combine a cross-cohort, within school design with controls for friendship options and friendship choices using school-student type-friendship pattern fixed effects. We demonstrate that our design is successful in eliminating spurious correlations between exogenous attributes and friend's health behaviors. 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.

Further, our model of friendship formation suggests that our approach can be applied under quite general circumstances. For example, our model allows for a very general nonlinear 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 onset where the simultaneity between own health behavior and friendship behavior is less of a concern. The key assumption required to apply this identifications strategy is that individuals experience some type of shock in exposure to health behavior of potential friends. This shock assures that some variation

²⁰ Tables E2, E3, E4 report results examining the sensitivity to stratifying these sample based on whether respondents are members of "small" versus "large" clusters and whether respondents have "small" versus "large" numbers of friends.

remains in friends' health behavior even after asymptotically eliminating variation across individuals in friendship outcomes. Supporting this assumption for our application, 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 appear 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 effects. There continues to be little evidence of bias in the school FE models when we focus on a subsample of students that have larger numbers of friends or on subsamples of clusters where we observed the largest number of students, where our analysis should have the greatest potential for identifying bias. Further, counterfactual analyses of race, maternal and paternal education, and maternal and paternal nativity find at most only weak evidence of an effect of friends' behavior on predetermined attributes. Finally, our estimates are robust to cohort level controls representing aggregate peer effects.

ACKNOWLEDGEMENTS

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, Midwest Health Economics Conference, and the Second Annual Economics of Risky Behaviors (AMERB) conference. We thank Yonghong An, Michael Anderson, Tao Chen, Ethan Cohen-Cole, Chad Cotti, Bill Evans, Don Kenkel, Brian Krauth, Anna Mueller, Bruce Sacerdote, Rusty Tchernis, Gautam Tripathi, and the editor and two anonymous reviewers 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 was supported by core grants to the Center for Demography and Ecology at the University of Wisconsin-Madison (P2C HD047873) and to the Center for Demography of Health and Aging at the University of Wisconsin-Madison (P30 AG017266).

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (<http://www.cpc.unc.edu/addhealth>). No direct support was received from grant P01-HD31921 for this analysis.

Funding information

Eunice Kennedy Shriver National Institute of Child Health and Human Development, Grant/Award Number: P01-HD31921; Center for Demography of Health and Aging at the University of Wisconsin-Madison, Grant/Award Number: P30 AG017266; Center for Demography and Ecology at the University of Wisconsin-Madison, Grant/Award Number: P2C HD047873; NICHD, Grant/ Award Number: 1R21 HD066230-01A1

APPENDIX A

PARTIAL EQUILIBRIUM MODEL OF FRIENDSHIP FORMATION

We begin this appendix by repeating Equation (1):

$$H_{is} = \beta_s + \tilde{H}_{is}\beta_1 + \tilde{X}_{is}\beta_2 + X_i\beta_3 + \varepsilon_{is} + \mu_{is}, \quad (A1)$$

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_j 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 nonnegative.²¹ Further, we assume that μ_{is} and ε_{is} are idiosyncratic errors such that

Assumption 1.

$$E[\mu_{is} | \tilde{H}_{is}, \tilde{X}_{is}, X_i] = 0 \text{ and } E[\varepsilon_{is} | X_i] = 0.$$

Although μ_{is} affects the individual's own health behavior H_{is} , it is assumed to not directly affect the health behavior of the individual's friends $\tilde{H}_{is}$. It is important, however, to put this assumption into context. Given the presence of ε_{is} in Equation (A1), the first part of Assumption 1 simply implies that the unobservable in Equation (A1) can be decomposed into additively separable components: one endogenous component that affects choices of friends over unobserved attributes X_j and a second exogenous, idiosyncratic disturbance to health behavior.

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

$$\Pr_{is}[X_j = x, H_j = h | X_j, \varepsilon_{is}, j \in K_i] = f_{sxh}(X_i, \varepsilon_{is}, \pi_{is}), \quad (\text{A2})$$

where π_{is} is an additional unobservable that does not enter Equation (A1) and so does not directly affect health behaviors 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 because the social environment varies across schools. The four probabilities must sum to one for a given school for any value of the functions' arguments. In summary, own health behaviors H and friends' types X depend upon a common error component ε_{is} , but friend's health behaviors are affected by an additional unobservable π_{is} .

Assumption 1 requires a source of variation in the health behavior of friends that cannot be explained by student friendship choices. Everyone in a cluster has the same demographic composition of friends, and as the number of friends becomes large, the average friends' health behavior for everyone in the cluster should converge to the expected incidence of health behavior for that composition unless there is some shock to health behavior that cannot be explained by friendship choices. We rewrite Equation (4) to allow each individual to belong to a different cohort or grade g in the same school and divide peer health behavior

²¹We also note that an alternative specification might involve a single unobservable each for determining health behavior and friendship outcomes. The specification in 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) in the main text and a friendship formation model, $\tilde{\mu}_{is}$ and ε_{is} , are correlated, and then we can define μ_{is} as $\tilde{\mu}_{is} - E[\tilde{\mu}_{is} | \varepsilon_{is}]$ where we assume that the $E[\tilde{\mu}_{is} | \varepsilon_{is}] = \alpha_0 + \alpha_1 \varepsilon_{is}$ 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.

into a predicted component based on observables ($\widehat{H}_{js}$), the unobservable attributes that affect health behavior (ε_{jgcs}), and a cohort specific shock to health behavior (π_{gcs}):

$$H_{igcs} \approx \left(\frac{1}{n_i} \sum_{j \in \Omega_{is}} (\widehat{H}_{js} + \varepsilon_{jgcs} + \pi_{gcs}) \right) \beta_1 + \rho_{cs} + \mu_{igcs}. \quad (A3)$$

In Equation (A1), the cohort specific term π_{gcs} creates variation in the health behavior of friends that unlike ε_{igcs} does not limit to zero as Ω_{is} grows in size.

In practical terms, imagine two similar individuals (on observables and unobservables) in the same school and in different grades, but grades that have very similar alcohol usage. As these two individuals make more friendship choices, they will reveal their similarity on unobservables, but as those friendship choices increase, any random variation in the health behaviors of friends will also be averaged out so that asymptotically, these individuals are also exposed to the same health behaviors. In terms of Equations (3) and (A1), if unobservable attributes ε_{jgcs} are the primary source of unexplained variation in H_{js} , then the inclusion of the school by friendship choice by student-type fixed effects ρ_{cs} will simultaneously eliminate both the source of bias, unexplained variation in own health behavior ε_{js} , and the source of signal, unexplained variation in ε_{jgcs} , through averaging the health behavior over many friends. Only when a portion of the variation arises from environmental factors π_{gcs} that are persistent regardless of the number of friendship choices can the inclusion of the fixed effects isolate a meaningful signal associated with the effect of exposure to friends.²² The data section in the main text provides a careful discussion of the variation in health behavior across cohorts in our sample.

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) and 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 unobservables might be correlated with X_i , some variation must remain in the unobservables associated with forming friends with specific health behaviors that does not enter own health behavior after conditioning on X_i . These assumptions can be summarized as follows

Assumption 2.

$$\frac{\partial f_{s11}}{\partial H} = 0, \frac{\partial f_{s11}}{\partial H} = 0 \frac{\partial f_{s11}}{\partial H} = 0 \frac{\partial f_{s11}}{\partial H} = 0$$

²²Student's unobservable attributes and cohort environmental factors may be correlated. Consistency only requires variation in the environmental factors that cannot be entirely explained by the unobservable attributes.

Assumption 2 is that individuals cannot select into friendships based on their own health behaviors. Obviously, this assumption is quite strong, and we are highlighting this assumption to explain how we will address concerns about violations, as opposed to attempting to justify this assumption. The direct implication is that our estimates provide insights into the bias arising from sorting on unobservable student traits. In order to address reverse causality, later in the paper, we estimate a model of the onset of health behaviors using information on the friendships that were in place prior to onset. Specifically, we model the adoption of drinking or smoking in a second period for a subsample of students who were nondrinkers or nonsmokers in the first period. Onset of a health behavior in the second period depends in the first period on exposure to friends who drink or smoke conditional on a control for unobserved variables, the estimated first period fixed effects from Equation (4).

$$H_{igcst+1} \approx \left(\frac{1}{n_{it}} \sum_{j \in \Omega_{ist}} H_{jst} \right) \beta_1 + \hat{\rho}_{cst} + \mu_{igcst+1} \text{ where } H_{igscst} = 0 \quad (A4)$$

Because the friendships had formed prior to the onset of the health behavior, the health behavior could not have caused the friendship. The classic concern about using time lags to identify social interaction effects has always been that common unobservables cause both the social interactions and any behavioral changes. In our case, identification from lags is less problematic because the primary identification strategy is designed to mitigate the impact of correlation unobservables.

Assumption 3.

$$\frac{\partial f_{s11}}{\partial \varepsilon} + \frac{\partial f_{s10}}{\partial \varepsilon} > 0 \text{ and } \frac{\partial f_{s11}}{\partial \varepsilon} + \frac{\partial f_{s01}}{\partial \varepsilon} > 0.^{23}$$

Assumption 3 is monotonicity, which is also a central assumption in Dale and Krueger (2002) and Fu and Ross (2013). Unobservables that affect health behavior must have a monotonic relationship with friendship choices over friends' attributes X_j . Without monotonicity, multiple values of the unobservable might be consistent with the observed friendship choices for observationally equivalent individuals. For example, risk taking could be an unobservable that drives health behaviors but is nonmonotonic if, for example, the least risk averse and the most risk averse students both end up with fewer friends than individuals who take some risks in order to make friends, but do not take so many risks that they end up driving friends away.

Although monotonicity assumptions are often strong and untestable, our general model is designed to allow for substantial nonmonotonic sorting over observable attributes. Our

²³The assumption of a positive relationship between good type and the individual friendship formation propensity y_{js} 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_{js} 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 e_{js} in Equation (A1)

friendship clusters are defined to assure that observationally equivalent individuals in the same cluster have made the same friendship choices on key observables. Therefore, the probability of a specific friendship composition arising is a nonparametric function of the full set of attributes used to define student type for that particular school, and monotonicity is only imposed on the unobserved attributes. In this way, our monotonicity assumption is no stronger than the monotonicity assumptions needed to estimate traditional sample selection models when exclusion restrictions are available. Unlike our risk-taking example, it is harder to explain situations where two White students both are likely to have friendships with African-American students because one student has unobserved attributes that contribute strongly to unhealthy behaviors and because the other has unobserved attribute that contribute strongly to healthy behaviors, and these two extremes both contribute to forming across race or across social class friendships.

Assumption 4.

$$\frac{\partial f_{s11}}{\partial \pi} = -\frac{\partial f_{s10}}{\partial \pi} > 0, \frac{\partial f_{s01}}{\partial \pi} = -\frac{\partial f_{s00}}{\partial \pi} > 0 \text{ and } \text{Var}[\pi_{is}|X_i] = 0$$

As discussed above, the monotonicity requirement in Assumption 3 is critical for all results and will be maintained throughout. The requirement that own health behavior does not affect friendship formation in Assumption 2 is necessary in order to be able to impose Assumption 1: Because without this assumption, μ_{is} would affect friends' health behavior. We will examine the implications of relaxing Assumption 2 in the next section 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. The implications of violating Assumption 4 are further examined in the appendix.

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

$$\frac{1}{n_i} \sum_{j \in \Omega_{is}} X_{js} = \frac{1}{n_k} \sum_{j \in \Omega_{ks}} X_{js} \text{ for all } i \text{ and } k \text{ in cluster } c \text{ and there exist no individuals } l$$

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 of friends because students choose over many demographic attributes including

race, ethnicity, and maternal education. Further, later in the paper, we demonstrate that our empirical results are robust to focusing on the subsample of students with the largest number of friends.²⁴

Proposition 1.

Under Assumptions 1 through 4 plus Definition 1, the within-cluster estimator from Equation (4) is asymptotically identified, and the bias arising from this cluster fixed-effects model limits to zero as n_i becomes large for all i in the sample.

Proof. See final section of the appendix.

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

$$\phi_1^c = \frac{\text{Cov}[\varepsilon_{is} - \bar{\varepsilon}_c(\bar{H}_{is} - \bar{H}_c)]}{\text{Var}[(\bar{H}_{is} - \bar{H}_c)]} \quad (\text{A5})$$

where $\bar{\varepsilon}_c$ and $\bar{H}_c$ are the cluster specific mean of ε_{is} and $\bar{H}_{is}$, 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, as the within cluster variation in ε limits to zero, the within-cluster variance of $\bar{H}_{is}$ will also limit to zero unless $\bar{H}_{is}$ contains variation associated with π_{is} . In other words, there will be no within-cluster variation in the variable of interest without some shock to health behavior that is independent of friendship formation. Therefore, the nonzero 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 $\bar{H}_{is}$

$$\bar{H}_{is} = \lambda_s + \bar{X}_{is}\lambda_1 + X_i\lambda_2 + V_{is} \quad (\text{A6})$$

such that $V_{is} = V(\bar{X}_{is}, X_i, \varepsilon_{is}, \pi_{is})$. 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.

$$E[V_{is} | \bar{X}_{is}, X_i] = 0 \text{ and } \text{Var}[V_{is} | \delta_c] \leq \text{Var}[V_{is}]$$

The first part of Assumption 5 implies that

$$E[\widetilde{H}_{is} | \widetilde{X}_{is}, X_i] = \lambda_s + \widetilde{X}_{is}\lambda_1 + X_i\lambda_2 \quad (A7)$$

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 $\widetilde{H}_{is}$ and $\widetilde{X}_{is}$ for a given X_i .

This assumption would be standard if $\widetilde{X}_{is}$ 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 empirically 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 (A1).

Proof. See final section of the 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 individuals 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. Assumption 4 is the reason that our identification strategy requires across cohort variation in the exposure to the health behaviors of other students. The across cohort variation is what assures that different individuals in the same cluster exogenously face independent shocks π_{is} .

A.1 | 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 Assumptions 1 and 2 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

$$\Pr_{is}[X_j = x, H_j = h | H_i, X_i, \varepsilon_{is}, \pi_{is}, j \in K_i] = f_{sxh}(H_i, X_i, \varepsilon_{is}, \pi_{is}) \quad (A8)$$

with

Assumption 6.

$$E[\mu_{is} | \widetilde{X}_{is}, X_i] = 0 \text{ and } E[\varepsilon_{is} | X_i] = 0$$

Assumption 7.

$$\frac{\partial f_{s11}}{\partial H} = -\frac{\partial f_{s10}}{\partial H} > 0 \text{ and } \frac{\partial f_{s01}}{\partial H} = -\frac{\partial f_{s00}}{\partial H} > 0$$

Therefore, the idiosyncratic error μ_{is} does not have a conditional expectation of zero because it influences the health behavior of friends $\widetilde{H}_j$ 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 3 through 7 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 a second term that captures the correlation between μ_{is} and $\widetilde{H}_{is}(\varphi_1)$ as n_i becomes large for all i in the sample.

Proof: See final section of the 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 nonzero, 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.

APPENDIX B**PROOF OF PROPOSITIONS AND GENERALIZATION OF THE SHOCK TO FRIENDSHIP FORMATION****Proposition 1.**

Under Assumptions 1 through 4 plus Definition 1, the within-cluster estimator from Equation (4) is asymptotically identified and the bias arising from this cluster fixed-effects model limits to zero as n_i becomes large for all i in the sample.

Proof: First, based on Equations (A1) and (A2), 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 (A1), we write the expectation of Equation (A1) as

$$E[H_{is}|\widetilde{H}_{is}, \widetilde{X}_{is}, X_i] = \beta_s + \widetilde{H}_{is}\beta_1 + \widetilde{X}_{is}\beta_2 + X_{it}\beta_3 + E[\varepsilon_{is}|\widetilde{H}_{is}, \widetilde{X}_{is}, X_i] \quad (B1)$$

and substitute the linear projection of ε_{is} on the conditioning variables, $\varphi_s + \widetilde{H}_{is}\varphi_1 + \widetilde{X}_{is}\varphi_2 + X_{it}\varphi_3$, into Equation (A1).²⁵ This yields

$$E[H_{is}|\widetilde{H}_j, \widetilde{X}_{is}, X_i] = (\beta_s + \varphi_s) + \widetilde{H}_{is}(\beta_1 + \varphi_1) + \widetilde{X}_{is}(\beta_2 + \varphi_2) + X_{it}(\beta_3 + \varphi_3) \quad (B2)$$

Based on this linear projection, we define the bias in the estimated coefficient on peer health behavior, $\widetilde{H}_j$, as²⁶

$$\phi_1 = \frac{Cov[\varepsilon_{is}, \widetilde{H}_{is} - E[\widetilde{H}_{is}|\widetilde{X}_{is}, X_i]]}{Var[\widetilde{H}_{is} - E[\widetilde{H}_{is}|\widetilde{X}_{is}, X_i]]} \quad (B3)$$

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

$$\delta_c = \overline{H}_c\beta_1 + \widetilde{X}_{is}\beta_2 + X_{it}\beta_3 + \bar{\varepsilon}_c + \bar{\mu}_c \quad (B4)$$

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

After controlling for cluster fixed effects in Equation (A1), the health behavior model takes the following mean differenced form:

$$(H_{is} - \overline{H}_c) = (H_{is} - \overline{H}_c)\beta_1 + (\varepsilon_{is} - \bar{\varepsilon}_c) + (\mu_{is} - \bar{\mu}_c) \quad (B5)$$

because the cluster mean of individual health behaviors equals the cluster mean of the friends health behaviors by construction. The bias associated with the estimated coefficient on $(\widetilde{H}_{is} - \overline{H}_c)$ in this model is

$$\phi_1^c = \frac{Cov[\varepsilon_{is} - \bar{\varepsilon}_c, (\widetilde{H}_{is} - \overline{H}_c) - E[(\widetilde{H}_{is} - \overline{H}_c)|\delta_c]]}{Var[(\widetilde{H}_{is} - \overline{H}_c) - E[(\widetilde{H}_{is} - \overline{H}_c)|\delta_c]]} = \frac{Cov[\varepsilon_{is} - \bar{\varepsilon}_c, (\widetilde{H}_{is} - \overline{H}_c)]}{Var[(\widetilde{H}_{is} - \overline{H}_c)]} \quad (B6)$$

²⁵This 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 $\widetilde{H}_{is}$, $\widetilde{X}_{is}$, X_{it} and ε_{is} , and ϕ_1 is the relative bias in those estimates if one is unable to condition on ε_{is} .

²⁶This arises from the standard omitted variables formula for a regressor that is orthogonal to all other regressors, and orthogonality is obtained using a conditioning argument where $y = \alpha_1 Z + \alpha_2 X + u$ can be rewritten as the following conditional regression $y = \alpha_1(Z - E[Z|X]) + \alpha_1 E[Z|X] + \alpha_2 X + u$.

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, that is, observed attributes (X_i) or friendship choices that proxy 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

$$\Pr_{is}[X_j = 1] = f_{s11}(X_i, \varepsilon_{is}, \pi_{is}) + f_{s10}(X_i, \varepsilon_{is}, \pi_{is}) = f_s^X(X_i, \varepsilon_{is}) \quad (\text{B7})$$

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

$$\lim_{n_i \rightarrow \infty} \bar{X}_{is} = f_s^X(X_i, \varepsilon_{is}) \quad (\text{B8})$$

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

Because all individuals in cluster c have the same observable type X_{is} and the same fraction of good type friends, $\bar{X}_{is}$, Equation (B8) implies that

$$f_s^X(X_{is}, \varepsilon_{is}) = f_s^X(X_{is}, \varepsilon_{ks}) \quad \text{for all } i, k \in c \quad (\text{B9})$$

when the number of friends is large.

However, under the monotonicity assumption (Assumption 2), Equation (B9) can only hold if $\varepsilon_{is} = \varepsilon_{ks}$ for all i and k in the cluster, and so from Equation (B6)

$$\lim_{n_i \rightarrow \infty} \phi_1^c = \frac{\lim_{n_i \rightarrow \infty} \text{Cov}[\varepsilon_{is} - \bar{\varepsilon}_c, (\tilde{H}_{is} - \bar{H}_c)]}{\lim_{n_i \rightarrow \infty} \text{Var}[(\tilde{H}_{is} - \bar{H}_c)]} = 0 \quad (\text{B10})$$

because the within cluster variation in ε limits to zero while the within-cluster variance of $\tilde{H}_{is}$ contains variation associated with π and so is strictly positive.²⁷

The parameter estimate of β_1 will be asymptotically identified only if there is within-cluster variation in friends' health behaviors or $\text{Var}[\tilde{H}_j | \delta_c] \neq 0$. Assumption 4 requires that $\text{Var}[\pi_{is}]$

²⁷As the number of friends becomes large, the cluster fixed-effect model acts as a nonparametric 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 (A1) as $(\varepsilon_{is}, \mu_{is}) \perp (\tilde{H}_{is}, \tilde{X}_{is}, X_i) | \delta_c$, 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.

$X_{ij} = 0$ as a result for any individual i and k in the same cluster, but different cohorts, will have different patterns of friendship health behavior, that is,

$$\lim_{n_i \rightarrow \infty} f_{sxh}(X_i, \varepsilon_{is}, \pi_{is}) \neq \lim_{n_i \rightarrow \infty} f_{sxh}(X_i, \varepsilon_{is}, \pi_{ks}) = \lim_{n_i \rightarrow \infty} f_{sxh}(X_k, \varepsilon_{ks}, \pi_{ks})$$

As the number of friendship choices become large, the two individuals in the same cluster have the same value of ε_{is} and the probability limits to a continuous frequency. The continuity of the distribution of π_{is} assumes that $\pi_{is} = \pi_{ks}$ except over a space of measure zero and so variation in friends' health behaviors will exist between any two individuals in the same cluster.

Proposition 2.

Under Assumptions 1 through 5 plus Definition 1, the bias arising from estimating the cluster fixed-effects model in Equations (5) and (6) has the same sign and is smaller than the bias that arises for the OLS model described in Equation (A1).

Proof: Using Equation (A6), the bias from the cohort fixed effect model in Equation (B6) reduces to

$$\phi_1^c = \frac{\text{Cov}[\varepsilon_{is} - \bar{\varepsilon}_c, (\bar{H}_{is} - \bar{H}_c)]}{\text{Var}[(\bar{H}_{is} - \bar{H}_c)]} = \frac{\text{Cov}[\varepsilon_{is} - \bar{\varepsilon}_c, V_{is} - \bar{V}_c]}{\text{Var}[V_{is} - \bar{V}_c]} \quad (\text{B11})$$

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

$$\text{Var}[\bar{V}_c] = \frac{1}{m_i} \text{Var}[V_{is}] + \frac{m_i - 1}{m_i} \text{Cov}[V_{is}, V_{ks} | i, k \in c] = \quad (\text{B12})$$

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

Similarly,

$$\text{Cov}[V_{is}, \bar{V}_c] = \frac{1}{m_i} \text{Var}[V_{is}] + \frac{m_i - 1}{m_i} \text{Cov}[V_{is}, V_{ks} | i, k \in c] \quad (\text{B13})$$

so that the denominator of Equation (B11) takes the form

$$\text{Var}[V_{is} - \bar{V}_c] = \left(1 - \frac{1}{m_i}\right) (\text{Var}[V_{is}] - \text{Cov}[V_{is}, V_{ks} | i, k \in c]) \quad (\text{B14})$$

Turning to the numerator of Equation (B11), 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

$$Cov[\varepsilon_{is}, \bar{V}_c] = \frac{1}{m_i} Cov[\varepsilon_{is}, V_{is}] + \frac{m_i - 1}{m_i} Cov[\varepsilon_{is}, V_{ks} | i, k \in c] \quad (B15)$$

Using all three covariance terms,

$$Cov[\varepsilon_{is} - \bar{\varepsilon}_c, V_{is} - \bar{V}_c] = \left(1 - \frac{1}{m_i}\right) (Cov[\varepsilon_{is}, V_{is}] - Cov[\varepsilon_{is}, V_{ks} | i, k \in c]) \quad (B16)$$

and Equation (B11) can be rewritten using Equations (B14) and (B16) as

$$\phi_1^c = \frac{(Cov[\varepsilon_{is}, V_{is}] - Cov[\varepsilon_{is}, V_{ks} | i, k \in c])}{(Var[V_{is}] - Cov[V_{is}, V_{ks} | i, k \in c])} \quad (B17)$$

Next, using Equations (A6) and (A7), the OLS bias in Equation (B3) reduces to

$$\phi_1 = \frac{Cov[\varepsilon_{is}, \bar{H}_{is} - E[\bar{H}_{is} | \bar{X}_{is}, X_i]]}{Var[\bar{H}_{is} - E[\bar{H}_{is} | \bar{X}_{is}, X_i]]} = \frac{Cov[\varepsilon_{is}, V_{is}]}{Var[V_{is}]} \quad (B18)$$

Note that the first terms in the numerator and denominator in Equation (B17) are the same as the numerator and denominator in Equation (B18). Equation (B17) will be smaller in magnitude than Equation (B18) or $|\phi_1^c| < |\phi_1|$ if the relative or percentage reduction in the first numerator term caused by the second numerator term in Equation (B17) is smaller than the equivalent reduction in the denominator or if²⁸

$$\frac{Cov[\varepsilon_{is}, V_{ks} | i, k \in c]}{Cov[\varepsilon_{is}, V_{is}]} > \frac{Cov[V_{is}, V_{ks} | i, k \in c]}{Var[V_{is}]} \quad (B19)$$

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

$$V_{is} = \xi_s + \varepsilon_{is} \xi_1 + U_{is} \quad (B20)$$

where $U_{is} = U(\bar{X}_{is}, X_i, \varepsilon_{is}, \pi_{is})$ and $Cov[\varepsilon_{is}, U_{is}] = 0$.

Further, $Cov[\varepsilon_{is}, U_{ks} | i, k \in c] = 0$ and $Cov[U_{is}, U_{ks} | i, k \in c] = 0$ both also equal zero because the all sources of a linear relationship between the $\bar{H}$'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.

Using Equation (B20) and the above results, we can rewrite Equation (B19) as

$$\frac{Cov[\varepsilon_{is}, \varepsilon_{ks} | i, k \in c]}{Var[\varepsilon_{is}]} > \frac{Cov[\varepsilon_{is}, \varepsilon_{ks} | i, k \in c]}{Var[\varepsilon_{is}] + \frac{1}{\xi_1^2} Var[U_{is}]} \quad (B21)$$

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

From Equation (B1) 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 $\tilde{X}_{is}$ must also increase monotonically²⁹ with ε_{is} . Therefore, we can express the fraction of good type friends as a monotonic function of ε_{is} and a stochastic variable of unknown form

$$\tilde{X}_{is} = g_s^X(X_i, \varepsilon_{is}, v_{is}). \quad (B22)$$

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

$$g_s^X(X_i, \varepsilon_{is}, v_{is}) = g_s^X(X_i, \varepsilon_{ks}, v_{ks}), \quad (B23)$$

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

The implicit function theorem and monotonicity assumption allows us to rewrite (B13) as

$$\varepsilon_{is} = g_e^{-1}(X_i, v_{is}, g_s^X(X_k, \varepsilon_{ks}, v_{ks})) \equiv \tilde{g}(\varepsilon_{ks}, X_i, v_{is}, v_{ks}), \quad (B24)$$

where g_e^{-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 because 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

$$Cov[\varepsilon_{is}, \varepsilon_{ks} | i, k \in c] = Cov[\tilde{g}(\varepsilon_{ks}, X_i, v_{is}, v_{ks}), \varepsilon_{ks} | i, k \in c] > 0 \quad (B25)$$

²⁹The following argument also holds for a monotonically decreasing function.

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

In order to sign the cohort fixed effects bias in Equation (B17) relative to the OLS bias in Equation (B18), we substitute Equation (B20) in the numerators of the bias expressions. For OLS, the expression reduces to

$$\text{sign}(\phi_1) = \text{sign}\left(\frac{\xi_1 \text{Var}[\varepsilon_{is}]}{\text{Var}[V_{is}]}\right) = \text{sign}(\xi_1) \quad (\text{B26})$$

which takes the same sign as ξ_I .

For the cohort fixed effects model,

$$\text{sign}(\phi_1^c) = \text{sign}\left(\frac{\xi_1(\text{Var}[\varepsilon_{is}] - \text{Cov}[\varepsilon_{is}, \varepsilon_{ks} | i, k \in c])}{(\text{Var}[V_{is}] - \text{Cov}[V_{is}, V_{ks} | i, k \in c])}\right) = \text{sign}(\xi_1) \quad (\text{B27})$$

which will take the same sign as ξ_I 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,

$$\text{Var}[V_{is}] \geq \text{Var}[V_{is} | \delta_c] \geq \text{Cov}[V_{is}, V_{ks} | i, k \in c] \quad (\text{B28})$$

B.1 | 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: (a) To illustrate that Assumption 4 is crucial for our identification strategy and (b) 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_j as

$$\Pr_{is}[X_j = x, H_j = h | X_i, \varepsilon_{is}, \pi_{is}, j \in K_i] = f_{sxh}(X_i, y_{is} = \varepsilon_{is} + \alpha\pi_{is}, \pi_{is}) \quad (\text{B29})$$

and replace Assumptions 3 and 4 with

Assumption 7.

$$\frac{\partial f_{s11}}{\partial y} + \frac{\partial f_{s10}}{\partial y} > 0, \frac{\partial f_{s11}}{\partial y} + \frac{\partial f_{s01}}{\partial y} > 0, \frac{\partial f_{s11}}{\partial \pi} = -\frac{\partial f_{s10}}{\partial \pi} > 0 \text{ and } \frac{\partial f_{s01}}{\partial \pi} = -\frac{\partial f_{s01}}{\partial \pi} > 0$$

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,

$$\lim_{n_i \rightarrow \infty} \bar{X}_{is} = f_S^X(X_i, \varepsilon_{is} + \alpha\pi_{is}) \quad (\text{B30})$$

and

$$f_S^X(X_{is}, \varepsilon_{is} + \alpha\pi_{is}) = f_S^X(X_{is}, \varepsilon_{ks} + \alpha\pi_{ks}) \quad \text{for all } i, k \in c \quad (\text{B31})$$

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

$$\varepsilon_{is} + \alpha\pi_{is} = \varepsilon_{ks} + \alpha\pi_{ks} \quad \text{for all } i, k \in c \quad (\text{B32})$$

Further, Equation (B32) implies that

$$\varepsilon_{is} - \bar{\varepsilon}_c = \alpha(\pi_{is} - \bar{\pi}_c) \quad (\text{B33})$$

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 (A6) in terms of the relevant disturbances as

$$V_{is} = \zeta_0 + \varepsilon_{isS_1} + \pi_{isS_2} + \bar{U}_{is} \quad (\text{B34})$$

And using Equation (B23)

$$V_{is} - \bar{V}_c = (\varepsilon_{is} - \bar{\varepsilon}_c)\zeta_1 + (\pi_{is} - \bar{\pi}_c)\zeta_2 + (\bar{U}_{is} - \bar{U}_c) = (\varepsilon_{is} - \bar{\varepsilon}_c)(\zeta_1 - \zeta_2/\alpha) \quad (\text{B35})$$

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

The bias from the cohort fixed effect model as shown in Equation (B11) can be rewritten using Equation (B35) as

$$\phi_1^c = \frac{\text{Cov}[\varepsilon_{is} - \bar{\varepsilon}_c, V_{is} - \bar{V}_c]}{\text{Var}[V_{is} - \bar{V}_c]} = \frac{(\zeta_1 + \zeta_2/\alpha)\text{Var}[\varepsilon_{is} - \bar{\varepsilon}_c]}{(\zeta_1 + \zeta_2/\alpha)\text{Var}[\varepsilon_{is} - \bar{\varepsilon}_c] + \text{Var}[\bar{U}_{is} - \bar{U}_c]} \quad (\text{B36})$$

The same substitution into the OLS bias expression from Equation (B19) yields

$$\varphi_1 = \frac{\text{Cov}[\varepsilon_{is}, V_{is}]}{\text{Var}[V_{is}]} = \frac{\zeta_1 \text{Var}[\varepsilon_{is}]}{\zeta_1 \text{Var}[\varepsilon_{is}] + \text{Var}[\bar{U}_{is}]} \quad (\text{B37})$$

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

In general, Proposition 1 will not hold for arbitrary values of the underlying parameters because the presence π_{is} of 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 ($\zeta_1 > 0$), 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 (ζ_2), then $\zeta_1 + \zeta_2/\alpha$ is opposite sign of ζ_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 noncohort 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 a second term that captures the correlation between μ_{is} and $\bar{H}_{is}(\varphi_1)$ 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

$$E[H_{is} | \bar{H}_{is}, \bar{X}_{is}, X_i] = \beta_s + \bar{H}_{is}\beta_1 + \bar{X}_{is}\beta_2 + X_{it}\beta_3 + E[\varepsilon_{is} + \mu_{is} | \bar{H}_{is}, \bar{X}_{is}, X_i] \quad (\text{B38})$$

$$E[\mu_{is} | \bar{H}_{is}, \bar{X}_{is}, X_i] = \phi_s + \bar{H}_{is}\phi_1 \quad (\text{B39})$$

$$E[H_{is} | \bar{H}_{is}, \bar{X}_{is}, X_i] = (\beta_s + \varphi_s + \phi_s) + \bar{H}_{is}(\beta_1 + \varphi_1 + \phi_1) + \bar{X}_{is}(\beta_2 + \varphi_2) + X_{it}(\beta_3 + \varphi_3) \quad (\text{B40})$$

where using the expansion in Equation (A6), the new bias term may be expressed as

$$\phi_1 = \frac{\text{Cov}[\mu_{is}, \bar{H}_{is} - E[\bar{H}_{is} | \bar{X}_{is}, X_i]]}{\text{Var}[\bar{H}_{is} - E[\bar{H}_{is} | \bar{X}_{is}, X_i]]} = \frac{\text{Cov}[\mu_{is}, V_{is}]}{\text{Var}[V_{is}]} \quad (\text{B41})$$

Taking the expectation of the cluster fixed-effects model in Equation (B5) yields

$$E[H_{is} | \bar{H}_{is} - \bar{H}_c, \delta_c] = (\bar{H}_{is} - \bar{H}_c)\beta_1 + \delta_c + E[(\varepsilon_{is} - \bar{\varepsilon}_c) | \bar{H}_{is} - \bar{H}_c, \delta_c] + E[(\mu_{is} - \bar{\mu}_c) | \bar{H}_{is} - \bar{H}_c, \delta_c] \quad (\text{B42})$$

The form of the bias in the estimated coefficient on $(\bar{H}_{is} - \bar{H}_c)$ that is associated with the expectation over $(\varepsilon_{is} - \bar{\varepsilon}_c)$ has been previously defined in Equation (B6). Again, exploiting the expansion in Equation (A6), the bias associated with the expectation over $(\mu_{is} - \bar{\mu}_c)$ is

$$\phi_1^c = \frac{\text{Cov}[\mu_{is} - \bar{\mu}_c, (\bar{H}_{is} - \bar{H}_c) - E[(\bar{H}_{is} - \bar{H}_c) | \delta_c]]}{\text{Var}[(\bar{H}_{is} - \bar{H}_c) - E[(\bar{H}_{is} - \bar{H}_c) | \delta_c]]} = \frac{\text{Cov}[\mu_{is} - \bar{\mu}_c, V_{is} - \bar{V}_c]}{\text{Var}[V_{is} - \bar{V}_c]} \quad (\text{B43})$$

By Equation (B40), the expectation of the estimate of the effect of friends' health behavior in the cluster fixed effects model is $(\beta_1 + \varphi_1^c + \phi_1^c)$, and Proposition 1 establishes that φ_1^c limits to zero with the number of friends.

Following the derivations in Equations (B12) through (B17) except for μ instead of ε yields

$$\phi_1^c = \frac{(\text{Cov}[\mu_{is}, V_{is}] - \text{Cov}[\mu_{is}, V_{ks} | i, k \in c])}{(\text{Var}[V_{is}] - \text{Cov}[V_{is}, V_{ks} | i, k \in c])} \quad (\text{B44})$$

However, membership in the cluster c only depends upon X_{is} and $\bar{X}_{is}$ and so provides no information concerning the expectation of either μ_{is} or V_{is} because 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,

$$\phi_1^c = \frac{\text{Cov}[\mu_{is}, V_{is}]}{\text{Var}[V_{is}]} = \varphi_1 \quad (\text{B45})$$

and

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

APPENDIX C

SAMPLE CONSTRUCTION AND DATA NOTES

Beginning with the 90,000 Add Health³⁰ Wave 1 in school sample, we created our analysis sample in several steps. Almost 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 based on incomplete information (nicknames rather than names, or name not on the school roster). Finally, 11,000 students are dropped because they have no same-sex/same-grade level friends.³¹ Most previous studies of friendship networks limit the network definition to same-sex friends because opposite-sex nominations may represent romantic relationships. We limit our analysis to same-grade friends in order to use cross-cohort variation in friendship opportunities. These issues yield a sample of approximately 55,000 students with friends. In Table C1, 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. Table C2 presents an analysis of the correlates of being dropped from the sample for these reasons.³²

Tables on Sample Composition

TABLE C1

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	2,651	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

³⁰The larger study contains an in-school questionnaire administered to a nationally representative sample of students in Grades 7 through 12 in 1994–1995 and three in-home surveys that focus on a subsample of students in 1995 (Wave 1), and approximately 1 year later (Wave 2) and then 6 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.

³¹4,300 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).

³²Briefly, race, gender, family structure, and missingness on other variables predict sample selection in to the original 66,000 observations to some extent; however, health behaviors are not robust, important predictors. In regard 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% point increase in the probability).

% Smoke	Freq.	Percent	Cum.	% Drink	Freq.	Percent	Cum.
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	

Note. The results identify the number and frequency of students who have friends with an average incidence of health behavior falling into a particular bin. The bins run from the number listed to any number less than the number in the next row, but because individuals at most have five same-gender friends, the top and bottom rows capture the number of people with friends where either none smoke or drink or all smoke or drink. Similarly, the rows associated with 0.20–0.29 and 0.80–0.89 both capture individuals with five friends where four exhibit one behavior and the other does not, but 0.20–0.29 also capture individuals with four friends where one smokes. Sample sizes are based on the total number of individuals with at least one same-grade/same-gender friend who reported their drinking or smoking behavior.

TABLE C2

Predictors of dropped sample

Outcome	Means	No ID	No friend nominations	No found nominations	Any drop	No same grade	No same-grade/same
Means		0.047	0.14	0.20	0.24	0.065	0.129
Age	15	0.005*** (0.001)	0.026*** (0.003)	0.016*** (0.002)	0.040*** (0.003)	0.032*** (0.002)	0.024*** (0.002)
Male	0.502	0.007*** (0.001)	0.072*** (0.005)	0.016*** (0.002)	0.083*** (0.006)	0.030*** (0.003)	0.007*** (0.002)
Hispanic	0.155	0.004 (0.003)	0.014** (0.006)	0.002 (0.004)	0.020*** (0.007)	−0.006 (0.006)	−0.001 (0.004)
Black	0.19	0.007*** (0.002)	0.037*** (0.006)	0.027*** (0.005)	0.060*** (0.007)	0.017*** (0.005)	0.013*** (0.004)
Asian	0.056	−0.003 (0.004)	0.002 (0.009)	0.001 (0.009)	0.004 (0.013)	−0.038*** (0.008)	−0.017*** (0.005)
Native born	0.9	−0.003 (0.003)	−0.023*** (0.005)	−0.027*** (0.006)	−0.042*** (0.007)	−0.007 (0.008)	−0.018*** (0.006)
Live with mom	0.92	0.037*** (0.011)	0.292*** (0.019)	0.059*** (0.017)	0.306*** (0.023)	0.152*** (0.008)	0.067*** (0.006)
Mom education	13.36	−0.000 (0.000)	−0.001* (0.001)	−0.001 (0.000)	−0.002*** (0.001)	−0.001** (0.000)	−0.001*** (0.000)
Mom care	4.76	−0.002* (0.001)	−0.009*** (0.002)	−0.005*** (0.002)	−0.013*** (0.002)	−0.006*** (0.002)	−0.006*** (0.002)
Smoke	0.36	0.005** (0.002)	0.007** (0.003)	0.011*** (0.002)	0.017*** (0.003)	0.024*** (0.003)	0.013*** (0.002)
Drink	0.55	0.000 (0.001)	−0.020*** (0.003)	−0.001 (0.002)	−0.017*** (0.003)	0.001 (0.003)	−0.004 (0.002)
% Black	0.19	−0.010 (0.031)	−0.110 (0.084)	0.015 (0.051)	−0.097 (0.107)	−0.046 (0.036)	−0.043 (0.035)
% Hispanic	0.16	0.036 (0.070)	0.089 (0.084)	0.067 (0.054)	0.162* (0.097)	−0.065 (0.049)	−0.035 (0.049)
% Mom grad	0.32	0.040 (0.035)	−0.064 (0.052)	−0.032 (0.032)	−0.062 (0.065)	0.002 (0.042)	−0.035 (0.036)

Outcome	Means	No ID	No friend nominations	No found nominations	Any drop	No same grade	No same-grade/same
% Smoke	0.36	0.060 (0.042)	0.042 (0.055)	0.011 (0.037)	0.072 (0.074)	−0.092 ^{**} (0.039)	−0.061 (0.032)
% Drink	0.55	−0.000 (0.027)	−0.089 (0.056)	0.012 (0.039)	−0.047 (0.066)	0.103 ^{***} (0.039)	0.058 (0.038)
Constant		−0.081 ^{**} (0.032)	−0.458 ^{***} (0.049)	−0.161 ^{***} (0.040)	−0.542 ^{***} (0.060)	−0.576 ^{***} (0.039)	−0.317 [*] (0.039)
Observations		88995	84789	73363	88995	59470	59470
R ²		0.173	0.188	0.210	0.245	0.246	0.122

Note. The table presents regression coefficients on demographics for models of selection into our sample where 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 for the subsample of those who received an ID, 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) for those who made friendship nominations, Any drop is being dropped for any reason from the full sample, and no same-grade and no same-grade/same-gender friends are indicators for having no friends in the same grade or of the same gender and in the same grade in the sample of students that was not dropped for any reason. The regression models also include school fixed effects, grade fixed effects, and missing indicators for demographic controls. Standard errors (shown in parentheses) are clustered at the school level.

p < .01

**
p < .05

*
p < .1.

TABLE C3

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
Age	−0.007 ^{***} (0.002)	−0.033 ^{***} (0.004)	−0.035 ^{***} (0.004)	−0.036 ^{***} (0.004)	−0.042 ^{***} (0.004)
Male	0.006 ^{**} (0.003)	−0.005 (0.006)	−0.014 ^{**} (0.006)	−0.015 ^{***} (0.006)	−0.017 ^{***} (0.006)
Hispanic	0.020 ^{***} (0.005)	0.014 (0.011)	0.013 (0.012)	0.044 ^{***} (0.016)	0.078 ^{***} (0.021)
Black	0.015 ^{**} (0.007)	0.029 (0.021)	0.030 (0.023)	0.041 (0.025)	0.043 (0.026)
Asian	0.079 ^{***} (0.012)	0.091 ^{***} (0.018)	0.097 ^{***} (0.018)	0.132 ^{***} (0.025)	0.192 ^{***} (0.027)
Native born	−0.013 ^{**} (0.006)	0.000 (0.010)	0.006 (0.009)	0.005 (0.010)	0.006 (0.011)
Live with mom	−0.007 [*] (0.004)	−0.049 ^{***} (0.008)	−0.069 ^{***} (0.009)	−0.153 ^{***} (0.011)	−0.162 ^{***} (0.012)
Mom education	−0.000 (0.000)	−0.000 (0.001)	−0.000 (0.001)	0.012 ^{***} (0.002)	0.013 ^{***} (0.002)
Mom care	0.000 (0.001)	0.000 (0.003)	0.002 (0.003)	0.003 (0.003)	0.003 (0.003)

Outcome	Single cluster	Single cluster	Single cluster	Single cluster	Single cluster
Smoke	−0.009 *** (0.002)	−0.030 *** (0.005)	−0.033 *** (0.005)	−0.030 *** (0.005)	−0.031 *** (0.005)
Drink	−0.003 (0.003)	−0.006 (0.005)	−0.007 (0.004)	−0.004 (0.005)	−0.001 (0.005)
% Black	0.012 (0.028)	−0.023 (0.055)	−0.094 (0.072)	−0.090 (0.064)	−0.151 * (0.085)
% Hispanic	0.035 (0.050)	0.097 (0.087)	0.049 (0.094)	0.077 (0.087)	−0.051 (0.097)
% Mom college grad	0.042 (0.038)	−0.081 (0.069)	−0.099 (0.084)	−0.038 (0.070)	−0.085 (0.078)
% Smoke	0.024 (0.028)	0.037 (0.058)	−0.001 (0.067)	0.064 (0.074)	0.040 (0.076)
% Drink	−0.058 ** (0.028)	−0.125 *** (0.044)	−0.109 ** (0.053)	−0.144 *** (0.054)	−0.099 * (0.057)
Constant	0.207 *** (0.042)	0.972 *** (0.063)	1.084 *** (0.070)	1.040 *** (0.070)	1.242 *** (0.071)
Observations	59470	59470	59470	59470	59470
R ²	0.068	0.087	0.078	0.096	0.092

Note. The table presents regression coefficients on demographics for models of whether an observation belongs to a cell where the only observation is itself (singleton cell) based on cell definitions based on friendship, friendship by school, friendship by school by own demographics, friendship by school by cohort, and friendship by school by own demographics by cohort. The regression models also include school fixed effects, grade fixed effects, and missing indicators for demographic controls. Standard errors (shown in parentheses) are clustered at the school level.

p < .01

**
p < .05

*
p < .1.

TABLE C4

Analysis of the change in composition of the sample due to singleton clusters (smoking)

Restricted sample results				
Outcome	Smoke	Smoke	Smoke	Smoke
Group	None	School	School-cluster	School-cluster-X
% Smoke	0.350 *** (0.013)	0.332 *** (0.013)	0.306 *** (0.016)	0.306 *** (0.016)
Observations	20,570	20,570	20,570	20,570
R ²	0.135	0.147	0.339	0.380
OLS estimates with sample relevant to fixed effects				
Outcome	Smoke	Smoke	Smoke	Smoke
Group	None	School	School-cluster	School-cluster-X
% Smoke	0.385 *** (0.010)	0.383 *** (0.011)	0.365 *** (0.012)	0.358 *** (0.011)

Restricted sample results				
Observations	49,845	46,493	31,803	24,159
R^2	0.140	0.140	0.140	0.139
Fixed effects vary with full sample				
Outcome	Smoke	Smoke	Smoke	Smoke
Group	None	School	School-cluster	School-cluster-X
% Smoke	0.385 *** (0.010)	0.366 *** (0.011)	0.308 *** (0.019)	0.301 *** (0.023)
Observations	49,845	49,845	49,845	49,845
R^2	0.140	0.147	0.582	0.703

p < .01
**
p < .05
*
p < .1.

TABLE C5

Analysis of the change in composition of the sample due to singleton clusters (drinking)

Restricted Sample results				
Outcome	Drink	Drink	Drink	Drink
Group	None	School	School-cluster	School-cluster-X
% Drink	0.297 *** (0.014)	0.271 *** (0.014)	0.250 *** (0.017)	0.262 *** (0.016)
Observations	20,486	20,486	20,486	20,486
R^2	0.133	0.150	0.347	0.389
OLS estimates with sample relevant to fixed effects				
Outcome	Drink	Drink	Drink	Drink
Group	None	School	School-cluster	School-cluster-X
% Drink	0.329 *** (0.011)	0.325 *** (0.011)	0.310 *** (0.012)	0.305 *** (0.013)
Observations	49,656	46,311	31,672	24,062
R^2	0.153	0.153	0.146	0.142
Fixed effects vary with full sample				
Outcome	Drink	Drink	Drink	Drink
Group	None	School	School-cluster	School-cluster-X
% Drink	0.329 *** (0.011)	0.302 *** (0.011)	0.253 *** (0.019)	0.256 *** (0.022)
Observations	49,656	49,656	49,656	49,656
R^2	0.153	0.163	0.609	0.726

Note. Each column and row in Tables C4 and C5 display coefficients from separate regressions. The first row in both Tables C4 and C5 repeats the results from Table 4 where the sample is restricted to only clusters that have observations in at least two cohorts consistent with the one observation per cohort in each cluster sample in column 5 of Table 4. The second row

presents OLS (first column) or school FE results using the nonsingleton sample associated with the fixed-effect specification listed for each column. The third row reproduces the model specifications relevant to each column using the full sample so that the impact of the inclusion of the singleton clusters on the R^2 is illustrated. All regressions control for grade-level fixed effects, standard demographic controls shown above and indicators for observations where variables are missing. Standard errors (shown in parentheses) are clustered at the school level.

 $p < .01$

**
 $p < .05$

*
 $p < .1$.

TABLE C6

Effect of individual attributes on health behaviors

Outcome	Smoke	Drink
Specification	School-cluster-X FE	School-cluster-X FE
Maternal Education	-0.058 (0.041)	-0.047 (0.048)
Mom Born US	0.009 (0.006)	0.017 ** (0.007)
Paternal Education	-0.078 (0.069)	-0.119 * (0.071)
Dad Born US	0.000 (0.008)	0.002 (0.009)
Live with Mom	-0.008 ** (0.004)	-0.008 * (0.004)

Note. The estimates presented here are the coefficients on the individual controls used in the balancing when those variables are included as regressions in a model equivalent to the models estimated in column 4 of Table 5.

**
 $p < .05$

*
 $p < .1$.

APPENDIX D

BALANCING TEST RESULTS AND DISCUSSION

In Tables D1, D2, ^{D3}, D4, we present evidence from these diagnostic tests for the effect of friends' smoking (Tables D1 and D2) and drinking (^{Tables D3} and D4). Each set of rows examines the correlation between a different outcome and friend's health behavior. Columns add controls from left to right. The first column and row shows the traditional conditional correlation between maternal education and the average level of smoking or drinking among the individual's

TABLE D1

Balancing tests of friendship sorting: friends' smoking

Outcome	White	White	White	White	White
Specification	OLS	OLS	OLS	OLS	OLS
Fixed effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort

% Smoke	0.171 *** (0.027)	0.077 *** (0.015)	0.007 (0.009)	NA	NA
Observations	53,564	53,564	53,564		
R^2	0.017	0.365	0.808		
Outcome	Maternal education	Maternal education	Maternal education	Maternal education	Maternal education
Specification	OLS	OLS	OLS	OLS	OLS
Fixed Effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort
% Smoke	-0.292 *** (0.075)	-0.324 *** (0.048)	-0.227 *** (0.060)	-0.088 * (0.050)	-0.077 (0.050)
Observations	53,564	53,564	53,564	53,564	43,236
R^2	0.002	0.112	0.602	0.883	0.951
Outcome	Nativity of mother	Nativity of mother	Nativity of mother	Nativity of mother	Nativity of mother
Specification	OLS	OLS	OLS	OLS	OLS
Fixed effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort
% Smoke	0.056 *** (0.018)	0.028 *** (0.006)	0.010 * (0.005)	0.004 (0.006)	0.002 (0.006)
Observations	52,322	52,322	52,322	52,322	42,197
R^2	0.010	0.163	0.698	0.822	0.880

 $p < .01$

**
 $p < .05$

*
 $p < .1$.

friends based on OLS estimation.³³ Column 2 controls for school fixed effects. Column 3 controls for school by friendship cluster fixed effects. Column 4 controls for fixed effects associated with school, friendship cluster, and the individual's race and maternal educational attainment, which is obviously not identified for the case of race but is identified for maternal years of education. Column 5 presents the results for Column 4 selecting one observation per cohort per cluster and weighting clusters back up to their original size, except for race, where the one observation per cohort sample is estimated using the model in column 3 that does not include own race in creating the fixed effects.

Rows 1 through 6 present the conditional correlation between whether White, maternal education, mothers' nativity, whether lives with mother, paternal education and father's nativity, and the share friend's smoking or drinking. The OLS and school fixed-effect models are sizable and highly significant for all variables except for live with mother, which is statistically significant, but noticeably smaller. However, the inclusion of school by friendship cluster by individual-type fixed effects (columns 4 and 5) substantially erodes the school fixed-effect estimates. Focusing on the one per cluster estimates, even for the small effects on lives with mother, the estimates are reduced to half the magnitude of the school fixed-effect estimates. For the other more sizable estimates, the effects are reduced by between 59% and 100%. Of the 12 counterfactuals in column 5, the estimate on percent

friends' drinking is just significant at the 5% level for nativity of mother and just misses 5% significance for maternal education, but as noted above, our method is expected to only reduce bias with a finite number of choices, and these estimates are substantially smaller than the school fixed-effect estimates, 61% and 68%, respectively.

As noted above, our balancing tests can only be conducted for variables that are not included in the construction of our observationally equivalent students. If variables such as paternal education or live with mother do not conditionally correlate with smoking or drinking, then students may have less incentive to sort into friendship clusters based on those variables, and as a result, our balancing tests could be of limited value. In order to assess the importance of the variables that we use in our balancing tests, we regress whether the individual smokes or drinks on those five variables controlling for school by friendship cluster by student-type fixed effects. We find that these variables are important, with students having a lower likelihood of smoking and drinking if they live with their mother, and having lower levels of drinking if they are immigrants or if they have higher levels of paternal education.³⁴

TABLE D2

Balancing tests of friendship sorting: friends' smoking

Outcome	Live with mom	Live with mom	Live with mom	Live with mom	Live with mom
Specification	OLS	OLS	OLS	OLS	OLS
Fixed effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort
% Smoke	-0.010** (0.004)	-0.012*** (0.003)	-0.012** (0.006)	-0.008* (0.005)	-0.006 (0.005)
Observations	52,761	52,761	52,761	52,761	42,474
R ²	0.001	0.011	0.537	0.890	0.927
Outcome	Paternal education	Paternal education	Paternal education	Paternal education	Paternal education
Specification	OLS	OLS	OLS	OLS	OLS
Fixed effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort
% Smoke	-0.299*** (0.061)	-0.269*** (0.045)	-0.216*** (0.067)	-0.124 (0.082)	-0.110 (0.082)
Observations	40,517	40,517	40,517	40,517	32,450
R ²	0.004	0.087	0.619	0.775	0.853
Outcome	Nativity of father	Nativity of father	Nativity of father	Nativity of father	Nativity of father
Specification	OLS	OLS	OLS	OLS	OLS
Fixed effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort
% Smoke	0.082*** (0.023)	0.028*** (0.007)	0.015* (0.008)	0.005 (0.008)	0.005 (0.008)

³⁴The estimates are shown in Appendix C (Table C6).

Observations	40,760	40,760	40,760	40,760	32,662
R^2	0.008	0.312	0.779	0.882	0.920

 $p < .01$,**
 $p < .05$,*
 $p < .1$ **TABLE D3**

Balancing tests of friendship sorting: friends' drinking

Outcome	White	White	White	White	White
Specification	OLS	OLS	OLS	OLS	OLS
Fixed effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort
% Drink	0.071 *** (0.024)	0.027 ** (0.010)	-0.001 (0.008)	NA	NA
Observations	53,445	53,445	53,445		
R^2	0.003	0.361	0.809		

Outcome	Maternal education	Maternal education	Maternal education	Maternal education	Maternal education
Specification	OLS	OLS	OLS	OLS	OLS
Fixed Effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort
% Drink	-0.239 *** (0.057)	-0.276 *** (0.034)	-0.176 *** (0.048)	-0.074 (0.045)	-0.088 * (0.045)
Observations	53,445	53,445	53,445	53,445	43,095
R^2	0.002	0.111	0.603	0.882	0.910

Outcome	Nativity of mother	Nativity of mother	Nativity of mother	Nativity of mother	Nativity of mother
Specification	OLS	OLS	OLS	OLS	OLS
Fixed effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort
% Drink	0.067 *** (0.017)	0.049 *** (0.010)	0.023 *** (0.008)	0.020 ** (0.009)	0.019 ** (0.009)
Observations	52,212	52,212	52,212	52,212	42,085
R^2	0.012	0.166	0.700	0.822	0.868

 $p < .01$ **
 $p < .05$ *
 $p < .1$.**TABLE D4**

Balancing tests of friendship sorting: friends' drinking

Outcome	Live with mom	Live with mom	Live with mom	Live with mom	Live with mom
Specification	OLS	OLS	OLS	OLS	OLS

Fixed effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort
% Drink	-0.009 ** (0.004)	-0.009 ** (0.004)	-0.009 (0.006)	-0.007 (0.006)	-0.005 (0.006)
Observations	52,647	52,647	52,647	52,647	42,397
R ²	0.001	0.011	0.538	0.890	0.926

Outcome	Paternal education	Paternal education	Paternal education	Paternal education	Paternal education
% Drink	-0.259 *** (0.063)	-0.275 *** (0.043)	-0.159 ** (0.074)	-0.091 (0.091)	-0.070 (0.091)
Observations	40,412	40,412	40,412	40,412	32,417
R ²	0.003	0.088	0.619	0.775	0.853

Outcome	Nativity of father	Nativity of father	Nativity of father	Nativity of father	Nativity of father
% Drink	0.066 *** (0.019)	0.036 *** (0.007)	0.014 * (0.007)	0.007 (0.009)	0.005 (0.009)
Observations	40,665	40,665	40,665	40,665	32,659
R ²	0.006	0.313	0.779	0.882	0.918

Note. Each set of rows and each column in Tables D1, D2, D3, D4 display coefficients from separate regressions. Cluster and X refer to cells associated with friendship patterns and student demographics, respectively. Paternal and maternal attributes are based on student reports where education is years of schooling and nativity is whether not born in the United States. For the balancing test on the dummy variable for White, the model used for the one student per cluster per cohort analysis is school-cluster fixed-effect model. All regressions control for grade-level fixed effects and standard errors are clustered at the school level.

$p < .01$

**

$p < .05$,

*

$p < .1$.

Finally, the rich structure of friendship-type clusters creates singleton clusters of students—those students who have unique or unusual friendship preferences for their school. These singleton clusters do not contribute to the identification of the friendship effects, as there is no within-cluster variation. Tables C3, C4, and C5 examine the significance of excluding the variation associated with these observations from our estimates of the effects of friends' health behaviors. Although 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 (Tables C3). In Tables C4 and C5, we repeat the substantive analyses presented below for subsamples excluding observations associated with singleton clusters or varying the fixed effect structure, and singleton clusters have little effect on the pattern of estimates observed.³⁵

³⁵ Tables D1, D2, D3, D4 show the results for smoking and drinking. Subpanel 2 shows the OLS estimates imposing the implicit sample attrition arising from each fixed effect structure, and subpanel 3 estimates the fixed-effect models with the full sample allowing the implicit sample to adjust with the vector of fixed effects. Sample attrition only reduces the OLS and fixed-effect estimates by approximately 10%. Further, the school-cluster fixed-effect estimates are virtually unchanged by the sample restrictions even though the effective school-cluster sample as shown in panel 2 is much larger.

APPENDIX E

ADDITIONAL ROBUSTNESS RESULTS

E.1 | Additional evidence on identification

In addition to the models described in the main text, we pursue a variety of robustness exercise, described below.

First, we replicate our analysis with an alternative set of fixed effects intended to minimize the number of singleton clusters. Specifically, instead of requiring individuals to have the same number of friends with a specific demographic attribute, we require that individuals in a cluster match in terms of whether they have one or more friends of that type for each demographic attribute. Table E1 presents these results. The first two columns repeat the estimates from Table 4, columns 2 and 4. The third and fourth columns present estimates applying the less restrictive condition for matching on friendship choices, and the final two columns present estimates only applying the less restrictive condition to the subsample of observations that belong to singleton clusters using the baseline conditions for matching in Tables 3 and 4.

The qualitative conclusions are unchanged. The inclusion of our fixed effect controls for friendship choices result in at most modest declines in the estimated effects of friend's health behavior. However, quantitatively, the controls for friendship choice result in smaller reductions in the estimated effect sizes in comparison with the school fixed-effect estimates. In comparison, controlling for school by friendship cluster (not including student type) fixed effects in column 3 of Tables 3 and 4 yields estimates that are very similar to the column 4 results from the same table using school by student type by friendship fixed effects. Therefore, the detailed controls for friendship patterns in the fixed-effect structure appear to be more important than the controls for student type in practice.

A possible threat to identification in our models of the onset of smoking and drinking is that the estimated cluster fixed effects may contain considerable noise due to the small number of observations used to estimate many fixed effects. Obviously, these cluster sizes 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 patterns capture unobservables that influence smoking or drinking, we should find larger estimated parameters on the predicted fixed effects when those fixed effects are based on more observations and more accurately measured, and so these estimated fixed effects should provide a superior control for student unobservables.

Table E2 presents results separately for subsamples of students in clusters that are smaller than or larger than median cluster size. The average cluster size is 16 and 2.5 for the above and below median subsamples, respectively. Columns 1 and 2 present the estimates from the school FE models of health behavior onset for the small and large cluster subsamples, respectively. Columns 3 and 4 present the estimates from models that include the estimated cluster FE from the Wave 1 model for the same subsamples. Although there is heterogeneity

of effect sizes across the clusters, a comparison of columns 1 to 3 and columns 2 to 4 continues to illustrate only modest declines in the estimated effect for smoking, 15% and 21% for small and large clusters, respectively, and no declines in the estimated effects for drinking. We do not observe large differences in the impact of controlling for the predicted fixed effects between the small and the large cluster samples.

TABLE E1

Friendship effects using alternative cluster definitions

Outcome	Smoke	Smoke	Smoke	Smoke	Smoke	Smoke
Cluster definition	Baseline number of friends		Clusters based on any friends of a given type		Any friends for singleton clusters only	
Specification	School FE	School-Clstr-X	School FE	School-Clstr-X	School FE	School-Clstr-X
% Smoke	0.332 *** (0.013)	0.305 *** (0.016)	0.366 *** (0.012)	0.358 *** (0.014)	0.358 *** (0.012)	0.339 *** (0.016)
Observations	20,569	20,569	28,691	28,691	26,121	26,121
R ²	0.147	0.381	0.148	0.353	0.147	0.406
Outcome	Drink	Drink	Drink	Drink	Drink	Drink
Cluster definition	Baseline number of friends		Clusters based on any friends of a given type		Any friends for singleton clusters only	
Specification	School FE	School-Clstr-X	School FE	School-Clstr-X	School FE	School-Clstr-X
% Drink	0.271 *** (0.014)	0.261 *** (0.016)	0.300 *** (0.013)	0.304 *** (0.015)	0.291 *** (0.013)	0.289 *** (0.015)
Observations	20,485	20,485	28,761	28,761	26,187	26,187
R ²	0.150	0.390	0.161	0.367	0.159	0.418

Note. Each column displays coefficients from separate regressions. The first and second columns repeat the school fixed effect and school by cluster by type fixed effect results from Table 4 where friendship clusters are defined based on number of friends of each demographic type. The third and fourth columns use a less restrictive condition that individuals who have friendship (one or more reciprocal links) with the same demographic types of students are placed in the same cluster. The final two columns use this less restrictive condition for observations that belong to singleton clusters when clusters are based on number of friends of each type but use the original cluster definition for all other observations. Cluster and X refer to cells associated with friendship patterns and student demographics, respectively. All regression include controls for the composition of each cohort/grade in each school on race, ethnicity, gender, age, maternal education, whether live with mother, and whether native born. All regressions control for grade-level fixed effects and indicators for observations where control variables are missing. Standard errors (shown in parentheses) are clustered at the school level.

 $p < .01$

**
 $p < .05$

*
 $p < .1$.

TABLE E2

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. without own cohort	School FE, Cluster FE Est. without own cohort
Sample	Restricted	Restricted	Restricted	Restricted
	Small cluster	Large cluster	Small cluster	Large cluster

% Smoke	0.055 [*] (0.029)	0.052 ^{**} (0.021)	0.047 (0.031)	0.041 ^{**} (0.021)
Predicted FE without own cohort			−0.025 (0.029)	0.054 (0.036)
Guryan et al. Control			9.439 (11.739)	18.294 ^{**} (8.465)
Observations	833	1,062	833	1,062
R ²	0.190	0.125	0.191	0.133

Outcome	Drink	Drink	Drink	Drink
% Drink	0.151 ^{***} (0.048)	0.079 ^{**} (0.037)	0.151 ^{***} (0.047)	0.082 ^{**} (0.038)
Predicted FE without own cohort			−0.015 (0.044)	−0.034 (0.070)
Guryan et al. Control			−1.374 (18.174)	−24.185 [*] (14.256)
Observations	571	657	571	657
R ²	0.238	0.175	0.238	0.179

Note. Each column and panel display coefficients from separate regressions. The restricted sample is estimated only using students who belong to school–cluster–X cells where there are students in that cell from at least two grades. Small and large clusters refer to subsamples with the number of observations in school–cluster–X cell above or below the median. The predicted fixed effect is the weighted mean over all grades of the grade–school–cluster–X residual from the model in column 4 of Table 7 omitting the fixed effect contribution from a student's own grade. The Guryan control is the same weighted mean over all grade–cluster–X cells in a school again omitting the contribution from a student's own grade–cluster–X cell. All regressions control for grade-level fixed effects, standard demographic controls shown above, and indicators for observations where variables are missing. Standard errors (shown in parentheses) are clustered at the school level.

 $p < .01$

**
 $p < .05$

*
 $p < .1$.

TABLE E3

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	1–3 Friends	4–5 Friends	1–3 Friends	4–5 Friends
% Smoke	0.306 ^{***} (0.013)	0.588 ^{***} (0.025)	0.288 ^{***} (0.016)	0.561 ^{***} (0.046)
Observations	17,478	3,092	17,478	3,092
R ²	0.141	0.244	0.362	0.500

Outcome	Drink	Drink	Drink	Drink
Specification	School FE	School FE	School–cluster–X FE	School–cluster–X FE
Sample	1–3 Friends	4–5 Friends	1–3 Friends	4–5 Friends
% Drink	0.246 ^{***}	0.548 ^{***}	0.243 ^{***}	0.536 ^{***}

	(0.013)	(0.031)	(0.015)	(0.048)
Observations	17,405	3,081	17,405	3,081
R^2	0.140	0.268	0.367	0.531

Note. Each column and panel display coefficients from separate regressions. Each column represents a subsample of individuals with either 1–3 or 4–5 same grade/same gender friends. Cluster and X refer to cells associated with friendship patterns and student demographics, respectively, and the results shown in columns 3 and 4 are equivalent to models presented in column 4 of Table 5. All regressions control for grade-level fixed effects, standard demographic controls shown above, and indicators for observations where variables are missing. Standard errors (shown in parentheses) are clustered at the school level.

$p < .01$

**

$p < .05$

*

$p < .1$.

TABLE E4

Friendship effects on onset 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. without own cohort	School FE–cluster FE Est. without 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 *** (0.015)	0.073 *** (0.017)	0.067 *** (0.020)	0.001 (0.029)	0.066 *** (0.021)	0.001 (0.029)
Observations		6,137		1,895		1,895
R^2		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. without own cohort	School FE–Cluster FE Est. without 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 *** (0.016)	0.151 *** (0.029)	0.083 *** (0.028)	0.220 *** (0.071)	0.081 *** (0.029)	0.081 *** (0.029)
Observations		4,310		1,228		1,228
R^2		0.075		0.124		0.124

Note. Each pair of columns in each panel displays coefficients from separate regressions. The restricted sample is estimated only using students who belong to school–cluster–X cells where there are students in that cell from at least two grades. Each column represents a subsample of individuals with either 1–3 or 4–5 same grade/same gender friends. Columns 5 and 6 include controls for a predicted fixed effect based on a weighted mean over all grades of the grade–school–cluster–X residual from the model in column 4 of Table 7 omitting the fixed effect contribution from a student's own grade and a Guryan-style control that is the same weighted mean over all grade–cluster–X cells in a school again omitting the contribution from a student's own grade–cluster–X cell. All regressions control for grade-level fixed effects, standard demographic controls shown above, and indicators for observations where variables are missing. Standard errors (shown in parentheses) are clustered at the school level.

$p < .01$

**

$p < .05$

*

$p < .1$.

An additional 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. Our

positive estimates of friendship effects may be driven by students who have a very small number of friends and for whom the number of choices being made is insufficient to eliminate the bias from student unobservables. In Table E3, we divide the Wave 1 sample by the number of friends, three or less versus four or more. 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 6% relative to the school fixed effects for smoking and by less than 3% for drinking. The changes in effect size due to the inclusion of cluster FE's is very similar across the two subsamples.

In Table E4, we present estimates for the effect of friends' behavior on the onset of smoking and drinking for the 1- to 3-friend and 4- to 5-friend subsamples. Because the 4- to 5-friend subsample is quite small (<300), we estimate a joint model interacting dummies for number of friends with the share of friends smoking or drinking.³⁶ In the first two columns, we present the school FE estimates. 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 E2, the relationship for onset is also stronger for the subsample with 4–5 friends. 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- to 5-friend subsample is very near zero. It is important to note that this zero estimate does not provide evidence in either support of or against the validity of our identification strategy. Turning to columns 5 and 6, 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 3% and changing the zero estimate for friends' smoking for 4–5 friends by only 0.002.³⁷

REFERENCES

- Aizer Anna. 2009 Peer effects and human capital accumulation: The externalities of ADD. NBER Working Paper #14354.
- Aizer A, & Currie J (2004). Networks or neighborhoods? Correlations in the use of publicly funded maternity care in California. *Journal of Public Economics*, 88(12), 2573–2585.
- Aliprantis Dionissi and Richter Francisca G.C.. 2012 Marginal neighborhood effects from moving to opportunity. Cleveland Federal Reserve Working Paper.
- Angrist J (2014). The perils of peer effects. *Labour Economics*, 30(C), 98–108.
- Babcock P (2008). From ties to gains? Evidence on connectedness and human capital acquisition. *Journal of Human Capital*, 2(4), 379–409.
- Bayer Patrick and Ross Stephen L.. 2008 Identifying individual and group effects in the presence of sorting: A neighborhood effects application. NBER Working Paper #12211.
- Bayer P, Ross SL, & Topa G (2008). Place of work and place of residence: Informal hiring networks and labor market outcomes. *Journal of Political Economy*, 116(6), 1150–1196.
- Bertrand M, Luttmer E, & Mullainathan S (2000). Network effects and welfare cultures. *Quarterly Journal of Economics*, 115(3), 1019–1055.

³⁶It is not realistic to estimate 60–70 school fixed effects in samples with only a couple of hundred observations.

³⁷Further, 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 FEs, the estimated effects of friends smoking are 0.050 and 0.068 for 1–3 and 4–5 friends, respectively, and those reductions compared with the estimates in columns 1 and 2 are nearly identical in magnitude at about 6%.

- Bifulco R, Fletcher J, & Ross SL (2011). The effect of classmate characteristics on individual outcomes: Evidence from the Add Health. *American Economic Journal: Economic Policy*, 3(1), 25–53.
- Bramoullé Y, Djebbari H, & Fortin B (2009). Identification of peer effects through social networks. *Journal of Econometrics*, 150(1), 41–55.
- Brock WA, & Durlauf SN (2001). Discrete choice with social interactions. *Review of Economic Studies*, 68(2), 235–260.
- Calvó-Armengol A, Patacchini E, & Zenou Y (2009). Peer effects and social networks in education. *Review of Economic Studies*, 76(4), 1239–1267.
- Christakis NA, & Fowler JH (2008). The collective dynamics of smoking in a large social network. *New England Journal of Medicine*, 358, 249–258.
- Conti G, Galeotti A, Muller G, & Pudney S (2013). Popularity. *Journal of Human Resources*, 48(4), 1072–1094.
- Cutler D, & Glaeser E (1997). Are Ghettos good or bad? *Quarterly Journal of Economics*, 112(3), 827–872.
- Dale SB, & Krueger AB (2002). Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *Quarterly Journal of Economics*, 117(4), 1491–1527.
- Eisenberg D, Golberstein E, & Whitlock JL (2014). Peer effects on risky behaviors: New evidence from college roommate assignments. *Journal of Health Economics*, 33, 126–138. [PubMed: 24316458]
- Fletcher JM (2010). Social interactions and smoking: Evidence using multiple student cohorts, instrumental variables, and school fixed effects. *Health Economics*, 19(4), 466–484. [PubMed: 19382102]
- Fletcher JM (2012). Peer influences on alcohol consumption. *Journal of Population Economics*.
- Fletcher Jason M., Ross Stephen L. and Zhang Yuxiu. 2013 The determinants and consequences of friendship composition. NBER Working Paper #19215.
- Fletcher J, & Marksteiner R (2017). Causal spousal health spillover effects and implications for program evaluation. *American Economic Journal: Economic Policy*, 9(4), 144–166. [PubMed: 30057688]
- Friesen J, & Krauth B (2011). Ethnic enclaves in the classroom. *Labour Economics*, 18(5), 656–666.
- Fu S, & Ross SL (2013). Wage premia in employment clusters: How important is worker heterogeneity? *Journal of Labor Economics*, 31(2), 271–304.
- Glaeser EL, Sacerdote B, & Scheinkman JA (1996). Crime and social interactions. *Quarterly Journal of Economics*, 111(2), 507–548.
- Guryan J, Kroft K, & Notowidigdo MJ (2009). Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics*, 1(4), 34–68. [PubMed: 20454555]
- Halliday TJ, & Kwak S (2012). What is a peer? The role of network definition in the estimation of endogenous peer effects. *Applied Economics*, 44(3), 289–302.
- Heckman JJ (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.
- Hoekstra ML, & Carrell S (2010). Externalities in the classroom: How children exposed to domestic violence affect everyone. *American Economic Journal: Applied Economics*, 2(1), 211–228.
- Lavy V, Passerman D, & Schlosser A (2012). Inside the black box of ability peer effects: Evidence from variation in low achievers in the classroom. *Economic Journal*, 122(559), 208–237.
- Manski CF (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, 60(3), 531–542.
- Mayer A, & Puller SF (2008). The old boy (and girl) network: Social network formation on university campuses. *Journal of Public Economics*, 92(1–2), 329–347.
- Nathan Anil. 2008 The effects of racial and extracurricular friendship diversity on achievement. Working Paper.
- Patacchini E, & Zenou Y (2012). Juvenile delinquency and conformism. *Journal of Law, Economics and Organization*, 28(1), 1–31.

- Renna F, Grafova IB, & Thakur N (2008). The effects of friends on adolescent body weight. *Economics & Human Biology*, 6(3), 377–387. [PubMed: 18672412]
- Ross SL (2011). Social interactions within cities: Neighborhood environments and peer relationships. In Brooks N, Donaghy K, & Knapp G (Eds.), *Handbook of urban economics and planning*. Oxford: Oxford University Press.
- Sacerdote B (2001). Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics*, 116(2), 681–704.
- Trogon JG, Nonnemaker J, & Pais J (2008). Peer effects in adolescent overweight. *Journal of Health Economics*, 27(5), 1388–1399. [PubMed: 18565605]
- Weinberg Bruce. 2008 Social interactions with endogenous associations. Ohio State University Working Paper.

TABLE 1

Descriptive statistics

Variable	Mean	Std Dev	Min	Max
Smoke	0.34	0.47	0	1
Drink	0.54	0.50	0	1
Get drunk	0.29	0.45	0	1
Exercise	2.29	1.19	0	4
Any exercise	0.95	0.22	0	1
Age	14.87	1.69	10	19
Male	0.45	0.50	0	1
White	0.62	0.49	0	1
Hispanic	0.13	0.33	0	1
Black	0.16	0.37	0	1
Asian	0.06	0.23	0	1
Live with mom	0.94	0.25	0	1
Maternal years of education	13.49	2.33	0	18
Maternal caring scale	4.79	0.58	1	5
Native born	0.92	0.26	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	0.39	0	1
Grade = 11	0.16	0.37	0	1
Grade = 12	0.15	0.36	0	1
Missing	0.34	0.47	0	1
Number of nominations	2.78	1.30	1	5
Proportion White	0.61	0.43	0	1
Proportion Black	0.17	0.35	0	1
Proportion Hispanic	0.13	0.28	0	1
Proportion Asian	0.06	0.20	0	1
Proportion other race	0.04	0.14	0	1
Proportion mom less high school	0.14	0.25	0	1
Proportion mom some college	0.18	0.25	0	1
Proportion of mom college grad	0.35	0.33	0	1
Proportion native	0.93	0.22	0	1
Proportion live with mom	0.94	0.18	0	1

Note. This table presents the mean, standard deviation, minimum and maximum for the approximately 54,000 individuals sampled in Wave 1 who nominated at least one friend of the same gender who could be confirmed to attend the same grade in the same school and reported whether they smoke or drank.

TABLE 2

Variation in friendship options

Peer variable	R^2
% Maternal college graduate	92.8
% Black	97.8
% Hispanic	97.8
% White	97.7
% Asian	95.0
% Native born	96.0
Mean maternal caring scale	55.1
% Smoke cigarettes	76.9
% Drink alcohol	81.4
With peer attribute controls	
% Smoke cigarettes	78.0
% Drink alcohol	82.3

Note. The results reported indicate the R^2 from a regression of the grade-level measure of peer characteristics or peer health behaviors on a complete set of school-level and grade-level dummy variables for the Wave 1 sample of 65,000 students. The final two rows include controls for the peer variables in the six top rows.

TABLE 3

Friendship network effects

Outcome	Smoke	Smoke	Smoke	Smoke	Smoke
Specification	OLS	OLS	OLS	OLS	OLS
Fixed effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort
% Smoke	0.350 *** (0.013)	0.332 *** (0.013)	0.306 *** (0.016)	0.306 *** (0.016)	0.320 *** (0.023)
Observations	20,570	20,570	20,570	20,570	20,570
R ²	0.135	0.147	0.339	0.380	NA
Outcome	Drink	Drink	Drink	Drink	Drink
Specification	OLS	OLS	OLS	OLS	OLS
Fixed effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort
% Drink	0.297 *** (0.014)	0.271 *** (0.014)	0.250 *** (0.017)	0.262 *** (0.016)	0.287 *** (0.022)
Observations	20,486	20,486	20,486	20,486	20,486
R ²	0.133	0.150	0.347	0.389	NA

Note. Each column displays coefficients from separate regressions. Cluster and X refer to cells associated with friendship patterns and student demographics, respectively. Additional controls include age, gender, race and ethnicity dummy variables, maternal education, whether live with mother, whether native born, and indicators for whether an observation has data missing on a given variable. All regressions in the table and future tables also include controls for grade-level fixed effects and indicators for observations where control variables are missing, and standard errors (shown in parentheses) are clustered at the school level.

p < .01

**
p < .05

*
p < .1.

TABLE 4

Friendship network effects with controls for cohort composition

Outcome	Smoke	Smoke	Smoke	Smoke	Smoke
Specification	OLS	OLS	OLS	OLS	OLS
Fixed effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort
% Smoke	0.347 *** (0.013)	0.332 *** (0.013)	0.305 *** (0.016)	0.305 *** (0.016)	0.302 *** (0.016)
Observations	20,569	20,569	20,569	20,569	20,569
R ²	0.137	0.147	0.339	0.381	NA
Outcome	Drink	Drink	Drink	Drink	Drink
Specification	OLS	OLS	OLS	OLS	OLS
Fixed effects	None	School	School-cluster	School-cluster-X	School-cluster-X One per cohort
% Drink	0.294 *** (0.013)	0.271 *** (0.014)	0.249 *** (0.017)	0.261 *** (0.016)	0.258 *** (0.016)
Observations	20,485	20,485	20,485	20,485	20,485
R ²	0.135	0.150	0.347	0.390	NA

Note. Each column displays coefficients from separate regressions. Cluster and X refer to cells associated with friendship patterns and student demographics, respectively. All regression include controls for the composition of each cohort/grade in each school on race, ethnicity, gender, age, maternal education, whether live with mother, and whether native born.

 $p < .01$

**
 $p < .05$

*
 $p < .1$.

TABLE 5

Two stage estimates of friendship effects on smoking and drinking

Outcome	Smoke	Smoke	Smoke	Smoke
Specification	School FE	School-Cluster-X One per cohort	School FE, Cluster FE Est. with own cohort	School FE, Cluster FE Est. without own cohort
Sample	Individual	Individual	Individual	Individual
% Smoke	0.340 *** (0.014)	0.320 *** (0.023)	0.318 *** (0.010)	0.328 *** (0.011)
Predicted FE with or without own cohort			1.037 *** (0.003)	-0.002 (0.009)
Guryan et al.				-171.477 ***
Control				(16.991)
Observations	17,500	17,500	17,500	17,500
R ²	0.147	NA	0.377	0.392

Outcome	Drink	Drink	Drink	Drink
Sample	Individual	Individual	Individual	Individual
% Drink	0.289 *** (0.014)	0.287 *** (0.022)	0.287 *** (0.010)	0.288 *** (0.012)
Predicted FE with or without own cohort			1.037 *** (0.004)	0.005 (0.011)
Guryan et al.				-179.230 ***
Control				(15.565)
Observations	17,435	17,435	17,435	17,435
R ²	0.154	NA	0.385	0.403

Note. Each column and panel display coefficients from separate regressions. The individual sample is estimated using student level data for which a school by type by friendship pattern fixed effect is identified and the first two columns are based on the one student per cohort per fixed-effect sample. The predicted fixed-effect regressor in column 3 is the estimated school-cluster-X fixed effect from the model shown in column 4 of Table 4. The predicted fixed effect in column 4 is the weighted mean over all grades of the grade-school-cluster-X residual from the model in column 4 omitting the fixed effect contribution from a student's own grade. The Guryan control is the same weighted mean over all grade-cluster-X cells in a school again omitting the contribution from a students' own grade-cluster-X cell.

 $p < .01$

**
 $p < .05$

*
 $p < .1$.

TABLE 6

Longitudinal analysis of the onset of smoking and drinking

Outcome	Smoking onset	Smoking onset	Smoking onset	Smoking onset	Smoking onset
Specification	OLS	School FE	School FE	School FE, Cluster FE Est. without own cohort	School FE, Cluster FE Est. with own cohort
Sample	Full sample	Full sample	Restricted	Restricted	Restricted
% Smoke	0.054 *** (0.013)	0.052 *** (0.015)	0.053 *** (0.017)	0.051 *** (0.017)	0.045 *** (0.017)
Predicted FE with or without own cohort				−0.017 (0.035)	0.003 (0.021)
Guryan et al.					11.688 **
Control					(5.394)
Observations	2,278	2,278	1,891	1,891	1,891
R ²	0.027	0.086	0.093	0.093	0.095

Outcome Sample	Drinking onset Full sample	Drinking onset Full sample	Drinking onset Restricted	Drinking onset Restricted	Drinking onset Restricted
% Drink	0.110 *** (0.022)	0.109 *** (0.025)	0.108 *** (0.027)	0.106 *** (0.029)	0.113 *** (0.028)
Predicted FE with or without own cohort				−0.041 (0.056)	−0.026 (0.028)
Guryan et al.					−11.571
Control					(11.310)
Observations	1,501	1,501	1,223	1,223	1,223
R ²	0.033	0.116	0.120	0.121	0.122

Note. Each column and panel display coefficients from separate regressions. The first two columns use the full Wave 2 sample. The restricted sample used in columns 3 through 5 only contains students who belong to school-cluster-X cells where there are students in that cell from at least two grades (fixed effects identified). The predicted fixed effects in columns 4 and 5 and the Guryan control are defined as described in Table 5 for columns 3 and 4.

p < .01

**
p < .05

*
p < .1.