Abstract Title Page

Title:

Patterns of Change in U.S. Gender Achievement Gaps during Elementary and Middle School

Authors and Affiliations:

Erin Fahle, Stanford University

Abstract Body

Background / Context:

Research on gender achievement gaps shows they exist, and are largest in the tails of the distribution, starting as early as Kindergarten and persisting through eighth grade.

In mathematics, studies find small average gender achievement gaps and larger systematically male-favoring gaps among the highest achieving students. Using the Early Childhood Longitudinal Study-Kindergarten (ECLS-K) 1998-99 cohort data, multiple studies find that an average mathematics gap favoring boys emerges by the end of 1st grade, grows to approximately 0.20 SD by the end of 3rd grade, and persists through 5th grade (Lee, Moon and Hegar, 2011; Husain and Millimet, 2009; Fryer and Levitt, 2009; Robinson and Lubienski, 2011; Sohn, 2012). Additionally, work using ECLS-K finds small, but significant gaps exist at the upper tail of the mathematics score distribution starting as early as Kindergarten, and that these gaps grow and extend to the rest of the distribution by 3rd grade (Husain & Millimet, 2009; Robinson & Lubienski, 2011; Fryer & Levitt, 2009).

In contrast, in ELA, studies show that a substantial average gap favoring females exists in ELA as early as the fall of Kindergarten and that this gap remains fairly static through fifth grade using the ECLS-K data (Husain and Millimet, 2009; Robinson and Lubienski, 2011; Fryer and Levitt, 2009; Chatterji 2006). Similar to mathematics, prior work also find that it is critical to examine the tails of the distribution. The gap among the lowest achieving students appears to be larger than the average gap for all grades (Robinson and Lubienski, 2011).

Despite showing evidence that the average gaps and gaps among the highest and lowest achieving students do exist across grades, potentially changing in magnitude and significance as students matriculate from Kindergarten through eighth grade, prior research does not explicitly model these changes in the gaps across grades.

Purpose / Objective / Research Question / Focus of Study:

This paper seeks to understand underlying patterns in how gender achievement gaps grow, shrink or stay constant as students move through elementary school and middle school. The focus of this paper is not only to capture these trends for average gender achievement gaps, but also for gaps in the tails of the achievement distribution as those have been demonstrated critical areas of analysis by prior work in the field and may exhibit different patterns than those at the mean. Understanding when gaps change, and in what direction, provides important information about the mechanisms that may produce the changes the gaps and highlight critical time-points for intervention, sparking future lines of research in this field.

Setting:

This study focuses on gender achievement gaps within the United States, with particular attention to how these gaps change as students progress from first through eighth grades. It is a longitudinal study of changes in the gaps over grades, terms and cohorts. The gaps are estimated at the district level, providing a higher resolution analysis of patterns of change.

Population / Participants / Subjects:

The population of study is U.S. school districts serving first through eighth grade students that use the Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) assessment. The MAP assessment is a low-stakes formative assessment designed to help teachers better evaluate and assist their students throughout the school year. Approximately 3,700 school districts across the U.S. use the MAP assessment. The data include fall and spring test scores for nine grades (Kindergarten through 8th grade) over nine school years (2005 through 2013) with test records for the approximately 7 million students, 15,000 schools, and 3,700 districts in the dataset from all U.S. states. The Kindergarten school year is not used in this study due to the significantly reduced administration of the test in that grade across districts.

This sample is not representative of all U.S. elementary and middle school districts; however, due to the large number of participating districts it is a useful starting point to understand changes in the gaps across grades, particularly in states where a majority of districts participate. Additionally, this data offers the unique advantage of having both fall and spring test scores for a subset of the districts, which enables the investigation of how gaps change during the summer as compared with the school year that has not been the focus on gender achievement research to date.

Intervention / Program / Practice:

This study is a secondary data analysis of the MAP assessment data. It focuses on characterizing how district-level gender achievement gaps change as students matriculate from first through eighth grade.

Research Design:

To enable comparability across grades, this study uses metric-free measures of the mean and tail gender achievement gap, specifically the V-Statistic (Ho, 2009; Ho & Haertel, 2006; Ho & Reardon, 2012) and the Proportion-Adjusted Relative Difference (Robinson & Lubienski, 2011). The V-statistic is calculated as: $V = \sqrt{2} * \Phi^{-1}(P_{a<b})$ where Φ^{-1} is the inverse cumulative density function for the standard normal distribution. V is interpretable as the gap in common standard deviation units between two groups, and is invariant to monotone scale transformations. It is further equal to Cohen's d when the group distributions are normal (with either equal or unequal variances), yielding an interpretation as a scale-invariant effect size that is comparable across tests, times or grades (Ho, 2009; Ho and Haertel, 2006, Ho and Reardon, 2012; Reardon and Ho, 2015).

The Proportion-Adjusted Relative Difference (λ_θ) is percentile-specific proportion, which provides information about the representation of males compared with females at different points (or percentiles) of the distribution.

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

where $\Phi_m(\theta)$ and $\Phi_f(\theta)$ are the cumulative density functions for males and females at the percentile, θ. For percentiles less than 50 (greater or equal to 50), λ_θ is the proportion of students scoring at or below (above) the given percentile who are male (female). Note that $\lambda_\theta =$

0.5 indicates gender parity, $\lambda_\theta > 0.5$ indicates a female advantage, and $\lambda_\theta < 0.5$ indicates a male advantage, where advantage is defined as underrepresentation in the lower tail and over representation in the upper tail of the achievement distribution.

The use of metric free measures enables the comparison of gender achievement gaps across grades without concerns about vertical-scaling or comparable test metrics, that unmet would distort the trend analysis.

Data Collection and Analysis:

This study leverages the longitudinal nature of the NWEA MAP assessment data in order to track the district-level gaps over grades and terms for multiple cohorts. Three different types district gaps for each subject-by-grade-by-term-by-cohort case are estimated: (1) a mean gap; (2) a lower-tail gap ($\theta = 10$); and (3) an upper-tail gap ($\theta = 90$) using the V-statistic and proportion-adjusted relative difference, described above. To analyze the patterns in these three metric-free gaps across grades, this study adopts a change in gap framework, modeling changes in the subject-by-grade-by-term-by-cohort gaps using Hierarchical Linear Modeling (HLM) with multiple functional forms for the grade and term variables. The baseline, linear grade-trend model is as follows:

$$G^s_{dgtc} = \pi_{dgtc} + \epsilon_{dgtc}$$
$$\pi^s_{dgtc} = \beta_{00d} + \beta_{100}g^* + \beta_{200}c^* + \beta_{300}f + \epsilon_{dgtc}$$
$$\beta_{00d} = \gamma_{000} + u_{00d}$$

where G^s_{fgtc} is the V gap or λ_θ statistic in subject, s, for a district, d, grade, g, term, t, and cohort, c; g^* is the student grade level grand mean centered at 5.5; c^* is the year in which the students were in Kindergarten grand mean centered at 2004.5; and, f is an indicator variable for whether the test was administered in the fall term. The models are precision-weighted to account for the fact the gaps are estimated. Additional specifications of this model include: (1) a fully non-parametric specification for grade-by-term (i.e. indicators for each grade-by-term combination) and (2) including random coefficients on the linear and non-parametric grade-term trends.

Investigating multiple functional forms (i.e. linear, non-parametric) stems from trying to narrow the hypothesized mechanisms by understanding if gaps consistently increase or decrease over grades (linear trend) or if the changes are more sporadic (nonparametric trend). Including random slopes enables the characterization in the variability across districts in these trends.

Findings / Results:

This study finds that gender achievement gaps in both mathematics and reading change meaningfully as students progress through grades (please insert figure 3 here). Specifically, changes in the gaps are best captured using a fully nonparametric trend on grade-by-term. Use of a linear trend masks the period of increase followed by a period of decrease in the gaps at different grades. In both subjects, the trend in the average gaps that emerges is that gaps change in favor of males in elementary school (growing the mathematics gap, shrinking the reading gap), and in favor of females in middle school (shrinking the mathematics gap, growing the reading gap). These trends hold for the upper and lower tail gaps, as well, with the exception of the lower tail in reading where male students continuously fall behind.

Conclusions:

This paper adds to the current literature through systematically analyzing and characterizing the changes in the gaps as students move through elementary and middle school. Particularly it offers two key advantages: (1) the use of district-level longitudinal data with multiple cohorts across eight grades provides robustness to the observed results; and (2) the use of metric-free measures ensures comparability of the gap-sizes across grades. It further provides intuition as to the potential mechanisms driving these changes, sparking critical lines of future research. Planned extensions include jointly modeling the subjects due to similar trends across the gaps, uncovered in the separate subject analyses, and further investigating the differences in summer and school year gap trends.

Appendices

Appendix A. References

Andreescu, T., Gallian, J. A., Kane, J. M., & Mertz, J. E. (2008). Cross-Cultural Analysis of Students with Exceptional Talent in Mathematical Problem Solving. *Notices of the AMS*, *55*(10), 1248–1260.

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.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender Differences in Multiple-Choice Tests: the Role of Differential Guessing Tendencies. *Journal of Educational Measurement*, *28*(1), 23-35.

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.

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.

Fryer Jr, R. G., & Levitt, S. D. (2009). *An empirical analysis of the gender gap in mathematics* (No. w15430). National Bureau of Economic Research.

Gamer, M., & Engelhard Jr, G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12*(1), 29-51.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science 320*(5880), 1164.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics*, *34*(2), 201-228.

Ho, A. D., & Haertel, E. H. (2006). Metric-free measures of test score trends and gaps with policy relevant examples (CSE Report No. 665). Los Angeles, CA: *Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies.*

Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal “proficiency” categories. *Journal of Educational and Behavioral Statistics*, *37*(4), 489-517.

Husain, M., & Millimet, D. L. (2009). The mythical “boy crisis”? *Economics of Education Review*, (28), 38–48. doi:10.1016

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences*, *106*(22), 8801–8807.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science, 321*(5888), 494-495.

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.

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.

Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research*, (37), 239–253.

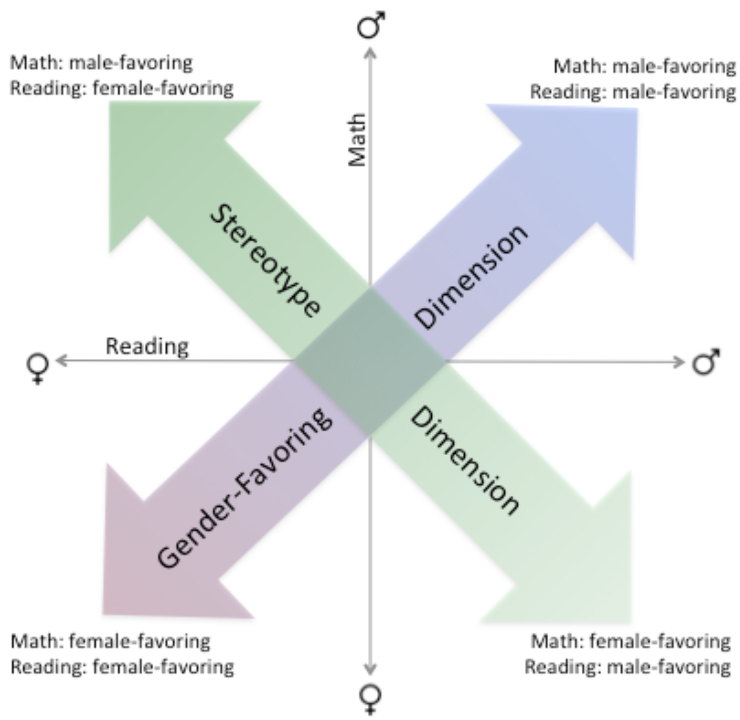
Pope, D. G., & Sydnor, J. R. (2010). Geographic variation in the gender differences in test scores. *The Journal of Economic Perspectives*, *24*(2), 95-108.

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.

Sohn, K. (2012). A new insight into the gender gap in math. *Bulletin of Economic Research, 64*(1), 135-155.

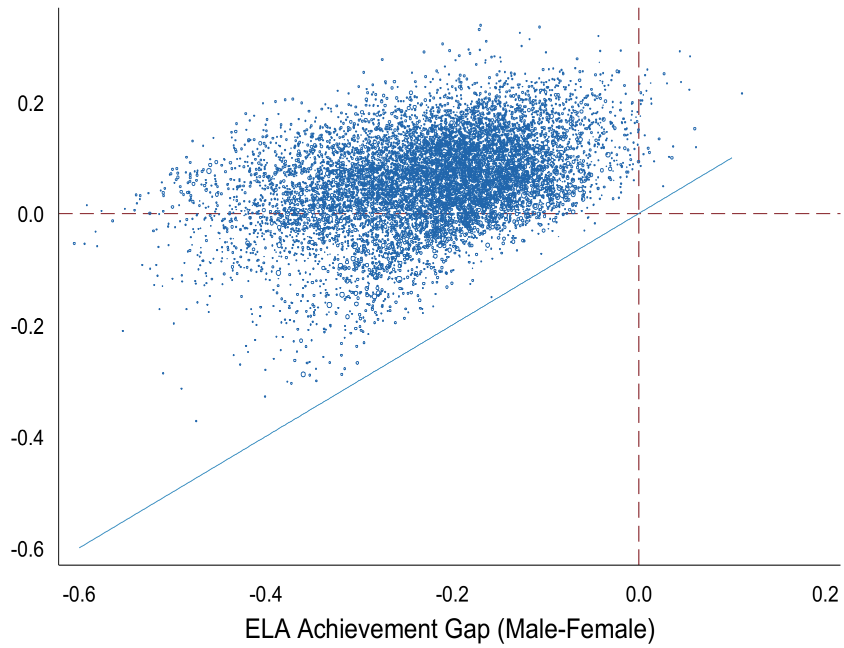
Appendix B. Tables and Figures

Figure 1: Dimensions of Gender Achievement Gaps



Notes: Reading gaps are plotted on the x-axis and corresponding mathematics gaps on the y-axis. Positive (negative) values indicate gaps are male-favoring (female-favoring). Gender equality in the gaps is at the origin.

Figure 2: Male-Female Mathematics and ELA Achievement Gaps, School Districts 2009-2012



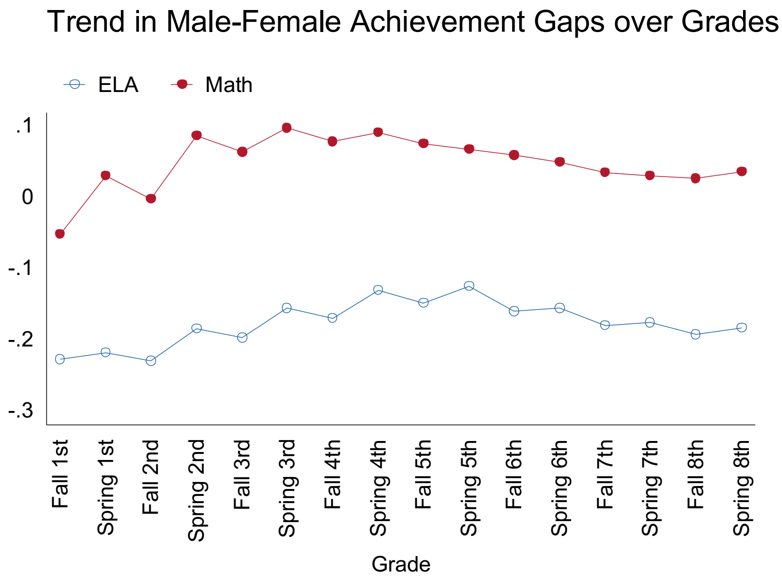
Notes: Reading gaps are plotted on the x-axis and corresponding mathematics gaps on the y-axis. Positive (negative) values indicate gaps are male-favoring (female-favoring). The model used to estimate average gaps includes state fixed effects and adds the average state NAEP gap to the Empirical Bayes estimate.

Table 1: Relationship between Proportion Multiple-Choice Items on State Tests and the Size of Gender Gaps, State-Level

	Mathematics				ELA			
	Grade 4		Grade 8		Grade 4		Grade 8	
Model 1: State-Level NAEP Audit Test								
Proportion Short Response+	-0.135	**	-0.109		-0.223	*	-0.376	**
Extended Response	(0.041)		(0.075)		(0.084)		(0.113)	
Model 2: District-Level NWEA Audit Test								
Proportion Short Response+	-0.126		-0.151	*	-0.296	**	-0.389	***
Extended Response	(0.122)		(0.068)		(0.098)		(0.101)	

All models are weighted by $1/se^2$. Standard errors that are clustered by state. Model 1 includes data from 2009 Ed Facts and NAEP data from grades 4 and 8. Model 2 data from 2009 Ed Facts and NWEA data sources from grades 4 and 8. The models are restricted to state or district by grade cells with gap data from both Ed Facts and NAEP/NWEA. Both models also include the proportion of "other" (not shown) items.

Figure 3: Trend in Male-Female Achievement Gaps over Grades



Notes: Estimated trends are taken from 3-level precision weighted HLM models with a non-parametric grade-by-term