Supplementary Information
Metagenomic effects in human behavior: The case of adolescent smoking

Ramina Sotoudeh, Kathleen Mullan Harris, Dalton Conley

Correspondence may be addressed to sotoudeh@princeton.edu, kathie_harris@unc.edu, or dconley@princeton.edu

www.pnas.org/cgi/doi/10.1073/pnas.1806901116

Table S1: Descriptive statistics

	Individual level	N	Grade level	N	Classroom level	N	Friend level	N	In-degree friend level	N	Out-degree friend level	N
Cigarettes Smoked Per Day (mean)	1.7275	9694	1.731	9666	1.7936	3505	1.5683	4692	1.4429	3506	1.6915	3293
	(5.0975)		(1.6667)		(3.3139)		(3.7709)		(3.7351)		(3.9124)	
Smoking PGS (mean)	0	9743	0.0006	9666	0.0419	3511	0.0392	4703	0.0833	3515	0.0678	3303
	(1)		(0.6318)		(0.7721)		(0.9251)		(0.9343)		(0.9283)	
Bad Apple (proportion)	0.1022	9735	0.1024	9666	0.104	3510	0.1042	4702	0.1091	3514	0.1024	3302
	(0.3029)		(0.0782)		(0.2012)		(0.2454)		(0.2654)		(0.2511)	
Shining Star (proportion)	0.1032	9740	0.1027	9666	0.1034	3511	0.1008	4703	0.0956	3515	0.1009	3303
	(0.3042)		(0.0781)		(0.1999)		(0.2478)		(0.2531)		(0.258)	
Male (proportion)	0.4758	9743	0.4756	9666	0.4711	3511	0.4603	4703	0.4481	3515	0.4568	3303
	(0.4994)		(0.1266)		(0.3345)		(0.4085)		(0.4353)		(0.4189)	
White (proportion)	0.599	9729	0.5989	9666	0.6198	3510	0.6309	4701	0.6498	3513	0.6525	3301
	(0.4901)		(0.3532)		(0.4205)		(0.4625)		(0.4631)		(0.4624)	
Black (proportion)	0.212	9740	0.2118	9666	0.1723	3511	0.1984	4703	0.1832	3515	0.1884	3303
	(0.4088)		(0.2658)		(0.2967)		(0.3843)		(0.3762)		(0.3824)	
Hispanic (proportion)	0.1451	9729	0.1451	9666	0.1476	3510	0.1206	4701	0.116	3513	0.1119	3301
	(0.3523)		(0.2222)		(0.2829)		(0.3051)		(0.3059)		(0.3001)	
Other (proportion)	0.0804	9737	0.0809	9666	0.0949	3511	0.0928	4703	0.0888	3515	0.0903	3303
	(0.2719)		(0.1219)		(0.2024)		(0.264)		(0.2646)		(0.2666)	
Family Income (mean)	0.462	9743	0.4621	9666	0.4849	3511	0.4639	4703	0.4653	3515	0.47	3303
	(0.4257)		(0.1743)		(0.3222)		(0.3563)		(0.3528)		(0.3576)	
Maternal Education (mean)	5.4146	9743	5.4149	9666	5.5447	3511	5.4452	4703	5.4541	3515	5.4698	3303
	(2.2336)		(0.8979)		(1.5595)		(1.8911)		(1.97)		(1.9184)	
Older Sibling (proportion)	0.4916	9728	0.4916	9666	0.5058	3509	0.508	4702	0.5086	3515	0.5151	3301
	(0.5)		(0.1328)		(0.3307)		(0.406)		(0.4308)		(0.4204)	
Household Smoker (proportion)	0.4624	9743	0.4621	9666	0.4397	3511	0.4507	4703	0.4469	3515	0.454	3303
	(0.4659)		(0.1504)		(0.3075)		(0.3874)		(0.406)		(0.3985)	
PC 1 (mean)	0	9743	-0.0005	9666	-0.0916	3511	-0.0484	4703	-0.0881	3515	-0.077	3303
	(1)		(0.6829)		(0.7529)		(0.9346)		(0.9233)		(0.9308)	
PC 2 (mean)	0	9743	0.0008	9666	0.0596	3511	0.0088	4703	0.0043	3515	-0.0193	3303
	(1)		(0.6256)		(0.8845)		(1.0092)		(1.0204)		(0.9895)	

Notes: Standard deviations in parentheses

Results as visualized in the main text

The tables below correspond to the figures reported in the main text of the article in the order that they appear in the text.

Table S2: Metagenomic effects of smoking (Figures 1 and 2)

	(1) Grade-mates	(2) Classmates	(3) Friends	(4) In-degree	(5) Out-degree
Mean Smoking PGS (residualized)	0.0525*	-0.00902	0.0316+	0.0393*	0.0231
	(0.0243)	(0.0168)	(0.0177)	(0.0199)	(0.0254)
Controls					
Smoking PGS (residualized)	0.0193+	0.00792	0.0116	-0.0000797	0.0322*
	(0.0115)	(0.0149)	(0.0144)	(0.0175)	(0.0157)
Black	-0.118***	-0.102***	-0.0492*	-0.0464+	-0.0334
	(0.0162)	(0.0157)	(0.0224)	(0.0252)	(0.0257)
Hispanic	-0.0324	-0.0279	0.0101	-0.0135	0.0207
	(0.0314)	(0.0215)	(0.0368)	(0.0340)	(0.0401)
Other Race	-0.0116	-0.00798	-0.000590	0.0153	0.00244
	(0.0160)	(0.0213)	(0.0219)	(0.0244)	(0.0237)
Household Smoker	0.134***	0.166***	0.123***	0.119***	0.111***
	(0.0164)	(0.0169)	(0.0170)	(0.0179)	(0.0200)
Older Sibling	0.0242	0.0166	0.0114	0.0123	0.0117
	(0.0159)	(0.0190)	(0.0149)	(0.0171)	(0.0182)
Maternal Education	-0.00883	-0.0333+	0.000918	-0.0232	-0.00277
	(0.0261)	(0.0180)	(0.0229)	(0.0218)	(0.0263)
Male	0.0205	-0.0000544	0.0215	0.00923	0.0195
	(0.0150)	(0.0162)	(0.0163)	(0.0169)	(0.0206)
Family Income	-0.0231*	-0.0200	0.00355	0.0193	-0.00638
	(0.00997)	(0.0175)	(0.0159)	(0.0198)	(0.0109)
Proportion Black	-0.0882	0.0362	-0.0571*	-0.0719**	-0.0542+
	(0.0598)	(0.0237)	(0.0253)	(0.0250)	(0.0308)
Proportion Hispanic	-0.0572	0.0105	-0.0733*	-0.0732*	-0.0663*
	(0.0749)	(0.0231)	(0.0314)	(0.0289)	(0.0330)
Proportion Other Race	0.0127	0.00543	0.0120	-0.0266+	0.0244
	(0.0522)	(0.0133)	(0.0184)	(0.0151)	(0.0235)
Proportion Household Smoker	0.00292	0.0276+	0.0684***	0.0879***	0.0477**
	(0.0338)	(0.0167)	(0.0142)	(0.0202)	(0.0161)
Proportion Older Sibling	0.0251	-0.0225	0.0140	0.00955	0.0220
	(0.0243)	(0.0155)	(0.0149)	(0.0165)	(0.0166)
Mean Maternal Education	0.117*	-0.0660***	-0.0511*	-0.0286	-0.0521*
	(0.0464)	(0.0181)	(0.0203)	(0.0175)	(0.0257)
Proportion Male	-0.0397	-0.00346	-0.0129	-0.0192	-0.000866
	(0.0285)	(0.0170)	(0.0167)	(0.0221)	(0.0159)
Mean Family Income	0.0223	0.0253	-0.0354**	-0.0502***	-0.0232+
	(0.0593)	(0.0243)	(0.0136)	(0.0147)	(0.0129)
Constant	-0.134+	-0.443***	0.0691	-0.471***	0.336
	(0.0689)	(0.0571)	(0.415)	(0.109)	(0.540)
N	3853	2820	3709	2743	2609
R^2	0.108	0.128	0.117	0.139	0.109
adj. R^2	0.086	0.096	0.081	0.091	0.057

Notes: Standard errors in parentheses

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S3: The effects of bad apples and shining stars on smoking outcomes (Figure 3)

	(1) Grade-Mates	(2) Grade-Mates	(3) Classmates	(4) Classmates	(5) Friends	(6) Friends	(7) In-degree	(8) In-degree	(9) Out-degree	(10) Out-degree
Proportion Bad Apples	0.0584**		-0.00697		0.0181		0.0296		0.0135	
	(0.0210)		(0.0182)		(0.0193)		(0.0249)		(0.0211)	
Proportion Shining Stars		-0.0149		0.00326		-0.0196		-0.0174		-0.00876
		(0.0218)		(0.0153)		(0.0130)		(0.0143)		(0.0254)
Controls										
Smoking PGS (residualized)	0.0181	0.0162	0.00847	0.00816	0.0107	0.0110	-0.00112	0.000253	0.0316*	0.0315*
	(0.0115)	(0.0116)	(0.0146)	(0.0149)	(0.0143)	(0.0143)	(0.0173)	(0.0176)	(0.0156)	(0.0155)
Black	-0.117***	-0.117***	-0.102***	-0.102***	-0.0488*	-0.0504*	-0.0445+	-0.0468+	-0.0331	-0.0346
	(0.0162)	(0.0161)	(0.0157)	(0.0156)	(0.0224)	(0.0221)	(0.0255)	(0.0251)	(0.0255)	(0.0257)
Hispanic	-0.0321	-0.0308	-0.0280	-0.0279	0.00932	0.00983	-0.0143	-0.0140	0.0204	0.0203
	(0.0312)	(0.0314)	(0.0215)	(0.0216)	(0.0365)	(0.0365)	(0.0339)	(0.0341)	(0.0399)	(0.0400)
Other Race	-0.0110	-0.0111	-0.00797	-0.00795	-0.00181	-0.000963	0.0134	0.0143	0.00225	0.00221
	(0.0160)	(0.0160)	(0.0213)	(0.0215)	(0.0219)	(0.0219)	(0.0240)	(0.0242)	(0.0238)	(0.0238)
Household Smoker	0.135***	0.135***	0.166***	0.166***	0.124***	0.123***	0.119***	0.119***	0.111***	0.110***
	(0.0164)	(0.0164)	(0.0168)	(0.0169)	(0.0171)	(0.0171)	(0.0179)	(0.0179)	(0.0198)	(0.0198)
Older Sibling	0.0247	0.0239	0.0167	0.0167	0.0116	0.0111	0.0127	0.0122	0.0116	0.0114
	(0.0159)	(0.0159)	(0.0190)	(0.0190)	(0.0150)	(0.0149)	(0.0171)	(0.0171)	(0.0184)	(0.0183)
Maternal Education	-0.00923	-0.00863	-0.0334+	-0.0334+	-0.00009	0.000337	-0.0240	-0.0246	-0.00351	-0.00305
	(0.0262)	(0.0262)	(0.0179)	(0.0180)	(0.0229)	(0.0228)	(0.0218)	(0.0216)	(0.0263)	(0.0262)
Male	0.0235	0.0203	-0.0000938	-0.000209	0.0220	0.0212	0.00933	0.00923	0.0193	0.0192
	(0.0149)	(0.0151)	(0.0162)	(0.0161)	(0.0165)	(0.0163)	(0.0169)	(0.0169)	(0.0206)	(0.0207)
Family Income	-0.0239*	-0.0232*	-0.0201	-0.0200	0.00395	0.00360	0.0194	0.0190	-0.00617	-0.00654
	(0.0100)	(0.00996)	(0.0175)	(0.0175)	(0.0159)	(0.0161)	(0.0198)	(0.0200)	(0.0109)	(0.0109)
Proportion Black	-0.0595	-0.0601	0.0362	0.0360	-0.0574*	-0.0555*	-0.0737**	-0.0712**	-0.0548+	-0.0526+
	(0.0525)	(0.0581)	(0.0237)	(0.0237)	(0.0254)	(0.0250)	(0.0258)	(0.0249)	(0.0312)	(0.0311)
Proportion Hispanic	-0.0572	-0.0282	0.0101	0.0105	-0.0743*	-0.0735*	-0.0745*	-0.0733*	-0.0669*	-0.0657*
	(0.0720)	(0.0730)	(0.0233)	(0.0232)	(0.0317)	(0.0314)	(0.0294)	(0.0291)	(0.0331)	(0.0329)
Proportion Other Race	0.0234	0.0207	0.00538	0.00512	0.0128	0.0135	-0.0263+	-0.0245+	0.0253	0.0256
	(0.0503)	(0.0523)	(0.0133)	(0.0132)	(0.0186)	(0.0183)	(0.0151)	(0.0146)	(0.0240)	(0.0238)
Proportion Household Smoker	0.00762	0.0122	0.0279+	0.0277+	0.0689***	0.0689***	0.0887***	0.0894***	0.0481**	0.0489**
	(0.0336)	(0.0349)	(0.0168)	(0.0168)	(0.0143)	(0.0143)	(0.0205)	(0.0201)	(0.0162)	(0.0166)
Proportion Older Sibling	0.0290	0.0216	-0.0225	-0.0225	0.0151	0.0146	0.0105	0.00998	0.0227	0.0224
	(0.0247)	(0.0248)	(0.0155)	(0.0156)	(0.0147)	(0.0148)	(0.0164)	(0.0165)	(0.0166)	(0.0167)
Mean Maternal Education	0.111*	0.119*	-0.0657***	-0.0658***	-0.0508*	-0.0512*	-0.0292+	-0.0295+	-0.0518*	-0.0515*
	(0.0486)	(0.0497)	(0.0181)	(0.0181)	(0.0202)	(0.0202)	(0.0175)	(0.0175)	(0.0255)	(0.0254)
Mean Male	-0.0212	-0.0412	-0.00410	-0.00412	-0.0135	-0.0122	-0.0199	-0.0191	-0.00101	-0.000319
	(0.0293)	(0.0291)	(0.0167)	(0.0167)	(0.0167)	(0.0167)	(0.0223)	(0.0222)	(0.0159)	(0.0157)
Mean Family Income	0.0104	0.0190	0.0256	0.0254	-0.0365**	-0.0361**	-0.0499***	-0.0509***	-0.0244*	-0.0244+
	(0.0601)	(0.0622)	(0.0244)	(0.0240)	(0.0132)	(0.0138)	(0.0144)	(0.0147)	(0.0124)	(0.0129)
Constant	-0.124+	-0.0984	-0.447***	-0.440***	0.0404	0.0771	-0.512***	-0.531***	0.315	0.363
	(0.0666)	(0.0676)	(0.0595)	(0.0581)	(0.421)	(0.427)	(0.108)	(0.111)	(0.543)	(0.540)
N	3853	3853	2820	2820	3709	3709	2743	2743	2609	2609
R^2	0.109	0.108	0.128	0.128	0.116	0.116	0.139	0.138	0.109	0.109
adj. R^2	0.086	0.085	0.096	0.096	0.080	0.080	0.091	0.090	0.057	0.056

Notes: Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Genes as an instrument for estimating peer effects

Having shown a robust association between peer genotypes and individual smoking outcomes, we can utilize peer genotypes to answer age-old questions of peer effects. The study of peer effects, that is, the causal effect of peers on individuals' behaviors, beliefs or outcomes, suffers from a set of difficulties that thwart easy estimation: 1) the reflection problem, where it is unclear whether peers are influencing the individual or vice versa; 2) contextual bias, where a common factor may be leading to the observed similarity between two peers, 3) the selection problem (or homophily), where individuals with similar attributes or genotypes tend to become friends, and 4) exclusion bias, or the fact that one cannot be friends with oneself, which constrains peer mixing patterns and induces mechanical negative correlation between individual and peer behavior. This has led to the development of a range of models designed to be robust to these biases, many of which instrument peer behavior using an exogenous factor to ensure an identified estimate.[i,ii] Questions remain however about how exogenous these instruments are with respect to peer behavior.

Using two-stage least squares model we are able to quantify the effect of peer polygenic score on individual smoking behavior as it operates through peer behavior. Here, the distribution of smoking PGS in a grade provides a good opportunity for the estimation of the effect of peer smoking on individual smoking. Conditional on school-level variation, the distribution of genetic propensity to smoke can be thought of as "as-if-random" in a given grade given that genes are assigned at birth and being in a given grade is determined, more or less, by birth date.

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

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

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 additive error term is ξ.

Equation 2 portrays the second stage regression of our model:

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

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 arguably causally identified, it also reveals a substantively interesting quantity. It provides a point estimate of the impact of one's peers genetic risk of smoking on an individual's smoking as it operates through peer behavior. The results can be found in Table S4 and they show a positive and significant relationship between peer and individual smoking, with a point estimate similar to that of the metagenomic effects models presented in Table S2, signaling that the effect of peer genes operates through behavior.

Like any other instrumental variable analysis, we make a series of assumptions, including that the instrument is sufficiently strong in the first state, monotonicity (that there are not individuals for whom having peers that smoke more make them smoke less), and that the exclusion restriction assumption is not violated (peer PGS does not affect ego smoking other than through the smoking behavior of peers). We undertook a series of analyses to test whether these assumptions are reasonably true. Though we cannot completely rule out the possibility of unaccounted pleiotropy (i.e. an exclusion restriction violation), we rule out obvious hypotheses regarding some potential pleiotropic effects of smoking risk (see Pleiotropy and Network Structure section in this SI Appendix).

A series of tests were also undertaken to ensure the strength and robustness of the instrument. A Lagrange-Multiplier (LM) test for under-identification using the Kleibergen and Paap (2006) rk statistic was used to test that the instrument was not under-identified, while an F statistic was used to test that the instrument was not weak. Because we are using clustered standard errors, we cannot use the standard F-statistic of the first stage as a test of weak or strong instrument. The null was rejected for the under-identification test. A traditional Cragg-Donald Wald test returns an F-statistic of 217.5, while a Kleibergen-Paap rk Wald F statistic, which is more robust when i.i.d is not met, returns an F-statistic of 6.7. This discrepancy signals that the instrument may be weak, so the results should be interpreted cautiously.

On the other hand, individual smoking PGS is positively and significantly associated with mean cigarettes per day in the first stage of the 2SLS. This association could signal either residual genetic clustering or reflection bias. We would expect genetic clustering to be the cause if the association is due to the correlation between individual PGS and grade-level smoking PGS. This is true generally at the school level, which is to say, genetic similarity is higher within schools than between them. However, as we demonstrate later in the SI appendix, conditional on being in the same school, students appear to be assigned randomly to grades with respect to their genes. Since our identification depends on comparing students in grades in the same school, we would not expect the identification of peer effects (i.e. the second stage of the 2SLS), nor the social genetic effects, to be confounded by population stratification.

That individual smoking PGS predicts grade-level smoking behavior could also result from reflection bias. Even though we remove a given individual in the calculation of the mean smoking PGS, we cannot completely account for the possibility that the focal individual might have, in a previous time point, influenced the smoking behavior of his or her grade-mates. However, because our instrument is mean smoking PGS which cannot be impacted by ego's PGS (given an individual's genes cannot affect the genes of others), this should not be a concern when considering the final stage of the 2SLS, which is driven by the variation in mean smoking behavior due to mean smoking PGS. Nonetheless, because of the stronger assumptions inherent to 2SLS, our preferred specification is the reduced form model presented in the main text.

Table S4: Metagenomic effects as instrumental variable

	(1) 2SLS – First Stage Mean Cigarettes Per Day	(2) 2SLS – Second Stage Cigarettes Per Day
Mean grade-mate Smoking PGS (residualized)	0.164*	
	(0.0633)	
Mean grade-mate Cigarettes Per Day		0.322***
		(0.0598)
Individual Controls:		
Smoking PGS (residualized)	0.0169*	0.0141
	(0.00811)	(0.0119)
Black	-0.0182*	-0.112***
	(0.00883)	(0.0159)
Hispanic	-0.00896	-0.0294
	(0.0102)	(0.0306)
Other Race	-0.00651	-0.00958
	(0.00849)	(0.0160)
Household Smoker	-0.000978	0.135***
	(0.00964)	(0.0164)
Older Sibling	0.0107	0.0206
	(0.00773)	(0.0158)
Maternal Education	0.0183*	-0.0147
	(0.00852)	(0.0260)
Male	-0.0116	0.0243+
	(0.0104)	(0.0146)
Family Income	0.00652	-0.0251*
	(0.0141)	(0.0103)
Grade-level controls:		
Proportion Black	-0.542**	0.0855*
	(0.177)	(0.0418)
Proportion Hispanic	-0.293	0.0370
	(0.216)	(0.0374)
Proportion Other Race	-0.0771	0.0373+
	(0.148)	(0.0216)
Proportion Household Smoker	0.138	-0.0418*
	(0.0943)	(0.0191)
Proportion Older Sibling	0.134*	-0.0183
	(0.0624)	(0.0137)
Mean Maternal Education	0.197+	0.0547*
	(0.113)	(0.0263)
Mean Male	-0.0718	-0.0163
	(0.0843)	(0.0168)
Mean Family Income	-0.00834	0.0249
	(0.166)	(0.0206)
Constant	0.707***	0.227***
	(0.172)	(0.0638)
N	3875	3853
R^2	0.841	0.074
adj. R^2	0.837	0.051

Notes: Standard errors in parentheses
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Metagenomic effects for whites

We include individuals of all races but control for the individual's own race and the racial composition of the grade. The concern with this approach is that the instrument might not work as well when applied to other races because the gene discovery was conducted on individuals of European ancestry. However, other studies have shown that—unlike the case for other phenotypes[iii]—the smoking PGSs from GWAS based on European ancestry populations predict smoking behavior in individuals of other ancestries.[iv] Additionally, a GWAS of individuals of African ancestry has shown overlapping SNPs with European ancestry populations.[v]

We also conduct additional analyses where we test whether the findings hold for same race egos and peers (i.e., individuals affected by their same race peers). We only test the within-race effects for white respondents because the number of grades with substantial non-white peers is too low to allow for this test. Principle Components were constructed on the white subsample and the first two are included in the analyses. Table S5 shows that the effect of white peers on white egos is again positive and significant at the grade-mate level, even if the effect size is attenuated from what we find for the whole sample. Classroom and friend regressions (including in- and out- degree nominations), both likely underpowered, fail to reach significance.

Table S5: Metagenomic effects for whites only

	(1) Grade-mates	(2) Classmates	(3) Friends	(5) In-degree	(6) Out-degree
Mean Smoking PGS (residualized)	0.0697*	-0.00884	0.0208	0.00764	0.0391
	(0.0281)	(0.0247)	(0.0207)	(0.0242)	(0.0262)
Controls:					
Smoking PGS (residualized)	-0.00347	-0.0252	-0.00344	-0.0216	0.0333
	(0.0188)	(0.0230)	(0.0219)	(0.0269)	(0.0221)
Household Smoker	0.166***	0.199***	0.125***	0.130***	0.0914***
	(0.0209)	(0.0240)	(0.0217)	(0.0267)	(0.0245)
Older Sibling	0.0379+	0.00700	0.0220	0.0291	0.00606
	(0.0194)	(0.0284)	(0.0199)	(0.0278)	(0.0218)
Maternal Education	-0.0455*	-0.0559+	-0.0304	-0.0578*	-0.0402
	(0.0207)	(0.0290)	(0.0224)	(0.0276)	(0.0257)
Male	-0.0109	-0.0192	-0.0137	0.0000170	-0.0239
	(0.0170)	(0.0247)	(0.0196)	(0.0265)	(0.0227)
Family Income	-0.0387**	-0.0411+	0.00725	0.0374	0.0147
	(0.0135)	(0.0232)	(0.0249)	(0.0330)	(0.0239)
Proportion Household Smoker	-0.0607	0.0242	0.0642**	0.0600*	0.0435+
	(0.0499)	(0.0257)	(0.0240)	(0.0297)	(0.0263)
Proportion Older Sibling	-0.0246	-0.00361	0.000652	0.00773	-0.00146
	(0.0401)	(0.0279)	(0.0229)	(0.0262)	(0.0261)
Mean Maternal Education	0.000236	-0.127***	-0.0748**	-0.0773**	-0.0531+
	(0.0549)	(0.0306)	(0.0257)	(0.0269)	(0.0314)
Proportion Male	-0.0681	-0.0123	-0.0185	-0.0218	-0.0206
	(0.0461)	(0.0276)	(0.0267)	(0.0386)	(0.0242)
Mean Family Income	-0.0450	0.0416	-0.0538***	-0.0518**	-0.0412
	(0.0669)	(0.0283)	(0.0160)	(0.0157)	(0.0279)
Constant	-0.105	-0.341**	-0.320**	-0.271*	-0.317***
	(0.123)	(0.108)	(0.110)	(0.118)	(0.0741)
N	2160	1348	1805	1380	1303
R^2	0.127	0.122	0.148	0.159	0.144
adj. R^2	0.092	0.068	0.090	0.084	0.063

Notes: Standard errors in parentheses
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In Table S6, we replicate the findings for Table S3 on the white sample. The direction and magnitude of the coefficients tell a similar story as the results for the whole sample and again only the proportion of bad apples in one's grade is statistically significant.

Table S6: Bad apples and shining stars for whites

	(1) Grade-Mates	(2) Grade-Mates	(3) Classmates	(4) Classmates	(5) Friends	(6) Friends	(7) In-degree	(8) In-degree	(9) Out-degree	(10) Out-degree
Proportion Bad Apples	0.0613*		-0.0364+		0.00828		0.00703		0.0287	
	(0.0288)		(0.0199)		(0.0195)		(0.0258)		(0.0233)	
Proportion Shining Stars		0.00302		0.0189		-0.0208		-0.0187		-0.0277
		(0.0315)		(0.0258)		(0.0183)		(0.0212)		(0.0196)
Controls										
Smoking PGS (residualized)	-0.00566	-0.00857	-0.0259	-0.0250	-0.00440	-0.00341	-0.0219	-0.0208	0.0324	0.0325
	(0.0187)	(0.0190)	(0.0228)	(0.0228)	(0.0220)	(0.0219)	(0.0268)	(0.0267)	(0.0222)	(0.0221)
Household Smoker	0.168***	0.167***	0.200***	0.199***	0.126***	0.125***	0.130***	0.130***	0.0935***	0.0900***
	(0.0210)	(0.0212)	(0.0240)	(0.0240)	(0.0216)	(0.0216)	(0.0267)	(0.0267)	(0.0243)	(0.0246)
Older Sibling	0.0377+	0.0378+	0.00763	0.00639	0.0221	0.0213	0.0292	0.0280	0.00579	0.00473
	(0.0192)	(0.0192)	(0.0284)	(0.0283)	(0.0199)	(0.0200)	(0.0278)	(0.0279)	(0.0218)	(0.0216)
Maternal Education	-0.0463*	-0.0477*	-0.0542+	-0.0561+	-0.0310	-0.0303	-0.0580*	-0.0576*	-0.0418	-0.0404
	(0.0207)	(0.0206)	(0.0289)	(0.0291)	(0.0224)	(0.0224)	(0.0276)	(0.0274)	(0.0258)	(0.0255)
Male	-0.00901	-0.00987	-0.0207	-0.0186	-0.0132	-0.0135	0.000007	-0.00023	-0.0242	-0.0237
	(0.0169)	(0.0170)	(0.0246)	(0.0249)	(0.0196)	(0.0196)	(0.0265)	(0.0265)	(0.0227)	(0.0227)
Family Income	-0.0377**	-0.0377**	-0.0414+	-0.0415+	0.00748	0.00672	0.0379	0.0371	0.0147	0.0137
	(0.0134)	(0.0132)	(0.0230)	(0.0234)	(0.0249)	(0.0248)	(0.0333)	(0.0330)	(0.0238)	(0.0235)
Proportion Household Smoker	-0.0440	-0.0557	0.0262	0.0251	0.0652**	0.0637**	0.0602*	0.0593*	0.0450+	0.0454+
	(0.0482)	(0.0544)	(0.0259)	(0.0255)	(0.0239)	(0.0238)	(0.0297)	(0.0295)	(0.0264)	(0.0268)
Proportion Older Sibling	-0.0229	-0.0198	-0.00426	-0.00440	0.00110	0.00156	0.00783	0.00808	-0.00228	-0.000577
	(0.0394)	(0.0399)	(0.0277)	(0.0278)	(0.0229)	(0.0230)	(0.0262)	(0.0264)	(0.0261)	(0.0260)
Mean Maternal Education	-0.00393	-0.0186	-0.126***	-0.126***	-0.0744**	-0.0760**	-0.0769**	-0.0780**	-0.0528+	-0.0536+
	(0.0548)	(0.0576)	(0.0307)	(0.0307)	(0.0256)	(0.0259)	(0.0270)	(0.0269)	(0.0313)	(0.0316)
Mean Male	-0.0523	-0.0598	-0.0131	-0.0116	-0.0187	-0.0183	-0.0221	-0.0216	-0.0224	-0.0199
	(0.0492)	(0.0497)	(0.0273)	(0.0278)	(0.0265)	(0.0267)	(0.0386)	(0.0386)	(0.0240)	(0.0243)
Mean Family Income	-0.0405	-0.0373	0.0458	0.0416	-0.0540***	-0.0547***	-0.0517**	-0.0524***	-0.0413	-0.0434
	(0.0660)	(0.0673)	(0.0294)	(0.0281)	(0.0161)	(0.0159)	(0.0158)	(0.0157)	(0.0288)	(0.0281)
Constant	-0.149	-0.140	-0.358***	-0.343**	-0.325**	-0.330**	-0.272*	-0.280*	-0.323***	-0.328***
	(0.127)	(0.132)	(0.104)	(0.113)	(0.113)	(0.110)	(0.119)	(0.117)	(0.0755)	(0.0734)
N	2160	2160	1348	1348	1805	1805	1380	1380	1303	1303
R^2	0.126	0.125	0.124	0.123	0.148	0.148	0.159	0.159	0.144	0.144
adj. R^2	0.092	0.091	0.070	0.069	0.090	0.090	0.084	0.085	0.062	0.062

Notes; Standard errors in parentheses
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Pleiotropy and network structure

We evaluated whether there are possible pleiotropic effects between smoking PGS and social network structure. For example, the polygenic score for smoking may also reflect individuals' friendliness or agreeableness, which would artificially induce a correlation between grade-level smoking and grade-level network density (i.e. the proportion of the possible friendships in a grade which are in fact observed). To evaluate whether this was the case, we performed simple correlation analyses. We correlated grade-level network density with the grade-level mean polygenic score for smoking, produced by residualizing each individual's polygenic score on their first four PCs and then averaging over each grade. The resulting grade-level polygenic scores were further residualized by school fixed effects before estimating the correlation. This returned a very low correlation (0.03), which was insignificant at 0.05 level. We performed similar analyses at the individual (ego-centric) level and found similarly trivial and insignificant correlations.

Evaluating the assumption that PGS is exogenous given grade and school fixed effects

We evaluated our central assumption that PGS is exogenous given grade and school fixed effects. This assumption amounts to the idea that, within a school, the assignment to grade is random with respect to genes and, specifically, smoking PGS. Another way to conceptualize this is that there is some gene pool for the specific school but selection into grade, in particular, is random, so that we can unbiasedly test between-grade differences.

To test whether this is true, we ran a series of simulations where for each school, we randomly re-assign students' grades, holding the number of students in each grade constant, effectively jumbling the students' assignments to grade within each school, and as a result, the assignment of smoking PGS to grade. We do this 1000 times for each school, at each iteration calculating each grade's mean smoking PGS, giving each grade within a school 1000 bootstrapped mean smoking PGS scores for comparison.

We then test if the observed grade mean smoking PGS for each grade falls within the bounds of that grade's specific simulation results. To do this, we establish a 95% interval for each grade within each school by ordering the simulations for that grade according to their estimated mean smoking PGS value and set the bounds as the value at the upper and lower 2.5% of this ordered distribution. If, within a school, a grade's true smoking value falls within its simulated 95% interval, then we say that it is reasonable that it was drawn from the school superset at random.

Thus, for each school and grade combination, we can evaluate whether its observed mean smoking PGS value is reasonably random given the school PGS distribution. We then calculate what proportion of school-grade combinations appear randomly drawn. If our assumption, that grade PGS are drawn at random from school PGS, is true, this proportion should be close to 95% (i.e. the size of the interval).

Overall, we find that 96.2% of grades fall within 95% bounds of their simulation, giving us confidence that our central assumption holds.

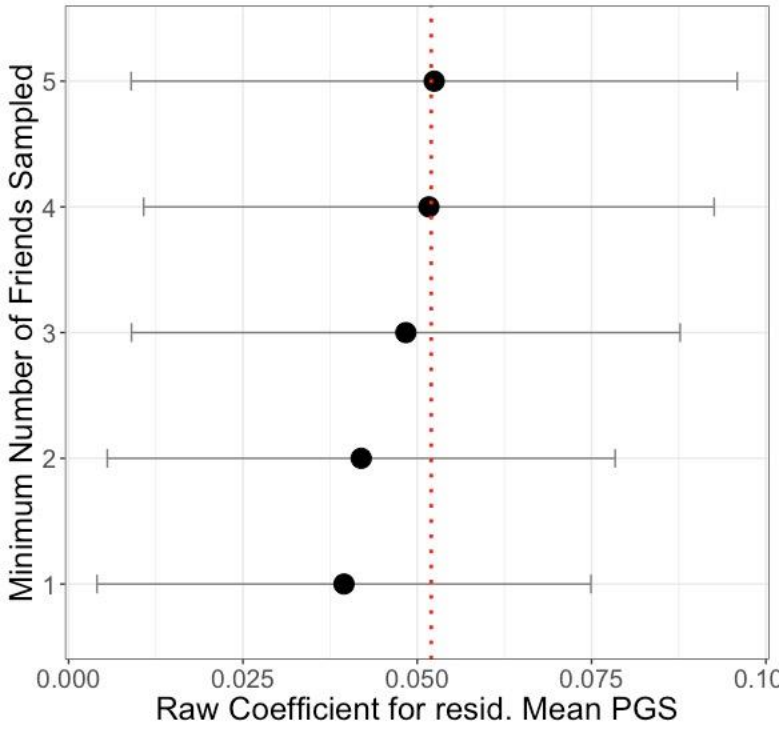
Understanding the magnitude of the grade-level, compared to friend-level, coefficients

The friend-level coefficient is lower than the grade-level coefficient in the main results, despite the likely presence of homophily at the friend-level and the general assumption from sociological literature that closer relationships should be more influential than distant ones.

We hypothesized that this slight difference in coefficient size was due to the fact that friend groups were usually smaller and therefore estimates of friend-level PGS were noisier, which was compounded by the 2-to-9 sampling rate of adolescents (and thus 2-to-9 sampling of friends they nominated during the in-school interview) from the in-school sample for in-home interviews in which DNA was collected[vi], all resulting in a higher noise-to-signal ratio in the friend-group when compared to the grade.

We tested this hypothesis by evaluating the estimates from friend groups which included varying numbers of respondents. Since the intended outcome of this exercise was to explain both homophily and influence, we use coefficients from a naïve model that contains school and grade level fixed effects as the only controls. Figure S1 shows below that as the number of friends increases, the coefficient approaches the estimate we obtained using the grade-level models marked by the dashed red line.

Figure S1: Coefficient size as a function of number of peers sampled



References

[i] Manski C.F. (1993). Identification of endogenous social effects: The reflection problem. *Rev Econ Stud* 60:531-542.

[ii] O'Malley A.J., Elwert F., Rosenquist J.N., Zaslavsky A.M. & Christakis N.A. (2014). Estimating peer effects in longitudinal dyadic data using instrumental variables. *Biometrics* 70:506-15.

[iii] Bamshad M. (2005). Genetic influences on health: does race matter? *Jama* 294:937-946.

[iv] Saccone N.L., et al. (2009). The CHRNA5-CHRNA3-CHRNB4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in African-Americans and in European-Americans. *Cancer Res*. 69:6848-56.

[v] Chen L.S., et al. (2012). Smoking and genetic risk variation across populations of European, Asian, and African American ancestry--a meta-analysis of chromosome 15q25. *Genet Epidemiol* 36(4):340-51.

[vi] Harris, K.M., et al. (2019). Cohort Profile: The National Longitudinal Study of Adolescent to Adult Health (Add Health). *International Journal of Epidemiology*. Forthcoming.