Stanford cepa | Center for Education Policy Analysis

CEPA Working Paper No. 21-09

The Economic Achievement Gap in the US, 1960-2020: Reconciling Recent Empirical Findings

AUTHORS

sean f. reardon

Stanford University

ABSTRACT

Has the gap in average standardized test scores between students from high- and low-income families widened, narrowed, or remained stable over the last 3 decades? The question is important both because the achievement gap is measure of how (un)equally educational opportunities are distributed in the US, and because the disparity in educational outcomes is a leading indicator of the degree of economic mobility. If the gap is widening, it suggests that children's educational experiences and opportunities in early and middle childhood – in their homes, neighborhoods, childcare and preschool programs, and K-12 schools – are becoming increasingly unequal, a sign that the growing economic inequality in the US has led to a parallel growth in educational inequality. A narrowing gap, however, would suggest the opposite: changes in early childhood or K-12 schooling have been equity-enhancing, even in the face of increased economic inequality among families. And because test scores and the skills they measure are valued in college admissions and the labor market, the trend in the test score gap may predict the trend in economic mobility several decades later.

VERSION

November 2021

Suggested citation: Reardon, S.F. (2021). The Economic Achievement Gap in the US, 1960-2020: Reconciling Recent Empirical Findings. (CEPA Working Paper No. 21.09). Retrieved from Stanford Center for Education Policy Analysis: https://cepa.stanford.edu/wp21-09

The Economic Achievement Gap in the US, 1960-2020: Reconciling Recent Empirical Findings

sean f. reardon
Stanford University

November 2021

Preliminary/Partial Draft: For Discussion

Direct correspondence to sean f. reardon (sean.reardon@stanford.edu).

The Economic Achievement Gap in the US, 1960-2020: Reconciling Recent Empirical Findings

Has the gap in average standardized test scores between students from high- and low-income families widened, narrowed, or remained stable over the last 3 decades? The question is important both because the achievement gap is measure of how (un)equally educational opportunities are distributed in the US, and because the disparity in educational outcomes is a leading indicator of the degree of economic mobility. If the gap is widening, it suggests that children's educational experiences and opportunities in early and middle childhood – in their homes, neighborhoods, childcare and preschool programs, and K-12 schools – are becoming increasingly unequal, a sign that the growing economic inequality in the US has led to a parallel growth in educational inequality. A narrowing gap, however, would suggest the opposite: changes in early childhood or K-12 schooling have been equity-enhancing, even in the face of increased economic inequality among families. And because test scores and the skills they measure are valued in college admissions and the labor market, the trend in the test score gap may predict the trend in economic mobility several decades later.

A number of recent papers have attempted to measure the trend in the income achievement gap. Sirin (2005), in a meta-analysis of published studies on the association between academic achievement and socioeconomic status, finds that the average correlation between the two *declined* from studies published prior to 1980 to those published in the 1990s. Reardon (2011), using data from a dozen different nationally-representative studies, estimated that the gap *grew* by 30-40% between cohorts born in the 1970s and 1990s. Reardon and Portilla (2016) (hereafter RP2016) used three nationally representative samples of kindergarteners born from 1993-2005 to estimate that the income achievement gap *narrowed* by 10-15% over that period. Chmielewski (2019), using several decades of international assessments, estimates that the gap between students from high- and low-SES families has *changed little*, though the direction of the change depends somewhat on the measure of socioeconomic status. Likewise, Hanushek, Peterson, Tapley, and Woessman (2020) (hereafter HPTW2020) find that gap

between students from high- and low-SES families was largely *unchanged* SES Q4-Q1 gap between cohorts born in the 1950s through 2000. And most recently, and in contrast to the other studies, Hashim, Kane, Kelley-Kemple, Laski, & Staiger (2020) (hereafter HKKLS2020) find that the income achievement gap *declined sharply* between cohorts born around 1980 and those born in 2000.

The differences in the estimated trends among these are very large in some cases: at the extremes, Reardon (2011) estimates the gap grew 30-40%, while HKKLS2020 estimate it narrowed by roughly 40-50%, on average across grades and subjects. In this paper, I attempt to understand and reconcile the divergent estimates in these papers. I reexamine their methods and data and provide some additional new estimates. I focus primarily on three papers: Reardon (2011), HPTW 2020, and HKKLS 2020. I begin by reviewing the primary findings of the papers.

Key Findings of Prior Studies

Reardon (2011)

Reardon (2011) uses data from 13 nationally representative samples of students who were born from 1943 through 2001; in some studies the students were in kindergarten or elementary school when tested; in others they were in middle or high school. Reardon estimates the "90-10 income achievement gap," the gap in average scores between students whose family incomes were at the 90th and 10th percentile of the national contemporaneous family income distribution. Reardon standardizes the gap relative to the national student test score distribution in each study-year and adjusts the estimates to account for measurement error in reported family income and in standardized test scores. Figures 1 and 2 report Reardon's key findings, showing that the 90-10 income achievement gap grew significantly over the six decades. Because of data quality concerns in the early studies (particularly because family income was reported by students rather than parents), Reardon focuses on the trend for cohorts born from 1970 onward. During this period, he concludes, the 90-10 income achievement gap grew by 30-40% in both

math and reading.

Figure 1

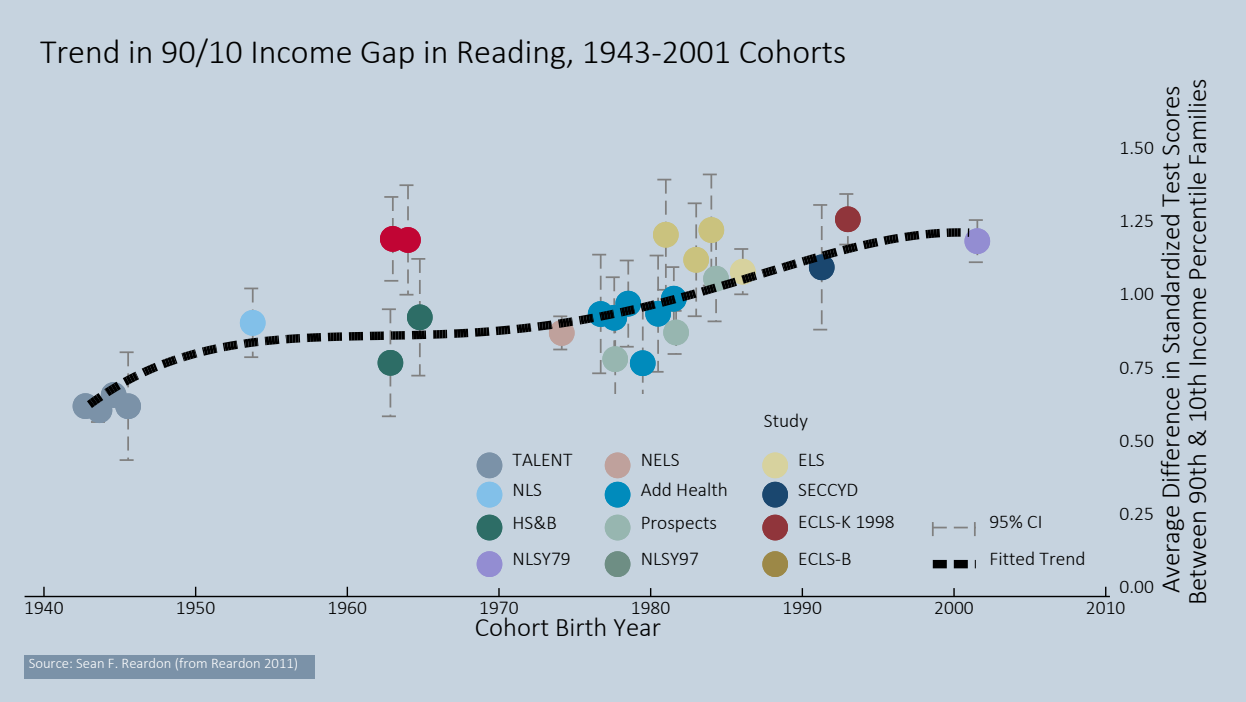
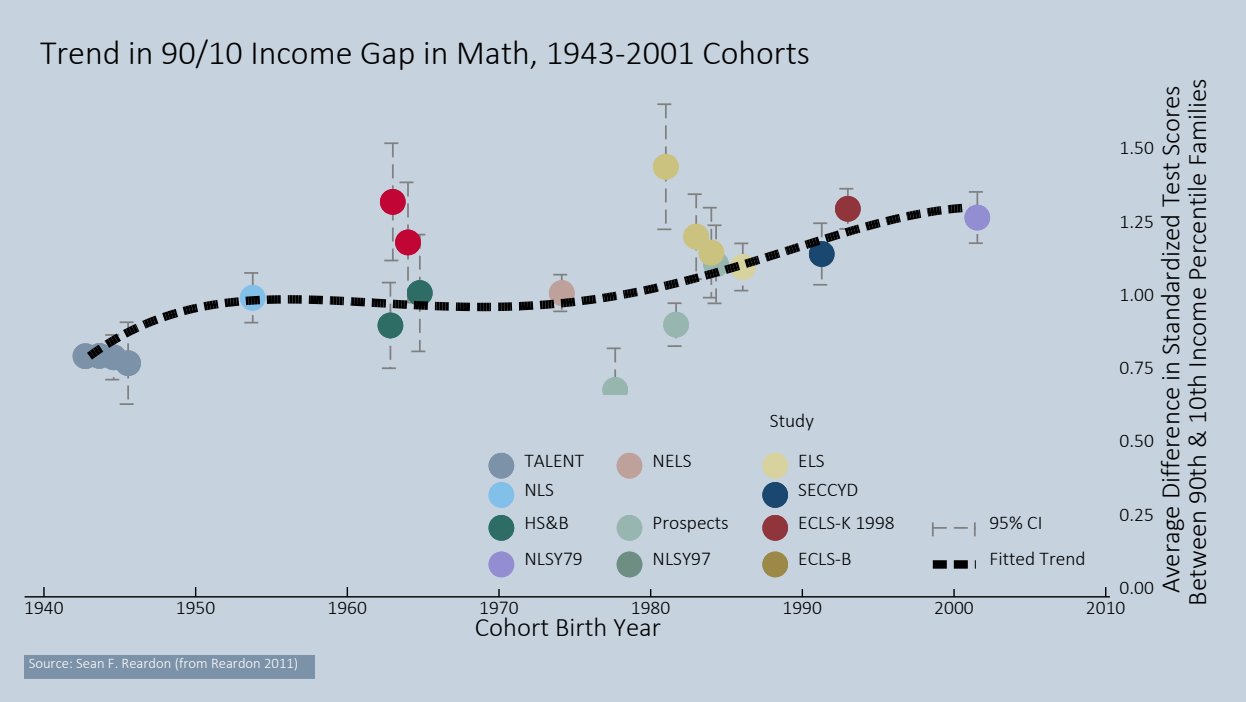


Figure 2



Hanushek, Peterson, Tapley, & Woessman (2020)

HPTW use data from 4 repeated studies: the Long Term Trend National Assessment of Educational Progress (NAEP-LTT); the Main National Assessment of Educational Progress (NAEP); the Trends in International Mathematics and Science Study (TIMSS), and the Programme for International Assessment (PISA). The studies include cohorts born from 1961-2001. Each of these studies has multiple waves, which allows comparison to achievement gaps on similar tests over time. The studies do not include information on family income, however, so HPTW construct indices of family socioeconomic status based on student-reported information about their parents' educational attainment and possessions in their homes. HPTW estimate the average test score difference between students in the top and bottom quartile of these SES measures (what I will call the "Q4-Q1 SES gap"). Unlike Reardon (2001), HPTW do not standardize gaps within each study-year; rather they standardize the gaps relative to the national standard deviation of the tests' scores in the 2000 wave (or the closest year available). They do not adjust the gaps to account for measurement error in the test scores or the SES measures.

Figure 3

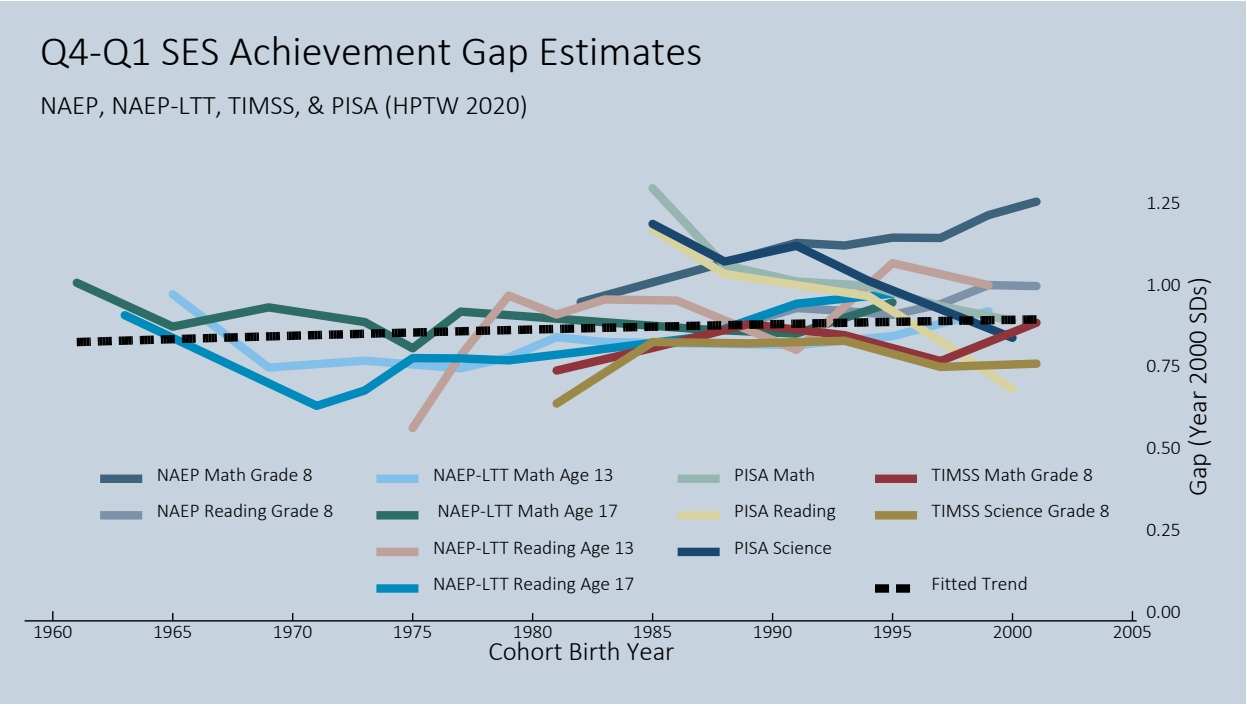


Figure 3 displays the estimated Q4-Q1 gaps from HPTW and their fitted quadratic trend. The linear and quadratic coefficients trend are not jointly significant. HPTW conclude that the SES achievement gap has not changed for cohorts born from the 1960s to 2000.

Hashim, Kane, Kelley-Kemple, Laski, & Staiger (2020)

HKKLS 2020 use data from the Main NAEP assessments administered from 1990 through 2015 to estimate the 90-10 income achievement gap trend in math and reading in 4th and 8th grades. Students in their samples were born between 1976 and 2005. Because NAEP does not include information on family income, they use multiple-step estimation approach that uses information on the income distribution among families living near sampled schools to construct estimates of the achievement gap. Like HPTW, HKKLS do not standardize achievement gaps within test years. Rather, they report gaps in NAEP test score units.

Figure 4

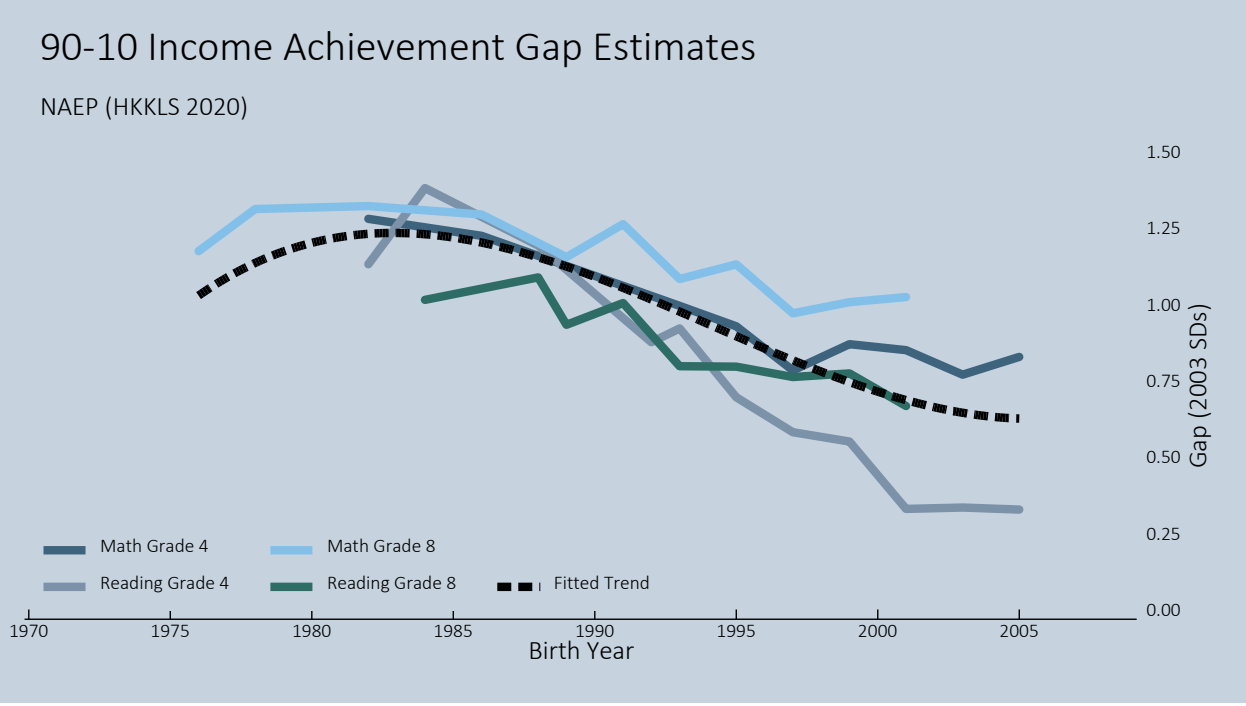


Figure 4 shows the HKKLS estimates of the 90-10 income achievement gap. The gaps show a

substantial decline in the gap from those born in 1980 to 2000.[1] The decline ranges from roughly 20% in 8th grade math to 75% in 4th grade reading.

Summary of Prior Findings

As is evident from Figures 1-4, the three studies come to very different conclusions regarding the trend in the income or SES achievement gap. It is not immediately clear why this is. Table 1 summarizes some of the key features of the 3 papers. As is clear, they differ in many ways: they use different studies, different measures, different standardizations, and different analytic methods. In principle, the estimates based on within-study trends are preferable, because they hold the test instrument (relatively) constant. But none of the studies with multiple waves have measures of family income, which is why HPTW uses an SES index rather than income and why HKKLS use estimated income distribution data from schools' neighborhoods rather than individual family income in their estimator. Those studies that have family income (those used by Reardon) vary considerably in sample age, data collection procedures, and test instruments, leading to concerns (articulated in HPTW) that their gap estimates are not comparable.

Table 1: Comparison of Study Measures and Methods

	Reardon 2011	HPTW 2020	HKKLS 2020
SES Measure			
Dimension	Family Income	SES Index	Household Income
Reliability Adjusted	Yes	No	No
Gap Measure			
Difference	90-10	Q4-Q1	90-10
Standardization	Contemporaneous	2000 SDs	NAEP scale
Reliability Adjusted	Yes	No	Yes[a]
Studies Used			
NAEP-LTT		Yes	
NAEP		Yes	Yes
TIMSS		Yes	
PISA		Yes	
NLSY	Yes		
NCES HS Studies	Yes		
ECLSK Studies	Yes		
Other Studies	Yes		
Birth Cohort Range	1943-2001	1960-2000	1976-2005
Analysis Method	between study, within subject	Pooled within study, using age and subject FE	within study-grade-subject

[a] NAEP-reported SD deviation estimates take into account measurement error, so when I standardize their estimates, they are implicitly adjusted for measurement error in the tests.

[1] I have standardized the estimates in Figure 4 by dividing them by the national standard deviation of NAEP scores in 2003, within grade and subject, for comparability with one another and the gaps reported in other papers.

Below I discuss each study in more detail to determine what accounts for their very conflicting findings. I find that differences in the standardization and reliability adjustments used or not used in the studies accounts for little or none of the different findings. I find that the HKKLS estimator appears to produce highly biased estimates in recent years: two alternate approaches, one using their same data but a different estimator, and one using their estimator but less error-prone data, yield very different results then those they report, suggesting that their estimates are flawed by a combination of their estimation approach and the data they use to measure school income distributions. I find that the HPTW results are biased downward by the influence of several high-leverage early cohort studies and by their overweighting the PISA samples in their analysis. When their gap estimates are analyzed without this overweighting, their data show a clear upward trend in the SES achievement gap over the last three decades. Finally, a reanalysis of the Reardon 2011 data that focuses on trends among a subset of similar studies within the larger pool of 13 studies, suggests that the income gap has grown, though perhaps not by as much as reported in Reardon (2011). I detail the analyses that lead to these conclusions below.

Hashim, Kane, Kelley-Kemple, Laski, and Staiger (2020)

a. The HKKLS estimator

To start, it is useful to briefly summarize the HKKLS estimation approach. HKKLS (2020) want to estimate the coefficient β from a regression of test scores (y) on ln(income) (denoted inc):

$$y = \alpha + \beta(inc) + e.$$

[1]

From this they can estimate the achievement gap between students with incomes at the 90th percentile of the income distribution ($inc90$) and those at the 10th percentile ($inc10$) as

$$\hat{G}^{9010}_{HKKLS} = \hat{\beta}(inc90 - inc10).$$

[2]

But the NAEP data do not include a measure of income for individual students, so they cannot directly estimate β from Equation [1]. Instead they note that

$$\beta = ICC \cdot \beta_b + (1 - ICC)\beta_w$$

[3]

where ICC is the between-school proportion of variance in income, and where β_b and β_w are the coefficient from the following between- and within-school regressions (where j indexes schools), respectively:

$$\bar{y}_j = \alpha_b + \beta_b \overline{inc}_j + \bar{e}_j,$$

[4a]

$$y - \bar{y}_j = \beta_w \left(inc - \overline{inc}_j\right) + e.$$

[4b]

Given this, they turn their attention to estimating ICC, β_b, and β_w. With data on the mean and variance of log income in each school, they can estimate ICC. With data on the average test scores and average log incomes in each school, they can directly estimate β_b from [4a]. But they cannot fit [4b] without student-level income data. Instead, they note that [4b] implies that, in each school j,

$$var\left(y - \bar{y}_j\right)\big|j \; = var\left(\beta_{wj}\left(inc - \overline{inc}_j\right) + e\right)\big|j$$

$$\tau_j = \beta_{wj}^2 \sigma_j + \omega_j,$$

[5]

where β_{wj} is the within-school slope in school j, and τ_j, σ_j, and ω_j are the within-school variances of test scores, log income, and residual error e, respectively. Using estimates of τ_j and σ_j, they fit the regression model:

$$\tau_j = \eta + \gamma \sigma_j + v_j,$$

[6]

and estimate the within-school slope as

$$\hat{\beta}_w = \sqrt{\hat{\gamma}}.$$

[7]

Given these estimate, HKKLS estimate the 90-10 income achievement gap as

$$\hat{G}^{9010}_{HKKLS} = \left[\widehat{ICC} \cdot \hat{\beta}_b + \left(1 - \widehat{ICC}\right)\sqrt{\hat{\gamma}}\right](inc90 - inc10)$$

[8]

Note that because the ICC is small (roughly 0.2) and changes little over time, the gap $\hat{G}^{9010}_{HKKLS}$ and its trend depend much more on $\hat{\gamma}$ than on the other terms in Equation [8].

The HKKLS estimator is a clever approach to fitting an individual-level association when only aggregate information on the regressor (income) is available. However, the estimator is subject to a number of potential sources of bias. Key among these are: 1) measurement-error induced bias in the estimates of $\hat{\beta}_b$ and $\hat{\beta}_w$, resulting from the fact that $\overline{inc}_j$ and σ_j are measured with error; 2) variance in the within-school slope β_{wj} across schools; 3) potential bias in $\hat{\gamma}$ as an estimator of β_w due to covariance between the within-school slope β_{wj} and the within-school income variance σ_j; and 4) potential covariances among the error components of the estimates used in the righthand side of Equation [8]. A more detailed derivation of potential bias in the HKKLS estimator is shown in the Appendix.

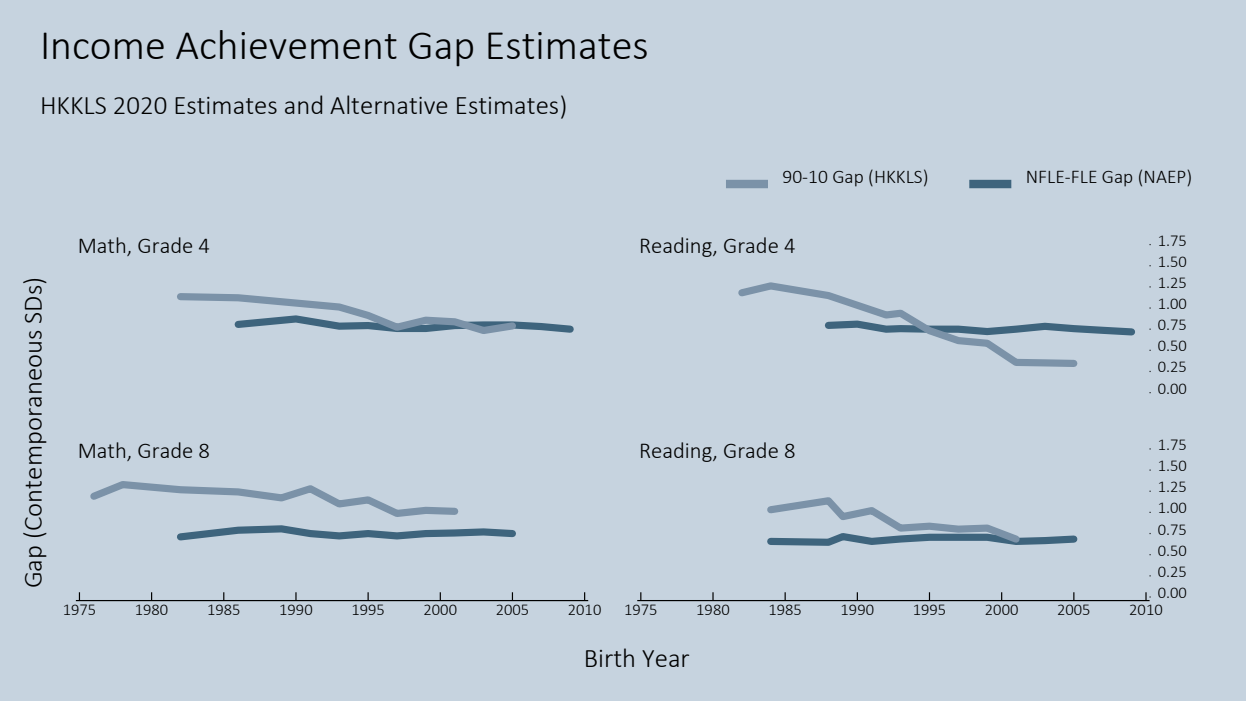
Rather than attempt to assess the bias in the HKKLS estimates directly, I first consider an alternate estimator of the 90-10 income gap using NAEP data; I also implement the HKKLS estimator using a data set with far less measurement error in the aggregate income distribution estimates.

b. An alternate 90-10 NAEP income gap estimator

One way of assessing the validity of the HKKLS estimates is to assess whether they comport with other information available in the NAEP data. In particular, while the NAEP data do not include detailed information about individual students' family income, they do include information about students' eligibility for the National School Lunch Program. As a result, we can easily use NAEP data to compute the

trend in the test score gap between students eligible and not-eligible for free school lunches. These trends are shown in Figure 5, along with the HKKLS estimates of the 90-10 income test score gaps. Note that I have standardized the HKKLS and NFLE-FLE gaps relative to the national standard deviation of test scores among public school students in the corresponding year-subject-grade.

Figure 5



Note: Both the HKKLS trend estimates and NFLE-FLE trend estimates are based on the same NAEP nationally representative samples of public school students. Gaps are standardized relative to the corresponding year-grade-subject national standard deviation of NAEP test scores among public school students. Note that NAEP data do not include FLE status prior to 1996, so NFLE-FLE gap estimates are not available for the earliest cohorts used in the HKKLS analysis. Likewise, HKKLS did not use data from the 2017 and 2019 NAEP studies, so the HKKLS estimates are missing for the most recent cohorts.

The gap between the non-free-lunch-eligible and free-lunch-eligible (NFLE-FLE) students has been very stable over the last two decades. A linear trend fit through the gap trend is in each grade-subject not significantly different than 0. A pooled model using all subjects and grades together (with subject-grade fixed effects) also shows no significant trend (beta = -0.0008; se = 0.0006; p = 0.20). Note that the estimates of the FLE gap from NAEP may be biased downward because of measurement error in free lunch eligibility status as a measure of income. This measurement error is likely increasing in recent years

because of the Community Eligibility Provision, which between 2012 and 2015 (depending on the state) and which changed the way that school districts certified and reported free lunch eligibility status.

The absence of a trend in the NFLE-FLE gaps is in striking contrast to the HKKLS estimated 90-10 trends, which show a clear average downward trend both individually and when pooled (beta = -0.0237; se = 0.0028; $p < 0.001$). Over 20 years, the estimated gaps declined by 0.47 SDs on average, a decline in the gaps of roughly 40-50%.

How should we make sense of these apparently conflicting findings? They are based on the same samples of students, and both measure a gap between students of different incomes, though they differ in their specification of the income gap. Are these estimates compatible? Under two simple assumptions, it is possible to derive a simple relationship between the 90-10 gap and the NFLE-FLE gap.

First, note that the HKKLS estimator is based on the assumption that the association between log income and achievement is linear:

$$y = \alpha + \beta(inc) + e.$$

[1]

Let us additionally assume that log income is normally distributed with standard deviation σ in the population. This is approximately true in the US. Now suppose instead of estimating β, we want to estimate the gap in average scores between students above and below some income threshold, call it inc^P, the value of log income corresponding to the $100 \cdot P^{th}$ percentile of the income distribution. It is straightforward to show that under these two assumptions, the gap between those with incomes above and below inc^P is

$$G^{NP-P} = \frac{1}{P(1-P)}\left[\beta\sigma\left(\phi\left(\Phi^{-1}(P)\right)\right)\right].$$

[9]

Under the same assumptions, the 90-10 income achievement gap is

$$G^{9010} = \beta(inc90 - inc10)$$

$$= \beta\sigma[\Phi^{-1}(.90) - \Phi^{-1}(.10)]$$

$$= 2.563\beta\sigma.$$

[10]

The relationship between G^{9010} and G^{NP-P} is therefore

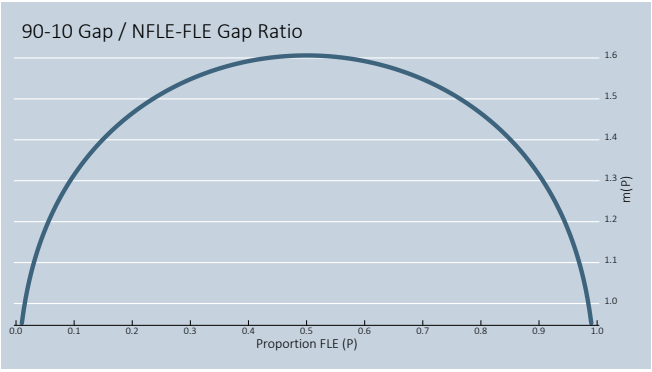
$$G^{9010} = \frac{2.563 \cdot P(1-P)}{\left(\phi\left(\Phi^{-1}(P)\right)\right)} G^{NP-P}$$

$$= m(P) \cdot G^{NP-P}.$$

[11]

Equation [11] shows that the 90-10 gap is a multiple of the NP-P gap, where the multiplier $m(P)$ depends on the proportion P. Figure 6 displays the multiplier function $m(P)$ as a function of P. Note that the multiplier is larger than 1 unless $P <\approx .01$ or $P >\approx .99$; for values of $P \in (.15, .85)$, the multiplier is roughly between 1.4 and 1.6. In other words, as long as $P \in (.15, .85)$, the 9010 gap is roughly 1.4-1.6 times the NP-P gap.

Figure 6

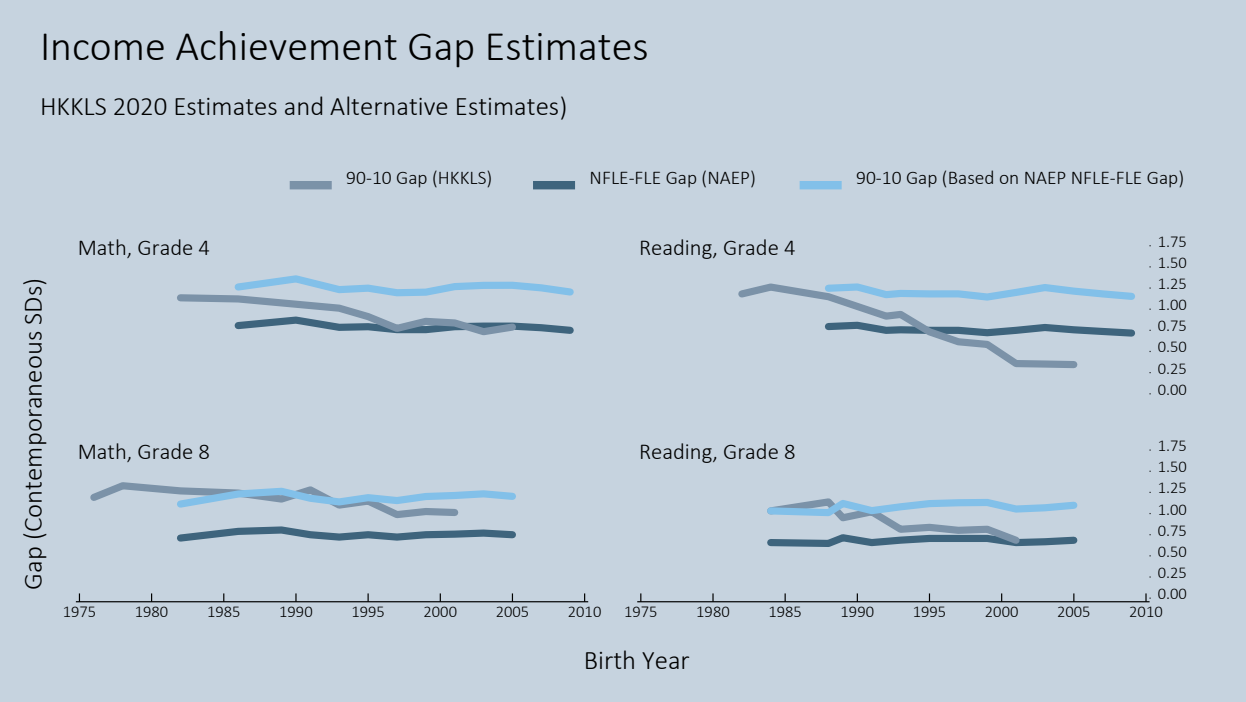


In the years 1996-2019 (the years when NAEP data include FLE status), the proportion of students in public schools grew from 33% in 1996 to 47% in 2019; therefore, we expect the 90-10 income achievement gap to be about 1.55-1.60 times the FLE gap during those years.[2] The estimated 90-10

[2] The percent FLE estimates come from the CCD. The proportion of the NAEP sample classified as FLE in grades 4 math and reading grew from 37 to 54% and 41 to 54%, respectively, over this period; the proportions in the 8th

income gaps implied by the NFLE-FLE gaps are shown in Figure 7.

Figure 7



The 90-10 gap estimates based on the NAEP data and the HKKLS estimator are often dramatically smaller, and their trends are very different, than those based on the NAEP NFLE-FLE gap estimates. The HKKLS estimates are in some cases smaller even than the NFLE-FLE gap estimates. While it is mathematically possible for the 90-10 gap to be smaller than the NFLE-FLE gap, it implies a very nonlinear relationship between income and test scores. In other words, the HKKLS estimates (and their trend) are inconsistent with the NAEP NFLE-FLE estimates. Importantly, both sets of estimates are based on the same (NAEP) test scores from the same samples of students; both are also based on an assumption that the association between test scores and log income is linear. The differences therefore lie in features of the two estimators. The estimates based on the NAEP NFLE-FLE gaps rely on an additional assumption

grade math and reading samples grew from 27 to 47% and 30 to 50%, respectively. I use the CCD proportions in my calculations because they are not sample-based, and because up to 15% of students in the NAEP samples are missing FLE status in the early years of the NAEP data. Because the multiplier function $m(P)$ is very flat in the range of .3 to .7, the results are very insensitive to error in the proportion FLE, so my estimates are insensitive to which source of FLE data I use.

that the income distribution is log normal (though the estimates are unlikely to change much given modest violations of that assumption). The HKKLS estimates are likely subject to more sources of error and bias, particularly because they rely on potentially very noisy measures of the income distribution in each school.

c. Flaws in the HKKLS estimator

The HKKLS estimator is clever, but not without complexity. Under the assumption that $\beta_{wj} = \beta_w$ for all schools and that σ_j is measured without error, this will yield unbiased estimates of β_b. But neither of those assumptions is likely valid.

The estimation of all three terms in Equation [3] (ICC, β_b, and β_w) depends on estimates of the school-specific means and variances of log income. HKKLS estimate these from Census and ACS data for the set of census tracts nearest each school. These estimates are likely to have considerable error for several reasons: students enrolled in a given school do not necessarily live in the nearest census tracts, and vice versa; the household income distribution of public school students' families may differ from the overall household income distribution in a tract; and tract level household income distributions are based on 6-16% samples, and so contain considerable sampling error. This error may lead to bias in the estimates. The bias will be larger, in general, the larger the error variance in the estimates of neighborhood income distributions. Moreover, if β_{wj} varies among schools or covaries with σ_j, $\hat{\gamma}$ will be biased.[3] The potential for bias raises some concerns, but it is difficult to assess the degree of bias.

d. Using the HKKLS estimator with better data

The HKKLS estimates are based on school-level data. HKKLS estimate the mean and variance of test scores within each school from NAEP data; these are typically based on small samples (within-school NAEP samples are roughly 20-30 students on average). HKKLS estimate the mean and variance of the

[3] See [to be added] Appendix for detail.

family income distribution for each school using Census or ACS data; they compute a weighted average of the income distributions in tracts near each school. Their estimates of the income distribution within each school are subject to considerable error for three reasons: not all students in a school live in a nearby census tract; not all households in a tract have children enrolled in the geographically linked public school; and Census and ACS income distribution estimates are based on samples within each tract. The samples in the ACS were considerably smaller than those in the Census (roughly half as large), so measurement error in the income estimates is larger after 2000 than before.

If measurement error in the estimated means and standard deviations of schools' log income distribution leads to bias in the HKKLS estimates, we might correct this using better data. In the analysis below, I implement the HKKLS estimator using school district data rather than school data. To do so, I use estimates of the mean and variance of test scores in each district from SEDA and estimates of the mean and variance of household income in each district from the EDGE (Education Demographic Geographic Estimates) data. The SEDA and EDGE estimates are much less error-prone than the data HKKLS use.

The Census and ACS do not provide tabulations of family income by school attendance zone (which is why HKKLS rely on a weighted average of the nearest tract income distributions), but NCES's EDGE program does provide tabulations of Census and ACS data by school district. The EDGE estimates of income distributions among families living within each school district is less error-prone than the tract-based estimates of school income distributions used by HKKLS for two reasons: first, districts are larger than schools, so the EDGE district estimates are based on larger samples of families than the school estimates; second, the population of families living within a school district boundary matches the population of children enrolled within the district much more closely than the population living in nearby tracts matches the population enrolled in a school. In the vast majority of cases, almost all students enrolled in a school district live within the district's geographic boundaries; likewise, very few students

attend public school outside their geographic school district.[4] As a result, the EDGE

The reason that HKKLS do not use the EDGE school district income distribution estimates is that NAEP samples students within schools, not districts, so while HKKLS can estimate the mean and variance of test scores within a nationally-representative sample of schools, they cannot do so for districts. The SEDA data, however, do provide estimates of the mean and standard deviation of test scores for every school district in the US, albeit over a shorter time frame than NAEP. These estimates are based on the test scores of the full population of students, rather than a sample, so have much greater precision than NAEP's sample-based estimates.

I construct a special version of the SEDA data that includes estimates of the mean and variance of test scores in each school district in the US from 2009-2018. I treat charter schools as part of the district in which they are geographically located, so that the distribution of test scores in each district includes all public school located in the districts. These estimates are available for math and reading in each grade 3-8. The estimates are standardized within each grade-year-subject to the estimated national public school student test score distribution. In grades 4 and 8 in odd years, the national distributions are available from NAEP; in the other grades and years, I interpolate the means and standard deviations from the NAEP data.

The SEDA estimates are not available for all grades, years, and subjects in each district. There are three primary reasons why some estimates are missing. First, in some state-year-grade-subjects, not all students take the same standardized test at the end of the year. For example, prior to 2014, 8th grade students in CA took standardized math tests based on their eighth grade math course rather than a common test (some took an Algebra test, some took a general math test, etc). Because not all students

[4] The match is not exact, because some students living in a geographic district attend private school or are homeschooled. In some years, however, the EDGE provides tabulations of family income among families who have school-age children enrolled in public school. In those years, the correlation between the median log income of all households and of families with children in public school is 0.93.

too the same test (and because selection into which test they took was related to their math skills), it is not possible to construct a measure of average performance in those year-grade-subjects. This is an issue only in 7th and 8th grade in several states (albeit some large states, including CA and TX) in some years. Second, in some state-year-grade-subjects or district-year-grade-subjects, fewer than 95% of students took the required test. This was most common in a few states, such as NY and CO, in 2015 and later, when the test "opt-out" movement was strong. In cases where fewer than 95% of students in a state took the test, I use no data for districts in that state-grade-year-subject, because an accurate estimate of the statewide distribution is needed to construct the district-level estimates. In addition, in cases where more than 95% of the students in the state took the test, but fewer than 95% in the district took the test, I exclude data for the district-grade-year-subject. Third, estimates are available only when there are at least 20 students tested in the relevant district-grade-year-subject.

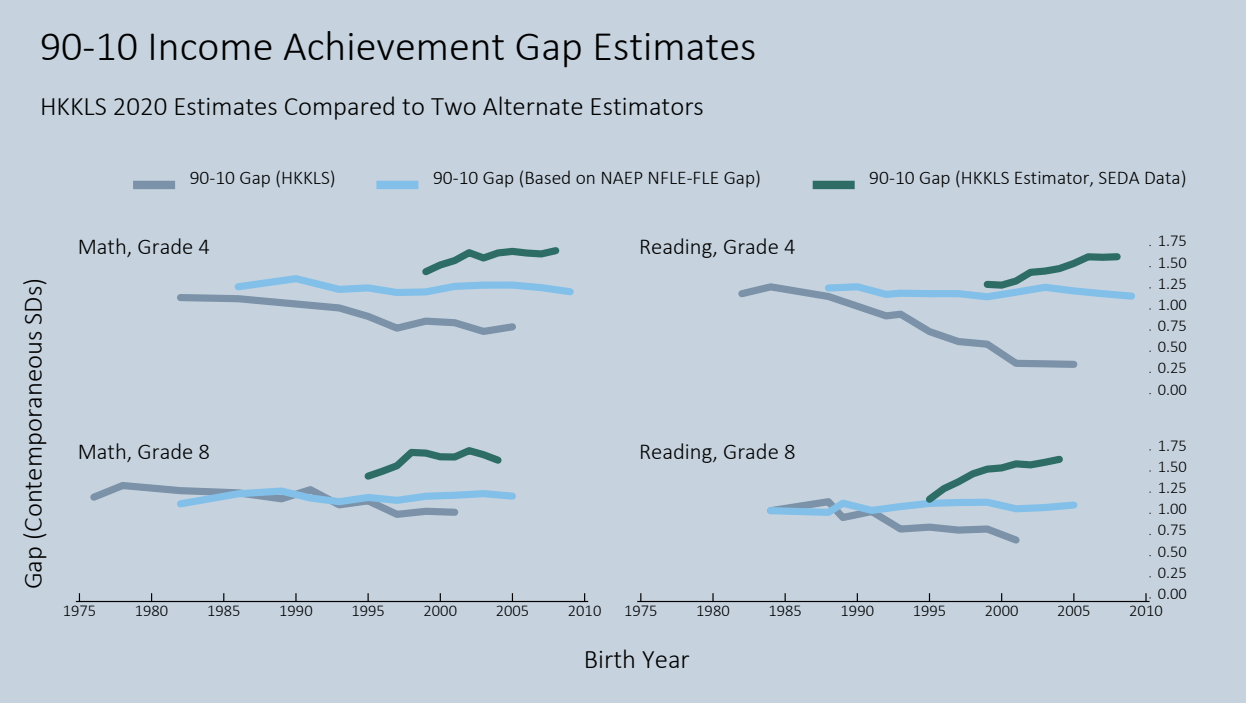
The SEDA data include estimates in at least one grade-year-subject for 11,797 school districts in the US. In each district there are up to 120 grade-year-subject estimates (10 years, 6 grades, 2 subject). The data I use include a total of 1.13 million district-grade-year-subject estimates, an average of 96 estimates per school district (80% of the possible 120 observations per district). In each subject and grade, roughly 5,000 school districts have test score data in all 10 years; roughly 9,000 districts have data in at least 8 of 10 years (See Table 2). The exception is in 7th and 8th grade math, where considerably fewer districts have all 10 or at least 8 years of data (in 8th grade math only 2,724 districts have all 10 years of data and only 6,411 districts have at least 8 years of data). In the analyses below, I fit the models using three samples: 1) using all 1.13 million district-year-grade-subject observations (11,797 districts); 2) using only district-grade-subject observations where a district has data for at least 8 of the 10 years; 3) using only district-grade-subject observations where a district has data for all 10 years. The last sample is completely balanced within grade-subject, so trends estimated from this sample are not biased by the unbalanced nature of the data (where some districts contribute to estimates of the 90-10 gap in some

years but not others), but the completely balanced panel is less representative of the US, because districts missing data are disproportionately concentrated in some states. For example, 20 states have no districts with data in all 10 years in any grade-subject; 5 states have no districts with at least 8 years of data in any grade-subject.

Table 2 here

Figure 8 below shows the estimated 90-10 income achievement gap estimates based on the HKKLS estimator applied to the SEDA and EDGE data. The SEDA-based estimates shown here are those from the balanced panel, though the trends are very similar when using all the data or when limiting the sample to districts with at least 8 years of data (see Appendix Figure A1). The ICC between districts is smaller than between schools (it declines modestly from roughly 0.12 to 0.11 from 2009 to 2018), so the SEDA-based estimates are based largely on the within-school slope estimates derived from the regression of test score variance on log income variance (Equation [6]). There is only a 6-year period when all three estimates are available (corresponding to the years 2009-2015). In those years, the gaps estimated from the SEDA and EDGE data are larger than those estimated by HKKLS from the NAEP and census tract data, and are also generally larger than those implied by the NAEP NFLE-FLE gap.

Figure 8



Moreover, the three estimators show divergent trends. If we limit analysis to grades 4 and 8 from 2009-2015, the pooled average annual change in the gap is negative (-0.023 SDs/year; se = 0.006; p<.01) for the HKKLS estimates; flat (0.006 SDs/year; se = 0.004; p = 0.18, ns) for the estimates based on the NAEP NFLE-FLE gap; and positive (0.048 SDs/year; se = 0.005; p<.001) for the estimates based on the SEDA data. The difference in the two trends estimated using the HKKLS estimator is quite large (a difference in the trend of 0.07 SDs/year), suggesting substantial bias in one or both estimates. Given that the data used to estimate β_w from Equation [6] is arguably much less error-prone in the SEDA and EDGE data than the NAEP and census tract data, there is little reason to trust the HKKLS estimates. That said, the SEDA-based estimates here are not unambiguously unbiased either.

e. Summary of HKKLS reanalysis

In sum, the HKKLS estimator, while clever, relies on several questionable assumptions and error-prone data. The trend estimates based on the estimator are in some cases wildly inconsistent with more reliable and valid estimates of the NFLE-FLE gap. While it is mathematically possible to have 90-10 gaps

smaller than FLE gaps, it is highly implausible. This suggests the HKKLS estimator, when used with the error-prone Census-tract and ACS data, is substantially biased. It is less clear why its bias would change over time, but one possibility is that the smaller samples in the ACS, relative to the Census, increase bias that is due to measurement error in the mean and variance of income associated with a school.

When I use the HKKLS estimator with much less error-prone district-level data, and with much larger samples of districts and students, the estimated 90-10 gap trend is quite different than that reported by HKKLS; the estimated gap is much larger and grows over time. Even so, the 90-10 gaps based on the SEDA data are larger than those reported in other studies, raising the question of whether they also are biased, albeit in a different direction. More analysis is needed before they should be interpreted as accurate.

Hanushek, Peterson, Tapley, and Woessman (2020)

HPTW report that the trends in the Q4-Q1 SES gap has been essentially flat for birth cohorts from 1960 to 2000. Their estimates and models differ in several ways from those in Reardon (2011). I reanalyze their data here to assess the sensitivity of their findings to their estimation and modeling choices.

First, HPTW standardize their gap estimates relative to the standard deviation of the test in the year 2000, or the nearest available year. Standardizing by a constant allows them to estimate trends in the absolute gap, rather than relative to a potentially changing population standard deviation. But it makes the interpretation of the gap different than in Reardon 2011. For comparability with Reardon 2011, I standardize the HPTW gaps relative to their contemporaneous standard deviation. I also adjust them for measurement error, diving the gaps by the square root of test reliability. I compare estimates of the trend using these contemporaneously-standardized, reliability-adjusted gaps to HPTW's gap estimates.

Second, Reardon (2011) found that the 90-10 income achievement gap grew for cohorts born after 1970 (cohorts born prior to that had lower-quality income data, so the trend was less certain).

HPTW's data include 6 NAEP-LTT waves for cohorts born in the 1960s. The NAEP-LTT gaps for these cohorts are modestly larger than for cohorts born in the 1970s. In some specifications, I restrict the sample to cohorts born after 1970 to compare the HPTW trend estimates to Reardon's estimates.

Third, HPTW pool estimates from reading, math, and science in a single model, using subject, study, and age fixed effects. One problem with this model is that, in the PISA and TIMSS studies, the same sample of students takes the tests in different subject, so the gap estimates in multiple subjects share common sampling error. In the NAEP and NAEP-LTT studies, a different sample of students takes the math and reading test in each year, so the gap observations are independent within year and age. In TIMSS, the estimated math and science gaps are nearly identical at each wave, a result of the fact that the gaps are based on the same sample of students. In PISA, the use of multiple subjects in the same model is more problematic. In each year, PISA denotes one subject the focal subject and administers students a full-length test in that subject. Reading was the focal subject in 2000 and 2009; math was the focal subject in 2003 and 2012; and science was the focal subject in 2006 and 2015. Students answer a small number of questions in the other two subjects. The small number of items in the non-focal subjects is not sufficient to provide reliable student-level test scores, so the scores in the non-focal subjects are estimated using a Bayesian model; they are shrunken toward the score in the focal subject. As a result, scores in the non-focal subject are based in large part on students' scores in the focal subject—not only is do they share sampling variance, but they share measurement error variance as well. Because there is little independent information in the three tests, including all three in the model gives the PISA results three times the weight they should have. I assess the impact this has on the trend estimates by including only the focal PISA test gap estimates in the model; in these model I also include only a single observation for each TIMSS wave, where the gap is the average of the math and science gaps in that wave. These models therefore include only a single observation per sample, rather than a single observation per gap estimate.

One other feature of the PISA data is worth noting. The math PISA assessment was substantially revised in 2003 and the science assessment was substantially revised in 2006. As a result, the test scales for these subjects are not stable. This is not a problem if the gaps are standardized relative to their contemporaneous standard deviation, but it does mean that the absolute gaps in PISA math 2000 and PISA science 2000 and 2003 are not comparable to later PISA gaps. HPTW include these gaps in their estimates, which may cause bias in the estimated trends. In some specifications, I exclude these three PISA observations to assess their impact on the trend estimates.

One other key difference between the Reardon (2011) estimates and the HPTW estimates is that HPTW construct an index of socioeconomic status based on student-reported information about parental education and home possessions. Reardon uses parent-reported income (or student-reported family income for several studies of birth cohorts born prior to 1970). It may be that the HPTW and Reardon trends differ because the association between income and test scores has changed differently over time than the association between SES and achievement. Reardon 2011 included some supplemental analyses that suggested the achievement gap with respect to parental education had not changed much over the last few decades, unlike the income achievement gap. The trends may also differ if changes in the measurement properties of the SES index differ from changes in the measurement of income over time. It would be useful to have more information about the extent to which the measurement properties of the HPTW SES index are stable over time, but I do not investigate that here. Rather, I take the HPTW Q4-Q1 gap estimates as reported in their Appendix A2 and reanalyze the trends in those. In other words, I focus on choices regarding the sample of test score gap estimates to use and the scaling of the test score gaps, but not choices regarding the construction of the SES measure or the estimation of gaps along that dimension.

Table 3 reports the fitted estimates from models of the form

$$G_{saty}^{Q4-Q1} = \alpha + \beta\left(birthyear_{ay} - 1970\right) + \Gamma_{sat} + e_{saty},$$

where s indexes studies, a indexes age, t indexes subjects, and y indexes test years. The models include study-age-subject fixed effects (Γ_{sta}). The full sample include the 81 gap estimates included in HPTW Appendix A2. In some specifications, I limit the sample to cohorts born after 1970 (models 6-10); in some I exclude the 3 PISA observations with noncomparable test scales (model 2 and 7); in some I include only one observation per sample of students, rather than multiple observations (of different test subjects) of the same sample (models 4, 5, 9, and 10); and in some models I use the contemporaneously-standardized and reliability-adjusted gap estimates (models 3, 5, 8, and 10). The model closest to that reported in HPTW is model 1.[5] The preferred model for comparison to Reardon (2011) is model 10: this model uses the adjusted gap estimates, includes only one observation per sample of students, and includes only post-1970 observations.

Table 3 here

In model 10, the estimated annual change in the gap is 0.0064 SDs/year (se = 0.0015; p<0.001), implying that the gap increased by 0.19 SDs from the 1970 to 2000 birth cohorts, an increase of 23% from its 1970 value of 0.83. The choice to use contemporaneously-standardized and reliability-adjusted measures is not responsible for this; in fact the trend in the unadjusted gaps is a little steeper (0.007 SDs/year).

The difference between this trend and the flat trend estimated by HPTW appears largely due to two factors: the difference in samples, and the exclusion of cohorts born prior to 1970. Roughly half the difference is due to the first factor. The PISA trend differs from the other three studies—unlike the other studies, the trend declines across PISA cohorts. Because the HPTW analyses overweight the PISA sample (by including it three times in the model), the models that include only one observation per sample place

[5] HPTW use a quadratic rather than a linear specification, though their fitted trend shows very little curvature (see Figure 3 above). In addition, I use study-subject-age fixed effects, while they use study, subject, and age fixed effects; the differences between their specification and model 1 do not appreciably affect the conclusion that there is no significant trend in the Q4-Q1 SES achievement gap.

less weight on the PISA trend, yielding a positive average trend. Comparing models 8 and 10, the estimated trend is nearly twice as large in model 10, which includes only one observation per sample, as in model 8, which includes all observations.

Roughly half of the difference is due to the restriction of the models to cohorts born after 1970. The NAEP-LTT study is the only study with samples of students born before 1970; these observations have considerable leverage in the linear trend. Because the gap estimates for those cohorts are modestly larger than for cohorts born in the 1970s, the linear trend is much flatter when the pre-1970 cohorts are included.

In sum, I conclude that the HPTW estimated trend is biased because the inclusion of multiple gap estimates per sample effectively overweights the PISA and TIMSS samples. The PISA trend, in particular, is quite different than the trends in the other 3 studies. Table 3 reports estimates of the post-1970 trend from each study separately. The contemporaneously-standardized, reliability-adjusted gap grew by 0.006 SDs/year in NAEP-LTT ($p<.001$), by 0.014 SDs/year in NAEP, by 0.010 SDs/year in TIMSS (ns), and declined by 0.015 SDs/year in PISA ($p<.10$).

Discussion

We have three studies with three very different conclusions about a very important question. What are we to make of this? My analyses here suggest that both the HPTW and HKKLS papers are flawed. When the analyses in HPTW are focused on the post-1970 cohorts and weighted appropriately, there is a clear upward trend in the SES achievement gap. The gap appears to have grown by roughly 25% over a 30-year period, from cohorts born in 1970 to 2000. The data from NAEP-LTT, Main NAEP, and TIMSS are each individually consistent with this. The PISA data, however, tell a very different story, and it is unclear how to reconcile the very different PISA trend from that in the 3 other studies. On the whole, however, the weight of the evidence in the HPTW estimates suggests that socioeconomic gaps are growing.

The picture is less clear from the HKKLS estimates. The magnitude of gaps they report for recent years is inconsistent with all other sources of data, including the NAEP data they use. Moreover, their same estimator applied to less error-prone data yields not only much larger gaps but a trend in the opposite direction. Thus, the HPTW analyses do not seem to provide informative evidence regarding the size and trends in the income achievement gap, despite the ingenuity of their estimation approach.

This is a very preliminary draft. In a future iteration of this paper, I will provide more detailed reanalysis of Reardon (2011) and a discussion of the Chmielewski paper as well. I appreciate all comments and suggestions.

Table 2

Table 1: Distribution of Districts by Number of Years of SEDA Test Score Data, by Grade and Subject

N Years With Math	Grade					
Test Score Estimates	3	4	5	6	7	8
0	336	347	343	347	606	635
1	234	232	342	360	244	287
2	157	160	170	164	224	260
3	194	153	178	258	364	375
4	173	210	268	199	735	726
5	414	398	430	770	1,516	1,872
6	793	839	743	427	441	496
7	371	342	386	416	496	735
8	895	908	885	886	953	1,485
9	3,205	3,156	3,028	3,150	1,642	2,202
10	5,025	5,052	5,024	4,820	4,576	2,724
Total	11,797	11,797	11,797	11,797	11,797	11,797

N Years With Reading	Grade					
Test Score Estimates	3	4	5	6	7	8
0	335	343	341	342	418	422
1	238	239	220	233	224	230
2	160	157	164	156	181	165
3	195	154	184	268	268	290
4	167	203	251	190	192	188
5	390	388	381	767	759	790
6	861	915	861	468	474	516
7	456	400	413	416	487	708
8	650	682	668	777	1,090	1,076
9	3,170	3,137	3,193	3,148	2,908	3,268
10	5,175	5,179	5,121	5,032	4,796	4,144
Total	11,797	11,797	11,797	11,797	11,797	11,797

Table 3: Estimated Linear Trends in the Q4-Q1 SES Test Score Gap, Various Model Specifications

	Models including all studies & all years					Models including all studies and post-1970 cohorts				
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Gap in 1970	0.8992 ***	0.8721 ***	0.9205 ***	0.8654 ***	0.8736 ***	0.8559 ***	0.8154 ***	0.9016 ***	0.7923 ***	0.8324 ***
	(0.0286)	(0.0272)	(0.0267)	(0.0256)	(0.0221)	(0.0359)	(0.0333)	(0.0345)	(0.0300)	(0.0281)
Annual Change	0.0013	0.0022	0.0026 +	0.0033 *	0.0041 **	0.0035 +	0.0050 **	0.0038 **	0.0070 ***	0.0064 ***
	(0.0015)	(0.0014)	(0.0014)	(0.0014)	(0.0012)	(0.0018)	(0.0017)	(0.0017)	(0.0016)	(0.0015)
Implied % Change in Gap 1970-2	4% ns	8% ns	8% +	**11%** *	**14%** **	12% +	**18%** **	**13%** **	**27%** ***	**23%** ***
Sample and model specifications										
Birth Cohorts Included	all		all	all	all	>1970	>1970	>1970	>1970	>1970
All Observations	x		x			x		x		
Excluding PISA 2000 M&S, 2003 S		x					x			
Non-redundant Observations				x	x				x	x
Gap Specification										
2000 SD	x	x		x		x	x		x	
Contemporaneous SD			x		x			x		x
Reliabilty Adjustment			x		x			x		x
Fixed Effects Included										
Study-Subject-Age FEs	x	x	x	x	x	x	x	x	x	x
N (gap estimates)	81	78	81	64	64	75	72	75	58	58
N (study-subject-ages)	11	11	11	11	11	11	11	11	11	11

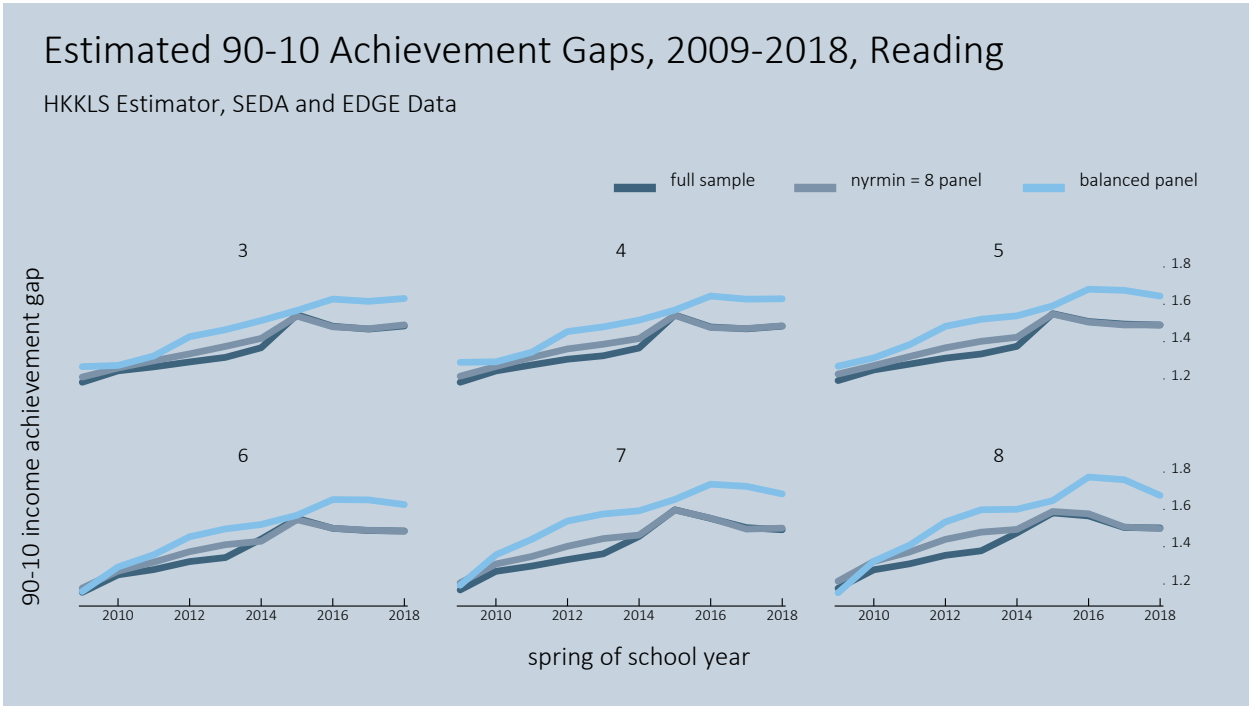
Note: + p<0.10; * p<0.05; ** p<0.01; *** p<0.001. The data include 81 estimates of SES 75-25 achievement gaps from Hanushek et al (2020), Appendix Table A2. The model regresses the gap estimate on a linear birth cohort term centered at 1970, and include study (PISA, TIMSS, NAEP, NAEP-LTT)-subject (math, reading, science)-age fixed effects. The intercept is the estimated average gap among the cohort born in 1970; the coefficient on cohort represents the average within-study-subject change in the gap across cohorts). The estimated percentage change from 1970-2000 is computed by multipling the annual change estimate by 30, and then dividing by the magnitude of the gap in 1970. The models with non-redundant observations include only one observation per sample of students; specifically they include only the focal PISA assessment results in each year (reading in 2000 and 2009; math in 2003 and 2012; science in 2006 and 2015) and a single observation for TIMSS in each year, in which the outcome is the average of the TIMSS math and science gaps in that year.

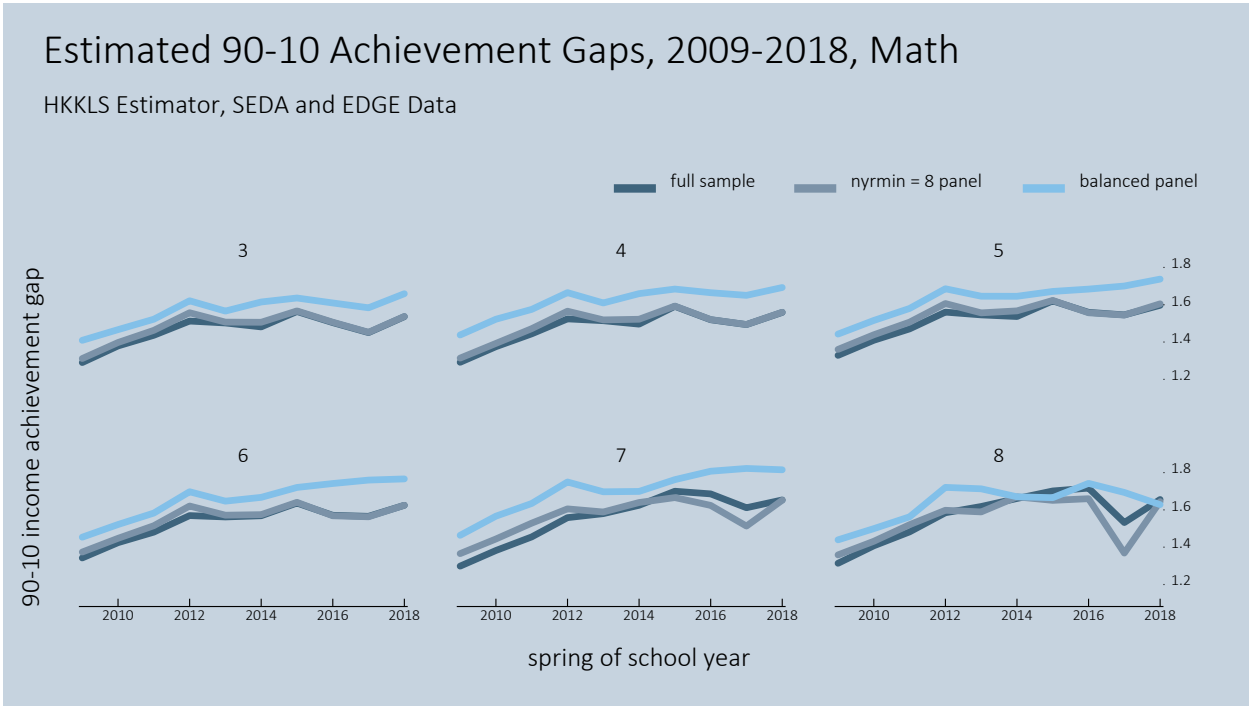
Table 4: Estimated Linear Trends in the Q4-Q1 SES Test Score Gap, by Study

	Study			
	NAEP-LTT	Main NAEP	TIMSS	PISA
Gap in 1970	0.7787 ***	0.7432 ***	0.7410 **	1.4368 **
	(0.0260)	(0.0409)	(0.1316)	(0.0990)
Annual Change	0.0059 ***	0.0139 ***	0.0100	-0.0150 +
	(0.0017)	(0.0016)	(0.0060)	(0.0043)
Implied % Change in Gap 1970-200(	**23%** ***	**56%** ***	40% ns	-31% +
Sample and model specifications				
Birth Cohorts Included	>1970	>1970	>1970	>1970
All Observations				
Excluding PISA 2000 M&S, 2003 S				
Non-redundant Observations	x	x	x	x
Gap Specification				
2000 SD				
Contemporaneous SD	x	x	x	x
Reliabilty Adjustment	x	x	x	x
Fixed Effects Included				
Study-Subject-Age FEs	x	x	x	x
N (gap estimates)	32	14	6	6
N (study-subject-ages)	4	2	2	3

Note: + p<0.10; * p<0.05; ** p<0.01; *** p<0.001. The data include 58 estimates of SES 75-25 achievement gaps from Hanushek et al (2020), Appendix Table A2. The models regress the gap estimate on a linear birth cohort term centered at 1970, and include subject-age fixed effects. The intercept is the estimated average gap among the cohort born in 1970; the coefficient on cohort represents the average within-study-subject change in the gap across cohorts). The estimated percentage change from 1970-2000 is computed by multipling the annual change estimate by 30, and then dividing by the magnitude of the gap in 1970. The models include only non-redundant observations: only one observation per sample of students; specifically they include only the focal PISA assessment results in each year (reading in 2000 and 2009; math in 2003 and 2012; science in 2006 and 2015) and a single observation for TIMSS in each year, in which the outcome is the average of the TIMSS math and science gaps in that year.

Figure A1


Estimated 90-10 Achievement Gaps, 2009-2018, Reading
HKKLS Estimator, SEDA and EDGE Data
full sample
nyrmin = 8 panel
balanced panel
3
4
5
6
7
8
90-10 income achievement gap
1.2
1.4
1.6
1.8
2010
2012
2014
2016
2018
spring of school year


Estimated 90-10 Achievement Gaps, 2009-2018, Math
HKKLS Estimator, SEDA and EDGE Data
full sample
nyrmin = 8 panel
balanced panel
3
4
5
6
7
8
90-10 income achievement gap
1.2
1.4
1.6
1.8
2010
2012
2014
2016
2018
spring of school year

Appendix: Sources of potential bias in the HKKLS estimator

1. The parameters of interest

HKKLS (2020) want to estimate the coefficient β from a regression of test scores (y) on ln(income) (denoted i):

$$y = \alpha + \beta(lnincome) + e,$$

(A1)

but they do not observe i for individual students. Instead they note that

$$\beta = ICC \cdot \beta_b + (1 - ICC)\beta_w$$

(A2)

where ICC is the between-school proportion of variance in $lnincome$, where β_b is the coefficient from the between-school regression

$$\bar{y}_j = \beta_b \bar{\imath}_j + \bar{e}_j,$$

(A3)

and β_w is the coefficient from the within-school regression

$$y - \bar{y}_j = \beta_w\left(i - \bar{\imath}_j\right) + e.$$

(A4)

Given this, they turn their attention to estimating the ICC, β_b, and β_w.

2. Estimating β_w from individual data

Given student-level data on income and achievement, β_w can be obtained from a FE model. If β_{wj} is the coefficient within school j, then β_w is an enrollment and variance-weighted average of the school-specific coefficients (where σ_j is the variance of i within school j, and σ is the total within-school variance of i):

$$\beta_w = \frac{\sum n_j \sigma_j \beta_{wj}}{\sum n_j \sigma_j} = \overline{\beta_{wJ}} + \frac{J}{\sum n_j \sigma_j} cov\left(\beta_{wj}, n_j \sigma_j\right)$$

(A5)

If β_{wj} doesn't vary, then this is simple. But if β_{wj} covaries with school size or school income variance, then the expected value of the FE estimate is not equal to the average of the β_{wj}'s. The point here is that the β_w needed for equation 2 above is not the average within-school slope if $cov\left(\beta_{wj}, n_j \sigma_j\right) \neq 0$. If we assume that n_j is roughly constant (as it is in the NAEP data used by HKKLS), we can simplify this as

$$\beta_w = \frac{\sum \sigma_j \beta_{wj}}{\sum \sigma_j}$$

$$= \overline{\beta_{wJ}} + \frac{1}{\bar{\sigma}} cov(\beta_{wj}, \sigma_j)$$
$$= \overline{\beta_{wJ}} \left(1 + CV(\beta_{wj}) \cdot CV(\sigma_j) \cdot corr(\beta_{wj}, \sigma_j)\right)$$

(A6)

3. Estimating $\boldsymbol{\beta_w}$ from a variance-on-variance regression

HKKLS do not have student-level income, so cannot estimate the FE model above. Instead they do the following:

Compute the variance of both sides of equation 4 within each school:

$$var(y - \bar{y}_j)|j \; = var(\beta_{wj}(i - \bar{\imath}_j) + e)|j$$
$$\tau_j = \beta_{wj}^2 \sigma_j + \omega_j$$

(A7)

Now write this as

$$\tau_j = \overline{\beta_{wJ}^2}\sigma_j + (\beta_{wj}^2 - \overline{\beta_{wJ}^2})\sigma_j + \omega_j$$
$$= \overline{\beta_{wJ}^2}\sigma_j + (\beta_{wj}^2 - \overline{\beta_{wJ}^2})\sigma_j + \omega_j$$
$$= \gamma\sigma_j + v_j$$

(A8)

where $\gamma = \overline{\beta_{wJ}^2}$ (and note that $\overline{\beta_{wJ}^2} = \overline{\beta_{wJ}}^2 + var(\beta_{wj})$) and

$$v_j = (\beta_{wj}^2 - \gamma)\sigma_j + \omega_j.$$

(A9)

Now the OLS regression of equation (A8) yields:

$$\hat{\gamma} = \frac{cov(\sigma_j, \tau_j)}{var(\sigma_j)}$$
$$= \frac{cov(\sigma_j, \gamma\sigma_j + [(\beta_{wj}^2 - \gamma)\sigma_j + \omega_j])}{var(\sigma_j)}$$
$$= \gamma + \frac{cov(\sigma_j, (\beta_{wj}^2 - \gamma)\sigma_j)}{var(\sigma_j)}$$
$$\approx \gamma + \bar{\sigma}\frac{cov(\sigma_j, \beta_{wj}^2)}{var(\sigma_j)}$$
$$\approx \overline{\beta_{wJ}}^2 + var(\beta_{wj}) + 2\overline{\beta_{wJ}}\bar{\sigma}\frac{cov(\sigma_j, \beta_{wj})}{var(\sigma_j)}$$

$$\approx \overline{\beta_{wj}}^2 \left(1 + CV^2(\beta_{wj}) + 2\frac{CV(\beta_{wj})}{CV(\sigma_j)} corr(\sigma_j, \beta_{wj})\right)$$

$$\approx \beta_w{}^2 \left[1 + CV(\sigma_j) CV(\beta_{wj}) corr(\beta_{wj}, \sigma_j)\right]^{-2} \left(1 + CV^2(\beta_{wj}) + 2\frac{CV(\beta_{wj})}{CV(\sigma_j)} corr(\sigma_j, \beta_{wj})\right) \quad \text{(A10)}$$

The last line follows from equation (A6) above. HKKLS take the square root of $\hat{\gamma}$ as an estimate of β_w.

4. Bias in the estimate of $\boldsymbol{\beta_w}$ from the variance-on-variance regression

There are two sources of bias in $\hat{\gamma}$ as an estimator or β_w^2: there will be bias if 1) $var(\beta_{wj}) \neq 0$; the bias will have additional terms if $corr(\sigma_j, \beta_{wj}) \neq 0$.

If $var(\beta_{wj}) = 0$, then $CV(\beta_{wj}) = 0$, and $E[\hat{\gamma}] = \overline{\beta_{wj}}^2 = \beta_w^2$; (but even then the $E[\sqrt{\gamma}] \neq \beta_w$).

In the absence of a correlation between σ_j and β_{wj}, $\hat{\gamma}$ will be biased upwards by a ratio of $1 + CV^2(\beta_{wj})$.

Equation (A10) shows that the estimate will be biased (roughly, ignoring the fact that $E\left[\sqrt{\hat{\beta}_w}\right] \neq \beta_w$) by a multiplicative factor

$$bias = \frac{\sqrt{1 + CV^2(\beta_{wj}) + 2\frac{CV(\beta_{wj})}{CV(\sigma_j)} corr(\sigma_j, \beta_{wj})}}{\left|1 + CV(\sigma_j) CV(\beta_{wj}) corr(\beta_{wj}, \sigma_j)\right|}. \quad \text{(A11)}$$

The bias ratio is not a simple expression. In the absence of a correlation between the within-school slope and the within-school variance ($corr(\sigma_j, \beta_{wj}) = 0$); variance in β_{wj} will bias $\hat{\gamma}$ upwards. But the bias becomes more complex when the correlation is not zero. In general, the estimate will be biased downward when $corr(\sigma_j, \beta_{wj}) < 0$ and $CV(\sigma_j) = \frac{sd(\sigma_j)}{\bar{\sigma}}$ is small (little variance across schools of the within-school income variance); the effect of increasing the $CV(\beta_{wj})$ is ambiguous; it may increase or decrease the bias multiplier, depending on the values of the other terms.

HKKLS note that sometimes their estimate of $\hat{\gamma}$ is negative (which makes it impossible to take its square root). That suggests potential downward bias or considerable sampling variance in $\hat{\gamma}$. Equation (A11) shows that the bias factor could make $\hat{\gamma}$ negative: the bias will be downward if $Var(\beta_{wj}) > 0$ and

$$\frac{corr(\sigma_j, \beta_{wj})}{CV(\sigma_j)} < -\frac{1}{2} CV(\beta_{wj});$$

and the bias will be sufficiently downward to make $E[\hat{\gamma}]$ negative if

$$CV(\beta_{wj})\left[CV(\beta_{wj}) + 2\frac{corr(\sigma_j, \beta_{wj})}{CV(\sigma_j)}\right] < -1.$$

5. Measurement-error induced bias

A further complication is introduced because σ_j is not known, but is estimated with error (considerable error in all likelihood, given that it's inferred from ACS estimated income distributions in tracts surrounding a school). This will bias γ toward 0, assuming the error in $\hat{\sigma}_j$ is classical. If the error in $\hat{\sigma}_j$ is not classical – if the measurement error is correlated with the within-school slope – the estimator will have additional bias, whose direction will depend on that correlation.