

Gender Achievement Gaps in U.S. School Districts

AUTHORS

Sean F. Reardon

Stanford University

Erin M. Fahle

Stanford University

Demetra Kalogrides

Stanford University

Anne Podolsky

Learning Policy Institute

Rosalía C. Zárate

Stanford University

ABSTRACT

In the first systematic study of gender achievement gaps in U.S. school districts, we estimate male-female test score gaps in math and English Language Arts (ELA) for nearly 10,000 school districts in the U.S. We use state accountability test data from third through eighth grade students in the 2008-09 through 2014-15 school years. The average school district in our sample has no gender achievement gap in math, but a gap of roughly 0.23 standard deviations in ELA that favors girls. Both math and ELA gender achievement gaps vary among school districts and are positively correlated – some districts have more male-favoring gaps and some more female-favoring gaps. We find that math gaps tend to favor males more in socioeconomically advantaged school districts and in districts with larger gender disparities in adult socioeconomic status. These two variables explain about one fifth of the variation in the math gaps. However, we find little or no association between the ELA gender gap and either socioeconomic variable, and we explain virtually none of the geographic variation in ELA gaps.

VERSION

June 2018

Suggested citation: Reardon, S.F., Fahle, E.M., Kalogrides, D., Podolsky, A., & Zárate, R.C. (2018). Gender Achievement Gaps in U.S. School Districts (CEPA Working Paper No.18-13). Retrieved from Stanford Center for Education Policy Analysis: <http://cepa.stanford.edu/wp18-13>

Gender Achievement Gaps in U.S. School Districts

Sean F. Reardon (Stanford University)

Erin M. Fahle (Stanford University)

Demetra Kalogrides (Stanford University)

Anne Podolsky (Learning Policy Institute)

Rosalía C. Zárate (Stanford University)

Version: June, 2018

Direct correspondence to Sean F. Reardon, Stanford University, 520 Galvez Mall, #536, Stanford, CA 94305. Email: sean.reardon@stanford.edu. This research was supported by grants from the Institute of Education Sciences (R305D110018 and R305B090016), the Spencer Foundation (Award #201500058), and the William T. Grant Foundation (Award #186173) to Stanford University (Sean F. Reardon, Principal Investigator). We thank Richard DiSalvo and Joseph Van Matre for excellent research assistance. Some of the data used in this paper was provided by the National Center for Education Statistics (NCES). The opinions expressed here are ours and do not represent views of NCES, the Institute of Education Sciences, or the U.S. Department of Education.

Sean F. Reardon (sean.reardon@stanford.edu), Erin M. Fahle (efahle@stanford.edu), Demetra Kalogrides (dkalo@stanford.edu), and Rosalía C. Zárate (rzarate@stanford.edu), Stanford University, 520 Galvez Mall, Stanford, CA 94305, (650.736.8517); Anne Podolsky (apodolsky@learningpolicyinstitute.org), Learning Policy Institute, 1530 Page Mill Road, Suite 200, Palo Alto, CA 94304 (650.332.9797).

Gender Achievement Gaps in U.S. School Districts

Abstract:

In the first systematic study of gender achievement gaps in U.S. school districts, we estimate male-female test score gaps in math and English Language Arts (ELA) for nearly 10,000 school districts in the U.S. We use state accountability test data from third through eighth grade students in the 2008-09 through 2014-15 school years. The average school district in our sample has no gender achievement gap in math, but a gap of roughly 0.23 standard deviations in ELA that favors girls. Both math and ELA gender achievement gaps vary among school districts and are positively correlated – some districts have more male-favoring gaps and some more female-favoring gaps. We find that math gaps tend to favor males more in socioeconomically advantaged school districts and in districts with larger gender disparities in adult socioeconomic status. These two variables explain about one fifth of the variation in the math gaps. However, we find little or no association between the ELA gender gap and either socioeconomic variable, and we explain virtually none of the geographic variation in ELA gaps.

1. Introduction

Most national studies find that, on average, males outperform females on math tests and females outperform males on reading or English Language Arts (ELA) tests in the U.S. (Chatterji, 2006; Cimpian, Lubienski, Timmer, Makowski, & Miller, 2016; Fryer & Levitt, 2010; Husain & Millimet, 2009; Lee, Moon, & Hegar, 2011; Penner & Paret, 2008; Robinson & Lubienski, 2011; Sohn, 2012). These gender achievement gaps vary among states (Hyde, Lindberg, Linn, Ellis, & Williams, 2008; Pope & Sydnor, 2010), but there is little systematic research on variation in the gaps at a smaller geographic scale. Recent studies on the relationship between socioeconomic status and gender achievement, however, provide evidence that suggests gender achievement gaps may differ substantially among local communities. In particular, they indicate that community and family socioeconomic contexts differentially affect male and female academic achievement and educational attainment (Autor, Figlio, Karbownik, Roth, & Wasserman, 2016, 2017; Chetty, Hendren, Lin, Majerovitz, & Scuderi, 2016b). Thus, the large variability in local socioeconomic contexts within the U.S. may produce variation in gender achievement gaps when measured locally.

We do two things in this paper. First, we provide a high-resolution description of the patterns of gender differences in academic performance across the U.S., using scores on roughly 260 million standardized tests taken by public school students. We estimate the mean math and ELA test scores for male and female students for each of roughly ten thousand U.S. school districts in grades three through eight from the 2008-09 to 2014-15 school years. These data enable us to estimate male-female test scores gaps, as well as changes in the gaps over grades and cohorts within districts, providing a description of gender differences in academic performance at an unprecedented level of detail.

Second, we investigate the associations between the district-level gender achievement gaps and (1) local socioeconomic conditions and (2) local gender disparities in adult income, educational attainment, and occupation. We use a composite measure of the average socioeconomic status (SES) of

adults living in a school district (computed from median household income, average adult educational attainment, household poverty rates, and other measures) to determine the robustness of the association found in prior literature across the U.S. We then examine whether there is an association between gender achievement gaps and a composite measure of local adult gender disparities in SES (computed from male-female differences in individual income, educational attainment, and other measures) – a coarse proxy measure of local gendered norms, expectations, and role models. We examine the associations of these variables with both math and ELA achievement gaps to understand whether SES and local socioeconomic gender inequality influence gender achievement gaps in subject-specific ways or similarly for both subjects.

Our data include estimates of gender achievement gaps in math and ELA for nearly 10,000 school districts in the U.S. The average school district has no gender achievement gap in math; the average ELA gap is roughly -0.23 standard deviations (about three quarters of a grade level in favor of females). Both math and ELA gender gaps vary among school districts and are positively correlated – districts in which female students’ average math scores are higher than males’ tend to be districts where females’ average ELA scores are much higher than males’ scores, and vice versa. Further, we find that math gaps tend to favor males more in socioeconomically advantaged school districts and in districts with larger gender disparities in individual SES. These two variables explain about one fifth of the variation in the math gaps. However, the associations between these variables and ELA gender gap are small and inconsistent across models; socioeconomic variables explain virtually none of the geographic variation in ELA gaps.

2. Background

Previous research demonstrates the existence of math and ELA gender achievement gaps in the U.S. National studies using the Early Childhood Longitudinal Study-Kindergarten (ECLS-K) data find that a significant average math achievement gap of approximately one tenth of a standard deviation (depending

on the gap measure and the sample) in favor of males emerges by the end of kindergarten. This gap grows through fifth grade to approximately 0.15 standard deviations (Cimpian et al., 2016; Fryer & Levitt, 2010; Husain & Millimet, 2009; Lee et al., 2011; Penner & Paret, 2008; Robinson & Lubienski, 2011; Sohn, 2012). From fifth through eighth grade, the trend is reversed and the male-favoring math gap narrows (Robinson & Lubienski, 2011). ELA gaps, in contrast, favor females by approximately 0.15-0.20 standard deviations in kindergarten (Chatterji, 2006; Fryer & Levitt, 2010; Husain & Millimet, 2009; Robinson & Lubienski, 2011). The ELA gap narrows modestly (becomes less female-favoring) through fifth grade, but widens again by eighth grade (Robinson & Lubienski, 2011). These gaps have changed little in recent decades (Fahle & Reardon, 2018).

These national findings, however, mask significant variation in gender achievement gaps among states. Using achievement test data collected from 10 states, Hyde and colleagues (2008) report state-level gender gaps in performance on second through eleventh grade math assessments. Their results show that the male-female math gaps vary among states and grades but are generally near zero (they range from -0.13 SD to 0.10 SD). Using the National Assessment of Educational Progress (NAEP) data, Pope and Sydnor (2010) investigate the ratios of male-to-female students scoring above the 95th percentile in math and reading in eighth grade pooling data from the 2000, 2003, and 2005 assessments. Their results align with national studies, showing math gaps generally favor males and reading gaps favor female. They also find considerable variation in these upper tail ratios in both subjects. In math, they vary from 0.81 in Hawaii (indicating that 45 percent of high-scoring students were male; 55 percent were female) to 2.07 in Kentucky (67 percent of high-scoring students were male; 33 percent were female). In reading, the upper tail male-female ratio varies from 0.57 in Massachusetts (indicating that 36 percent of high-scoring students were male; 64 percent were female) to 0.22 in Utah (18 percent of high-scoring students were male; 82 percent were female). Pope and Sydnor further note that the male-female ratio in math is strongly negatively correlated with the male-female ratio in reading. In states where males are

overrepresented among high achieving students in math, females tend to be overrepresented among high achieving students in reading.¹

This evidence of variation raises the question: Why might gender achievement gaps differ across geographic contexts? Prior research – discussed below – suggests that both the gender stereotypes and the availability of socioeconomic resources within a community shape gender disparities in academic interests and achievement among children. Insofar as these factors vary across local contexts, we might expect the gender achievement gaps to also vary.

2.1 Gender Stereotypes

Gender stereotypes encapsulate conventional beliefs about the household roles, expected behavior, and academic talents of males and females. Traditional conservative gender stereotypes in the U.S.² generally maintain that men should be the primary breadwinners, while women should be the primary homemakers; that males are assertive, while females are demure; and that males are talented in math and science, while females are talented in languages. When widely accepted in a community, such stereotypes may affect male and female students' personal beliefs, interests, or actions. There is evidence that children become aware of gender stereotypes as early as second grade (Cvencek, Meltzoff, & Greenwald, 2011; Gunderson, Ramirez, Levine, & Beilock, 2012) and that their educational opportunities can be impeded by negative stereotypes.

In particular, stereotypes may contribute to shaping students' beliefs about their academic capability (Eccles, Jacobs, & Harold, 1990; Eccles, Wigfield, Harold, & Blumenfeld, 1993; Jacobs et al., 2002), their interest in different subjects (Cech, 2013; Charles & Bradley, 2009), and their academic

¹ The negative correlation that they observe between the gender gaps in representation in the upper tails of states' math and reading distributions, however, is not evident in other measures of gender achievement gaps. We compute gender gaps between the means of the male and female NAEP test score distributions in each state; the correlations between mean math and reading state gender gaps are generally positive and small (analyses not shown here). NAEP data can be retrieved here: <https://www.nationsreportcard.gov/ndecore/landing>.

² These stereotypes have remained relative stable over the past 30 years, despite the large cultural shifts in women's roles (Haines, Deaux, & Lofaro, 2016).

performance (Ambady, Shih, Kim, & Pittinsky, 2001; Spencer, Steele, & Quinn, 1999; Tomasetto, Alparone, & Cadinu, 2011). Female students may experience stereotype threat in math, resulting in lower test scores that reinforce the negative stereotype (Ambady et al., 2001; Tomasetto et al., 2011). Stereotypes may also be reinforced by parents' or teachers' differential encouragement of male and female children to pursue subject-specific activities (Eccles et al., 1990; Upadyaya & Eccles, 2015; Witt, 1997). But, interestingly, parents' rejection of these stereotypes can also moderate their negative effects: Tomasetto and coauthors (2011) show that the performance of female students whose mothers rejected the "male-math" stereotype did not decrease under stereotype threat.

Therefore, the extent to which community attitudes endorse these gender stereotypes may produce variation in gender performance across contexts. There is some evidence that gender stereotypes or norms differ regionally or among states in the U.S., but no large scale research on the extent to which they vary among smaller local communities (Carter & Borch, 2005; Kågesten et al., 2016). At the state level, there is evidence that stereotypes are associated with gender differences in academic performance. Pope and Sydnor (2010) examine the associations between stereotypical gender achievement disparities among high-performing students (the gender gap stereotype index) and adults' and children's gender stereotypes. They show that adults' stereotypes about gender roles, as measured by the General Social Survey, explain up to 40% of the regional variation in the gender disparities among high performing students – census divisions with more traditional stereotypes about gender roles had larger, more stereotypical achievement gaps. They further replicate these findings using student survey questions on NAEP, finding a positive association between students' self-reported agreement with the statement "math is for boys" and the gender gap stereotype index.

2.2 Parental Resources

Some research suggests that parental socioeconomic status and education may influence the development of gender differences in performance among children through parental spending. Although

there is not strong evidence that parents spend more money on male or female children (Hao & Yeung, 2015), there is evidence that parents invest their resources (time and money) in their children in gendered ways (Raley & Bianchi, 2006). For example, parents engage in more reading, storytelling, and verbal activities with their female children as early as 9 months of age (Baker & Milligan, 2013), but believe that their sons are more talented in science and math (Raley & Bianchi, 2006). These gendered patterns of investment may arise from parents' own gendered stereotypes or because broader social norms lead children to develop gendered interests, which parents then respond to and reinforce. Either way, parents' investment in, and support of, gendered activities may create or reinforce children's gender stereotypical interests, identities or skills.^{3,4}

However, the variability of parental resources and of spending on children may lead to variability in the extent to which gender stereotypes are reinforced. Affluent, highly educated parents spend more money and more time with their children than their peers (Dotti Sani & Treas, 2016; Duncan & Murnane, 2011; Guryan, Hurst, & Kearney, 2008; Hao & Yeung, 2015; Kornrich, 2016; Ramey & Ramey, 2010). Therefore, if investments are gendered and if they exacerbate children's gendered interests/identities/skills, then greater investments of rich families may lead to greater gender differences in children's interests/identities/skills. As a result, gender achievement gaps may be larger and more stereotypically patterned in higher-SES communities.

Empirical evidence to some extent supports this hypothesis. Pope and Sydnor (2010) find that states with higher median income have more stereotypical upper tail gender achievement gaps in math

³ Alternatively, sociologists have hypothesized that children in affluent families have the opportunity to indulge in their gendered interests which may exacerbate gender differences (Charles, Harr, Cech, & Hendley, 2014). In other words, affluent students are able to pursue academic activities that align with their gendered interests (whereas poor students must make decisions more economically), such that gender differences in achievement will be magnified among affluent students.

⁴ Gender differences in parental spending, however, may not all be stereotypical – they invest in different aspects of their children's education (unrelated to field). In particular, parents tend to have higher educational expectations for their daughters and invest more time in educational activities with their daughters, but are more involved in school activities for their sons and save more for their sons' college education (Raley & Bianchi, 2006).

and ELA; however, they do not find significant associations for parental education. Penner and Paret (2008) find that the achievement gap between the highest achieving males and females in math is greatest for students from families with high parental education. Lubienski, Robinson, Crane, and Ganley (2013) find that gender gaps in mathematics performance on the ECLS-K are larger among high-SES than low-SES students beginning in third grade. Together, these findings suggest that higher socioeconomic status exacerbates gender achievement gaps, leading to more male-favoring gaps in math and, to some extent, to more female-favoring gaps in ELA.

A competing hypothesis about the influence of SES suggests a different possible pattern. Trivers-Willard (1973) contends that in poorer conditions, including lower socioeconomic conditions, parents will invest more in their daughters because in such contexts daughters will have higher returns to education (and higher likelihood of finding a high-status spouse) compared with sons. In contrast, in better conditions, parents will invest more in their sons because they have higher potential for economic success than their daughters (Trivers & Willard, 1973).

Again, there is some evidence in support of this hypothesis: sons of higher status fathers are more likely to attend private school than daughters (Hopcroft & Martin, 2016) and tend to achieve higher degrees of educational attainment than daughters (Hopcroft, 2005; Hopcroft & Martin, 2016). Conversely, daughters from lower-income families are more likely to attend private school than sons (Hopcroft & Martin, 2016) and more likely to have higher educational attainment than sons (Cox, 2003).⁵ Among children from low-income families and those raised by a single-parent or a working mother, male students have lower average academic, behavioral, and economic outcomes than females, relative to the gender differences among children from more advantaged families (Autor et al., 2017; Bertrand & Pan,

⁵ The types of differential investments parents make in male and female children may also vary by social class (Hao & Yeung, 2015). Hao and Yeung find that although parents invest more in girls than boys at all levels of SES, lower SES parents invest more in school-related (tuition, school supplies, tutoring) and socio-cultural expenditures (drawing, music, sports, community activities, toy/presents, vacations), whereas high-SES parents invest more in status-signaling expenditures (clothes, shoes, cars) for female children.

2013; Chetty et al., 2016b; Fan, Fang, & Markussen, 2015). Entwisle, Alexander, & Olson (2007) find that males receiving meal subsidies perform lower on reading tests than similar females, in part because of parents' lower expectations for males' school achievement. In another study, the same authors find that the math reasoning skills of male students were more strongly influenced by the education level and median household income in the neighborhood than were the skills of their female peers (Entwisle, Alexander, & Olson, 1994). Moreover, recent evidence suggests that living in high-poverty and high-crime communities more negatively affects males' achievement than females' achievement (Chetty & Hendren, 2016; Chetty, Hendren, Lin, Majerovitz, & Scuderi, 2016a).

3. Research Aims & Framework

Given the lack of analyses of gender achievement gaps at geographies smaller than states and their association with local socioeconomic conditions, the primary goal of this paper is to provide detailed information about local patterns of male-female test score gaps. First, we provide a description of the patterns of gender differences in academic performance among nearly 10,000 U.S. school districts; our data span 6 grades and 7 years, covering 12 unique student cohorts. Of particular interest is the joint distribution—the variances and the covariance—of gender achievement gaps in math and ELA; this describes how gaps reflect subject-specific gender stereotypes or favor one gender over the other.

Figure 1 presents a stylized illustration of the dimensions of this joint distribution. School districts will fall into one of the four quadrants of the figure, each of which represents a different stereotypical or gender-favoring average pattern. Districts in the upper left quadrant have stereotypical gender gap patterns – males outperform females in math, on average (positive math gap), and females outperform males in ELA (negative ELA gap). In contrast, districts in the lower right quadrant have gender achievement patterns that are opposite in direction to common stereotypes. Districts in the lower left

quadrant have gender gaps favoring female students in both subjects, while those in the upper right quadrant are places where male students outperform females in both subjects.

[Figure 1 here]

Once we estimate the math and ELA gender gaps in each school district, we can plot them on a figure similar to Figure 1, examining how the location of districts in the figure varies by grade, cohort, and local socioeconomic characteristics. The correlation between the math and ELA gender gaps in this figure will illustrate the extent to which districts vary primarily along the stereotype dimension (the northwest-southeast dimension) or along the gender-favoring dimension (the southwest-northeast dimension). Variation or change along the stereotype dimension indicates that districts differ in the extent to which gender achievement gaps conform to the conventional stereotype that males outperform females in math, and females outperform males in ELA. Variation or change along the gender-favoring dimension indicates the extent to which gender achievement patterns in both subjects are more male- or female-favoring.

Second, we investigate the associations between district-level gender achievement gaps and two aspects of local communities: average adult socioeconomic status and gender disparities in individuals' income, educational attainment, and occupations. The latter serves as a proxy measure of local gender role models, norms, stereotypes, and expectations.⁶ Our goal is to understand whether gender achievement gaps vary systematically with these district characteristics, and to classify that variation along the dimensions in Figure 1. In other words, we seek to answer the question: Are gaps more/less

⁶ Ideally, we would have assessed whether gender achievement gaps are related to a direct measure of stereotypes, such as a survey; however, this type of data is unavailable at the local level. As a proxy, researchers have used composites of differences between males and females in their economic participation and opportunity, educational attainment, health and survival, and political participation as a measure of collective attitudes. This is most often seen in the international literature (e.g., Guiso, Monte, & Sapienza, 2008; Hyde & Mertz, 2009). In so much as adult socioeconomic gender disparities signal that a community has more traditional gender stereotypes about occupational roles, we might expect them to also indicate a community has academic gender stereotypes. However, the measure may also reflect a gender advantage. For example, large male-female disparities may indicate that males in a community have more opportunity, both economically and educationally.

stereotypical or more male/female-favoring in high SES districts and in districts with larger gender SES disparities? Through answering this question, we hope to shed light on the different possible influences of gender achievement gaps described above, and how they shape local variation in the gaps and trends in the gap over grades and cohorts.

4. Data

4.1 Achievement Data

The student achievement data used in this study comes from the *EDFacts* database, a federal database that includes aggregated state accountability test scores for every school in the U.S. The *EDFacts* data includes counts of students scoring at each state-defined proficiency level (e.g., “Below Basic,” “Basic,” “Proficient,” and “Advanced”) on state accountability tests in grades three through eight for both math and ELA. The counts are disaggregated by school, grade, year, test subject, and gender. These data are available for the 2008-09 through 2014-15 school years. In our analysis we include all public schools serving any students in grades three through eight, regardless of whether they are part of an elementary (K-8) or unified (K-12) school district. We aggregate data from all schools in a school district and use these aggregated data to measure gender achievement gaps in each school district. We focus on districts rather than schools for several reasons: school districts more closely correspond to local communities than schools; detailed socioeconomic data from the American Community Survey are available at the district but not the school level; and our estimated gender achievement gaps are much more precise for districts than for individual schools. In aggregating school data within school districts, we assign charter schools to either (1) the public school district chartering them, or (if they are not chartered by a traditional public school district) (2) the public school district in which they are geographically located. As a result, a “school district” in our analysis is a geographic unit, rather than strictly an

administrative unit, and so corresponds to the population of public school students living in a geographic region.

Using the counts of male and female students in each proficiency category within each geographic school district, we estimate the means and standard deviations of the underlying male and female test score distributions in each district using the heteroskedastic ordered probit (HETOP) model introduced by Reardon, Shear, Castellano, and Ho (2016).^{7,8} We link these estimates to a common scale and standardize them relative to the student-level national distribution of scores within their respective subject, grade, and year, using the methods described by Reardon, Kalogrides, and Ho (2017).

There are roughly 12,000 school districts serving grades three through eight in the U.S.; the data allow us to estimate both male and female mean achievement in at least one grade-year-subject for 9,799 school districts.⁹ On average, we have 132 separate grade-year-subject-gender estimates of mean achievement per district in the analytic sample, a total of almost 1.3 million observations.¹⁰ These 1.3 million observations are based on nearly 260 million test score records across the subjects, grades, and years in our sample (roughly 200 test scores on average, per district-grade-year-subject-gender observation in our analytic sample). We denote the estimated mean test score for a given gender

⁷ Note that we exclude 381 state-grade-year-subject cases (of the possible 4,284 state-subject-grade-year cases) where: (1) not all students in a state take a common test; (2) data was incomplete due to pilot testing; (3) the number of tests reported was more than 10% higher than the enrollment (typically because some students took multiple tests in a subject, such as in 8th grade math in some states); and, (4) state testing participation rates were lower than 95%. In each of these state-grade-year-subject cases, we cannot meaningfully compare the test score distributions across gender and districts. For a list of the omitted cases, see Fahle et al. (2018; Table A1).

⁸ For a complete description of the methodology used to estimate the district-subgroup-subject grade-year test score distributions see Fahle et al., 2018.

⁹ In addition to excluding a small number of state-grade-year-subject cases where state-level student participation rates were less than 95%, we exclude individual districts with lower than 95% participation rates. Such exclusions were rare except in 2015, when some states and districts experienced high non-participation rates. We also do not report estimates for district-subject-grade-year-gender cells in which there are fewer than 20 students because estimated means are very imprecise in such cases. For a detailed description of these issues, Fahle et al., 2018. Finally, we restrict our analytic sample to district-grade-year-subject cases where both male and female test score mean estimates are available.

¹⁰ Given that we have 7 years, 6 grades, 2 subjects, and 2 genders, the maximum possible number of estimates per district is 168.

subgroup s in district d , subject b , grade g , and year y as $\hat{\mu}_{sbgyd}$. Summaries of the district test score mean estimates by gender and male-female differences in the means are shown in Table 1.¹¹

[Table 1 here]

Table 1 shows that, on average across districts, female test score means in ELA are higher than male test score means in every grade and year in our sample (the average difference across grades and years is -0.23SD, or roughly two-thirds of a grade-level). The difference in the average district gender means is larger (in favor of females) both in later grade and later years. In math, districts' male and female student mean test scores are nearly equal on average across grades and years (the average male-female mean difference over grades and years is -0.01SD). In the early grades, the male means are higher than the female means, but by later grades that pattern is reversed. Over time, the male-female difference in math test score means has stayed relatively constant.

4.2 Covariate Data

We include two primary categories of covariates as correlates of district-level male-female achievement gaps: (1) average socioeconomic characteristics of all adults in the district; and (2) differences in individual socioeconomic status between males and females in the district. We construct these measures of socioeconomic averages and differences from the 2006-2010 Education Demographic and Geographic Estimate (EDGE) detailed tables.¹² EDGE tabulates the demographic and socioeconomic

¹¹ Note that the average of the male and female test score means is not exactly zero in Table 1, as might be expected given the data is standardized to the national student test score distribution within subject, grade, and year and given that male and female students each compose roughly 50% of the population in each district. In fact, the average of the district male and female test score means is closer to 0.04 (across subjects, grades, and years). This happens for two reasons: (1) the reported averages are unweighted, and (2) not all districts are represented in the sample. When weighting by the sample size in the district – to recover the average student achievement for male and females in each subject, grade, and year – the average of the male and female mean test scores (not shown) is very close to zero, as expected. The small remaining deviations from zero then result from the fact that not all districts (or students) are represented in our sample as they did not have sufficient numbers of male or female students. We report the unweighted district-average male and female means because the unit of analysis for this paper is the school district (we are focused on the gap in each district), rather the student.

¹² We use the 2006-2010 EDGE data because the later years' EDGE tabulations do not include all of the variables of interest for the relevant population (students attending public schools in the district; parents of children attending public schools in the district).

characteristics of adults who live in each school district in the U.S. using American Community Survey (ACS) data. The measures of socioeconomic characteristics include median income, educational attainment, occupation, poverty rates, unemployment rates, labor force participation rates, proportion in business/management occupations, and proportion in science occupations. These measures are available in the EDGE data for all adults, as well as separately by gender. Although data on other occupation categories are available in EDGE, we use only measures of participation in business/management and science occupations, as those are stereotypically male-dominated sectors (and the inclusion of other occupational categories did not improve the reliability of our measures).

To simplify the analyses, we construct an overall SES composite from these seven variables using principal component analysis. The factor loadings for each variable are shown in Table 2. We use the overall SES composite as our measure of average socioeconomic status in a school district. We then construct gender-specific versions of the SES composite using gender-specific variables (e.g., male income, female income, etc.) and the same factor loadings as shown in Table 2. The difference between the male and female SES composites is then our measure of the gap in SES between adult males and females in a school district.

[Table 2 here]

We also include controls for student demographics using data from the Stanford Education Data Archive (SEDA) and the Common Core of Data (CCD). The CCD is an annual survey of all public elementary and secondary schools and school districts in the U.S. SEDA uses CCD data to create district-level measures of the percent black, Hispanic, other race, and white students in public school districts, averaged over the grades and years in our sample.¹³ Table 3 provides a summary of the SES measures and

¹³ The demographic data provided in SEDA includes imputed data and will not match the raw CCD data. The multiple imputation process is described in detail in the SEDA Technical Documentation (Fahle et al., 2018).

demographic covariates from the various data sources. The average district in our sample has a male-female SES gap of one-standard deviation and is 9 percent black and 14 percent Hispanic.

[Table 3 here]

5. Methods

In this paper, we aim to provide a description of gender achievement gaps across U.S. school districts and to generate unbiased estimates of the association between district covariates and subject-specific gender achievement gaps and growth rates of achievement gaps. Complicating these aims is the issue that measurement of gender achievement gaps may be confounded by differences among the standardized tests used in different states, grades, and years. Specifically, gender achievement gaps measured using tests with more multiple-choice items (vs. constructed-response items) are more male-favoring than tests with fewer multiple-choice items (Beller & Gafni, 2000; Bielinski & Davison, 2001; DeMars, 1998; Garner & Engelhard, 1999; Lindberg, Hyde, Petersen, & Linn, 2010; Reardon, Kalogrides, Fahle, Podolsky, & Zárata, 2018).

Because state accountability tests vary in format (as well as other factors such as content that may influence the measurement of gaps), this poses an issue for generating district-level average gap estimates that are comparable across states, and possibly grades and years. Any gap comparisons will be biased by differences in test format across states, and it is not clear how biased they will be given that information on many tests' item composition is not readily available. Moreover, this complicates estimating unbiased coefficients if the item format of state accountability tests is related to the average SES levels within the state. For example, if states that have higher average SES also have tests with more multiple-choice questions, this will bias the estimated association between SES and gender achievement gaps upward, leading to the possibly erroneous conclusion that the gender gaps are more male-favoring in higher SES school districts.

To address this issue, we adopt the following procedure to purge our district gap estimates of systematic differences that might arise because of differences across states in the content or format of their tests (see Appendix A for more detail). We first residualize all the test score means within state, subject, grade, year, and gender—we subtract the statewide average score within a gender, subject, grade, and year from each district’s corresponding estimate. Note, however, that the resulting residualized gender-specific district means do not contain any information about the average magnitude of the gender mean or achievement gap within each state (because they constrain average male and female scores within each state to be zero). This will limit our ability to provide an accurate description of the variation in gender gaps across states, and will also lead to an underestimation of the true variance of test score means and gaps in the U.S. To remedy this, we add the average NAEP scores in the corresponding state-subject-grade-year-gender to the residualized state means. That is, if $\hat{\mu}_{sbgyf}$ and $\hat{\mu}_{sbgyf}^{naep}$ denote the average standardized test scores on the state test and the NAEP test, respectively, in state f for gender s in subject-grade-year $bg y$, then we compute:

$$\hat{\mu}'_{sbgyd} = \hat{\mu}_{sbgyd} - \hat{\mu}_{sbgyf} + \hat{\mu}_{sbgyf}^{naep} \tag{1}$$

The NAEP-adjusted residualized estimate of the mean, $\hat{\mu}'_{sbgyd}$, is purged of between-state differences in the tests used and therefore of bias due to differences in the content or format of those tests. We use these NAEP-adjusted residualized means in all of the analyses that follow. These estimates provide comparable measures of performance and gender achievement gaps across states, grades, and years. Note that Equation 1 implies that all between-state, -subject, -grade, and -cohort variation in the NAEP-adjusted residualized estimates comes from the information provided by the NAEP assessments; all within-state-subject-grade-cohort information comes from the state accountability tests.

A summary of the NAEP-adjusted mean residuals is shown in Table 4. The differences in the NAEP-adjusted district gender mean test scores are broadly consistent with those based on our raw means from the state accountability tests (shown in Table 1). On average across districts, female students

have higher mean test scores than male students in ELA (gap of -0.23 SD, averaged across grades/years).

In math, the differences in the district-average test score means are close to zero, but consistently favor males in nearly every grade and year. Both ELA and math male-female differences tend to favor females more in later grades and later years.

[Table 4 here]

6. Models

We fit the following model to construct estimates of the average math and ELA gaps within each district:

$$\begin{aligned}\hat{\mu}'_{sbgyd} = & [\beta_{00d} + \beta_{01d}(\text{grade} - 5.5) + \beta_{02d}(\text{cohort} - 2006.5)] \cdot \text{math} \\ & + [\beta_{10d} + \beta_{11d}(\text{grade} - 5.5) + \beta_{12d}(\text{cohort} - 2006.5)] \cdot \text{math} \cdot (\text{male} - 0.5) \\ & + [\beta_{20d} + \beta_{21d}(\text{grade} - 5.5) + \beta_{22d}(\text{cohort} - 2006.5)] \cdot \text{ela} \\ & + [\beta_{30d} + \beta_{31d}(\text{grade} - 5.5) + \beta_{32d}(\text{cohort} - 2006.5)] \cdot \text{ela} \cdot (\text{male} - 0.5) \\ & + e_{sbgyd} + r_{sbgyd}\end{aligned}$$

$$\beta_{00d} = \gamma_{000} + \mathbf{X}_d \boldsymbol{\Gamma}_{000} + u_{00d}$$

$$\beta_{01d} = \gamma_{010} + \mathbf{X}_d \boldsymbol{\Gamma}_{010} + u_{01d}$$

$$\beta_{02d} = \gamma_{020} + \mathbf{X}_d \boldsymbol{\Gamma}_{020} + u_{02d}$$

$$\beta_{10d} = \gamma_{100} + \mathbf{X}_d \boldsymbol{\Gamma}_{100} + u_{10d}$$

$$\beta_{11d} = \gamma_{110} + \mathbf{X}_d \boldsymbol{\Gamma}_{110}$$

$$\beta_{12d} = \gamma_{120} + \mathbf{X}_d \boldsymbol{\Gamma}_{120} + u_{12d}$$

$$\beta_{20d} = \gamma_{200} + \mathbf{X}_d \boldsymbol{\Gamma}_{200} + u_{20d}$$

$$\beta_{21d} = \gamma_{210} + \mathbf{X}_d \boldsymbol{\Gamma}_{210} + u_{21d}$$

$$\beta_{22d} = \gamma_{220} + \mathbf{X}_d \boldsymbol{\Gamma}_{220} + u_{22d}$$

$$\beta_{30d} = \gamma_{300} + \mathbf{X}_d \boldsymbol{\Gamma}_{300} + u_{30d}$$

$$\beta_{31d} = \gamma_{310} + \mathbf{X}_d \boldsymbol{\Gamma}_{310}$$

$$\beta_{32d} = \gamma_{320} + \mathbf{X}_d \boldsymbol{\Gamma}_{320} + u_{32d}$$

$$e_{sbgyd} \sim N(0, \hat{\omega}_{sbgyd}^2); r_{sbgyd} \sim N(0, \sigma^2); \mathbf{U}_d \sim MVN(0, \boldsymbol{\tau}^2)$$

(2)

where $\hat{\mu}'_{sbgyd}$ is the NAEP-adjusted residualized estimated mean test score for subgroup s (male or female), in subject b (math or ELA), district d , grade g , and year y ; math is an indicator variable equal to 1 if the tested subject is math; ela is an indicator variable equal to 1 if the tested subject is ELA; male is

an indicator variable equal to 1 if the tested subgroup is male; *grade* is a continuous variable indicating the tested grade; *cohort* is a continuous variable indicating the tested cohort (where cohort is defined as the tested *year* – *grade*, and so indicates the year in which a cohort was in their spring kindergarten semester); and \mathbf{X}_d is a vector of (time- and grade-invariant) district-level covariates. The $\mathbf{u}_{..d}$ are multivariate normally-distributed mean-zero residuals with variance-covariance matrix $\boldsymbol{\tau}^2$ to be estimated; r_{sbgyd} is a normally-distributed mean-zero residual with variance σ^2 to be estimated; and e_{sbgyd} is a normally-distributed mean-zero sampling error term with known variance $\hat{\omega}_{sbgyd}^2$ equal to the sampling variance of $\hat{\mu}_{sbgyd}^*$.¹⁴ Model estimation is performed using HLM.

In other words, this model uses up to 168 (2 genders, 6 grades, 7 years, 2 subjects) estimates of gender-grade-year-subject NAEP-adjusted means in each district to estimate each districts' average performance in math, the male-female gender achievement gap in math, the average performance in ELA, the male-female gender achievement gap in ELA, the growth over grades and cohorts of each of those terms, and a residual error term. The average performance and gaps ($\beta_{.0d}$) are then modeled as a function of a vector of district covariates and a residual error term indicating the difference between the true average/gap that is predicted by the covariates and the national average/gap. Similarly, the grade ($\beta_{.1d}$) and cohort slopes ($\beta_{.2d}$) of the average performance and gaps are modeled as functions of district covariates and district-specific residual error terms. We exclude the district-level error terms on the grade slopes of the gender achievement gaps in math and ELA (β_{11d}, β_{31d}) because our initial models including them indicated their variance was not statistically distinguishable from zero. That is, we cannot reject the null hypothesis that the gender gaps change at the same rate from third to eighth grade in all districts.

¹⁴ Note that we use the sampling variance of $\hat{\mu}_{sbgyd}^*$ (the residualized means) as an estimate of $\hat{\omega}_{sbgyd}^2$ (the sampling variance of the NAEP-adjusted residualized means). We do not add in the sampling variance of the NAEP state mean because the state level error is common to all districts in a state (and is similar across states, since the NAEP sample sizes are the same across states).

In our null model (Model 1), we do not include any district-level covariates, \mathbf{X}_d . From this model, we recover an estimate of the true variance in the gender achievement gaps among U.S. school districts. We then formally test whether SES and adult male-female SES disparities are associated with overall achievement, gender achievement gaps, and the growth across grades or cohorts in both of these measures by adding these measures, in \mathbf{X}_d , to the null model (Model 2). Next, we add racial composition variables to test whether the SES and male-female SES disparity associations hold after controlling for student demographics (Model 3). Finally, we estimate a variant of Model 3 that includes state-level random effects on the four β_{0d} terms (shown as Model 4). In this model, we state-mean center the district covariate vector \mathbf{X}_d ; as a result, the coefficient estimates from Model 4 are the same as we would obtain had we added state fixed effects in each of the $\beta_{..d}$ equations (but the model is more computationally efficient than the fixed effects model). Note that centering the vector \mathbf{X}_d changes the interpretation of the estimated coefficient vector $\hat{\boldsymbol{\Gamma}}_{..0}$. It now represents average within-state associations. This has the advantage of ensuring that the associations are not biased by between-state differences in the standardized tests.

7. Results

Table 5 reports selected coefficients from the fitted models. Model 1 (the null model) shows that the average district male-female math gap is 0.03 standard deviations; the average ELA gap is -0.23 standard deviations. In other words, in the average school district, there is essentially no gender achievement gap in math, but two-thirds of a grade-level difference in favor of females in ELA (one grade level in grades 3-8 is roughly equal to 0.33SD). Because we center grade and cohort at the midpoints of the grades and cohorts contained in our data (grade 5.5 and cohort 2006.5), these can be interpreted as the average gaps halfway through fifth grade for the average cohort in our sample. On average, both the math and ELA gaps change in favor of females from grade three to eight (by roughly -0.06SD in math and

-0.10SD in ELA in five grades). Across cohorts, the gaps change relatively little per year; the models imply that the average math gap changes by roughly -0.05SD (in favor of females) and by roughly 0.02SD (in favor of males) over 11 cohorts. These results are largely consistent with the less parametric patterns evident in Table 4.

[Table 5 here]

The gender gaps vary significantly among districts in both math and ELA. From Model 1, the estimate of the true variance of the ELA gaps is 0.0024 (SD = 0.049) and of the math gaps is 0.0025 (SD = 0.050). The estimated distribution of the ELA gaps (mean = -0.23; SD = 0.049) implies (assuming the gaps are normally distributed) that 95% of school districts have ELA gaps between -0.13 and -0.33 standard deviations (i.e., favoring females by between one-third and one grade level); in no district do males perform, on average, as well or higher than females in ELA. A district's gap would have to be almost 5 standard deviations from the mean ELA gap in this case. The distribution of math gaps (mean = 0.029; SD = 0.050) implies that 95% of districts have math gaps that are between -0.07 and +0.13 standard deviations, favoring males in 72% of school districts and females in 28%. In Figure 2, we map estimates of the ELA and math achievement gaps across U.S. school districts to provide a visual representation of this variation.¹⁵ In the maps, orange indicates that female students outperform male students on average, and blue the opposite; white indicates missing data. Darker shades signify larger average gaps. In both subjects, the maps confirm that there is clear variation in the gaps among and within states, although some states have less between-district variation in the gaps than others.

[Figure 2 here]

Figure 3 shows the Empirical Bayes estimate of the male-female math achievement gap plotted against the Empirical Bayes estimate of the ELA achievement gap for each district in our sample. Note

¹⁵ The maps are shown to illustrate the variation; however, note that they understate the true variance of the gaps in ELA and math as a result of shrinkage. The reliabilities of the gaps are 0.59 in both subjects from Model 1.

that the gaps fall predominantly in the upper left quadrant indicating that in most school districts, gender achievement gaps align with subject-specific gender stereotypes. In contrast, the math and ELA gaps are positively correlated: districts with more male-favoring math gaps tend to also have more male-favoring (less female-favoring) ELA gaps. This suggests that variation among districts is gender-favoring, despite the fact that gaps on average are stereotypical. Indeed, Table 6 shows that the estimated correlation between the math and ELA gaps is 0.85. In other words, in terms of the dimensions described in Figure 1, districts' gender achievement gaps vary primarily along the male/female-favoring dimension rather than the stereotype dimension.

[Figure 3 here]

Table 6 further shows that there is a moderate correlation between the average performance and the male-female gap in math (0.46), but that there is almost no association in ELA (0.05). Districts with higher math performance tend to have more male-favoring math gaps, but the average performance of students in ELA is unrelated to the size of the ELA gap.

[Table 6 here]

Model 2 in Table 5 provides estimates of the associations between the socioeconomic variables and gender achievement gaps. It shows that both district SES and male-female SES differences are positively associated with the male-female math gap. In wealthier school districts and in school districts with more economic gender inequality, math gaps favor males more, on average. The associations are weaker for the ELA gender gaps. There is no significant relationship between overall SES and the ELA gender gap in Model 2. But there is a small positive association between local gender SES disparities and the ELA gender gap: ELA gaps favor males more in communities with large gender SES disparities. Overall, these results suggest meaningful associations between math achievement gaps and both local SES and local gender SES disparities, but no or very small associations for ELA. This is evident in the proportion of

variance explained by the two SES variables (Model 2): together they explain 20% of the variance in math gaps but none of the variance in the ELA gaps.

We illustrate these results in Figures 4 and 5. Figure 4 shows the male-female achievement gaps in math plotted against the SES composite. For the ELA gaps, the slope of the line is nearly flat; the gaps are uncorrelated ($r = 0.03$) with SES, (as is evident in the multivariate models). However, in math, the slope of the line is positive ($r = 0.43$), indicating that the gap is more male favoring in high-SES places compared to low-SES places. Figure 5 shows positive relationships in both math and ELA between achievement gaps and our measure of economic gender inequality (SES composite difference). However, the relationship is much steeper ($r = 0.32$) in math (as seen in Model 3), indicating that in places with larger male-female disparities in SES among adults, math, and to a much smaller extent ELA, gaps tend to favor male students more (relative to average districts).

[Figures 4 and 5 here]

After controlling for racial composition (Model 3) or estimating the association within states (Model 4), the associations between the math gaps and SES and gender SES disparities remain statistically significant (see Table 5). In ELA, the results are inconsistent across models: the association with SES is positive and statistically significant in Model 3, and negative and statistically significant in Model 4. However, the coefficients in ELA are very small and of little practical significance. The association between gender SES disparities and ELA gender gaps is smaller and no longer statistically significant in both Models 3 and 4.

Although it is not the focus of our analysis here, Model 3 indicates that racial composition is also associated with the gender gaps. Both math and ELA gaps are more female-favoring in districts with a larger proportion of black students. Math gaps are also more female-favoring in districts with more Hispanic students, but the opposite is true for ELA gaps. Figure 6 illustrates the bivariate associations between the gender gaps in math and ELA and the proportion of students in a district who are black.

Although this suggests that within-district gender gaps are smaller (more female-favoring) among black students than among white students, the pattern in Figure 6 does not prove this. A similar pattern would result if white and black gender gaps were similar within any given district but both white and black gender gaps were negatively correlated with the proportion of black students in a district. Nonetheless, the figure does indicate that black children are, on average, in school districts where gender gaps are more female-favoring; while white children are disproportionately in school districts where gender gaps favor males more. These associations are not large, but they are quite clear in the data nonetheless. Moreover, these associations persist even after controlling for district socioeconomic characteristics (see Model 3 of Table 5).

[Figure 6 here]

Using the framework described in Figure 1, we can characterize where (in what kinds of districts) and when (in what grades/cohorts) gaps are more stereotypical vs. gender favoring. In Figure 7, we plot the stylized results from Model 3, illustrating the joint distribution of math and ELA gaps in the same two-dimensional space as in Figures 1 and 3. These stylized results are derived from the estimates in Models 1 and 3 of Table 5. In each panel of the figure, we plot the estimated 95% coverage ellipses (the ellipses in which 95% of districts lie) for two sets of estimated gaps. These illustrate the relative direction and magnitude of the differences in gaps associated with different grades, cohorts, or types of districts.

[Figure 7 here]

In the first panel of the figure (upper left), we show that gender achievement gaps favor females more in later grades than earlier ones: the ellipse representing the joint distribution of the gaps shifts down and to the left (in the more female-favoring direction) between third and eighth grade. A comparison of the implied distribution of gaps across cohorts (upper right panel) reveals that the distribution of gaps in the earliest cohort in our sample is located down and to the right relative to the last cohort in our sample. Gender gaps are smaller and less stereotypical (math gaps are less male-

favoring; ELA gaps are less female-favoring) in more recent cohorts. Finally, in districts with high average SES (lower left) or large male-female SES disparities (lower right), the distribution of gaps is shifted upward in the figure relative to poorer districts and those with smaller gender SES disparities, indicating that math gaps are more male-favoring in more socioeconomically advantaged and unequal district. Interestingly, the magnitude of the SES-related differences in achievement gap patterns is much smaller than the magnitude of the average differences in gaps across grades or cohorts.

In fitting Model 1 in Table 5, we found no statistically significant variation in the gap grade slopes among districts (which is why the random coefficients on the math and ELA gap grade slopes are omitted from Model 1). This is somewhat surprising; we might anticipate that factors that produce variation in the gaps among districts would also have cumulative effects as students age, which would be reflected as differences between districts in the gap grade slopes in our model. However, the test of between-district variance on this interaction term is a low-power test; there is likely some (modest) degree of variation among district that our models have insufficient power to detect. Indeed, we do find that the district characteristics predict changes in the gender achievement gaps from third to eighth grade. In high-SES districts, math achievement gaps grow more female-favoring than they do in lower-SES districts (Model 3); however, this pattern is reversed in the within-state models (Model 4). Further, in communities with larger male-female economic disparities, the growth in the math and ELA gaps is more male-favoring (Models 3 and 4).

The cohort slopes, in contrast, show that math gaps have changed in favor of females and ELA gaps in favor of males over the cohorts in our sample – in opposition to the common stereotypes, as illustrated in Figure 7. There is significant variance in the gap cohort slopes among districts and the change in the math and ELA gaps over cohorts is more female favoring in high-SES communities. In math, we also find that the change in gaps over cohorts is more male-favoring in communities with high male-female SES disparities but find no significant results in ELA (Model 3). These results largely hold when

looking within states (Model 4) with the exception that in high SES communities the change in the math gaps over cohorts is more male (rather than more female) favoring.

8. Discussion

No prior study has examined gender achievement gaps at the local level and with the level of detail we have here. Given this, our primary goal in this paper is to establish a set of stylized facts regarding the size, variation, and correlates of gender achievement gaps in math and ELA. Five particular patterns—and their implications—are worth noting.

First, in virtually every school district in the U.S., female students outperformed male students on ELA tests in grades 3-8 during the 2008-09 to 2014-15 school years. In the average district, the gap is roughly one-quarter of a standard deviation, though the gaps vary somewhat among districts. A quarter of a standard deviation is a substantial gap; it corresponds to roughly two-thirds of a grade level and is larger than the effects of most large-scale educational interventions. On math tests, in contrast, the gender gap in the average district is quite small – roughly 0.03 standard deviations in favor of males. Again, this varies among districts. Female students modestly outperform males in a quarter of districts; males modestly outperform females in the others. But in only a small number of districts are the gaps larger than a third of a grade level. Although the general absence of large, male-favoring math gaps would seem at odds with public perceptions of math gender achievement gaps, these patterns generally align with those estimated at the state or national level in other studies (Fryer & Levitt, 2010; Husain & Millimet, 2009; Lee et al., 2011; Penner & Paret, 2008; Pope & Sydnor, 2010; Robinson & Lubienski, 2011; Sohn, 2012).

Second, districts' math and ELA gender gaps are strongly positively correlated: school districts vary largely on the gender-favoring dimension and very little on the subject-specific stereotype dimension (as shown in Figure 1). The fact that gender gaps vary among school districts suggests that local

conditions and processes – in addition to larger societal forces – play a role in shaping them. Moreover, these local processes appear to primarily operate on the gender-favoring dimension and differ among communities largely in the extent to which they produce more male- or more female-favoring achievement patterns in both math and ELA. That implies that if these differences are driven by, say, local norms, the content of these norms must be primarily about relative general academic expectations for male and female students, rather than about subject-specific differential expectations.

Third, on average gender achievement gaps become more female-favoring from grade three to eight in both math and ELA. In third grade, male students outperform females in math in most districts and female students outperform males by roughly half a grade level on ELA tests. By eighth grade, in the average district male and female students score equally well on math tests, but females are a nearly a grade level ahead of their male classmates in ELA. It is important to note that other studies, using national data, find that this pattern reverses in high school. Male students outperform females in math and the ELA gap in favor of females is smaller in high school than in eighth grade (Fahle & Reardon, 2018). This suggests that the forces that shape gender achievement gap patterns vary during the child and adolescent developmental period as well as among local communities.

Fourth, the data are relatively silent with regard to what local processes produce these gaps – most of the variation among districts in gender achievement gaps is unaccounted for by socioeconomic and demographic district characteristics. In our models, gender gaps in math appear to be related to local socioeconomic conditions; many of the communities with the largest math achievement gaps are affluent, predominantly white, suburban communities in which adult gender employment and income disparities tend to be particularly large. This same pattern is not true of gender gaps in ELA performance, however.

Moreover, our data do not directly describe the mechanisms that account for the association between math achievement gaps and socioeconomic conditions. One possibility is that parents in affluent

and highly-educated communities invest more time and resources in their sons' educations than their daughters', and that these investments either are directed toward or particularly affect male students' math skills. Another possibility is that local gender stereotypes, norms, and expectations are correlated with socioeconomic conditions and may lead to male and female students receiving different educational opportunities or internalizing different beliefs about their capacity or roles. A third possibility is that parental and community investments in children's education magnify small gender differences in interests that develop in early childhood. If some children develop slightly stereotyped interests in early childhood and if parents and schools respond to those differential interests by encouraging children to do more of what they are interested in, small early differences in math interest may be exacerbated. This effect would be most pronounced in high-resource environments and in environments where gendered roles and status are most stereotypical – places where men work and earn much more than women.

Fifth, gender achievement gaps in grades three to eight have been trending toward gender parity over recent cohorts of students, though this trend is more pronounced in math than in ELA. Our estimates indicate that the math gap has declined by roughly 0.05 standard deviations (about a sixth of a grade level) from the 2001 to the 2012 spring kindergarten cohort. In the most recent cohorts, there is no gender gap in math in the average school district. The trend in ELA gaps has been much slower: the ELA gap declined, on average, by roughly 0.02 standard deviations across the same set of cohorts. The combination of these trends indicates that gender gaps in middle school achievement have become somewhat less pronounced and less aligned with subject-specific gender stereotypes in recent years.

It is not clear, however, what has driven the changes over time in the gender gaps. One possibility is that gender norms have changed in recent years, but there is little evidence to suggest that there has been a significant change in gender norms in the last decade. Another possibility is that the recession played a role. The recession generally lowered family incomes, and affected male workers more than female workers, thereby reducing the male-female SES disparity in many communities. Given that lower

community SES and smaller gender SES disparities are both associated with less male-favoring math achievement gaps, it is possible that the recession led to reduced gender disparities in math. One might test this hypothesis, by examining whether gender gaps changed more in communities hardest hit by the recession, but such an analysis is beyond the scope of this paper.

This paper has several limitations. One is that the tests used to measure achievement vary in format and content among states, grades, and years, and these differences may lead to differences in measured gender achievement gaps. We address this issue by using the NAEP assessments to adjust the gender-specific scores on each states' tests. This method is not perfect, however, and may not fully correct for the differences in content and format of states' tests. Another is that we do not have good measures of local norms, expectations, stereotypes, or of how boys and girls are treated in school and home and community. Because of this, we cannot rule out the potentially important influence these factors may have on gender achievement gaps that we may be unable to observe with our coarse proxy measure. Third, the patterns described here apply to grades 3 through 8. We cannot speak to the existence of gaps or trends in later grades, and how they may differ from what we see here. Finally, we find some suggestive relationships between race and gender achievement gaps but are unable to estimate gender gaps by race. Prior evidence shows that gender achievement gaps are characteristically different among students from different racial groups (e.g., Penner & Paret, 2008) and we are limited in our ability to explore that using our data.

There is clearly more work to do. Our analyses cannot account for a large amount of the variance in the gender achievement gaps among school districts. Moreover, we cannot answer why the gaps in math and ELA do not appear to be related to the same local factors, despite the strong correlation between them. Future work should consider other features of local context that may influence gender achievement gaps in both subjects, such as political representation and affiliation, teacher's stereotypes and biases, and behavioral differences between male and female students. These extensions, as well as

the collection of new data that allows for the estimation of gender gaps in high school or the estimation of gender by race/ethnicity at high resolution, will be important lines of future inquiry.

References

- Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science, 12*(5), 385–390. <https://doi.org/10.1111/1467-9280.00371>
- Autor, D., Figlio, D., Karbownik, K., Roth, J., & Wasserman, M. (2016). School Quality and the Gender Gap in Educational Achievement. *American Economic Review, 106*(5), 289–295. <https://doi.org/10.1257/aer.p20161074>
- Autor, D., Figlio, D., Karbownik, K., Roth, J., & Wasserman, M. (2017). *Family Disadvantage and the Gender Gap in Behavioral and Educational Outcomes* (NBER Working Paper Series No. 22267). *NBER Working Paper Series*. Cambridge, MA. <https://doi.org/10.3386/w22267>
- Baker, M., & Milligan, K. (2013). *Boy-Girl Differences in Parental Time Investments: Evidence from Three Countries* (NBER Working Paper Series No. 18893). Cambridge, MA. <https://doi.org/10.3386/w18893>
- Beller, M., & Gafni, N. (2000). Can Item Format (Multiple Choice vs. Open-Ended) Account for Gender Differences in Mathematics Achievement? *Sex Roles, 42*(1/2), 1–21. <https://doi.org/10.1023/A:1007051109754>
- Bertrand, M., & Pan, J. (2013). The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior. *American Economic Journal: Applied Economics, 5*(1), 32–64. <https://doi.org/10.1257/app.5.1.32>
- Bielinski, J., & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. *Journal of Educational Measurement, 38*(1), 51–77. <https://doi.org/doi:10.1111/j.1745-3984.2001.tb01116.x>
- Carter, J. S., & Borch, C. A. (2005). Assessing the Effects of Urbanism and Regionalism on Gender-Role Attitudes, 1974-1998*. *Sociological Inquiry, 75*(4), 548–563. <https://doi.org/10.1111/j.1475-682X.2005.00136.x>
- Cech, E. A. (2013). The Self-Expressive Edge of Occupational Sex Segregation. *American Journal of Sociology, 119*(3), 747–789. <https://doi.org/10.1086/673969>
- Charles, M., & Bradley, K. (2009). Indulging Our Gendered Selves? Sex Segregation by Field of Study in 44 Countries. *American Journal of Sociology, 114*(4), 924–976. <https://doi.org/10.1086/595942>
- Charles, M., Harr, B., Cech, E., & Hendley, A. (2014). Who likes math where? Gender differences in eighth-graders' attitudes around the world. *International Studies in Sociology of Education, 24*(1), 85–112. <https://doi.org/10.1080/09620214.2014.895140>
- Chatterji, M. (2006). Reading Achievement Gaps, Correlates, and Moderators of Early Reading Achievement: Evidence From the Early Childhood Longitudinal Study (ECLS) Kindergarten to First Grade Sample. *Journal Of Educational Psychology, 98*(3), 489–507. <https://doi.org/10.1037/0022-0663.98.3.489>
- Chetty, R., & Hendren, N. (2016). *The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects*. Cambridge, MA. <https://doi.org/10.3386/w23001>
- Chetty, R., Hendren, N., Lin, F., Majerovitz, J., & Scuderi, B. (2016a). Childhood Environment and Gender Gaps in Adulthood. NBER Working Paper No. 21936, *106*(5), 282–288. Retrieved from <http://www.nber.org/papers/w21936>
- Chetty, R., Hendren, N., Lin, F., Majerovitz, J., & Scuderi, B. (2016b). Childhood environment and gender gaps in adulthood. *American Economic Review, 106*(5), 282–288. <https://doi.org/10.1257/aer.p20161073>
- Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have Gender Gaps in Math Closed? Achievement, Teacher Perceptions, and Learning Behaviors Across Two ECLS-K Cohorts. *AERA Open, 2*(4), 233285841667361. <https://doi.org/10.1177/2332858416673617>
- Cox, D. (2003). Private transfers within the family: mothers, fathers, sons and daughters. In A. H. Munnell

- & A. Sundren (Eds.), *Death and dollars: The role of gifts and bequests in America* (pp. 168–197). Washington, DC: The Brookings Institution Press.
- Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math-Gender Stereotypes in Elementary School Children. *Child Development, 82*(3), 766–779. <https://doi.org/10.1111/j.1467-8624.2010.01529.x>
- DeMars, C. E. (1998). Gender Differences in Mathematics and Science on a High School Proficiency Exam: The Role of Response Format. *Applied Measurement in Education, 11*(3), 279–299. https://doi.org/10.1207/s15324818ame1103_4
- Dotti Sani, G. M., & Treas, J. (2016). Educational Gradients in Parents' Child-Care Time Across Countries, 1965–2012. *Journal of Marriage and Family, 78*(4), 1083–1096. <https://doi.org/10.1111/jomf.12305>
- Duncan, G. J., & Murnane, R. J. (Eds.). (2011). *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*. Russell Sage Foundation. <https://doi.org/10.7758/9781610447515>
- Eccles, J., Jacobs, J. E., & Harold, R. D. (1990). Gender Role Stereotypes, Expectancy Effects, and Parents' Socialization of Gender Differences. *Journal of Social Issues, 46*(2), 183–201. <https://doi.org/10.1111/j.1540-4560.1990.tb01929.x>
- Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and Gender Differences in Children's Self- and Task Perceptions during Elementary School. *Child Development, 64*(3), 830. <https://doi.org/10.2307/1131221>
- Entwisle, D. R., Alexander, K. L., & Olson, L. S. (1994). The Gender Gap in Math: Its Possible Origins in Neighborhood Effects. *American Sociological Review, 59*(6), 822. <https://doi.org/10.2307/2096370>
- Entwisle, D. R., Alexander, K. L., & Olson, L. S. (2007). Early schooling: The handicap of being poor and male. *Sociology of Education, 80*(2), 114–138. <https://doi.org/10.1177/003804070708000202>
- Fahle, E. M., & Reardon, S. F. (2018). Education. In Stanford Center on Poverty and Inequality (Ed.), *State of the Union: The Poverty and Inequality Report, Special Issue* (pp. 9–12). Pathways Magazine.
- Fahle, E. M., Shear, B. R., Kalogrides, D., Reardon, S. F., Disalvo, R., & Ho, A. D. (2018). *Stanford Education Data Archive: Technical Documentation v2.1*. Retrieved from https://cepa.stanford.edu/sites/default/files/SEDA_documentation_v21.pdf
- Fan, X., Fang, H., & Markussen, S. (2015). Mothers' Employment and Children's Educational Gender Gap. *National Bureau of Economic Research Working Paper Series, No. 21183*. <https://doi.org/10.3386/w21183>
- Fryer, R. G., & Levitt, S. D. (2010). An Empirical Analysis of the Gender Gap in Mathematics. *American Economic Journal: Applied Economics, 2*(2), 210–240. <https://doi.org/10.3386/w15430>
- Garner, M., & Engelhard, G. J. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12*(1), 29–51. <https://doi.org/10.1207/s15324818ame1201>
- Guiso, L., Monte, F., & Sapienza, P. (2008). Differences in Test Scores Correlated with Indicators of Gender Equality. *Science, 320*(May), 1–2. <https://doi.org/10.1126/science.1154094>
- Gunderson, E. A., Ramirez, G., Levine, S. C., & Beilock, S. L. (2012). The Role of Parents and Teachers in the Development of Gender-Related Math Attitudes. *Sex Roles, 66*(3–4), 153–166. <https://doi.org/10.1007/s11199-011-9996-2>
- Guryan, J., Hurst, E., & Kearney, M. (2008). Parental Education and Parental Time with Children. *Journal of Economic Perspectives, 22*(3), 23–46. <https://doi.org/10.1257/jep.22.3.23>
- Haines, E. L., Deaux, K., & Lofaro, N. (2016). The Times They Are a-Changing ... or Are They Not? A Comparison of Gender Stereotypes, 1983–2014. *Psychology of Women Quarterly, 40*(3), 353–363. <https://doi.org/10.1177/0361684316634081>
- Hao, L., & Yeung, W.-J. J. (2015). Parental Spending on School-Age Children: Structural Stratification and Parental Expectation. *Demography, 52*(3), 835–860. <https://doi.org/10.1007/s13524-015-0386-1>
- Hopcroft, R. L. (2005). Parental Status and Differential Investment in Sons and Daughters: Trivers-Willard Revisited. *Social Forces, 83*(3), 1111–1136. <https://doi.org/10.1353/sof.2005.0035>

- Hopcroft, R. L., & Martin, D. O. (2016). Parental Investments and Educational Outcomes: Trivers–Willard in the U.S. *Frontiers in Sociology*, 1(March), 1–12. <https://doi.org/10.3389/fsoc.2016.00003>
- Husain, M., & Millimet, D. L. (2009). The mythical “boy crisis”? *Economics of Education Review*, 28(1), 38–48. <https://doi.org/10.1016/j.econedurev.2007.11.002>
- Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender Similarities Characterize Math Performance. *Science (New York, N.Y.)*, 321(5888), 494–495. <https://doi.org/10.1126/science.1160364>
- Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences of the United States of America*, 106(22), 8801–7. <https://doi.org/10.1073/pnas.0901265106>
- Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in Children’s Self-Competence and Values: Gender and Domain Differences across Grades One through Twelve. *Child Development*, 73(2), 509–527. <https://doi.org/10.1111/1467-8624.00421>
- Kågesten, A., Gibbs, S., Blum, R. W., Moreau, C., Chandra-Mouli, V., Herbert, A., & Amin, A. (2016). Understanding factors that shape gender attitudes in early adolescence globally: A mixed-methods systematic review. *PLoS ONE*, 11(6), 1–36. <https://doi.org/10.1371/journal.pone.0157805>
- Kornrich, S. (2016). Inequalities in Parental Spending on Young Children. *AERA Open*, 2(2), 233285841664418. <https://doi.org/10.1177/2332858416644180>
- Lee, J., Moon, S., & Hegar, R. L. (2011). Mathematics skills in early childhood: Exploring gender and ethnic patterns. *Child Indicators Research*, 4(3), 353–368. <https://doi.org/10.1007/s12187-010-9088-9>
- Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, 136(6), 1123–1135. <https://doi.org/10.1037/a0021276>
- Lubienski, S., Robinson, J., Crane, C., & Ganley, C. (2013). Girls’ and Boys’ Mathematics Achievement, Affect, and Experiences: Findings from the ECLS-K. *Journal for Research in Mathematics Education*, 44(4), 634–645. <https://doi.org/10.5951/jresmetheduc.44.4.0634>
- Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research*, 37(1), 239–253. <https://doi.org/10.1016/j.ssresearch.2007.06.012>
- Pope, D. G., & Sydnor, J. R. (2010). Geographic Variation in the Gender Differences in Test Scores. *Journal of Economic Perspectives*, 24(2), 95–108. <https://doi.org/10.1257/jep.24.2.95>
- Raley, S., & Bianchi, S. (2006). Sons, Daughters, and Family Processes: Does Gender of Children Matter? *Annual Review of Sociology*, 32(1), 401–421. <https://doi.org/10.1146/annurev.soc.32.061604.123106>
- Ramey, G., & Ramey, V. A. (2010). The Rug Rat Race. *Brookings Papers on Economic Activity*, 2010(1), 129–176. <https://doi.org/10.1353/eca.2010.0003>
- Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárte, R. C. (2018). The Relationship Between Test Item Format and Gender Achievement Gaps on Math and ELA Tests in Fourth and Eighth Grades. *Educational Researcher*, XX(X), 0013189X1876210. <https://doi.org/10.3102/0013189X18762105>
- Reardon, S. F., Kalogrides, D., & Ho, A. D. Linking U.S. school district test score distributions to a common scale (2017). Working Paper, Stanford Center for Education Policy Analysis.
- Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2016). Using Heteroskedastic Ordered Probit Models to Recover Moments of Continuous Test Score Distributions From Coarsened Data. *Journal of Educational and Behavioral Statistics*, 42(16-02), 3–45. <https://doi.org/10.3102/1076998616666279>
- Robinson, J. P., & Lubienski, S. T. (2011). 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*, 48(2), 268–302.
<https://doi.org/10.3102/0002831210372249>
- Sohn, K. (2012). A new insight into the gender gap in math. *Bulletin of Economic Research*, 64(1), 135–155. <https://doi.org/10.1111/j.1467-8586.2010.00358.x>
- Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype Threat and Women’s Math Performance. *Journal of Experimental Social Psychology*, 35, 4–28. <https://doi.org/10.1006/jesp.1998.1373>
- Tomasetto, C., Alparone, F. R., & Cadinu, M. (2011). Girls’ Math Performance Under Stereotype Threat: The Moderating Role of Mothers’ Gender Stereotypes. *Developmental Psychology*, 47(4), 943–949. <https://doi.org/10.1037/a0024047>
- Trivers, R. L., & Willard, D. E. (1973). Natural Selection of Parental Ability to Vary the Sex Ratio of Offspring. *Science*, 179(4068), 90–92. <https://doi.org/10.1126/science.179.4068.90>
- Upadaya, K., & Eccles, J. (2015). Do teachers’ perceptions of children’s math and reading related ability and effort predict children’s self-concept of ability in math and reading? *Educational Psychology*, 35(1), 110–127. <https://doi.org/10.1080/01443410.2014.915927>
- Witt, S. D. (1997). Parental influence on children’s socialization to gender roles. *Adolescence*, 32, 253–259. <https://doi.org/10.1080/0300443001620101>

Tables

Table 1. Mean Achievement Estimates by Gender, Subject, Grade, and Year

	ELA							Math						
	2009	2010	2011	2012	2013	2014	2015	2009	2010	2011	2012	2013	2014	2015
	<i>Male</i>													
3	-0.02	-0.03	-0.03	-0.03	-0.05	-0.03	-0.05	0.08	0.08	0.07	0.07	0.07	0.08	0.09
4	-0.03	-0.03	-0.04	-0.05	-0.06	-0.05	-0.07	0.06	0.06	0.05	0.05	0.05	0.07	0.06
5	-0.02	-0.05	-0.05	-0.05	-0.06	-0.06	-0.09	0.05	0.05	0.04	0.04	0.03	0.04	0.03
6	-0.03	-0.04	-0.04	-0.05	-0.06	-0.07	-0.10	0.06	0.05	0.04	0.04	0.03	0.03	0.02
7	-0.05	-0.05	-0.06	-0.06	-0.08	-0.08	-0.11	0.05	0.05	0.04	0.02	0.01	0.01	0.00
8	-0.06	-0.06	-0.06	-0.06	-0.08	-0.10	-0.11	0.07	0.05	0.05	0.02	0.01	0.01	0.00
	<i>Female</i>													
3	0.18	0.16	0.16	0.17	0.16	0.17	0.18	0.05	0.04	0.03	0.04	0.04	0.05	0.07
4	0.17	0.15	0.15	0.16	0.15	0.16	0.17	0.03	0.03	0.03	0.05	0.03	0.04	0.05
5	0.17	0.16	0.15	0.16	0.15	0.15	0.17	0.03	0.03	0.02	0.04	0.03	0.04	0.06
6	0.19	0.19	0.18	0.19	0.17	0.17	0.20	0.07	0.06	0.06	0.05	0.06	0.07	0.08
7	0.20	0.20	0.20	0.20	0.18	0.18	0.22	0.08	0.07	0.06	0.06	0.06	0.07	0.06
8	0.22	0.21	0.21	0.20	0.19	0.19	0.23	0.08	0.07	0.06	0.05	0.06	0.07	0.09
	<i>Male-Female</i>													
3	-0.19	-0.19	-0.20	-0.20	-0.21	-0.20	-0.22	0.03	0.04	0.03	0.03	0.03	0.03	0.02
4	-0.20	-0.18	-0.20	-0.21	-0.21	-0.21	-0.24	0.03	0.03	0.02	0.01	0.02	0.03	0.01
5	-0.19	-0.21	-0.20	-0.20	-0.21	-0.21	-0.26	0.02	0.02	0.01	0.00	0.00	-0.01	-0.03
6	-0.22	-0.22	-0.22	-0.23	-0.24	-0.25	-0.29	-0.01	-0.01	-0.02	-0.01	-0.03	-0.05	-0.06
7	-0.25	-0.25	-0.26	-0.26	-0.26	-0.26	-0.33	-0.03	-0.02	-0.03	-0.04	-0.04	-0.06	-0.06
8	-0.27	-0.27	-0.27	-0.27	-0.27	-0.29	-0.35	-0.01	-0.02	-0.02	-0.03	-0.05	-0.06	-0.10

Notes: Table is based on the mean achievement estimates, standardized to the National NAEP distribution within subject, grade and year. To account for the fact that the data are unbalanced (not all districts have estimates in each grade, year, and subject), we obtain the estimated average test score means for each subject, grade, and year by gender from a model regressing the average test scores in a gender-subject-grade-year-district cell on a set of district and gender-subject-grade-year fixed effects. The averages reported in each cell in Table 1 are the estimated coefficients from the gender-subject-grade-year dummy variables in this model. Note that these averages are not weighted by sample size, and thus reflect the mean test score for the average district in each subject, grade, year, and gender.

Table 2. SES Factor Loadings

	Factor Loadings
Median Household Income (in 10,000s)	0.186
Proportion of Adults with BA+	0.182
Poverty Rate, 5-17 Year Olds	-0.176
Percent of 25-64 Year Olds in Labor Force & Unemployed	-0.142
Percent of 25-64 Year Olds Not in Labor Market	-0.154
Proportion in Management, Business and Financial Occupations	0.200
Proportion in Computer, Engineering and Science Occupations	0.176

Notes: The factors were generated in the overall sample of 12,954 districts. The same factor loadings were applied to the male and female specific versions of the SES composite.

Table 3. Means and Standard Deviations of Covariates

	<i>Mean</i>	<i>Standard Deviation</i>	<i>Source</i>
Percent Black in Public Schools	0.09	0.17	CCD
Percent Hispanic in Public Schools	0.14	0.21	CCD
Percent Other Race in Public Schools	0.04	0.10	CCD
Percent White in Public Schools	0.73	0.28	CCD
SES Composite, All Adults	-0.18	1.02	ACS, 2006-2010
SES Composite, Male Adults	0.33	1.22	ACS, 2006-2010
SES Composite, Female Adults	-0.69	0.90	ACS, 2006-2010
Male-Female SES Composite Difference	1.02	0.66	ACS, 2006-2010

Notes: In total, the summary includes 9,799 school districts. Means and standard deviations are unweighted. The SES composite includes income, education, employment, poverty, proportion in business occupations, and proportion in science occupations. The male and female specific SES composites use the factor loadings for all adults.

Table 4. NAEP-Adjusted Mean Achievement Estimates by Gender, Subject, Grade, and Year

	ELA							Math						
	2009	2010	2011	2012	2013	2014	2015	2009	2010	2011	2012	2013	2014	2015
	<i>Male</i>													
3	-0.05	-0.06	-0.06	-0.05	-0.06	-0.06	-0.06	0.07	0.07	0.06	0.06	0.05	0.05	0.05
4	-0.06	-0.07	-0.06	-0.06	-0.07	-0.07	-0.07	0.07	0.07	0.06	0.06	0.05	0.05	0.05
5	-0.07	-0.08	-0.07	-0.07	-0.08	-0.08	-0.08	0.07	0.06	0.06	0.06	0.05	0.05	0.04
6	-0.08	-0.08	-0.08	-0.08	-0.09	-0.09	-0.09	0.07	0.06	0.06	0.05	0.04	0.05	0.04
7	-0.09	-0.09	-0.09	-0.09	-0.10	-0.10	-0.10	0.07	0.07	0.06	0.04	0.03	0.03	0.02
8	-0.09	-0.10	-0.09	-0.09	-0.10	-0.11	-0.10	0.07	0.06	0.05	0.04	0.03	0.03	0.02
	<i>Female</i>													
3	0.13	0.13	0.12	0.13	0.12	0.11	0.11	0.01	0.01	0.01	0.01	0.01	0.02	0.02
4	0.14	0.14	0.14	0.14	0.13	0.12	0.12	0.01	0.02	0.02	0.02	0.01	0.02	0.02
5	0.15	0.15	0.15	0.15	0.14	0.13	0.14	0.02	0.02	0.02	0.02	0.02	0.03	0.03
6	0.17	0.17	0.17	0.17	0.16	0.15	0.15	0.03	0.03	0.03	0.03	0.02	0.04	0.03
7	0.18	0.18	0.18	0.18	0.17	0.16	0.16	0.04	0.04	0.04	0.02	0.02	0.03	0.02
8	0.20	0.20	0.20	0.20	0.18	0.17	0.18	0.04	0.04	0.04	0.03	0.02	0.03	0.02
	<i>Male-Female</i>													
3	-0.18	-0.18	-0.18	-0.18	-0.18	-0.17	-0.18	0.06	0.06	0.05	0.05	0.04	0.04	0.03
4	-0.21	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	0.06	0.05	0.05	0.04	0.04	0.03	0.02
5	-0.23	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	0.05	0.04	0.04	0.03	0.03	0.02	0.02
6	-0.25	-0.25	-0.25	-0.25	-0.24	-0.24	-0.24	0.04	0.04	0.03	0.03	0.02	0.01	0.01
7	-0.27	-0.27	-0.27	-0.27	-0.26	-0.26	-0.26	0.03	0.03	0.02	0.02	0.01	0.01	0.00
8	-0.29	-0.29	-0.29	-0.29	-0.29	-0.28	-0.28	0.03	0.02	0.01	0.01	0.00	0.00	0.00

Notes: Table is based on NAEP-adjusted mean residual estimates. Gender-subject-grade-year averages are estimated using a district fixed effects model with subject-grade-year-gender fixed effects, coefficients from this model are reported in the table. No adjustments are made for precision.

Table 5. Selected Multivariate Regression Results

	Math				ELA			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Male-Female Gap (Intercept)	0.0288 *** (0.0007)	0.0286 *** (0.0006)	0.0297 *** (0.0006)	0.0278 *** (0.0032)	-0.2304 *** (0.0006)	-0.2303 *** (0.0006)	-0.2308 *** (0.0006)	-0.2401 *** (0.0036)
SES Composite		0.0177 *** (0.0007)	0.0137 *** (0.0007)	0.0160 *** (0.0009)		0.0003 (0.0008)	0.0019 * (0.0008)	-0.0028 ** (0.0009)
Male-Female SES Composite Difference		0.0096 *** (0.0012)	0.0085 *** (0.0011)	0.0072 *** (0.0012)		0.0049 *** (0.0012)	0.0018 (0.0012)	0.0014 (0.0012)
Percent Black			-0.0811 *** (0.0034)	-0.0779 *** (0.0040)			-0.0288 *** (0.0037)	-0.0282 *** (0.0041)
Percent Hispanic			-0.0096 ** (0.0030)	-0.0175 *** (0.0042)			0.0423 *** (0.0032)	0.0199 *** (0.0042)
Percent Other Race			0.0081 (0.0068)	-0.0066 (0.0082)			-0.0157 * (0.0073)	-0.0056 (0.0082)
Grade	-0.0120 *** (0.0003)	-0.0119 *** (0.0003)	-0.0122 *** (0.0003)	-0.0125 *** (0.0003)	-0.0193 *** (0.0003)	-0.0194 *** (0.0003)	-0.0196 *** (0.0003)	-0.0198 *** (0.0003)
SES Composite -x- Grade		-0.0022 *** (0.0004)	-0.0019 *** (0.0004)	0.0020 *** (0.0004)		-0.0002 (0.0003)	0.0000 (0.0004)	0.0008 + (0.0004)
Male-Female SES Composite Difference -x- Grade		0.0040 *** (0.0006)	0.0043 *** (0.0006)	0.0012 + (0.0007)		0.0021 *** (0.0006)	0.0021 *** (0.0006)	0.0018 ** (0.0006)
Percent Black -x- Grade			0.0120 *** (0.0017)	0.0069 ** (0.0020)			0.0053 *** (0.0016)	0.0156 *** (0.0019)
Percent Hispanic -x- Grade			0.0025 (0.0015)	-0.0002 (0.0022)			0.0029 * (0.0014)	0.0083 *** (0.0020)
Percent Other Race -x- Grade			0.0135 *** (0.0036)	0.0146 ** (0.0043)			0.0110 ** (0.0035)	0.0143 ** (0.0041)
Cohort	-0.0044 *** (0.0002)	-0.0044 *** (0.0002)	-0.0046 *** (0.0002)	-0.0048 *** (0.0002)	0.0018 *** (0.0002)	0.0018 *** (0.0002)	0.0020 *** (0.0002)	0.0018 *** (0.0002)
SES Composite -x- Cohort		-0.0010 *** (0.0003)	-0.0005 + (0.0003)	0.0017 *** (0.0003)		-0.0003 (0.0003)	-0.0006 * (0.0003)	-0.0005 (0.0003)
Male-Female SES Composite Difference -x- Cohort		0.0021 *** (0.0004)	0.0021 *** (0.0004)	-0.0001 (0.0005)		0.0001 (0.0004)	0.0002 (0.0004)	0.0003 (0.0005)
Percent Black -x- Cohort			0.0084 *** (0.0012)	0.0025 + (0.0015)			-0.0023 + (0.0013)	0.0017 (0.0015)
Percent Hispanic -x- Cohort			0.0036 *** (0.0011)	-0.0016 (0.0016)			-0.0022 * (0.0011)	-0.0030 + (0.0016)
Percent Other Race -x- Cohort			0.0075 ** (0.0026)	0.0021 (0.0032)			0.0044 (0.0027)	0.0007 (0.0032)
Within-District SD	0.093	0.093	0.093	0.093	0.093	0.093	0.093	0.093
District-Level Gap SD	0.050	0.045	0.042	0.042	0.049	0.049	0.048	0.043
District-Level Grade Slope SD								
District-Level Cohort Slope SD	0.010	0.010	0.010	0.010	0.011	0.011	0.011	0.011
State-Level Gap SD				0.022				0.024
District Level Correlation between Math and ELA Gaps	0.85	0.92	0.95	0.99	0.85	0.92	0.95	0.99
District-Level Gap Reliability	0.59			0.51	0.59			0.53
District-Level Grade Slope Reliability								
District-Level Cohort Slope Reliability	0.31			0.30	0.35			0.35
State-Level Gap Reliability				0.89				0.91
District-Level Relative R ²		0.20	0.29	0.29		0.00	0.05	0.24

Notes: Selected coefficients shown. Standard errors are in parentheses; + p<0.10 * p<0.05 ** p<0.01 *** p<0.001. The number of observations in all models is 1,296,824 and the number of districts in all models is 9,799. All models use the NAEP-adjusted within state-subject-grade-year-gender mean residuals.

Table 6. Correlation Matrix

	Average Math Score	Male-Female Math Gap	Average ELA Score	Male-Female ELA Gap	SES Composite	Male-Female SES Composite Difference	Proportion of Black Students
Average Math Score	1.00						
Male-Female Math Gap	0.46	1.00					
Average ELA Score	0.93	0.50	1.00				
Male-Female ELA Gap	0.09	0.85	0.05	1.00			
SES Composite	0.70	0.43	0.73	0.03	1.00		
Male-Female SES Composite Difference	0.43	0.32	0.43	0.06	0.52	1.00	
Proportion of Black Students	-0.40	-0.41	-0.38	-0.13	-0.18	-0.26	1.00

Notes: The correlations between the average scores and gaps, and between these and the covariates are disattenuated to account for measurement error in the gaps and average scores. The correlations among the covariates are based on the observed data, with no disattenuation. Correlations among the average test score and test score gap measures are estimated from Model 1. Correlations between the average performance/gaps and the covariates (SES Composite, the Male-Female SES Composite Difference, and the Percentage of Black Students) are derived from the explained variance between a model that including each covariate separately (models not shown in paper) and a model with no covariates (Model 1).

Figures

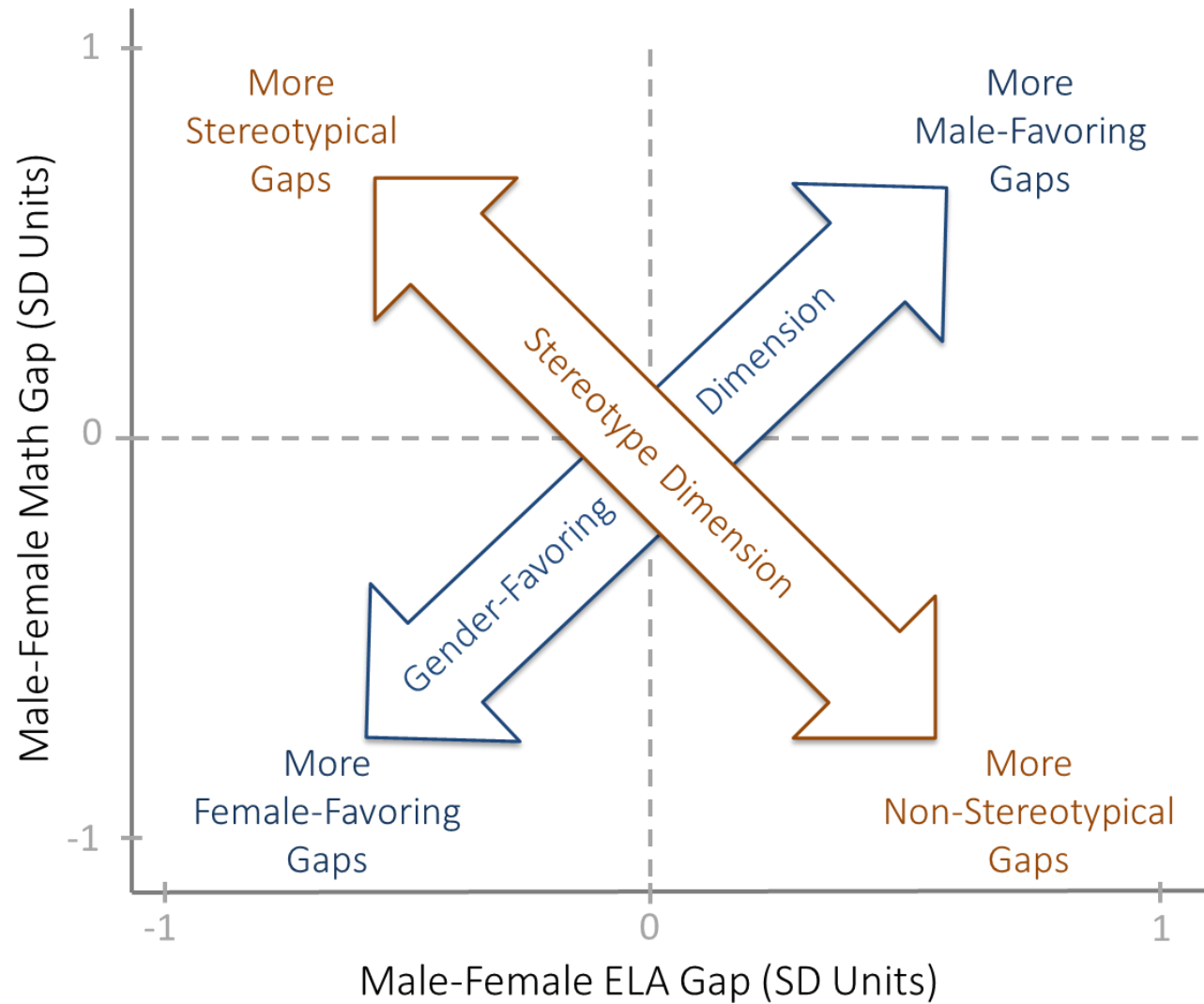
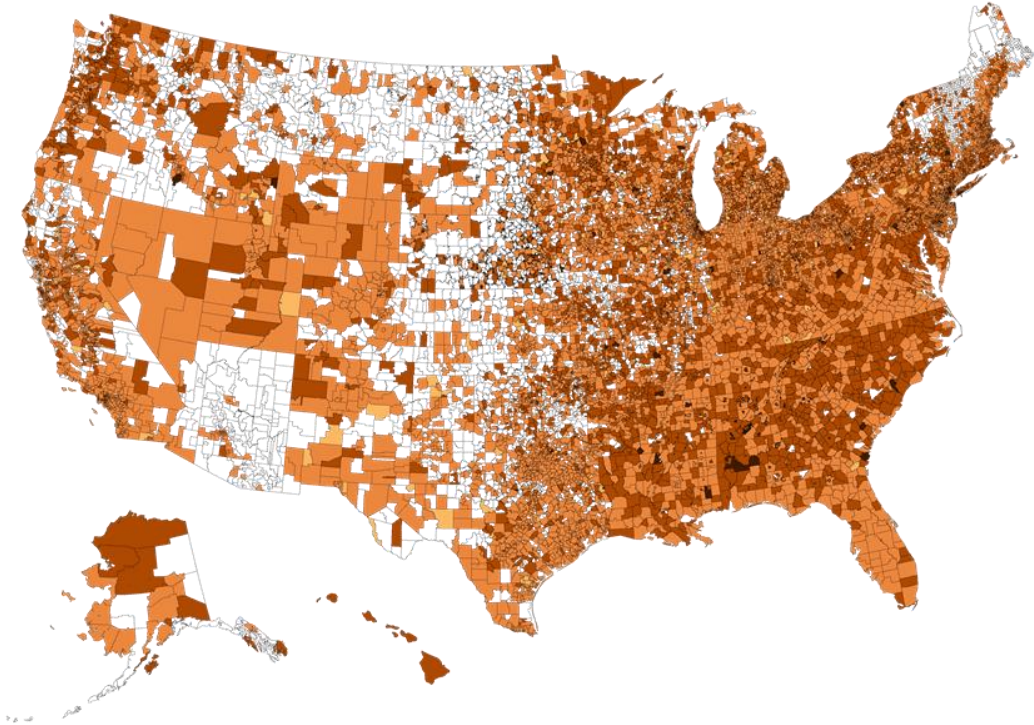
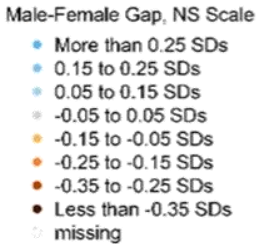
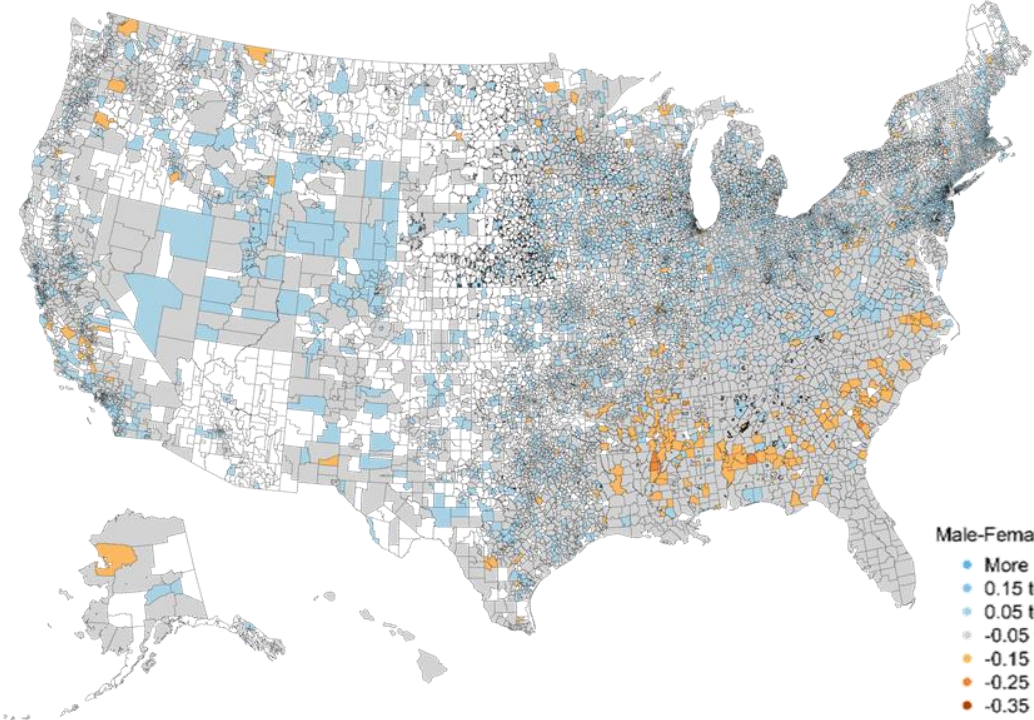


Figure 1. Dimensions of Gender Achievement

ELA

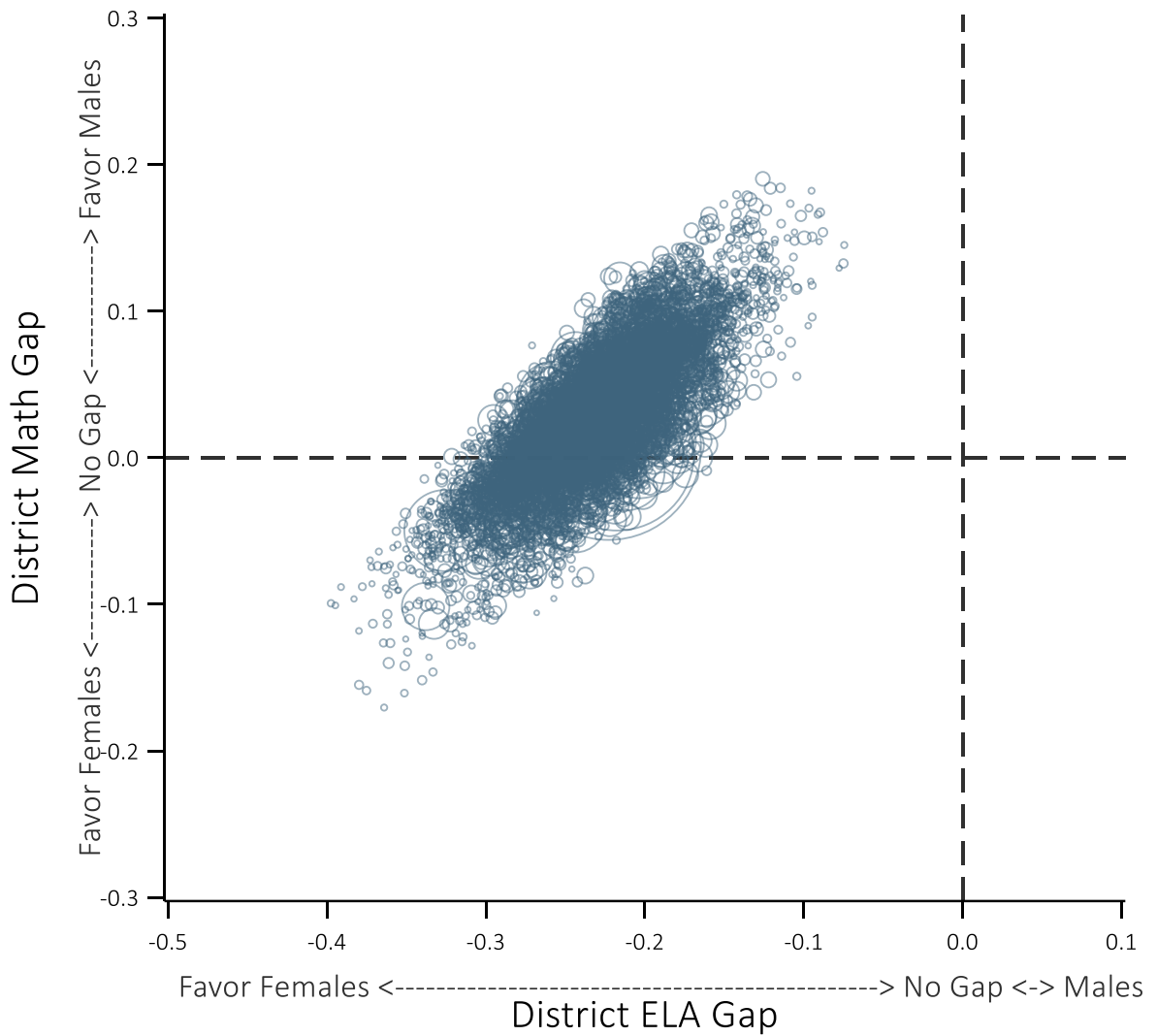


Math



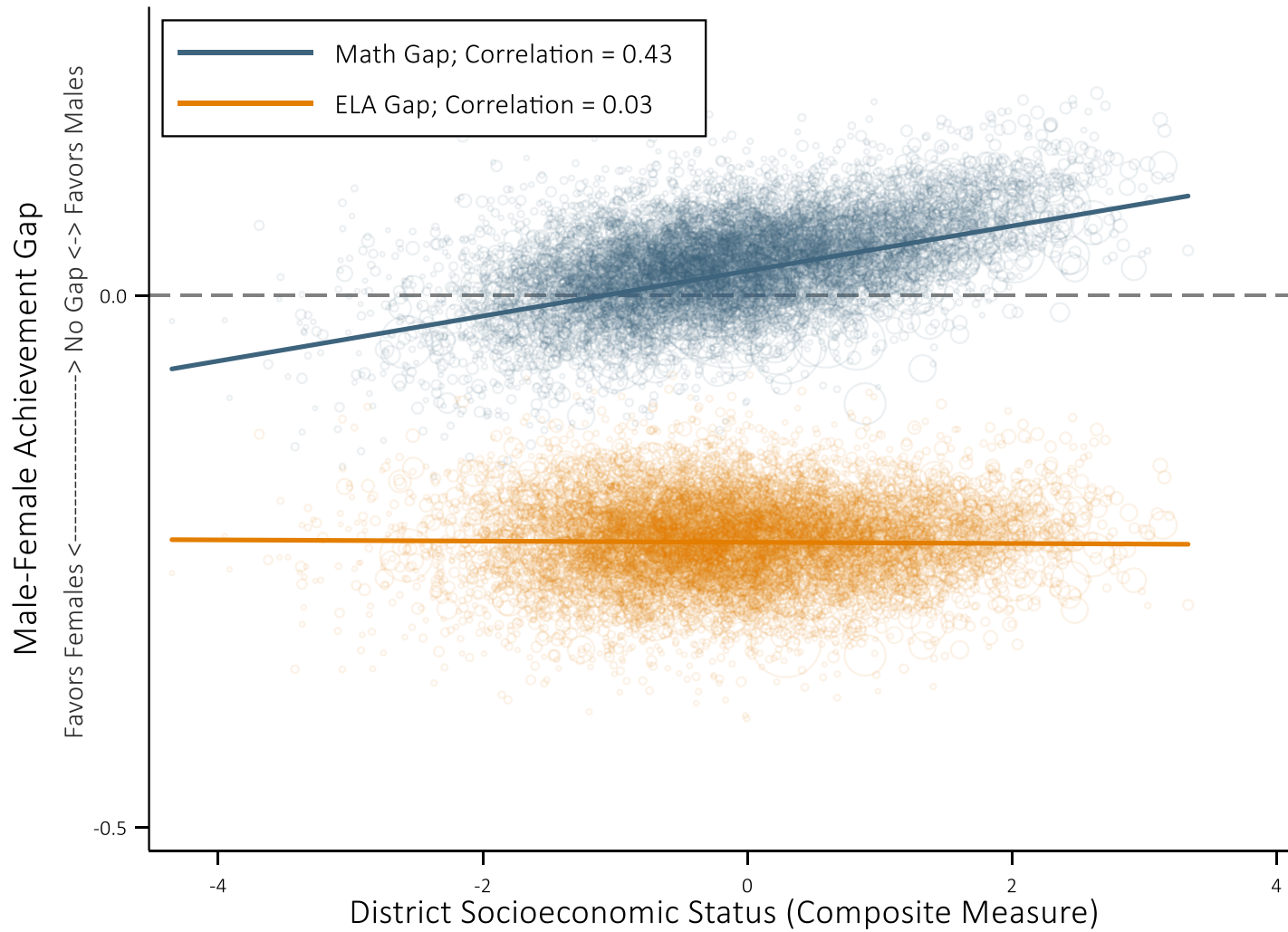
Note: The Empirical Bayes estimates (from Model 3) displayed.

Figure 2. Average Gender Achievement Gaps in U.S. School Districts



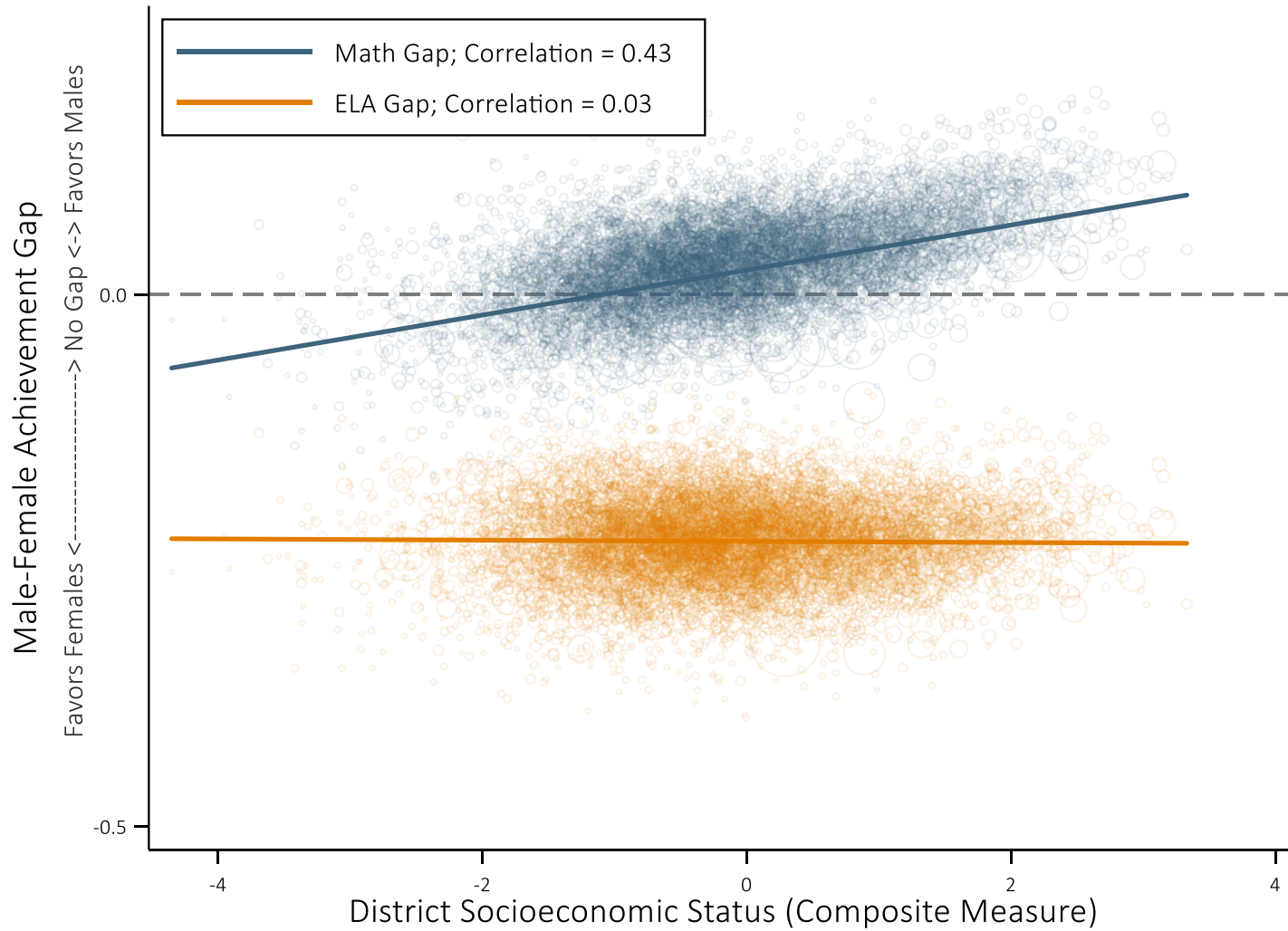
Observations are Empirical Bayes test score gap estimates from Model 3, weighted by the average number of students in a district-subject-grade-year. They are in standard deviation units, standardized to the NAEP national distribution within each subject, grade, and year.

Figure 3. The Relationship of Math and ELA Gender Achievement Gaps, across U.S. School Districts



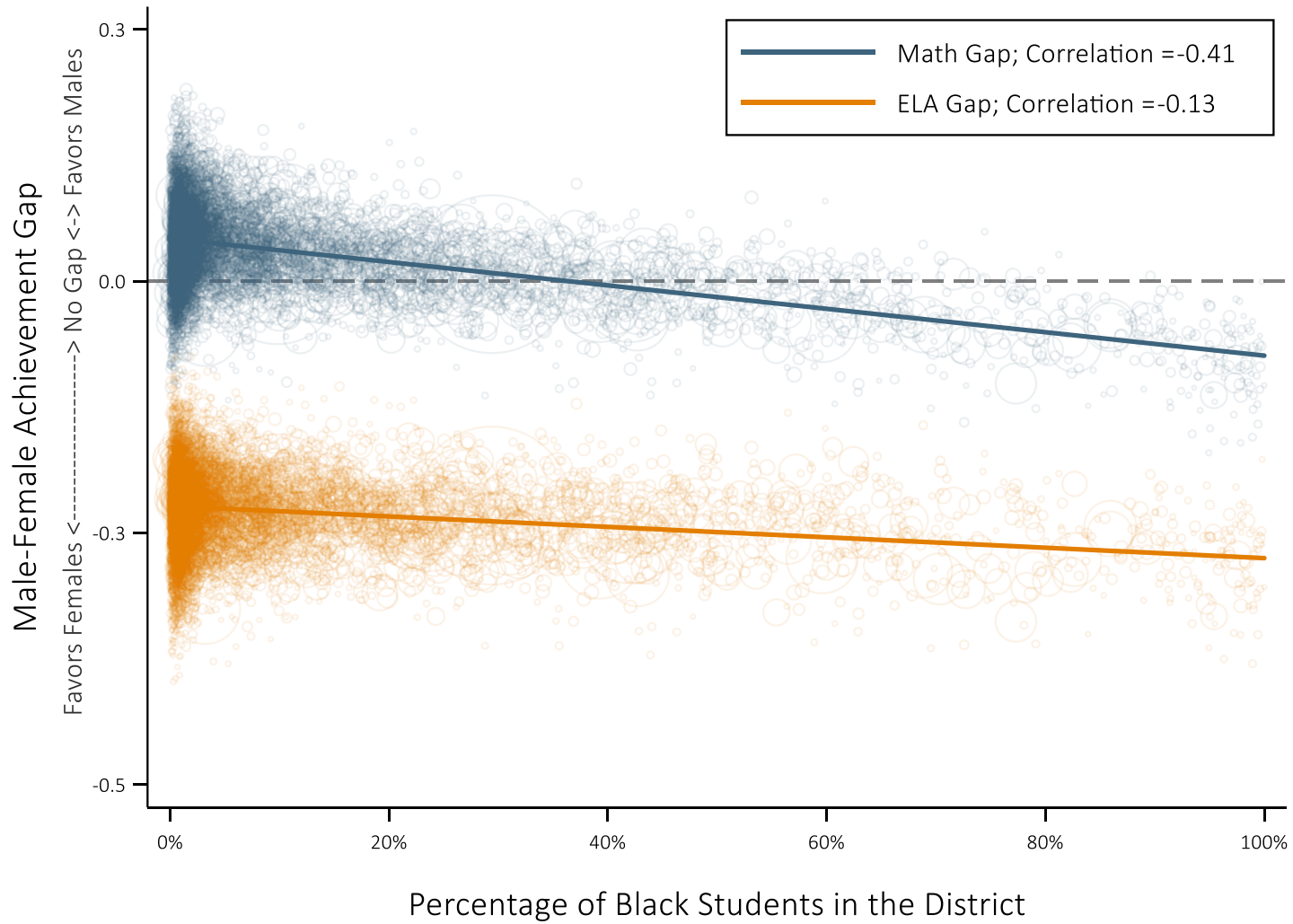
Note: The Empirical Bayes estimates shown underestimate the true variance in the male-female achievement gaps.

Figure 4. Male and Female Achievement in ELA and Math vs. SES, in U.S. School Districts



Note: The Empirical Bayes estimates shown underestimate the true variance in the male-female achievement gaps.

Figure 5. Male and Female Achievement in ELA and Math vs. SES Differences, in U.S. School Districts



Note: The Empirical Bayes estimates shown underestimate the true variance in the male-female achievement gaps.

Figure 6. Male and Female Achievement in ELA and Math vs. Percent of Black Students, in U.S. School Districts

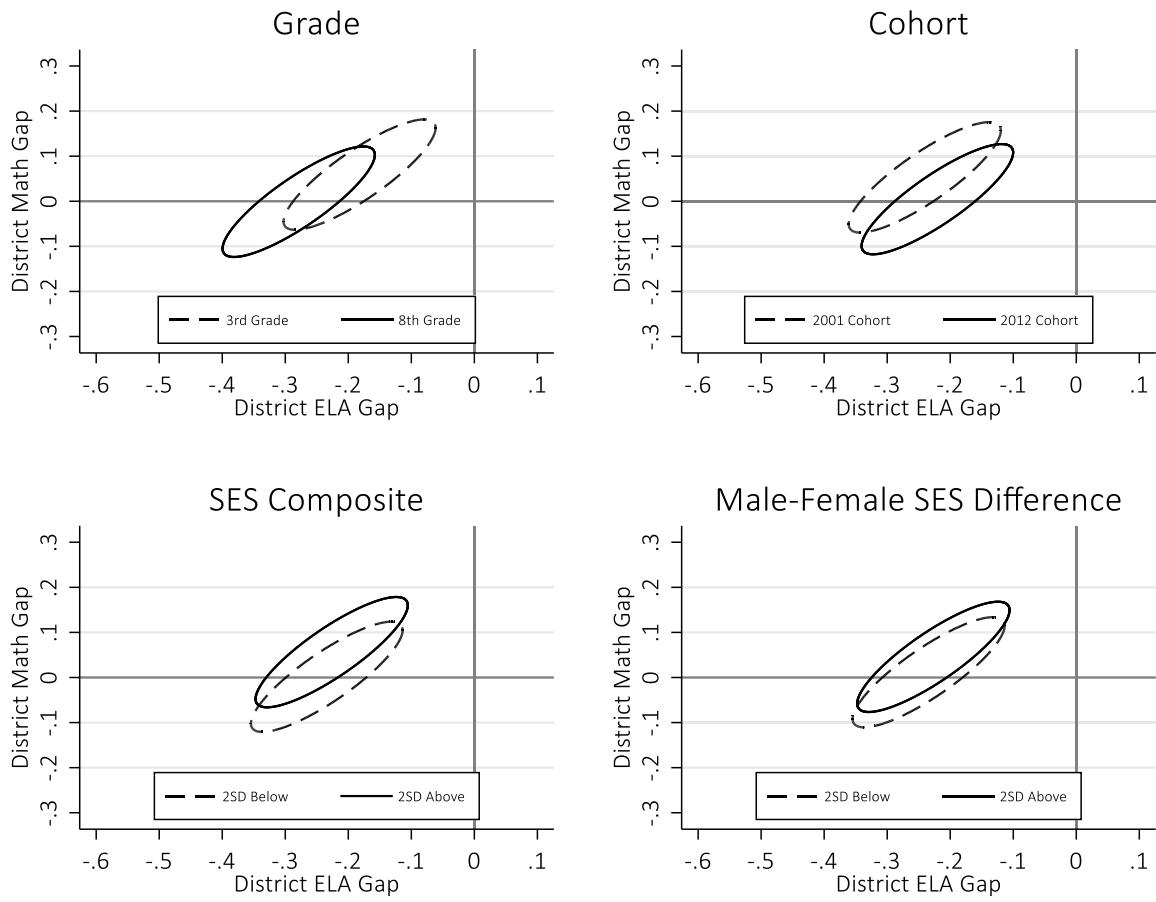


Figure 7. Stylized Model Results

Appendix A: Constructing NAEP-Adjusted Residualized Achievement Measures

We can write the estimated mean test score for gender s in subject b , grade g , year y , and district d in state f as

$$\hat{\mu}_{sbgyd} = \mu_{sbgyd}^{naep} + v_{sbgyf} + \hat{\epsilon}_{sbgyd},$$

where μ_{sbgyd}^{naep} is the true mean score for that population on the NAEP test; v_{sbgyf} is an error term specific to gender s on the test in state f in subject-grade-year bgy (that is, it is the difference in gender gaps as measured by the state test and the NAEP test in state f ; we assume here that this difference is constant across districts within state f , but not common across states); and $\hat{\epsilon}_{sbgyd}$ is sampling error specific to the estimated test score mean of gender s in subject b , grade g , year y , and district d . The estimated male-female gap in district d will then be

$$\begin{aligned} \hat{g}_{bgyd} &= [u_{[male]bgyd} - u_{[female]bgyd}] + [v_{[male]bgyf} - v_{[female]bgyf}] \\ &\quad + [\hat{\epsilon}_{[male]bgyd} - \hat{\epsilon}_{[female]bgyd}] \\ &= [g_{bgyd}^{naep}] + [\Delta v_{bgyf}] + [\Delta \hat{\epsilon}_{bgyd}], \end{aligned}$$

where g_{bgyd}^{naep} is the true male-female gap that we would observe if the NAEP test in subject b , grade g , and year y were administered in district d ; Δv_{bgyf} is the difference between the gender gap on NAEP and the gender gap on the state f test in subject b , grade g , and year y ; and $\Delta \hat{\epsilon}_{bgyd}$ is measurement error in the gap.

If Δv_{bgyf} is not constant across the tests used in different states, grades, years, and subjects, the measured gaps will not be comparable across states, grades, year and subject. Prior research suggests that Δv_{bgyf} varies considerably among states (Reardon et al., 2018). To address this, we residualize $\hat{\mu}_{sbgyd}$ by subtracting the average score of gender s in subject b , grade g , year y , and state f . We first estimate μ_{sbgyf} , the mean test score for students of gender s in subject b , grade g , year y , and state f , by taking a precision weighted average of the $\hat{\mu}_{sbgyd}$'s. Then we residualize $\hat{\mu}_{sbgyd}$ to obtain

$$\hat{\mu}_{sbgyd}^* = \hat{\mu}_{sbgyd} - \hat{\mu}_{sbgyf}$$

Finally, we add the average NAEP score for gender s in subject b , grade g , year y , and state f to the corresponding residualized district means¹⁶

$$\hat{\mu}'_{sbgyd} = \hat{\mu}_{sbgyd}^* + \hat{\mu}_{sbgyf}^{naep}$$

Note that we can rewrite this as

$$\begin{aligned} \hat{\mu}'_{sbgyd} &= \hat{\mu}_{sbgyd}^* + \hat{\mu}_{sbgyf}^{naep} \\ &= [\mu_{sbgyd}^{naep} + v_{sbgyf} + \hat{\epsilon}_{sbgyd}] - E[\mu_{sbgyd}^{naep} + v_{sbgyf} + \hat{\epsilon}_{sbgyd} | s, b, g, y, f] + \hat{\mu}_{sbgyf}^{naep} \\ &= [\mu_{sbgyd}^{naep} + v_{sbgyf} + \hat{\epsilon}_{sbgyd}] - \mu_{sbgyf}^{naep} - v_{sbgyf} + \hat{\mu}_{sbgyf}^{naep} \\ &= \mu_{sbgyd}^{naep} + \hat{\epsilon}_{sbgyd}. \end{aligned}$$

Thus $\hat{\mu}'_{sbgyd}$ is an estimate of the mean test scores of gender s in district d that has the bias due to the the difference between the state's test and NAEP removed. The $\hat{\mu}'_{sbgyd}$'s, and the resulting gap estimates, are therefore comparable across states within a grade, year, and subject.

¹⁶ NAEP does not report means in every grade and year, so we use interpolation to recover estimates of $\hat{\mu}_{sbgyf}^{NAEP}$ in non-tested grades and years. We use NAEP data from 2009, 2011, 2013, and 2015 NAEP tests in grades 4 and 8. We first standardize the gender-subject-grade-year estimates within subject, grades, and years using the standard deviation of achievement for all students in the same subject, grade, and year. We then interpolate to obtain estimated gender- and subject-specific means for the non-tested grades and years.