Legewie, Joscha; DiPrete, Thomas A.

The High School Environment and the Gender Gap in Science and Engineering

Sociology of Education

Sociology of Education
Online Appendix
DOI: 10.1177/0038040714547770

Online Appendix A – Coding of STEM Fields

The coding of the STEM variables (planed major in 12th grade and major of bachelor's degree) separates science and engineering fields that are traditionally gender-typed and usually require a pre-college preparation in math and science from other fields. Based on this general motivation, we exclude nursing and other health care majors from STEM fields. Sensitivity analyses that include clinical and health sciences such as nursing in the STEM category generally confirm our findings.

Field of study classifications *Non S/E fields*: no bachelor's degree earned, agricultural business/production, agricul/ture/animal/plant sci, conservation/natural resources, forestry, architect/environmental design, graphic/industrial design, drama, speech, film arts, music, fine arts/art history, fpa: other, accounting, finance, ops research/administrative science, business admin/management, hrd/labor relations, other business, other business support, medical office support, marketing/distribution, journalism, communications, radio/tv/film, communication technologies, early childhood education, elementary education, secondary education, special education, physical education, education: other, med/vet lab tech/assist, dental assist/hygiene, hper, allied health: other, physical therapy, occupational therapy, other therapies, speech path/audiology, clinical health sci, nursing, health/hospital admin, public health, oth health sci/profess, para-legal/pre-law, law, psychology, anthropol/archaeology, economics, geography, history, sociology, political science, internat relations, other, amer studies/civiliz, area studies, ethnic studies, retailing, hospitality mgmnt, real estate, information technols, other personal service, engin tech: el/electron, computer technology, foreign languages, nutrition/food sci, textiles/fashion, fcs and oth human ecology, child study/guidance, culinary arts/food mgt, english/amer literature, writing: creative/tech, letters: other, liberal/general studies, library/archival sci, womens studies, environ studies, biopsychology, integrated/gen science, interdisc humanities, social sci: general, interior design, recreation/sports, philosophy, religious studies, theology, bible studies, clin/counsel psych, admin of justice, social work, public administration, human/community serv, graphic/print communic, and air transport.

DOI: 10.1177/0038040714547770

STEM fields: biochemistry, biological science: other, math sciences/statistics, chemistry, geology/earth science, physics, phys sci: other, computer programming, data/information management, computer science, electrical/communication engineer, chemical engineering, civil engineering, mechanical engineering, engineering: other, computer engineering, and engineering tech: non-elect

Online Appendix B – Analysis for the Section 'Is the High School Effect Lasting and How Big Is the Effect?'

In the section "Is the High School Effect Lasting and How Big Is the Effect?", we reported results about post–high school transition rates for students from different high schools as well as for the reduction in the gender gap in STEM BAs if all high schools would encourage women to study science and engineering. To conduct this assessment, we group high schools by the size of the gender gap in science and engineering orientation. In particular, we use the empirical Bayes estimates of the gender gap from the value-added multilevel model (NELS 88-92) in the main text to group schools into those with a small gender gap (bottom terciles) and those with a big gender gap (top terciles). We then match this newly created school-level variable to the students in NELS 88-2000. Table A1 reports post–high school transition rates for the full school sample and for high schools with a small and a big gender gap. The post–high school transition rates include the rate at which students change their orientation to a different field (leakage rate), persist in pursuing their STEM major plans (persistence rate), and enter a STEM major without having developed such plans in high school (late entry rate). Results show that the post–high school transition rates are remarkably constant across the three samples, indicating that high schools have a lasting effect on gender differences.

[Table A1 about here]

For the second part of the analysis, we again group high schools into terciles according to the size of their gender gap in STEM orientation. We then calculate the gender gap in STEM BA

DOI: 10.1177/0038040714547770

degrees assuming the same 8th-grade orientation and post–high school transition rates across all three samples. In other words, we assume that differences in the gender gap across the three samples only emerge because of differences in the transition rates within high schools, and not from group differences in 8th-grade orientation and transition rates after high school. As shown in Table A2, in the full sample, boys are 1.7 times as likely as girls to graduate from college with a STEM BA degree. However, this substantial male advantage is reduced to 1.3 (male/female odds ratio) in the subsample of students who attend high schools with a small gender gap. Accordingly, the gender gap would be reduced by about 25 percent if the environment in all schools would encourage girls to study science and engineering at the same rates as do the top third of schools.

[Table A2 about here]

Online Appendix C – High School Effect for STEM Subfields

[Table A3 about here]

[Table A4 about here]

Online Appendix D – Sensitivity Analysis: Robustness to Violations of Conditional Independence Assumption

Although we control for a large set of pre-treatment control variables that are directly related to the selection process, unobserved confounding variables might nonetheless bias our estimated effects for the two high school variables. To estimate their potential impact, we conduct a simulation based sensitivity analysis. Other sensitivity analyses have been proposed for estimating the effect of confounding variables in propensity score matching analyses (Rosenbaum 2002), linear regression models (Frank 2000), or instrumental variable regression (DiPrete and Gangl 2004). Here we apply a simulation-based sensitivity analysis proposed by Ichino, Mealli,

DOI: 10.1177/0038040714547770

and Nannicini (2008; see also Nannicini 2007) for matching methods to the case of regression analysis with a binary dependent variable. The starting point of this and similar sensitivity analyses is to posit an unobserved variable U (here assumed to be binary) that violates the conditional independence assumption. The binary covariate U can be simulated based on different assumptions and added to the regression model as an additional covariate to get an understanding about the robustness of the results to specific failures of the independence assumption.

For simplicity, consider a set of observed pre-treatment covariates X and three binary variables: the treatment variable T, the outcome Y, and an unobserved confounding variable U. To qualify as a confounding variable, U has to be associated with both the treatment and the outcome variable after controlling for X. If we make the further simplifying (and conservative) assumption that U and X are independent, conditional on T and Y,[1] the distribution of U can be characterized with a set of four probabilities p_{ij} that define U depending on the treatment and outcome status (see Ichino et al. 2008:317)

$$\begin{aligned} p_{ij} &= P(U{=}1|T{=}i, Y{=}j, X) \\ &= P(U{=}1|T{=}i, Y{=}j) \\ with &\quad i,j \in \{0,1\} \end{aligned}$$

Hence, p_{ij} defines the probability that $P(U{=}1)$ when $T{=}i$ and $Y{=}j$. Following Ichino and colleagues (2008), we focus our sensitivity analysis on two statistics based on these four parameters p_{ij} that reflect different assumptions about the unobserved confounding variable. In particular, "the real threat to the baseline estimate is coming from a potential confounder that has both a positive effect on the untreated outcome ($p_{01}-p_{00}>0$), [hereafter, "*d*"] and on the selection into treatment ($p_{1.}-p_{0.}>0$) [hereafter, "*s*"]" (Ichino et al. 2008:318). The two statistics, *d* and *s,* together with the marginal probability $P(U{=}1)$ and the difference $p11-p10$ determine the four values of p_{ij}. Accordingly, fixing two secondary statistics $P(U{=}1){=}.4$ and setting $p11-p10{=}0$ allows us to simulate U for each observation in our dataset using random draws from a binomial

DOI: 10.1177/0038040714547770

distribution with p_{ij} as the probability parameter so that $U \sim Binominal(p_{ij})$.[2] After simulating U, we reestimate Model II in Tables 4 and 5 for the curriculum index and gender segregation with the additional control variable U and confined to the female respondents. We then compare the observed effects with the one obtained with the additional simulated confounder U. Changing the parameter d and s in this simulation and comparing the obtained effects helps us to understand how robust the estimated effect is to additional unobserved covariates. The parameter d is associated with, but not the same as, the effect of U on the untreated Y, and the parameter s is associated with, but not the same as, the effect of U on s. For each d and s, we can compute the average odds ratio in the data of the effect of U on Y, conditional on X (hereafter, Γ), and also the average odds ratio of U on T, conditional on X (hereafter, Λ). We can thereby produce unobserved confounding variables that have effects similar to those of observed covariates. For further details about this sensitivity analysis, we refer the reader to Ichino and colleagues (2008) and Nannicini (2007).

[Figure A1 about here]

Figure A1 shows the results of our simulation-based sensitivity analysis for the curriculum Index, which has the smallest observed effect (similar results were obtained for the other treatment indicators and are available from the authors). We show the results for values of d and s that range from 0 to .5 and reflect different relations between the simulated cofounder U and the treatment and outcome variable. To restate, our treatment variable is the index value of the number of AP math and science courses in the high school, and our outcome variable is the STEM orientation in 12th grade. As indicated by the Γ and Λ values in the figure, these values of d and s correspond to an odds-ratio effect of U on Y of between 1.1 and 11.1, and of between 1 and 10.7 on the treatment indicator after conditioning on all the covariates used in the main regression. The shading of each square indicates how the estimated effect size changes depending on the two parameters d and s, with black indicating the observed effect size of T in the sample,

DOI: 10.1177/0038040714547770

and white indicating a zero or negative effect. For most of the observed pre-treatment covariates, the odds ratio for the outcome effect (conditional on other covariates) is between .8 and 1.2, while some have slightly higher values. The estimated effect of a specific covariate *x* on *T* conditional on the other observed covariates (the selection effect of *x*), is generally smaller than is the estimated outcome effect. Accordingly, most estimated covariates conditional on observable variables are equivalent to a confounding variable that would be located in the four squares in the top-left corner of Figure A1, where *d* and *s* are both in the range [0, .05]. Confounding variables that had a similar strength relationship with *Y* and *T,* as do nearly all of the observed covariates, would fall in the slightly wider range of the upper-left nine squares where both *d* and *s* are in the range [0, .10]. For these values of *d* and *s*, the estimated effect with the simulated confounder *U* is still substantial. For example, 8th-grade math performance, which is one of the most important control variables, has an odds-ratio outcome effect of 1.06 and a selection effect of 1.212 and therefore lies in the square region defined by $d \in [0, .05]$ and $s = .05$. As the diagram shows, confounding variables as powerful as 8th-grade math performance still leave a substantial portion of the positive curriculum effect on STEM orientations intact. Accordingly, our estimates are relatively robust to an additional confounder that is similar to the currently used control variables and unrelated to any of the covariates in the current model.[3]

References

DiPrete, T.A. and M. Gangl. 2004. "Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments." Sociological Methodology pp. 271–310.

Frank, Kenneth. 2000. "Impact of a Confounding Variable on the Inference of a Regression Coefficient." Sociological Methods & Research 27:147–194.

Ichino, Andrea, Fabrizia Mealli, and Tommaso Nannicini. 2008. "From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity? " *Journal of Applied Econometrics* 23:305–327.

DOI: 10.1177/0038040714547770

Nannicini, Tommaso. 2007. "Simulation–based sensitivity analysis for matching estimators." Stata Journal 7:334–350.

Rosenbaum, Paul. 2002. Observational Studies. New York: Springer Verlag.

DOI: 10.1177/0038040714547770

Tables and Figures

Table A1: Post–High School Transition Rates for Full Sample and Schools with Small/Big Gender Gap

Post-HS Transition Rates	Gender	Full Sample	Schools with Small Gender Gap	Schools with Big Gender Gap
Leakage Rate	male	0.669	0.615	0.692
	female	0.649	0.686	0.637
Late Entry Rate	male	0.078	0.084	0.082
	female	0.051	0.06	0.039
Persistence Rate	male	0.331	0.385	0.308
	female	0.351	0.314	0.363

Note: National Education Longitudinal Study 88-2000. The sample uses multiple imputation for missing data. It excludes drop-out students and students who did not participate in the base year or 1st, 2nd, or 4th follow-up.

DOI: 10.1177/0038040714547770

Table A2: Gender Gap in STEM BAs for Full Sample, Schools with Small Gender Gap, and Schools with Big Gender Gap

Proportion of Students with STEM bachelor's degree			Gender Gap	
	Male	Female	Difference	Odds Ratio
Full Sample	0.098	0.060	0.038	1.713
Schools with Small Gender Gap	0.124	0.095	0.029	1.349
Schools with Big Gender Gap	0.077	0.039	0.039	2.089

Note: National Education Longitudinal Study 88-2000. The sample uses multiple imputation for missing data. It excludes drop-out students and students who did not participate in the base year or 1st, 2nd, or 4th follow-up.

DOI: 10.1177/0038040714547770

Table A3: Logistic Regression Estimates for the Effect of High Schools' Math and Science Curricula for Different STEM Subfields

	STEM		Physical Science and Engineering		Biological and Life Science	
	Coef	(se)	Coef	(se)	Coef	(se)
Intercept	-3.243***	(0.232)	-3.976***	(0.268)	-3.706***	(0.405)
Male	0.961***	(0.073)	1.331***	(0.129)	-0.183	(0.129)
Curriculum Index (CI)	0.145*	(0.057)	0.126	(0.087)	0.153	(0.087)
Curriculum Index (CI) x Male	-0.247***	(0.067)	-0.236**	(0.118)	-0.180	(0.118)
Pre–High School Control Variables						
Std Demographic Control Variables	yes		yes		yes	
Urban/Region Control Variables	yes		yes		yes	
8th-Grade Control Variables	yes		yes		yes	
Students	11,270		11,270		11,270	

** p < 0.05, ** p < 0.01, *** p < 0.001; Continuous variables are standardized; Clustered standard errors in parentheses*

Note: Data from the National Education Longitudinal Study. The sample uses multiple imputation for missing data. It excludes drop-out students and students who did not participate in all survey waves (base year and 1st and 2nd follow-up).Control variables are described in Table 2.

DOI: 10.1177/0038040714547770

Table A4: Logistic Regression Estimates for the Effect of Gender Segregation of Extra-Curricular Activities

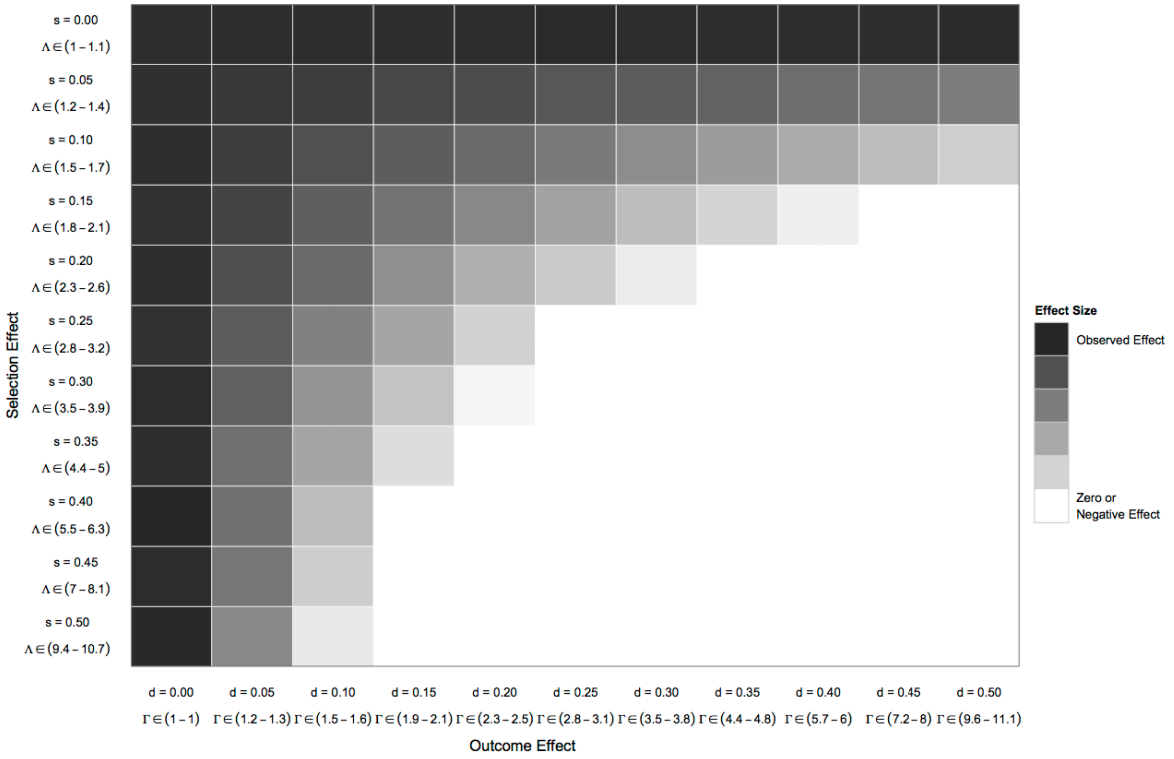
	STEM		Physical Science and Engineering		Biological and Life Science	
	Coef	(se)	Coef	(se)	Coef	(se)
Intercept	-3.050***	(0.450)	-3.282***	(0.505)	-5.170***	(0.872)
Male	0.540**	(0.182)	0.829***	(0.212)	-0.415	(0.361)
Gender Segregation	-0.302*	(0.140)	-0.112	(0.170)	-0.557*	(0.235)
Gender Segregation x Male	0.311*	(0.158)	0.195	(0.188)	0.135	(0.302)
Pre–High School Control Variables						
Std Demographic Control Variables	yes		yes		yes	
Urban/Region Control Variables	yes		yes		yes	
8th-Grade Control Variables	yes		yes		yes	
Students	2,350		2,350		2,350	

** p < 0.05, ** p < 0.01, *** p < 0.001; Continuous variables are standardized; Clustered standard errors in parentheses*

Note: Data from the High School Effectiveness Study combined with pre–high school information from the National Education Longitudinal Study (combined HSES sample). The sample uses multiple imputation for missing data. It excludes drop-out students and students who did not participate in all survey waves (base year and 1st and 2nd follow-up).Control variables are described in Table 2.

DOI: 10.1177/0038040714547770

Figure A1: Sensitivity of Estimate to Additional Confounding Variable



Note: For each combination of d and s*, we conducted 100 simulation runs so that the whole graph is based on 12,100 simulations. A Stata implementation of the simulation-based sensitivity analysis for matching procedures is available from Nannicini (2007). Our R implementation for both matching procedures and regression methods together with the graphical presentation of the results are available from the first author.*

DOI: 10.1177/0038040714547770

[1] The marginal association of U and X will be nonzero in the sample because of the association between U, T, and Y along with the association between X, T, and Y.
[2] The results are consistent across different values for $P(U=1)$ and $p11-p10$.
[3] Note that the assumption that the confounding variable is unrelated to the pre-treatment control variables in the current model produces a conservative sensitivity analysis considering that we use a large set of variables that are directly related to the selection process and highly relevant for the outcome.