The Development of Gender Achievement Gaps in Mathematics and Reading During Elementary and Middle School: Examining Direct Cognitive Assessments and Teacher Ratings

Joseph Paul Robinson
Sarah Theule Lubienski
University of Illinois at Urbana-Champaign

Using K–8 national longitudinal data, the authors investigate males’ and females’ achievement in math and reading, including when gender gaps first appear, whether the appearance of gaps depends on the metric used, and where on the achievement distribution gaps are most prevalent. Additionally, teachers’ assessments of males and females are compared. The authors find no math gender gap in kindergarten, except at the top of the distribution; however, females throughout the distribution lose ground in elementary school and regain some in middle school. In reading, gaps favoring females generally narrow but widen among low-achieving students. However, teachers consistently rate females higher than males in both subjects, even when cognitive assessments suggest that males have an advantage. Implications for policy and further research are discussed.

KEYWORDS: achievement gaps, distributional analysis, gender, longitudinal data, metric-free gap analysis, teacher ratings

JOSEPH PAUL ROBINSON is an assistant professor of quantitative and evaluative research methodologies at the University of Illinois at Urbana-Champaign, Department of Educational Psychology, 210F Education Building, 1310 S. Sixth St., Champaign, IL 61820; e-mail: jpr@illinois.edu. His research focuses on causal inference and quasi-experimental designs, economics of education, patterns and causes of educational inequality, and the effects of educational practices and policies on equity and access.

SARAH THEULE LUBIENSKI is a professor of mathematics education at the University of Illinois at Urbana-Champaign; e-mail: stl@illinois.edu. She studies mathematics achievement, instruction, and reform, focusing on inequities in students’ mathematics outcomes and the policies and practices that shape those outcomes.
Recent debates about gender and education have focused on whether males or females are more “shortchanged” in school. Scholars interested in gender equity have traditionally been primarily concerned about females, but others now argue that males are actually disadvantaged. Males score lower in elementary reading assessments, tend to get worse grades, and are less likely to complete high school and attend college than females (Riordan, 1999; Sommers, 2000). After reviewing the evidence about gender and educational outcomes, Riordan (1999) concluded that “males are not flourishing in schools” (p. 47) and called for schools to more carefully monitor the needs of males. However, a recent American Association of University Women (AAUW; 2008) report counters claims of a “boys’ crisis.” Drawing on data from fourth grade through college, they argue that both females’ and males’ achievement has improved over the past few decades and that females’ gains have not come at males’ expense.

These arguments raise the question of whether our schools are, indeed, shortchanging one gender group or another. All too often, though, this question is addressed by comparing the achievement of groups in one school subject, at a single point in time, usually some time after they entered school. However, to determine whether one group is losing ground relative to another group, we should begin measuring student achievement at the start of kindergarten and then follow the same children throughout their school careers. Additionally, given that gender patterns in math performance tend to run counter to those in reading, examinations of both subjects together provide a more complete picture of girls’ and boys’ learning.

This study examines gender patterns in student achievement using data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K). These analyses follow students from kindergarten through eighth grade, the highest grade level that will be included in this data set. The study investigates the unique achievement trends of males and females in math and reading, if and when gender gaps develop, where on the achievement distribution the gaps are most prevalent, and whether the answers to these questions depend upon the metric used to measure achievement. Additionally, teachers’ own assessments of males and females are compared to the gender patterns on direct cognitive assessments. The (dis)similarity in the teacher trends and direct cognitive trends is discussed as one potential source of the gender gap, suggesting the importance of a heightened awareness of the needs of particular student groups.

Background

Recent concerns about gender equity, as well as more general education policies (e.g., No Child Left Behind [NCLB]), have tended to focus on the subjects of math and reading. The ECLS-K data set also gives primary focus
to these two subjects. Hence, math and reading are the two academic subjects considered here.

Math and Gender

Most national analyses of gender disparities in U.S. school achievement have used data from the National Assessment of Educational Progress (NAEP). According to the NAEP Long-Term Trend (LTT), the gender gap in 17-year-olds’ math achievement was 8 points (favoring males) in 1973, or approximately 0.3 standard deviations (SDs). Gender disparities in high school course taking and related issues began to receive attention in the 1970s (Fennema & Hart, 1994), and the high school achievement gap narrowed over the next decade. Since 1990, the LTT math gap for 17-year-olds has remained between 3 and 6 points (AAUW, 2008; Perie, Moran, & Lutkus, 2005; Rampey, Dion, & Donahue, 2009). In contrast, there were small but significant LTT math gender gaps favoring females for both 9- and 13-year-olds in 1973. However, by the early 1990s, these gaps had reversed to favor males and have remained generally around 0.1 SDs or less (Perie et al., 2005). Over the past decade, results from the Main NAEP (which is more responsive to curricular trends than the LTT) have shown small but persistent math gender disparities favoring males at fourth, eighth, and twelfth grades, with gaps of roughly 0.1 SDs, or the equivalent of a few months of schooling (McGraw, Lubienski, & Strutchens, 2006; Perie et al., 2005).

Hence, gender disparities in U.S. math achievement have been relatively small and have varied over time. Gender gaps have also been found to vary—in both magnitude and direction—by country (Else-Quest, Hyde, & Linn, 2010; Mullis, Martin, & Foy, 2008). Despite the variation in gender patterns in math achievement across nations, both TIMSS (Trends in International Mathematics and Science Study) and PISA (Program for International Student Assessment) data reveal that boys express more positive attitudes toward math in almost all participating countries (Else-Quest et al., 2010; Ginsburg, Cooke, Leinwand, Noell, & Pollock, 2005).

When and for whom does the U.S. math gender gap develop? Although recent NAEP and TIMSS data consistently indicate that U.S. males outscore females at fourth grade, these data sets do not allow for examining whether such gaps exist before fourth grade, including whether gaps are present when children begin school. Hence, the ECLS-K database has recently been used to examine gender-related patterns in early achievement. Using ECLS-K, researchers have found math gender gaps as early as kindergarten or first grade. Math gaps favoring males have also been found to increase between kindergarten and third grade (Husain & Millimet, 2009; LoGerfo, Nichols, & Chaplin, 2006; Rathbun, West, & Germino-Hauskin, 2004).

Denton and West (2002) found no overall gender differences at first grade but found that males tended to be more proficient in advanced
The Development of Gender Achievement Gaps

math skills than females. These findings echo an earlier, smaller-scale study reported by Fennema, Carpenter, Jacobs, Franke, and Levi (1998) as well as NAEP analyses revealing a larger gender gap at the top of the achievement distribution (McGraw et al., 2006). Most relevant to this study, Penner and Paret (2008) examined the development of gender gaps in math achievement from kindergarten through fifth grade. They found that gender gaps begin as early as kindergarten in the top of the achievement distribution and then appear throughout the rest of the distribution by third grade.

However, most recently, Hyde, Lindberg, Linn, Ellis, and Williams (2008) found that gender gaps in math were not significant on NCLB tests given in second through eleventh grades in 10 states, raising questions about whether there really is a gender gap in math achievement anymore. They also attempted to examine gaps on more challenging test items, given that males have been found to outperform females on such items, but they found that the tests lacked such items. However, they did find gaps favoring males at the upper end of the achievement distribution.

Gender disparities among the highest-achieving students appear to have implications for later career choices. Over the past decade, women earned only 18% of engineering bachelor’s degrees (Dey & Hill, 2007). The latest U.S. census data indicate that women who work full time still earn only 77% of men’s salaries, or 69% when comparing men and women 10 years out of college. Much of this wage gap is attributable to the fact that more men pursue math-related careers, as both women and men in those fields earn more than their counterparts in other fields. (Dey & Hill, 2007). Moreover, the lack of women in such careers diminishes the pool of high-quality U.S. students who contribute to those fields.

There have been laudable attempts to boost females’ interest in math through special programs (e.g., Karp & Niemi, 2000; Morrow & Morrow, 1995). Many of these programs have targeted females during their middle and high school years, which have traditionally been considered a critical time for the formation of females’ mathematical attitudes and aptitudes. However, disparities in men’s and women’s career choices remain.

Reading and Gender

In contrast to math, females tend to outscore males in various reading assessments. In fact, reading scores have played a primary role in arguments that schools are shortchanging males academically (e.g., Riordan, 1999; Sommers, 2000). However, gender gaps in reading achievement are not new. Almost 50 years ago, Gates (1961) found that females in second through eighth grades outscored males in reading. This gap has persisted over the past several decades but narrowed significantly for 9-year-olds, from 13 points (0.3 SDs) in 1971 to 7 points on the 2008 NAEP LTT (Ramphey et al., 2009). Similarly, reading achievement data from the 2005
and 2007 Main NAEPs reveal that females outscored males by less than 0.2 SDs at fourth grade but more than 0.3 SDs at eighth and twelfth grades. Gender gaps in reading tend to be larger and more pervasive in countries around the world than math gaps. For example, gaps in 15-year-olds’ reading performance measured by PISA consistently favor females, averaging more than 0.3 SDs (Organisation for Economic Co-operation and Development, 2009). Additionally, fourth-grade females significantly out-scored males in 38 of the 40 the countries that participated in the 2006 Progress in International Reading Literacy Study (PIRLS), with the difference averaging roughly 0.2 SDs. The U.S. gender gap of 0.1 SDs on PIRLS was below the international average.

When and for whom does the reading gender gap develop? As with math, the ECLS-K data have provided the basis for several recent studies of gender differences in elementary school reading. According to Denton and West (2002), gender disparities in ECLS-K reading performance appear in first grade, where females tend to be slightly more proficient in some advanced reading skills. Similarly, Rathbun et al. (2004) found that third-grade females were more likely than their male peers to derive meaning from reading text and to make literal inferences. However, they found no substantive gender differences in the overall gains students made from kindergarten through third grade. Husain and Millimet (2009) found that low-achieving males tend to lose ground in reading between kindergarten and third grade.

The ECLS-K results through third grade raise the question of whether boys are simply “late bloomers” who will eventually catch up with their female peers or whether reading gaps will persist or even widen in later grades and require targeted interventions. Patterns in Main NAEP suggest that reading gaps between males and females do not narrow over time. However, again, NAEP does not follow the same students over time. Additionally, it could be that gender gaps narrow for most students but widen at the top or bottom of the achievement distribution. The availability of the newly released K–8 ECLS-K data allows for a new examination of this question.

Why Are There Gender Gaps in Achievement?

Investigations of the causes of gender differences in achievement have spanned over three decades and have involved a variety of disciplines, including psychology, sociology, biology, and education. Some researchers have examined the role of parents’ beliefs and practices in shaping gender patterns in academic outcomes (Jacobs, 1991; Lubienski & Crane, 2009). Others have examined gender differences in affective factors, including students’ attitudes toward, and self-confidence in, reading (Baker & Wigfield, 1999; Rathbun et al., 2004) and math (Eccles, 1986; Fennema & Sherman, 1977; Leder, 1992). Researchers have also examined the field of math itself.
and its subtle, multifaceted barriers to females’ participation (e.g., see review by Lacampagne, Campbell, Herzig, Damarin, & Vogt, 2007).

Studies in these and other areas have informed theories regarding the root causes of gender disparities. Traditionally, these theories fell into “nature” or “nurture” camps, with the former attributing gender differences to genetics and the latter subscribing to gender role socialization theory, or the idea that parents, teachers, and others teach girls and boys to conform to their expected gender roles (Block, 1973). However, more recently, scholars have argued that each of these perspectives is too simplistic. For example, psychobiosocial theorists suggest an interplay among biology, psychology, and socialization, with achievement differences between boys and girls originating from small biological differences, which can be reinforced and magnified in their particular cultural context (Halpern, Wai, & Saw, 2005; Lytton & Romney, 1991; Wood & Eagly, 2002). Other scholars emphasize individual agency, arguing that women make informed choices based on their perceptions, values, and beliefs (e.g., Eccles, 1986). However, despite their differences in emphases, scholars from these various orientations recognize the importance of environmental factors in the formation of gender differences in experiences, values, beliefs, and ultimately, achievement.

This study does not test the merits of competing explanations of gender gaps but is rooted in the perspective that socializing agents—especially teachers—play an important role in shaping girls’ and boys’ achievement. Indeed, the fact that math gender gaps vary by time and place indicates the central role that environment and socialization play in the formation of these gaps (Else-Quest et al., 2010). Although reading gaps appear to be more persistent across contexts than math gaps, there is ample evidence that teachers shape males’ and females’ achievement in both reading and math (e.g., Beilock, Gunderson, Ramirez, & Levine, 2010; Good, 1987). Factors underlying achievement patterns undoubtedly exist at home and in society at large, but classroom teachers are arguably the points of greatest leverage within the education community. Hence, the issue of teacher expectations and socialization of males and females merits further discussion, as it is directly relates to this study’s focus on teacher assessments of their male and female students.

Teacher expectations. Prior research indicates that teachers’ beliefs about students’ knowledge and abilities vary by gender and are important influences of classroom processes and student achievement in both reading and math. According to Good’s (1987) review of the literature, teachers generally demand more from students they view as higher achievers and treat them with more respect. More recently, Tach and Farkas (2006) examined ECLS-K data and found that being placed into a higher-ability reading group was positively related to learning behaviors and achievement. Hence, if
teachers underestimate males' reading abilities, this might negatively affect their learning, particularly if ability grouping is used.

In a review of teacher expectations and beliefs about math and gender, Li (1999) concluded that teachers tend to view math as a male domain and also tend to have higher expectations for, and better attitudes toward, their male students. Similarly, in a study of 38 first-grade teachers, Fennema, Peterson, Carpenter, and Lubinski (1990) found that teachers’ beliefs about females and males differed, with teachers more often naming males as the “best math students” and attributing males’ success to ability and females’ success to their effort. Still, it is unclear from this study whether the teachers would tend to rate males as higher achieving than females in general, as opposed to only among the highest achievers. In contrast, in a study of 56 Michigan math teachers, Madon et al. (1998) found that teachers tended to rate seventh-grade females’ performance and effort as higher than that of males but tended to rate their abilities equally. They also found that while teachers’ perceptions of males’ and females’ achievement were accurate, males and females actually reported similar levels of effort.

McKown and Weinstein (2002) studied relationships between teacher expectations and student performance in the classrooms of 30 San Francisco area teachers of first, third, and fifth grades. They found that in math, females were more likely than males to be harmed by teachers’ underestimates of their abilities and were less likely to benefit from teachers’ overestimates of their abilities. However, no such pattern was found in reading.

Overall, there is conflicting evidence about whether teachers tend to rate males’ or females’ math performance higher and whether teacher assessments are consistent with direct cognitive assessment (e.g., standardized exams). Much of the existing evidence is rather dated and from a relatively small number of classrooms. There is even less evidence available pertaining to teachers’ expectations of males and females in reading.

“Good girls.” A factor that may underlie teachers’ discrepant views of males and females is the socialization of females into “good-girl” roles. The effects of this socialization may be evident in several ways in school (Forgasz & Leder, 2001). For example, females tend to earn higher grades, even in math and science (AAUW, 2008). Ready, LoGerfo, Lee, and Burkam (2005) found that the majority of the gender gap in kindergarten literacy learning could be explained by the tendency for females to exhibit more positive learning approaches (e.g., on-task behavior) than males. Additionally, more young males than females report that they engage in “problem behaviors,” such as fighting at school (Rathbun et al., 2004).

According to a study by Flynn and Rahbar (1994), teachers tend to refer males for special education services twice as often as females, despite the fact that roughly equal numbers of males and females fall into the “reading-disabled” category, according to test results. Similarly, Hibel, Farkas,
and Morgan (2006) found that even after accounting for reading and math test scores, males are disproportionately referred to special education. Flynn and Rahbar hypothesize that such differences are likely due to males’ more disruptive behaviors, and they conclude that females might be noticed only when they are severely struggling, which, they argue, is unfair to females.

Correll (2001) analyzed the National Educational Longitudinal Study of 1988 and found that males were almost 4 times more likely to choose a quantitative college major than females with similar math achievement. Consistent with the hypothesis that girls strive to please the teacher, Correll found that teachers’ feedback (e.g., grades) was a greater influence of females’ self-perceptions than of males’. She also found that males view themselves as better in math relative to females with equal test scores, but the opposite was true for reading, further indicating that cultural beliefs influence students’ self-perceptions. Drawing from literature of teacher expectations published in the early 1990s, Correll argued that teachers judge males as more competent in math than academically similar females and that such judgments contribute to females’ perceptions of their own competence and later career choices. However, it is unclear whether teachers today actually do hold males’ math abilities in higher regard.

Some Unanswered Questions

Overall, scholars have drawn attention to the existence of math and reading gender gaps and have highlighted possible causes. We know from cross-sectional, international data that math gender gaps appear earlier in the United States than in most other countries and that gender gaps in both math (favoring males) and reading (favoring females) seem to be larger at twelfth grade than at earlier grades. Studies have also pointed to the importance of teachers’ perceptions and treatment of students; however, findings regarding teacher expectations of males and females tend to be dated and based on limited samples.

Some initial research using ECLS-K has confirmed the existence of gender gaps in math and reading achievement in early elementary school (LoGerfo, Nichols, & Chaplin, 2006; Rathbun et al., 2004). However, Reardon’s (2008) finding that the size of the gap and the direction of its growth can depend on the metric used for the analysis (e.g., scale scores, standardized scores) suggests that we explore whether such gaps hold up regardless of the metric used to measure them as well as whether patterns that exist in early elementary grades persist through the middle school years. This is particularly important to examine in the case of math, with recent research suggesting that gender gaps in U.S. school achievement are no longer significant (Hyde et al., 2008). Additionally, we do not know whether U.S. teachers’ assessments of males’ and females’ academic achievement
mirror students’ test performance or whether teachers might systematically under- or overestimate females’ proficiency in math or reading, relative to what direct cognitive assessments suggest. For this study, the specific set of research questions is as follows:

1. What are the achievement scores of males and females in reading and math from kindergarten through eighth grade? What types of skills does each group demonstrate at various time points?
2. When do math and reading gender gaps first appear in elementary school, and do they widen or narrow as children progress from kindergarten to eighth grade? Are gender gaps concentrated in a particular achievement range (e.g., among low-achieving students), or are they consistent across the score distribution? Does the metric of the achievement measure (scale score, standardized score) affect the answers to these questions?
3. Are teachers’ assessments of the relative progress of males and females similar to those of formal cognitive assessments?
4. How do K–8 patterns in gender gaps in reading achievement and teacher assessments compare to those in math? And what does this comparison suggest for future research into the causes of these gaps?

Data

The ECLS-K data set—collected by the U.S. Department of Education—is used for these analyses. ECLS-K includes data on a nationally representative sample of about 21,400 kindergarten students in academic year 1998–1999.

Sample Sizes and Attrition

The number of students in the ECLS-K sample decreased over time, from a high of 20,578 in spring of kindergarten to a low of 9,725 in spring of eighth grade. However, this study involved only 7,075 of the 9,725 eighth graders for several reasons. Of the full sample, 7,803 had nonzero longitudinal weights. Of that group, 7,248 had valid Wave 1 math and reading scores. The majority of the students dropped due to nonvalid scores were students not assessed at the start of kindergarten due to limited English proficiency. From the 7,248-student sample, each successive full-sample wave lost the following number of students due to missing test-score data: 6 (in Wave 2), 5 (in Wave 4), 69 (in Wave 5), 22 (in Wave 6), and 66 (in Wave 7). This yields 7,080 students, 5 of whom were subsequently dropped for missing assessment date information, resulting in the final analytic sample of 7,075 students. Use of the ECLS-K longitudinal sample weights makes the analyses representative of the population of English-proficient students in kindergarten in 1998–1999. For completeness, we ran our analyses using the full sample at each cross-sectional wave as well; though not presented here, the results were very similar.
Finally, to lessen teacher burden, math teacher survey data were collected for only half of the ECLS-K fifth and eighth graders (the other half were assigned to science). However, the sample was split randomly, so the estimates are unbiased.

ECLS-K Assessments and Metrics

The ECLS-K assessment items were created in consultation with state and national standards, elementary content specialists, and multicultural experts. Items were field-tested and their construct validity confirmed by verifying that student performance consistently correlated with the established Woodcock-McGrew-Werder Mini-Battery of Achievement (Pollack, Najarian, Rock, & Atkins-Burnett, 2005; Tourangeau, Nord, Lê, Pollack, & Atkins-Burnett, 2006). Reliabilities were consistently high, ranging from .89 to .96 (Tourangeau et al., 2006).

ECLS-K provides several types of assessment scores for math and reading, which can be divided into two broad categories: direct cognitive assessments and teacher ratings. Direct cognitive assessment scores come from the assessments based on item response theory (IRT) that are administered to students in each wave. Although students completed only a subset of the full test battery, the National Center for Education Statistics converted students’ scores into a metric that reflects the number of questions the students would have answered correctly if they received the full test battery. These scores are called the IRT scale scores. The ability scores were converted into another metric, which standardized the assessment scores within each wave of data collection. These $T$ scores have a mean of 50 and SD of 10; we converted the metric to a $z$ score, standardized to have a mean of 0 and pooled SD of 1, so that gaps can be interpreted as effect sizes. We follow Cohen’s (1988) suggestion for interpreting effect sizes of 0.2 SDs as small, 0.5 as medium, and 0.8 as large; however, Valentine and Cooper (2003) caution that in education, effects are likely to be small, which may lead to interpretations that minimize the importance of smaller effect sizes when strictly following Cohen’s guidelines.

The second type of assessment is based on teacher evaluations of students’ proficiency. ECLS-K refers to these scores as the “academic rating scale” scores; for simplicity, however, we will refer to them as teacher ratings. Teachers were asked to rate “the degree to which a child has acquired and/or chooses to demonstrate” a variety of reading and math skills, knowledge, and behaviors. The 5-point teacher rating scale ranged from 1 = *not yet*, which was defined as “child has not yet demonstrated skill, knowledge or behavior,” to 5 = *proficient*, indicating that the “child demonstrates skill, knowledge, or behavior competently and consistently.” The specific areas rated within reading and math varied by grade. The fifth-grade reading teacher rating questionnaire, for example, included 11 areas related to...
reading, writing, and speaking, including “reads fluently,” “conveys ideas clearly when speaking,” “composes multi-paragraph stories/reports,” and “reads and comprehends expository text.” The math domains spanned number, measurement, geometry, and statistics, with specific items including “models, reads, writes and compares fractions,” and “recognizes properties of shapes and relationships among shapes.” Teachers were instructed to rate only those aspects that had been introduced in the class and to otherwise select not applicable.

ECLS-K performed Rasch analyses (similar to those used in the direct cognitive assessments) on the teacher rating scale in an effort to (a) create a measure for modeling growth in the teacher ratings, (b) make the ratings more comparable to the direct cognitive assessment, and (c) estimate values for students whose teachers did not complete some items because those skills had not been taught yet. We standardized the teacher rating scores, just as we did for the direct cognitive assessments, meaning that these gaps can be interpreted as effect sizes as well.

**Method**

Our analyses explore achievement scores and gaps using a variety of strategies to provide a more complete picture of the development of gaps. First, we explore the achievement scores at the 10th, 50th, and 90th percentiles of males and females separately. We then turn to achievement gaps and begin by asking the traditional question, “Do achievement gaps exist on average, and how big are they?” This question, however, may depend on (a) when the gaps are measured (e.g., fall of kindergarten, spring of eighth grade), (b) the metric used (e.g., scale scores, standardized scores), and (c) who does the rating (i.e., the ECLS-K test-administrators or teachers). We then turn to questions of where in the achievement distribution gaps exist, grow, and shrink over time. A portion of the distributional achievement gap analyses is devoted to “metric-free” analyses, so named because these analyses rely not on the magnitude of the gap but only on the ordinal rank. We use these metric-free analyses because there is concern that the ECLS-K IRT scale scores are not “interval scaled” (Reardon, 2008).

This concern about the ECLS-K metrics merits some explanation. A test is said to be interval scaled if a 1-point difference between groups means the same magnitude of difference in true cognitive skills regardless of where in the score distribution the gap is measured and if the meaning of a 1-point difference is stable across time. However, Reardon (2008) notes the ECLS-K IRT scale scores are meant to be interpreted as the number of items correct on a test and are therefore sensitive to the relative proportion of “easy” to “difficult” items.
Achievement Scores Over Time by Gender

Examining the achievement scores of females and males separately helps us identify if one group is gaining new skills while the other is stagnating, thereby providing additional context to the subsequent gap analyses. Since we are interested in achievement throughout the distribution, we will plot the achievement of the 10th, 50th, and 90th percentiles of males and females at each wave of data collection. To provide additional context as to which skills students are learning, we map the ECLS-K-provided skill proficiencies onto the achievement scores. In this way, we can see, for example, that the 10th percentile of females in eighth grade is learning skills related to place value, while the 50th percentile of females is learning higher-lever skills related to rate, measurement, and fractions. One limitation of these proficiency levels is that they convey a hierarchy of math or reading knowledge that might not always hold. For example, students might learn a great deal about fractions before learning about measurement or place value. Hence, caution is warranted in interpreting these proficiency levels.5

Achievement Gaps on Average

In the existing literature, average achievement gaps—in general, not just gender gaps—have been measured using three different approaches: mean achievement differences (in the original test metric), mean standardized differences, and metric-free (or rank-based) measures (Reardon & Robinson, 2008). To address our first question regarding average achievement gaps, we explore mean differences in the scale score metric and standardized score metric. These two metrics are used because each has strengths and weaknesses: In particular, the original metric is more sensitive to assumptions about interval scaling, while the standardized score metric is more susceptible to measurement error biasing estimated gaps toward zero (Reardon, 2008).

Our analyses of average achievement gaps involve a series of weighted least squares regressions, where each child’s observation is weighted by the appropriate longitudinal child weight, provided by ECLS-K. Separate analyses are conducted at each of six waves of data collection (from fall of kindergarten through spring of eighth grade) by subject (reading and math) and metric (scale scores and standardized scores). In addition, we present similar analyses for the teacher ratings of students’ proficiency levels. A more technical description of this analysis, as well as the other analyses discussed below, can be found in the supplementary materials, accessible through the online version of this article on the journal’s Web site.

Achievement Gaps Throughout the Distribution

Our remaining questions concern achievement gaps in direct cognitive assessments and teacher ratings throughout the distribution rather than average differences. As a first approach to these questions, we used quantile
regression (Koenker & Bassett, 1978), which was similarly used by Penner and Paret (2008) in their study of the math achievement gender gap. In addition, we develop and apply a metric-free method for studying gaps throughout the distribution; this contribution is significant, as we discuss below.

**Quantile regression.** Using quantile regression, we estimate metric-based gaps at specified quantiles (e.g., the median, the 10th percentile, the 75th percentile; Koenker & Bassett, 1978). For instance, one of our quantile regression analyses will tell us the difference between the 90th percentile of males’ math achievement and the 90th percentile of females’ math achievement. See the online supplementary materials for this article for more details on this approach.

**Metric-free gender gaps throughout the score distribution.** Recent research on racial-ethnic achievement gaps has called for metric-free measures of achievement gaps (Ho & Haertel, 2006; Reardon, 2008; Reardon & Galindo, 2009). Rather than relying on psychometric scaling assumptions, metric-free measures rely only on the ordered rank of students. For example, a metric-free analysis might ask the question, “What is the probability that a randomly selected girl scores higher than a randomly selected boy?” (as in Reardon & Galindo, 2009, except they are interested in Hispanic and White students).

Although achievement gaps measured on a traditional metric are affected by the addition or deletion of difficult or easy items, metric-free measures are not affected unless such items are differentially difficult based on gender. For example, prior research suggests that males outperform females on math questions involving measuring instruments, such as speedometers (McGraw & Lubienski, 2007). If such items were added to a test, both the metric-free and the metric-based comparisons of males and females would be affected. However, if items were added that were generally more difficult or easy for males and females alike, the metric-free comparison would not be affected, while the metric-based comparisons could be heavily influenced.

Since we are interested in the gender gap throughout the distribution, we require a measure that reflects the metric-free gap at different points in the achievement distribution. Ho and Haertel (2006) used a “proportional difference” measure, which in our case would consist of subtracting the proportion of males observed from the proportion of females observed by a given percentile. Although this measure has appeal for its simple interpretation, its calculation obscures the magnitude of the relative differences in the tails of the overall distribution (see the article’s supplementary materials for an example). Given our particular interest in achievement gaps in the tails of the distribution, we develop and implement a different metric-free measure for assessing ordinal gaps throughout the distribution—our measure, which we call $\lambda_\theta$, provides an index of the relative difference between the genders, adjusting for the proportions of each group observed, where $\theta$ indicates the percentile at which $\lambda$ is evaluated.
Let $\Phi_m(\theta)$ and $\Phi_f(\theta)$ be the cumulative distribution functions for males and females observed by the $\theta$th percentile of the overall distribution. For percentiles below the median (i.e., $\theta < 50$), $\lambda_\theta$ reflects the proportion of males at or below a specific percentile, relative to the sum of the separate proportions of males and females at or below that percentile. For percentiles at or above the median, $\lambda_\theta$ reflects the proportion of females above a specific percentile of the overall distribution, relative to the sum of the separate proportions of males and females above that percentile. The scale for $\lambda_\theta$ ranges from 0 (favoring males) to 1 (favoring females). For example, if $\lambda_\theta = 0.5$ at each percentile of the distribution (i.e., for each value of $\theta$), this signifies that males and females are equally represented throughout the distribution (i.e., their individual cumulative density functions overlap perfectly). The supplementary online materials provide further details for constructing $\lambda_\theta$ as well as an illustration of the difference between the proportional difference measure and $\lambda_\theta$.

When the comparison groups are equally represented in the population and sample (as males and females are), we can take advantage of this fact and simplify our interpretation of $\lambda_\theta$. In our case, $\lambda_\theta$ is simply the proportion of the group that is female (for achievement score values above the median) or male (for values below the median). For instance, if $\lambda_{90} = 0.35$ (at the 90th percentile), that means that the top 10% of students is composed of 35% females and 65% males. If $\lambda_{10} = 0.6$ (at the 10th percentile), that means that in the bottom 10% of students, 60% are male and 40% are female.

Results

Achievement Scores of Males and Females

Figure 1 plots the math achievement scores of the 10th, 50th, and 90th percentiles of males and females at kindergarten entry through eighth grade. The left axis presents the numeric value of the IRT scale score (which ECLS-K recommends for measuring growth; Tourangeau et al., 2006), while the right axis lists the item-cluster proficiency level associated with that value. For example, according to these levels, in the spring of kindergarten, the 10th percentile of males is learning about relative size, while the 50th percentile of males is learning about sequences, and the 90th percentile of males is working on addition and subtraction. Although the male and female score profiles are similar within each percentile shown, males gain an early advantage at the top of the distribution. Over time, males pull away from
their female counterparts, first at the 90th percentile (by the spring of first grade), then the 50th percentile (by the spring of third grade), and finally at the 10th percentile (by the spring of fifth grade). At the 10th and 90th percentiles, however, the gaps appear to reduce between fifth and eighth grades.

Turning to the reading achievement scores, Figure 2 suggests that males and females at each of the percentiles presented are learning similar skills. Notably, as students progress through the grade levels, the 90th percentiles of males and females track each other quite well, but the emergence of a gender gap can be seen developing at the 50th and 10th percentiles.

As demonstrated by these figures, both males and females are learning new and advanced skills as they progress through the grade levels. Moreover, this skill acquisition is occurring at the lower, middle, and upper portions of the distributions. Therefore, if gaps do emerge, it is not because one group’s learning has stalled; rather, the other group has acquired more knowledge in a given time period. We now turn to exploring the gaps.

**Average Differences in Direct Cognitive Assessments**

Consistent with previous analyses of the early waves of ECLS-K data, this study finds no significant gap between females’ and males’ overall mean math scores at the start of kindergarten, regardless of whether we look at the T scores or scale scores (see Table 1). As students progress through
elementary school, the math gender gap widens and peaks—in favor of males—at third and fifth grades before actually reversing its growth trajectory during middle school. The peak average advantage for males is 0.24 SDs—a small but nontrivial effect size—in the standardized score metric and almost 6 points in the scale score metric. By the spring of eighth grade, this reduces to 0.12 SDs (half of its peak value) and 2.5 points in the respective metrics.

The pattern is different in reading and depends somewhat on which metric is used for the analysis. Using the IRT-based scale scores, LoGerfo, Nichols, and Reardon (2006) found that males and females both learn considerable amounts (about 80 points in the scale score metric) from kindergarten through third grade in reading, but females learn even more (about 3 points more). Table 1 shows a consistent story using the IRT scale scores for reading. However, standardized scores convey a different story. Though females are increasing their advantage in the IRT scale scores, females are losing relative ground to males. That is, at each successive wave of assessments, the scale score distribution widens (i.e., the standard deviation increases), so that converting the scale scores into a standardized metric reveals that females had an average advantage of 0.20 SDs when they began kindergarten but only about a 0.13-SD advantage by the end of fifth grade. By the end of eighth grade, females are about 0.21 SDs ahead of males, which is similar to their advantage at the end of first grade, although their scale score advantage is about 1.5 points greater in eighth grade than in first grade. From both methodological and

![Figure 2. Reading achievement scores by gender and at different points in distribution.](image)

*Note.* The top pair of lines represents the 90th percentiles of males and the 90th percentile of females. The middle and bottom pairs represent the 50th and 10th percentiles, respectively.
policymaking perspectives, this emphasizes the importance of exploring the gaps in a number of ways.

### Table 1
Mean Male-Female Differences by Subject, Assessment Type, and Wave of Data Collection

<table>
<thead>
<tr>
<th>Wave</th>
<th>Mathematics by Assessment Type</th>
<th>Reading by Assessment Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T$ Score</td>
<td>IRT Scale</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>Fall K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male-female</td>
<td>$-0.003$</td>
<td>$0.169$</td>
</tr>
<tr>
<td></td>
<td>$(0.024)$</td>
<td>$(0.211)$</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
<tr>
<td>$N$</td>
<td>$7,075$</td>
<td>$7,075$</td>
</tr>
<tr>
<td>Spring K</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male-female</td>
<td>$0.040$</td>
<td>$0.720^{**}$</td>
</tr>
<tr>
<td></td>
<td>$(0.024)$</td>
<td>$(0.279)$</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>$0.000$</td>
<td>$0.001$</td>
</tr>
<tr>
<td>$N$</td>
<td>$7,075$</td>
<td>$7,075$</td>
</tr>
<tr>
<td>Spring 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male-female</td>
<td>$0.075^{**}$</td>
<td>$2.065^{***}$</td>
</tr>
<tr>
<td></td>
<td>$(0.024)$</td>
<td>$(0.417)$</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>$0.001$</td>
<td>$0.003$</td>
</tr>
<tr>
<td>$N$</td>
<td>$7,075$</td>
<td>$7,075$</td>
</tr>
<tr>
<td>Spring 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male-female</td>
<td>$0.236^{***}$</td>
<td>$5.794^{***}$</td>
</tr>
<tr>
<td></td>
<td>$(0.024)$</td>
<td>$(0.573)$</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>$0.014$</td>
<td>$0.014$</td>
</tr>
<tr>
<td>$N$</td>
<td>$7,075$</td>
<td>$7,075$</td>
</tr>
<tr>
<td>Spring 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male-female</td>
<td>$0.240^{***}$</td>
<td>$5.694^{***}$</td>
</tr>
<tr>
<td></td>
<td>$(0.024)$</td>
<td>$(0.577)$</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>$0.014$</td>
<td>$0.013$</td>
</tr>
<tr>
<td>$N$</td>
<td>$7,075$</td>
<td>$7,075$</td>
</tr>
<tr>
<td>Spring 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male-female</td>
<td>$0.124^{***}$</td>
<td>$2.495^{***}$</td>
</tr>
<tr>
<td></td>
<td>$(0.024)$</td>
<td>$(0.529)$</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>$0.004$</td>
<td>$0.003$</td>
</tr>
<tr>
<td>$N$</td>
<td>$7,075$</td>
<td>$7,075$</td>
</tr>
</tbody>
</table>

*Note.* Robust standard errors appear in parentheses below estimated male-female gaps. IRT = item response theory.

*p < .05. **p < .01. ***p < .001.
Table 2

Standardized Score Quantile Regression Results by Subject and Wave of Data Collection

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Fall K</th>
<th>Spring K</th>
<th>Spring 1</th>
<th>Spring 3</th>
<th>Spring 5</th>
<th>Spring 8</th>
<th>Fall K</th>
<th>Spring K</th>
<th>Spring 1</th>
<th>Spring 3</th>
<th>Spring 5</th>
<th>Spring 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>10th</td>
<td>-0.048</td>
<td>-0.009</td>
<td>-0.036</td>
<td>0.129**</td>
<td>0.284***</td>
<td>0.043</td>
<td>-0.121***</td>
<td>-0.276***</td>
<td>-0.355***</td>
<td>-0.292***</td>
<td>-0.090*</td>
<td>-0.243***</td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td>(0.037)</td>
<td>(0.044)</td>
<td>(0.059)</td>
<td>(0.058)</td>
<td>(0.044)</td>
<td>(0.035)</td>
<td>(0.040)</td>
<td>(0.049)</td>
<td>(0.061)</td>
<td>(0.035)</td>
<td>(0.042)</td>
</tr>
<tr>
<td>25th</td>
<td>-0.018</td>
<td>0.043</td>
<td>0.036</td>
<td>0.193***</td>
<td>0.296***</td>
<td>0.190***</td>
<td>-0.213***</td>
<td>-0.235***</td>
<td>-0.108***</td>
<td>-0.243***</td>
<td>-0.147***</td>
<td>-0.175***</td>
</tr>
<tr>
<td></td>
<td>(0.035)</td>
<td>(0.035)</td>
<td>(0.037)</td>
<td>(0.041)</td>
<td>(0.039)</td>
<td>(0.035)</td>
<td>(0.029)</td>
<td>(0.032)</td>
<td>(0.026)</td>
<td>(0.059)</td>
<td>(0.048)</td>
<td>(0.032)</td>
</tr>
<tr>
<td>50th</td>
<td>0.016</td>
<td>0.031</td>
<td>0.133***</td>
<td>0.195***</td>
<td>0.218***</td>
<td>0.206***</td>
<td>-0.182***</td>
<td>-0.125***</td>
<td>-0.130***</td>
<td>-0.223***</td>
<td>-0.186**</td>
<td>-0.138***</td>
</tr>
<tr>
<td></td>
<td>(0.032)</td>
<td>(0.030)</td>
<td>(0.025)</td>
<td>(0.028)</td>
<td>(0.027)</td>
<td>(0.034)</td>
<td>(0.026)</td>
<td>(0.023)</td>
<td>(0.028)</td>
<td>(0.048)</td>
<td>(0.058)</td>
<td>(0.032)</td>
</tr>
<tr>
<td>75th</td>
<td>0.048</td>
<td>0.099**</td>
<td>0.190***</td>
<td>0.308***</td>
<td>0.305***</td>
<td>0.132***</td>
<td>-0.178***</td>
<td>-0.135***</td>
<td>-0.190***</td>
<td>-0.132***</td>
<td>-0.061</td>
<td>-0.125***</td>
</tr>
<tr>
<td></td>
<td>(0.051)</td>
<td>(0.030)</td>
<td>(0.027)</td>
<td>(0.081)</td>
<td>(0.057)</td>
<td>(0.026)</td>
<td>(0.042)</td>
<td>(0.019)</td>
<td>(0.055)</td>
<td>(0.024)</td>
<td>(0.054)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>90th</td>
<td>0.124**</td>
<td>0.195***</td>
<td>0.285***</td>
<td>0.286***</td>
<td>0.294***</td>
<td>0.119***</td>
<td>-0.127**</td>
<td>-0.212***</td>
<td>-0.085</td>
<td>-0.059**</td>
<td>-0.019</td>
<td>-0.105**</td>
</tr>
<tr>
<td></td>
<td>(0.045)</td>
<td>(0.041)</td>
<td>(0.031)</td>
<td>(0.085)</td>
<td>(0.037)</td>
<td>(0.032)</td>
<td>(0.041)</td>
<td>(0.055)</td>
<td>(0.046)</td>
<td>(0.018)</td>
<td>(0.043)</td>
<td>(0.035)</td>
</tr>
<tr>
<td>N</td>
<td>7,075</td>
<td>7,075</td>
<td>7,075</td>
<td>7,075</td>
<td>7,075</td>
<td>7,075</td>
<td>7,075</td>
<td>7,075</td>
<td>7,075</td>
<td>7,075</td>
<td>7,075</td>
<td>7,075</td>
</tr>
</tbody>
</table>

Note. Standard errors, appearing in parentheses below estimated male-female gaps, were calculated using 500 bootstrapped replications. The outcome variable standardized score (in the table title) refers to the T scores rescaled to be interpreted as effect sizes with a standard deviation of 1.

*p < .1. *p < .05. **p < .01. ***p < .001.
The Development of the Math Achievement Gender Gap Over the Distribution

Although there is no math achievement gap on average at the start of kindergarten, our analyses reveal that males in the uppermost portions of the distribution are outperforming their female counterparts. Interestingly, this gap at the top when students begin kindergarten creeps its way farther down the achievement distribution as grade level increases, such that significant achievement gaps exist throughout the upper 90% of the distribution by the spring of third grade.

For the standardized score metric of the direct cognitive assessments, Table 2 presents the results of quantile regressions, which estimate the math achievement gap between males and females at the 10th, 25th, 50th, 75th, and 90th percentiles of the overall distribution. The first column presents the estimated math gap at each of these percentiles when students are in the fall of kindergarten and shows that the gap at the 90th percentile is 0.12 SDs (in favor of males) and is not significantly different from zero at the lower percentiles. By the spring of kindergarten, the gap has grown in favor of males throughout the distribution, and the gap is significantly different from zero at the 90th percentile (where the gap is 0.20 SDs) and the 75th percentile (where the gap is 0.10 SDs). The gap continues to spread further down the distribution, becoming significant at the median by the spring of first grade and at the 25th and 10th percentiles by spring of third grade. By the spring of fifth grade, the gap has widened or remained steady—between 0.22 and 0.30 SDs—at each of the percentiles examined. Yet by eighth grade, the gap reduced at each of the percentiles, though the reductions were largest at the ends.

Thus far, we have described metric-based gaps, either on average or at specific points in the distribution. Figure 3 presents metric-free measures of the math gap throughout the achievement distribution at each wave of data collection. Recall that the index ($\lambda_\theta$) is a measure of the groups' relative (not absolute) proportions observed, where values below 0.5 favor males and values above 0.5 favor females. In each panel, we draw a line through the value of 0.5—the value representing equal proportions above (or below) a given percentile—and present 95% confidence intervals around the index value to evaluate the statistical significance of a value. The first panel of Figure 3 shows a significant rank-based gap (in favor of males) beginning just above the 75th percentile of the overall distribution. For example, at the 99th percentile, the value of $\lambda_{90} = 0.25$ indicates that in the fall of kindergarten, the top 1% of students comprises 25% females and 75% males.

Interestingly, the value of $\lambda_{90}$ (i.e., the top 1% of students) moves in the direction of 0.5 (equality) in math as grade level increases—in the spring of kindergarten, $\lambda_{90} = 0.15$ (15% of the top 1% are females); in the spring of third grade, $\lambda_{90} = 0.25$ (25% are female); and by the spring of eighth grade,
$\lambda_{90} = 0.37$ (37% are female). While the trend toward equality among the top 1% is promising, we do not observe an equal (50%) representation at the 99th percentile. In addition, the narrowing of the gender gap at the extreme top of the distribution should not overshadow the widening of the gender gap at virtually all points below the 99th percentile: For example, the trend for proportion of students at or above the 75th percentile who are female ($\lambda_{75}$) progresses from 0.5 in fall of kindergarten to 0.4 in spring of first grade and remains about 0.4 through eighth grade.
Consistent with the quantile regression results, the gap favoring males spreads to the lower percentiles as grade level rises. By the spring of fifth grade, the metric-free graphics show that the gap significantly favors males at all percentiles except the very lowest. Also consistent with the metric-based gaps, the gap reduces—throughout the distribution but perhaps more in the tails—during the middle school years.

A Reduction in the Reading Achievement Gender Gap During the Elementary School Years?

When children enter kindergarten, females exhibit an advantage in reading at all points in the test score distribution. Over the course of elementary school, though, the reading gap appears to ultimately shrink for most students, except those at the lowest parts of the distribution. This finding is consistent across the various approaches to distributional gaps, as detailed below.

The right half of Table 2 presents quantile regression results for standardized scores in reading. This metric-based set of estimates shows consistently significant gaps (favoring females) throughout the distribution in the fall of kindergarten, continuing through the spring of third grade. The gap when students enter school is about 0.12 SDs at both the 10th and 90th percentiles and is closer to about 0.18 to 0.21 SDs in the interquartile range. After the fall of kindergarten, the gaps at the top quantiles generally decrease, while the lowest quantile presented (i.e., the 10th percentile) exhibits the largest gap (except in spring of fifth grade), ranging from 0.24 to 0.36 SDs (except in fifth grade). The spring of fifth grade is also notable for the lack of a significant difference in the upper portion of the distribution. By eighth grade, however, the gap favoring females is again significant—between 0.1 and 0.24 SDs—at each of the percentiles in the table.

As with the math trends, the metric-free results (see Figure 4) parallel the metric-based results; for reading, both types of analyses reveal that (a) the gap favors females throughout the distribution when students begin kindergarten, (b) the gap becomes more concentrated in the lowest percentiles, and (c) the gap in the top quarter of the distribution reduces to nonsignificant levels temporarily in fifth grade before again widening somewhat during middle school. Later, we discuss the importance of looking at the gender representation among struggling readers (e.g., the bottom 5%), those most at risk of becoming illiterate adults. By the spring of eighth grade, λ₅ = 0.67: In the bottom 5% of readers, 67% are male. In other words, males outnumber females 2-to-1 among the bottom 5% of readers.

Differences Between Teacher Assessments and Direct Cognitive Assessments

Two interesting findings emerge when comparing the teacher ratings and the direct cognitive assessments: (a) Teachers rate females higher
(relative to males) than would be suggested by the direct cognitive assessments, and (b) for math, the trends in the average gaps track each other very closely (despite teachers’ beliefs that females perform relatively better than the cognitive scores suggest).

First, teachers rate females as demonstrating a greater level of skills than males in both subjects at the beginning of kindergarten, as shown in Table 1. In reading, the average gender gap (favoring females) in teachers’ ratings is 0.194, virtually identical to the gap of 0.196 in the direct cognitive assessments.

Figure 4. Metric-free gender reading achievement gaps through distribution by wave.
Note. Values above 0.5 favor females; values below 0.5 favor males. The gray region in each panel is a 95% confidence interval, generated by 500 bootstrapped replications.
assessments. In math, however, the direct cognitive assessments find no average gender difference at the beginning of kindergarten, yet on average, teachers rate females higher by 0.14 SDs. In math, judging by the teacher ratings, males have “caught up” to females by spring of first grade; however, by that period, the standardized direct cognitive assessments show a significant advantage for males of 0.08 SDs. However, a caveat is necessary here: ECLS-K designed the teacher ratings to complement the direct cognitive assessments but not to serve as an exact teacher rating analog to the direct assessments. We return to the implications of these assessment differences in the Discussion section.

Despite the questionable comparability between the teacher ratings and direct cognitive assessments within a given wave, we can compare the trends in the teacher ratings to the trends of the direct cognitive assessments. This brings us to our second point: There is a striking similarity between the direct cognitive assessment and teacher rating trends in math. Figure 5 displays the average gender differences (i.e., males-females) for the standardized scores from the direct cognitive assessments and teacher rating scale at each wave of data collection by content area. The location of the teacher rating score trend toward the bottom of the graphic illustrates that teachers on average rated females’ performance higher than did the direct cognitive assessments in both math and reading.

In math, both lines exhibit the same trend: Males’ relative performance improves through the end of fifth grade, followed by improved relative performance of females between fifth and eighth grades. Although not presented in this graphic, the scale scores (shown in Table 1) reveal the same pattern. This is compelling evidence that in math—regardless of the assessment or metric used—males’ relative performance improves in elementary school, and females’ relative performance improves during middle school.

In reading, teachers and the direct cognitive assessments evaluate females as performing at 0.2 SDs above males at the fall of kindergarten. Beyond that, however, the trend lines show no systemic similarities. By the spring of eighth grade, the direct cognitive assessment suggests that the gap is where it began in kindergarten (about 0.2 SDs), yet teacher reports suggest that the gap has widened to 0.4 SDs in favor of females.

Since the same concerns about the lack of interval scaling for the direct cognitive scores also apply to the teacher rating scores, we calculated metric-free distributional gaps for the teacher rating scores as well. Figures 6 and 7 give values of $\lambda_g$ throughout the distribution of teacher ratings. If teachers rated the skills and abilities of females higher than those of males, the value of $\lambda_g$ is greater than 0.5.

In math, teachers perceive males to close the achievement gap at the lower end of the distribution and, in fifth grade, to pull away from females at the upper end of the distribution. Interestingly, and mirroring the findings of the direct cognitive score gaps, teachers perceive females to close the
achievement gap during middle school, which results in the eighth-grade gaps looking remarkably similar to the initial gaps at the beginning of kindergarten.

In reading, teachers consistently view females as outperforming males. Additionally, the $\lambda_\theta$ values for the teacher ratings are generally slightly higher than those of the direct cognitive reading assessments (in Figure

**Figure 5.** Average gender achievement gaps by subject and rater type. 
*Note.* The gray bars on each estimated gap are 95% confidence intervals.
4). That is, teachers perceive females as more highly concentrated among the top readers than would be suggested by their direct cognitive scores. Resembling the direct cognitive reading score trend, teachers perceive females to gain ground in reading during middle school, as evidenced by the line moving upward from fifth to eighth grade. Yet across the waves, the direct cognitive scores showed that low-achieving males were

---

**Figure 6.** Metric-free gender math teacher rating gaps through distribution by wave.

*Note.* Values above 0.5 favor females; values below 0.5 favor males. The gray region in each panel is a 95% confidence interval, generated by 500 bootstrapped replications.
particularly falling behind, while this pattern is less consistent in the teacher ratings.

Limitations and Other Considerations

We found no evidence that certain potential confounders (of race, socioeconomic status, or grade level) contributed to differential achievement.
growth for males and females. Additional information on this robustness check is contained in the supplementary online materials.

We acknowledge potential dangers of highlighting gender differences in studies such as this. Readers should be mindful that achievement disparities by gender are generally small, and the distributions are far more similar than different. This study reports disparities as they exist today, and the findings do not suggest that females cannot score as well as males in math or that males cannot score as well as females in reading.

Our analyses did not attempt to explain why differential achievement exists between the genders. Our goal was to (a) highlight areas in the achievement distribution where gaps grow, (b) identify the grade levels when gaps begin to appear, and (c) determine if teachers’ reports of gender gap trends are similar to those of cognitive assessments. We hope that future researchers can use our analyses as they explain the mechanisms of these gaps.

**Discussion**

**Math Results Summary**

This study confirms that there is no overall gender gap in math achievement as students begin kindergarten, regardless of whether one looks at the ECLS-K standardized or scale scores. However, teachers rated kindergarten females’ mathematical skills significantly higher than those of males. By first grade, a math gender gap favoring males appears in the standardized and scale scores, and this gap is particularly pronounced among higher-scoring students. Yet teachers continue to rate first-grade females’ academic proficiency higher than that of males. The math gap then continues to widen between first and third grades, spreading throughout the achievement distribution. This gap holds steady from third to fifth grade at about 0.24 SDs. However, teachers’ ratings do not match this gap, and even by fifth grade, teachers do not rate males higher than females.

Consistent with our finding that females gain ground in middle school on the direct cognitive assessments, teachers’ ratings of females also increased relative to those of males during this time. Specifically, while teachers rated males and females similarly in fifth grade, they rated females 0.2 SDs higher in eighth grade. Overall, by the end of middle school, males still scored significantly ahead of females on the direct math assessment, but teachers rated females’ math achievement significantly higher than that of males.

Traditionally, math education researchers have considered middle and high school an especially important time for the formation of gender disparities. Just over a decade ago, Fennema et al. (1998) and others (e.g., Sowder, 1998) expressed surprise when gender differences in females’ and males’
math strategies were found in early elementary grades. These researchers noted that the males studied tended to use more abstract, invented strategies than did females. Guerra (2009) suggests that differential teacher treatment of males and females—with males more likely to “get away with” breaking classroom rules—might contribute to males’ being less likely to follow given algorithms and to do more creative mathematical problem solving. Forgasz and Leder (2001) note similar patterns in Australian classrooms. Fennema et al. and Sowder (1998) hypothesized that gender disparities in strategy use could contribute to increasing math achievement disparities in later grades. Hence, it is particularly surprising that in this study, females actually gain ground in math in middle school, with the gender gap in achievement halving between fifth and eighth grades. Additionally, although the math achievement gap began as most prominent among the highest-scoring students (those most likely to use abstract strategies), by eighth grade, the disparity was actually largest in the middle of the achievement range.

However, despite that narrowing at the top of the distribution, females composed only 40% of the top 10% of math students in eighth grade, the same as they did in the spring of kindergarten. Although the fact that the top 1% comprised 37% females by eighth grade is a marked improvement from the spring of kindergarten, where only 15% of the top 1% of scorers were female, this persistent disparity at the top of the distribution likely has implications for the representation of females in math-related careers.

Reading Results Summary

In reading, females scored significantly higher than males at the start of kindergarten and were also rated higher by teachers. This study examined if males were simply “late bloomers” who would soon catch up to their female peers in elementary school.

From kindergarten through fifth grade, the gap shrinks according to the standardized scores yet widens according to the scale scores. However, these differences were relatively minor. According to both the metric-based and metric-free distributional analyses, the reading gap favoring females starts out evenly distributed but becomes increasingly pronounced at the lower end of the distribution, with some narrowing of the gap in the upper end through fifth grade. To illustrate, in the fall of kindergarten, the 10th and 90th percentiles of females had an advantage of about 0.12 SDs over their male counterparts. By eighth grade, the gap at the 90th percentile of males and females is one tenth of an SD but grew to about one quarter of an SD for the 10th percentile.

Our metric-free analyses further illustrate these disparities. Assuming that 5% of U.S. adults are illiterate (Kutner, Greenberg, & Baer, 2005), we examine the gender representation below the 5th percentile of the eighth-grade reading distribution, finding that males make up about 67% of this
group. Overall, focusing on the extremes of the distributions allows us to obtain vital information. Specifically, knowing that females at the top of the distribution need attention in math and males at the bottom of the distribution need attention in reading is crucial to supporting students’ lifelong goals and prospects.

**Implications**

The patterns uncovered in this study hold several important implications for researchers, educators, and policymakers. One surprising finding of this study is the narrowing of the math gap during the middle school years. Not only was this narrowing evident throughout the entire achievement distribution on all metrics of the direct cognitive assessments, but even teachers’ assessments suggest that females were gaining ground relative to males between fifth and eighth grades. If we could identify the factors contributing to the narrowing of that middle school gap, we might find ways to capitalize on girls’ momentum during middle school years and close the remainder of the gap. Traditionally, gender intervention programs, such as Summermath, have targeted middle or high school females. It is possible that such interventions and additional teacher education about gender disparities have contributed to the narrowing of math achievement disparities between fifth and eighth grades. However, another possibility—which could also explain the increasing advantage of females in reading—is that females gain ground in both math and reading during middle school because homework becomes a more important method of learning in the middle grades, and females report spending a greater amount of time on homework than do males (Lubienski, McGraw, & Strutchens, 2004). These hypotheses require further research.

Regardless of whether these hypotheses are true, the results of this study suggest that future math-focused interventions with females might be better targeted toward elementary grades than previously thought. Most TIMSS nations exhibit no gender gap at fourth grade, and therefore, these early gender disparities in the United States should not be viewed as inevitable (Mullis et al., 2008). Given Beilock et al.’s (2010) findings that mathematically anxious first- and second-grade female teachers adversely affect their female students’ gender stereotypes and math learning, it is possible that math anxiety among U.S. elementary math teachers is one cause of the early gender disparities identified here. The employment of subject matter specialists in earlier grades could be one approach to addressing this issue.

As a caution, though, we must acknowledge that despite the middle school gap reduction in math, the eighth-grade gaps remain significant at many points in the distribution. Additionally, evidence from PISA and NAEP suggests that math gaps between males and females do not continue to narrow in high school (Ginsburg et al., 2005; Perie et al., 2005). One interesting finding in Correll’s (2001) study is that high school females with strong
math and verbal abilities are less likely to take calculus than other females
with similar math skills but weaker verbal abilities, illustrating Eccles’s
(1986) argument that women are making reasoned choices among options
and not simply avoiding math. Given females’ relatively strong verbal skills
exhibited in the current study, perhaps one promising intervention would be
to help females understand ways in which a combination of math and verbal
skills can be a powerful asset in STEM (science, technology, engineering,
and mathematics) related careers.

Another important finding from this study is that across both math and
reading, females tended to be rated as more knowledgeable than males by
their teachers. In math, this assessment is incompatible with the direct cog-
nitive assessments. Given prior research suggesting that teachers view males
as more mathematically talented than females and that teachers’ views of
girls’ competence is an important factor in shaping girls’ own confidence
in and pursuit of math (e.g., Correll, 2001), it may seem encouraging that
teachers do not appear to view females as mathematically inferior to males.
However, if teachers perceive males as needing additional support to catch
up to females, then the extra attention could give males an added advantage.
Prior studies of teachers’ special education referrals (Hibel et al., 2006) sug-
gest that teachers’ assessments of students’ cognitive abilities are influenced
by how well students behave. This study may suggest that the math struggles
of females, who may be socialized to be less disruptive in class, might be
more often ignored. However, additional studies are needed before causal
conclusions can be reached. In the meantime, however, school personnel
and teacher educators should be aware of the possible tendency for teachers
to mistake females’ compliance with understanding and thereby underesti-
mate females’ need for assistance in math.

On the other hand, the discrepancy between teacher ratings and direct
cognitive assessments could stem from females’ possession of mathematical
problem-solving skills that are somehow not captured by the direct assess-
ments. Future analyses should explore this possibility, as it suggests that
schools should capitalize on females’ particular skills rather than allow
them to fall behind males. Still, given that teachers were asked to rate the
degree to which students “have acquired and/or choose to demonstrate”
knowledge and skills that are taught in the class, it is possible that responses
were biased in favor of females, who might more often display skills taught
by their teachers, as opposed to males, who, according to prior research,
might tend to draw more upon strategies not taught in the math classroom
(e.g., Fennema et al., 1990).

Finally, in terms of the “gender wars,” this study indicates that the ques-
tion of whether schools are “shortchanging” females versus males is not
a simple one. The answer depends upon which grade and subject one exam-
ines, which students one considers, and which outcome variables are used.
According to this ECLS-K analysis, females generally enter kindergarten on
par with males mathematically, but they are significantly behind within a few short years (across the entire achievement distribution), and females are still significantly behind at eighth grade. Hence, despite some encouraging findings, this study suggests that females lose ground in math during their elementary school years. However, in reading, females begin school ahead of their male counterparts, but males at the top of the distribution gain at least as much ground as females, while males in the lowest decile fall substantially behind their female peers. Therefore, for reading, the data suggest that during elementary and middle school, we should focus more attention on the lowest-achieving males.

Despite these patterns, it would be remiss to suggest that schools alone are the cause of achievement differences between the genders—that is, given the limitations of an observational study such as this, we cannot be certain that schools are not trying their hardest to remediate gender differences but that nonschool forces act to exacerbate differences. For this reason, we do not endorse conclusions that “schools shortchange” one group over another. As the literature reviewed earlier indicates, there are many potential contributors to gender disparities, and those aspects should not be assumed to be unimportant. For example, this study indicates that a substantial gender gap in reading exists as students enter kindergarten. Clearly, policymakers and educators need to look both within and beyond the school walls when addressing males’ and females’ early learning needs. And finally, it is worth emphasizing again that disparities within genders are much larger than disparities between genders, and students of both genders deserve instruction that will optimize their reading and math learning.

Notes

This research was funded through a grant from the Institution of Education Sciences, and a prior draft of this article was presented at the 2009 American Educational Research Association annual conference in San Diego, California. The authors thank Andrei Cimpian for helpful comments.

1To reduce the likelihood of “ceiling effects,” the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K) designed an adaptive test that hones in on the underlying ability of the students. As part of this, three different test forms were designed (low, middle, and high). Less than one tenth of 1% of test takers attained perfect scores on the high form in any wave from fall of kindergarten through spring of fifth grade (see the eighth-grade psychometric report, Tables 4–6, at http://nces.ed.gov/pubs2009/2009002.pdf). In eighth grade, less than 1% of all test takers attained a perfect score on the highest test form.


3The majority of ECLS-K attrition was the result of students’ moving to other schools. The National Center for Education Statistics (NCES) followed a stratified random subsample of movers each year. ECLS-K sampling weights take these factors into account, adjusting the weights of the remaining students to more accurately represent all students, including those whom ECLS-K did not follow. Differential attrition (between males and females) on the basis of characteristics not adjusted for by NCES could bias estimates in...
an unknown direction. However, given the equal representation of males and females across race groups and socioeconomic groups, such bias is unlikely.

4For example, if ECLS-K were to add more “difficult” items, then the item response theory scale score would show a large gap between those students who could and could not answer those items, thus stretching the scale at the upper end. Of course, this also illustrates the inherent difficulty in believing that any measure is truly interval scaled.

5More details can be found in chapter 4 of the ECLS-K eighth-grade user’s manual at http://nces.ed.gov/pubs2009/2009002.pdf.

6These lines represent the scores of males (or females) at a given percentile at a given wave (e.g., the 90th percentile of males in the spring of kindergarten, the 90th percentile of males in the spring of first grade, etc.). They are not “trajectories” tracking a student who begins kindergarten at, say, the 90th percentile and then mapping out his or her achievement through eighth grade.

7The placement of the labels on the right axis is at the point where students have a 50% chance of being proficient in that skill level (which is also the approach taken by LoGerfo, Nichols, and Reardon, 2006). Thus, the placement of the skill-level label reflects where students are in the process of acquiring that skill, not where they have mastered it.

8These analyses compared differences in mean teacher ratings for males and females with differences in mean standardized cognitive assessment scores. Curious if teachers’ ratings were more closely aligned with males’ or females’ direct cognitive assessment scores, we compared the correlations between teacher ratings and students scores for males and females. We found no significant differences.

9National Assessment of Educational Progress results mirror the ECLS-K results in that the gender gap of the 2003 fourth-grade sample was roughly double that of the 2007 eighth-grade sample (0.1 standard deviations at fourth grade versus 0.06 standard deviations at eighth grade). Data from the NAEP Data Explorer: http://nces.ed.gov/nationsreportcard/naepdata/dataset.aspx.

References


Manuscript received June 12, 2009
Final revision received April 18, 2010
Accepted April 19, 2010
Supporting Online Materials

Weighted least squares (WLS) regression. To estimate average achievement differences we use WLS regression, in which students are weighted by their longitudinal sampling weight (ECLS-K variable c1_7fc0):

\[ Y_{iw} = \alpha + \beta M_i + \epsilon_{iw}, \]  

where \( Y \) is the assessment score (e.g., scale score, standardized score, teacher rating score) of child \( i \) at wave \( w \). This score is predicted by a constant term and an indicator variable \( M \), taking on the value 1 for males and 0 for females. Therefore, \( \beta \) is the average amount by which males outscore females (where negative values indicate females outscore males on average).

Potential confounders. We explored the data for potential confounders that could contribute to differential achievement growth for males and females. The main variables we examined were race, socioeconomic status, and grade-level. The results of these tests suggested the gap development trends for reading and math in this paper are not due to differences between males and females in any of these dimensions.

Chi-squared tests were performed on the final analytic sample to see if there existed evidence of variation in the percentage of males and females by race group (\( \chi^2(8, N = 7,075) = 8.20, p = 0.41 \)) or socioeconomic status quintile (\( \chi^2(5, N = 7,075) = 3.17, p = 0.67 \)). Neither exhibited any suggestion of differences between the genders. The chi-squared test for grade-level in wave-7, however, suggested highly significant differences between males and females: \( \chi^2(5, N = 7,075) = 42.53, p < 0.001 \). Specifically, males were far more likely to be in grades 6 or 7. As a result, we performed an additional set of analyses, restricting the sample to only the students who were in 8th grade at the wave-7 data collection (\( N = 6,471 \)). These analyses were remarkably similar to the original table of average differences presented in Table 1. That is, the
gap changes favored males during elementary school and favored females during middle school; moreover, the estimates from these new analyses were statistically indistinguishable from the original estimates in Table 1.

Readers interested in race interactions with gender can refer to Husain and Millimet’s (2009) analysis of K-3rd grade ECLS-K data.

Quantile regression. Because we are concerned with the gender achievement gap across the achievement distribution, we need a method that will yield estimated gaps at specific achievement levels, such as the 10th percentile and the 75th percentile—this is contrasted with the WLS approach above, which yields conditional mean differences.\(^1\) To obtain these quantile-specific estimates of the gap, we employ quantile regression (Koenker & Bassett, 1978). The gender achievement gap \(\beta\) in a content area is obtained by solving the minimization problem:

\[
\min_{\beta} \left[ \sum_{\{i|y_{iw} \geq \beta M_i\}} \theta |y_{iw} - \beta M_i| + \sum_{\{i|y_{iw} < \beta M_i\}} (1 - \theta) |y_{iw} - \beta M_i| \right],
\]

where \(\theta\) is the quantile at which we are estimating the gender gap \(\beta\). That is, quantile regression balances the sum of the absolute residuals above and below the regression line by applying different weights to values above and below a given quantile. In our case of only one dependent indicator variable (i.e., being male), quantile regression simply estimates the value of \(Y\) at, e.g., the 90th percentile for males and the 90th percentile for females, and then presents the difference. That is, we can think of this very simple case of quantile regression as \(\beta_{.9} = Y_{.9}^m - Y_{.9}^f\), where .9 indicates the 90th percentile of each separate (male and female) distribution of outcome \(Y\).

Similarly, we estimate this for the 10th, 25th, 50th, and 75th percentiles in each subject and wave.

\(^1\) Note that in Equation A1, there are no covariates other than the indicator for being male, so this is a specific case which yields just the unconditional average difference between males and female.
Estimating the metric-free distributional measure ($\lambda_\theta$). Let $\phi_m(\theta)$ and $\phi_f(\theta)$ be the cumulative distribution functions for males and females observed by the $\theta^{th}$ percentile of the overall distribution. Males and females are represented in equal proportions; therefore, if the lowest-scoring female scored above the highest-scoring male, then only females would be found above $\theta = 50$, and only males would be found below that point. Thus, our metric-free index tells us the proportion of females (relative to the proportion of males) concentrated above or below a given percentile of the overall distribution. In addition, we want index values above a given point (say, 0.5) to indicate an advantage for one group (say, females), and index values below that point to indicate an advantage for the other group (say, males).

For females, it is desirable to have a larger proportion of your group above the median. Therefore, for values of $\theta \geq 50$, we divide the proportion of females above $\theta$ by the sum of the proportion of males above it and proportion of females above it. If, proportionately, more females have scores above any of these values of $\theta$, then our index will be greater than 0.5. Conversely, for females, it is desirable to have a relatively lower proportion of the group represented for any value below the median (i.e., for values of $\theta < 50$). For index values above 0.5 to still favor females, the index numerator is the proportion of males below a given value of $\theta$, and the denominator is the sum of the proportion of males below that value and the proportion of females below it. Therefore, our index ($\lambda$) is represented by the conditional

---

2 Note that the continuity of the index at the median (where Equation A3 changes conditionally) is due to the fact that males and females are represented in roughly equal proportions in the sample (and population). If this measure were applied to study a gap where the groups were less equally represented, we suggest conditioning Equation A3 at the value of $\theta$ equal to the proportion of the group whose favor is indicated by values below 0.5. For example, if studying the black-white gap and blacks were 30% of the sample (and whites were 70%), and with values below 0.5 favoring whites, use the top equation in A3 for values below $\theta = 70$ and switch to the bottom equation for values above it. The interpretation would remain the same. The reason for changing the value on which equation A3 is conditioned is simple: if the lowest-scoring black student scored above the highest-scoring white student, then the value of the index should be 1 throughout—this is only achieved if conditioning A3 on $\theta = 70$ in this case (or where $\theta = 50$, in the male-female case of equally-represented groups).
equation:

\[
\lambda_{\theta} = \begin{cases} 
\frac{\Phi_m(\theta)}{\Phi_m(\theta) + \Phi_f(\theta)} & \text{if } \theta < 50 \\
1 - \frac{\Phi_f(\theta)}{2 - [\Phi_m(\theta) + \Phi_f(\theta)]} & \text{if } \theta \geq 50
\end{cases}
\]  

(A3)

**Difference between “proportional difference” values and \( \lambda_{\theta} \).** An alternative option to \( \lambda_{\theta} \) is Ho and Haertel’s (2006) “proportional difference” (PD):

\[
PD(\theta) = \Phi_f(\theta) - \Phi_m(\theta),
\]  

(A4)

which here would subtract the proportion of males observed from the proportion of females observed by a given percentile of the overall distribution. The interpretation is simple, an obvious appeal to this method. The drawback is that it diminishes the perceived importance of differences at the extremes of the distribution by not accounting for the reduced proportion of data in the tails of the distribution. The PD is a measure of absolute difference, whereas our index \( \lambda_{\theta} \) is a measure of relative difference, scaling for the potential to observe a score by a certain point. For example, if by the 5th percentile 10% of females were observed, but 0% of males were observed, the difference would be 10% in the PD metric. Assume then that males and females were observed in equal proportions such that by the 50th percentile, 45% of males and 55% of females were observed, maintaining the 10% difference in the PD measure. In our measure, the difference in proportion observed by the 5th percentile would mean the numerator is .1 and the denominator is (.1 + 0), resulting in a value of 1. Since the groups are then represented in equal proportions for values between the 5th and 50th percentiles, the value of the measure at the 50th percentile is .55 \([= .55 / (.55 + .45)]\), reflecting the fact that the gap in the proportion of each group observed reduced. The choice to use the PD measure or \( \lambda_{\theta} \) depends on the particular research question, including whether or not the tails of the distribution are of
This builds off of recent research in achievement gaps (Penner, 2008; Xie & Shauman, 2003), which used logits (log odds, from a logistic regression model) to represent gaps. In our case, for any value of $\theta$, the logistic regression results for males and females (predicting the probability $p$ that individual $i$ has an achievement score at or below the $\theta^{th}$ percentile of achievement aggregated over males and females) are:

$$[p_i(\theta) | \text{male}] = \Phi_m(\theta) = \left[ 1 + \exp[-(\beta_0 + \beta_1)] \right]^{-1}$$

$$[p_i(\theta) | \text{female}] = \Phi_f(\theta) = \left[ 1 + \exp[-\beta_0] \right]^{-1}$$

Thus, using the logistic regression as the basis for $\lambda_\theta$, we can rewrite (A3) as:

$$\lambda_\theta = \begin{cases} 
\left[ 1 + \exp[-(\beta_0 + \beta_1)] \right]^{-1} & \text{if } \theta < 50 \\
\frac{1 - \left[ 1 + \exp[-\beta_0] \right]^{-1}}{2 - \left[ 1 + \exp[-(\beta_0 + \beta_1)] \right]^{-1} + \left[ 1 + \exp[-\beta_0] \right]^{-1}} & \text{if } \theta \geq 50
\end{cases}$$

(A5)

This logistic-regression-based formulation of $\lambda_\theta$ will likely be of greatest help to researchers wishing to condition on other factors (e.g., to estimate the black-white achievement gap conditioned on socioeconomic status). In such cases, the variables conditioned on would simply become part of the right-hand-side logistic regression, and the gaps would be estimated at some common value (e.g., the mean of each conditioned variable). Yet, the benefits of converting logistic-regression estimates to $\lambda_\theta$ (instead of just the logistic regression log-odds) are that (a) log-odds can be difficult to interpret and (b) $\lambda_\theta$ is bounded by (0,1), with a value of 0.5 signifying parity, whereas log-odds can take on values $[0,\infty]$. 
References


