NBER WORKING PAPER SERIES

THE INFLUENCE OF PEER GENOTYPES AND BEHAVIOR ON SMOKING OUTCOMES: EVIDENCE FROM ADD HEALTH

Ramina Sotoudeh
Dalton Conley
Kathleen Mullan Harris

Working Paper 24113
http://www.nber.org/papers/w24113

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2017

This paper benefited from comments and discussions with Mark Hoffman, Parijat Chakrabarti, and the instructors of RSF Summer Institute in Social-Science Genomics. Funding for this study was provided by the Russell Sage Foundation (grant #83-15-29) and by the John D. and Catherine T. MacArthur Foundation Connected Learning Research Network. This paper uses data from The National Longitudinal Study of Adolescent Health (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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w24113.ack

The Influence of Peer Genotypes and Behavior on Smoking Outcomes: Evidence from Add Health

Ramina Sotoudeh, Dalton Conley, and Kathleen Mullan Harris

NBER Working Paper No. 24113
December 2017
JEL No. D79,I12,I20

ABSTRACT

We introduce a novel use of genetic data for studying social influences on behavior: Using data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), we deploy the distribution of genotypes in a given grade within a school to instrument the influence of peer smoking on an individual's own smoking behavior. We argue that this design alleviates many problems inherent to estimating peer effects. Using this approach, we find the relationship between peer smoking and individual smoking to be larger than that estimated by prior studies. Further, we explore the reduced form relationship between peer genotypes and ego smoking and find that the impact of peers' genetic risk for smoking on ego's smoking behavior is at least half as large as the effect of individual's own genotype and sex, and 30% the effect of age. Moreover, peer influence on smoking appears heterogeneous by race: although whites and non-whites are equally susceptible to peer influence with respect to smoking, white egos are more likely to be influenced by white alters. This analysis suggests a promising way that genetic information can be leveraged to identify peer effects that avoids the reflection problem, contextual effects and selection into peer groups.

Ramina Sotoudeh
125 Wallace Hall
Princeton University
Princeton, NJ 08544
sotoudeh@princeton.edu

Kathleen Mullan Harris
155 Hamilton Hall
Chapell Hill
North Carolina, 27599-3210
kathie_harris@unc.edu

Dalton Conley
Department of Sociology
Princeton University
153 Wallace Hall
Princeton, NJ 08544
and NBER
dconley@princeton.edu

Introduction

Scholars of peer effects have long recognized two difficulties in their attempts to ascertain the impact of friends, colleagues and other social connections on our own behavior and attitudes. First, in the absence of random assignment of peers, it is impossible to determine whether any observed associations between peers' and ego's outcomes are the result of "true" peer effects—i.e. the causal impact of alters on ego—or, alternatively, whether they merely reflect homophily (the fact that individuals who are similar on relevant observed or unobserved dimensions tend to form ties). That is, individuals' associations on a given trait may result from the fact that they actively select friends who are similar to themselves on either observed or unobserved factors which may then result in similarities among friends. In fact, using naive models and improper extrapolation may lead the researcher to conclude that traits such as height, which cannot possibly spread through networks, exhibit peer effects and as a result, many of the findings from this literature have been called into question (Cohen-Cole & Fletcher, 2008). Another problem in peer effects research comes from the common exposure of individuals in a given social group (e.g., a friendship network or a classroom) to common environmental characteristics of that social group – i.e. similar teachers or community characteristics – which may, in turn, confound interpretations of peer effects. To address these challenges, a growing number of studies have leveraged quasi-natural experiments that generate randomly assigned peers such as freshman college roommates (Guo et al. 2015), class assignment (Figlio 2002; Legewie 2012) and even golf foursomes among executives (Guryan, Kroft, and Notowidigo, 2009).

Even in cases where random assignment of peers is leveraged, there remains the "reflection problem" to use Manski's term (1993). That is, it is hard for the observer to know

who is causing whose behavior in a dyad or network of relationships. Take for example the case of Figlio's (2002) study of disruptive classmates on academic achievement. He uses the assignment of a boy with a feminine name to a classroom as an instrument for a disruptive peer, since these "boys named Sue" appear to be randomly assigned to classes and tend to get into more disciplinary problems. He finds that being assigned a disruptive peer engenders deleterious consequences for the other students in terms of behavioral and academic outcomes. However, one cannot be certain that Sue's mere presence as a gender-norm-disrupting influence is not itself the treatment. Namely, Sue may get into more trouble because he is picked on, not as the instigator. This implies that while the reduced form effect of Sue's presence on the academic achievement of his peers is still accurately estimated, the precise nature of the treatment remains ambiguous due to the reflection problem. Many scholars seek to use pre-treatment variables such as high school GPA in college roommate studies in order to get around this dilemma (Guo et al. 2015).

With these concerns in mind, the present paper deploys a novel approach to identify peer effects on adolescent smoking: We use the genetic propensity to smoke of as-if-randomly-assigned same-grade schoolmates as an exogenous source of influence on ego's smoking behavior that avoids the reflection problem. Namely, since genotype is assigned at conception, we can identify the biological source of behavioral contagion and hence overcome the reflection problem. Meanwhile, by deploying school and year fixed effects, we identify variation in peer genetic propensity to smoke (i.e. classmates in a given grade in a given school) that, we argue, is essentially randomly assigned (solving the homophily problem). Lastly, since genotype is not a plastic, labile trait, we avoid contextual effects that might influence actual measures of peer behavior.

Aside from using genetic data as a tool to better understand school-based social dynamics, we are also substantively interested in the way peer genes affects a person's outcome (beyond the person's own genetic contributions) and the forces that may mediate this association. In this regard, we are proposing a novel way to integrate social and genetic data. That is, we explore the possibility of considering genes not only as attributes of individuals but also as attributes of an individual's larger social environment. Since genes increase an individual's propensity towards a behavior or a state of being, the genetic makeup of an individual's surroundings may well be affecting the behavior or outcome of the proband individual as well, through social channels. This approach calls into question the sharp line between nature and nurture, between genes and social environment.

With this in mind, the contribution of the present paper is fourfold. First, we make a contribution to methodology and research design by showing how peer genes can be used as an instrument for peer behavior. Second, we use this method to re-evaluate the relationship between peer smoking and individual smoking, which we find to be positively and significantly related. Third, we make substantive contributions to the social genetics literature by exploring the extent to which others' genes, net of one's own genes, affect one's outcome. These social genetic effects are large and merit consideration by both social and biological scientists. Namely, we find that even by the most conservative estimate, the magnitude of the effect of peers' genes on an individual's smoking outcome is at least half as large as that of individual's own genes, sex and 30% that of age on his or her smoking behavior. Finally, we consider the ways in which an individual's own genes, or social characteristics such as race, may moderate the effect of peer smoking on the ego. We find modest evidence of gene by environment (GxE) interaction, individuals with higher genetic propensity to smoke are more susceptible to peer influence of

smoking. There also seems to be heterogeneous peer influence in smoking by race: although both whites and non-whites are as susceptible to peer influence of smoking, we show that white egos are more likely to be influenced by white alters.

Using genes in social science

The genetic makeup of organisms is determined at conception and remains unchanged throughout life[1]. Genes, though temporally unvarying, have been shown to influence a variety of human, behavioral outcomes. These two properties of genes – the fact that they are determined at conception, and are associated with individual outcomes – makes them great candidates for sources of exogenous variation which can be leveraged to elucidate causal relationships in social research. This kind of research is made possible with the integration of genetic data into large-scale, multidisciplinary surveys. In this line of research, studies have identified genetic variants linked to many social and complex outcomes including: educational outcomes (Okbay et al. 2016), smoking behaviors (TAG 2010), fertility (Barban 2016) and psychopathological outcomes (PGC 2013; PGC 2009)—among many other social and health-related outcomes.

That said, significant methodological hurdles remain for research that uses genetic data to predict observational outcomes in population-based samples. The first hurdles relate to measuring genetic contributions to social outcomes. Initial studies of genetic effects on social outcomes used candidate genes; that is, genes of pre-specified theoretical interest that are measured with particular outcomes in mind. These studies failed to account for the non-random distribution of candidate genes across environments, because the genes may act as proxies for

[1] An exception to this rule is somatic mutation, genetic changes passed down to progeny via cell division, as in the case of cancer (Murphy et al. 1995).

unmeasured environmental influences (Fletcher & Conley, 2013) and were also underpowered to detect main and interaction effects. However, more recently, the advent of dense microarrays that measure common genetic variants known as Single Nucleotide Polymorphisms (or SNPs) has made it possible and relatively inexpensive to measure millions of genetic loci in a single study.

As a result, researchers have moved away from single gene studies and towards using genome-wide association studies (GWASs) to measure genetic risk. Using results from a GWAS, researchers can construct a polygenic score (PGS) for a phenotype by aggregating the effects of thousands of SNPs across the genome and weighting them by the strength of their association with that phenotype. In essence, a polygenic score is a weighted average or composite score that takes into account information across an individual's entire genome to measure his/her genetic predisposition or risk for a particular outcome. A polygenic score for individual *i* is a weighted average across *n* SNPs of the number of reference alleles *x* (0,1 or 2) at loci *j* multiplied by the score for that SNP β:

$$(1) \qquad PGS_i = \sum_{j=1}^{n} \frac{(\beta_{ij} x_{ij})}{n}$$

Polygenic scores have several attractive features. First, complex behavioral outcomes that are the focus of socio-genetic research are usually highly polygenic; that is, they reflect the aggregate influence of many different genes (Visscher, Hill, & Wray, 2008). Individuals fall somewhere on a continuum of genetic risk that reflects small contributions from many genetic loci—even clinically dichotomous outcomes may reflect a shift along a phenotypic continuum known as decanalization (Gibson, 2009). Second, individual genetic loci influencing the etiology of complex phenotypes have low penetrance: often no single gene produces a symptom or trait at

a detectable level, making it difficult to distinguish between environmental and genetic factors (Gibson, 2012). Finally, they are "hypothesis-free" measures—i.e. *ex ante* knowledge about the biological processes involved is not needed to estimate a score for a particular phenotype.[2] A polygenic scores casts a wide net across an individual's entire genome to yield a single quantitative measure of genetic risk, allowing researchers to explore how genes operate within environments where the biological mechanisms behind the outcome are not yet fully understood (Belsky & Israel, 2014).

Using these more comprehensive measures of genetic influence, researchers have turned their attention to understanding more complex paradigms of how genes and social forces interact to explain individual outcomes (Baud et al. 2017; Cawley et al. 2017; Domingue et al. 2017). Rather than focusing specifically on the relationship between a gene or a set of genes and an outcome, these studies explore the complex ways by which genes come to bear on social outcomes—for example, at the interaction between a person's and a group's genetic makeup, on the one hand, and an individual's behavior, on the other. Considering these issues represent new approaches within the paradigms of "social genetic effects" (SGA) and "gene-by-environment" interactions (GxE). In this paper, we explore these two paradigms as they relate to smoking behavior.

The *social genetic effects* paradigm rests on the observation that the genotypes of our compatriots serve as part of our own "social" environment (Domingue & Belsky 2017). Since other individuals' behaviors and states of being are (however partially) influenced by their genes, and since the extensive literature on social psychology and social networks have shown us that

[2] The approach of hypothesis-free gene discovery is akin to many other realms of social science made possible by computational power and digitized data, such as text analysis (Chakrabarti and Frye 2017; Hoffman et al. 2017)

others' behaviors likely matter for our behaviors and outcomes, then these so-called "social genetic effects" are to be expected. As with the peer effects literature, however, causality is hard to establish in a non-experimental setting because of issues of selection and context.

Recent experimental studies on animals have established that these social genetic effects explain a significant part of animal's health and psycho-social outcomes. Baud et al. (2017) find that murine cage-mates have significant effects on the outcomes of individual mice for more than a third of phenotypes they considered. Together, what they termed "social genetic effects" accounted for 29% of the phenotypic variance in the mice. The authors also show that not accounting for these social genetic effects, may, in many cases, lead to inflated estimates of heritability (Baud et al. 2017).

However, while intriguing, scholars have only just begun exploring social genetic effects and gene-by-environment interactions in humans. This is mainly due to the fact that unlike mice, we cannot randomly assign individuals' social or metagenomic environments. Indeed, even in ethnically homogenous samples, it turns out that friends (Christakis & Fowler, 2014) and spouses (Conley et al., 2016; Domingue et al., 2014; Guo, Wang, Liu, & Randall, 2014; Domingue et al., 2017) tend to be more genotypically similar than randomly matched individuals and that even environmental measures such as urbanity are correlated with genetic population structure (Conley et al., 2014). Thus, while other individuals' genotypes or phenotypes might influence the expression of one's genes, it is also possible that the individual's genes or traits influences the kind of social or genetic environment they select into (Domingue & Belsky, 2017). It is therefore difficult to separate out genetic homophily from true peer effects. If not for careful study design, genes and environments can act as mutual proxies, confounding the researcher's attempts to parse their effects.

Because of these difficulties, the few empirical studies that do consider the question of social genetics are, at best, suggestive. Using data from the National Longitudinal Study of Adolescent to Adult Health, Domingue et al. (2017) find an association between an individual's downstream educational attainment and the education attainment polygenic score of his or her schoolmates and self-reported friends, controlling for the focal individual's own polygenic score. They interpret this as evidence in favor of social genetic effects, but they also note that this association cannot be solely attributed to social genetic effects since families self-select into neighborhoods with their desired schools, and individuals actively select their friends. In the same study, Domingue et al. (2017) show a positive genotypic association between individuals and their friends as well as their schoolmates, which points to the possibility that selection based on genes may be driving their phenotypic results.

Another strategy employed to capture social genetic effects leverages the slight genetic variation between full siblings as treatments for explaining sibling outcomes.[3] Using a candidate gene approach, Rauscher, Conley and Siegel (2015) show that the expression of genetic variants linked to general health in an individual are moderated by the individual's sibling's variants. While a promising result initially, candidate genes were later shown to suffer from multiple testing bias and can lead to many false positive findings. Alternatively, using a similar full-sibling design, Cawley et al. (2017) use genome-wide data to look at "peer effects" between full siblings. They find a significant and positive association between a sibling's BMI polygenic

[3] A child inherits 50% of their genome from their mother and 50% from their father. The actual genes which make up those proportions, however, are randomly drawn from each parent's genome. It follows that, in expectation, full siblings will share 50% of their genes; but in a given case, full siblings can range from being almost genetically unrelated to nearly identical, depending entirely on how the random assignment of parental genes plays out. In practice most siblings share about 35 to 65 percent of their genome (Visscher et al. 2009).

score and ego's weight, controlling for ego's own polygenic score. Using siblings helps overcome selection problems, since an individual cannot choose their sibling nor their sibling's genotype, but it introduces a number of additional assumptions about within-family dynamics that limit the external validity of their findings. One threat to the validity of their findings is that they do not account for parents' genotype and so a sibling's higher polygenic score may act as a proxy for higher parental polygenic score. Since parents' eating habits play an important role in shaping household consumption patterns (and other aspects of a child's behavior such as body image), their genes, even when not transmitted to one of their offspring, may have a strong influence on their offspring's outcomes. Recent studies have already established the importance of these "non-transmitted alleles" on offspring outcomes (Kong et al. 2017). This contextual effect poses a threat to Cawley et al.'s research design in identifying true peer effects, because parent genotype may be causing both siblings to end up with higher weights than expected – i.e. evidence of common environmental exposure rather than peer effects between siblings. Even if the authors had accounted for parental genotype, the effects they report may be going through pathways other than direct sibling-to-sibling influence. For example, parental reaction to one child with a higher genetic risk for BMI may induce them to make specific kinds of foods available, which may then affect the eating habits and BMI of other children in the family. If this were the mechanism driving phenotypic association between siblings, the finding would be meaningful in a household setting, but it would not inform research on peer effects more broadly.

As the examples above illustrate, even with good measures of the genetic contributions to a trait, such as a PGS, existing efforts to find associations between genetic variation and social behavior in large, multidisciplinary surveys are often unable to support causal inferences because they use endogenous measures of genotype and social environment. To overcome these

estimation issues, new methods that provide adequate identification of exogenous genetic (G), environmental (E) and gene by environment interaction (GxE) effects are needed to provide a comprehensive way forward in understanding how the social factors influencing developmental and health outcomes interact with the biological factors that may also influence outcomes of interest. That is the task of the present study.

Data

The data for this study comes from the National Longitudinal Study of Adolescent to Adult Health (Add Health), a nationally representative cohort study on the health and behavior of adolescent school children first interviewed in 1994-95 across a sample of more than 130 middle and high schools across the United States (Harris 2010). An in-school survey was administered to every student present at each of the 130 schools asking them to self-report their social and health behaviors, friends, and family and school context. A second in-home survey was administered to a random sample of the students in each school, which entailed a far more detailed questionnaire about their behaviors, attitudes, health, parents, siblings, and home life. During the fourth wave of data collection in 2007-2008, 96% of the respondents who participated in the original in-home surveys (15,159 out of 15,701 individuals) agreed to provide biospecimen (in the form of saliva and capillary whole blood samples) to be immediately genotyped for specific SNPs and candidate genes; and, of those individuals, 80% (12,254) agreed to have their biospecimen archived for future use. With the more recent development of genotyping technology and the resulting decrease in cost, the archived samples were genotyped on about one million genetic markers, which provide genome-wide data for polygenic score construction (Harris et al. 2013). A number of standard genetic quality control screening

procedures were performed on the data, after which genotyping data was available for 9,975 individuals.[4] Further, following Fletcher (2010), we limit the sample to students in schools with a 12th grade and exclude students not assigned sample weights or who attended a grade with fewer than 20 total students in our sample. This leaves us with 4,950 respondents in total, and with genotyping data in classrooms for a reasonable number of peers.

The outcome variable, which measures individual smoking habits, is the response to the following question: "During the past 30 days, on the days you smoked, how many cigarettes did you smoke each day?" It is coded as 0 for individuals who did not smoke in the past thirty days, 1 for individuals who smoked 1 cigarette on the days that they smoked, and so on. This question was asked in the first wave of the data, between 1994 and 1995, when respondents were in middle or high school. We construct polygenic scores using the TAG consortium's "cigarettes per day" to build a PGS (we also test the PGS based on age at first smoking as a robustness check). The "cigarettes per day" measure for smoking both accords with our dependent variable as well as captures an aspect of smoking likely linked to nicotine dependence. For every individual in the data, we construct a smoking PGS, characterize their smoking behavior, construct measures of their exposure to peers' smoking behavior and peers' smoking PGSs.

Methods

[4] Two Illumina platforms were used for genotyping Add Health genetic data. Illumina Human Omni1-Quad BeadChip was used for roughly 80% of the sample and includes over 1.1 million genetic markers. Illumina Human Omni-2.5 Quad BeadChip was used for the remainder of the sample and includes 2.5 million markers. A series of quality control procedures were performed on the SNP and individual level. Genetic markers with call rates < 90% and minor allele frequency < 0.5% were excluded. SNPs with Hardy-Weinberg Equilibrium P-value < $5x10^{-5}$ were also excluded.

To overcome problems commonly associated with identifying peer effects (which also plague social genetic effects), we use the *linear-in-means model of social interactions* proposed by Manski (1993). The approach of this statistical model is to model an individual's outcome as a function of the average outcomes of their peers iteratively excluding each individual. We combine this strategy with an instrumental variable/fixed effects model in order to get at causal social genetic effects that go through peer behavior. Our key identifying assumption is that conditional on school-level variation (i.e. school fixed effects), the distribution of genetic propensity to smoke is as-if-random in a given grade. Since genes are assigned at birth, and being in a given grade within a school is roughly defined by the year and month of birth, it is reasonable to assume the absence of selection into the grade environment by the individual based on his or her genes.[5] We include a number of control variables (such as sex and socioeconomic indicators at the grade level) to ensure that educational mechanisms, such as holding boys and individuals with lower socioeconomic status back a grade, are not leading to systematic compositional differences in the gene pool. Using this method we can circumvent the reflection problem (where it is unclear whether peers are influencing the individual or vice versa) and contextual bias (where a common factor may be leading to the observed similarity between two peers).

Our units of analysis for the outcome are individual smoking behaviors in adolescence whereas the independent variable captures the exposure of the focal individual to their grade-level peers. Following previous work on peer effects (Fletcher, 2010; Fletcher, 2012a), we look

[5] This would not be true for example if there was a sudden demographic change in a given school district, which led to compositional differences between grades, or radical year to year shifts in behavior. This is largely not the case: 9th graders and 10th graders in most schools are the same, other than being a year apart.

at the composition of individuals in a given grade because grades are relevant ecological contexts for adolescent behaviors and social interactions both empirically as well as theoretically. Theoretically, considering the distribution of a behavior in a context larger than self-selected friends, reveals something about the social meaning of the behavior which may lead to variation in engagement rates (Bearman and Brückner, 2015; Frank et al., 2008; McFarland et al., 2014). Empirically, most friendships in the Add Health dataset are within, rather than between, grades. For instance, the network segregation of friendship within grades is far higher than the network segregation of friendship according to race or sex (see Figures B1-B3 in Appendix B). High segregation of friendship by grade shows that grade is the relevant ecological setting for social interaction (Moody, 2001). Grade-mates' statuses and behaviors have also been found to be salient predictors of students' outcomes. For example, researchers have shown effects of gender, race and socioeconomic status of grade-mates on individuals' outcomes (Hanushek et al., 2003; Hoxby, 2000). Grade-mates' behaviors have also been linked to individual behaviors such as drinking, smoking, dropping out of high school, church attendance, and illegal drug use in previous studies (Gaviria and Raphael, 2001; Fletcher, 2010; Fletcher, 2012a).

Because we know that individuals actively select their friends, rather than examine friendships directly, we study the effects of the larger group within which individuals form friendships (i.e. grade) on their outcomes. This "intent-to-treat" strategy allows us to circumvent common problems of peer effect identification that arise because of peer selection. To our knowledge only one scholarly paper to date have used this model to address causal peer influence on smoking. Fletcher (2010) used classmates' parents' smoking status and the number of older siblings an individual has as instruments to estimate the peer effects of smoking. However, these instruments pose a number of methodological problems. First, it is possible that

using household smoker as an instrument violates the monotonicity assumption underlying an instrumental variable design. This assumption holds that the treatment influences everyone whom it affects in the same way. It implies that there will be no "defiers" – i.e. individuals who act in opposition to the treatment. In the case of having a parent who smokes, however, the effect on the child, in terms of making them more likely to smoke, may actually not hold. A child of a smoker may well learn about the negative health and social consequences of smoking firsthand, and actively avoid smoking. Similar acts of defiance have been shown for some children of alcoholics and drug addicts (Mannella et al., 2000; Zimmerman et al., 2013) and it is reasonable that they hold for smoking as well. Second, Fletcher's exclusion restriction – that parent's smoking behavior and having older siblings could only influence peer smoking through the individual's own smoking behavior – may be violated if an individual is directly exposed to the smoking behaviors of their peers' parents or older siblings. An adolescent visiting a classmate's house after school may observe her classmate's parent or sibling smoking, and may be directly influenced by their behavior, even if the classmate does not smoke. Finally, it is feasible that parental smoking and having older siblings are associated with an individual's smoking behavior through channels other than the direct transmission of smoking behavior itself. An individual in a household with older siblings and smokers may be more likely to engage in a number of deviant behaviors, and this engagement in deviant behaviors, rather than the learned behavior of smoking, may be the pathway linking household characteristics to adolescent behavior.

By controlling for peers' family characteristics (i.e. peer parental smoking and older siblings), rather than using them as instruments, and, alternatively, using peer smoking PGS as the instrument, this paper addresses the previously mentioned shortcomings of the peer effects literature and provides a non-biased estimate of the peer effects of smoking. Beyond the more

consistent identification strategy of peer effects of smoking, the instrument used in this paper is itself of substantive interest. Following the social genetic effects literature, our model reveals the quantity of social genetic effects directly: we estimate the effect of peer genes through peer behavior on an individual's outcome. The contribution of other's genes to an individual's outcome, controlling for individual's own genes, highlights the importance of the social environment in the role that genes play in predicting our outcomes. Genes are often conceptualized as being a non-social force in shaping an individual's outcome. The literature on gene-by-environment effects has long established that the influence of genes on an outcome is dependent on and mediated by the social-environmental landscape. Here, we will show that the aggregate influence of others' genes on an individual's outcome is considerably larger than that of her own genes.

In the models that follow, we control for the sex, race, maternal education, familial smoking behavior, number of siblings and family income of the ego as well as the mean of those variables at the grade-level. As discussed before, the possibility of the direct influence of parent or sibling smoking on peers is a potential violation of the exclusion restriction assumption. By controlling for the presence of household smoker (and the average household smokers of peers in a grade), we block the direct pathway between parental/sibling smoking on peers. In addition, we control for the ego's own smoking PGS, so our point estimates show the effect of peers above and beyond one's own genetic risk for smoking. We initially include individuals of all races in the analyses, but control for the individual's own race and the racial composition of the grade.[6]

[6] The main worry with this formulation is that smoking polygenic smoke was trained on white individuals and therefore might not be a strong instrument for non-white respondents. In Appendix A, we show that the polygenic score is positively correlated with and predicts smoking behavior in non-whites. This is confirmed by other studies that show overlapping SNPs in smoking GWAS conducted on African and European ancestry populations (Otto et al., 2016).

We later conduct subgroup analyses by race to detect whether peer effects are racially heterogeneous.

The first stage of the two stage least square is the following:

(1) $$\bar{Y}_{-igs} = \alpha + \gamma_1 \bar{Z}_{-igs} + \gamma_2 G_{igs} + \gamma_3 W + \xi_{igs}$$

where *igs* indexes an individual *i* in grade *g* in school *s* and -*igs* signifies that the estimate excludes the individual. *Y* is smoking behavior of the individual, $\bar{Y}_{-igs}$ is the average smoking level of peers excluding the individual, *G* is the smoking polygenic score for a given individual, $\bar{Z}_{-igs}$ is the average level of the smoking polygenic score within a school and grade excluding the individual and *W* is the vector of controls (including individual level and grade level variables). The intercept is α, and ξ is the additive error term.

Equation 2 portrays the second stage regression of our model:

(2) $$Y_{igs} = \alpha + \rho_1 \hat{\bar{Y}}_{-igs} + \rho_2 G_{igs} + \rho_3 W + v_{igs}$$

where $\hat{\bar{Y}}_{-igs}$ indexes the fitted values from the first stage regression, and v is the error term.

Using this model we can identify the peer effect of smoking as instrumented by the genes of peers. Not only is the model causally identified, it also reveals a substantively interesting quantity. It gives us a point estimate of the impact of one's peers' genetic risk of smoking on an individual's smoking as it goes through peer behavior. In addition to these linear models, we try alternative model specifications and functional forms including instrumental variable Poisson regression and substituting a logistic-in-means model for the linear-in-means model testing a sigmoid function of peer influence rather than a linear one; both alternative functional forms yield substantively similar results to our results presented below (see Appendix D).

We extend the model further to explore whether the effect of peer genes is moderated by an individual's own genes. We conceive of this extended model as a gene-by-environment

interactional framework in which the environment is defined as the genetic makeup of an individual's peers, or what Domingue et al. (2017) refer to as "social epistasis"[7]. We might expect individuals who have higher polygenic scores for smoking to be more influenced by their friends' smoking behaviors. However, it is also possible that these individuals' smoking habits are more inelastic to their friends' behaviors than those of their lower-PGS counterparts. To test whether an individual's own genes mediate the influence of peers on their own behavior, we subset the sample by polygenic score levels and run three separate models for each polygenic level.

Results

We begin by examining whether there is any evidence of social genetic effects (i.e. effect of others' genes on one's own outcome) in these data. Specifically, we examine the effect of peer polygenic score on individual smoking behavior, net of one's own polygenic score and a vector of controls, in addition to school and grade fixed effects. This model is the reduced form of the two stage least square model. Although this specific model does not specify the mechanisms through which social genetic effects work, it is still robust to selection, reflection and contextual biases and therefore identifies the causal effect of peer genes on individual behavior. Since being in a specific grade, within a given school, is determined by birth year, it is unlikely that active selection into grades, on the part of individuals and/or parents, or gene-environment correlations, are driving the results (see school by grade descriptive plots in Appendix A). Meanwhile, genes are determined at birth and therefore ward against reverse causation and the reflection problem;

[7] Epistasis is the technical term for a non-additive, joint effect of more than one genetic locus—a GxG interaction effect.

in addition, we remove each individual when calculating their peer exposure score. Contextual biases – in which a common factor makes both the individual as well as the peer group engage in a behavior –are also unlikely, because we examine how the gene pool within a grade, compared to other grades in that school, affects an individual's smoking behavior.

The results show a positive effect of average classmates' smoking polygenic score on individual smoking outcome, net of the individual's own polygenic score. In other words, we find that the genes of one's peers contribute to one's outcomes, even when controlling for one's own genes. A standard deviation increase in peer smoking polygenic score results in a 0.27 standard deviation increase in the number of cigarettes an individual smokes in a day. To contextualize this effect, we compare it to the magnitude of the effect of an individual's own genes on his or her smoking behavior. However, as Figure 1 shows, the standard errors of the estimate for mean classmate polygenic scores increase considerably. This is because there is only minimal variation in individual exposures to grade-mates within the same grade (the only source of variation is that the individual is excluded from their own measures of exposure), and thus our unit of analysis is essentially the grade, in contrast to models that treat sex, age or polygenic score at the individual level. With about 300 included grades across all schools, we have about 53% statistical power. Given the difficulty of identifying causal effects of peer behavior, we opt for sacrificing statistical efficiency for statistical consistency. Even with the loss of some statistical efficiency when moving the mean grade polygenic score as the independent variable, we can be confident that the magnitude of the effect of other's genes on an individual's smoking outcome is at least half as large as the individual's own smoking polygenic score. The lowest bound of the estimate, at a 95% confidence interval, corresponds to 0.03 - an estimate which 60% as large as the individual's own smoking polygenic score (0.05) and sex (0.05), and 30% of

the effect of age (0.1). [8] While being older and being male have been shown to be two of the strongest predictors of smoking (Conard et al. 1992; U.S. Department of Health and Human Services 2012, 2014), our results suggests that the magnitude of the effect of peers' PGS may have an even larger effect, with the lowest bound showing an effect size that is 60% as large as the effect of being male, and 30% as large as an additional standard deviation of age. The literature on adolescent smoking consistently points to the importance of peers in determining when an adolescent starts and ceases smoking and the number of cigarettes they will consume, but no research to date has identified peers' *genes* as an important, alternative source of influence.

[Figure 1 about here]

The reduced form model shows that the magnitude of the effect of peer genes on smoking is larger than the effect of one's sex, age and smoking polygenic score. The implication is that a large part of the effect of genes on outcomes, is in fact, social, due to the genetic composition of the peer group. In the next set of analyses, we incorporate peer genes into an instrumental variable framework thereby restricting the pathway of the effect of genes to the smoking behavior of peers. We use a two stage least squares using mean peer smoking PGS as an instrument for peer smoking behavior. Mean smoking polygenic score is a strong instrument, meaning (in the instrumental variables literature) that it is strongly correlated with the outcome, individual smoking behavior. The F statistic from the first stage regression is 53, well beyond the conventional desirable threshold of 10 (Bound, Jaegar and Baker, 1995). This gives us confidence that average peer smoking PGS is an ideal instrumental variable.

[8] For the last two models (age and sex) we did not include grade fixed effects, since much of the effect of age would be picked up by the grade.

The results, shown in Figure 2, show that a one standard deviation increase in peer smoking behavior (as instrumented by peer smoking polygenic score), results in a 0.29 standard deviation increase in the number of cigarettes an individual smokes a day. The magnitude of our coefficient is about twice as large as that of Fletcher's (2010).[9] In addition, the magnitude of the effect of peer smoking is very similar to that of peer polygenic score from the reduced form model in Figure 1, which suggests that the effect of peer genes goes almost entirely through peer behavior. This is both in line with what we would expect, given that the phenotype used in the gene discovery phase was smoking behavior, and it also makes us more confident that the exclusion restriction is not violated. Considering the unstandardized coefficients reveals that an increase of one cigarette on average by one's grade-mates, leads an individual to smoke 0.89 more cigarettes a day: the effect of average peer behavior on an individual's behavior is very large.[10]

As a robustness check for our modeling choices, we conduct the same analyses on height.[11] Height is a biological/developmental trait and we do not expect it to be influenced by peer height polygenic scores or actual peer height. In the figure below, we plot the coefficients

[9] Fletcher (2010) used a binary indicator for smoking. We replicated Fletcher's models, and standardized the variables in his models in order to enable comparison. Fletcher's standardized coefficients show that a one standard deviation increase in peer smoking behavior results in a 0.14 standard deviation increase in being a current smoker. This is in comparison to 0.29 using our model.

[10] As a sensitivity check, we conducted the same set of analyses (the reduced form in addition to the two-staged least squares) for a set of 14 schools where all students were followed was included in the in-home follow up sample (and therefore eligible for genotyping). The results are broadly consistent with what we find here, but lack statistical significance on account of lower power due to smaller sample size.

[11] The results are robust to clustered standard errors. We also provide additional robustness checks on functional form specifications in Appendix D.

for height alongside smoking. The results show that the social genetic effect as well as peer effects on height are not significantly different from zero.

[Figure 2 about here]

We next turn to the moderating role an individual's own PGS may play in the effect of his peer environment in explaining his smoking behavior. If we consider other's genes as an individual's environment, we can refer to this effect as a gene-by-environment (GxE) interaction. Studies that consider GxE effects in smoking have looked at the mediating role of psychosocial factors (Koval et al., 2000), birth cohort (Domingue et al., 2015), variation in state tobacco taxes (Fletcher, 2012b) and natural experiments of exposure to tobacco, such as the Vietnam War era draft lottery (Schmitz & Conley, 2015) on smoking outcomes. However, no study to date has identified the way an individual's genes interact with the behaviors of her peers in determining her smoking outcome. Does the influence of peer's smoking behavior differ for individuals with different polygenic scores? There are three plausible hypotheses. First, having a high polygenic score may make individuals more prone to influence by their smoking peers. Studies of other traits point to this possibility. A study of the polygenic score for alcohol dependence finds that the contribution of genes to developing an alcohol dependence is moderated by peer deviant behavior (Salvator et al., 2014). Peer substance use has also been shown to moderate the effect of individual genes on individual mental health (Salvator et al., 2015). Although these studies are not causally identified, and therefore cannot tell us whether the actual effects are working through peer behavior (as opposed to say, selection), they do provide suggestive evidence for the association between higher genetic propensity and higher peer influence.

Secondly, it could be the case that those with high polygenic scores are in fact less susceptible to their environmental contexts. The argument here is that those with high genetic risk of engaging in a behavior, especially if that behavior has a biological basis such as nicotine dependence, are going to engage in that behavior regardless of their environment. For instance, high risk adolescents may smoke regardless of their peer environment thanks to the intrinsic rewards they get from nicotine. Fletcher (2012b) shows that individuals with a protective polymorphism on the nicotinic acetylcholine receptor (those with lower genetic propensity to become dependent on nicotine) reduced their smoking habits in response to state tobacco tax policies, while others did not. Other studies have found that the polygenic score for smoking is more predictive of smoking behavior in more recent birth cohorts, after the dangers of smoking had become widely known (Domingue et al., 2016). The authors hypothesize that as the availability of information about the health risks of smoking, and subsequently the stigmatization of smoking behavior, increased over the course of the twentieth century, the biology of tobacco use became a more salient predictor of tobacco consumption. Guo et al. (2015) generalize this framework and look for differential effects of peers on binge drinking by low, medium and high levels of polygenic score. They show that individuals in the middle of the distribution are susceptible to peer influence but not those in the low or high ranges. They propose a "swing theory" of gene by environment interactions, whereby individuals with too high a genetic predisposition are likely to engage in a behavior regardless of their peers, and those with too low a genetic predisposition are not likely to engage in a behavior anyway. The group that is likely to be swayed by their peers is the individuals who fall in the middle of the polygenic score distribution. These dynamics are plausible given that 90% of U.S. adolescents try cigarettes before age 18 (U.S. Department of Health and Human Services 2012, 2014) and whether or not

they become smokers may be highly dependent on both their hard-wired response to that experimentation as well as the behavioral context surrounding them.

Third, it is possible that peers influence an individual regardless of her genetic disposition towards the behavior. We have shown that genes influence the smoking outcome of an individual (see Appendix A), and we have shown that the environment influences individual behavior above and beyond one's polygenic risk of smoking (see the reduced form of the model in Figure 1). The two however, may be independent from one another such that an individual's genetic predisposition to smoke is not affecting how likely they are to adopt their peers' behaviors. The susceptibility to adapting others' behaviors may in fact have a different genetic architecture yet to be discovered.

We test these three hypotheses in two different ways. First, we include an interaction term for individual and peer PGS in the reduced form model as in equation 3. Next following Guo et al.'s (2015), we test the swing theory of gene-by-environment interaction, splitting the sample into equal groups with low, medium and high polygenic scores for smoking. Although none of the results reach statistical significance, individuals with medium and high smoking PGS have a positive coefficient for the influence of peer genes on their smoking behavior, while those with low smoking PGS tend to have a negative one. This is modest evidence that individuals with a low genetic propensity to smoking are negatively influenced by their peer's smoking habits. Although the confidence intervals are slightly overlapping between the estimates for the low group on the one hand and medium and high on the other, the diverging point estimates provide suggestive evidence of the first hypothesis put forth in this section: that those with higher genetic propensity towards the behavior are more likely to be influenced by their peers.

[Figure 3 about here]

Heterogeneous effect by Race

We now examine racial dynamics of the peer influence of smoking. Who is influencing whom? The literature has established a robust positive association between being white and smoking cigarettes (Harris et al. 2006; Harrell et al., 1998; Johnson & Hoffman, 2000). This trend holds in our sample. The average cigarettes smoked per day is 2.3 for whites, 1.1 for Hispanics and 0.4 for blacks. Even when controlling for socioeconomic factors such as family income and maternal education, the individual's smoking polygenic score, sex and grade and school fixed effects, whites are significantly more likely than both blacks and Hispanics to smoke (analyses available upon request).

However, the literature has not thoroughly explored the differential influence of and susceptibility to peers of different races. In the analyses that follow, we split the sample according to ego's race and re-run the previous analyses to analyze the differential susceptibility of white and non-white respondents to peer influence.[12] This analysis looks at the *susceptibility* of the egos by race to peer influence in general and therefore includes peers of all races.

[Figure 4 about here]

As Figure 4 shows, both whites and non-whites are positively affected by peer smoking. Although whites tend to have a higher point estimate for the effect of peer smoking, it is not significantly different from the susceptibility of non-whites to the smoking behavior of their peers.

[12] Because the sample consists primarily of white respondents, splitting it into blacks and Hispanics separately would strain the statistical power of the analyses. In addition, Hispanics and blacks are similarly less likely to smoke than whites. We therefore, combine the two racial categories into a "non-white" category.

We next look at race-specific and interracial peer effects. Are whites more susceptible to the effect of their white, as opposed to nonwhite, peers and vice versa? Here, our data only allows us to examine the effect of white peers on white and nonwhite individuals because there are not enough school-grades with sufficient populations of nonwhites.

[Figure 5 about here]

We find that white peers have a positive effect on white egos, but we do not find an effect significantly different from zero for non-white egos and egos of all races. Given the large confidence intervals (not fully shown in the figure due to scaling) however, and the fact that the sample size reduces to 310 respondents, this may be due to the loss of statistical power. What we do find, however is a robust within-race influence and susceptibility for white respondents.

Discussion

In this paper, we introduce a novel strategy that uses genes to estimate peer effects. We show that genes can be used as instrumental variables in a linear-in-means estimation model with school and grade fixed effects. Genes are determined at birth, are associated with the trait of interest, and their distribution in a grade, controlling for grade and school level characteristics and fixed effects is as-if random. These properties make genes good candidates for instrumental variables in social science research. This strategy helps overcome three common problems in estimating peer effects: 1) the selection or homophily problem whereby individuals select their peers based on observed or unobserved common characteristics, 2) contextual endogeneity, where an attribute of the context (i.e. teacher, or a health class offered to individuals in the same

grade of a school) affects both the ego and the peers, and 3) the reflection problem, whereby it is unclear whether ego is affecting peers or vice versa.

We find evidence of peer effects on adolescent smoking behavior. One additional cigarette smoked by one's grade-mates leads, on average, to an individual to smoke 0.89 more cigarettes a day. This is larger than estimates of the peer effect of smoking identified in previous studies. In fact, the magnitude of the effect is twice as large as those of Fletcher (2010) who estimates the peer effect of smoking using a similar estimation strategy but with different instrumental variables potentially susceptible to violation of a number of assumptions.

The endeavor undertaken here, however, is more than a simple estimation task. Using genes as instruments allows us to estimate the effect peer genes have on an individual's outcome, through their behavior. In the reduced form analysis, we show that the magnitude of the effects of peer genes on an individual's smoking habit, is at least half the magnitude of the effect of sex and the individual's own genes, and 30% the magnitude of the effect of age on smoking behavior.

Conceptually, our approach treats others' genes as part of the social environment of ego. This is a break from the view that genes are forces which act solely on the individual. It is true that genes encode proteins in an individual's body, but individuals' actions and bodies are shaped by society and have social implications (Freese et al. 2003; also see Zerubavel et al. 2015). Our results reiterate the social nature of our biology. The boundaries of genetic influences are not necessarily contained in our own bodies.

The social implications of social genetic effects are important for policy, especially as they relate to adolescent smoking. That an individual's smoking behavior, and ultimately health outcomes, are affected by his or her peers' genes is important to consider for understanding

social multiplier effects. Our findings also show that grade is indeed a relevant setting for peer influence. Our focus on the grade, as opposed to the friend group, as an ecological environment makes our results applicable and more readily actionable to smoking prevention efforts in schools. We therefore contribute to broadening the understanding of "peer effects". Humans are social creatures and take cues from members of their social environment beyond dyadic relations and interactions, which are the dominant way of operationalizing interpersonal dynamics in the social network literature. Though a behavior like smoking could hypothetically imply a one-on-one relationship between a smoker and his cigarettes, devoid of social meaning or identity for that individual, lay knowledge and scholarship tell us that being a smoker, even occasionally smoking a cigarette, entails a meaningful social identity (Falomir & Invernizzi 1999; Lennon et al. 2005; Haslam et al. 2009). That social identity, however, is context dependent. Being one of five smokers in a grade full of adolescents who look down on smoking is qualitatively different than being a smoker in a grade where smoking is common place, even if, in both cases, four of your best friends are smokers. In this regard, using the grade as the relevant peer group allows us to analyze the more complex cultural dynamics of smoking behavior, which are lost if we narrowly focus on dyadic or friend group relations.

We establish that genes influence individual-level smoking outcomes, and peers influence individual smoking behavior above and beyond one's polygenic risk to smoke; next we looked at the mediating role of the former (individual PGS) on the latter (individual's susceptibility to peer influence in terms of smoking behaviors). Even though our results did not reach statistical significance, the magnitudes of the estimates provide suggestive evidence that individual polygenic scores in the middle or higher terciles are more susceptible to the influence of peer smoking genes, and by extension peer smoking behavior. Further research is required to

ensure the robustness of this finding, and it may challenge theories of genetic inelasticity – whereby individuals with high polygenic scores are thought to engage in a given behavior regardless of their social environment. It also shows that individual genetic predisposition toward a behavior may not be independent from the individual's susceptibility towards that behavior.

We also examined differential peer effects of smoking by race. We establish that, as in other studies, whites tend to smoke more than non-whites. However, in terms of being influenced by peer smoking behavior, non-white egos are as likely as white egos to smoke more cigarettes in response to peer smoking behavior. We show that white peers have a positive impact on white *and* non-white egos, and that this is not the case for non-white peers, though we may be underpowered to adequately measure the effect of non-white peer effects on non-white individuals' smoking behavior.

Although the only outcome we consider in this paper is smoking, there is nothing about the model that would constrain this to be so. Future scholars may seek to apply this method to other socially driven outcomes, including health outcomes, that display an element of contagion and are in part influenced by genetic disposition — be that actual communicable disease as driven by the immunological profiles of peers around us, to depression and suicide.

Work Cited:

Alexander, C., Piazza, M., Mekos, D., & Valente, T. (2001). Peers, schools, and adolescent cigarette smoking. *Journal of adolescent health*, *29*(1), 22-30.

Barban, N., Jansen, R., De Vlaming, R., Vaez, A., Mandemakers, J. J., Tropf, F. C., ... & Tragante, V. (2016). Genome-wide analysis identifies 12 loci influencing human reproductive behavior.

Baud, A., Mulligan, M. K., Casale, F. P., Ingels, J. F., Bohl, C. J., Callebert, J., ... & Stegle, O. (2017). Genetic variation in the social environment contributes to health and disease. *PLoS genetics*, *13*(1), e1006498.

Beach, S. R., Brody, G. H., & Kogan, S. M. (2016). 4 Molecular Genetics and Epigenetics in the Context of Family and Couple Problems. *The Oxford Handbook of Relationship Science and Couple Interventions*, 39.

Bearman, P. S., & Brückner, H. (2015). Promising the Future: Virginity Pledges and First Intercourse1. *American journal of Sociology*.

Bearman, P. S., Moody, J., & Stovel, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *American journal of sociology*, *110*(1), 44-91.

Belsky, D. W., & Israel, S. (2014). Integrating genetics and social science: Genetic risk scores. *Biodemography and social biology*, *60*(2), 137-155.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, *90*(430), 443-450.

Cawley, J., Han, E., Kim, J. J., & Norton, E. C. (2017). *Testing for Peer Effects Using Genetic Data* (No. w23719). National Bureau of Economic Research.

Centers for Disease Control and Prevention (2008). Smoking-attributable mortality, years of potential life lost, and productivity losses--United States, 2000-2004. *MMWR. Morbidity and mortality weekly report*, *57*(45), 1226.

Chakrabarti, P., & Frye, M. (2017). A mixed-methods framework for analyzing text data: Integrating computational techniques with qualitative methods in demography. *Demographic Research*, *37*, 1351-1382.

Conley, D., & Fletcher, J. (2017). *The Genome Factor: What the social genomics revolution reveals about ourselves, our history, and the future*. Princeton University Press.

Conley D, Laidley T, Belsky D, Fletcher J, Boardman J, Domingue B. (2016). Assortative mating and differential fertility by phenotype and genotype across the 20th century. Proc Natl Acad Sci.

Conrad, K. M., Flay, B. R., & Hill, D. (1992). Why children start smoking cigarettes: predictors of onset. *Addiction*, *87*(12), 1711-1724.

Cook, C. J., & Fletcher, J. M. (2017). *High School Genetic Diversity and Later-life Student Outcomes: Micro-level Evidence from the Wisconsin Longitudinal Study* (No. w23520). National Bureau of Economic Research.

Cross-Disorder Group of the Psychiatric Genomics Consortium. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*, *381*(9875), 1371-1379.

Delph, L. F., & Kelly, J. K. (2014). On the importance of balancing selection in plants. *New Phytologist*, *201*(1), 45-56.

Domingue, B. W., & Belsky, D. W. (2017). The social genome: Current findings and implications for the study of human genetics. *PLoS genetics*, *13*(3), e1006615.
Domingue, B. W., Belsky, D., Fletcher, J., Conley, D., Boardman, J. D., & Harris, K. M. (2017). Analysis of genetic similarity among friends and schoolmates in the National Longitudinal Study of Adolescent to Adult Health (Add Health). *bioRxiv*, 107045.
Domingue, B. W., Conley, D., Fletcher, J., & Boardman, J. D. (2016). Cohort effects in the genetic influence on smoking. *Behavior genetics*, *46*(1), 31-42.
Domingue, B. W., Fletcher, J., Conley, D., & Boardman, J. D. (2014). Genetic and educational assortative mating among US adults. *Proceedings of the National Academy of Sciences*, *111*(22), 7996-8000.
Falomir, J. M., & Invernizzi, F. (1999). The role of social influence and smoker identity in resistance to smoking cessation. *Swiss Journal of Psychology/Schweizerische Zeitschrift für Psychologie/Revue Suisse de Psychologie*, *58*(2), 73.
Fletcher, J. M. (2010). Social interactions and smoking: Evidence using multiple student cohorts, instrumental variables, and school fixed effects. *Health Economics*, *19*(4), 466-484.
Fletcher, J. M. (2012a). Peer influences on adolescent alcohol consumption: evidence using an instrumental variables/fixed effect approach. *Journal of Population Economics*, *25*(4), 1265-1286.
Fletcher, J. M. (2012b). Why have tobacco control policies stalled? Using genetic moderation to examine policy impacts. *PloS one*, *7*(12), e50576.
Fletcher, J. M., & Conley, D. (2013). The challenge of causal inference in gene–environment interaction research: Leveraging research designs from the social sciences. *American journal of public health*, *103*(S1), S42-S45.
Fowler, J. H., Settle, J. E., & Christakis, N. A. (2011). Correlated genotypes in friendship networks. *Proceedings of the National Academy of Sciences*, *108*(5), 1993-1997.
Frank, K. A., Muller, C., Schiller, K. S., Riegle-Crumb, C., Mueller, A. S., Crosnoe, R., & Pearson, J. (2008). The social dynamics of mathematics coursetaking in high school. *American Journal of Sociology*, *113*(6), 1645-1696.
Freese, J., Li, J. C. A., & Wade, L. D. (2003). The potential relevances of biology to social inquiry. *Annual Review of Sociology*, *29*(1), 233-256.
Gaviria, A., & Raphael, S. (2001). School-based peer effects and juvenile behavior. *The review of economics and statistics*, *83*(2), 257-268.
Gibson, G. (2009). Decanalization and the origin of complex disease. *Nature Reviews Genetics*, *10*(2), 134-140.
Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, *13*(2), 135-145.
Guo, G., Li, Y., Wang, H., Cai, T., & Duncan, G. J. (2015). Peer influence, genetic propensity, and binge drinking: a natural experiment and a replication. *American journal of sociology*, *121*(3), 914-954.
Guo, G., Wang, L., Liu, H., & Randall, T. (2014). Genomic assortative mating in marriages in the United States. *PloS one*, *9*(11), e112322.
Guryan, Jonathan, Kory Kroft, and Matthew J. Notowidigdo. 2009. "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments." *American Economic Journal: Applied Economics*, 1(4): 34-68.
Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement?. *Journal of applied econometrics*, *18*(5), 527-544.

Harrell, J. S., Bangdiwala, S. I., Deng, S., Webb, J. P., & Bradley, C. (1998). Smoking initiation in youth: the roles of gender, race, socioeconomics, and developmental status. *Journal of Adolescent Health*, *23*(5), 271-279.

Harris, K.M., Halpern, C.T., Whitsel, E.E., Hussey, J., Killeya-Jones, L.A., Tabor, J., Elder, G.H., Hewitt, J., Shanahan, M., Williams, R.B., Siegler, I.C., & Smolen, A. (2013). Social, behavioral, and genetic linkages from adolescence into adulthood. *American Journal of Public Health*, *103 (S1)*, S25-S32.

Harris, K.M. (2010). An integrative approach to health. *Demography*, *47(1)*,1-22.

Harris, K.M., Gordon-Larsen, P., Chantala, K., & Udry, J.R. (2006). Longitudinal trends in race and ethnic disparities in leading health indicators from adolescence to young adulthood. *Archives of Pediatrics and Adolescent Medicine*, *160*,74-81.

Haslam, S. A., Jetten, J., Postmes, T., & Haslam, C. (2009). Social identity, health and well-being: an emerging agenda for applied psychology. *Applied Psychology*, *58*(1), 1-23.

Hoffman, M. A., Cointet, J., Brandt, P., Key, N. & Bearman, P. (2017). The (Protestant) Bible, the (Printed) Sermon, and the Word(s): The Semantic Structure of the Conformist and Dissenting Bible, 1660-1780. *Forthcoming in Poetics.*

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics*, *115*(4), 1239-1285.

Hugh-Jones D, Verweij KJ, Pourcain BS, Abdellaoui A. Assortative mating on educational attainment leads to genetic spousal resemblance for causal alleles. Intelligence. 2016

International Schizophrenia C, Purcell SM, Wray NR, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009;460(7256):748-752.

Jamal, A., Homa, D. M., O'Connor, E., Babb, S. D., Caraballo, R. S., Singh, T., ... & King, B. A. (2015). Current cigarette smoking among adults—United States, 2005–2014. *MMWR Morb Mortal Wkly Rep*, *64*(44), 1233-1240.

Johnson, R. A., & Hoffmann, J. P. (2000). Adolescent cigarette smoking in US racial/ethnic subgroups: findings from the National Education Longitudinal Study. *Journal of Health and Social Behavior*, 392-407.

Kong, A., Thorleifsson, G., Frigge, M. L., Vilhj.lmsson, B. J., ... & Stefansson K., (2017) Non-Transmitted Alleles and Genetic Nurturing. *Working paper*.

Koval, J. J., Pederson, L. L., Mills, C. A., McGrady, G. A., & Carvajal, S. C. (2000). Models of the relationship of stress, depression, and other psychosocial factors to smoking behavior: a comparison of a cohort of students in grades 6 and 8. *Preventive Medicine*, *30*(6), 463-477.

Legewie, J., & DiPrete, T. A. (2012). School context and the gender gap in educational achievement. *American Sociological Review*, *77*(3), 463-485.

Lennon, A., Gallois, C., Owen, N., & McDermott, L. (2005). Young women as smokers and nonsmokers: a qualitative social identity approach. *Qualitative Health Research*, *15*(10), 1345-1359.

McFarland, D. A., Moody, J., Diehl, D., Smith, J. A., & Thomas, R. J. (2014). Network ecology and adolescent social structure. *American sociological review*, *79*(6), 1088-1121.

McGue, M., Vaupel, J. W., Holm, N., & Harvald, B. (1993). Longevity is moderately heritable in a sample of Danish twins born 1870–1880. *Journal of gerontology*, *48*(6), B237-B244.

Melamede, R. (2005). Cannabis and tobacco smoke are not equally carcinogenic. *Harm Reduction Journal*, *2*(1), 21.

Mennella, J. A., & Garcia, P. L. (2000). Children's hedonic response to the smell of alcohol: effects of parental drinking habits. *Alcoholism: Clinical and Experimental Research*, *24*(8), 1167-1171.

Moody, J. (2001). Race, school integration, and friendship segregation in America. *American journal of Sociology*, *107*(3), 679-716.

Mounts, N. S., & Steinberg, L. (1995). An ecological analysis of peer influence on adolescent grade point average and drug use. *Developmental Psychology*, *31*(6), 915.

Murphy, G. P., Lawrence, W., & Lenhard, R. E. (1995). *American Cancer Society textbook of clinical oncology*. The Society.

Okbay, A., Baselmans, B. M., De Neve, J. E., Turley, P., Nivard, M. G., Fontana, M. A., ... & Gratten, J. (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature genetics*, *48*(6), 624-633.

Otto, J. M., Gizer, I. R., Bizon, C., Wilhelmsen, K. C., & Ehlers, C. L. (2016). Polygenic risk scores for cigarettes smoked per day do not generalize to a Native American population. *Drug and alcohol dependence*, *167*, 95-102.

Pallonen, U. E., Prochaska, J. O., Velicer, W. F., Prokhorov, A. V., & Smith, N. F. (1998). Stages of acquisition and cessation for adolescent smoking: an empirical integration. *Addictive behaviors*, *23*(3), 303-324.

Polderman, T. J., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics*, *47*(7), 702-709.

Rauscher, E., Conley, D., & Siegal, M. L. (2015). Sibling genes as environment: Sibling dopamine genotypes and adolescent health support frequency dependent selection. *Social science research*, *54*, 209-220.

Reuter, M., & Montag, C. (2016). Genes and Human Decision-Making. In *Neuroeconomics* (pp. 67-83). Springer Berlin Heidelberg.

Robinson MR, Kleinman A, Graff M, Vinkhuyzen AAE, Couper D, Miller MB, et al. Genetic evidence of assortative mating in humans. Nat Hum Behav. 2017 Jan 9;1:0016.

Ryan, A. M. (2000). Peer groups as a context for the socialization of adolescents' motivation, engagement, and achievement in school. *Educational Psychologist*, *35*(2), 101-111.

Salvatore, J. E., Aliev, F., Bucholz, K., Agrawal, A., Hesselbrock, V., Hesselbrock, M., ... & Edenberg, H. J. (2015). Polygenic risk for externalizing disorders: Gene-by-development and gene-by-environment effects in adolescents and young adults. *Clinical Psychological Science*, *3*(2), 189-201.

Salvatore, J. E., Aliev, F., Edwards, A. C., Evans, D. M., Macleod, J., Hickman, M., ... & Latvala, A. (2014). Polygenic scores predict alcohol problems in an independent sample and show moderation by the environment. *Genes*, *5*(2), 330-346.

Schmitz, L., & Conley, D. (2016). The long-term consequences of Vietnam-era conscription and genotype on smoking behavior and health. *Behavior genetics*, *46*(1), 43-58.

Smeland, O. B., Wang, Y., Lo, M. T., Li, W., Frei, O., Witoelar, A., ... & Chen, C. H. (2017). Identification of genetic loci shared between schizophrenia and the Big Five personality traits. *Scientific Reports*, *7*.

Spielman, D., Brook, B. W., Briscoe, D. A., & Frankham, R. (2004). Does inbreeding and loss of genetic diversity decrease disease resistance?. *Conservation Genetics*, *5*(4), 439-448.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545-15550.

Tenconi, E., Degortes, D., Clementi, M., Collantoni, E., Pinato, C., Forzan, M., ... & Favaro, A. (2016). Clinical and genetic correlates of decision making in anorexia nervosa. *Journal of clinical and experimental neuropsychology*, *38*(3), 327-337.

Tobacco and Genetics Consortium. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature genetics*, *42*(5), 441-447.

True, W. R., Heath, A. C., Scherrer, J. F., Waterman, B., Goldberg, J., Lin, N., ... & Tsuang, M. T. (1997). Genetic and environmental contributions to smoking. *Addiction*, *92*(10), 1277-1287.

U.S. Department of Health and Human Services. Preventing Tobacco Use Among Youth and Young Adults: A Report of the Surgeon General. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2012.

U.S. Department of Health and Human Services. The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2014.

Visscher, P. M., Hill, W. G., & Wray, N. R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics*, *9*(4), 255-266.

Visscher, P. M., Medland, S. E., Ferreira, M. A., Morley, K. I., Zhu, G., Cornes, B. K., ... & Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS genetics*, *2*(3), e41.

Zerubavel, N., Bearman, P. S., Weber, J., & Ochsner, K. N. (2015). Neural mechanisms tracking popularity in real-world social networks. *Proceedings of the National Academy of Sciences*, *112*(49), 15072-15077.

Zimmerman, M. A., Stoddard, S. A., Eisman, A. B., Caldwell, C. H., Aiyer, S. M., & Miller, A. (2013). Adolescent resilience: Promotive factors that inform prevention. *Child development perspectives*, *7*(4), 215-220.

Zou JY, Park DS, Burchard EG, Torgerson DG, Pino-Yanes M, Song YS, et al. Genetic and socioeconomic study of mate choice in Latinos reveals novel assortment patterns. Proc Natl Acad Sci. 2015;112(44):13621–13626.

Appendix A: Descriptive statistics

Smoking Behavior

	Mean	Std. Deviation	Minimum	Maximum
Cigs Per Day – Full Sample	1.59	4.82	0.00	95.00
Cigs Per Day – 8th Graders	0.90	3.92	0.00	60.00
Cigs Per Day – 9th Graders	1.62	4.25	0.00	40.00
Cigs Per Day – 10th Graders	1.65	4.97	0.00	70.00
Cigs Per Day – 11th Graders	2.11	5.27	0.00	60.00
Cigs Per Day – 12th Graders	2.54	6.83	0.00	95.00

Individual Variables

	Mean	Std. Deviation	Minimum	Maximum
Smoking PGS	-0.00094	0.00029	-0.00172	-0.00028
Black	0.24	0.43	0.00	1.00
Hispanic	0.16	0.37	0.00	1.00
Maternal Education	5.28	2.24	0.00	9.00
Household Smoker	0.47	0.46	0.00	1.00
Male	0.48	0.50	0.00	1.00
Older Siblings	0.85	1.18	0.00	12.00
Family Income ($1000s)	44	39	0.00	999
Height	5.53	0.48	4.00	6.75
Height PGS	-0.00006	0.00006	-0.00019	0.00004

Classmate Variables

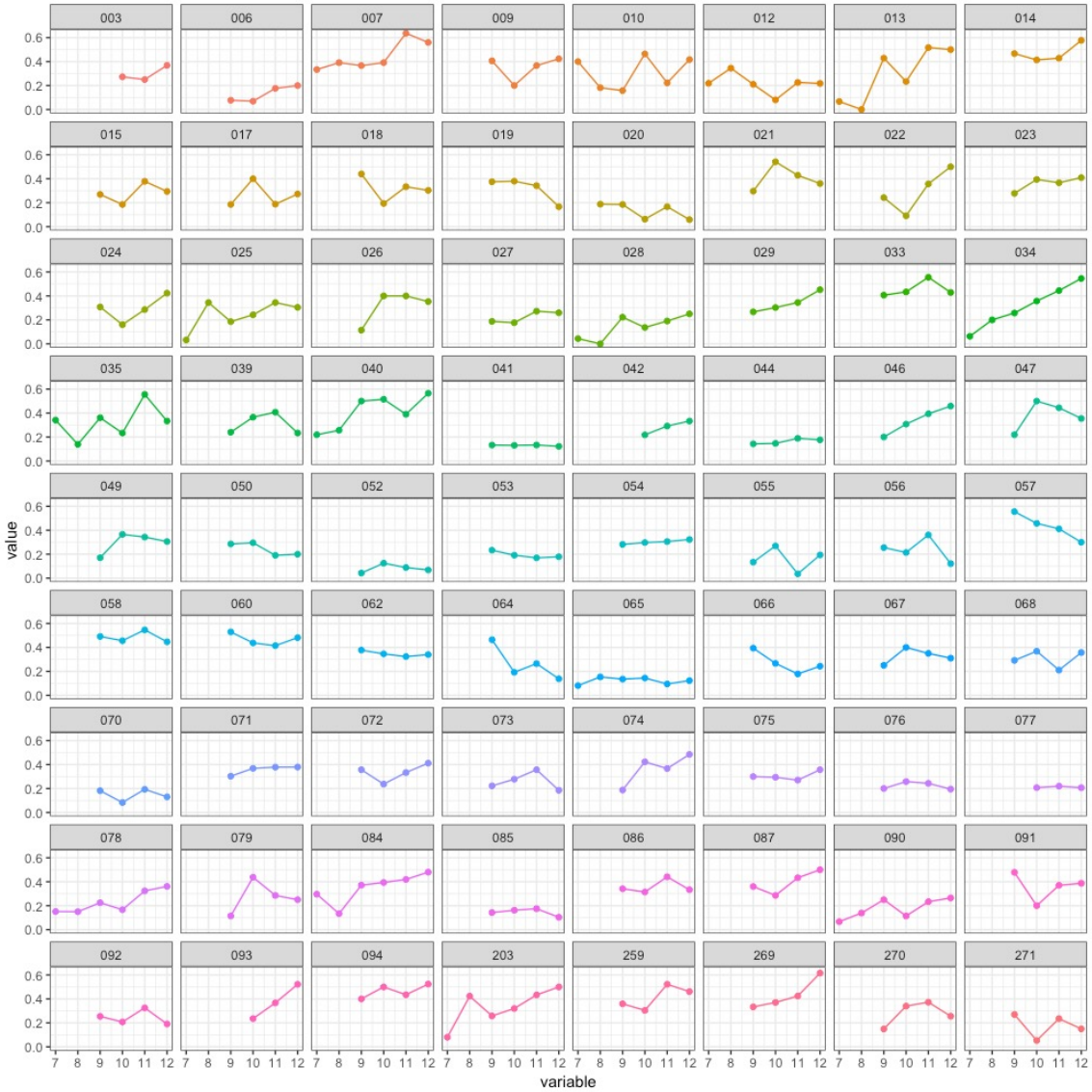
	Mean	Std. Deviation	Minimum	Maximum
Mean Cigs Per Day	1.59	1.37	0.00	7.27
Mean Smoking PGS	-0.0009	0.0002	-0.0014	-0.0007
% Family Smokes	0.47	0.12	0.13	0.86
% has Older Sibling	0.50	0.11	0.16	0.93
% Black	0.24	0.26	0.00	1.00
% Hispanic	0.16	0.23	0.00	0.94
Mean Maternal Education	5.28	0.67	3.41	7.83
% Male	0.48	0.08	0.15	0.72
Mean Family Income ($1000s)	44	10	19	93
Number of Classmates	61.71	66.29	20.00	231.00
Mean Height	5.53	0.14	5.01	6.00

Mean Height PGS	-0.00006	0.00004	-0.00015	0.00000

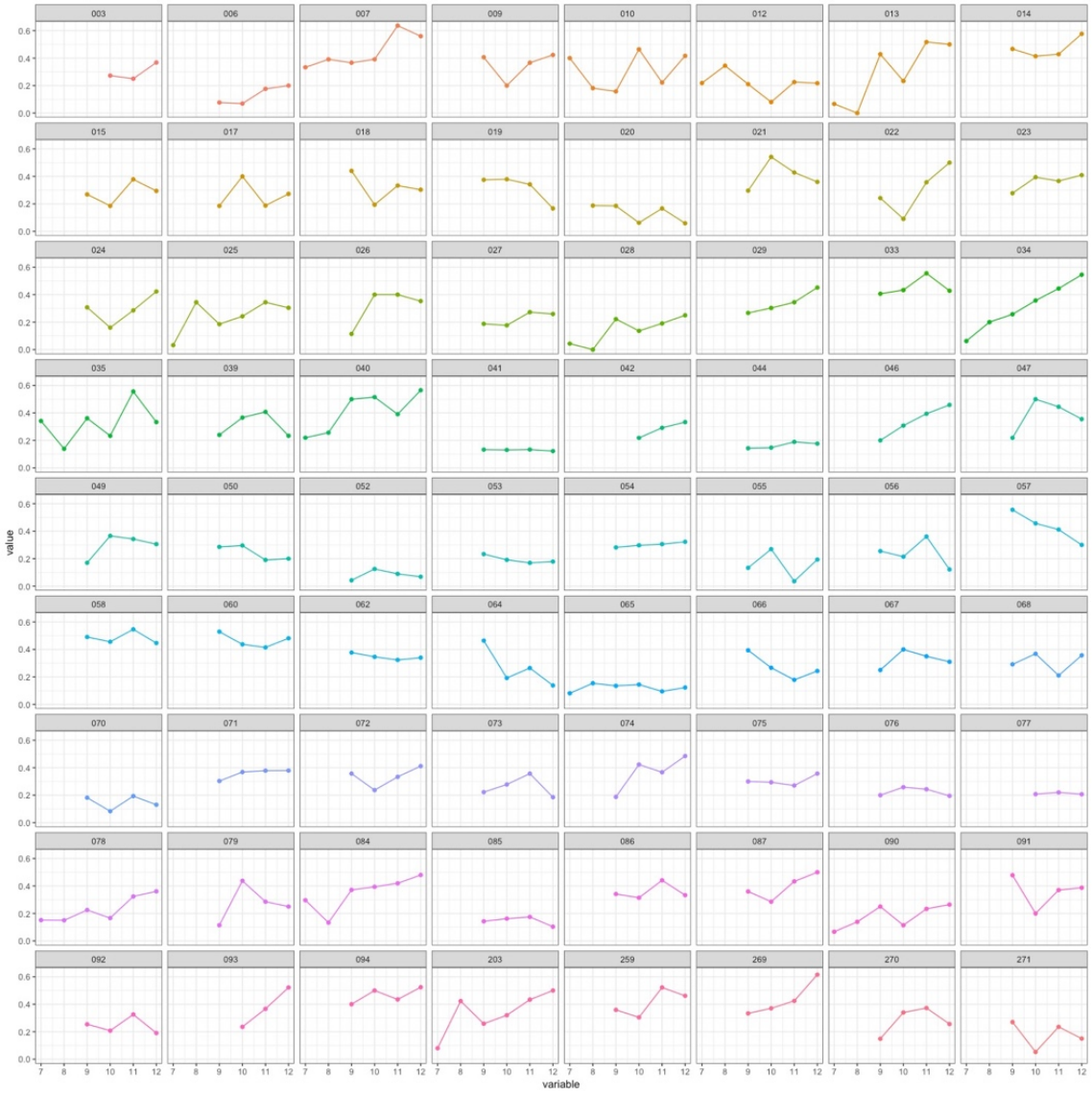
Cigs Per Day ~ Smoking PGS

	White	Black	Hispanic
Correlation	0.0287	0.0289	0.1035
Incremental R-Squared	0.00003	0.0001	0.0011

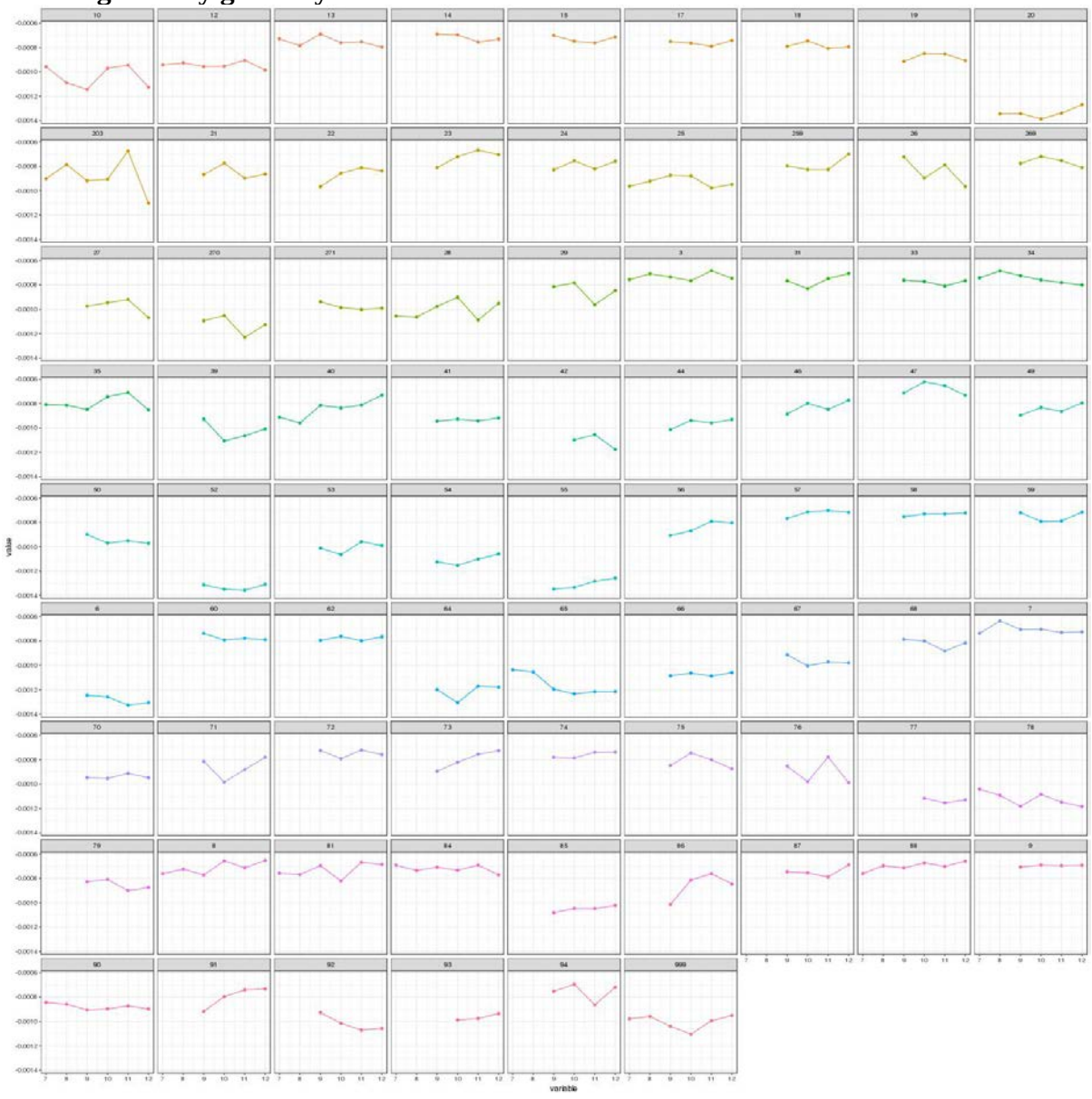
Smoking (cigs per day) by grade by school



Mean Smoking (cigs per day) by grade by school

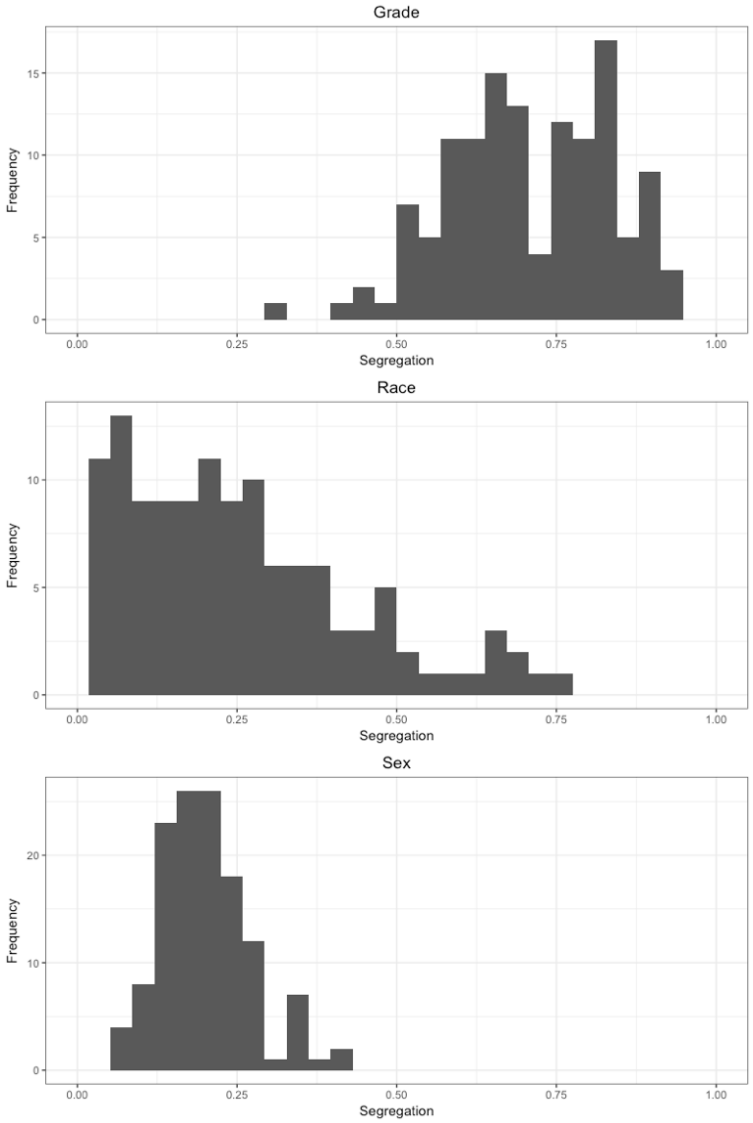


Smoking PGS by grade by school



Appendix B: Assortativity

Figure A1 - A3: Assortativity by Grade, Race and Sex



The histograms above show the distribution of segregation, a measure of the extent to which students have a friendship preference for others who share the same attribute, for a set of attributes known to shape friendships in high schools (Bearman et al. 2004; McFarland et al. 2014). Segregation is calculated using the equation below:

$$Segregation = \frac{Expected(Cross\ Trait\ Ties) - Observed(Cross\ Trait\ Ties)}{Expected(Cross\ Trait\ Ties)}$$

The value of segregation goes from -1 to 1, where -1 is a pure out group preference and 1 is pure in group preference. *Expected Cross Trait Ties* is the number of cross-trait ties we would expect to see in a random graph, where the relative group sizes (i.e. number of 8th, 9th, 10th, etc. graders) are held constant.

Grade shows the highest value for segregation; this shows a higher in-group preference for grades as compared to friendships based on sex or race.

Appendix C – Regression tables corresponding to figures in the text

Table C1: Effect of peer genes on individual outcome compared to individual attributes

	(1) Individual Genes	(2) Reduced Form	(3) Age and sex
	Cigs per Day	Cigs Per Day	Cigs Per Day
Smoking PGS	0.0522*	0.0607**	
	(0.0206)	(0.0209)	
Mean Smoking PGS		0.272*	
		(0.124)	
Black	-0.0868	-0.203***	-0.293***
	(0.0485)	(0.0513)	(0.0397)
Hispanic	-0.210***	-0.0879	-0.103*
	(0.0506)	(0.0488)	(0.0480)
Maternal Education	-0.00262	-0.00168	0.00664
	(0.0140)	(0.0141)	(0.0140)
Household Smoker	0.129***	0.131***	0.131***
	(0.0134)	(0.0137)	(0.0134)
Male	0.0724**	0.0698**	0.0552*
	(0.0260)	(0.0267)	(0.0259)
Older Siblings	0.0165	0.0176	0.0129
	(0.0129)	(0.0130)	(0.0129)
Age			0.0966***
			(0.0122)
Family Income	-0.0295*	-0.0286	-0.0270
	(0.0146)	(0.0149)	(0.0145)
% with Smoking Family Member		0.0300	
		(0.0399)	
% with Older Siblings		0.0241	
		(0.0260)	
% Black		0.0941	
		(0.127)	
% Hispanic		0.00648	
		(0.0862)	
Mean Maternal Education		0.0386	
		(0.0461)	
% Male		-0.0118	
		(0.0329)	

Mean Family Income		0.0230	
		(0.0484)	
Constant	0.159	0.178	-1.134***
	(0.207)	(0.224)	(0.265)
N	4909	4909	4907
R^2	0.097	0.099	0.102
adj. R^2	0.081	0.081	0.087
F	6.008	5.599	6.772

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
School and grade fixed effects are included in all models

Table C2: Instrumental Variable Regressions

	(1) 2SLS – First stage Mean Cigs Per Day	(2) 2SLS – Second stage Cigs Per Day	(4) 2SLS – Second Stage Height	(5) Reduced Form Height
Mean Cigs Per Day		0.290*		
		(0.133)		
Smoking PGS	0.0438***	0.0482*		
	(0.00788)	(0.0208)		
Mean Smoking PGS	0.944***			
	(0.0468)			
Mean Height			0.00749	
			(0.305)	
Mean Height PGS				0.00369
				(0.152)
Height PGS			0.193***	0.194***
			(0.0236)	(0.0239)
Black	0.0237	-0.209***	0.417***	0.417***
	(0.0193)	(0.0515)	(0.0546)	(0.0550)
Hispanic	-0.00908	-0.0851	0.0364	0.0360
	(0.0184)	(0.0491)	(0.0514)	(0.0503)
Maternal Education	0.00957	-0.00446	-0.0138	-0.0138
	(0.00533)	(0.0143)	(0.0145)	(0.0145)
Household Smoker	0.00720	0.129***		
	(0.00516)	(0.0138)		
Male	-0.00954	0.0727**	0.822***	0.822***
	(0.0101)	(0.0270)	(0.0276)	(0.0276)
Older Siblings	0.00806	0.0152	0.00163	0.00170
	(0.00486)	(0.0131)	(0.0135)	(0.0133)
Family Income	0.00536	-0.0301*	0.0158	0.0157
	(0.00561)	(0.0150)	(0.0156)	(0.0153)
% with Smoking Family	0.198***	-0.0275		

Member				
	(0.0150)	(0.0454)		
% with Older Siblings	0.114***	-0.00908	0.0313	0.0319
	(0.00981)	(0.0287)	(0.0353)	(0.0267)
% Black	0.0859	0.0709	-0.0524	-0.0498
	(0.0479)	(0.120)	(0.0908)	(0.149)
% Hispanic	-0.0922**	0.0326	-0.146	-0.148
	(0.0325)	(0.0880)	(0.130)	(0.0885)
Mean Maternal Education	0.0888***	0.0133	-0.0218	-0.0229
	(0.0174)	(0.0495)	(0.0541)	(0.0483)
% Male	-0.0208	-0.00547	0.0485	0.0511
	(0.0124)	(0.0337)	(0.105)	(0.0345)
Mean Family Income	0.0204	0.0171	-0.0637	-0.0645
	(0.0183)	(0.0487)	(0.0594)	(0.0500)
Constant	0.476***	0.00787	-0.990*	-0.105
	(0.0846)	(0.120)	(0.475)	(0.229)
N	4938	4909	4895	4895
R^2	0.833	0.068	0.233	0.234
adj. R^2	0.829	0.050	0.219	0.220
F	256.2			15.99

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
School and grade fixed effects are included in all models

Table C3: Gene by Environment Interaction models

	(1) Base Model	(2) Interaction Model	(3) Low Smoking PGS	(4) Medium Smoking PGS	(5) High Smoking PGS
	Cigs Per Day	Cigs Per Day	Cigs Per Day	Cigs Per Day	Cigs Per Day
Mean Cigs Per Day	0.272*	0.270*	-0.212	0.414	0.489
	(0.124)	(0.124)	(0.134)	(0.251)	(0.263)
Smoking PGS	0.0607**	0.0582**	0.0239	-0.0666	-0.0217
	(0.0209)	(0.0216)	(0.0289)	(0.0999)	(0.0882)
Smoking PGS x Mean Smoking PGS		-0.00922			
		(0.0198)			
Hispanic	-0.0879	-0.202***	-0.102*	-0.0566	-0.119
	(0.0488)	(0.0514)	(0.0491)	(0.0904)	(0.163)
Black	-0.203***	-0.0912	-0.139***	-0.253	-0.473
	(0.0513)	(0.0493)	(0.0407)	(0.145)	(0.626)
Maternal Education	-0.00168	-0.00143	-0.00251	0.0275	-0.0354
	(0.0141)	(0.0141)	(0.0131)	(0.0294)	(0.0307)
Household Smoker	0.131***	0.131***	0.0186	0.150***	0.199***
	(0.0137)	(0.0137)	(0.0136)	(0.0278)	(0.0281)

Male	0.0698**	0.0697**	0.0851***	0.0755	0.0291
	(0.0267)	(0.0267)	(0.0254)	(0.0538)	(0.0562)
Older Siblings	0.0176	0.0176	0.0150	0.00276	0.0364
	(0.0130)	(0.0130)	(0.0106)	(0.0289)	(0.0306)
Family Income	-0.0286	-0.0287	-0.000242	-0.0393	-0.0424
	(0.0149)	(0.0149)	(0.0155)	(0.0353)	(0.0263)
% Hispanic	0.00648	0.00667	0.0708	-0.167	0.115
	(0.0862)	(0.0862)	(0.0812)	(0.176)	(0.190)
% Black	0.0941	0.0947	-0.146	0.385	0.0159
	(0.127)	(0.127)	(0.134)	(0.274)	(0.250)
% Older Siblings	0.0241	0.0239	-0.0128	0.0431	0.00969
	(0.0260)	(0.0260)	(0.0272)	(0.0548)	(0.0545)
% with Smoking Family Member	0.0300	0.0305	-0.00410	0.0720	0.0149
	(0.0399)	(0.0400)	(0.0415)	(0.0839)	(0.0841)
Mean Maternal Education	0.0386	0.0393	0.0394	0.111	-0.0606
	(0.0461)	(0.0461)	(0.0459)	(0.0960)	(0.0955)
% Male	-0.0118	-0.0118	0.00522	0.0441	-0.117
	(0.0329)	(0.0329)	(0.0364)	(0.0693)	(0.0659)
Mean Family Income	0.0230	0.0231	0.0276	-0.0217	0.0221
	(0.0484)	(0.0484)	(0.0580)	(0.0962)	(0.0946)
Constant	0.178	0.176	-0.0286	-0.604	0.679
	(0.224)	(0.224)	(0.204)	(0.774)	(0.410)
N	4909	4909	1635	1634	1641
R^2	0.099	0.099	0.066	0.095	0.119
adj. R^2	0.081	0.081	0.016	0.039	0.067
F	5.599	5.541	1.313	1.710	2.300

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
School and grade fixed effects included in all models

Appendix D – Alternative Functional Forms and model specifications

As a robustness check, we performed the same analyses with different model specifications and functional forms. First, we tried Instrumental Variable Poisson regression, since the outcome variable of cigarettes per day is count data. The results are substantively similar to what obtains under traditional two stage least squares. Next, we tested what we refer to as "logit-in-means" estimates of peer smoking as opposed to the "linear-in-means" (Manski, 1993). For the logit-in-means estimate, we calculated the $log(\frac{p}{1-p})$ for every individual where p is the proportion of smokers in a grade excluding the individual. We then predicted cigarettes per day smoked by an individual in a grade. The idea here is that the effect of others' smoking behavior might take the form of a sigmoid function rather than a linear one. Having a few individuals in a grade who smoke may not be influential on the overall smoking behavior of the grade; on the other hand, an

additional smoker in a grade will not matter as much if the majority of the grade is already smoking. We find that the two-stage least squared model using logit-in-means yields very similar results to the 2SLS using linear-in-means estimates. Overall, the alternative functional form specifications all yield substantively similar results, signals the robustness of our findings.

Table D1: Alternative Functional Forms as Robustness Checks

	(1) IV Poisson Cigs Per Day	(2) 2SLS with Logistic-in-means Cigs Per Day
Mean Cigs per day	0.462*	
	(0.202)	
Logistic function of Classroom Smokers		0.120*
		(0.0565)
Smoking PGS	529.9*	0.0523*
	(259.5)	(0.0212)
Black	-1.290***	-0.249***
	(0.205)	(0.0549)
Hispanic	-0.222	-0.105*
	(0.240)	(0.0511)
Maternal Education	-0.00756	0.00678
	(0.0257)	(0.0150)
Household Smoker	0.928***	0.120***
	(0.103)	(0.0148)
Male	0.219*	0.0764**
	(0.0865)	(0.0278)
Older Siblings	0.0458	0.0173
	(0.0355)	(0.0133)
Family Income	-0.392**	-0.0308*
	(0.142)	(0.0153)
% with Older Siblings	-0.0479	0.00713
	(0.736)	(0.0270)
% with Household Smoker	-0.628	-0.0818
	(1.198)	(0.0626)
% Black	0.328	-0.202
	(1.141)	(0.130)
% Hispanic	0.447	-0.302*
	(1.230)	(0.131)
Mean Maternal Education	0.0114	0.130*
	(0.177)	(0.0589)
% Male	-0.281	0.00328
	(0.853)	(0.0356)
Mean Income	-0.700	0.00816
	(0.964)	(0.0500)
Constant	0.125	0.869*

	(1.507)	(0.425)
N	4909	4909
R^2		0.026
adj. R^2		0.007

Standard errors in parentheses

$^{\sim}$ p <0.1, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

School and grade fixed effects are included in all models

Figures and plots

Figure 1: Effect of peer genes on individual outcome compared to individual attributes

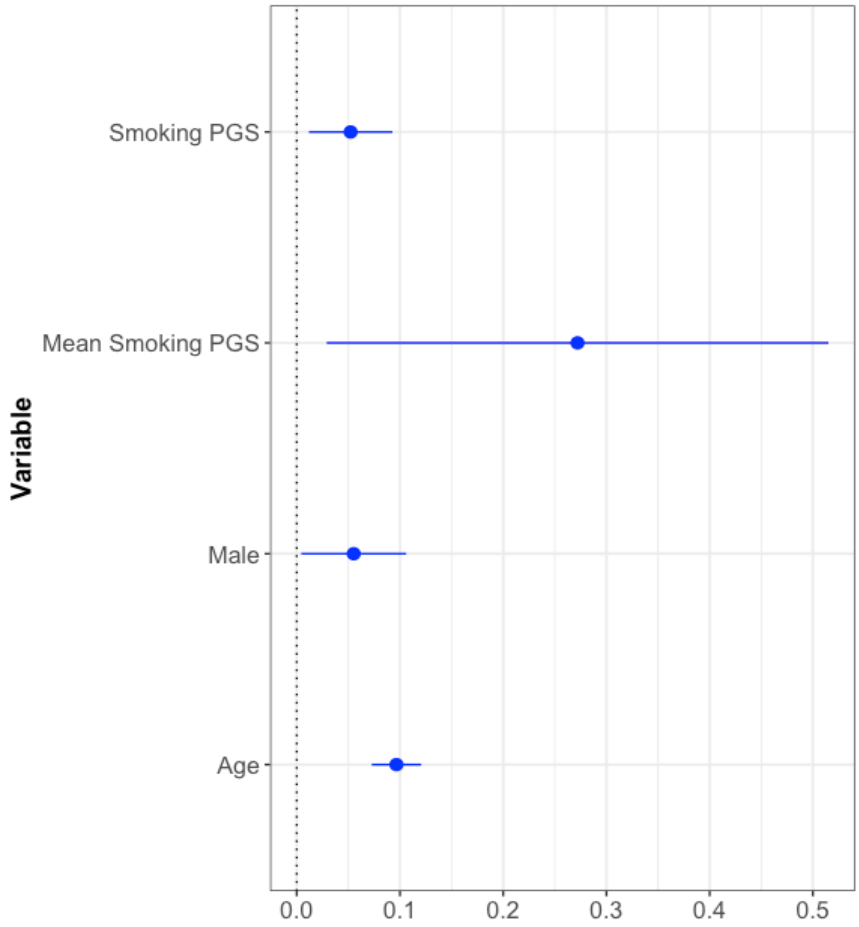


Figure 2: Effect of Peer Smoking and Peer Height on Ego Smoking and Ego Height

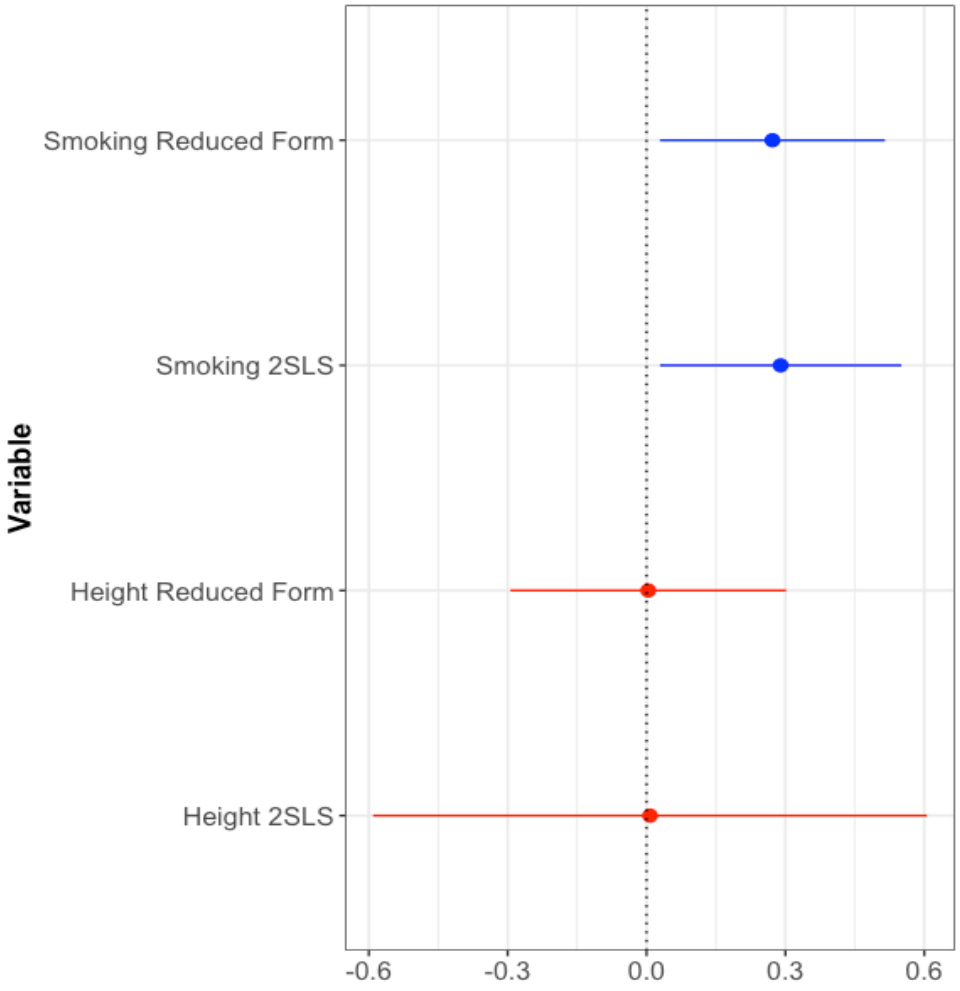


Figure 3: Effect of Peer Genes on Individual's Smoking Outcome by Polygenic Score Level

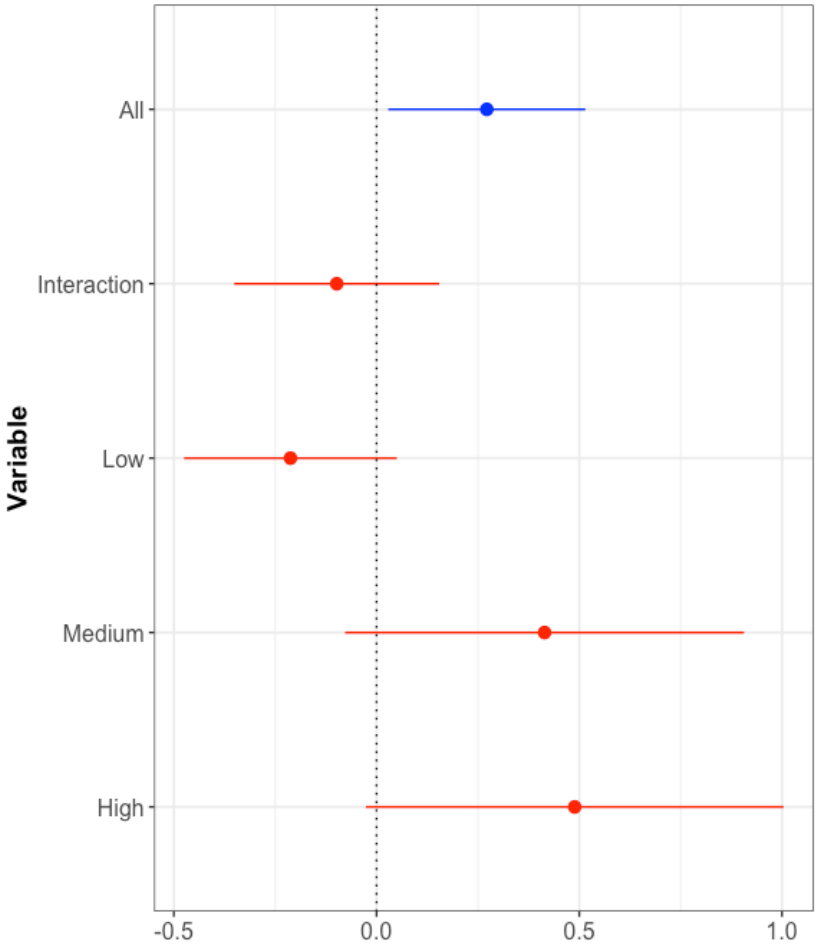


Figure 4: Effect of Peer Smoking on Ego Smoking Behavior by Race of Ego

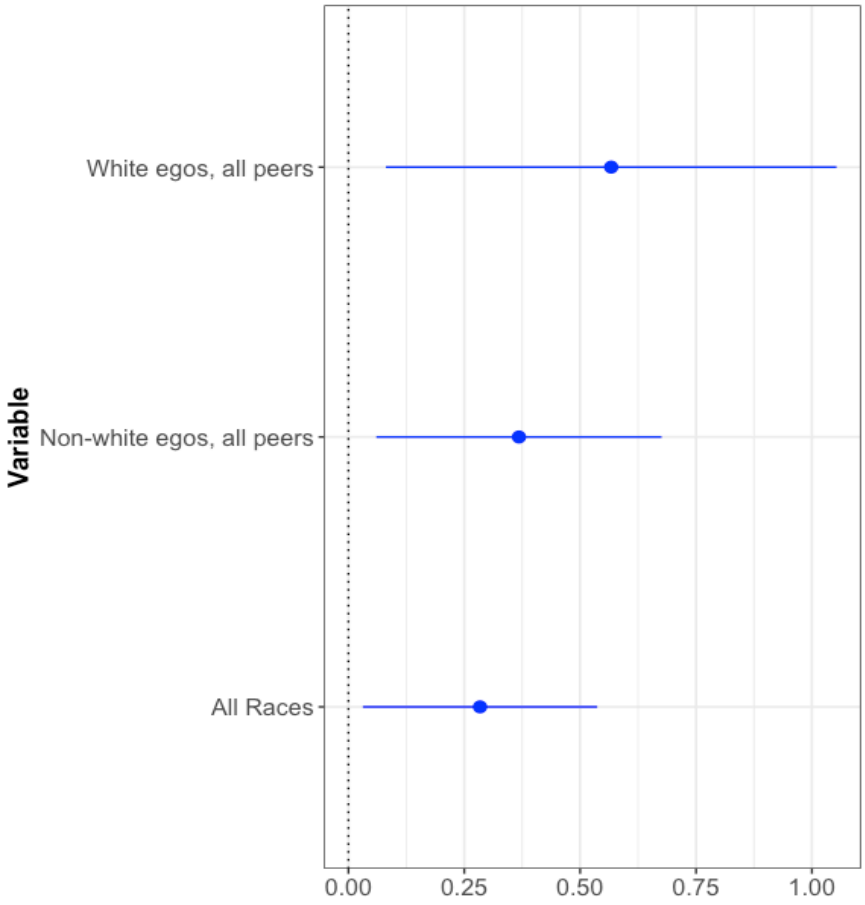


Figure 5: Effect of White Peer Smoking on Ego Smoking Behavior by Race of Ego