Northwestern | INSTITUTE FOR POLICY RESEARCH

Working Paper Series

WP-20-22

Males at the Tails: How Socioeconomic Status Shapes the Gender Gap

David Autor
Ford Professor of Economics
Massachusetts Institute of Technology

David Figlio
Orrington Lunt Professor of Education and Social Policy,
Dean of the School of Education and Social Policy, and IPR Fellow
Northwestern University

Krzysztof Karbownik
Assistant Professor of Economics
Emory University

Jeffrey Roth
Research Professor of Pediatrics
University of Florida

Melanie Wasserman
Assistant Professor of Economics
University of California, Los Angeles

Version: May 26, 2020

DRAFT
Please do not quote or distribute without permission.

ABSTRACT

Analyzing Florida birth certificates matched to school records, the researchers document that the female advantage in childhood behavioral and academic outcomes is driven by gender gaps at the extremes of the outcome distribution. Using unconditional quantile regression, they investigate whether family socioeconomic status (SES) differentially affects the lower tail outcomes of boys. They find that the differential effects of family SES on boys' outcomes are concentrated in the parts of the distribution where the gender gaps are most pronounced. Accounting for the disproportionate effects of family environment on boys at the tails substantially narrows the gender gap in high school dropout.

The authors thank seminar participants at University of California at Davis. Autor acknowledges support from the Russell Sage Foundation (Grant #85-12-07). Figlio and Roth acknowledge support from the National Science Foundation and the Institute for Education Sciences (CALDER grant), and Figlio acknowledges support from the National Institute of Child Health and Human Development and the Bill and Melinda Gates Foundation. Wasserman acknowledges support from the National Institute on Aging, Grant #T32-AG000186. They are grateful to the Florida Departments of Education and Health for providing the de-identified, matched data used in this analysis. The conclusions expressed in this paper are those of the authors and do not represent the positions of the Florida Departments of Education and Health or those of our funders.

Introduction

Women now outpace men in many measures of educational achievement, including the propensity to graduate from high school, enroll in post-secondary education, and graduate from college. In 2010, for example, the high school graduation rate among U.S. women was 87 percent, while it was 81 percent among U.S. men (Murnane, 2013). These female-favorable gaps have motivated a burgeoning literature examining their potential determinants, with recent papers tracing the evolution of gender disparities during childhood (Buchmann and DiPrete, 2006; DiPrete and Jennings, 2012; Lundberg, 2017; Autor et al., 2019). While realized gender gaps in adult educational outcomes are large, a wrinkle in the examination of the precursor childhood gender gaps is that these differences appear *on average* to be relatively modest. For example, recent work documents that among U.S. eighth grade students during the early to mid-2000s, boys and girls exhibited modest differences in their mean test scores, with boys maintaining a small advantage in math and girls maintaining a more robust advantage in reading (Pope and Sydnor, 2010; Bertrand and Pan, 2013). In behavioral outcomes, where boys have long experienced a higher incidence of disciplinary problems than girls, the average female-favorable gap is larger but still modest (Bertrand and Pan, 2013; Autor et al., 2019). In the Florida public school system that we study here, the gender gap in school absences is a mere 0.45 percentage points, with the average boy and girl both attending more than 94 percent of school days.

In this paper, we document and analyze two fact patterns that jointly help to explain why *modest* mean gaps between boys and girls in early academic and behavioral outcomes translate into large differences in realized educational attainment. We first show that female-favorable gaps in behavioral and academic outcomes during childhood—where present—stem largely from the overrepresentation of boys in the lower tails of the academic and behavioral outcome distributions. This is visible in the upper two panels of Figure 1, which plots the fraction of students who are boys at each percentile of the academic and behavioral outcome distributions among children born in Florida in 1992-1993 who attended Florida public schools in grades five to eight. Evident from this figure is the substantial overrepresentation of boys in the bottom quintile of attendance rates (panel A) and reading and math scores (panel B). Boys make up 49 percent of the sample (the dark horizontal line in each panel). But at the 10^{th} percentile of the attendance and reading score distributions, boys comprise 55 percent of the population, and their overrepresentation rises convexly at lower percentiles. For math test scores, males are overrepresented at both the lower and upper tails of the distribution, yielding a small mean math advantage for boys.

These childhood lower tail behavioral and academic outcomes are highly predictive of subsequent high school dropout. As shown in panels C and D of Figure 1, high school dropouts are drawn disproportionately from the lower tails of the test score and attendance distributions. Children at the 10^{th} percentile of the math and reading score distributions are almost four times as likely to leave high school without a degree as those at the 90^{th} percentile. Poor attendance in school is even more predictive: the dropout ratio among 10^{th} percentile attendees exceeds that of 90^{th} percentile attendees by a factor of six.

Figure 1 Panels C and D additionally reveal the second pattern that undergirds our analysis: boys and girls at the lower tails of the behavioral and academic distributions have a similar likelihood to drop out of high school; thus, conditional on childhood school performance, there is only a small remaining gender gap in high school dropout. These two patterns—boys and girls at the tails have comparable dropout rates, but boys are substantially overrepresented at the tails—suggests that better understanding *why* males are overrepresented at the lower tails of the childhood outcome distribution may help to illuminate why they are more prone to high school dropout.

This paper tests the hypothesis that these boy-girl differences in tail outcomes stem in part from differential susceptibility of boys to adverse child-rearing conditions—specifically, that less favorable home environments differentially raise the prevalence of adverse outcomes among boys relative to girls. Because these adverse outcomes are determinative of high school dropout, this differential sensitivity could help explain the large gender gap in dropout. Our analysis employs the universe of Florida birth records for years 1992-2000, matched to public school test score and disciplinary outcomes, to assess the effect of childhood environmental influences—family, neighborhoods, and schools—on the gender gap throughout the distribution of behavioral, academic, and high school attainment outcomes. To quantify the effect of family socioeconomic status (SES) on the gender gap throughout the outcome distribution, we use the unconditional quantile regression method of Firpo et al. (2009).

Our empirical strategy rests on two assumptions. The first is that child gender is as good as randomly assigned to family types due to naturally occurring chance determination of the gender of a child at the time of conception. The second assumption is that the gender gap in *potential* outcomes is independent of family SES throughout the outcome distribution. We assess the validity of these assumptions, first, by testing whether boys are more or less likely than girls to be born to disadvantaged families. Second, we examine whether family SES differentially affects boys' neonatal health throughout the distribution. From these tests, we conclude that family environment does *not* exert a meaningful differential effect on boys' relative to girls' initial conditions, as measured both by their initial allocation to family types and by their health at birth—consistent with the supposition that its impact on outcomes arises postnatally.

Our empirical analysis has two main results. First, we show that the differential adverse effects of family disadvantage (proxied by SES) for boys appear to be concentrated in precisely the parts of the distribution where the gender gaps are most pronounced. These relationships are evident from unconditional quantile regressions, and they are robust to including detailed controls for confounding factors. Second, by extrapolating the effects of family environment on the gender gap in grade-school behavioral and academic outcomes to high school dropout and on-time graduation decisions, we show that a substantial fraction of the gender gap in high school outcomes can be explained by the differential effect of family SES on boys' medium-run outcomes. For the lowest decile of the behavioral and academic outcome distributions, a one standard deviation increase in family SES—equivalent to the difference between a family with a married high school graduate mother and a family with an unmarried high school dropout mother—would eliminate over 40 percent of the

decile-specific gender gap in high school dropout. This impact is due to (1) the differential benefits of family environment for behavioral and academic outcomes of boys at low quantiles and (2) the positive relationship between these intermediate outcomes and high school completion for both boys and girls.

Our paper contributes to a growing literature examining the effect of childhood environment on the gender gap in educational outcomes. Thus far, this literature has focused on conditional mean differences. Bertrand and Pan (2013), Autor et al. (2019), and Aucejo and James (2019) demonstrate that growing up in a more disadvantaged family environment disproportionately harms boys' childhood outcomes, including disciplinary infractions, standardized test scores, and high school graduation rates.[1] Another strand of the literature characterizes gender gaps throughout the academic outcome distribution but does not explore their determinants. Contini et al. (2017) use Italian data and Fryer and Levitt (2010) and Robinson and Lubienski (2011) use U.S. data to document that the gender gaps in math (favoring males) in kindergarten through eighth grade are largest at the upper end of the achievement distribution. Ellison and Swanson (2010) show that boys outscore girls at the far right tail of advanced mathematics. Robinson and Lubienski (2011) further document that gender gaps in reading (favoring females) are largest at the bottom of the achievement distribution. This paper departs from the prior literature by, first, documenting the critical role of gender gaps at the tails of the distribution in explaining large gender disparities in educational attainment, and second, by exploring the determinants of these tail gaps.

1 Data and empirical approach

1.1 Florida birth certificates linked to school records

Our data come from the universe of birth certificates from Florida for years 1992-2000 linked to public school records from 2002-03 through 2009-10 school years.[2] For each year that a child attends a Florida public school, our data report the child's Florida Comprehensive Assessment Test (FCAT) math and reading scores, as well as records on daily absences and suspensions for grades three through eight. To measure children's academic performance, we use standardized math and reading test scores. To measure children's behavioral performance, we compute a "combined attendance rate" by totaling the number of absences and suspensions in a given academic year, dividing by the number of school days in the year, and subtracting this absence rate from one. We account for the higher rates of absences and suspensions during middle school relative to elementary school grades by normalizing each child's combined attendance rate by the average combined attendance rate in their grade. To reduce computational demands of the quantile regression models, we average grade three through eight outcomes, so each child contributes one observation to the analysis for each

[1] For college attendance, employment, and earnings, Chetty et al. (2016) find a differential advantage of growing up in a higher income family for boys relative to girls. These longer-run effects documented in the U.S. using administrative data stand in contrast with findings from Danish administrative data (Brenøe and Lundberg, 2018) and U.S. survey data (Lundberg, 2017; Lei and Lundberg, 2020).

[2] Details on the matching procedure are reported in Figlio et al. (2014).

outcome.

We construct a family SES index from a principal components analysis of demographic variables reported on the child's birth certificate: maternal education (in years); maternal age at birth (in years); family structure (married, not married); and Medicaid receipt at the time of birth.[3] We compute a measure of neighborhood SES by aggregating the family SES index to the zip code level (excluding the child's own family) and assigning children the neighborhood SES of their zip code of birth. Our measure of school quality is from the Florida Department of Education's school-level gain scores, measuring schools' estimated average contribution to student outcomes. For each school, we average the gain scores between 2002 and 2013 and then convert this average into a percentile rank in the gains distribution across Florida public schools. We assign each child the cumulative quality of schools attended from grades three through eight, by computing a years-weighted average of the schools' percentile ranks (Autor et al., 2016). The main sample comprises all children born in Florida 1994-2000 who attended Florida public schools. To compute the relationship between childhood behavioral and academic outcomes and subsequent high school completion, we additionally analyze children born in 1992 and 1993, for whom we observe high school graduation outcomes. Summary statistics are found in Appendix Tables A.2 and A.3.

1.2 Empirical strategy

Our empirical framework for estimating the effect of family SES on the gender gap in children's outcomes throughout the distribution follows Havnes and Mogstad (2015). Let $F_{Y_{g,s}}(y)$ represent the cumulative distribution function of educational outcome Y for children with gender $g \in \{m, f\}$, whose family SES at the time of the children's birth is $s \in \{0, 1\}$, where one denotes the family is high SES and zero denotes the family is low SES.[4]

First we construct the distributional contrast γ_g, which captures for children of gender g, the difference in the shares from high and low SES backgrounds who score above a given outcome level y.

$$\gamma_g(y) = (1 - F_{g,1}(y)) - (1 - F_{g,0}(y))$$

This quantity will be positive if a higher fraction of high SES than low SES children of gender g score above level y.

Since family SES is not randomly assigned to children, γ_g will incorporate not only differences between children due to SES but also differences due to the correlation of SES and other child and environmental characteristics. For this reason, γ_g does not reflect the pure quantile treatment effect of SES on the outcomes of children of gender g. We can write $1 - F_{Y_{g,s}}(y)$ as the sum of two distributions, $\delta_{g,s}(y) + \eta_{g,s}(y)$, which allows us to write:

$$\gamma_g(y) = (\delta_{g,1}(y) - \delta_{g,0}(y)) + (\eta_{g,1}(y) - \eta_{g,0}(y)),$$

[3] Details on the construction of this measure are in Table A.1.

[4] We use a discrete SES measure here for expositional clarity. The SES measure in our empirical application is continuous.

where $\delta_{g,1}(y) - \delta_{g,0}(y)$ is the causal effect of SES on the share of children of gender g who score above level y, while $\eta_{g,1}(y) - \eta_{g,0}(y)$ is the difference in the latent distributions of the educational outcome among children born to high and low SES households. Note that the subscript $g \in \{m, f\}$ indicates that we are holding gender constant while contrasting across levels of SES. In our definition of $\delta_{g,s}(y)$, we include both the direct effect of family SES and the indirect causal effect of SES that may operate through other correlated channels, such as schools and neighborhoods. We address these potential environmental confounds in our empirical analysis.

We can then write the difference-in-differences contrast as follows:

$$\begin{aligned}\tau_{DiD}(y) &= \gamma_m(y) - \gamma_f(y) \\ &= [(\delta_{m,1}(y) - \delta_{f,1}(y)) - (\delta_{m,0}(y) - \delta_{f,0}(y))] \\ &\quad - [(\eta_{m,1}(y) - \eta_{f,1}(y)) - (\eta_{m,0}(y) - \eta_{f,0}(y))]\end{aligned}$$

The first bracketed expression in this equation corresponds to the causal effect of SES on the share of boys relative to girls who score above outcome level y. The second bracketed expression contains terms reflecting the potentially confounding difference in the latent distributions of the educational outcome among high and low SES children. In addition, it contains the potentially confounding difference in the latent distributions of the educational outcome among boys and girls of a given SES level. To permit identification of the causal effect of SES on the gender gap in educational outcomes throughout the distribution, we make the following two assumptions:

Assumption 1. *The latent gender gap in $\eta_{g,s}(y)$ is independent of SES. Hence $E[\eta_{m,1}(y) - \eta_{f,1}(y)] = E[\eta_{m,0}(y) - \eta_{f,0}(y)]$ for all y.*

Under this assumption, the double difference of the shares of high-SES boys and girls relative to low-SES boys and girls who score above outcome level y eliminates the bias terms. This assumption does *not* require that the latent distribution of educational outcomes is independent of SES for either sex. Rather, akin to the standard parallel trends assumption in a differences-in-differences (DD) setting (augmented for a quantile setting), this assumption requires that any gender difference in latent outcomes is independent of SES and hence can be eliminated by differencing the gender gap across SES levels. We test the validity of this assumption in section 2.3.[5]

If SES were perfectly observed in our data, then assumption 1 would be sufficient to identify the causal effect of SES on the gender gap throughout the distribution. In practice, we proxy SES with detailed information from birth certificates. Even under assumption 1, our usage of a proxy for SES could lead to a spurious correlation between the gender gap in $\eta_{g,s}$ and the proxy, if there is gender imbalance among family types. This would potentially confound the causal effect of SES

[5]In a standard DD setting, the assumption that the contrast in outcomes between untreated (here, low SES) boys and girls is an additive constant permits identification. In the quantile setting, we allow the latent relationship between SES and the educational outcome to vary without restriction throughout the distribution while imposing that it does so identically by gender. The double difference of the distributions of outcomes between high-SES boys and girls relative to low-SES boys and girls eliminates the confounding effect of the $\eta_{g,s}$ terms. Due to the quantile analysis, this additivity assumption is invariant to monotone transformations of the dependent variable (Athey and Imbens, 2006), which is not true in a standard DD model.

on the gender gap with the non-random assignment of genders to family types. This confound is eliminated with the following assumption, the validity of which we test in section 2.3:

Assumption 2. *The gender of children is as good as randomly assigned to family SES.*

Bringing this empirical approach to the data, we construct the following contrast:

$$\hat{\tau}_{DiD}(y) = (F_{f,0}(y) - F_{m,0}(y)) - (F_{f,1}(y) - F_{m,1}(y)), \tag{1}$$

The estimand $\hat{\tau}_{DiD}$ corresponds to differences in *shares* of children scoring above each educational outcome level. To convert these shares into quantile treatment effects, we employ the method developed by Firpo et al. (2009) to estimate unconditional quantile regression using the recentered influence function (RIF). This approach estimates the quantile treatment effect by contrasting γ_m and γ_f and scaling this contrast by the kernel density of the joint distribution of SES and test scores at a given y. Thus, the RIF analysis recovers the quantile treatment effect by inverting the cumulative distribution function of the outcome variable in the neighborhood of the treatment variable.

2 Family environment and the gender gap throughout the distribution

2.1 Main results

We estimate baseline gender gaps using the following specification:

$$Y_i = \alpha + \beta \text{Boy}_i + e_i \tag{2}$$

where Y_i is an academic or behavioral outcome for child i, and Boy_i is an indicator for a male child. To characterize the gender gap throughout the outcome distribution, we replace the dependent variable in equation (2) with the recentered influence function for each quantile of the academic or behavioral outcome distribution.

Next we incorporate the interaction of child gender and family environmental factors into the specification:

$$Y_i = \alpha + \beta_1 \text{Boy}_i + \beta_2 \text{SES}_i + \beta_3 \left(\text{Boy}_i \times \text{SES}_i\right) + \mathbf{X'}_i\gamma + e_i \tag{3}$$

where SES_i is an index of the family's socioeconomic status at birth, and $\mathbf{X'}_i$ is a vector of other controls including child race and ethnicity, year and month of birth, and birth order; and we again replace the dependent variable with the recentered influence function. The coefficient β_3 on the interaction term $(\text{Boy}_i \times \text{SES}_i)$ permits the relationship between family SES and outcomes to differ by child sex. Under our identifying assumptions, this coefficient corresponds to the differential causal effect of family SES on outcomes of boys relative to girls.

The first outcome we consider is attendance in grade school, a behavioral outcome. On average, girls are absent 5.1 percent of school days while boys are absent 5.6 percent of school days (Table A.3). This mean difference masks considerable heterogeneity in the gender gap throughout

the attendance distribution, however. Figure 2 Panel A reveals this heterogeneity by plotting the raw gender gap (which we have reversed for expositional purposes to be the girl-boy gap) at 50 percentiles of the attendance distribution.[6] The positive gender gap indicates that boys have a lower attendance rate (higher absence rate) at essentially every point in the distribution, consistent with the overrepresentation of boys at the lower end of the distribution depicted in Figure 1. Relevant to our analysis, the female-favorable gender gap in attendance is considerably larger at the lower tail of the distribution. At the 10^{th} percentile, boys miss 1.06 percentage points more school days than do girls.[7] This gap shrinks monotonically as one moves upward in the distribution. The gap at the 90^{th} percentile of the distribution of 0.10 percentage points is less than one-tenth as large as that at the 10^{th} percentile of the distribution.

To characterize the effects of family SES on the gender gap throughout the attendance distribution, we estimate unconditional quantile regressions corresponding to equation (3), which we plot for 50 percentiles. The coefficients on the interaction term $\text{Boy}_i \times \text{SES}_i$ plotted in Panel A of Figure 2 indicate that the differential effect of family environment on boys' attendance rates is largest at the lower end of the distribution, precisely where the gender gap is most pronounced. As the female advantage in attendance attenuates, the estimated effect of family SES on the gender gap also declines. We estimate that a 1σ fall in SES increases the 10^{th} percentile boy-girl gap in absences by 0.63 percentage points. At the median, this effect is only 0.14 percentage points, and at the 90^{th} percentile, it is a mere 0.03 percentage points.

These estimated effects can be compared to the observed gender gaps at these percentiles, equal to 1.06, 0.30, and 0.10, respectively. The estimates imply that a 1σ decline in SES is predicted to expand the lower-tail (10/50) gender gap in attendance by roughly two-thirds of its observed magnitude.[8] Similar to Autor et al. (2019), we find a substantial differential effect of family SES on boys' behavioral outcomes *on average* as well. The OLS coefficient on the interaction of boy and family SES is positive and highly significant, at 0.26 percentage points. Our findings here, however, indicate that this mean effect is driven almost entirely by the lower half of the distribution. A 1σ rise in SES is predicted to close the lower-tail 10/50 boy-girl attendance gap by 0.49 percentage points while compressing the upper tail 50/90 gap by only 0.11 percentage points.

The estimates for the academic outcomes—reading and math scores—tell a more nuanced story. For reading, there is a female advantage, both on average and throughout the distribution. The female-favorable gap in reading test scores plotted in Figure 2 Panel B largely mirrors that for attendance. The gender gap in reading narrows from 0.27σ at the 10^{th} percentile to 0.09σ at the 90^{th} percentile of the score distribution (coefficients reported in Appendix Table A.5). As with attendance, we estimate that boys differentially benefit from a more advantaged family environment. At the 10^{th} percentile, a 1σ increase in the family SES index corresponds to a 0.14σ closure of the gender gap in reading, half of the raw gap.

Math test scores report a small mean male *advantage* of 0.03 standard deviations (Appendix

[6] Specifically, we estimate a RIF regression of equation (2) and plot the coefficients on Boy_i multiplied by -1.
[7] RIF regressions estimates corresponding to this figure are reported in Appendix Table A.4.
[8] $0.64 = (0.63 - 0.14) / (1.06 - 0.30)$

Table A.6). This mean difference stems from boys outperforming girls above approximately the 20^{th} percentile of the score distribution. Meanwhile, there is a substantial male disadvantage in math scores below the 20^{th} percentile. As depicted in Panel C of Figure 2, family SES exerts a differential effect on boys' versus girls' math scores, but this effect is non-monotone throughout the outcome distribution. At the tails of the distribution, boys differentially benefit from higher family SES, while in the middle of the distribution, family SES confers a slight differential benefit to girls.

In summary, for all three outcomes, the differential effect of SES on boys is amplified at the lower tail of the distribution, precisely where there are large female-favorable gaps. If we provisionally interpret these as causal estimates, we would conclude that the differential benefits of a one standard deviation increase in family environment for boys can explain nearly half of the male disadvantage at the lower tail of the academic and behavioral distributions.

2.2 The role of other environmental factors

Our hypothesis is that family environment—proxied by SES—is differentially consequential for the behavioral and academic outcomes of boys relative to girls. While we have so far considered family SES in isolation, family SES is correlated with other environmental factors that may also affect behavioral and achievement outcomes, such as the demographics of neighborhoods and the quality of schools that children attend. To distinguish the role of SES from other correlated factors, we estimate augmented models of the form:

$$\begin{aligned} Y_i = &\alpha + \beta_1 \text{Boy}_i + \beta_2 \text{SES}_i + \beta_3 \left(\text{Boy}_i \times \text{SES}_i\right) + \\ &\beta_4 \text{NBD}_i + \beta_5 \left(\text{Boy}_i \times \text{NBD}_i\right) + \\ &\beta_6 \text{SCH}_i + \beta_7 \left(\text{Boy}_i \times \text{SCH}_i\right) + \mathbf{X'}_i \gamma + e_i, \end{aligned} \tag{4}$$

where we additionally include controls for a child's neighborhood SES at birth, NBD_i, and school quality, SCH_i, as well as the interactions of these attributes with Boy_i. Figure 3 displays how the inclusion of these additional controls affects our main results by plotting the interaction term $\text{Boy}_i \times \text{SES}_i$ with and without controls for school quality and neighborhood SES (corresponding OLS and RIF regression results are reported in Appendix Tables A.7, A.8, and A.9). Panel A shows that the inclusion of school and neighborhood controls modestly reduces the differential effect of family SES on boys' attendance, though more than two-thirds of the unadjusted effect remains at the 10^{th}, 25^{th}, and 50^{th} percentiles. Similarly, the effect of family SES on the gender gap in math and reading scores is modestly attenuated by these additional environmental controls (Panels B and C), and its estimated impact remains substantial at the lower tail, where these effects are largest.

2.3 Testing the identifying assumptions

Family SES and the gender gap in potential outcomes: We probe the validity of assumption 1–that the gender gap in potential outcomes is independent of family SES throughout the outcome distribution–by examining the relationship between family SES and neonatal health. Using equation

(3), we test the relationship between SES and the gender gap in both log birthweight and a health index based on a principal component analysis of all observed birth outcomes, including birthweight in grams, gestational length in weeks, one and five minutes Apgar scores, congenital disorders, labor and delivery complications, abnormal conditions at birth, maternal health problems, and adequacy of prenatal care. This analysis jointly tests whether family SES differentially affects sex at conception and the viability of male versus female embryos (as per Trivers and Willard, 1973)—which could occur if the sensitivity of the fetus to maternal stress levels during pregnancy differs by sex.

The results of this analysis are presented in Appendix Table A.10. Unsurprisingly, boys weigh three to four percent more than girls at the time of birth, both on average and throughout the birthweight distribution. Importantly, we detect almost no economically meaningful relationship between family SES and the gender gap in birthweight. The mean effect is a precisely estimated zero, whereas the estimated quantile effects range between plus and minus one percent. To interpret the magnitude of this relationship, we draw on the fact that, using the same Florida data as employed here, Figlio et al. (2014) find that a ten percent increase in birthweight generates a 0.05σ gain in children's academic outcomes. Our point estimates for the effect of SES on the gender gaps in birthweight ($+/-1\%$) are an order of magnitude *smaller* than this effect, implying that any confounding influence of SES on birthweight could add no more than $+/-0.005\sigma$ to the gender gap in academic outcomes. This is minute relative to the effects we estimate in Figure 2. The conclusion is similar for the specification examining the neonatal health index.

We also directly test the sensitivity of our main results to the inclusion of controls for birthweight and the at-birth health index and their interactions with Boy_i. The results are plotted in Appendix Figure A.1, where we observe that these additional controls have almost imperceptible effects on the estimated relationship between family SES and the gender gap in academic and behavioral outcomes throughout the distribution. These findings suggest that the effect of family SES on the gender gap manifests after birth, lending validity to the assumption that the gender gap throughout the distribution of potential outcomes is independent of family SES.

Exogeneity of gender: We probe the validity of Assumption 2, that boys and girls are as good as randomly assigned to families, by testing whether child gender is related to family socioeconomic status. We estimate the following linear probability model for child sex:

$$Boy_i = \alpha + \beta \text{SES}_i + e_i \tag{5}$$

where the variables have the same definition as those in the above specifications. Appendix Table A.11 reports the results. In the full sample of children born in the state of Florida, there is no relationship between family SES and the probability that a newborn child is male. In our analysis sample, which matches births to public schooling records, we note that there is a small, positive, and statistically significant relationship between the likelihood of a male child and family SES. A 1σ increase in SES is associated with a 0.007 increase in the fraction of children who are male. Given its statistical significance, we benchmark the economic magnitude of this relationship in two ways, both of which indicate that it is unlikely to bias our results.

First, we compare the relationship between SES and the sex ratio in our matched sample to sex ratios in the literature. Our coefficient of 0.007 shifts the sex ratio from the sample mean of 1.021 to 1.050. This is within the range of the biological norm in the white U.S. population, which does not exhibit sex selective behavior (Almond and Edlund, 2008; Almond and Sun, 2017).

Second, we examine the sensitivity of our results to the imbalance of male children throughout the family SES distribution by implementing the bounding approach of Lee (2009). As an upper bound, we assume the excess gender is positively selected and drop the excess males (females) throughout the SES distribution that are the highest performing in each of our outcomes. As a lower bound, we assume the excess gender is negatively selected and drop the excess males (females) throughout the SES distribution that are the lowest performing. We then re-estimate the main specification from equation (3) using the upper and lower bound samples.[9] The results are depicted in Appendix Figure A.2. For all outcomes, the lower bounds are just below the 95 percent confidence intervals of the original point estimates in the lower half of the distribution. This pattern supports the notion that, if the excess gender is negatively selected, the relative prevalence of boys in advantageous family circumstances does not impact the effects of family SES on the gender gap. In the upper half of the distribution, the lower bounds are a bit lower than the 95 percent confidence intervals of the original point estimates, indicating that we may slightly overestimate the differential effects of family SES on the highest quantiles of the male distribution. The upper bounds for all outcomes substantially amplify the $\text{Boy}_{ig} \times \text{SES}_{is}$ coefficients at the lower tails. Thus, if the overrepresented gender is positively selected—meaning the highest performing boys are missing from our sample—our main results *understate* the differential benefits of family SES for boys.

3 Implications for the Gender Gap in Educational Achievement

What do the effects of family environment on the gender gap throughout the childhood outcome distribution imply for the gender gap in high school graduation? In our sample, we observe high school outcomes for cohorts born in 1992 and 1993. Among these cohorts, the gender gap in on-time high school completion is seven percentage points, while the gap in high school dropout is three percentage points, both favoring women (Appendix Table A.3).[10] To compute the contribution of family environment to the gender gap in high school outcomes, we scale our unconditional quantile estimates of family environment on the gender gap in academic and behavioral outcomes by the relationship between each outcome and high school dropout/completion.

First, we estimate the relationship between intermediate (attendance, reading scores, math scores) and high school outcomes (on-time graduation and dropout), using a cubic specification:

$$HighSchoolOutcome_i = \alpha + f(\text{Attendance}_i) + f(\text{Reading}_i) + f(\text{Math}_i) + \mathbf{X'}_i\gamma + e_i \qquad (6)$$

[9] See Appendix B for the details of the bounding procedure.

[10] The remaining four percentage points are due to gender differences in the fraction of students who remain in high school past the on-time completion date. We do not observe whether these students ultimately complete high school.

where $f(\cdot)$ are cubic polynomials of intermediate academic and behavioral outcomes. The regression results, reported in Appendix Table A.12, confirm the robust predictive power of all three elementary/middle school outcomes for high school completion. Next, we multiply the quantile-specific estimates of the coefficient on $\text{Boy}_i \times \text{SES}_i$ from equation (3) by the marginal effect on the high school outcome from equation (6), corresponding to the value at the same quantile of the intermediate outcome. The results of this extrapolation exercise are presented in Figure 4, where the bars denote the contribution of a 1σ increase in family environment to the gender gap in high school outcomes through its effects on the gender gap in math, reading, and attendance, respectively. In our setting, a one standard deviation shift represents, for example, shifting from growing up in a family with a mother who is an unmarried high school dropout to one in which the mother is a married high school graduate.

Panel A depicts the effects for high school dropout, while Panel B plots the effects for on-time graduation. In Panel A, of the three intermediate outcomes, family environment makes the largest contribution to the gender gap in dropping out through its effects on attendance. This is due to the substantial differential effect of family environment on boys' attendance—particularly at the lower tail of the distribution—and the sizable predictive relationship between elementary/middle school attendance and subsequent high school dropout.[11] Family environment does not contribute meaningfully to high school outcomes through math or reading test scores. At the 10^{th} percentile of the attendance distribution, we observe that 2.9 percentage points, or 42 percent, of the decile-specific gender gap in high school dropout is explained by the contribution of family SES through attendance alone. When we instead focus on the contribution of family SES to the gender gap in high school dropout rates through its mean effect on attendance, the contribution is just 1.07 percentage points. Focusing on average effects thus obscures the substantial explanatory power of SES at the lower tail. Logically, this contribution is much smaller higher in the distribution, where family environment makes little contribution to gender gaps in childhood school outcomes (Figure 2).

The results for on-time high school graduation are comparable. At the lowest deciles, family environment can explain 3.5 percentage points, or 32 percent, of the decile-specific gender gap, again operating exclusively through its effects on attendance. In contrast, the contribution of family environment to the gender gap in high school graduation, through its *mean* differential effect on boys' attendance, is a mere 1.35 percentage points.

4 Conclusion

Leveraging rich longitudinal data, we document and analyze two patterns that in part explain why *modest* mean gaps between boys and girls in early academic and behavioral outcomes translate into

[11] From Appendix Table A.12, we note that the marginal effect of attendance on high school outcomes is *smaller* in magnitude at the lower tail than at the upper tail of the attendance distribution. Due to the substantial differential effect of SES on boys' attendance at the lower tail of the attendance distribution, however, the overall contribution of SES to the gender gap in high school outcomes through attendance remains larger at the lower tail.

large differences in educational attainment: first, female-favorable gaps in childhood behavioral and academic outcomes are driven primarily by a preponderance of boys at the lower tails of the respective outcome; and, second, conditional on these outcomes, there is only a small remaining gender gap in high school dropout. Using unconditional quantile regression, we find that family disadvantage differentially reduces the outcomes of boys at low quantiles, thus expanding the gender gap at the tails of the distribution. These tail differences translate into substantial gender gaps in the likelihood of high school non-completion. Family SES is correlated with other environmental attributes, such as neighborhood SES and school quality, but these correlates only modestly attenuate the differential effects of family environment on boys' outcomes. We unfortunately cannot directly analyze the effect of family SES on gender gaps after high school. The disproportionate effects of family circumstances on the lowest performing boys, however, paired with the predictive power of behavioral outcomes for high school completion, positions family environment to contribute meaningfully to gender gaps in long-term educational achievement and likely to adult earnings as well.

References

Almond, Douglas and Lena Edlund, "Son-biased sex ratios in the 2000 United States Census," *Proceedings of the National Academy of Sciences*, 2008, *105* (15), 5681–5682.

_ and Yixin Sun, "Son-biased sex ratios in 2010 US Census and 2011-2013 US natality data," *Social Science & Medicine*, 2017, *176*, 21–24.

Athey, Susan and Guido W Imbens, "Identification and inference in nonlinear difference in differences models," *Econometrica*, 2006, *74* (2), 431–497.

Aucejo, Esteban M. and Jonathan James, "Catching up to girls: Understanding the gender imbalance in educational attainment within race," *Journal of Applied Econometrics*, 2019, *34* (4), 502–525.

Autor, David, David Figlio, Krzysztof Karbownik, Jeffrey Roth, and Melanie Wasserman, "School Quality and the Gender Gap in Educational Achievement," *American Economic Review Papers and Proceedings*, 2016.

_ , _ , _ , _ , and _ , "Family Disadvantage and the Gender Gap in Behavioral and Educational Outcomes," *American Economic Journal: Applied Economics*, 2019, *11* (3), 338–81.

Bertrand, Marianne and Jessica Pan, "The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior," *American Economic Journal: Applied Economics*, jan 2013, *5* (1), 32–64.

Brenøe, Anne Ardila and Shelly Lundberg, "Gender gaps in the Effects of Childhood Family Environment: Do They Persist into Adulthood?," *European Economic Review*, 2018, *109*, 42–62.

Buchmann, Claudia and Thomas A DiPrete, "The Growing Female Advantage in College Completion: The Role of Family Background and Academic Achievement," *American Sociological Review*, 2006, *71* (4), 515–541.

Chetty, Raj, Nathaniel Hendren, Frina Lin, Jeremy Majerovitz, and Benjamin Scuderi, "Childhood Environment and Gender Gaps in Adulthood," *American Economic Review Papers and Proceedings*, 2016.

Contini, Dalit, Maria Laura Di Tommaso, and Silvia Mendolia, "The gender gap in mathematics achievement: Evidence from Italian data," *Economics of Education Review*, 2017, *58*, 32–42.

DiPrete, Thomas A. and Jennifer L Jennings, "Social and Behavioral Skills and the Gender Gap in Early Educational Achievement," *Social Science Research*, 2012, *41* (1), 1–15.

Ellison, Glenn and Ashley Swanson, "The Gender Gap in Secondary School Mathematics at High Achievement Levels: Evidence from the American Mathematics Competitions," *Journal of Economic Perspectives*, 2010, *24* (2), 109–128.

Figlio, David N, Jonathan Guryan, Krzysztof Karbownik, and Jeffrey Roth, "The Effects of Poor Neonatal Health on Children's Cognitive Development," *American Economic Review*,

2014, *104* (12), 3921–3955.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux, "Unconditional Quantile Regressions," *Econometrica*, 2009, *77* (3), 953–973.

Fryer, Roland G and Steven D Levitt, "An Empirical Analysis of the Gender Gap in Mathematics," *American Economic Journal: Applied Economics*, 2010, *2* (2), 210–240.

Havnes, Tarjei and Magne Mogstad, "Is universal child care leveling the playing field?," *Journal of Public Economics*, 2015, *127*, 100–114.

Lee, David S, "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 2009, *76*, 1071–1102.

Lei, Ziteng and Shelly Lundberg, "Vulnerable Boys: Short-term and Long-term Gender Differences in the Impacts of Adolescent Disadvantage," *IZA Discussion Paper No. 12944*, 2020.

Lundberg, Shelly, "Father Absence and the Educational Gender Gap," *IZA Discussion Paper No. 10814*, 2017.

Murnane, Richard J, "U.S. High School Graduation Rates: Patterns and Explanations," *Journal of Economic Literature*, jun 2013, *51* (2), 370–422.

Pope, Devin G and Justin R Sydnor, "Geographic Variation in the Gender Differences in Test Scores," *Journal of Economic Perspectives*, 2010, *24* (2), 95–108.

Robinson, J. P. and S. T. Lubienski, "The Development of Gender Achievement Gaps in Mathematics and Reading During Elementary and Middle School: Examining Direct Cognitive Assessments and Teacher Ratings," *American Educational Research Journal*, 2011, *48* (2), 268–302.

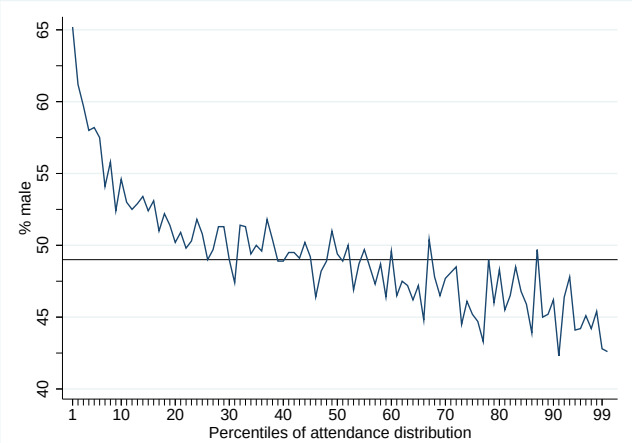
Trivers, Rorbert L and Dan E Willard, "Natural Selection of Parental Ability to Vary the Sex Ratio of Offspring," *Science*, 1973, *179* (4068), 90–92.

Figures and Tables

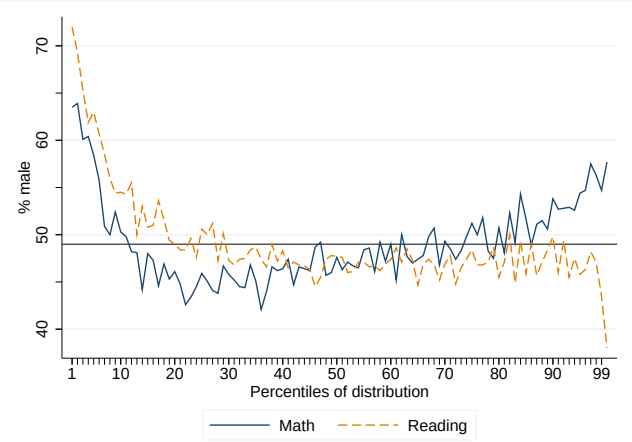
Figure 1: Males at the Tails

I. Fraction Male throughout the Educational and Behavioral Outcome Distribution

A. Attendance

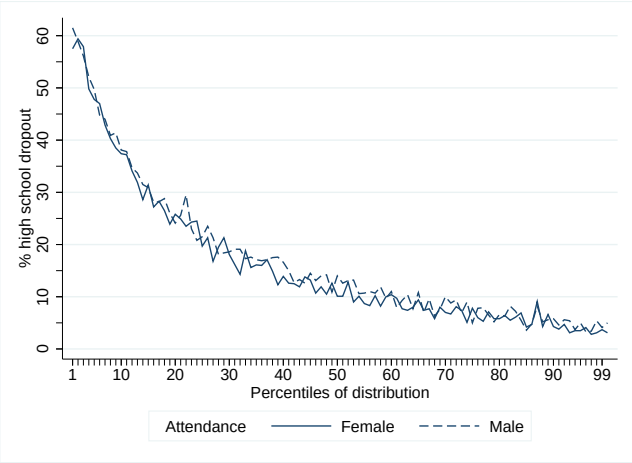


B. Test Scores

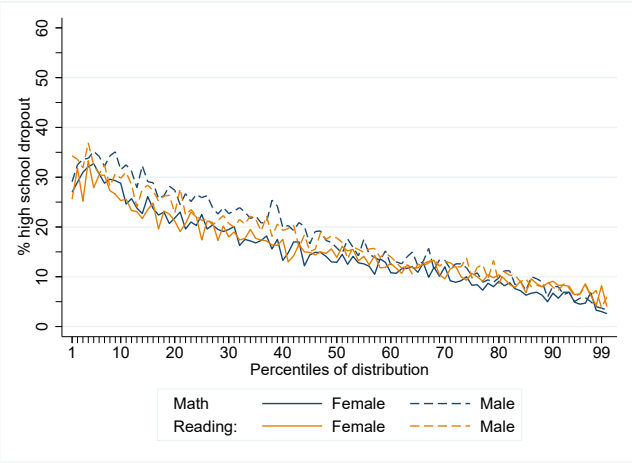


II. High School Dropout Rate throughout the Educational and Behavioral Outcome Distribution

C. Attendance

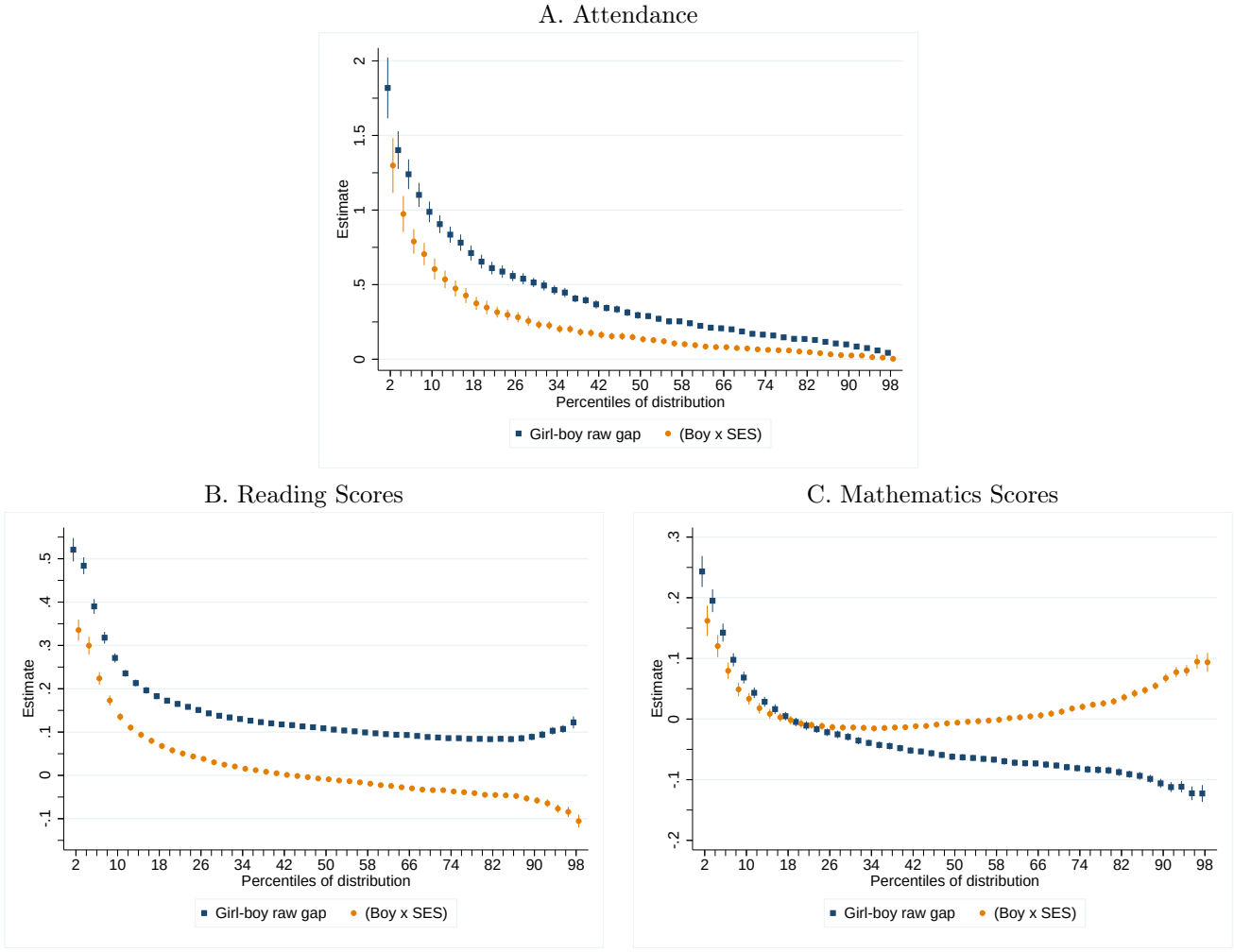


D. Test Scores



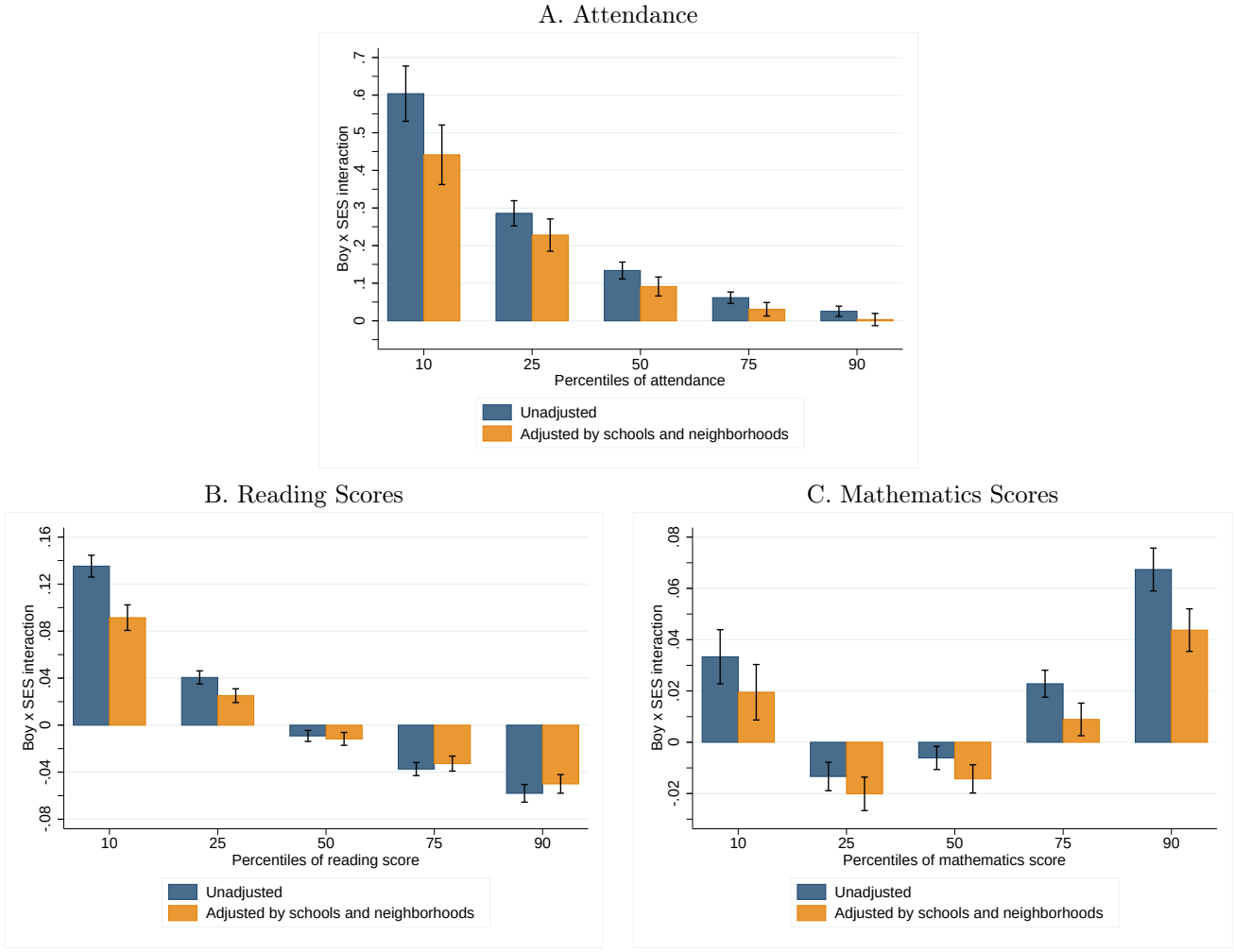
Note: These figures use data for birth cohorts 1992 and 1993 for whom we can observe both high school graduation outcomes as well as intermediate educational outcomes. The intermediate outcomes are: attendance rate, mathematics test scores, and reading test scores, all recorded for grades 5 to 8. Panel A plots the fraction male in each percentile of the attendance (averaged grades 5 to 8 attendance) distribution. Panel B plots the fraction male in each percentile of the test score (averaged grades 5 to 8 test scores) distributions separately for mathematics (solid navy line) and reading (dashed orange line). Panel C plots the fraction of high school dropouts at each attendance percentile (averaged grades 5 to 8 attendance) separately for females (solid navy line) and males (dashed navy line). Panel D plots the fraction of high school dropouts at each test score percentile (averaged grades 5 to 8 test scores) separately for females (solid lines) and males (dashed lines) and by testing domain (navy for mathematics and orange for reading).

Figure 2: Effect of SES on the Gender Gap throughout the Distribution



Note: The sample is individuals born between 1994 and 2000 in Florida. These figures plots girl-boy gender gaps (navy squares) and the differential effects of SES on boys (orange circles) for every other percentile of the attendance (Panel A), reading test score (Panel B), and mathematics test score (Panel C) distributions. The estimates are from RIF regressions (Firpo et al., 2009) and implemented using the rifreg command in Stata. Each scatterplot series contains 49 estimates. Raw gender gap estimates come from regressing one of the three outcome variables on a Boy indicator and multiplying these coefficients by -1. Boy x SES interactions come from regressing one of the three outcome variables on Boy indicator, SES index, interaction between Boy and SES (the plotted coefficients of interest), race/ethnicity indicators, month of birth indicators, year of birth indicators, and birth order indicators. Spikes represent 95% confidence intervals based on bootstrapped standard errors with 200 replications.

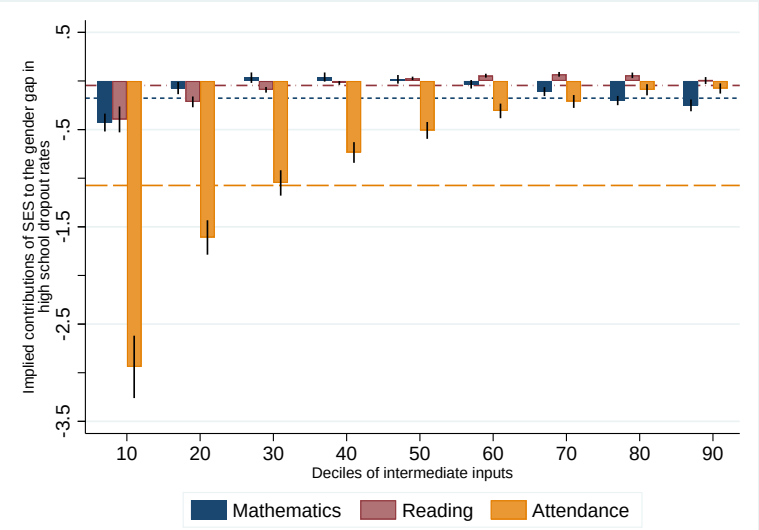
Figure 3: The Role of Schools and Neighborhoods in the Gender Gap throughout the Distribution



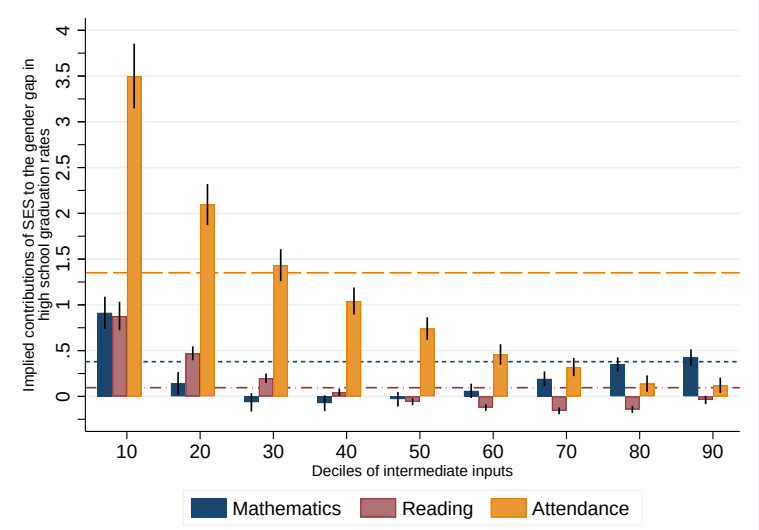
Note: The sample is individuals born between 1994 and 2000 in Florida. These figures plots the coefficients on the interaction of Boy x SES without (navy bars) and with (orange bars) controls for Boy x school quality and Boy x neighborhood SES measures. Each figure plots estimates from ten unconditional quantile regressions, estimated using RIF regressions (Firpo et al. 2009) and implemented using the rifreg command in Stata. The quantiles of interest are 10, 25, 50, 75, and 90. Panel A presents results for attendance, panel B presents results for mathematics test scores, and panel C presents results for reading test scores. Estimates represented by navy bars come from regressing one of the three outcome variables on Boy indicator, SES index, interaction between those two (plotted parameter of interest), school quality, neighborhood quality, race/ethnicity indicators, month of birth indicators, year of birth indicators, and birth order indicators. Estimates represented by orange bars include the same set of control variables but further add Boy x school quality and Boy x neighborhood SES interactions. Whiskers represent 95% confidence intervals based on bootstrapped standard errors with 200 replications.

Figure 4: Implications for the Gender Gap in High School

A. High School Dropout



B. High School Graduation

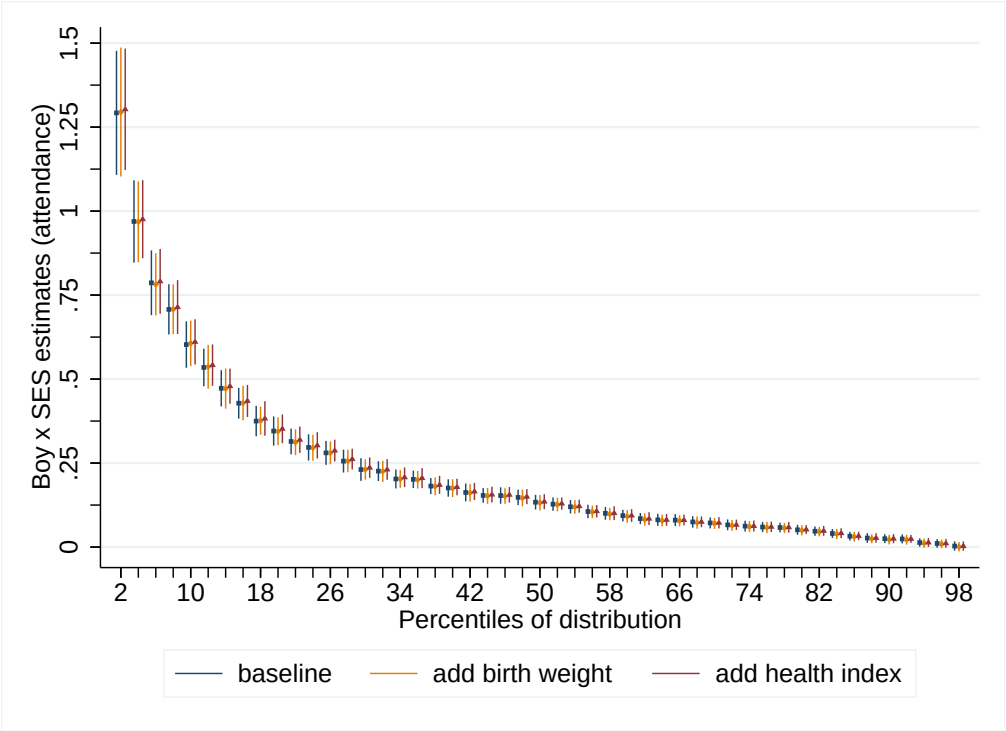


Note: The graphs depict an extrapolation of the intermediate attendance and test score results for high school dropout (Panel A) and high school graduation (Panel B). We first estimate the relationship between high school dropout or graduation and a cubic polynomial in attendance, reading test scores, and math test scores in grades 5 to 8, additionally controlling for gender, SES, race/ethnicity, month of birth dummies, year of birth dummies, and birth order dummies. We compute marginal effects of attendance, reading scores, and math scores on the two high school outcomes, separately at each decile (10 to 90) of these independent variables' distributions. Separately, we compute unconditional quantile effects of Boy x SES for 10 to 90 deciles from our main specifications as described in Figure 2. This figure plots the implied effects of a one standard deviation change in the SES index (1.51) on high school dropout or high school graduation, operating through the differential effect of SES on boys relative to girls (Boy x SES interaction) at each decile of the distribution of intermediate inputs. Each bar represents the decile-specific effect of a one standard deviation change in SES obtained by multiplying marginal effect of a given intermediate input (attendance, mathematics, or reading) and the Boy x SES coefficient for this input. The dashed lines represent the contribution of family SES to the gender gap through its average effect on attendance (orange), mathematics (navy), and reading (maroon). For high school graduation, the implied contributions of the mean effects are 1.35, 0.38, and 0.10 for attendance, mathematics and reading, respectively. For high school dropout, the implied contributions of the mean effects are -1.07, -0.18, and -0.05 for attendance, mathematics and reading, respectively. Standard errors are obtained by bootstrapping the procedure 500 times.

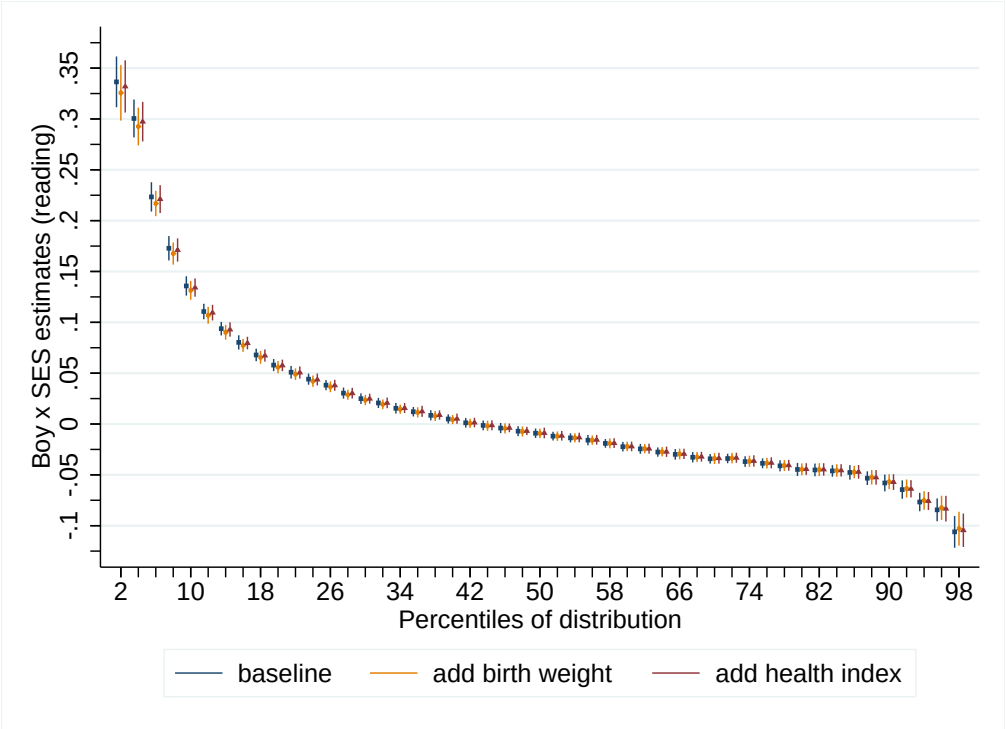
A Supplemental Figures and Tables

Figure A.1: Sensitivity of Main Results to the Inclusion of Neonatal Health

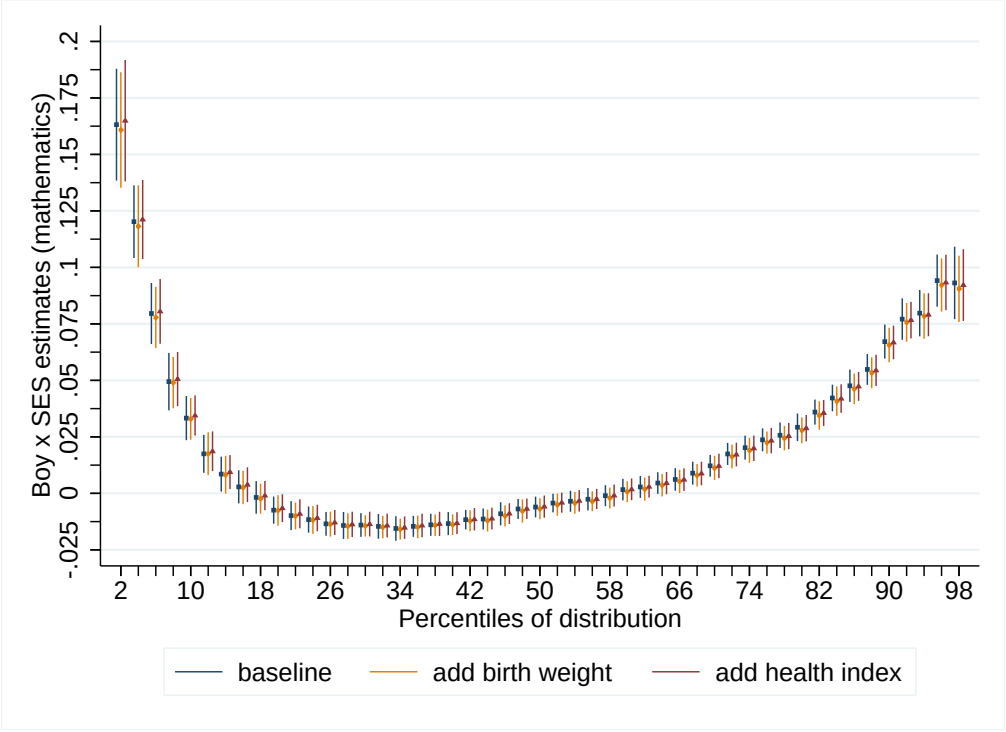
A. Attendance



B. Reading Scores



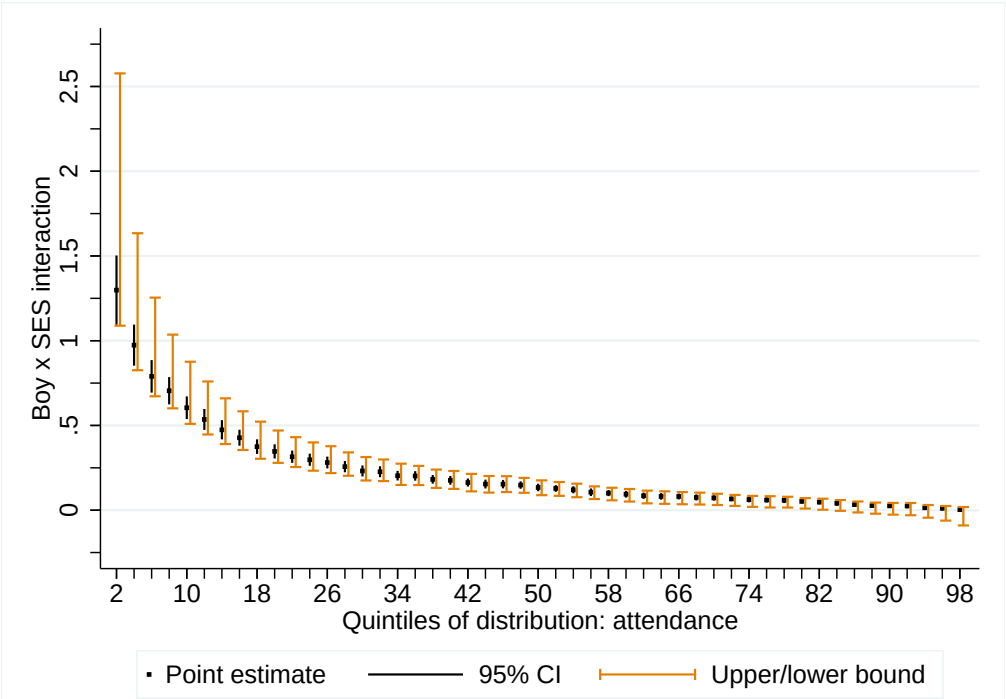
C. Mathematics Scores



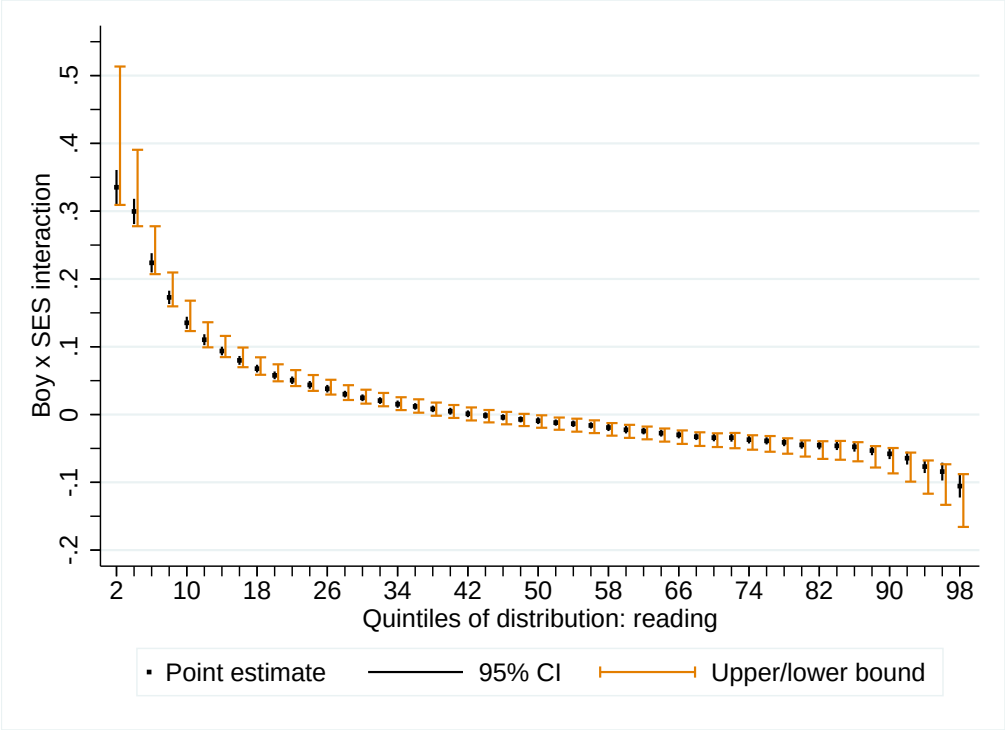
Note: This figure plots the coefficients on the interaction term Boy x SES, from three separate specifications. The navy squares depict the main results, estimated from equation (3). The orange diamonds plots the estimates from a specification that additionally controls for log birth weight and its interaction with Boy. The maroon triangles plot the estimates from a specification that additionally controls for the health index and its interaction with Boy. The health index is based on the first component of a principal components analysis (PCA) using the following variables: birth weight, gestational age, one and five minutes Apgar scores, indicator for adequate prenatal care, indicator for maternal health problems in pregnancy, indicator for congenital disorders, indicator of labor and delivery complications, and indicator for abnormal conditions at birth. Spikes represent 95% confidence intervals based on bootstrapped standard errors with 200 replications.

Figure A.2: Bounding the Effects of Gender Imbalance

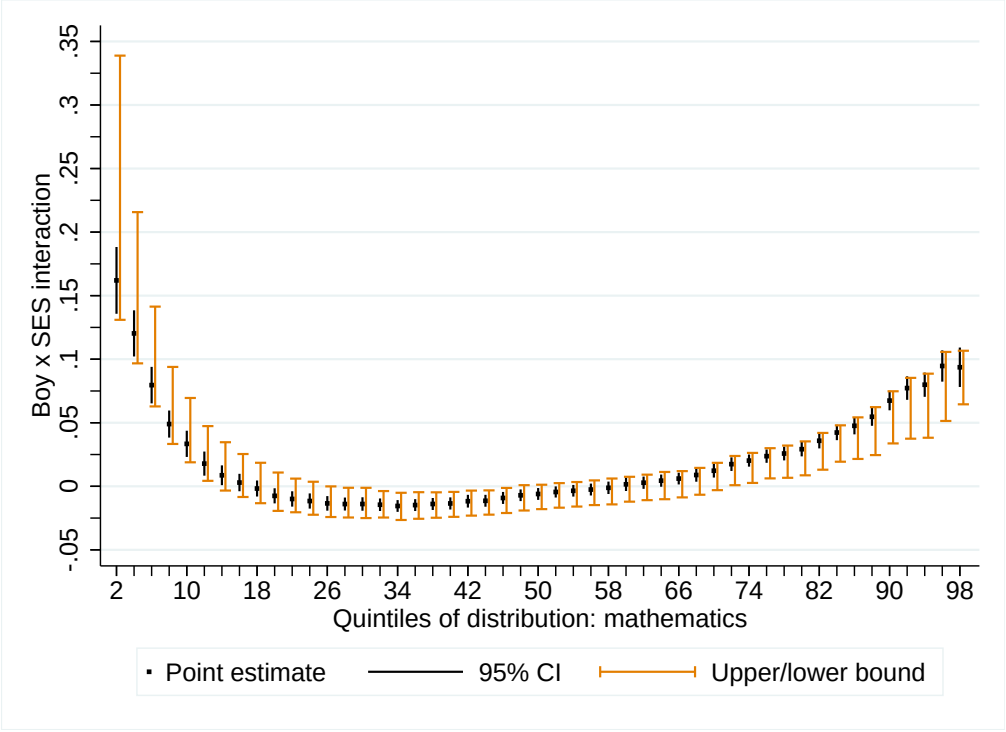
A. Attendance



B. Reading Scores



C. Mathematics Scores



Note: The sample is individuals born between 1994 and 2000 in Florida. This figure plots the coefficients on the interaction term Boy x SES, estimated from equation (3), and provides Lee bounds for the estimates. See Appendix B for details of the bounding procedure.

Table A.1: Construction of Principal Components SES Index

	First component	*Second component*
Mother's years of education	0.49	0.71
Married	0.50	-0.67
Non-medicaid birth	0.53	-0.18
Mother's age at birth	0.49	0.17
Eigenvalue	2.27	0.64
Summary statistics for the first component		
Mean	0.00	
Standard deviation	(1.51)	
Mean boys	0.00	
Standard deviation boys	(1.51)	
Mean girls	0.00	
Standard deviation girls	(1.51)	

Note: This table reports the results of a principal components analysis of mother's education (in years), mother's age at birth (in years), a non-Medicaid birth indicator, and an indicator for parents married at the time of birth. The eigenvectors associated with the first and second components are reported, as well as their associated eigenvalues. The bottom panel reports summary statistics of the SES index, defined as the first component of the principal components analysis, for the overall sample, and separately for boys and girls.

Table A.2: Sample Selection: Matched Florida Birth and Public School Records

	White, Black and Hispanic Births (1)	With Complete Data (2)	Matched to Florida School Records (3)	Matched to Outcomes (4)
White non-Hispanic	68.4	68.4	66.0	64.4
Black non-Hispanic	21.5	21.6	23.9	25.6
Hispanic	10.0	10.0	10.1	10.0
High school dropout	18.9	18.8	20.6	21.9
High school graduate	60.5	60.6	62.0	62.6
College graduate	20.3	20.6	17.4	15.5
Age 21 or below	24.1	24.0	26.1	27.4
Age between 22 and 29	42.1	42.1	42.2	42.0
Age between 30 and 35	24.9	25.0	23.5	22.7
Age 36 or above	8.9	8.9	8.3	7.9
Married at time of birth	64.3	64.5	61.4	59.2
Boy	51.3	51.3	51.0	50.5
Birth weight (grams)	3,343	3,344	3,331	3,326
N	940,609	900,801	734,074	552,819

Note: This table reports summary statistics (means) for the Florida statewide data for individuals born between 1994 and 2000. The first column is the full sample of Florida births 1994-2000, excluding immigrant mothers; the second column drops the 4.2% of records that are missing key variables; the third column contains the approximately 81% of column 2 records that were matched to Florida school records; and the fourth column is the subset of column 3 for children who remained in Florida public schools through third grade and had at least one test score. All demographic characteristics are derived from the birth certificate.

Table A.3: Summary Statistics

	(1) All	(2) Males	(3) Females	(4) Difference (M-F)
Attendance (raw)	94.64	94.42	94.86	-0.45
	(4.51)	(4.75)	(4.24)	(0.01)
Attendance (adjusted)	-0.32	-0.54	-0.08	-0.46
	(4.52)	(4.75)	(4.25)	(0.01)
Mathematics	0.07	0.08	0.05	0.04
	(0.90)	(0.93)	(0.86)	(0.00)
Reading	0.09	0.01	0.16	-0.15
	(0.88)	(0.92)	(0.84)	(0.00)
Observations	552,819	279,352	273,467	552,819
High school graduate	70.67	66.93	74.26	-7.33
	(45.53)	(47.05)	(43.72)	(0.24)
High school dropout	16.49	18.03	15.00	3.03
	(37.11)	(38.44)	(35.71)	(0.20)
Observations	144,945	71,140	73,805	144,945

Note: This table reports summary statistics (means, standard deviations, and standard errors) for outcomes of interest in the empirical samples used in the main analyses. Rows 1 to 5 are based on individuals born in Florida between 1994 and 2000 whom we observe with at least one year of outcomes (column 4 of Appendix Table A.2). Rows 6 to 8 are based on individuals born in Florida between 1992 and 1993 for whom we observe high school graduation outcomes. Column 1 presents means and standard deviations for children of both genders; column 2 presents means and standard deviations for males; column 3 presents means and standard deviations for females; while column 4 presents differences in means between columns 2 and 3 with standard errors.

Table A.4: Distributional Effects of Family SES on the Gender Gap: Attendance

	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
	A. OLS			*B. Q10*			*C. Q25*		
Boy × Family SES			25.88			63.28			30.06
			(1.27)			(3.67)			(1.98)
Family SES		135.19	122.05		258.23	226.09		188.17	172.90
		(0.71)	(0.91)		(3.15)	(3.41)		(1.71)	(1.78)
Boy	-48.68	-51.57	-51.54	-106.19	-112.57	-112.51	-59.48	-63.72	-63.69
	(1.32)	(1.27)	(1.27)	(3.55)	(3.47)	(3.77)	(1.83)	(1.86)	(1.84)
	D. Q50			*E. Q75*			*F. Q90*		
Boy × Family SES			13.57			6.10			2.60
			(1.06)			(0.76)			(0.68)
Family SES		115.01	108.11		66.17	63.07		38.80	37.48
		(0.84)	(0.94)		(0.50)	(0.68)		(0.38)	(0.54)
Boy	-30.25	-32.46	-32.45	-16.56	-17.63	-17.62	-9.99	-10.50	-10.50
	(1.18)	(1.10)	(1.11)	(0.86)	(0.82)	(0.77)	(0.72)	(0.69)	(0.69)
Child & mother controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes

Note: This table reports estimates from OLS and unconditional quantile regression models, where the dependent variable is the attendance rate, from grades 3 through 8, multiplied by 100. Unconditional quantile regression estimates for 10, 25, 50, 75 and 90 percentiles are obtained using the rifreg command in Stata, developed based on Firpo et al. (2009). Columns 1 reports the raw gender gap in attendance; columns 2 include the following control variables: family SES index, dummies for maternal race and ethnicity, child's month and year of birth dummies, and birth order dummies; columns 3 further include the Boy x SES interaction. For OLS, heteroskedasticity robust standard errors are in parentheses. For unconditional quantile regression, bootstrapped standard errors with 200 replications are in parentheses.

Table A.5: Distributional Effects of Family SES on the Gender Gap: Reading Scores

	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
	A. OLS			*B. Q10*			*C. Q25*		
Boy × Family SES			1.75			13.75			4.19
			(0.21)			(0.48)			(0.29)
Family SES		29.81	28.93		27.53	20.57		26.17	24.05
		(0.12)	(0.15)		(0.34)	(0.36)		(0.21)	(0.22)
Boy	-15.02	-16.46	-16.46	-27.42	-28.99	-29.00	-15.57	-17.08	-17.08
	(0.24)	(0.21)	(0.21)	(0.52)	(0.50)	(0.49)	(0.30)	(0.31)	(0.29)
	D. Q50			*E. Q75*			*F. Q90*		
Boy × Family SES			-0.91			-3.79			-5.86
			(0.23)			(0.26)			(0.42)
Family SES		30.30	30.75		31.64	33.56		32.16	35.13
		(0.20)	(0.23)		(0.23)	(0.26)		(0.32)	(0.40)
Boy	-10.99	-12.52	-12.52	-8.68	-10.05	-10.05	-8.96	-10.20	-10.19
	(0.29)	(0.26)	(0.25)	(0.32)	(0.27)	(0.27)	(0.35)	(0.37)	(0.36)
Child & mother controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes

Note: This table reports estimates from OLS and unconditional quantile regression models, where the dependent variable is the standardized reading score, from grades 3 through 8, multiplied by 100. Unconditional quantile regression estimates for 10, 25, 50, 75 and 90 percentiles are obtained using the rifreg command in Stata, developed based on Firpo et al. (2009). Columns 1 reports the raw gender gap in attendance; columns 2 include the following control variables: family SES index, dummies for maternal race and ethnicity, child's month and year of birth dummies, and birth order dummies; columns 3 further include the Boy x SES interaction. For OLS, heteroskedasticity robust standard errors are in parentheses. For unconditional quantile regression, bootstrapped standard errors with 200 replications are in parentheses.

Table A.6: Distributional Effects of Family SES on the Gender Gap: Math Scores

	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
	A. OLS			*B. Q10*			*C. Q25*		
Boy × Family SES			2.34			3.61			-1.20
			(0.21)			(0.52)			(0.29)
Family SES		30.07	28.88		28.20	26.37		27.38	27.99
		(0.12)	(0.15)		(0.35)	(0.41)		(0.20)	(0.25)
Boy	3.37	1.91	1.90	-7.22	-8.97	-8.98	1.79	0.22	0.22
	(0.24)	(0.21)	(0.21)	(0.51)	(0.50)	(0.48)	(0.33)	(0.31)	(0.31)
	D. Q50			*E. Q75*			*F. Q90*		
Boy × Family SES			-0.56			2.28			6.68
			(0.24)			(0.29)			(0.38)
Family SES		29.98	30.27		32.05	30.90		33.02	29.64
		(0.18)	(0.23)		(0.22)	(0.25)		(0.27)	(0.33)
Boy	6.11	4.63	4.63	8.11	6.75	6.75	10.53	9.30	9.29
	(0.27)	(0.25)	(0.24)	(0.30)	(0.27)	(0.27)	(0.41)	(0.36)	(0.41)
Child & mother controls	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes

Note: This table reports estimates from OLS and unconditional quantile regression models, where the dependent variable is the standardized mathematics score, from grades 3 through 8, multiplied by 100. Unconditional quantile regression estimates for 10, 25, 50, 75 and 90 percentiles are obtained using the rifreg command in Stata, developed based on Firpo et al. (2009). Columns 1 reports the raw gender gap in attendance; columns 2 include the following control variables: family SES index, dummies for maternal race and ethnicity, child's month and year of birth dummies, and birth order dummies; columns 3 further include the Boy x SES interaction. For OLS, heteroskedasticity robust standard errors are in parentheses. For unconditional quantile regression, bootstrapped standard errors with 200 replications are in parentheses.

Table A.7: Determinants of the Gender Gap throughout the Distribution: Attendance

	OLS			Q10			Q25			Q50			Q75			Q90		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Boy × Family SES	25.86	19.07		63.24	46.67		30.03	24.07		13.56	9.39		6.09	3.18		2.60	0.52	
	(1.26)	(1.47)		(3.57)	(4.31)		(1.73)	(1.91)		(1.07)	(1.22)		(0.74)	(0.87)		(0.71)	(0.81)	
Boy × School Quality		10.77	15.83		21.99	34.38		8.99	15.38		5.77	8.26		3.86	4.70		2.94	3.08
		(1.45)	(1.41)		(3.99)	(3.82)		(2.00)	(2.10)		(1.34)	(1.32)		(0.95)	(0.94)		(0.86)	(0.84)
Boy × Neighborhood SES		4.73	10.85		15.55	30.55		4.59	12.33		3.69	6.71		2.76	3.78		1.78	1.95
		(1.53)	(1.45)		(4.29)	(4.17)		(2.06)	(2.20)		(1.32)	(1.39)		(1.00)	(0.86)		(0.73)	(0.81)
Family SES	106.66	110.12	119.80	192.96	201.39	225.10	150.63	153.66	165.89	95.60	97.72	102.49	57.28	58.76	60.37	35.14	36.19	36.46
	(0.96)	(1.02)	(0.78)	(2.91)	(3.57)	(3.23)	(1.59)	(1.79)	(1.44)	(0.91)	(1.03)	(0.80)	(0.63)	(0.72)	(0.57)	(0.54)	(0.60)	(0.44)
School Quality	39.83	34.37	31.80	83.47	72.32	66.03	53.37	48.82	45.57	31.07	28.14	26.88	15.84	13.88	13.46	8.46	6.97	6.90
	(0.74)	(0.99)	(0.98)	(2.07)	(2.47)	(2.48)	(1.21)	(1.41)	(1.42)	(0.68)	(1.05)	(1.01)	(0.52)	(0.68)	(0.72)	(0.42)	(0.62)	(0.63)
Neighborhood SES	2.64	0.20	-2.93	8.01	0.04	-7.63	8.18	5.82	1.87	3.50	1.60	0.06	0.12	-1.29	-1.81	-2.05	-2.96	-3.05
	(0.78)	(1.05)	(1.03)	(1.96)	(2.80)	(2.83)	(1.18)	(1.54)	(1.58)	(0.71)	(1.04)	(1.01)	(0.49)	(0.71)	(0.74)	(0.42)	(0.60)	(0.61)
Boy	-50.95	-50.96	-50.98	-111.29	-111.30	-111.35	-62.92	-62.92	-62.94	-32.00	-32.00	-32.01	-17.39	-17.39	-17.39	-10.37	-10.37	-10.37
	(1.26)	(1.26)	(1.26)	(3.55)	(3.48)	(3.76)	(1.84)	(1.86)	(1.84)	(1.11)	(1.09)	(1.11)	(0.85)	(0.81)	(0.76)	(0.71)	(0.69)	(0.69)

Note: This table reports estimates from OLS and unconditional quantile regression models, where the dependent variable is the attendance rate, from grades 3 through 8, multiplied by 100. Unconditional quantile regression estimates for 10, 25, 50, 75 and 90 percentiles are obtained using the rifreg command in Stata, developed based on Firpo et al. (2009). Columns 1 report the coefficients on Boy x SES, from estimation of equation (3), but further controlling for school quality and neighborhood SES. Columns 2 add Boy x school quality and Boy x neighborhood SES interactions to specification from columns 1. Columns 3 exclude the Boy x Family SES interaction term. For OLS, heteroskedasticity robust standard errors are in parentheses. For unconditional quantile regression, bootstrapped standard errors with 200 replications are in parentheses.

Table A.8: Determinants of the Gender Gap throughout the Distribution: Reading

	OLS			Q10			Q25			Q50			Q75			Q90		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Boy × Family SES	1.74	0.63		13.74	9.30		4.18	2.61		-0.91	-1.17		-3.79	-3.34		-5.86	-5.05	
	(0.20)	(0.24)		(0.46)	(0.50)		(0.29)	(0.31)		(0.23)	(0.27)		(0.26)	(0.34)		(0.36)	(0.43)	
Boy × School Quality		1.81	1.98		5.85	8.33		2.06	2.75		1.07	0.76		0.16	-0.73		-0.410	-1.76
		(0.24)	(0.24)		(0.47)	(0.51)		(0.30)	(0.29)		(0.30)	(0.28)		(0.34)	(0.33)		(0.49)	(0.44)
Boy × Neighborhood SES		0.72	0.92		4.19	7.17		1.50	2.34		-0.44	-0.82		-1.13	-2.21		-1.39	-3.01
		(0.25)	(0.24)		(0.47)	(0.53)		(0.35)	(0.33)		(0.29)	(0.27)		(0.33)	(0.32)		(0.42)	(0.39)
Family SES	23.53	24.09	24.41	15.00	17.25	21.97	18.88	19.67	21.00	25.36	25.49	24.90	28.24	28.01	26.32	29.67	29.26	26.70
	(0.16)	(0.17)	(0.13)	(0.32)	(0.35)	(0.33)	(0.22)	(0.24)	(0.19)	(0.23)	(0.24)	(0.19)	(0.27)	(0.29)	(0.21)	(0.34)	(0.41)	(0.28)
School Quality	11.84	10.92	10.84	11.96	9.00	7.74	11.61	10.57	10.22	11.83	11.30	11.46	11.68	11.61	12.06	12.17	12.38	13.07
	(0.12)	(0.17)	(0.16)	(0.25)	(0.30)	(0.32)	(0.16)	(0.20)	(0.22)	(0.16)	(0.23)	(0.21)	(0.20)	(0.25)	(0.24)	(0.26)	(0.35)	(0.38)
Neighborhood SES	3.07	2.70	2.59	3.42	1.28	-0.25	2.66	1.89	1.46	3.06	3.28	3.47	2.99	3.56	4.11	2.90	3.61	4.43
	(0.13)	(0.17)	(0.17)	(0.28)	(0.34)	(0.32)	(0.18)	(0.24)	(0.23)	(0.16)	(0.21)	(0.20)	(0.20)	(0.24)	(0.24)	(0.23)	(0.32)	(0.32)
Boy	-16.30	-16.31	-16.31	-28.84	-28.86	-28.86	-16.93	-16.93	-16.93	-12.37	-12.37	-12.37	-9.90	-9.90	-9.90	-10.03	-10.03	-10.03
	(0.20)	(0.20)	(0.20)	(0.51)	(0.50)	(0.49)	(0.30)	(0.30)	(0.29)	(0.27)	(0.26)	(0.25)	(0.28)	(0.27)	(0.27)	(0.34)	(0.37)	(0.36)

Note: This table reports estimates from OLS and unconditional quantile regression models, where the dependent variable is the standardize reading score, from grades 3 through 8, multiplied by 100. Unconditional quantile regression estimates for 10, 25, 50, 75 and 90 percentiles are obtained using the rifreg command in Stata, developed based on Firpo et al. (2009). Columns 1 report the coefficients on Boy x SES, from estimation of equation (3), but further controlling for school quality and neighborhood SES. Columns 2 add Boy x school quality and Boy x neighborhood SES interactions to specification from columns 1. Columns 3 exclude the Boy x Family SES interaction term. For OLS, heteroskedasticity robust standard errors are in parentheses. For unconditional quantile regression, bootstrapped standard errors with 200 replications are in parentheses.

Table A.9: Determinants of the Gender Gap throughout the Distribution: Math

	OLS			Q10			Q25			Q50			Q75			Q90		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
Boy × Family SES	2.33	0.85		3.60	2.18		-1.21	-1.90		-0.57	-1.39		2.27	0.89		6.67	4.30	
	(0.21)	(0.25)		(0.52)	(0.57)		(0.26)	(0.33)		(0.23)	(0.32)		(0.27)	(0.31)		(0.41)	(0.40)	
Boy × School Quality		1.95	2.17		1.55	2.13		0.25	-0.260		0.87	0.50		2.14	2.37		4.22	5.37
		(0.25)	(0.24)		(0.53)	(0.54)		(0.34)	(0.30)		(0.31)	(0.28)		(0.34)	(0.32)		(0.50)	(0.50)
Boy × Neighborhood SES		1.41	1.68		1.64	2.34		1.28	0.67		0.99	0.55		1.02	1.30		1.20	2.58
		(0.26)	(0.25)		(0.60)	(0.56)		(0.36)	(0.36)		(0.30)	(0.28)		(0.33)	(0.30)		(0.45)	(0.50)
Family SES	22.92	23.67	24.10	20.95	21.67	22.77	22.74	23.09	22.13	24.57	24.99	24.29	24.46	25.16	25.61	22.50	23.70	25.88
	(0.16)	(0.17)	(0.13)	(0.45)	(0.43)	(0.35)	(0.25)	(0.27)	(0.21)	(0.23)	(0.24)	(0.19)	(0.22)	(0.25)	(0.21)	(0.32)	(0.32)	(0.27)
School Quality	13.07	12.09	11.97	12.51	11.72	11.42	12.11	11.99	12.24	12.74	12.30	12.49	13.74	12.66	12.54	14.80	12.67	12.09
	(0.13)	(0.17)	(0.17)	(0.29)	(0.40)	(0.36)	(0.19)	(0.23)	(0.23)	(0.16)	(0.22)	(0.21)	(0.19)	(0.26)	(0.24)	(0.25)	(0.35)	(0.36)
Neighborhood SES	3.39	2.67	2.53	2.45	1.61	1.25	2.36	1.71	2.02	2.97	2.47	2.70	4.03	3.51	3.36	4.93	4.31	3.60
	(0.13)	(0.18)	(0.17)	(0.29)	(0.39)	(0.38)	(0.18)	(0.25)	(0.27)	(0.15)	(0.22)	(0.20)	(0.18)	(0.23)	(0.24)	(0.23)	(0.33)	(0.34)
Boy	2.08	2.07	2.07	-8.81	-8.81	-8.82	0.38	0.381	0.381	4.80	4.80	4.80	6.93	6.92	6.92	9.48	9.47	9.47
	(0.21)	(0.21)	(0.21)	(0.52)	(0.50)	(0.47)	(0.30)	(0.31)	(0.31)	(0.24)	(0.26)	(0.23)	(0.28)	(0.27)	(0.27)	(0.38)	(0.36)	(0.40)

Note: This table reports estimates from OLS and unconditional quantile regression models, where the dependent variable is the standardized mathematics score, from grades 3 through 8, multiplied by 100. Unconditional quantile regression estimates for 10, 25, 50, 75 and 90 percentiles are obtained using the rifreg command in Stata, developed based on Firpo et al. (2009). Columns 1 report the coefficients on Boy x SES, from estimation of equation (3), but further controlling for school quality and neighborhood SES. Columns 2 add Boy x school quality and Boy x neighborhood SES interactions to specification from columns 1. Columns 3 exclude the Boy x Family SES interaction term. For OLS, heteroskedasticity robust standard errors are in parentheses. For unconditional quantile regression, bootstrapped standard errors with 200 replications are in parentheses.

Table A.10: Distributional Effects of Family SES on the Gender Gap in Neonatal Health

	OLS		Q10		Q25		Q50		Q75		Q90	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
						A. Ln(Birth Weight)×100						
Boy×Family SES		-0.05		-0.98		-0.57		-0.03		0.55		0.90
		(0.05)		(0.10)		(0.07)		(0.05)		(0.05)		(0.06)
Family SES		1.67		2.42		2.07		1.70		1.22		0.87
		(0.04)		(0.09)		(0.05)		(0.04)		(0.04)		(0.04)
Boy	3.70	3.57	3.29	3.13	3.52	3.39	3.72	3.60	3.87	3.76	3.81	3.72
	(0.05)	(0.05)	(0.12)	(0.12)	(0.07)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.07)	(0.07)
Mean of Y						809.27						
# children						552,819						
						B. Health at Birth Index×100						
Boy×Family SES		-0.68		-2.51		-1.44		-0.29		0.26		0.99
		(0.26)		(0.81)		(0.31)		(0.20)		(0.18)		(0.23)
Family SES		4.83		9.30		6.49		4.07		2.58		1.23
		(0.20)		(0.60)		(0.24)		(0.14)		(0.13)		(0.16)
Boy	2.97	2.56	-3.60	-4.20	1.37	0.97	4.07	3.75	5.28	5.01	6.09	5.83
	(0.27)	(0.26)	(0.74)	(0.77)	(0.33)	(0.31)	(0.21)	(0.23)	(0.20)	(0.18)	(0.25)	(0.21)
Mean of Y						0.61						
# children						552,025						

Note: This table reports estimates from OLS and unconditional quantile regression models, where the dependent variables are log birthweight (in grams) and a neonatal health index at birth based on a principal component analysis of all birth outcomes, including birthweight in grams, gestational age in weeks, one and five minutes Apgar scores, indicator for adequate prenatal care, indicator for maternal health problems in pregnancy, indicator for congenital disorders, indicator of labor and delivery complications, and indicator for abnormal conditions at birth. Unconditional quantile regression estimates for 10, 25, 50, 75 and 90 percentiles are obtained using the rifreg command in Stata, developed based on Firpo et al. (2009). Columns 1 report the raw gender gap in either log birth weight or health index at birth; columns 2 include the following additional control variables: Boy x SES interaction, family SES index, dummies for maternal race and ethnicity, child's month and year of birth dummies, and birth order dummies. Abnormal conditions is a dummy variable equal to one if any of the following conditions is observed: anemia; birth injury; fetal alcohol syndrome; hyaline membrane disease; meconium aspiration syndrome; assisted ventilation <30 minutes; assisted ventilation>30 minutes; seizure; or other specified abnormal conditions. Mother health issues during pregnancy is equal to one if the mother suffered from any of a large set of chronic or pregnancy-related disorders (anemia; cardiac disease; acute or chronic lung disease; diabetes; genital herpes, hydramnios/oligohydramnios; hemoglobinopathy; chronic hypertension; pregnancy associated hypertension; eclampsia; incompetent cervix; previous infant 4000+ grams; previous preterm or small for gestational age infant; renal disease; RH sensitization; uterine bleeding; other specified health problem) during pregnancy or delivery. Congenital anomaly indicator is equal to one if a child has been diagnosed with any of a large set of congenital conditions :anencephalus, spina bifida, hydrocephalus, microcephalus, other central nervous system anomalies, head malformations, other circulatory/respiratory anomalies, rectal atresia, esophageal fistula, gastroschisis, other gastrointestinal anomalies, malformed genitalia, renal agenisis, other urogenital anomalies, cleft lip, dactyly issues, club foot, diaphragmatic hernia, other musculoskeletal anomalies, Downs Syndrome, or other chromosomal anomalies. Complications of labor and delivery indicator is equal to one if the birth suffered from any of the following conditions: fever, heavy meconium, premature rupture of membranes, abruptio placenta, placenta previa, other excessive bleeding, seizures during labor, precipitous labor, prolonged labor, dysfunctional labor, breech, cephalopelvic disproportion, cord prolapse, anesthetic complications, or fetal distress. Prenatal care adequacy is defined according to the Kessner Adequacy of Prenatal Care Utilization index (APCU), which is equal to one if the mother received standard prenatal care services during pregnancy. For OLS, heteroskedasticity robust standard errors are in parentheses. For unconditional quantile regression, bootstrapped standard errors with 200 replications are in parentheses.

Table A.11: Exogeneity of Gender

	(1)	(2)	(3)	(4)
	Outcome: Boy*100			
	Full population		Empirical sample	
SES index	0.045	-0.004	0.694	0.503
	(0.033)	(0.036)	(0.067)	(0.075)
Mean of Y	51.268		50.532	
Observations	939,810	939,810	552,819	552,819
Controls	No	Yes	No	Yes

Note: This table reports estimates from an OLS specification that regresses an indicator variable, multiplied by 100, for whether a child is a boy on family SES. Columns 1 and 2 use the full sample of births in the state of Florida. Columns 3 and 4 use the sample of births matched to public schooling records. Columns 1 and 3 do not include any additional controls. Columns 2 and 4 include controls for birth month, birth year, birth order, and maternal race/ethnicity dummies. Heteroskedasticity robust standard errors are in parentheses.

Table A.12: Relationship between Intermediate Outcomes and High School Outcomes

	Attendance	Attendance squared	Attendance cubed	Math	Math squared	Math cubed	Reading	Reading squared	Reading cubed
				Panel A. High school dropout					
Outcome: High school dropout	-11.980	-0.265	0.050	-5.234	0.206	0.422	-1.798	0.331	0.228
Mean of Y = 16.5; N = 144,945	(0.164)	(0.139)	(0.018)	(0.224)	(0.083)	(0.030)	(0.225)	(0.078)	(0.035)
Marginal effect at 10th percentile		-11.121			-3.900			-1.556	
Marginal effect at 25th percentile		-11.775			-5.108			-1.953	
Marginal effect at 50th percentile		-12.115			-5.190			-1.757	
Marginal effect at 75th percentile		-12.266			-4.404			-1.063	
Marginal effect at 90th percentile		-12.335			-3.096			-0.054	
				Panel B: High school graduation					
Outcome: High school graduation	16.820	1.515	0.018	9.541	-0.715	-0.673	3.977	-0.656	-0.460
Mean of Y = 70.7; N = 144,945	(0.172)	(0.082)	(0.009)	(0.259)	(0.095)	(0.035)	(0.260)	(0.088)	(0.039)
Marginal effect at 10th percentile		13.244			8.340			3.459	
Marginal effect at 25th percentile		15.759			9.739			4.276	
Marginal effect at 50th percentile		17.659			9.407			3.895	
Marginal effect at 75th percentile		18.852			7.706			2.510	
Marginal effect at 90th percentile		19.569			5.247			0.486	

Note: This table reports coefficients and marginal effects from OLS estimation of equation 6. We regress an indicator for high school dropout (Panel A) or high school graduation (Panel B) on cubic polynomials of standardized attendance, mathematics test scores, and reading test scores. Each panel reports estimates from a separate regression with 9 displayed variables and additional controls. Outcome variables are multiplied by 100. Each row reports marginal effects for the three intermediate inputs (attendance, mathematics, reading) at the 10th, 25th, 50th, 75th, and 90th percentiles. Additional controls include: SES index, and indicators for gender, race/ethnicity, month and year of birth, and birth order. Heteroskedasticity robust standard errors are in parentheses.

B Details of Bounding Procedure

Here we provide details regarding the procedure to bound the effects of differential sex selection into the analytical sample. In the full sample of births and in our analytical sample, we divide each sample into quintiles. For each of those quintiles we compute the difference in fraction of boys across the two samples. These five values are (starting from the first quintile) 1.4 percentage points, 1.3 percentage points, 1.2 percentage points, 0.4 percentage points, and -0.2 percentage points, implying excess boys in the bottom four quintiles and excess girls in the top quintile. Using our analytical sample we compute outcomes of girls at given percentiles for the bottom four quintiles and for boys in the top quintile. For example, since in the bottom quintile we observe 1.4 percentage points excess of girls we compute their outcomes at 1.4th and 98.6th percentile. We next define upper bound samples as females from the bottom four quintiles and males from the top quintile with outcomes greater than the upper percentile values defined in the prior step. Conversely, we define lower bound samples as females from the bottom four quintiles and males from the top quintile with outcomes less than the lower percentile values defined in the prior step. Finally, we run two regressions: (1) equation (3) using a sample excluding the upper bound as defined above; (2) equation (3) using a sample excluding the lower bound as defined above. The bounds in Figure A.2 report 95 percent confidence intervals for the coefficient ($\text{Boy}_i \times \text{SES}_i$) from these regressions.