Abstract Title Page

Title: Trends in Academic Achievement Gaps in the Era of No Child Left Behind

Authors and Affiliations:
Sean F. Reardon, Stanford University
Erica Greenberg, Stanford University
Demetra Kalogrides, Stanford University
Kenneth A. Shores, Stanford University
Rachel A. Valentino, Stanford University

Background / Context:

Achievement gaps between white students and their black and Hispanic peers, and between students from high- and low-income families, remain a stubborn problem in the U.S. One of the explicit goals of the 2001 No Child Left Behind (NCLB) was to narrow these gaps. While there is some evidence that NCLB produced modest increases in average levels of academic achievement (as measured by NAEP; see Wong, Cook, & Steiner, under review, and Dee and Jacob, 2009), it is less clear whether NCLB has narrowed achievement gaps.

Over the last two decades, most scholarship has relied on the National Assessment of Educational Progress (NAEP). NAEP comes in two forms: a long-term trend assessment (NAEP-LTT), administered roughly every four years to 9-, 13-, and 17-year-old students, and a "Main NAEP," taken by fourth- and eighth-graders every two years and twelfth-graders roughly every three years. Both NAEP-LTT and Main NAEP find that math and reading gaps between white and black students have narrowed over the last forty years (Grissmer, Flanagan, & Williamson, 1998; Hedges & Nowell, 1998; Neal, 2005; Vanneman, Hamilton, Baldwin Anderson, & Rahman, 2009). Hispanic-white gaps in the NAEP-LTT grew in the mid-1990s and have been closing since then (Reardon & Robinson, 2007; Hemphill & Vanneman, 2011).

There is a great deal of imprecision in NAEP estimates, however, due to their modest sample sizes and relatively infrequent collection. Moreover, although Main NAEP provides state-specific gap estimates, they are even more imprecise due to small within-state samples. As an alternative to NAEP, state accountability data, collected in accordance with NCLB, can be used to estimate state-specific achievement gaps. Because state tests are administered to virtually all students in grades 3-8 each year, they provide a much richer source of information for investigating trends in achievement gaps (although they are not without their problems, including issues of content comparability, possible cheating, etc; we discuss these concerns in the paper). Many analyses of such data rely on comparisons of the percentage of students scoring proficient (e.g., Kober, Chudowsky, & Chudowsky, 2010). Because states determine their own definitions of proficiency, however, these estimated differences in group achievement "are susceptible to striking distortions" that "are neither fully predictable nor easily remedied" (Ho, 2008, p. 352; cf. Furgol, Ho, & Zimmerman, 2010; Holland, 2002). Thus, in order to use state test score data, it is necessary to solve the problem of measuring gaps with proficiency rates. In this paper, we rely on the method developed by Ho and Reardon (forthcoming) to do so.

Purpose / Objective / Research Question / Focus of Study:

Our goals in this study are to use both NAEP and state accountability test score data to 1) provide a detailed description of the magnitude and trends of state-level academic achievement gaps among cohorts of students entering school in the 1990s and 2000s; 2) investigate the extent to which patterns and trends in gaps vary among states; and 3) provide preliminary evidence regarding the impact of NCLB on achievement gaps.

Setting:

This analysis covers all 50 states, and includes students in grades 3-8 in public schools.

Population / Participants / Subjects:

We use Main NAEP test score data from 4^{th}- and 8^{th}-graders in 1990-2009. We also use state-level categorical proficiency count data for students in grades 3-8 in 2001-2010, collected with the help of state department of education officials, as well as the Office of Planning,

Evaluation, and Policy Development at the U.S. Department of Education. We are primarily interested in white-black and white-Hispanic gaps, though we also report female-male achievement gaps.

Intervention / Program / Practice:

The intervention of interest is the NCLB legislation, though this paper is also generally interested in describing state-level patterns and trends in achievement gaps, as well as in characterizing the level or heterogeneity across states in achievement gaps and their trends.

Research Design:

This quantitative study is both descriptive and explanatory. First, we aim to chart achievement gap levels and trends across a variety of states, subjects, grades, and years. Using data from the Main NAEP 4^{th} and 8^{th} grade assessments as well as grade 3-8 data from state accountability systems, we estimate the achievement gap in each state-by-grade-by-year-by-subject for which we have available data. The first part of the paper describes these results, noting trends across cohorts of students, trends across grades, differences by subject (math or reading), and variation among states in each of these.

The second part of the paper investigates the extent to which patterns in our results suggest that NCLB may have had some effect on achievement gaps. We rely on three types of evidence for this analysis. First, we ask whether gaps narrowed during the NCLB era. Second, we examine the extent to which gap trends changed with the introduction of NCLB in ways that would suggest a causal effect. And third, we investigate whether states in which NCLB exerted more pressure to improve the achievement of low-performing groups showed larger changes/decreases in achievement gaps than states in which NCLB exerted less pressure. For this part of the analysis we draw on the state accountability typologies used by Wong, Cook, and Steiner (under review) and Dee and Jacob (2009) in their evaluations of the effects of NCLB on average levels of achievement. These analyses may not provide a rigorous test of the impact of NCLB on achievement gaps, but they will provide prima facie evidence of alignment—or misalignment—between observed changes and those we would expect if NCLB did, in fact, narrow achievement gaps.

Data Collection and Analysis:

We estimate achievement gaps using data from Main NAEP and categorical proficiency data from state-level accountability tests introduced under No Child Left Behind. We have state-level Main NAEP data from 1990-2009. We have state accountability test data for 2006-2007 through 2009-2010 for all 50 states, provided to us in a uniform file by the U.S. Department of Education. For 17 states, we have collected additional information for prior years from state departments of education. (We will likely have more states with earlier data within the next few months.) Table 1 below describes the state data we have currently.

(please insert Table 1 here)

NAEP and state accountability data are tabulated by subgroup, subject, grade, and year, allowing us to estimate gaps for subgroups defined by race/ethnicity, gender, and socioeconomic status. Later, we plan to collect and analyze data on additional subgroups, such as English Language Learners and their English-proficient counterparts.

To estimate achievement gaps, we compute the statistic V (Ho and Reardon, forthcoming) for each state-grade-year-subject cell. This measure is analogous to Cohen's d (and can be interpreted as the difference in mean scores between groups, divided by their pooled standard deviation) but is insensitive to the test metric in which scores are reported. This characteristic enables us to compute comparable gap measures across states even when we cannot be sure the test metrics used in different states' accountability systems, and in different years, are linearly equivalent. Moreover, the statistic V can be computed from data describing counts of students in each of 4 or 5 proficiency categories; that is, we do not need to know the means and standard deviations of the full test score distributions in order to compute V. As a result, we can estimate state-level gaps from data reported in NCLB-style accountability metrics (i.e., percents proficient, rather than mean scores). Ho and Reardon (forthcoming) show that V can be estimated very reliably (typically with errors less than 0.01-0.02 standard deviations of the true value) using proficiency count data.

We fit models of this form:

$$V_{scgt} = \Gamma_s + \Delta_c + \Lambda_g + \beta M_t + e_{scgt}, \;\; e_{scgt} \sim N\left[0, \sigma^2_{scgt}\right], \tag{1}$$

where V_{scgt} is the estimated achievement gap in subject t in state s, for students in cohort c and grade g. The Γ_s is a vector of state fixed effects; Δ_c is a set of cohort fixed effects; Λ_g is a set of grade fixed effects; and M_t is a dummy variable indicating whether the gap is for math or reading scores. We fit these models via WLS, using the inverse of the squared standard error of V as a precision weight. The parameters of interest here are the sets of fixed effects. In particular, we are interested in whether the vector Δ_c indicates that gaps narrow across cohorts.

Equation (1) presents our most general pooled model, using gap estimates from both reading and math, in all states, grades, and available years to estimate the trends in gaps across cohorts. In the paper, we fit a series of variations on this model, in some cases including linear trends across cohorts and grades, and in some case relaxing the assumption that cohort trends are constant across grades, states, or test subjects. We also fit models that include a dummy variable indicating whether a given cohort was subject to NCLB in a given grade. Using NAEP data, where we have a long time-series of of cohorts, we fit interrupted time-series models to examine whether the trend in achievement gaps changes abruptly in 2002, when NCLB takes effect.

We also fit random coefficient models, allowing the cohort trend to vary among states:

$$\begin{gathered} V_{scgt} = \alpha + (\gamma + v_s) COH_c + \Lambda_g + \beta M_t + u_s + e_{scgt} \\ e_{scgt} \sim N\left[0, \sigma^2_{scgt}\right]; \begin{bmatrix} v_s \\ u_s \end{bmatrix} \sim \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_v & \tau_{uv} \\ \tau_{uv} & \tau_u \end{pmatrix} \right], \end{gathered} \tag{2}$$

where COH_c is a linear cohort term. The key parameter of interest here is τ_v, the variance of cohort trends across the 50 states. These models allow us to test whether the cohort trends vary among states (and to then test hypotheses regarding predictors of state-specific trends). We use models of this form as one way of investigating whether reductions in achievement gaps were larger in states most likely to experience accountability pressures induced by NCLB.

Findings / Results:

To date, we have computed white-black, white-Hispanic, and female-male achievement gaps in math and reading for all 50 states from 2006-2007 through 2009-2010, with additional years for a smaller number of states, using the methods described above. The gaps we estimate

are much more precise than gaps generated from state-NAEP because they are based on the entire population of tested students within a state-grade, whereas the NAEP samples are limited to a few thousand students per state. Our estimated gaps correlate with estimated NAEP gaps above .80 for 4th and 8th grade math and reading in 2005 and 2007.

We estimate state gap trends in three ways. First, in Table 2, columns (1)-(3), we fit a regression model similar to the form shown in Equation (1) above, albeit with linear cohort and grade terms for parsimony, to gap estimates for 3rd-8th grade students in all 50 states from 2007 to 2010. Gaps are calculated in the order implied by their labels (i.e., white-black gaps are the difference in white and black achievement levels, and are generally positive); negative coefficients signify a trend in gap closure. The results of these simple pooled models indicate that the white-black and white-Hispanic gaps are getting smaller across cohorts, by 0.013-0.019 standard deviations per year. Gender gaps have not changed significantly across the nine cohorts of data included in these models. White-Hispanic gaps appear to narrow as cohorts progress through school. White-black gaps are somewhat larger in math than in reading, while white-Hispanic gaps are smaller in math than in reading.

(please insert Table 2 here)

Next, we examine cross-state variation in achievement gap trends. Table 2, columns (4)-(6), shows the results of a random coefficients model (like that in Equation (2)). The estimated standard deviation in the cohort trends is relatively large in comparison to the average trend (implying that gaps are widening in some states and narrowing in others), and we reject the null hypothesis of homogenous cohort trends across states.

Finally, in Figures 1-3 we plot cohort trends in achievement gaps, estimated using WLS regression with state, cohort, and grade fixed effects (as in Equation 1, above). By way of comparison, we overlay the gaps estimated from Main NAEP data using the same model. Figures 1 and 2 show a general decline in white-black and white-Hispanic achievement gaps in the last decade. The NAEP data, however, suggest that these declines began with cohorts who were in school prior to 2002, when NCLB was implemented, casting doubt on the possibility that NCLB is responsible for the recent declines. The female-male gap does not appear to have changed significantly over the same time period.

(please insert Figures 1-3 about here)

For 17 states, we have proficiency data covering a longer time span than those provided to us by the U.S. Department of Education. Accordingly, we re-fit the models from Table 1 using data for children enrolled in third through eighth grades between 2000 and 2010. With these longer panel data, the results are similar to those in Table 2 (results not shown).

Analyses that explicitly test the effect of NCLB on achievement gaps are currently being conducted and will be included in the final paper.

Conclusions:

Our findings to date indicate, first, that black-white and Hispanic-white achievement gaps have narrowed in the last decade or more. Male-female gaps appear largely unchanged over the same time period. Second, there is considerable variation across states in both the magnitude and trends in achievement gaps. Third, the patterns evident so far do not suggest a strong effect of NCLB on achievement gaps, though these analyses are not yet complete.

Appendices

Appendix A. References

Bollinger, C. R. (2003). Measurement error in human capital and the black-white wage gap. *Review of Economics and Statistics*, *85*(3), 578–585.

Carneiro, P., Heckman, J. J., & Masterov, D. (2003). *Labor market discrimination and racial differences in premarket factors*. Cambridge, MA: National Bureau of Economic Research.

Dee, T. & Jacob, B. (2009). *The impact of No Child Left Behind on student achievement*. Cambridge, MA: National Bureau of Economic Research. Working Paper 15531.

Furgol, K. E., Ho, A. D., & Zimmerman, D. L. (2010). Estimating trends from censored assessment data under No Child Left Behind. *Educational and Psychological Measurement*, *70*(5), 760-776.

Grissmer, D. W., Flanagan, A., & Williamson, S. (1998). Why did the black-white score gap narrow in the 1970s and 1980s? In C. Jencks & M. Phillips (Eds.), *The black-white test score gap* (pp. 182-228), Washington, DC: Brookings Institution Press.

Hedges, L. V., & Nowell, A. (1998). Black-white test score convergence since 1965. In C. Jencks & M. Phillips (Eds.), *The black-white test score gap* (pp. 149-181). Washington, DC: Brookings Institution Press.

Hemphill, F. C., & Vanneman, A. (2011). *Achievement gaps: How Hispanic and white students in public schools perform in mathematics and reading on the National Assessment of Educational Progress* (NCES 2011-459). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, *37*(6), 351-360.

Ho, A. D., Lewis, D. M., & Farris, J. L. M. (2009). The dependence of growth-model results on proficiency cut scores. *Educational Measurement: Issues and Practice, 28*, 15-26.

Ho, A. D., & Reardon, S. F. (forthcoming). Estimating achievement gaps from test scores reported in ordinal "proficiency" categories. *Journal of Educational and Behavioral Statistics*.

Holland, P. (2002). Two measures of change in the gaps between the CDFs of test score distributions. *Journal of Educational and Behavioral Statistics, 27*, 3-17.

Kober, N., Chudowsky, N., & Chudowsky, V. (2010). State test score trends through 2008-9, Part 2: Slow and uneven progress in narrowing gaps. Washington, DC: Center on Education Policy.

Neal, D. A. (2005). *Why has black-white skill convergence stopped?* Chicago: University of Chicago Press.

Reardon, S. F. (2011). The widening academic-achievement gap between the rich and the poor: New evidence and possible explanations. In R. M. Murnane & G. Duncan, *Whither Opportunity? Rising Inequality and the Uncertain Life Chances of Low-Income Children*. New York: Russell Sage Foundation.

Reardon, S. F., & Robinson, J. P. (2008). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 497-516). New York: Routledge.

Vanneman, A., Hamilton, L., Baldwin Anderson, J., & Rahman, T. (2009). *Achievement gaps: How black and white students in public schools perform in mathematics and reading on the National Assessment of Educational Progress* (NCES 2009-455). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Wong, M., Cook, T. D., & Steiner, P. M. (under review). No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series.

Appendix B. Tables and Figures

Table 1. Data Availability by Year

Prior to 2000-2010	2000-2010	2002-2010	2003-2010	2004-2010	2005-2010	2006-2010	2007-2010
CO DE LA	FL	IN PA	VA WI	AL GA NV	MT NM WY	CT MI MO	AK MA OH AR MD OK AZ ME OR CA MN RI DC MS SC HI NC SD IA ND TN ID NE TX IL NH UT KY NJ VT KS NY WA WV
*We have data for all 50 states from 2007-2010 at the minimum. Note that years represent spring of the academic year. States from which we may still be able to retrieve additional years of data include: AZ, NJ, OK, and OR.							

Table 2. Trends in Achievement Gaps, 2007-2010

	Fixed Effects Models						*Random Coefficient Models*					
	(1) White-Black		**(2) White-Hispanic**		**(3) Female-Male**		**(4) White-Black**		**(5) White-Hispanic**		**(6) Female-Male**	
Math (vs. Reading)	0.056	***	-0.095	***	-0.208	***	0.069	***	-0.046	***	-0.228	***
	(0.003)		(0.003)		(0.002)		(0.003)		(0.003)		(0.002)	
Kindergarten Cohort-1999	-0.013	***	-0.019	***	0.000		-0.005	*	-0.017	***	0.002	
	(0.001)		(0.001)		(0.001)		(0.002)		(0.002)		(0.001)	
Grade-3	-0.003	+	-0.013	***	0.014	***	0.002		-0.012	***	0.014	***
	(0.001)		(0.002)		(0.001)		(0.002)		(0.002)		(0.001)	
Constant	0.765	***	0.783	***	0.181	***	0.680	***	0.693	***	0.192	***
	(0.008)		(0.009)		(0.005)		(0.028)		(0.029)		(0.010)	
Random Effects Parameters												
SD of Cohort							0.013		0.012		0.005	
							(0.002)		(0.001)		(0.001)	
SD of Constant							0.189		0.197		0.057	
							(0.019)		(0.021)		(0.006)	
LR Chi-2							185***		154***		26***	
Number of States	50		50		50		50		50		50	
Number of Observations	1913		1912		2231		1913		1912		2231	

Notes: +p<.10; *p<.05; **p<.01; ***p<.001. Data include one observation for each state (all 50 states) by grade (grades 3-8) by subject combination (math and reading). Some cells are suppressed due to insufficient numbers of students in a given group. The fixed effects models are weighted by the inverse of the standard error of the gap squared. The likelihood ratio test tests the null hypothesis that the standard deviation of the cohort/grade coefficients is 0.

Figure 1

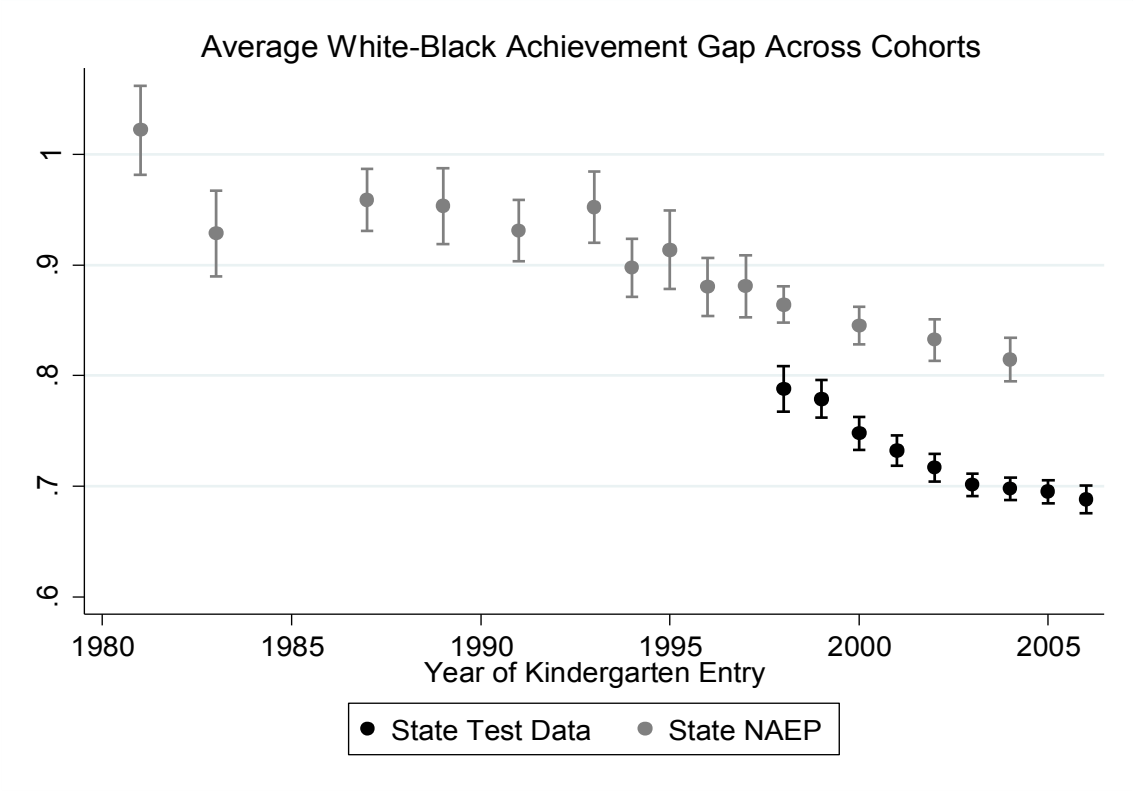


Figure 2

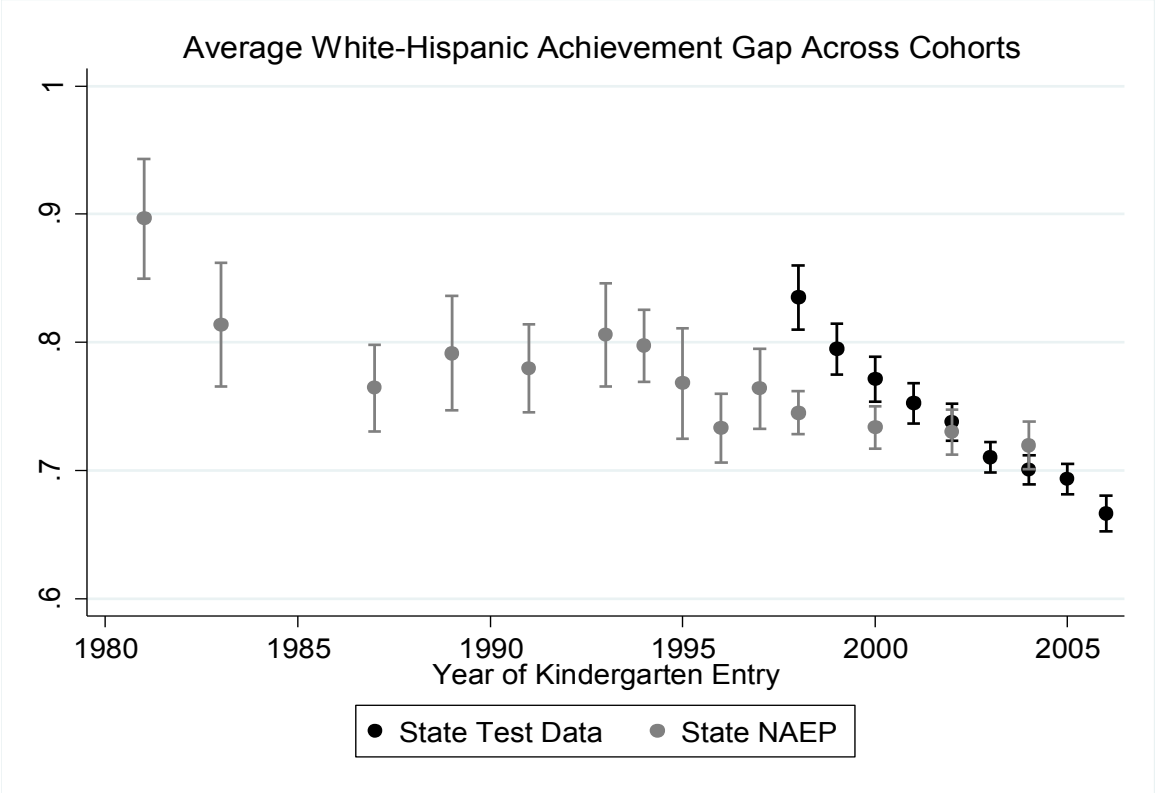
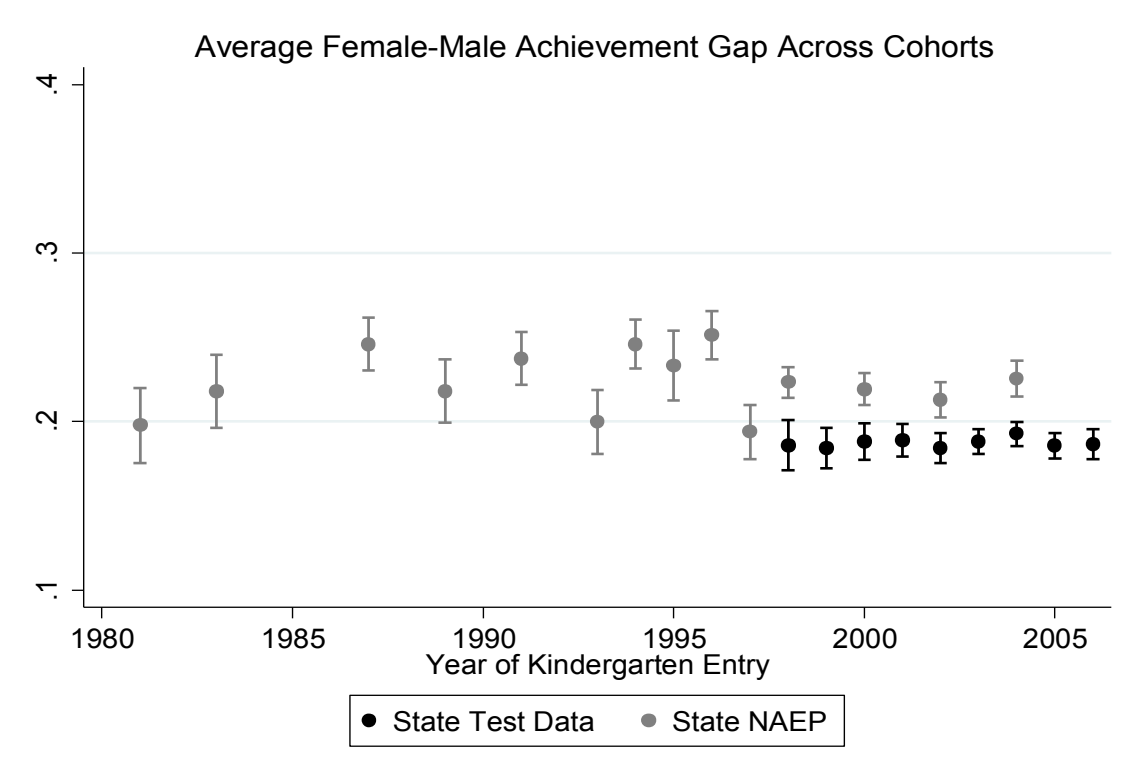


Figure 3


Average Female-Male Achievement Gap Across Cohorts
.1
.2
.3
.4
1980
1985
1990
1995
2000
2005
Year of Kindergarten Entry
State Test Data
State NAEP