THE INCOME-ACHIEVEMENT GAP AND ADULT OUTCOME INEQUALITY

ERIC R. NIELSEN

THE FEDERAL RESERVE BOARD

ABSTRACT. This paper documents a dramatic decrease in the achievement gap between youth from high- and low-income households from 1980 to 1997. It does so using ordinal methods that do not rely on two implausible assumptions pervasive in empirical work assessing achievement differences using test scores. In particular, the methods used in this paper do not assume that a given (normalized) test score has a fixed meaning over time, nor do they assume that test scores are cardinal measures of achievement. The paper shows that any weighting scheme that places more weight on higher test scores must conclude that the income-achievement gap in reading narrowed between the National Longitudinal Surveys of Youth 1979 and 1997 (NLSY79 and NLSY97). The situation for math achievement is more complex, but it is nonetheless clear that low-income youth in the middle and high deciles of the low-income math achievement distribution unambiguously gained relative to their high-income peers. Finally, an anchoring exercise suggests that the changes in the math and reading achievement distributions between the NLSY79 and the NLSY97 correspond to a convergence in the present discounted value of lifetime labor income of about $100,000 and in high school and college completion rates of about 0.06 to 0.07 for youth from high- versus low-income households. JEL Codes: I24, I21, J24, C18, C14.

1. INTRODUCTION

Many literatures in economics use test scores to measure group differences in achievement. For example, researchers employ test scores to measure changes in the black-white

Date: July 2016.

I thank Derek Neal, Gary Becker, Ali Hortaçsu, Sandy Black, Armin Rick, and seminar participants at the University of Chicago, the Federal Reserve Board, and the DC Education Working Group for helpful comments and feedback. The views and opinions expressed in this paper are solely those of the author and do not reflect those of the Board of Governors or the Federal Reserve System. Contact: Division of Research and Statistics, Board of Governors of the Federal Reserve System, Mail Stop 97, 20th and C Street NW, Washington, D.C. 20551. eric.r.nielsen@frb.gov. (202) 872-7591.

achievement gap over time, to assess school and teacher quality, and to quantify the importance of class size, funding, and other school inputs in generating student achievement. A fundamental question in all such applications is how best to use test-score data to measure student achievement. The typical methods employed in economics provide a deeply inadequate answer to this question.

In almost all empirical work, researchers first normalize test scores to have a mean of zero and a standard deviation of one within each year/age cohort. These "z-scores" are then compared across different cohorts using standard statistical techniques, such as mean differences and ordinary or instrumental least-squares regression. Unfortunately, such standard techniques make two very strong, unwarranted assumptions on the z-scores: cardinal and inter-group comparability. The cardinal comparability assumption (also referred to as the "interval scale" assumption) states that a score change of Δs represents the same change in achievement at all starting locations in the test scale. The inter-group comparability assumption states that a given test score has a fixed interpretation across all comparison groups, so that students from different groups (cohorts, grades, classrooms, etc.) may be ranked against each other consistently based on their test scores.

Neither cardinal nor inter-group comparability is well justified by either economic or psychometric theory. Cardinal comparability will be violated if improvements in some regions of the test-score distribution are more valuable than improvements in others. Since the notion of value is context-specific, a given test scale may be cardinal for some applications and not cardinal for others. As an example, in common economic applications, the presence of achievement thresholds or convex returns to skill will violate cardinal comparability. Inter-group comparability will be violated for achievement comparisons made over time in the very likely occurrence that the location of the underlying achievement distribution is shifting over time; inter-group comparability rules out by assumption secular increases or decreases in achievement. Emphasizing the implausibility of these assumptions is not mere pedantry; standard methods may produce severely biased estimates under even mild failures of either assumption. It is even possible for standard methods to misidentify the sign of the relevant achievement gap/change in such cases.

This paper studies changes in achievement inequality by parental income using methods that do not assume that standardized test scores satisfy either inter-group or cardinal comparability. In particular, I guarantee that inter-group comparability holds throughout my analysis by using only "crosswalked" test scores that have been explicitly and carefully constructed so that test scores are ordinally (though not necessarily cardinally) comparable over time. I next handle cardinal comparability using two complementary approaches: ordinal statistics and anchoring. First, the paper shows how to construct robust achievement gap-change measures using only the rank-order content of achievement test scores; these statistics are invariant under all order-preserving transformations of the test scores. Under intuitive, testable conditions, ordinal statistics allow one to unambiguously sign an achievement gap or gap change for all plausible cardinalizations of achievement. Ordinal methods cannot be used, however, to measure the magnitudes of achievement shifts. Therefore, rather than simply assume that observed test scores are cardinal measures of achievement, my second approach uses the reduced-form relationship between test scores and later-life economic outcomes to explicitly construct cardinally interpretable, anchored achievement scales.

Applied to the National Longitudinal Surveys of Youth (NLSY) 1979 and 1997 surveys, my ordinal estimates provide compelling evidence that the income-achievement gap narrowed substantially between 1980 and 1997. In particular, I show that any reasonable cardinalization of reading test scores will measure a smaller income-achievement gap in the NLSY97 than in the NLSY79 (which has test-score data for 1980). This conclusion follows because the low-income reading achievement distribution in 1997 is unambiguously higher than in 1980 in the sense of first-order stochastic dominance (FOSD), while the high-income distribution is unambiguously lower, again in the sense of FOSD. I cannot make such a strong claim about math achievement because low-achieving, low-income students, along with all high-income students, suffered an adverse shift in their math achievement between 1980 and 1997, while high-achieving, low-income students improved unambiguously. Nonetheless, any scheme that does not place too much weight on the bottom of the math achievement distribution would find a smaller math achievement gap in 1997 than in 1980. Additional ordinal measures, such as Cliff's δ and percentile-percentile

curves (PPCs), also all uniformly suggest that there was a large, statistically significant decrease in the income-achievement gap between 1980 and 1997. These estimates are robust to income and test-score measurement error, as well as to various methods for adjusting income to reflect household size and composition.

The anchored gap-change estimates further imply that the ordinal convergence in high- and low-income test scores between the NLSY79 and the NLSY97 corresponds to economically large shifts in both lifetime labor wealth and school completion rates at constant skill prices. Methodologically, I use longitudinal data in the NLSY79 to flexibly estimate the conditional distributions of labor wealth and school completion for a range of different achievement test scores. I then use these estimated conditional distributions as skill-pricing functions to understand the distributional consequences of the test-score shifts from the NLSY79 to the NLSY97. This approach implies that the decrease in the reading achievement gap between high- and low-income youth corresponds to a narrowing of the mean and median lifetime earnings gaps of \$50,000 to \$130,000 in present discounted dollars and decreases in the mean high school and college completion gaps of 0.05 to 0.07 probability units. In addition to these central tendency estimates, I also show that the anchored gap changes are highly heterogeneous across the supports of the test score distributions.

Both my ordinal and anchored findings stand in sharp contrast to recent work by Reardon[31], whose standard, cardinal approach finds no significant change in the income-achievement gap in the NLSY data and a large increase in the gap over the last several decades in a number of other data sources. My estimates are not directly comparable to Reardon's because we study different subsamples of the NLSY and because he covers a wider range of years using several additional data sources. Nonetheless, my results suggest that whether one treats test scores cardinally or not can matter critically in important empirical applications.

These empirical results are significant in their own right, aside from their unorthodox methodology. The income-achievement gap should be a major determinant of intergenerational mobility and adult outcome inequality. Rising income inequality in recent decades has led to widespread concern that children from different income classes may increasingly

face divergent life prospects. My estimates suggest that these concerns, though understandable, might not be born out in the data. Nonetheless, my results come with the significant caveat that they cover only one data source, the NLSY, and one specific time period (1980-1997). It is quite possible that these data are exceptional and that the long-term trend is toward increasing achievement inequality by family income. Moreover, a smaller income-achievement gap could still translate to greater adult economic inequality if the relationship between achievement and adult economic outcomes convexifies sufficiently.

My findings should give researchers employing standard methods pause. Ordinal methods disagree with traditional cardinal approaches in at least one important setting, and other findings that use standard methods may be similarly fragile. Ordinal methods are prima facie more credible, as the conditions needed for them to provide valid inference are substantially weaker than those required by cardinal methods. Moreover, my anchoring analysis suggests that the anchored gap/change estimates one obtains often depend crucially how one chooses to weight different moments of the anchoring distribution. When possible, researchers should use methods that do not rely on economically arbitrary cardinalizations of achievement.

The techniques used in this paper apply to any situation in which a researcher wishes to use test-score data to measure group differences in achievement; nothing about the analysis assumes that time is the dimension of comparison. For example, the methods in this paper could be used to study black-white achievement inequality in the American South versus the North at a given point in time, or the achievement of suburban students versus urban students across different metropolitan areas. More generally, these methods could be applied whenever the scale of the variable of interest is unknown. Happiness scales, poverty indexes, and much else could be fruitfully analyzed using ordinal methods or by anchoring the scales to some interpretable outcome.

The rest of this paper is organized as follows. Section 2 reviews the literatures assessing achievement differences using test scores. Section 3 describes the NLSY data. Section 4 presents a general framework for valuing achievement shifts using test scores. Section 5 presents the results of the ordinal analysis, while section 6 presents the anchoring

results. Sections 7-8 discuss the connection between my results and the larger literatures on parental investments and childhood human capital creation. Appendices A-D contain all the tables and figures, as well as additional technical details. An online appendix presents supplemental empirical and theoretical analysis.[1]

2. Literature Review

Virtually all empirical work in economics using test scores assumes that they are cardinal measures of achievement. The most intensely studied achievement gaps in the U.S. context are by race: Fryer and Levitt[21, 22], Clotfelter, Ladd, and Vigdor[8], Duncan and Magnuson[11], Hanushek and Rivkin[17], among many others, assume that z-scores are cardinal measures of achievement in order to assess changes in the black/white test-score gap over time. Cardinality assumptions are also pervasive in the literatures assessing the effects of school inputs (Krueger[23], Hoxby[19], ...), teacher quality (Raudenbush[29], ...), and parental income (Reardon[31]) on student achievement.

Of the many empirical papers that treat test scores cardinally, Reardon[31] is closest in subject matter to this paper. Reardon defines the income-achievement gap as the difference in average z-scores between students at the 90th vs. the 10th percentiles of the household income distribution. Using regression-based methods, he estimates that this gap is 30 to 40 percent larger for students born in 2001 than for those born three decades earlier. My paper does not directly address Reardon's larger thesis, as I study a much shorter time period while using a strict subset of the surveys that he employs. The large increases that Reardon documents come from cross-sectional income-achievement gaps estimated using surveys other than the NLSY79 and NLSY97. It is possible that my methods would also find an increasing income-achievement gap with these alternate data. However, Reardon does use both NLSY surveys in his analysis, and he calculates a negligible change in the income-achievement gap with these data, while I find a sharp decrease. In a parallel working paper, I extend my ordinal estimates to some of the other data sources used by Reardon. I find that the trends estimated using these other data

[1]This appendix can be found at the following url: `https://sites.google.come/site/ericnielsenecon/research`.

are substantially more muddled and that longer-term trends are only identified assuming particular, arbitrary cardinalizations of achievement.

I am not the first author to point out that test scores are not cardinal measures of achievement. Stevens[35] argues that most psychometric scales are inherently ordinal, while Lord[25] shows that the IRT scale scores commonly used as achievement measures in economics are ordinal in the sense that there are infinitely many rescalings of these scores that will fit any given set of raw test responses equally well. Jacob and Rothstein[20] caution again uncritically using test scores cardinally on the way to discussing many empirical challenges for economists wishing to use standard psychometric measures. Bond and Lang[6] and Reardon[30] show that it is often possible to flip the sign of standard achievement gap/change estimates using order-preserving transformations of test scores.[2] Nielsen[27] refines these arguments by demonstrating that even mild transformations of test scores are often sufficient to affect such sign flips.

Heckman and coauthors[10, 9], along with many others, argue for creating cardinally interpretable scales by anchoring test scores on life outcomes, an approach I take in Section 6 of this paper. The basic idea is to use longitudinal data to estimate the relationship between test scores and some outcome of interest, such as labor market earnings. This relationship can then be used to re-denominate test-score shifts into interpretable units. Anchoring offers an appealing solution to the arbitrariness of test scales. However, the approach has a number of significant downsides. Most importantly, anchoring requires longitudinal data on outcomes measured years or even decades after the test date in order to accurately capture lifetime differences between test-takers. For many pressing questions, waiting 20 to 30 years for an answer is not a viable option. In addition, estimates based on anchored scores may be sensitive to the particular outcome chosen as the anchor and to the functional forms used to implement the anchoring procedure. Finally, handling test-score measurement error plausibly is quite challenging for many anchoring methodologies. Therefore, ordinal methods and anchoring are best viewed as complements; ordinal approaches are feasible in wider variety of applications, but where

[2]Schroeder and Yitzhaki[32] make a similar argument using self-reported satisfaction measures.

anchoring analysis is possible, it will often produce more interpretable and compelling estimates.

3. Data

This paper uses the National Longitudinal Surveys of Youth 1979 and 1997 (NLSY79 and NLSY97). These are high-quality, nationally representative surveys of young adults with detailed data on respondents' family backgrounds, academic achievement, and later-life outcomes. This section briefly describes the construction of the analysis sample. Please refer to Appendix B for a more detailed discussion.

I define household income in both surveys using a comprehensive measure that sums together income from all sources (wages, capital, and government transfers) across all household members.[3] I obtain fairly similar estimates if I use instead parental wage income, which is more tightly related to parental education and human capital, to define my high and low categories. To reduce the influence of measurement error and transitory components of income on my estimates, I average the yearly household incomes over the first three years of each survey. Using only one year of income data yields qualitatively similar, though attenuated, gap-change estimates. Finally, my baseline estimates define high- and low-income to be the top and bottom quintiles of the cross-sectional household income distributions; different percentile cutoffs tell largely the same story.

My primary measures of achievement are the Armed Forces Qualifying Test (AFQT) and its math and reading subscores.[4] The AFQT is a high quality achievement test that has been shown to be predictive of later-life outcomes. Additionally, the AFQT and its constituent subtests have the very important, if unheralded, property that it is possible to put scores from the NLSY79 and NLSY97 versions on a common scale such that respondents from the two surveys can be ranked against each other consistently. Such a scaling is possible thanks to a percentile-equated crosswalk constructed from a sample

[3]This definition of household income includes income earned by the respondents themselves, in addition to the income earned by their parents. Since the respondents I study are less than 18 years old, their share of total household income is usually negligible.

[4]The math subtests used to define the math subscore of the AFQT changed in 1989; throughout, I will use the current, post-1989 definition. Using the old definition results in somewhat larger and more statistically significant estimated decreases in the math income-achievement gap.

of test takers who were randomly assigned to the NLSY79 and NLSY97 versions of the AFQT.[5]

The NLSY is an ideal setting for carrying out my anchoring analysis as both surveys collect a wealth of longitudinal data on income, education, employment, and many other outcomes annually or biennially through 2012/13. I construct college and high school completion indicators using the highest reported grade completed from any of the first 14 years of each survey. I construct the present discounted value of labor income (pdv_labor) using round-by-round data on labor earnings, employment status, and hours worked in the NLSY79.

The construction of pdv_labor is complicated by the fact that wage income data is often not available in a given round for a given respondent due to item non-response, non-interview, unemployment, or labor force non-participation. Rather than model selection into observing wage income explicitly, I handle these various forms of missing data by adopting extreme imputation rules. In particular, for each missing wage for each survey respondent I optimistically (pessimistically) impute a wage equal to the maximum (minimum) ever observed for that respondent. Although these imputation rules do not allow me to bound the estimands of interest, the fact that they produce qualitatively similar estimates suggests that selection is not driving my conclusions.[6]

Similarly, it is unclear how un(der)employment should be treated in the calculation of pdv_labor. If labor supply is always freely chosen, then full income (wage×full-time

[5]In particular, the AFQT changed from a pencil-and-paper format to a computer adaptive design between the two NLSY surveys. When this format change occurred, the military (the main user of the AFQT) randomly assigned a group of recruits to take one version of the test or the other. Segall[33] constructs the crosswalk between these two groups of test-takers by equating the component scores based on their percentile ranks. Altonji, Bhadarwaj, and Lange[2] take this crosswalk and add to it percentile-mapped crosswalk across different respondent ages in order to make inter-group comparable scores that are comparable both across NLSY surveys and across respondent ages. These authors make this crosswalk available at the following url: `http://www.econ.yale.edu/~fl88/data.html`. Creating a 1980-equivalent AFQT score by adding these crosswalked subscores together is not strictly valid because it ignores the covariance structure of the different ASVAB components. However, Segall[34] reports that the AFQT scores resulting from such a procedure are virtually identical to those obtained by crosswalking the AFQT scores directly.

[6]These optimistic and pessimistic imputation rules are quite extreme. Suppose that a respondent reports wage income of $100 in 1983 at the age of 20, and then reports wage income of $100,000 for 2000, 2002, and 2006, but has no wage income recorded for 2004. The pessimistic imputation rule will assign to this individual a wage income for 2004 of at most $100. Similarly, if the same individual has wage income missing in 1982, when she was 19 years old, the optimistic imputation rule will assign her a wage income of at least $100,000 for that year.

hours), regardless of actual hours worked, is the relevant measure of labor income, while if unemployment is always involuntary, observed annual earnings are the correct measure. As with the missing wage data, I avoid taking a stand on which view of the labor market is closer to the truth and instead estimate anchored gap changes under both assumptions. Fortunately, as Section 6 demonstrates, the anchored gap-change estimates under these polar opposite imputation procedures are quite similar.

I restrict my analysis to respondents who were between the ages of 15 and 17 when they took the AFQT. I select this age range for two reasons. First, the two NLSY surveys have fairly different age distributions, yet both have large numbers of respondents in this age range. Focusing on this age group therefore reduces the importance of whether and how I account for different respondent ages in my analysis. In principle, the crosswalked scores should be inter-group comparable across different ages, and qualitatively I find extremely similar results whether I estimate income-achievement gaps/changes separately for 15, 16, and 17 year-olds, if I use age/survey standardized z-scores, or if I ignore age entirely and estimate gaps/changes using the crosswalked scores pooled across these three age groups. Second, respondents in this age range were just on the cusp of becoming independent adults when they took the AFQT. Test scores for such students therefore provide a summary measure of the cumulative effect of parental income on achievement through the middle/end of high school.

Table 1 displays summary statistics for the NLSY79 and NLSY97 samples I analyze. My final samples include about 3,800 respondents in the NLSY79 and roughly 2,800 in the NLSY97. The two samples are quite similar in terms of race/sex composition and crosswalked achievement. Real household incomes are higher in the NLSY97, as expected. Notably, the variance of household income is much greater relative to its mean in the NLSY97, consistent with the well-documented widening of the income distribution in recent decades. Turning to life outcomes, about 88% of the NLSY79 sample has at least a high school education, and roughly 23% have at least 4 years of post-high school education. The average present discounted labor incomes in the NLSY79 range from about \$371,000 (pessimistic imputation + fixed labor supply) to \$1,017,000 (optimistic imputation + fully chosen labor supply).

The NLSY surveys are not the only sources for nationally representative data on student achievement and parental income. Surveys such as the National Education Longitudinal Study (1988) and High School and Beyond could be used to carry out the ordinal analysis undertaken in sections 4 and 5. I do not use these data for several reasons. First, these surveys do not have continuous, high-quality measures of parental income, which makes constructing comparable high- and low-income categories challenging. Second, these surveys do not contain achievement tests whose scores satisfy inter-group comparability. Third, the age ranges covered by these surveys vary widely, making the direct comparison of achievement scores between them difficult or impossible without taking a strong stand on how test scores should be compared across different ages. Fourth, these surveys do not contain detailed, data on long-run economic outcomes which rules out the anchoring analysis undertaken in Section 6.

4. Valuing Achievement Shifts

Consider a group of students with test scores distributed according to cumulative distribution function (cdf) F. How should the distribution F be evaluated? If F_A and F_B are the test-score distributions for student groups A and B, how should we determine whether A or B has more achievement? The standard approaches in economics to answering these questions assume that the scores themselves are cardinal measures of achievement. If test scores are cardinal, the value of a score distribution F can be consistently estimated by a sample mean of test scores. Analogously, the gap between A and B can be consistently estimated simply by the difference of sample means as long as cardinal comparability holds.

As pointed out in the introduction, using test scores in this way implicitly makes two very strong assumptions. First, it assumes inter-group comparability, that is, that a given normalized test score corresponds to the same underlying level of achievement across different comparison groups. This assumption is particularly implausible if the achievement tests used are very different from each other, or if the students being compared are very different in terms of age and background. Why would a calculus test score one standard

deviation above the mean for a high school senior correspond to the same level of achievement as a basic numeracy score one standard deviation above the mean for a first-grader? The second major assumption is cardinal comparability: that is, that the value of a test-score gain is constant throughout the range of possible scores. There is no particular reason to think that this is true; test scores are not designed to have this property, at least not when applied to economic questions. Moreover, I show in Nielsen[27] that even mild violations of this assumption are often sufficient to flip the sign of an achievement gap/change estimate.

A bit more formalism will make these points clearer. Let $s_{i,G}$ be the test score for student i in group G and let $\psi_G(s)$ be the monotone increasing map from group-G test scores to underlying achievement.[7] Denote by $W(a)$ the true value of achievement level a. Later, I will want to interpret W as the social welfare of the economic outcomes (income, health, etc.) associated with a, but for now suppose only that $W(a)$ is some non-decreasing function that converts achievement into cardinally-comparable units. In other words, W is a test scale such that that the true value of score distribution F_G is the expected value of $W(\psi_G(s))$; that is, such that $V(F_G) \equiv \mathbb{E}_{F_G}[W(\psi_G(s))]$.

Now consider the achievement gap between A and B. Measuring this gap by the difference in group means will only be valid if two conditions hold: (i) $\psi_A = \psi_B \equiv \psi$ and (ii) $W(\psi(s)) = \alpha + \beta s$ for some constants α and $\beta > 0$. The condition $\psi_A = \psi_B$ corresponds to the inter-group comparability assumption, while the condition that W be an affine transformation of observed test scores corresponds to the cardinal comparability assumption. Empirical methods that erroneously assume either (i) or (ii) will in general produce biased gap/change estimates.

[7]This formulation asserts that achievement is unidimensional. Allowing achievement to be multidimensional only slightly complicates the theoretical exposition. Empirically, however, such a modification presents significant difficulties because one must take a stand on which dimensions of achievement are measured by a given achievement test. The condition that ψ_G be monotone-increasing implies that test scores are perfect ordinal measures of achievement in the sense that $s_{i,G} > s_{j,G} \iff a_{i,G} > a_{j,G}$. In reality, of course, test scores are noisy measures of achievement – comparable tests will yield similar, but not identical, rank orderings of a given set of students. Measurement error is a thorny issue that I devote significant effort to handling in the empirical sections of this paper.

It is sometimes possible to guarantee that ψ is fixed across comparison groups. For example, the percentile-mapped crosswalk I use in this paper ensures (at least approximately) that test scores in the NLSY79 and NLSY97 satisfy inter-group comparability. Different versions of many standardized assessments share some items to facilitate the construction of comparable item-response theory scores. Although these common items do not generally allow one to create a cardinally comparable test scale, they can in principle be used to create a test scale that satisfies inter-group comparability.

When ψ is fixed, it is possible to unambiguously rank groups by achievement even when W is unknown, provided that certain stochastic dominance relationships hold between the relevant test-score distributions. The only requirement to affect such a comparison is that W be increasing, which is a natural assumption in this setting; test scores may not be cardinal measures of achievement, but a higher test score should still be better than a lower test score for any halfway sensible test. For example, if W is thought of as a social welfare function defined over various life outcomes, then W should be increasing in a because most positive life outcomes are increasing in achievement.

Consider first the task of assessing the achievement gap between groups A and B. Atkinson[3] showed in the context of social welfare functions that F_A will be preferred to F_B for any increasing W precisely when F_A first-order stochastically dominates ($\succ$) F_B. Under FOSD, changing W will affect the magnitude of the estimated achievement gap but cannot affect its sign. A similar condition guarantees the unambiguous sign identification of achievement gap changes. In particular, consider assessing the change in the gap between A and B from time t to $t+1$. If $F_{A,t} \succ F_{A,t+1}$ and $F_{B,t+1} \succ F_{B,t}$, then the gap between A and B must have narrowed for any weighting function W. These conditions are intuitive; $F_{A,t} \succ F_{A,t+1}$ implies that any weighting function W would measure a decrease in A's achievement over time, while $F_{B,t+1} \succ F_{B,t}$ implies that any W would measure an increase in B's achievement. Combined, these conditions therefore imply that any increasing weighting function would measure a decrease in the gap between A and B.

5. Ordinal Analysis

5.1. **Stochastic Dominance Tests.** The preceding section showed that a combination of FOSD tests can be used to construct ordinal tests of income-achievement gap changes. If high-income youth achievement declined unambiguously, in the sense of FOSD, while low-income achievement increased unambiguously, again in the sense of FOSD, then any cardinalization of achievement test scores will reach the same conclusion about the sign of the gap change. I jointly test these conditions using the procedure developed in Barrett and Donald[4], which is based on Kolmogorov-Smirnoff statistics.[8]

Tables 2 and 3 display the results of these tests for high- and low-income youth. The baseline tests show that low-income reading scores improved and high-income reading scores declined unambiguously. In contrast, the FOSD tests on the math test-score distributions show that while high-income youth suffered an unambiguous adverse shift in math achievement, low-income youth experienced neither unambiguous increases nor unambiguous decreases. The story using the white-only subsample is similar, although with some tests rejecting at 10% rather than 5%, the statistical evidence is somewhat weaker. Nonetheless, low-income white students appear to have improved unambiguously or held steady in reading, but not in math, between 1980 and 1997, while their high-income peers seem to have regressed unambiguously in both achievement measures. Interestingly, the estimates for black youth indicate just the opposite; low-income black youth regressed unambiguously in math, reading, and AFQT, while their high-income peers clearly improved.

The baseline estimates define the high- and low-income categories relative to the race and survey-specific income distributions. In Tables 2 and 3, the "pooled" estimates use the same real-dollar cutoffs to define the high- and low-income groups in both surveys while the "buckets first" estimates define the income percentile cutoffs in each survey prior to subsetting on race. The "pooled" and "buckets first" estimates show that the full-sample

[8] Suppose that we have independent samples $\{x_i\}_{i=1}^{N}$ and $\{y_j\}_{j=1}^{M}$ from two populations X and Y with the same bounded support. Consider testing the null $H_0 : F_y(z) \leq F_x(z) \forall z$ against the alternative $H_1 : \exists z \text{s.t.} F_y(z) > F_x(z)$. Barrett and Donald[4] define the following test statistic: $\hat{S}_1 = \left(\frac{NM}{N+M}\right)^{\frac{1}{2}} \sup_z \left(\hat{F}_y(z) - \hat{F}_x(z)\right)$ and show that the probability of observing $\hat{S}_1$ under H_0 is $exp\left(-2\hat{S}_1^2\right)$.

and white-only results are not sensitive to how the high- and low-income subgroups are defined. Turning to the black-only estimates, these alternative income group definitions still generally indicate that low-income achievement declined unambiguously. However, the FOSD tests now usually yield equality for the high-income distributions, rather than dominance by the NLSY97 distributions. Though weaker than the double-FOSD condition outlined in the previous section, an unambiguous decline for low-income black youth, coupled with no change for high-income black youth, still suggests that the black income-achievement gap widened. However, because the "equality" outcome of the FOSD tests consists of failing to reject two null hypotheses, one should not place too much emphasis on such results. Indeed, some of the Cliff's δ estimates discussed later in this section suggest a narrowing, rather than a widening, income-achievement gap for black youth.

These empirical results are remarkable. With only the assumption that observed, cross-walked test scores are ordinal measures of achievement, I have shown that any plausible test scale will find a decrease in the reading income-achievement gap. This result is much stronger than what one could ever hope to get out of a standard, cardinal estimate of the gap change because such estimates inherently hinge on the assumption that the true scale of achievement has been pinned down (up to affine transformations). A cardinal estimate can at most give a definite answer for the particular scale used, while the FOSD tests can give a definite answer for all possible test scales.

Tables 2 and 3 do not account for test-score measurement error in any way. My results are therefore conservative in the sense that classical measurement error will tend to bias the FOSD tests against a true null that one distribution dominates the other. To see why, suppose group A's true scores dominate group B's but that only noisy test scores are observed. Measurement error will make the observed test-score distributions less distinct from each other compared to the true distributions, which increases the likelihood of an erroneous rejection of the true null in finite samples because the empirical cdfs are more likely to be close to each other.

These FOSD results, while very strong in one sense, say nothing about the magnitudes of the high- and low-income test-score shifts and are also not informative about which parts of the relevant test-score distributions are driving the rejections (or lack thereof)

of stochastic dominance. Therefore, in the following sections, I present several additional ordinal estimates that do give a sense of the relative magnitudes of the test-score shifts that are driving the FOSD results. These estimates also provide much more information about how high- and low-income achievement distributions have shifted over time.

An additional shortcoming of FOSD tests is that they will yield inconclusive answers in many empirically relevant settings. If there is no way to construct a plausibly cardinal scale (say, through anchoring), then the ordinal methods discussed below provide an appealing alternative to both FOSD tests and standard cardinal approaches. These methods, like standard cardinal methods, will not be able to say anything about absolute changes in achievement, but they may still identify relative changes. Moreover, unlike cardinal approaches, these methods are robust to any order-preserving transformation of the test scores, and unlike FOSD tests, they are relatively likely to return a definite answer in real-world settings.

5.2. **Percentile-Percentile Curves.** This section uses percentile-percentile curves (PPCs) to document shifts in the income-achievement gap in the NLSY data. Let L and H denote youth from low- and high-income households. The PPC for L relative to H simply plots the percentiles of the group-L scores in the distribution of group-L scores against the percentiles of the group-L scores in the distribution of the group-H scores. Formally, let F_L and F_H be the cumulative distribution functions (cdfs) for low- and high-income students. The population PPC for L relative to H is given by $\{(p_i, q_i)\}$ $i \in L$, where $p_i = F_L(s_i)$ and $q_i = F_H(s_i)$. The PPC summarizes how the low-income scores compare to the high-income scores. If the scores in L tend to be lower than the scores in H, the corresponding PPC will lie below the 45-degree line, since the pth percentile in the group-L score distribution will correspond to the qth $< p$th percentile in the group-H score distribution. The further below the 45-degree line the PPC is, the more the scores in H dominate those in L.

Comparing PPCs from different surveys allows one to assess changes in the relative achievement of low-income youth. Shifts in the PPCs closer to (further from) the 45-degree line indicate decreasing (increasing) differences in the score distributions between

the high- and low-income youth. An important caveat is that only relative changes in test scores between groups L and H are detectable; if both groups are experiencing secular increases (or decreases) in their test scores, the PPCs will show no change. The empirical PPCs created using income in the NLSY surveys look very much like Lorenz curves. This is no accident, as the definition of a PPC is very similar to the definition of a Lorenz curve. Since high-income test scores always dominate low-income test scores in both NLSY surveys, the PPCs in this paper will all be below the 45-degree line. However, unlike Lorenz curves, this is not true by construction.

Figure 1 displays the math, reading, and AFQT income-achievement PPCs.[9] The large income-achievement gap in both surveys is clear from the great distance between each PPC and the 45-degree line. For example, the median reading score in the high-income distribution of the NLSY79 corresponds roughly to the 90th percentile in the low-income score distribution. Additionally, the NLSY97 reading and AFQT PPCs lie uniformly closer to the 45-degree line than the NLSY79 PPCs. This indicates that the score distributions for high- and low-income students are more similar to each other in the NLSY97 than in the NLSY79. The convergence in the math score distributions is more localized; the NLSY97 curve lies strictly above the NLSY79 curve only between the 50th and 90th test-score percentiles of the low-income distribution.

Figure 2 displays black-white achievement inequality by designating white respondents as the H group and black respondents as the L group. Both curves are very far below the 45-degree line, reflecting substantial black-white achievement inequality in both surveys. Furthermore, the NLSY97 curve is always above the NLSY79 curve, which implies that the black and white test-score distributions became more similar between the two NLSY surveys. This result is consistent with the findings on black-white achievement convergence documented in Altonji et al.[2], Neal[26], and elsewhere.

The narrowing of the black-white gap over this time period is a strong force pushing against a widening income-achievement gap. Since black students tend to come from more economically-disadvantaged families than white students, their relative improvement

[9]These PPCs are estimated using the obvious sample analogues $\hat{p}_i = \hat{F}_L(s_i)$ and $\hat{q}_i = \hat{F}_H(s_i)$, where $\hat{F}_L$ and $\hat{F}_H$ are the empirical cdfs of the high- and low-income score distributions.

implies that a large subpopulation of low-income families gained on relatively wealthy white families. Low-income white students would have to have fallen much farther behind their wealthier white peers in order for the change in the overall income-achievement gap to have been flat or positive.

The PPC framework can also be adapted to understand why some of the FOSD tests in Section 5.1 failed. Figure 3 plots the PPCs for high (low) income students in 1980 relative to high (low) income students in 1997. A necessary and sufficient condition for a 1997 distribution to dominate its 1980 counterpart is that this population PPC lies everywhere below the 45-degree line. Analogously, a 1980 distribution dominates if and only if its population PPC lies everywhere above the 45-degree line. The reading and AFQT PPCs in Figure 3 therefore simply confirm graphically the FOSD results. The high- and low-income PPCs for math present a more complex picture. The high-income math PPC is everywhere above the 45-degree line, suggesting that high-income math achievement deteriorated unambiguously between the two surveys. The low-income PPC for math is above the 45-degree line for scores below the 45th percentile and below the 45-degree line for scores above the 45th percentile. This suggests that the low end of the performance distribution shifted down among low-income students, while the high end shifted up. It is the downward shift at the bottom end of the low-income achievement distribution that is driving the rejection of FOSD; a weighting scheme that placed a lot of emphasis on the bottom end of the achievement distribution could potentially assess a larger math income-achievement gap in the NLSY97 than in the NLSY79.

5.3. **Cliff's δ.** The PPCs provide suggestive evidence that the income-achievement gap narrowed substantially between 1980 and 1997, but they do not readily admit formal hypothesis testing. I therefore seek a test statistic that allows me to conduct inference on shifts in the relative percentile distributions. Furthermore, since the relative percentiles are themselves not cardinally interpretable, the statistic should be ordinal in the relative percentiles. Cliff's δ is an easy-to-compute and easy-to-interpret ordinal statistic measuring the degree of overlap between two distributions that satisfies these requirements.

The definition of Cliff's δ is quite simple. Consider two randomly selected low-income students, one from the NLSY79 and one from the NLSY97. Let $q_{i,97} = F_{H,97}(s_i)$ and $q_{j,79} = F_{L,79}(s_i)$ be each student's test-score percentile relative to high-income students in her cohort. Cliff's δ is then defined by

$$\delta_{97,79} \equiv Pr(q_{i,97} \geq q_{j,79}) - Pr(q_{i,97} < q_{j,79}). \tag{1}$$

Equation 1 defines $\delta_{97,79}$ as the probability that a randomly selected low-income youth from the NLSY97 has a higher q than a randomly selected low-income youth from the NLSY79, minus the reverse probability. The subtraction is simply a normalization to ensure that Cliff's δ always lies between -1 and 1. A positive value of $\delta_{97,79}$ implies that the low-income respondent with higher achievement relative to her high-income peers is more likely to come from the NLSY97. In other words, $\delta_{97,79} > 0$ suggests that the income-achievement gap decreased.

Estimating $\delta_{97,79}$ requires two steps. First, one must estimate $q_{i,t}$ for each low-income student i in each survey t. Second, $\delta_{97,79}$ must be estimated from the estimated $\hat{q}$'s. Given consistent estimates $\{\hat{q}\}$ of the q's, a consistent estimator for $\delta_{97,79}$ is

$$\hat{\delta}_{97,79} = \frac{\sum_{i=1}^{N_{97,L}} \sum_{j=1}^{N_{79,L}} \left[\mathbb{I}\left(\hat{q}_i \geq \hat{q}_j\right) - \mathbb{I}\left(\hat{q}_i < \hat{q}_j\right)\right]}{N_{97,L} N_{79,L}}. \tag{2}$$

This estimator does not depend on the scale of the $\hat{q}$'s; it will be unaffected if the test scores in the two surveys are subjected to distinct, arbitrary rescalings.

I rely exclusively on bootstrapped confidence intervals to conduct inference on $\hat{\delta}_{97,79}$. Asymptotic formulas for the variance of $\hat{\delta}$ are available, but they do not account for the fact that both the $\hat{q}$'s and the high- and low-income thresholds are estimated from the data. Adjusting for this first-stage estimation is quantitatively important in this setting; the asymptotic formulas typically give standard errors that are about half as large as those obtained via the bootstrap.

Table 4 displays the baseline $\hat{\delta}$ estimates. I estimate large, positive $\hat{\delta}$'s for both reading and AFQT. Bootstrapped standard errors allow me to reject $\delta = 0$ at 5% for all comparisons and 1% for most. The point estimates for math are also positive, although they are smaller and only occasionally statistically significant at conventional levels. The race-specific $\hat{\delta}$'s show large decreases in the income-achievement gap among white youth and large increases among black youth, although the smaller sample sizes in the black-only comparisons mean that most of the estimates are not distinguishable from 0 at conventional levels.

Table 4 subdivides the sample by race before calculating the income thresholds; each comparison is between income categories defined relative to the race-specific income distribution. Since white respondents come from relatively wealthy households, the high- and low-income groups defined in this manner will be somewhat wealthier than their full-sample counterparts. Symmetrically, the high- and low-income groups in the black-only subsample will have lower incomes than their full sample counterparts. Table 5 presents estimates that set the income thresholds using the full sample before subsetting on race. The estimates for white respondents are quite similar to before. The black-only estimates now uniformly suggest a decrease in the income-achievement gap, with some of the reading and AFQT estimates attaining 5% significance. In other words, low-income black students lost ground relative to relatively high-income (and middle-income overall) black students, while gaining on the (far fewer) black students at the top of the overall income distribution.

There are many ways to slice the data prior to estimating the $\hat{\delta}$'s, and most of these methods suggest that the income-achievement gap decreased dramatically between 1980 and 1997. Table 5 also shows that using the same real dollar cutoffs in both surveys to define the high- and low-income groups does not change the story; if anything, this alternate method yields even larger estimated decreases in the income-achievement gap. Tables 6-7 show that various methods for adjusting test scores based on age yield very

similar $\hat{\delta}$ estimates, while Table 8 shows that the differences in the cross-sectional $\hat{\delta}$'s also point to a narrowing income-achievement gap.[10]

My baseline analysis treats all households equally regardless of their size and composition. Making no distinctions between households with the same total incomes but very different sizes and compositions ignores the fact that resources must be shared among household members. Therefore, I adjust for household size and composition by transforming income into equivalency units and then recomputing the δ estimates with the high- and low-income groups defined by percentiles in the transformed income distribution. In particular, I assume that the equivalency scale for household i with A_i adults and K_i children is given by $E_i = (A_i + \theta K_i)^{\gamma}$, where $\gamma \in (0, 1]$ gives the returns to scale in household production and $\theta \in [0, 1]$ gives the fraction of an adult's consumption used by a child. In my baseline specification, I follow Citro and Michael[7] and set $\gamma = \theta = 0.7$. I also estimate $\hat{\delta}$'s using $\theta = \gamma = 1$, which simply converts income into per capita units.

Table 9 displays the Cliff's $\hat{\delta}$'s calculated using these equivalency scales. The estimates are mostly quite similar to those calculated using unadjusted income; the math estimates are usually somewhat smaller and the reading and AFQT estimates a bit larger than their unadjusted counterparts. The standard errors are also quite similar to the unadjusted estimates for all three achievement tests. Thus, it does not appear that changes in household characteristics are driving the estimated convergence in high- and low-income achievement.[11]

Both test scores and household income are measured with error. Unfortunately, measurement error in either of these variables can create either positive or negative asymptotic

[10]The cross-sectional δ_k (in the population) for year k is defined by $\delta_k \equiv Pr(s_{i,k} > s_{j,k}) - Pr(s_{j,k} > s_{i,k})$ for $i \in H$ and $j \in L$. The cross-sectional δ for each survey can be consistently estimated by the obvious modification of equation 2. Conceptually, the downside to estimating achievement gap changes via $\hat{\delta}_{79} - \hat{\delta}_{97}$ is the implicit assumption that the cross-sectional δ's are cardinally comparable.

[11]I also test an alternative adjustment method in which I regress achievement test scores on a host of demographic variables such as race, sex, and age of parents and then use the estimated residuals as measures of "background-adjusted" achievement. Using regression-adjusted scores invariably results in smaller estimated shifts in the achievement gap between high- and low-income youth. In each case, however, the adjusted scores still show a sizable decrease in the income-achievement gap between the NLSY79 and the NLSY97, providing further evidence that household size and composition changes are not driving the main results.

bias in the $\hat{\delta}$'s. For sufficiently extreme measurement error distributions, it is even possible that the probability limit of $\hat{\delta}$ and δ will be of opposite signs.[12] Perverse outcomes like this generally require that the two surveys have very different amounts of measurement error. To see this, suppose that the relationship between household income and expected achievement is monotone increasing. Income measurement error will result in misclassifications at both ends of the income scale. These misclassifications will increase apparent achievement in the low-income group, since the misclassified youth will have higher average incomes and thus higher average achievement than their truly low-income peers. Symmetrically, the misclassifications in the high-income group will decrease the group's apparent achievement. Therefore, income measurement error will bias cross-sectional measures of achievement inequality toward 0. Now suppose that there is more actual achievement inequality and more measurement error in the NLSY97 than in the NLSY79. If the disparity in the amount of measurement error is sufficiently great, it may erroneously appear as though achievement inequality decreased between the two surveys. Similarly, measurement error in test scores will tend to bias cross-sectional measures of the income-achievement gap toward 0 but can bias gap-change estimates away from 0 if test scores in the two surveys have very different reliabilities.

I use the observed test-score and household income distributions, along with informed guesses about the reliabilities of both variables, to simulate the asymptotic bias stemming from each type of measurement error.[13] For the test scores, I extract measurement error variances by using the NLSY-reported reliabilities for each assessment. The NLSY surveys do not give reliability estimates for their income measures, so I use a range of reliabilities reported from other surveys and data sources. Table 11 has the results of these simulations, which suggest that both income and test-score measurement error lead to moderate attenuation bias. For a range of plausible reliabilities, the probability limits of the $\hat{\delta}$ estimates are 7 to 25 percent closer to 0 than the true population δ's. The only way to bias the estimates away from 0 is to assume that income in the NLSY79 is much more precisely measured than income in the NLSY97. To my knowledge, there is no good

[12]Please refer to the online appendix for a more formal presentation of these claims.
[13]Please refer to Appendix C for a detailed description of the simulation procedure.

reason to suppose that the two income measures differ so dramatically. Overall, then, I conclude that my baseline estimates are probably conservative, and, at the very least, that I have correctly identified the signs of the true δ's.

6. Anchoring Analysis

I have thus far been able to make a number of very strong claims about changes in the income-achievement gap using only ordinal methods. The ordinal analysis is limited, however, in that it cannot say whether a given test-score shift corresponds to an economically important change in achievement. Improvements in some parts of the test-score distribution may correspond to skills that have little real-world value, so that even large shifts in observed test scores may simply not be very valuable. The reading and AFQT achievement gaps narrowed unambiguously, but did they narrow by an interesting amount given a plausible set of weights? The FOSD analysis likewise shows that there exist achievement weights that would measure a larger math achievement gap in either NLSY survey. Given this ambiguity, would a realistic set of weights assess an increase or a decrease in the math gap?

This section estimates the economic importance of the convergence in achievement between high- and low-income youth by mapping crosswalked achievement test scores to various life outcomes. My basic approach uses the NLSY79 to flexibly estimate the reduced-form relationship between test scores and a particular later-life outcome. Holding this relationship constant, the empirical distributions of crosswalked test scores for low- and high-income youth in the NLSY97 can then be converted to counterfactual outcome distributions using this reduced-form relationship. These counterfactual distributions answer the following question: "If the relationship between achievement and the outcome were unchanged between the NLSY79 and the NLSY97, what would be the distribution of that outcome for the NLSY97 cohort given their observed test scores?"

I use a number of different methods to estimate the reduced-form relationship between test scores and either school completion or the present discounted value of lifetime labor income. These methods allow me to investigate different aspects of the anchored outcome distributions. The simplest anchoring approach uses regressions to approximate

the expected value of the outcome conditional on test scores. I use probit regressions to anchor on high school and college completion and polynomial regressions to anchor on labor income. I also use quantile regressions and various other numerical techniques to estimate the entire conditional distribution of lifetime income given test scores. With these distributional estimates in hand, I investigate how changes in test-score distributions correspond to changes to various percentiles of labor income.

It is important to emphasize that this approach does not allow me to make any causal claims. When I make statements like, "The improvement in achievement among low-income white men corresponds to an increase of $\$X$ of lifetime wage income," I am not arguing that the improvement in achievement caused an increase in wage income of $\$X$ for low-income white men. Rather, I am simply translating test-score shifts to income shifts using the same set of (plausible) skill prices for both surveys. Conceptually, differences in outcome inequality can come from either (or both) changes in the stocks of achievement held by high- and low-income youth and changes in the way skill is priced in the market. This paper looks only at changes in the stocks of achievement. Since the NLSY97 respondents are only around 30 years old, it is not really possible to estimate the relationship between achievement and their lifetime labor wealth with any accuracy. As time passes and more of the uncertainty in the NLSY97 respondents' lifetime outcomes is resolved, it will become possible to complete the full achievement-stock/skill-price decomposition.

6.1. **Formal Discussion.** Recall from Section 4 that the value of score distribution F_t is given by $V(F_t) = \mathbb{E}_{F_t}[W(\psi_t(s))]$, where $\psi_t(s) = a$ is the underlying achievement associated with test score s and W is some weighting function that converts achievement into cardinal units. I now modify this formalism to explicitly incorporate economic outcomes by decomposing W into two functions: $\Omega_t(a) : \mathbb{R} \to \mathbb{R}^N$, which maps achievement into an N-dimensional vector of different life outcomes, and $\mathcal{W}(\Omega_t(a)) : \mathbb{R}^N \to \mathbb{R}$, which is a standard social welfare function that takes the N outcomes from Ω_t as its argument. W is simply the composition of $\mathcal{W}$ and Ω_t, so that the value of F_t is given by $V(F_t; \Omega_t, \mathcal{W}, \psi_t) \equiv \mathbb{E}_{F_t}[\mathcal{W}(\Omega_t(\psi_t(s)))]$.

Now consider measuring changes in V between periods t and $t+1$. Using crosswalked test scores guarantees that $\psi_t = \psi_{t+1} \equiv \psi$; changes in V must come from changes in Ω (skill prices) or changes in F (skill distributions). There are two natural fixed-price comparisons that measure changes in V due to shifts in F: $\Delta(\Omega_j) = V(F_{t+1}; \Omega_j, \mathcal{W}, \psi) - V(F_t; \Omega_j, \mathcal{W}, \psi)$ for $j \in \{t, t+1\}$. In words, $\Delta(\Omega_j)$ measures the difference in value between test-score distributions F_{t+1} and F_t when scores from both distributions are translated to outcomes using either Ω_t or Ω_{t+1}. Similarly, there are two achievement-constant comparisons that quantify the value of shifts in Ω while holding the distribution of achievement fixed: $\Delta(F_j) = V(F_j; \Omega_{t+1}, \mathcal{W}, \psi) - V(F_j; \Omega_t, \mathcal{W}, \psi)$ for $j \in \{t, t+1\}$.

Although these expressions provide a convenient theoretical framework for thinking about evaluating changes in Ω and F, they are of little practical use, both because $\mathcal{W}$ is unknown and because a full list of the outcomes that plausibly enter into $\mathcal{W}$ will not be available in even the richest data sets. Given these difficulties, I pursue a much more modest objective: I ignore $\mathcal{W}$ altogether and focus on computing gap changes denominated in the units of some particular outcome y_n. Denote by $\omega_t^{(n)}(a) : \mathbb{R} \to \mathbb{R}$ the map from achievement to the nth outcome in $\Omega_t(a)$. The y_n-denominated value of distribution F is then given by $v\left(F; \omega^{(n)}, \psi\right) \equiv \mathbb{E}_F[\omega^{(n)}(\psi(s))]$. I define four gap changes denominated in the same units as y_n:

$$\Delta(\omega_j^{(n)}) \equiv v\left(F_{t+1}; \omega_j^{(n)}, \psi\right) - v\left(F_t; \omega_j^{(n)}, \psi\right), \ j \in \{79, 97\} \tag{3}$$

$$\Delta(F_j, n) \equiv v\left(F_j; \omega_{t+1}^{(n)}, \psi\right) - v\left(F_j; \omega_t^{(n)}, \psi\right), \ j \in \{79, 97\} \tag{4}$$

Equation (3) defines two fixed price gaps in which the change in test scores from F_t to F_{t+1} is valued using either $\omega_t^{(n)}$ or $\omega_{t+1}^{(n)}$. Analogously, equation (4) defines two fixed distribution gaps in which the change in the y_n-mapping from $\omega_t^{(n)}$ to $\omega_{t+1}^{(n)}$ is valued using either F_t or F_{t+1}. Once estimated, the changes defined in equations (3) and (4) can be combined to form anchored gap-change estimates for high- and low-income students. In practice, $\omega_{97}^{(n)}$ will be difficult to estimate because the NLSY97 respondents are not currently old enough to accurately measure differences in many economic outcomes. Therefore, I only report anchored gap changes using estimates of $\omega_{79}^{(n)}$.

6.2. **Regression-Based Labor Income Anchoring.** This section presents anchored gap/change estimates when the anchoring relationship is $\mathbb{E}_{79}[\text{pdv\_labor}|s]$, approximated using regressions of the form

$$\log(\text{pdv\_labor}) = \underbrace{\phi(s)}_{\text{polynomial}} + \gamma(\text{race/sex/age/income quintile dummies}) + \varepsilon. \tag{5}$$

I first estimate equation 5 on the NLSY79 respondents and then use the estimated coefficients to predict $\log(\widehat{\text{pdv\_labor}})$ for the NLSY97 respondents. These predicted log labor incomes can then be used to construct various anchored gap/change estimates. The main methodological subtlety lies in the treatment of test-score measurement error. I set $\phi(s) = \alpha + \beta s$ in my baseline specification and adjust $\hat{\beta}$ by the inverse of the estimated test reliability prior to calculating $\log(\widehat{\text{pdv\_labor}})$. Appendix D discusses the method in more detail.

Table 12 presents the regression-anchored mean gap-change estimates with bootstrapped standard errors for various demographic groups and achievement tests. These estimates generally suggest that the achievement shifts documented in Section 5 correspond to large and statistically-significant decreases in adult earnings inequality for both black and white men. For example, the narrowing of the math gap corresponds to a decrease in the adult earnings gap of between \$38,000 and \$63,000 in present-value dollars for white men, depending on the imputation method used. Similarly, the narrowing of the reading gap translates to a narrowing of the present-value lifetime earnings gap of between \$72,000 and \$132,000 for white men. For black men, the narrowing of the math achievement gap translates to small (≈\$15,000) and statistically insignificant decreases in labor earnings inequality, while for reading the corresponding estimates range from \$115,000 to \$208,000. These reading estimates are significant at 5% despite the comparatively small number of black men used in the analysis.

The results for women present a more muddled picture. For white women, the anchored gap changes for math hint at a widening but are statistically indistinguishable from 0 at conventional levels. The reading gap changes are generally significant at 5% and translate to a narrowing of outcome inequality of \$32,000 to \$42,000. The situation is reversed for

black women; the reading estimates suggest a widening of the outcome gap but are not distinguishable from 0, while for math the estimates again suggest a marginally significant widening.

I check the robustness of the regression methodology in a number of ways. First, I re-estimate a version of equation 5 using only white men and obtain substantially more negative gap-change estimates than those reported in table 12. Estimates that do not adjust for test-score measurement error as well as cubic estimates that set $\phi(s) = \alpha + \beta_1 s + \beta_2 s^2 + \beta_3 s^3$ produce qualitatively similar gap-change estimates to the baseline, although the magnitudes of the point estimates tend to be smaller using these alternate methodologies. Ignoring measurement error results in point estimates that are between 20 and 40 percent closer to 0 than those reported in table 12, while setting ϕ to be a cubic usually results in slightly smaller point estimates, although the differences with baseline are modest.[14]

6.3. **Distributional Labor Income Anchoring.** Regression-based estimates are easy to compute and easy to interpret, but they are also limited in that they cannot be used to study heterogeneity in the anchored effects. I now present an alternative anchoring methodology that allows me to estimate anchored gap changes at different points in both the test-score and anchored outcome distributions.

I first briefly outline my method; please refer to Appendix D for a more detailed, technical description of the approach. I stitch together and smooth a large number of quantile regressions in order to estimate $\hat{K}_{79}(y|s)$, the conditional distribution of outcome y given test scores s, on a grid of test scores spanning the range of observed scores. I then compute from these estimated conditional distributions the estimated conditional quantiles of y given s, which I use as my primary anchoring relationships.

Although this method can be used to compute mean gap-change estimates, its real value lies in its ability to estimate distributional effects.[15] Figures 4-5 plot the anchored gap changes and bootstrapped confidence intervals for various projected income percentiles for

[14]Tables with these alternative gap-change estimates can be found in the online appendix.

[15]Mean gap changes estimated using this method are similar to, but noisier than, regression-based mean gap-change estimates such as those reported in Section 6.2.

white male youth at a fixed test-score percentile in the high- and low-income distributions. That is, if $\hat{s}_{G,t}^{(p)}$ denotes the estimated pth percentile test score for group $G \in \{H, L\}$ students in year t, and $\hat{Y}_{G,t}^{(\tau,p)}$ denotes the estimated τth percentile of y given $\hat{s}_{G,t}^{(p)}$, these figures plot $(\hat{Y}_{H,97}^{(\tau,p)} - \hat{Y}_{L,97}^{(\tau,p)}) - (\hat{Y}_{H,79}^{(\tau,p)} - \hat{Y}_{L,79}^{(\tau,p)})$ for a fixed p and many different values of τ. I show estimates only for white males because this group had substantially higher labor force participation throughout their prime years than the other demographic groups in the NLSY data and so their anchored gap changes are more precisely estimated and depend less on the imputation method used.

Figures 4-5 show that the test-score shifts documented in section 5 correspond to economically large, statistically significant decreases in the lifetime earnings gap between high- and low-income white males for a wide range of test scores and projected incomes. For instance, the gap in expected median earnings for students at the medians of the high- and low-income distributions narrowed by \$50,000 to \$100,000, depending on the imputation method used and the assumptions made on labor supply. Comparing relatively high performing students (75th percentile) and relatively low performing students (25th percentile) yields qualitatively similar estimates, although the math gap changes for low performing students are quite a bit smaller. That the reading gap changes are consistently larger than the math gap changes for these low performing students is consistent with the declines in absolute math achievement for both high- and low-income students at the bottom of the achievement distribution. A final interesting pattern is that the math estimates are consistently below the reading estimates for both high and low performing students, while they are consistently above for median students. The baseline estimates depicted in these figures adjust for test-score measurement error. Ignoring measurement error entirely results in anchored gap changes that are 20-40% smaller, similar to what I found in the regression-based mean gap-change estimates.[16] Even if the observed test scores are free of measurement error, the observed distributional shifts correspond to large shifts in lifetime income.

[16]Please refer to the online appendix for the analogues of figures 4-5 estimated with no measurement error adjustments. Additionally, the online appendix shows that using cubic, rather than linear, quantile regressions produces qualitatively similar distributional gap-change estimates for white males.

Figures 6-9 break the distributional gap-change estimates out into separate changes over time for high- and low-income youth. The overall anchored gap changes are just the anchored changes for low-income youth minus the anchored changes for high-income youth. These figures therefore give some sense of whether it is declines among high-income youth or improvements among low-income youth (or both) that are driving the significantly positive overall estimates.

Figures 7 and 9 uniformly suggest that the positive overall reading gap changes are coming from both declines in anchored achievement for high-income youth and improvements in anchored achievement for low-income youth. Although the test-score changes for both income groups are contributing to the overall gap-change estimates, the magnitudes of the high-income decreases are typically much larger than the magnitudes of the low-income increases. In other words, holding skill prices fixed, the improvement in reading achievement among low-income youth was modestly valuable, while the decline in reading achievement among high-income youth was quite costly.

Figures 6 and 8 paint a more nuanced picture for math achievement. For students at the 50th and 75th percentiles of the high- and low-income math distributions, the story is the same as for reading: the overall gap-change estimates are driven by modest improvements for low-income youth and larger declines for high-income youth. However, relatively low-performing (25th percentile) high- and low-income youth both suffered declines in anchored math achievement. Since the declines for low-income youth are smaller in magnitude than the declines for high-income youth, the overall gap-change estimates remain positive.

The declines in anchored math achievement among low-performing youth mirror the ordinal declines documented in figure 3. In discussing the ordinal results, I argued that a score weighting scheme that placed a lot of emphasis on the bottom of the score distribution might measure an increase in the math income-achievement gap. Figures 6 and 8 demonstrate that using expected future wage income as test-score weights does not place sufficient emphasis on the bottom of the math achievement distribution to generate a negative gap-change estimate.

6.4. **School Completion.** In this section, I anchor test scores to high school and college completion rates. My approach uses a straightforward modification of equation 5 in which probit regression takes the place of polynomial least-squares regression. Table 13 displays the gap-change estimates with bootstrapped standard errors. Using NLSY79 skill prices, I calculate that the improvements in reading and AFQT achievement between 1980 and 1997 correspond to significant (at 5%) decreases of about 0.06 in the high school graduation gap for white men, while the changes in math achievement correspond to a smaller, insignificant decrease of 0.02. The changes in each of the three achievement measures correspond to highly significant (at 1%) decreases in the college completion gap of between 0.065-0.07. The estimates are inconclusive for other demographic groups, although there is some evidence that the changes correspond to large increases in both the high school and college completion gaps among black women and large decreases among black men. Because education differences are mostly fixed by age 30, it is also feasible to estimate anchored gap changes using NLSY97 skill prices. Interestingly, these estimates, omitted for brevity, are virtually identical to the NLSY79-anchored gap changes. The similarity between the two sets of estimates suggests that the reduced-form relationship between achievement measured around age 16 and school completion has not changed very much between 1980 and 1997.

7. A Puzzle: The Parental Income-Investment Gap

The gap in investment expenditures on children between high- and low-income parents increased dramatically over the last several decades. Data on parental time use and direct monetary expenditures show that while all parents substantially increased their investments since 1970, high-education and high-income parents increased their expenditures much more rapidly. For example, Duncan and Murnane[12] calculate that the parents in the top income quintile increased their enrichment expenditures per child by 150% between 1972 and 2006, while parents in the bottom quintile increased their expenditures "only" 57%. Looking at time diaries, Ramey and Ramey[28] estimate that

college-educated mothers increased their childcare time by almost 9 hours per week in the 1990s, while less-educated mothers increased their childcare time by only 4 hours.[17]

Given my finding that the income-achievement gap decreased between 1980 and 1997, these results are quite puzzling. High-income parents dramatically increased their investments relative to low-income parents but seem to have less than nothing to show for it. One caveat here is that the time use results do not necessarily contradict my results because high-income parents only began to differentially increase their time investments around 1993; the NLSY79 and NLSY97 youth could have received similar parental time investments at least through their early teen years. In contrast, the evidence on parental expenditures does imply that the gap in enrichment spending between high and low-income households is likely much larger in the NLSY97 than in the NLSY79. If parental investments are subject to decreasing returns, it is logically possible for the investment gap to increase and for the achievement gap to simultaneously decrease. However, my results using the crosswalked test scores show that the achievement of high-income youth actually decreased in absolute terms between 1980 and 1997. This is not consistent with a decreasing-returns explanation for achievement convergence, as such an explanation implies that both groups in 1997 should outperform their like-income peers in 1980.

There are a number of explanations that could rationalize the enrichment expenditure results with my estimates of the income-achievement gap. The parental expenditure data may be misclassifying consumption spending as enrichment spending. Art camp, trips to the science museum, and similar activities may simply not be effective at improving achievement test scores.[18] Alternatively, perhaps the kind of enrichment spending high-income parents differentially engage in has payoffs along dimensions not well-measured

[17]Gautier, Smeeding, and Furstenburg[13] find evidence in Canadian time-use surveys that educated mothers increased their time spent with children more than did low-education mothers. Guryan, Hurst, and Kearney [16, 15] estimate that college-educated mothers spend 16.5 hours per week on childcare tasks, while women with only a high school degree spend 12.1 hours. This finding is particularly surprising, as low-education mothers have higher fertility. Hill and Stafford [18] and Leibowitz [24] reach similar conclusions about cross-sectional differences in time investments. Aguiar and Hurst[1] find that parental time with children increased by roughly 2.0 hours per week between 1965 and 2003. Other papers reaching similar conclusions include Bianchi[5] and Ramey and Ramey[28].

[18]There are profound econometric problems associated with estimating achievement production functions. In an interesting recent paper, Caetano et al.[14] use a novel methodology to argue that napping is the most productive use of time for young children, while active time with parents is the most important for older children.

by achievement tests. For example, colleges like to see well-rounded students with diverse lists of extracurricular activities. Spending on these activities by parents may not improve achievement test scores, but may nevertheless provide a large benefit. These explanations are speculative, and without more research, my results have uncovered a genuine puzzle.

8. Discussion

Ordinal methods using test-score data show that the gap in academic achievement between youth from high-income and low-income households decreased dramatically between 1980 and 1997. These results are robust to measurement error, composition adjustments, and various data-inclusion criteria. Using percentile-equated test scales, I find strong evidence that the ordinal shifts in reading and AFQT test scores must correspond to unambiguous decreases in the underlying achievement gaps between high- and low-income youth. The ordinal shifts in math achievement do not necessarily correspond to a decrease in the underlying achievement gap, although low-income students above the 45th percentile of the low-income math-achievement distribution unambiguously gained. Anchoring reading and AFQT test scores on various later-life outcomes shows that these ordinal shifts correspond to economically-important shifts in achievement. For white men, the narrowing of the income-achievement gap translates to a narrowing in the lifetime wealth gap of roughly \$100,000 to \$200,000 and a narrowing of the high school and college completion gaps of 0.05 to 0.08 probability units. The estimates for math are smaller and less clear-cut, but they still suggest a sizable decrease in the wealth gap between 1980 and 1997.

My results should give pause to economists and policymakers who analyze achievement inequality using test-score data. The typical methods used to quantify differences in academic achievement between groups assume that test scores are cardinally comparable. This assumption is not well justified, and cardinal methods are often quite sensitive to order-preserving transformations of the test-score data. Cardinal methods can lead to conclusions about changes in achievement inequality that are not supported by the ordinal content of the test scores.

Given recent findings on changes in parental investments in children by income class, my finding that the income-achievement gap has narrowed is puzzling. High-income parents have increased their enrichment spending on their children much more rapidly than low-income parents have over the last three decades, yet my estimates imply that the distribution of high-income reading achievement shifted down while the low-income reading distribution shifted up. Even for math achievement, where the ordinal analysis leads to less clear-cut conclusions, I find no evidence that the achievement distribution for high-income youth shifted up between 1980 and 1997. Testing various hypotheses that could resolve this puzzle is a worthwhile avenue for future research.

Holding skill prices fixed, the anchoring estimates imply that the convergence in achievement between high- and low-income should have been a powerful force reducing adult outcome inequality. This does not imply, however, that inequality in outcomes between youth from high- and low-income households will be lower in the NLSY97 than in the NLSY79. If the returns to achievement become more convex over time, for example, smaller true achievement differences may well translate to larger absolute outcome differences than in the past. Unfortunately, the young age of the NLSY97 respondents precludes directly examining their lifetime labor market outcomes.

References

[1] Mark Aguiar and Erik Hurst. Measuring Trends in Leisure: The Allocation of Time Over Five Decades. *Quarterly Journal of Economics*, 122:969–1006, 2007.

[2] Joseph Altonji, Prashant Bharadwaj, and Fabian Lange. Changes in the Characteristics of American Youth: Implications for Adult Outcomes. *Journal of Labor Economics*, 2011.

[3] Anthony B. Atkinson. On the Measurement of Inequality. *Journal of Economic Theory*, 2:244–263, 1970.

[4] Garry Barret and Stephen Donald. Consistent Tests for Stochastic Dominance. *Econometrica*, 71:71–104, 2003.

[5] Suzanne M. Bianchi. Maternal Employment and Time with Children: Dramatic Change of Surprising Continuity? *Demography*, 37:401–414, 2000.

[6] Timothy Bond and Kevin Lang. The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results. *Review of Economics and Statistics*, forthcoming, 2013.

[7] Constance F. Citro and Robert T. Michael. Measuring Poverty: A New Approach. Technical report, The United States Census Bureau, 1995.

[8] Charles Clotfelter, Helen Ladd, and Jacob Vigdor. The Academic Achievement Gap in Grades 3-8. *The Review of Economics and Statistics*, 91:398–419, 2009.

[9] Flavio Cunha and James J. Heckman. Formulating, Identifying, and Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Journal of Human Resources*, 43:738–782, 2008.

[10] Flavio Cunha, James J. Heckman, and Susan Schennach. Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica*, 78:883–931, 2010.

[11] Greg Duncan and Katherine Magnuson. The Role of Family Socioeconomic Resources in the Black-White Test Score Gap Among Young Children. *Developmental Review*, 87:365–399, 2006.
[12] Greg Duncan and Richard Murnane. *Figure 1.6: Enrichment Expenditures on Children, 1972-2006*, chapter 1, page 11. Russell Sage, 2011.
[13] Anne Gauthier, Timothy Smeedeng, and Frank Furstenberg Jr. Are Parents Investing Less Time in Children? Trends in Selected Industrialized Countries. *Population and Development Review*, 30:647–671, 2004.
[14] Josh Kinsler Gregorio Caetano and Hao Teng. Toward Consistent Estimates of Children's Time Allocation on Skill Development. *Working Paper*, 2015.
[15] Jonathan Guryan, Erik Hurst, and Melissa Kearney. Parental Education and Parental Time Spent with Children. NBER Working Paper, 2008.
[16] Jonathan Guryan, Erik Hurst, and Melissa Kearney. Parental Education and Parental Time with Children. *Journal of Economic Perspectives*, 22:23–46, 2008.
[17] Eric Hanushk and Steven Rivkin. School Quality and the Black-White Achievement Gap. *NBER*, 12651, 2006.
[18] Russell Hill and Frank Stafford. Allocation of Time to Preschool Children and Educational Opportunity. *The Journal of Human Resources*, 9:323–341, 1974.
[19] Caroline Hoxby. The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *Quarterly Journal of Economics*, 115(4):1239–1285, 2000.
[20] Brian Jacob and Jesse Rothstein. The Measurement of Student Ability in Modern Assessment Systems. *Working Paper*, 2016.
[21] Roland G. Fryer Jr. and Steven D. Levitt. Understanding the Black-White Test Score Gap in the First Two Years of School. *The Review of Economics and Statistics*, 86(2):447–464, 2004.
[22] Roland G. Fryer Jr. and Steven D. Levitt. The Black-White Test Score Gap Through Third Grade. *American Law and Economics Review*, 8:249–81, 2006.
[23] Alan Krueger. Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics*, 115(2):497–532, 1999.
[24] Arleen Leibowitz. Education and the Allocation of Women's Time. In *Education, Income, and Human Behavior*. NBER, 1975.
[25] Frederic Lord. The 'Ability' Scale in Item Characteristics Curve Theory. *Psychometrika*, 40:205–217, 1975.
[26] Derek Neal. Why Has Black-White Skill Convergence Stopped? *Handbook of Economics of Education*, 1, 2006.
[27] Eric Nielsen. Achievement Gap Estimates and Deviations From Cardinal Comparability. *Finance and Economics Discussion Series, Board of Governor's of the Federal Reserve System*, 2015.
[28] Gary Ramey and Valerie Ramey. The Rug Rat Race. Working Paper 2010.
[29] Stephen Raudenbush. What Are Value-Added Model Estimating and What Does This Imply for Statistical Practice? *Journal of Educational and Behavioral Statistics*, 29 (1):121–129, 2004.
[30] Sean Reardon. Thirteen Ways of Looking at the Black-White Test Score Gap. CEPA Working Paper, Stanford University.
[31] Sean Reardon. The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations. *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances*, July 2011.
[32] Carsten Schroeder and Shlomo Yitzhaki. Revisiting the Evidence for a Cardinal Treatment of Ordinal Variables. *Working Paper*, 2015.
[33] D. Segall. Equating the CAT-ASVAB. In *Computerized Adaptive Testing: From Enquiry to Operation*. American Psychological Association, 1997.
[34] D. Segall. Chapter 18: Equating the CAT-ASVAB with the P&P-ASVAB. (from) CATBOOK, Computerized Adaptive Testing: From Enquiry to Operation. Technical report, United States Army Research Institute for the Behavioral and Social Sciences, 1999.
[35] S.Stevens. On the Theory of Scales of Measurement. *Science*, 103:677–680, 1946.

Appendix A. Tables and Figures

Table 1. Summary Statistics

Variable	Survey	N	Mean	S.D.	Min	Max
male	NLSY79	3,820	0.49	0.25	0	1
	NLSY97	2,824	0.50	0.25	0	1
white	NLSY79	3,820	0.79	0.58	0	1
	NLSY97	2,824	0.73	0.44	0	1
black	NLSY79	3,820	0.14	0.35	0	1
	NLSY97	2,824	0.14	0.35	0	1
income	NLSY79	3,820	\$43,853	\$26,728	0	\$165,753
	NLSY97	2,824	\$55,827	\$48,551	0	\$417,074
age	NLSY79	3,820	16.08	0.79	15	17
	NLSY97	2,824	15.73	0.70	15	17
math	NLSY79	3,820	96.89	18.34	55	134
	NLSY97	2,824	98.74	19.15	56	134
read	NLSY79	3,820	94.14	19.42	40	123
	NLSY97	2,824	93.62	19.96	40	123
afqt	NLSY79	3,820	142.58	27.12	67.5	190
	NLSY97	2,824	142.99	28.15	68.5	190
high school	NLSY79	3,820	0.88	0.32	0	1
college	NLSY79	3,820	0.23	0.42	0	1
pdv_pess_fixed	NLSY79	3,794	\$371,562	\$267,756	\$0	\$1,209,416
pdv_opt_fixed	NLSY79	3,795	\$672,697	\$436,490	\$0	\$2,442,542
pdv_pess_flex	NLSY79	3,795	\$451,766	\$251,922	\$0	\$1,209,416
pdv_opt_flex	NLSY79	3,772	\$1,017,437	\$739,481	\$0	\$7,060,312

Note: Respondent ages are restricted to 15-17 as of ASVAB test date. All dollars have been converted to a 1997 basis using the CPI-U. Sample statistics use each survey's respective base-year sampling weights.

TABLE 2. FOSD Probabilities of Crosswalked Score Distributions

Subject/Year	High/Low Defn.	Low	High	Low, White	High, White	Low, Black	High, Black
math 79	baseline	0.00	1.00	0.08	0.98	0.71	0.01
math 97	baseline	0.00	0.00	0.08	0.01	0.00	0.29
reading 79	baseline	0.01	0.88	0.07	0.82	0.49	0.05
reading 97	baseline	0.37	0.01	0.83	0.08	0.05	0.85
AFQT 79	baseline	0.01	0.98	0.13	0.98	0.80	0.04
AFQT 97	baseline	0.04	0.00	0.68	0.01	0.00	0.80
math 79	pooled	0.00	1.00	0.13	0.99	0.71	0.01
math 97	pooled	0.00	0.00	0.10	0.00	0.00	0.18
reading 79	pooled	0.01	0.89	0.09	0.85	0.51	0.28
reading 97	pooled	0.36	0.00	0.79	0.01	0.04	0.79
AFQT 79	pooled	0.01	0.97	0.20	0.98	0.81	0.21
AFQT 97	pooled	0.04	0.00	0.53	0.00	0.00	0.88
math 79	buckets first	0.00	1.00	0.12	0.98	0.19	0.22
math 97	buckets first	0.00	0.00	0.07	0.00	0.00	0.58
reading 79	buckets first	0.01	0.88	0.05	0.85	0.56	0.70
reading 97	buckets first	0.37	0.01	0.76	0.04	0.08	0.06
AFQT 79	buckets first	0.01	0.98	0.10	0.98	0.84	0.42
AFQT 97	buckets first	0.04	0.00	0.68	0.01	0.00	0.20

Note: Each cell represents the probability that the row distribution dominates the column distribution. The column distribution is just the same achievement test from the other NLSY survey. "Pooled" means that the income cutoffs are percentiles in the income distribution (in real dollars) pooled across both NLSY surveys. "Buckets First" means that the high- and low-income buckets are defined using the full, rather than the race-specific, sample income distributions.

TABLE 3. FOSD Tests of Crosswalked Score Distributions

Subject	High/Low Defn.	Low	High	Low, White	High, White	Low, Black	High, Black
math	baseline	crossing	$F_{79} \succ F_{97}$	equal	$F_{79} \succ F_{97}$	$F_{79} \succ F_{97}$	$F_{97} \succ F_{79}$
reading	baseline	$F_{97} \succ F_{79}$	$F_{79} \succ F_{97}$	equal	$F_{79} \succ F_{97}$	$F_{79} \succ F_{97}$	$F_{97} \succ F_{79}$
AFQT	baseline	crossing	$F_{79} \succ F_{97}$	equal	$F_{79} \succ F_{97}$	$F_{79} \succ F_{97}$	$F_{97} \succ F_{79}$
math	pooled	crossing	$F_{79} \succ F_{97}$	equal	$F_{79} \succ F_{97}$	$F_{79} \succ F_{97}$	$F_{97} \succ F_{79}$
reading	pooled	$F_{97} \succ F_{79}$	$F_{79} \succ F_{97}$	equal	$F_{79} \succ F_{97}$	$F_{79} \succ F_{97}$	equal
AFQT	pooled	crossing	$F_{79} \succ F_{97}$	equal	$F_{79} \succ F_{97}$	$F_{79} \succ F_{97}$	equal
math	buckets first	crossing	$F_{79} \succ F_{97}$	equal	$F_{79} \succ F_{97}$	$F_{79} \succ F_{97}$	equal
reading	buckets first	$F_{97} \succ F_{79}$	$F_{79} \succ F_{97}$	$F_{97} \succ F_{79}$	$F_{79} \succ F_{97}$	equal	equal
AFQT	buckets first	crossing	$F_{79} \succ F_{97}$	equal	$F_{79} \succ F_{97}$	$F_{79} \succ F_{97}$	equal

Note: Each test done at 5% using probabilities drawn from Table 2. "Pooled" means that the income cutoffs are percentiles in the income distribution (in real dollars) pooled across both NLSY surveys. "Buckets First" means that the high- and low-income buckets are defined using the full, rather than the race-specific, sample income distributions.

TABLE 4. Baseline $\hat{\delta}$ Estimates

Income Percentiles	Race	Math	Reading	AFQT
[80-100] vs [0-20]	all	0.11 (-0.01, 0.25)	0.25*** (0.14, 0.36)	0.23*** (0.13, 0.35)
[80-100] vs [20-40]	all	0.17*** (0.05, 0.27)	0.19*** (0.08, 0.29)	0.2*** (0.08, 0.3)
[90-100] vs [0-10]	all	0.1 (-0.13, 0.32)	0.28*** (0.09, 0.44)	0.25*** (0.07, 0.42)
[90-100] vs [10-20]	all	0.12 (-0.05, 0.31)	0.2** (0.06, 0.37)	0.18** (0.03, 0.36)
[80-100] vs [0-20]	white	0.11 (-0.02, 0.25)	0.18*** (0.06, 0.3)	0.17*** (0.05, 0.3)
[80-100] vs [20-40]	white	0.19*** (0.07, 0.32)	0.2*** (0.08, 0.34)	0.21*** (0.1, 0.35)
[90-100] vs [0-10]	white	0.16 (-0.11, 0.35)	0.22** (0.03, 0.39)	0.21** (0.01, 0.37)
[90-100] vs [10-20]	white	0.06 (-0.18, 0.26)	0.09 (-0.13, 0.27)	0.08 (-0.14, 0.25)
[80-100] vs [0-20]	black	-0.19 (-0.41, 0.06)	-0.13 (-0.38, 0.1)	-0.16 (-0.4, 0.09)
[80-100] vs [20-40]	black	-0.02 (-0.24, 0.23)	-0.06 (-0.33, 0.16)	-0.07 (-0.31, 0.16)
[90-100] vs [0-10]	black	-0.06 (-0.34, 0.33)	0.23 (-0.16, 0.5)	0.18 (-0.23, 0.5)
[90-100] vs [10-20]	black	-0.23 (-0.52, 0.16)	-0.03 (-0.41, 0.34)	-0.06 (-0.45, 0.31)

Note: Estimates use age-standardized, crosswalked test scores. 95% confidence intervals calculated using 2,500 bootstrap iterations shown in parentheses. (***) = two-sided hypothesis test significant at 1%; (**) = significant at 5%; (*) = significant at 10%.

TABLE 5. $\hat{\delta}$ Estimates Using Alternative High-/Low-Income Definitions

Income Percentiles	Race	Type	Math	Reading	AFQT
[80-100] vs [0-20]	all	pooled	0.16** (0.03, 0.28)	0.3*** (0.19, 0.4)	0.28*** (0.17, 0.38)
[80-100] vs [20-40]	all	pooled	0.19*** (0.07, 0.3)	0.21*** (0.1, 0.31)	0.23*** (0.11, 0.32)
[90-100] vs [0-10]	all	pooled	0.21* (-0.05, 0.45)	0.42*** (0.22, 0.56)	0.36*** (0.18, 0.55)
[90-100] vs [10-20]	all	pooled	0.2** (-0.02, 0.38)	0.31*** (0.14, 0.44)	0.28*** (0.1, 0.45)
[80-100] vs [0-20]	white	pooled	0.11 (-0.02, 0.26)	0.21*** (0.1, 0.33)	0.2*** (0.08, 0.33)
[80-100] vs [20-40]	white	pooled	0.22*** (0.09, 0.34)	0.24*** (0.11, 0.36)	0.26*** (0.13, 0.38)
[90-100] vs [0-10]	white	pooled	0.18 (-0.02, 0.4)	0.27*** (0.11, 0.46)	0.26*** (0.08, 0.46)
[90-100] vs [10-20]	white	pooled	0.07 (-0.12, 0.31)	0.16* (-0.03, 0.35)	0.14 (-0.04, 0.34)
[80-100] vs [0-20]	black	pooled	-0.17 (-0.34, 0.07)	-0.14 (-0.35, 0.12)	-0.18 (-0.39, 0.11)
[80-100] vs [20-40]	black	pooled	-0.02 (-0.21, 0.24)	-0.11 (-0.33, 0.14)	-0.12 (-0.32, 0.14)
[90-100] vs [0-10]	black	pooled	0.18 (-0.16, 0.53)	0.42*** (0.17, 0.7)	0.39*** (0.13, 0.7)
[90-100] vs [10-20]	black	pooled	0.12 (-0.22, 0.4)	0.28* (-0.04, 0.57)	0.25 (-0.1, 0.54)
[80-100] vs [0-20]	all	buckets first	0.11 (-0.01, 0.25)	0.25*** (0.14, 0.36)	0.23*** (0.13, 0.35)
[80-100] vs [20-40]	all	buckets first	0.17*** (0.05, 0.27)	0.19*** (0.08, 0.29)	0.2*** (0.08, 0.3)
[90-100] vs [0-10]	all	buckets first	0.1 (-0.13, 0.32)	0.28*** (0.09, 0.44)	0.25*** (0.07, 0.42)
[90-100] vs [10-20]	all	buckets first	0.12 (-0.05, 0.31)	0.2** (0.06, 0.37)	0.18** (0.03, 0.36)
[80-100] vs [0-20]	white	buckets first	0.09 (-0.05, 0.25)	0.2*** (0.08, 0.33)	0.19*** (0.07, 0.33)
[80-100] vs [20-40]	white	buckets first	0.13** (-0.02, 0.24)	0.13** (0.01, 0.25)	0.15** (0.01, 0.27)
[90-100] vs [0-10]	white	buckets first	0.19 (-0.07, 0.41)	0.3*** (0.1, 0.48)	0.28*** (0.08, 0.48)
[90-100] vs [10-20]	white	buckets first	0.03 (-0.15, 0.29)	0.11 (-0.04, 0.35)	0.09 (-0.07, 0.33)
[80-100] vs [0-20]	black	buckets first	0.06 (-0.19, 0.42)	0.31** (0.07, 0.59)	0.31** (0.02, 0.6)
[80-100] vs [20-40]	black	buckets first	0.23 (-0.05, 0.54)	0.4*** (0.15, 0.66)	0.43*** (0.14, 0.69)
[90-100] vs [0-10]	black	buckets first	0.08 (-0.27, 0.74)	0.34 (-0.16, 0.83)	0.26 (-0.32, 0.8)
[90-100] vs [10-20]	black	buckets first	0.24 (-0.1, 0.78)	0.51** (0.01, 0.82)	0.46** (-0.05, 0.81)

Note: Estimates use age-standardized, crosswalked test scores. "Pooled" means that the income cutoffs are percentiles in the income distribution (in real dollars) pooled across both NLSY surveys. "Buckets First" means that the high- and low-income buckets are defined using the full, rather than the race-specific, sample income distributions. 95% confidence intervals calculated using 2,500 bootstrap iterations shown in parentheses. (**) = two-sided hypothesis test significant at 5%; (*) = significant at 10%.

TABLE 6. $\hat{\delta}$ Estimates Under Different Treatments of Age

Income Percentiles	Race	Age Std	Math	Reading	AFQT
[80-100] vs [0-20]	all	no age std	0.11* (-0.02, 0.24)	0.25*** (0.14, 0.36)	0.24*** (0.12, 0.35)
[80-100] vs [20-40]	all	no age std	0.17*** (0.04, 0.27)	0.19*** (0.08, 0.3)	0.19*** (0.09, 0.3)
[90-100] vs [0-10]	all	no age std	0.11 (-0.1, 0.34)	0.28*** (0.11, 0.45)	0.25*** (0.05, 0.42)
[90-100] vs [10-20]	all	no age std	0.12 (-0.05, 0.31)	0.21*** (0.07, 0.38)	0.18** (0.02, 0.36)
[80-100] vs [0-20]	all	by age $\hat{q}$	0.11* (-0.01, 0.25)	0.26*** (0.16, 0.37)	0.24*** (0.13, 0.35)
[80-100] vs [20-40]	all	by age $\hat{q}$	0.17*** (0.04, 0.27)	0.19*** (0.09, 0.3)	0.2*** (0.09, 0.31)
[90-100] vs [0-10]	all	by age $\hat{q}$	0.07 (-0.1, 0.32)	0.28*** (0.14, 0.45)	0.22** (0.07, 0.43)
[90-100] vs [10-20]	all	by age $\hat{q}$	0.1 (-0.06, 0.3)	0.2** (0.06, 0.38)	0.17** (0.03, 0.36)

Note: "No age std" means that the crosswalked test scores were not adjusted for age prior to estimating $\hat{\delta}$. "By age $\hat{q}$" means that the relative percentiles used to calculate $\hat{\delta}$ via equation 2 were estimated off of each survey-age specific test-score distribution. 95% confidence intervals calculated using 2,500 bootstrap iterations shown in parentheses. (***) = two-sided hypothesis test significant at 1%; (**) = significant at 5%; (*) = significant at 10%.

TABLE 7. $\hat{\delta}$ Estimates at Specific Ages

Income Percentiles	Race	Age	Math	Reading	AFQT
[80-100] vs [0-20]	all	15	0.08 (-0.12, 0.32)	0.17 (0.00, 0.42)	0.16 (-0.03, 0.41)
[80-100] vs [20-40]	all	15	0.07 (-0.09, 0.31)	0.21** (0.04, 0.43)	0.18* (0.00, 0.41)
[90-100] vs [0-10]	all	15	0.09 (-0.16, 0.47)	0.42*** (0.19, 0.63)	0.33*** (0.12, 0.60)
[90-100] vs [10-20]	all	15	0.05 (-0.22, 0.37)	0.20 (-0.08, 0.45)	0.13 (-0.11, 0.43)
[80-100] vs [0-20]	all	16	0.08 (-0.15, 0.32)	0.23** (0.05, 0.41)	0.21** (0.01, 0.40)
[80-100] vs [20-40]	all	16	0.12 (-0.05, 0.31)	0.1 (-0.05, 0.29)	0.11 (-0.04, 0.32)
[90-100] vs [0-10]	all	16	0.81 (-0.21, 0.56)	0.44*** (0.07, 0.66)	0.34* (-0.04, 0.65)
[90-100] vs [10-20]	all	16	0.91 (-0.18, 0.43)	0.3** (-0.05, 0.49)	0.22 (-0.10, 0.49)
[80-100] vs [0-20]	all	17	0.09 (-0.16, 0.32)	0.25** (0.03, 0.49)	0.21* (-0.02, 0.47)
[80-100] vs [20-40]	all	17	0.32*** (0.12, 0.56)	0.29*** (0.09, 0.51)	0.31*** (0.10, 0.54)
[90-100] vs [0-10]	all	17	0.00 (-0.28, 0.39)	0.24 (-0.15, 0.55)	0.16 (-0.21, 0.58)
[90-100] vs [10-20]	all	17	0.35* (-0.13, 0.62)	0.35** (-0.05, 0.61)	0.31* (-0.10, 0.62)

Note: $\hat{\delta}$ estimated using crosswalked test scores separately by age at test administration. 95% confidence intervals calculated using 2,500 bootstrap iterations shown in parentheses. (***) = two-sided hypothesis test significant at 1%; (**) = significant at 5%; (*) = significant at 10%.

TABLE 8. Estimated Differences in Cross-Sectional $\hat{\delta}$'s

Income Percentiles	Race	Math	Reading	AFQT
[80-100] vs [0-20]	all	0.09** (0.02, 0.18)	0.16*** (0.08, 0.24)	0.14*** (0.07, 0.22)
[80-100] vs [20-40]	all	0.14*** (0.03, 0.23)	0.15*** (0.04, 0.25)	0.15*** (0.04, 0.25)
[90-100] vs [0-10]	all	0.10* (-0.01, 0.2)	0.14*** (0.03, 0.24)	0.13*** (0.03, 0.22)
[90-100] vs [10-20]	all	0.09 (-0.02, 0.22)	0.12** (0.01, 0.25)	0.11* (0.01, 0.24)
[80-100] vs [0-20]	white	0.09 (-0.01, 0.21)	0.13** (0.02, 0.24)	0.12** (0.01, 0.23)
[80-100] vs [20-40]	white	0.16*** (0.05, 0.29)	0.18*** (0.07, 0.32)	0.19*** (0.08, 0.33)
[90-100] vs [0-10]	white	0.13* (-0.03, 0.27)	0.18** (0.02, 0.32)	0.17** (0.01, 0.29)
[90-100] vs [10-20]	white	0.04 (-0.13, 0.19)	0.03 (-0.15, 0.18)	0.04 (-0.14, 0.18)
[80-100] vs [0-20]	black	-0.17* (-0.35, 0)	-0.1 (-0.28, 0.06)	-0.12 (-0.29, 0.04)
[80-100] vs [20-40]	black	-0.02 (-0.19, 0.18)	-0.04 (-0.25, 0.14)	-0.04 (-0.24, 0.14)
[90-100] vs [0-10]	black	-0.07 (-0.28, 0.16)	0.1 (-0.12, 0.34)	0.07 (-0.14, 0.3)
[90-100] vs [10-20]	black	-0.23* (-0.46, 0.04)	-0.08 (-0.3, 0.17)	-0.11 (0.13, -0.33)

Note: The cross-sectional $\hat{\delta}$ is defined as the probability that a randomly selected high-income youth has a larger test score than a randomly selected low-income youth from the same cohort, minus the reverse probability. The table shows cross-sectional $\hat{\delta}$'s from the NLSY79 minus the corresponding cross-sectional $\hat{\delta}$'s from the NLSY97. Estimates use age-standardized, crosswalked test scores. 95% confidence intervals calculated using 2,500 bootstrap iterations shown in parentheses. (***) = two-sided hypothesis test significant at 1%; (**) = significant at 5%.

TABLE 9. $\hat{\delta}$ Estimates Using Various Income Equivalency Scales

Income Percentiles	Race	Method	Math	Reading	AFQT
[80-100] vs [0-20]	all	per capita	0.05 (-0.08, 0.19)	0.19*** (0.08, 0.33)	0.19*** (0.05, 0.32)
[80-100] vs [20-40]	all	per capita	0 (-0.13, 0.1)	0.07 (-0.03, 0.2)	0.06 (-0.05, 0.17)
[90-100] vs [0-10]	all	per capita	0.05 (-0.15, 0.28)	0.31*** (0.12, 0.48)	0.27*** (0.06, 0.47)
[90-100] vs [10-20]	all	per capita	0.13 (-0.06, 0.33)	0.26*** (0.09, 0.41)	0.24*** (0.07, 0.4)
[80-100] vs [0-20]	white	per capita	0.03 (-0.13, 0.14)	0.12* (-0.02, 0.23)	0.11 (-0.05, 0.22)
[80-100] vs [20-40]	white	per capita	0.01 (-0.12, 0.15)	0.09 (-0.05, 0.21)	0.06 (-0.07, 0.2)
[90-100] vs [0-10]	white	per capita	0.18* (-0.03, 0.38)	0.28*** (0.11, 0.46)	0.27*** (0.1, 0.46)
[90-100] vs [10-20]	white	per capita	-0.04 (-0.27, 0.15)	0 (-0.21, 0.17)	0 (-0.22, 0.16)
[80-100] vs [0-20]	black	per capita	-0.17 (-0.4, 0.04)	-0.06 (-0.3, 0.17)	-0.09 (-0.33, 0.15)
[80-100] vs [20-40]	black	per capita	-0.04 (-0.27, 0.15)	-0.01 (-0.25, 0.22)	-0.02 (-0.25, 0.2)
[90-100] vs [0-10]	black	per capita	-0.08 (-0.38, 0.26)	0.17 (-0.23, 0.48)	0.18 (-0.2, 0.46)
[90-100] vs [10-20]	black	per capita	-0.27 (-0.56, 0.18)	0.03 (-0.3, 0.39)	0.04 (-0.28, 0.4)
[80-100] vs [0-20]	all	composition	0.04 (-0.08, 0.19)	0.21*** (0.11, 0.36)	0.2*** (0.08, 0.33)
[80-100] vs [20-40]	all	composition	0.05 (-0.07, 0.16)	0.14** (0.04, 0.26)	0.13** (0.01, 0.23)
[90-100] vs [0-10]	all	composition	0.16 (-0.08, 0.39)	0.39*** (0.22, 0.54)	0.36*** (0.18, 0.53)
[90-100] vs [10-20]	all	composition	0.16* (-0.03, 0.35)	0.24*** (0.09, 0.4)	0.23*** (0.08, 0.4)
[80-100] vs [0-20]	white	composition	0.03 (-0.1, 0.18)	0.15** (0.04, 0.29)	0.14** (0.01, 0.28)
[80-100] vs [20-40]	white	composition	0.1 (-0.06, 0.22)	0.09 (-0.04, 0.23)	0.11 (-0.04, 0.23)
[90-100] vs [0-10]	white	composition	0.22** (0.02, 0.42)	0.29*** (0.12, 0.47)	0.29*** (0.11, 0.48)
[90-100] vs [10-20]	white	composition	0.05 (-0.15, 0.29)	0.11 (-0.07, 0.31)	0.1 (-0.09, 0.31)
[80-100] vs [0-20]	black	composition	-0.2* (-0.42, 0.03)	-0.14 (-0.36, 0.11)	-0.17 (-0.38, 0.1)
[80-100] vs [20-40]	black	composition	-0.01 (-0.24, 0.2)	0 (-0.24, 0.23)	-0.02 (-0.24, 0.22)
[90-100] vs [0-10]	black	composition	-0.03 (-0.36, 0.27)	0.21 (-0.25, 0.47)	0.21 (-0.24, 0.44)
[90-100] vs [10-20]	black	composition	-0.28 (-0.58, 0.12)	0.02 (-0.35, 0.34)	0.01 (-0.37, 0.32)

Note: The per capita estimates divide the baseline family income measure by the number of household members reported in the base year of the survey. The composition-adjusted estimates adjust family income according to the equivalency scale $E_i = (A_i + \theta K_i)^\gamma$, where A_i is the number of adults in the household in the base year of the survey, K_i is the number of children, and $\theta = \gamma = 0.7$ are parameters that modulate the assumed economies of scale (γ) and child/adult consumption requirements (θ). Estimates use age-standardized, crosswalked test scores. 95% confidence intervals calculated using 2,500 bootstrap iterations shown in parentheses. (***) = two-sided hypothesis test significant at 1%; (*) = significant at 10%.

TABLE 10. NLSY Test-Score Reliabilities

Test	Unadjusted Reliability	Adjusted Reliability
math	0.85	0.82
reading	0.81	0.75
AFQT	0.89	0.86

Note: Unadjusted reliabilities are the midpoints of the ranges reported by NLS. The adjusted reliabilities are calculated from the unadjusted reliabilities R via $R_a = \frac{var(\hat{\varepsilon})-(1-R)var(s)}{var(\hat{\varepsilon})}$, where $\hat{\varepsilon}$ are the residuals of a regression of s on race/sex dummies, age dummies, and income quintile dummies from the NLSY79.

TABLE 11. Simulated Measurement Error Bias

Achievement Measure	Type	Minimum Bias	Maximum Bias
math	test score	-16%	-16%
reading	test score	-11%	-23%
AFQT	test score	-7%	-23%
math	income	-7%	-17%
reading	income	-8%	-17%
AFQT	income	-7%	-20%

Note: The lower and upper limits are calculated by taking optimistic and pessimistic estimates for the assessment reliabilities for each NLSY assessment. The narrow range of bias in math is due to the narrow range of reported reliabilities for the AFQT math subtest.

TABLE 12. Linear Regression Mean Achievement Gap Changes Adjusted for Test-Score Reliability

Group	Category	Math	Reading	AFQT
white men	opt_fix	-$63,414 (-$155,261, $14,263)	-$132,041** (-$247,917, -$36,460)	-$95,363** (-$181,332, -$19,585)
white men	opt_flex	-$40,385 (-$94,840, $5,929)	-$70,646** (-$133,717, -$14,232)	-$57,290** (-$114,168, -$9,152)
white men	pess_fix	-$41,491* (-$88,868, $1,236)	-$83,836*** (-$146,639, -$25,395)	-$58,886** (-$106,613, -$15,470)
white men	pess_flex	-$37,763* (-$84,664, $379)	-$72,734** (-$136,263, -$19,728)	-$53,577** (-$100,962, -$11,369)
white women	opt_fix	$11,715 (-$36,657, $59,615)	-$42,361 (-$99,652, $8,640)	-$18,858 (-$63,271, $24,066)
white women	opt_flex	-$423 (-$40,633, $39,746)	-$35,455 (-$78,782, $7,369)	-$21,454 (-$59,257, $16,633)
white women	pess_fix	$2,642 (-$24,594, $29,731)	-$32,487** (-$66,344, -$2,282)	-$15,653 (-$38,897, $9,013)
white women	pess_flex	$4,780 (-$28,376, $37,965)	-$34,960* (-$74,839, $2,424)	-$17,193 (-$49,370, $13,051)
black men	opt_fix	-$19,801 (-$179,407, $127,335)	-$207,763** (-$425,644, -$16,448)	-$120,309 (-$282,414, $31,031)
black men	opt_flex	-$11,082 (-$125,198, $80,949)	-$116,147* (-$238,259, $4,042)	-$74,244 (-$178,022, $21,708)
black men	pess_fix	-$15,138 (-$108,004, $66,541)	-$128,798** (-$265,301, -$20,463)	-$73,134 (-$169,645, $12,059)
black men	pess_flex	-$11,078 (-$105,178, $67,274)	-$115,435** (-$226,468, -$12,368)	-$67,413 (-$157,443, $17,502)
black women	opt_fix	$82,357 (-$21,675, $183,948)	$50,055 (-$122,341, $194,033)	$58,387 (-$73,050, $164,011)
black women	opt_flex	$69,617 (-$26,291, $155,729)	$35,147 (-$105,722, $151,131)	$47,209 (-$68,451, $138,421)
black women	pess_fix	$41,208 (-$28,080, $107,294)	$17,553 (-$98,920, $113,743)	$25,653 (-$56,014, $98,364)
black women	pess_flex	$53,133 (-$28,209, $127,421)	$26,769 (-$93,146, $124,807)	$35,454 (-$58,898, $113,831)

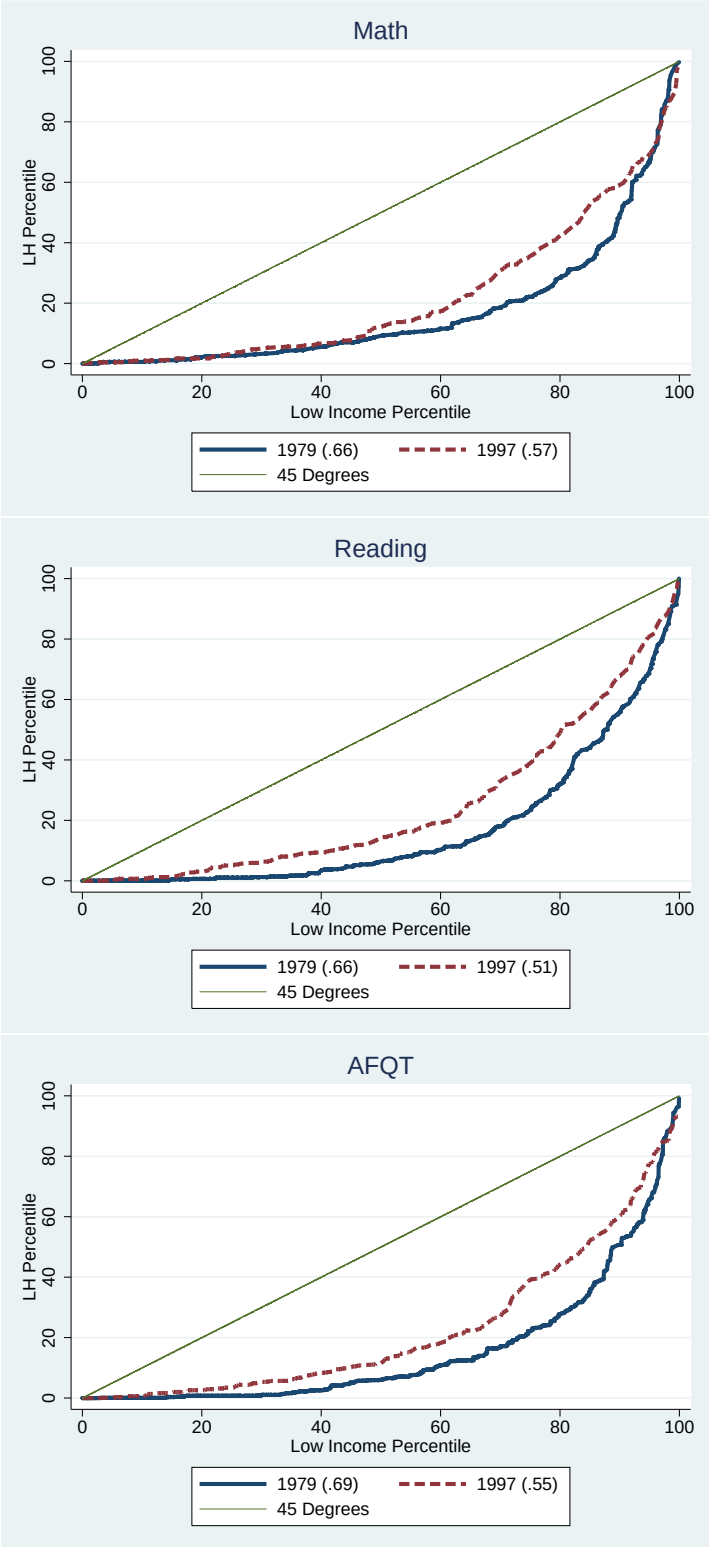
Note: Estimates use linear specifications of equation 5. The coefficients on the test scores are inflated by the appropriate inverse reliability drawn from Table 10. All dollar values deflated to 2015 basis using the CPI-U. PDV calculations use a 5% discount rate. Estimates use age-standardized, crosswalked test scores. 95% confidence intervals calculated using 2,500 bootstrap iterations shown in parentheses. (***) = two-sided hypothesis test significant at 1%; (**) = significant at 5%; (*) = significant at 10%.

TABLE 13. Probit Mean School Completion Gap-Change Estimates Using NLSY79 Skill Prices

	Group	Math	Reading	AFQT
college	white men	0.07*** (0.02, 0.12)	0.07*** (0.02, 0.12)	0.07*** (0.02, 0.12)
college	white women	0.01 (-0.05, 0.07)	0.04* (-0.01, 0.08)	0.04 (-0.01, 0.09)
college	black men	0.02 (-0.08, 0.12)	0.07 (-0.03, 0.17)	0.06 (-0.04, 0.17)
college	black women	-0.08 (-0.22, 0.05)	-0.05 (-0.14, 0.04)	-0.07 (-0.18, 0.05)
high school	white men	0.02 (-0.03, 0.08)	0.07** (0, 0.13)	0.06** (0.01, 0.12)
high school	white women	-0.05* (-0.09, 0)	0.01 (-0.05, 0.06)	0 (-0.06, 0.06)
high school	black men	-0.07 (-0.19, 0.05)	0.08* (-0.01, 0.17)	0.05 (-0.06, 0.15)
high school	black women	-0.06 (-0.18, 0.06)	-0.02 (-0.15, 0.11)	-0.01 (-0.13, 0.12)

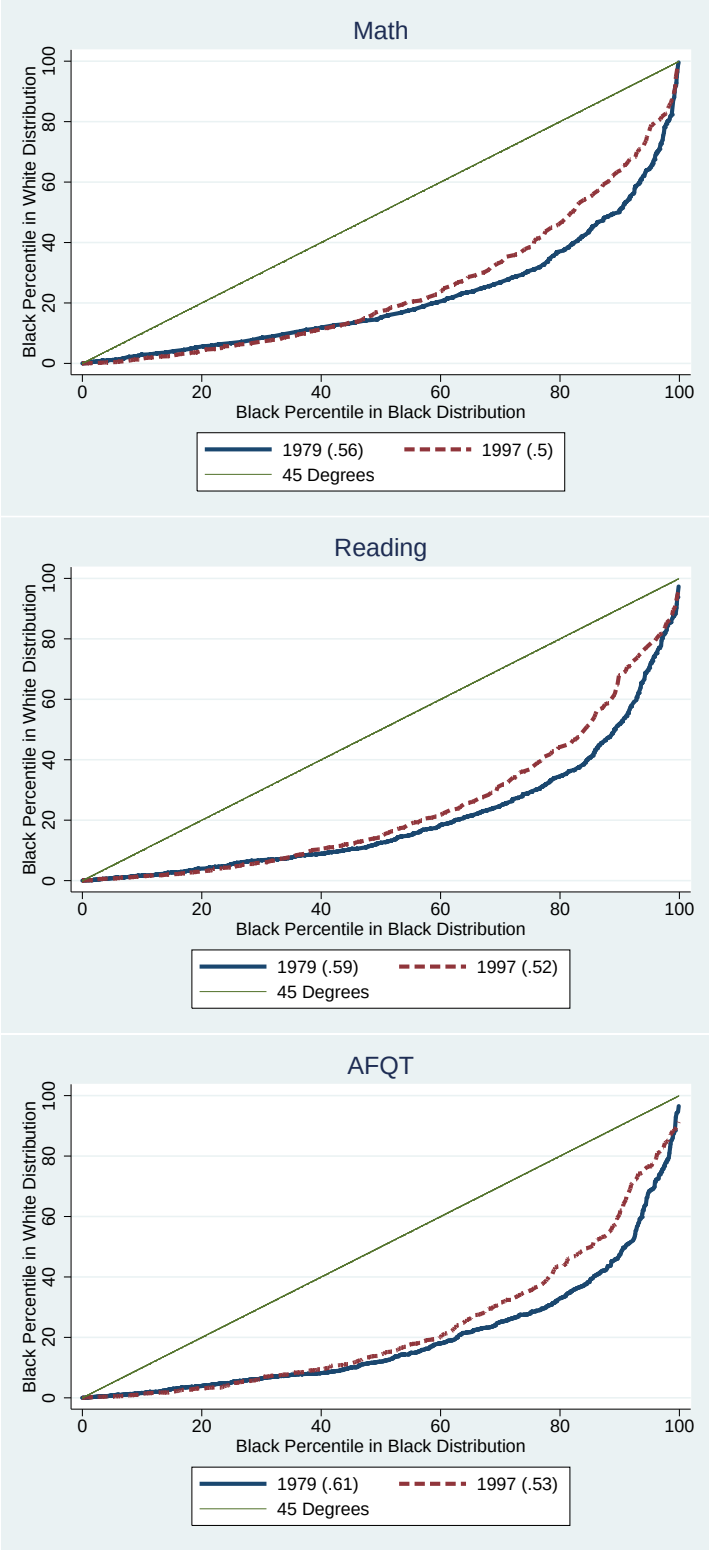
Note: Estimates based on probit models of the form $D_i = \Phi\left(g(s_i) + \gamma \text{Dummies}_i\right)$ where g is a cubic polynomial and Dummies_i is a vector of race/sex/income quintile dummies. Estimates do not adjust for measurement error and use age-standardized, crosswalked test scores. 95% confidence intervals calculated using 2,500 bootstrap iterations shown in parentheses. (***) = two-sided hypothesis test significant at 1%; (**) = significant at 5%; (*) = significant at 10%.

FIGURE 1. Top vs. Bottom Income Quintile PPCs



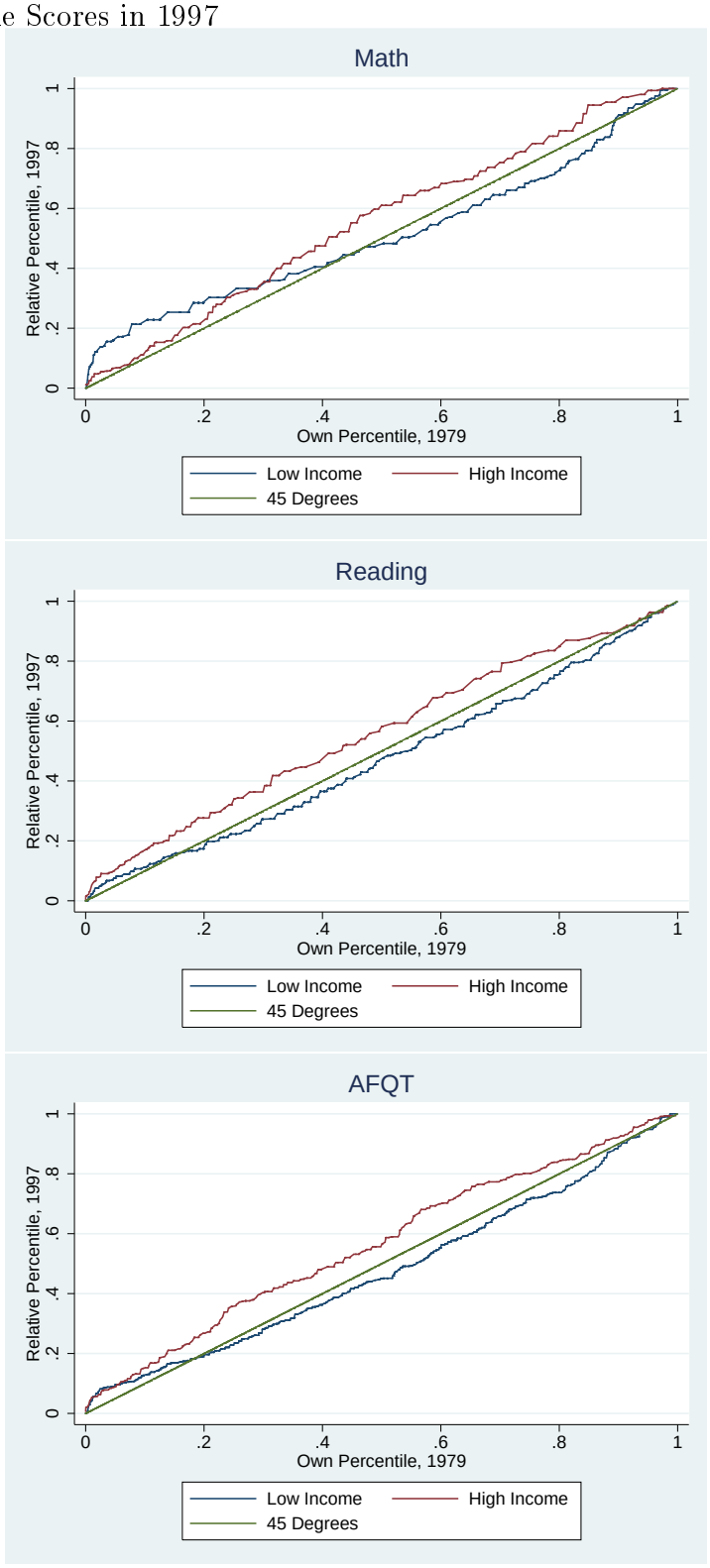
Note: Figures compare the top and bottom income quintiles. Relative percentiles are estimated using the relevant empirical cdfs. Green 45-degree line of equality plotted for reference. The pseudo-Gini coefficient is defined as $1 - 2\int_0^1 PPC(z)dz$. Pseudo-Gini Coefficients in parentheses after the line marker in the legend. Estimates use age-standardized, crosswalked test scores.

FIGURE 2. Black vs. White PPCs



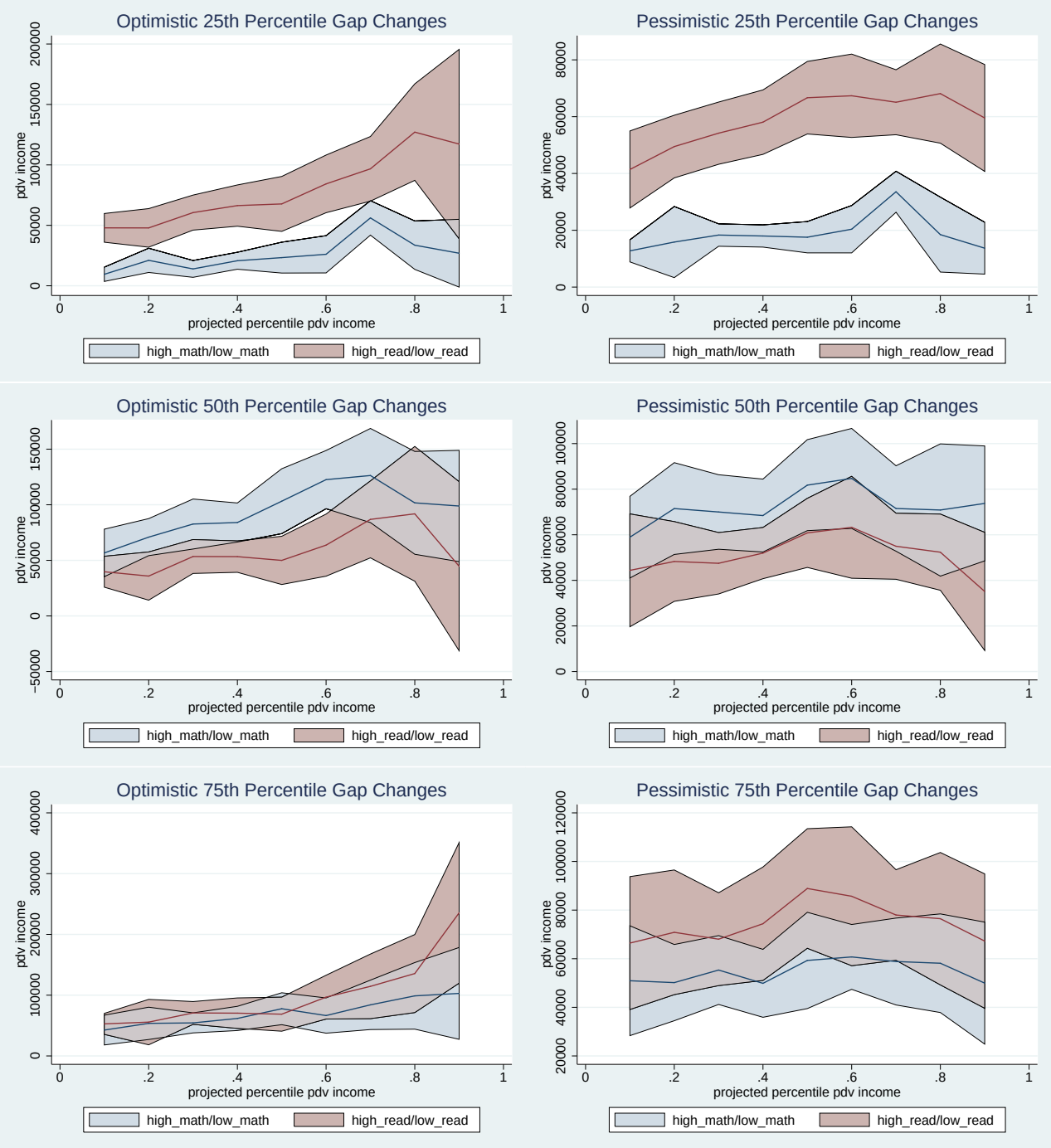
Note: Figures compare test scores from white and black respondents. Relative percentiles are estimated using the relevant empirical cdfs. Green 45-degree line of equality plotted for reference. The pseudo-Gini coefficient is defined as $1 - 2\int_0^1 PPC(z)dz$. Pseudo-Gini Coefficients in parentheses after the line marker in the legend. Estimates use age-standardized, crosswalked test scores.

FIGURE 3. High- and Low-Income Scores in 1980 Relative to High- and Low-Income Scores in 1997



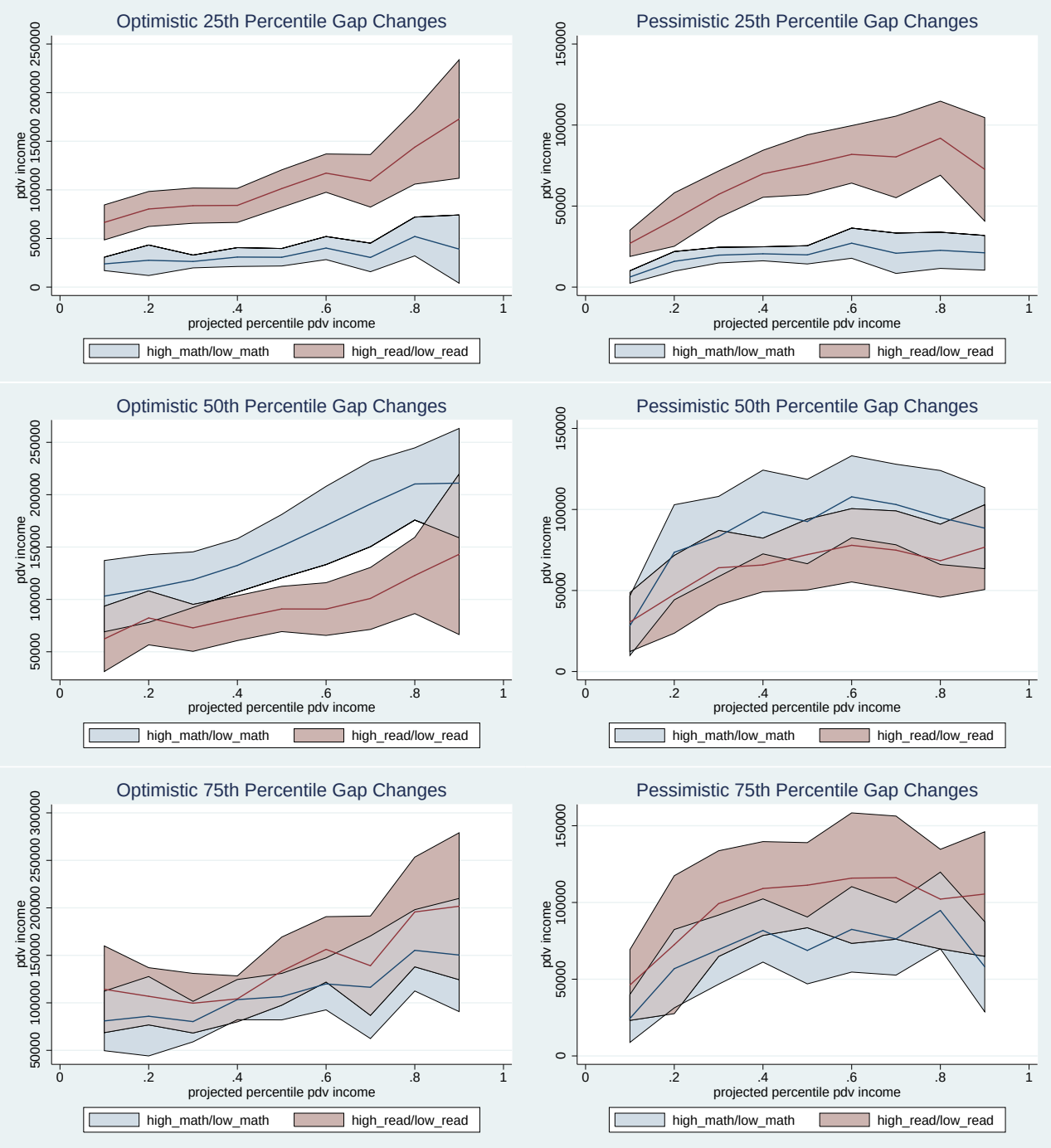
Note: Figures compare low (high) income youth in the NLSY79 to low (high) income youth in the NLSY97. Relative percentiles are estimated using the relevant empirical cdfs. Green 45-degree line of equality plotted for reference. The pseudo-Gini coefficient is defined as $1 - 2\int_0^1 PPC(z)dz$. Pseudo-Gini Coefficients in parentheses after the line marker in the legend. Estimates use age-standardized, crosswalked test scores.

FIGURE 4. White Male PDV Income Changes Assuming Flexible Labor Supply



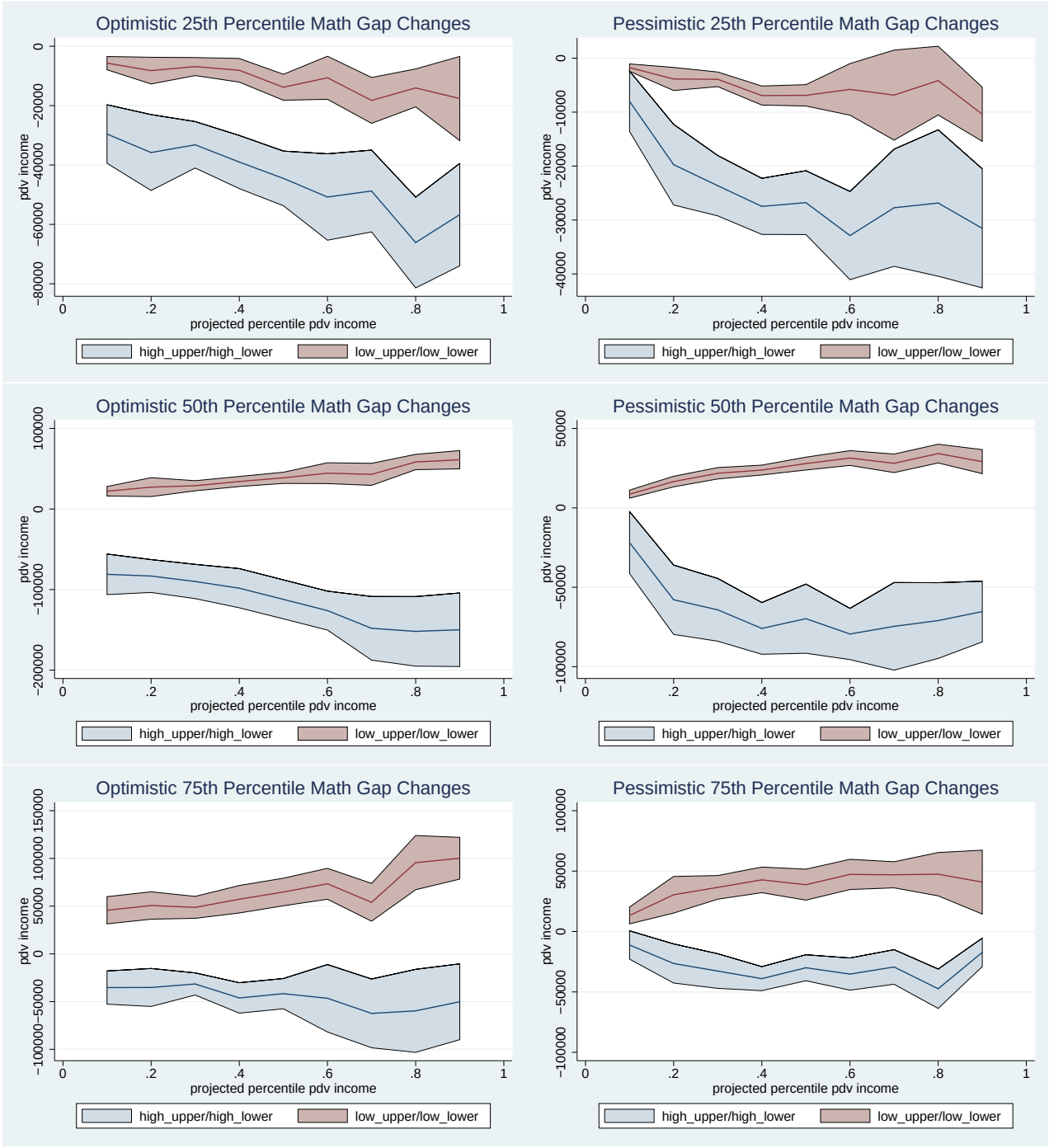
Note: Estimates based on 101 linear quantile regressions evenly spaced on $\tau \in [0, 100]$ of the form $y_i^{(\tau)} = \alpha^{(\tau)} + \beta^{(\tau)} s_i + \gamma^{(\tau)} \text{Dummies}_i$, where Dummies_i is a vector of age dummies and i is a white male youth from the NLSY79. All estimates are discrete approximations calculated on a grid of 1,500 evenly spaced points on the range of the test scores. The $\hat{\beta}^{(\tau)}$'s are inflated by the inverse of the relevant reliability from Table 10. Estimates shown are for white males who were 16 years old on the testing date. All dollar values deflated to 2015 basis using the CPI-U. PDV calculations use a 5% discount rate. Optimistic imputations assign missing incomes the maximum ever observed for each respondent, while pessimistic imputations assign the minimum. Estimates assume respondents have full control over their labor supply and therefore estimate labor income as (estimated $\text{wage}_i \times$ full time hours). Estimates use non-age-adjusted, crosswalked test scores. Figures show normal approximations to 95% confidence intervals based on 50 bootstrap iterations.

FIGURE 5. White Male PDV Income Changes Assuming Fixed Labor Supply



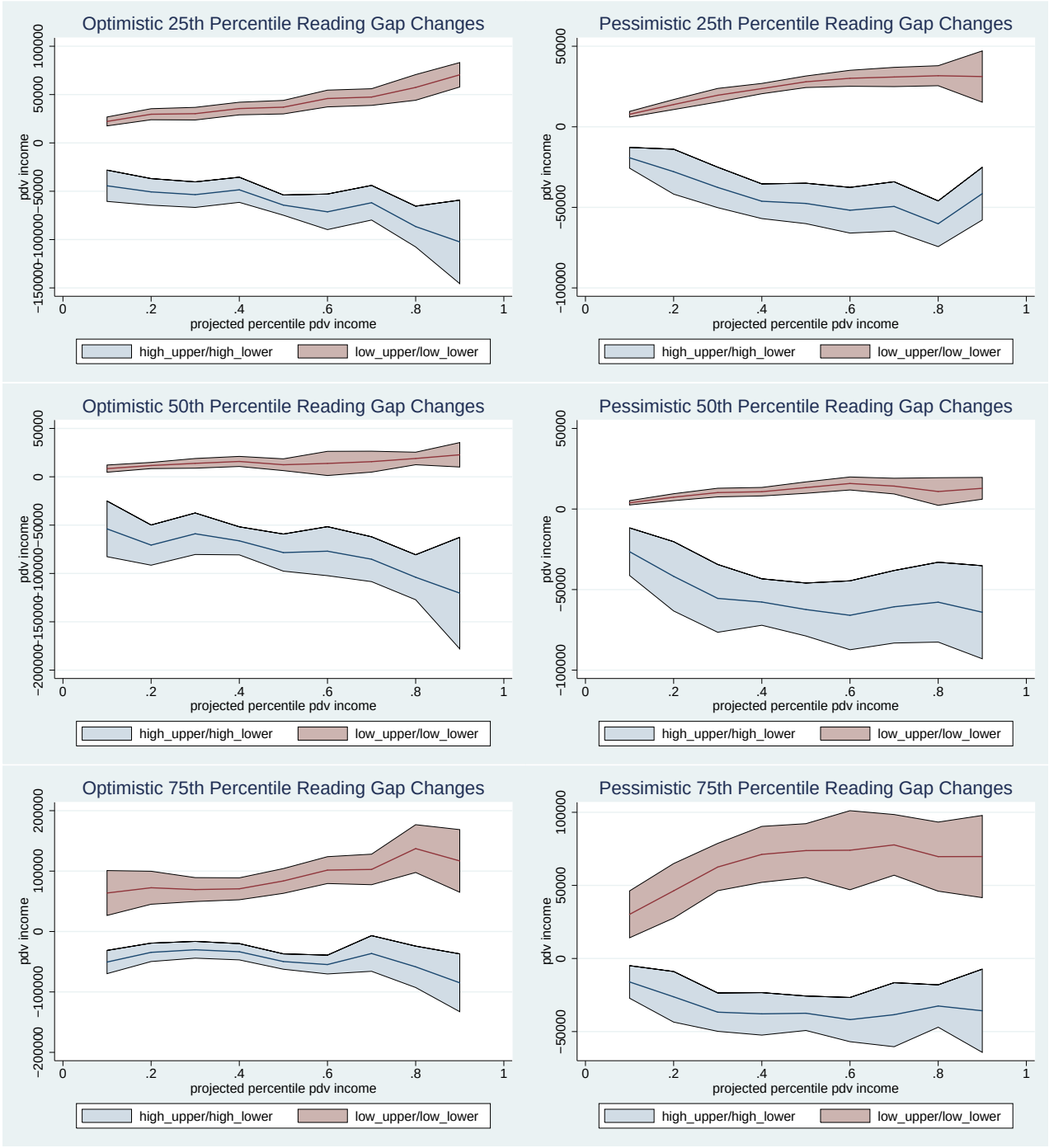
Note: Estimates based on 101 linear quantile regressions evenly spaced on $\tau \in [0, 100]$ of the form $y_i^{(\tau)} = \alpha^{(\tau)} + \beta^{(\tau)} s_i + \gamma^{(\tau)} \text{Dummies}_i$, where Dummies_i is a vector of age dummies and i is a white male youth from the NLSY79. All estimates are discrete approximations calculated on a grid of 1,500 evenly spaced points on the range of the test scores. The $\hat{\beta}^{(\tau)}$'s are inflated by the inverse of the relevant reliability from Table 10. Estimates shown are for white males who were 16 years old on the testing date. All dollar values deflated to 2015 basis using the CPI-U. PDV calculations use a 5% discount rate. Optimistic imputations assign missing incomes the maximum ever observed for each respondent, while pessimistic imputations assign the minimum. Estimates assume respondents have no control over their labor supply and therefore estimate labor income by the observed (or imputed) annual labor income. Estimates use non-age-adjusted, crosswalked test scores. Figures show normal approximations to 95% confidence intervals based on 50 bootstrap iterations.

FIGURE 6. PDV Math Changes Assuming Fixed Labor Supply for High- and Low-Income White Male Youth



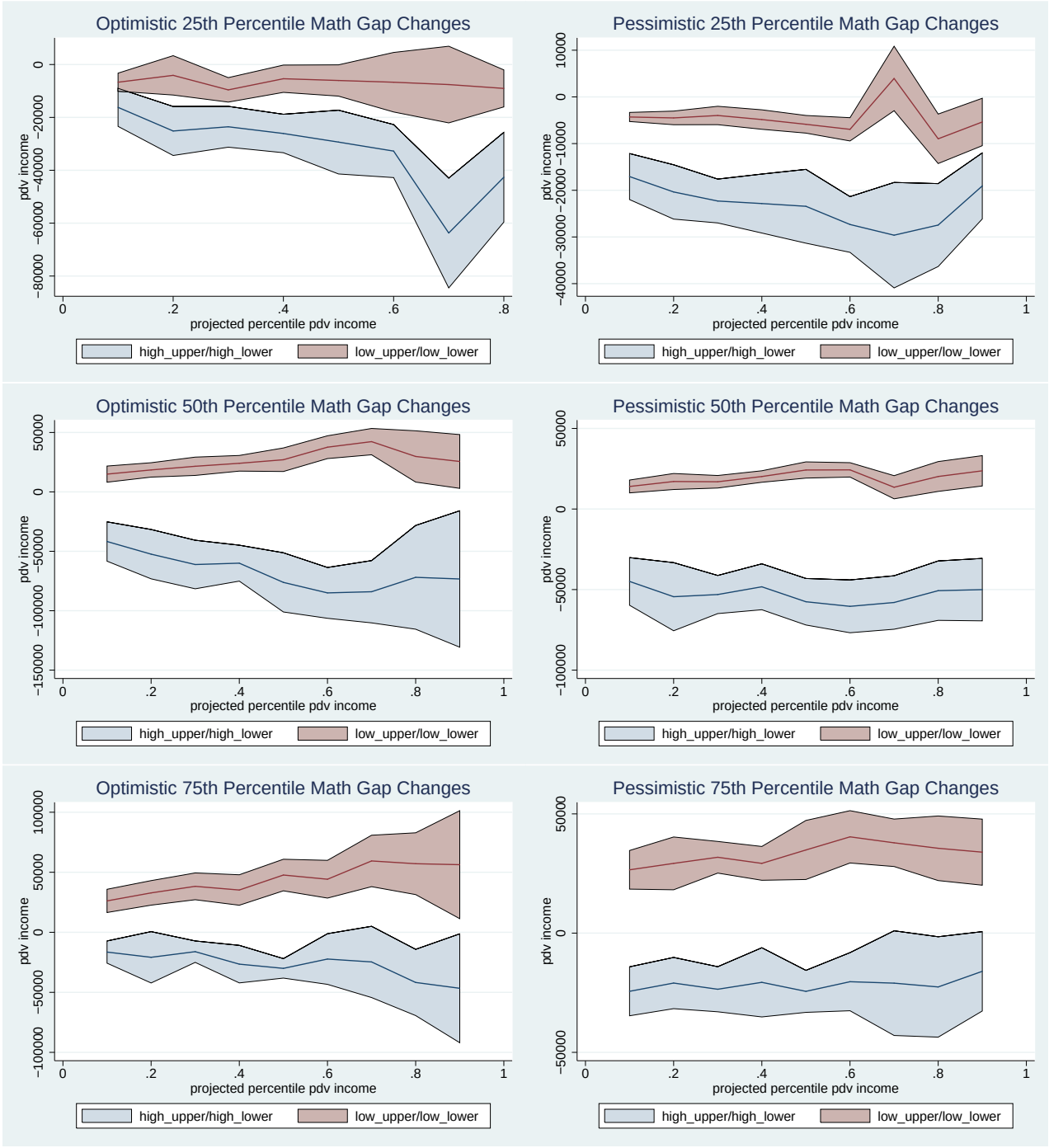
Note: Estimates based on 101 linear quantile regressions evenly spaced on $\tau \in [0, 100]$ of the form $y_i^{(\tau)} = \alpha^{(\tau)} + \beta^{(\tau)} s_i + \gamma^{(\tau)} \text{Dummies}_i$, where Dummies_i is a vector of age dummies and i is a white male youth from the NLSY79. All estimates are discrete approximations calculated on a grid of 1,500 evenly spaced points on the range of the test scores. The $\hat{\beta}^{(\tau)}$'s are inflated by the inverse of the relevant reliability from Table 10. Estimates shown are for white males who were 16 years old on the testing date. All dollar values deflated to 2015 basis using the CPI-U. PDV calculations use a 5% discount rate. Optimistic imputations assign missing incomes the maximum ever observed for each respondent, while pessimistic imputations assign the minimum. Estimates assume respondents have no control over their labor supply and therefore estimate labor income by the observed (or imputed) annual labor income. Estimates use non-age-adjusted, crosswalked test scores. Figures show normal approximations to 95% confidence intervals based on 50 bootstrap iterations.

FIGURE 7. PDV Reading Changes Assuming Fixed Labor Supply for High- and Low-Income White Male Youth



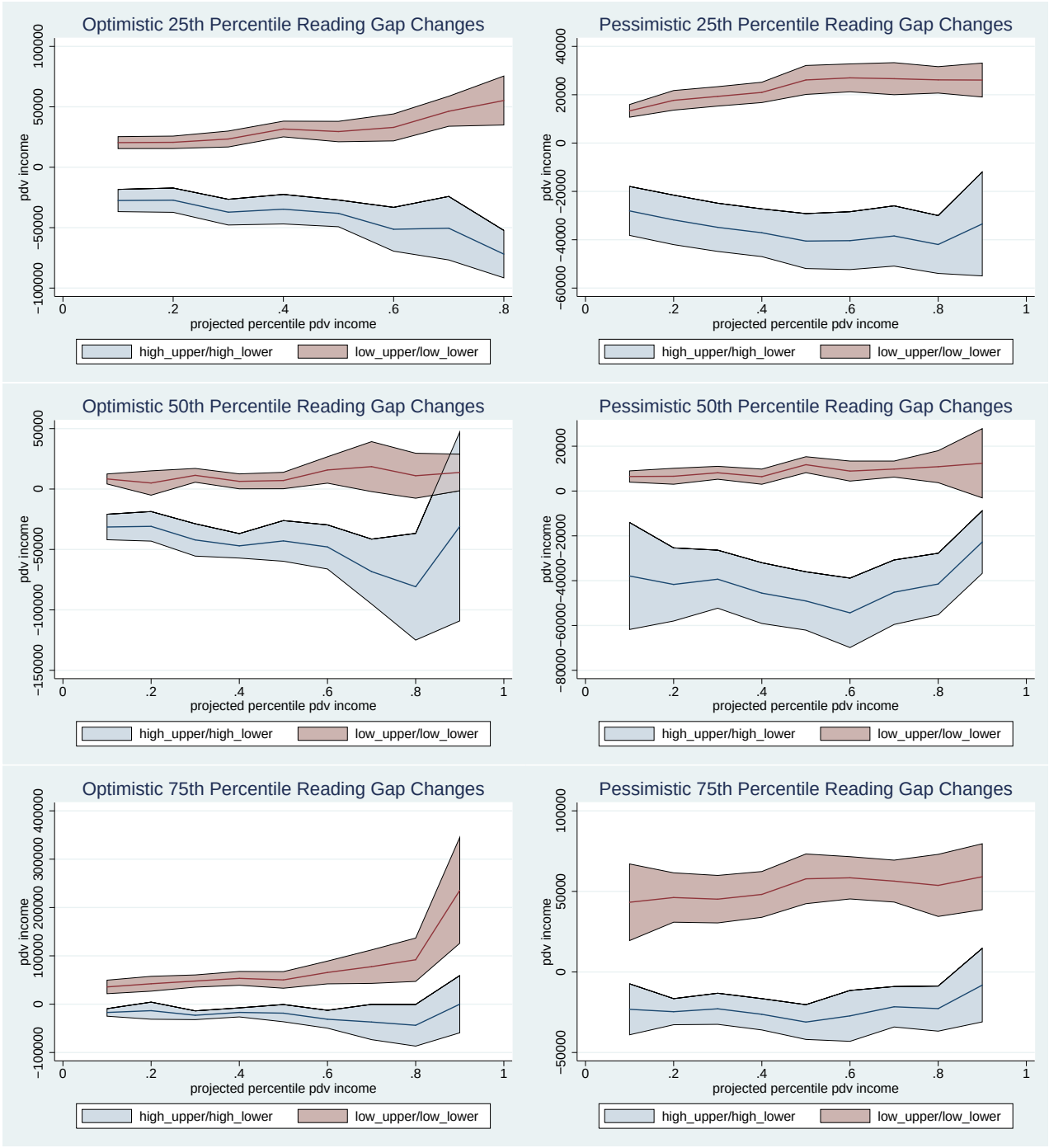
Note: Estimates based on 101 linear quantile regressions evenly spaced on $\tau \in [0, 100]$ of the form $y_i^{(\tau)} = \alpha^{(\tau)} + \beta^{(\tau)} s_i + \gamma^{(\tau)} \text{Dummies}_i$, where Dummies_i is a vector of age dummies and i is a white male youth from the NLSY79. All estimates are discrete approximations calculated on a grid of 1,500 evenly spaced points on the range of the test scores. The $\hat{\beta}^{(\tau)}$'s are inflated by the inverse of the relevant reliability from Table 10. Estimates shown are for white males who were 16 years old on the testing date. All dollar values deflated to 2015 basis using the CPI-U. PDV calculations use a 5% discount rate. Optimistic imputations assign missing incomes the maximum ever observed for each respondent, while pessimistic imputations assign the minimum. Estimates assume respondents have no control over their labor supply and therefore estimate labor income by the observed (or imputed) annual labor income. Estimates use non-age-adjusted, crosswalked test scores. Figures show normal approximations to 95% confidence intervals based on 50 bootstrap iterations.

FIGURE 8. PDV Math Changes Assuming Flexible Labor Supply for High- and Low-Income White Male Youth



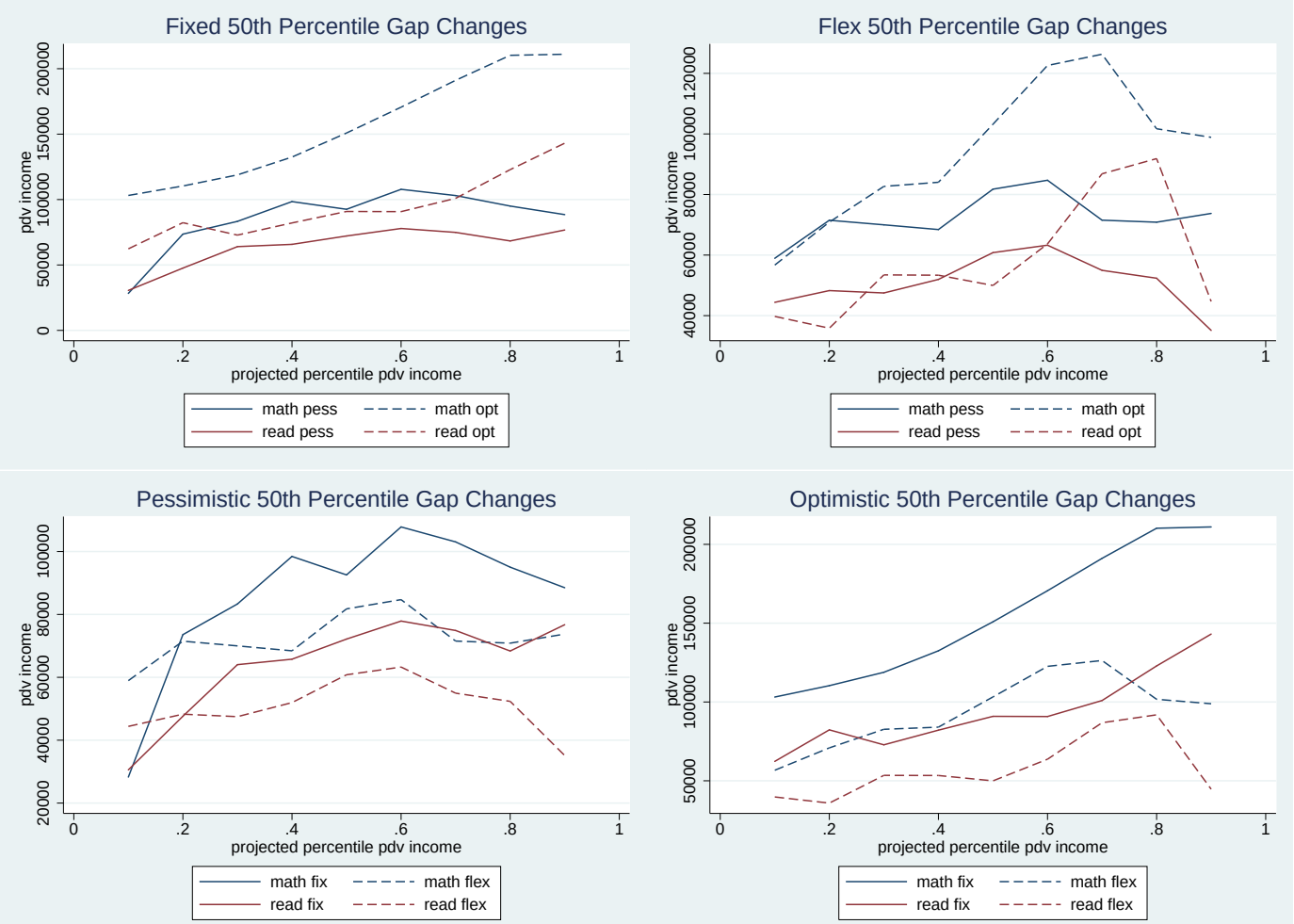
Note: Estimates based on 101 linear quantile regressions evenly spaced on $\tau \in [0, 100]$ of the form $y_i^{(\tau)} = \alpha^{(\tau)} + \beta^{(\tau)} s_i + \gamma^{(\tau)} \text{Dummies}_i$, where Dummies_i is a vector of age dummies and i is a white male youth from the NLSY79. All estimates are discrete approximations calculated on a grid of 1,500 evenly spaced points on the range of the test scores. The $\hat{\beta}^{(\tau)}$'s are inflated by the inverse of the relevant reliability from Table 10. Estimates shown are for white males who were 16 years old on the testing date. All dollar values deflated to 2015 basis using the CPI-U. PDV calculations use a 5% discount rate. Optimistic imputations assign missing incomes the maximum ever observed for each respondent, while pessimistic imputations assign the minimum. Estimates assume respondents have full control over their labor supply and therefore estimate labor income as (estimated $\text{wage}_i \times$ full time hours). Estimates use non-age-adjusted, crosswalked test scores. Figures show normal approximations to 95% confidence intervals based on 50 bootstrap iterations.

FIGURE 9. PDV Reading Changes Assuming Flexible Labor Supply for High- and Low-Income White Male Youth



Note: Estimates based on 101 linear quantile regressions evenly spaced on $\tau \in [0, 100]$ of the form $y_i^{(\tau)} = \alpha^{(\tau)} + \beta^{(\tau)} s_i + \gamma^{(\tau)} \text{Dummies}_i$, where Dummies_i is a vector of age dummies and i is a white male youth from the NLSY79. All estimates are discrete approximations calculated on a grid of 1,500 evenly spaced points on the range of the test scores. The $\hat{\beta}^{(\tau)}$'s are inflated by the inverse of the relevant reliability from Table 10. Estimates shown are for white males who were 16 years old on the testing date. All dollar values deflated to 2015 basis using the CPI-U. PDV calculations use a 5% discount rate. Optimistic imputations assign missing incomes the maximum ever observed for each respondent, while pessimistic imputations assign the minimum. Estimates assume respondents have full control over their labor supply and therefore estimate labor income as (estimated $\text{wage}_i \times$ full time hours). Estimates use non-age-adjusted, crosswalked test scores. Figures show normal approximations to 95% confidence intervals based on 50 bootstrap iterations.

FIGURE 10. Comparing Imputation Rules for White Male PDV Income Changes



Note: Estimates based on 101 linear quantile regressions evenly spaced on $\tau \in [0, 100]$ of the form $y_i^{(\tau)} = \alpha^{(\tau)} + \beta^{(\tau)} s_i + \gamma^{(\tau)} \text{Dummies}_i$, where Dummies_i is a vector of age dummies and i is a white male youth from the NLSY79. All estimates are discrete approximations calculated on a grid of 1,500 evenly spaced points on the range of the test scores. The $\hat{\beta}^{(\tau)}$'s are inflated by the inverse of the relevant reliability from Table 10. Estimates shown are for white males who were 16 years old on the testing date. All dollar values deflated to 2015 basis using the CPI-U. PDV calculations use a 5% discount rate. Optimistic imputations assign missing incomes the maximum ever observed for each respondent, while pessimistic imputations assign the minimum. Flexible labor supply estimates assume respondents have full control over their labor supply and therefore estimate labor income as (estimated $\text{wage}_i \times$ full time hours). Fixed labor supply estimates assume respondents have no control over their labor supply and therefore estimate labor income by the observed (or imputed) annual labor income. Estimates use non-age-adjusted, crosswalked test scores. Figures show normal approximations to 95% confidence intervals based on 50 bootstrap iterations.

APPENDIX B. DATA DESCRIPTION AND VARIABLE CONSTRUCTION

Calculating pdv_labor for each respondent is complicated by four forms of missing data. First, not every respondent has a valid income variable recorded in a given year. Second, not every survey respondent is in the labor force in a given year. Third, after 1994, the respondents were only interviewed every other year, so income data is missing for odd-numbered years between 1994 and 2012. Fourth, the NLSY79 respondents can only

be observed through ages 47-49, while the NLSY97 respondents can be seen only through ages 29-31. I address the first two kinds of missing data through the imputation rules described in Section 3. I address the third form of missing data by linearly interpolating wage income values for the odd-numbered years between 1995 and 2011 after applying one of my two imputation rules. I address the fourth form of missing data by using education-specific age-earnings profiles to extrapolate observed labor income profiles through to retirement.

I build the various pdv_labor estimates using NLS variables that measure total annual labor earnings and total hours worked across all jobs. I drop annual earnings above $250,000. Annual earnings are already truncated by the NLS, so this procedure removes very few observations from the sample. I do not adjust my anchored estimates for truncation, although the quantile regression-based estimates should not be sensitive to either truncation or the presence of outliers. I estimate wage rates by dividing annual earnings by annual hours worked. I perform this division after imputing missing wages using either the optimistic or pessimist rule outline in Section 3 and after filling in the alternate-year data using the procedure described above. This division results in a few unrealistically high wage estimates; I drop observations with implied wage rates above $500 (in 2015$); my estimates are not sensitive to this particular threshold choice. The NLS also provides estimated hourly rates of pay. Using these wage variables changes the anchored gap/change estimates very little. I convert hourly wages to annual earnings by assuming a full-time year of work consists of 2,087 hours, which is the Office of Personnel Management assumption for full-time work.

The age-earnings profiles of men with different education levels are not simply log-level shifts of each other. Highly educated men experience much more rapid wage income growth in percentage terms between the ages of 20 and 50. To account for these differences, I use Census Bureau data from 2005 to construct synthetic age-earnings profiles for men with different education levels. I use the mean earnings of men in several age buckets (18-24, 25-34, 35-44, 45-54, 55-64, and 65+) crossed with several education categories (<high school, high school, and college+). The results presented in this appendix use the same synthetic profile for both the optimistic and pessimistic imputations. Since the synthetic data are computed for full-time, year-round workers over the age of 18, they are more directly applicable to the estimates that assume no involuntary unemployment.

I use data bucketed into 5- and 10-year increments. Let $m_{e,a,a+1}$ be the slope of the earnings line connecting the labor income in age buckets a and $a+1$ for education category $e \in \{<\text{ high school},\ \text{high school},\ \text{college+}\}$, and let $\tilde{w}_{i,t,k}$ be the (imputed) annual wage income for respondent i in survey wave t using imputation rule $k \in \{\text{pess},\ \text{opt}\}$. Since most workers retire between the ages of 60 and 70, I assume that each NLSY respondent will work until age 65 and then retire. I calculate the expected annual wage income of i in year 2013, $\hat{w}_{i,2013,k}$ using a regression of $\hat{w}_{i,2012,k}$ on time trends and prior-year

income estimates. I assume that i's yearly income increases and decreases from $\hat{w}_{i,2013,k}$ between the ages of 47 and 65 in accordance with the slopes $\{m_{e(i),a,a+1}\}$, where $e(i)$ is the education level of i. Putting all of this together, the pdv of a youth who was 15 at the start of the NLSY79 is given by

$$\begin{aligned} PDV_{i,k} \equiv & \underbrace{\sum_{t=0}^{t=33} (0.95)^t \tilde{w}_{i,t,k}}_{\text{observed/imputed}} + \underbrace{\hat{w}_{i,2013,k} \sum_{j=1}^{8} (0.95)^{33+j} \left(1 + j m_{e(i),35,45}\right)}_{\text{projected, age 48-55}} \\ + & \underbrace{\hat{w}_{i,2013,k} \left(1 + 10 m_{e(i),35,45}\right) \sum_{j=1}^{10} (0.95)^{41+j} \left(1 + j m_{e(i),45,55}\right)}_{\text{projected, age 56-65}}. \end{aligned}$$

Both NLSY surveys record the highest grade completed for each respondent in each survey wave. Using these grade-completion variables, I construct a new variable for each survey wave t equal to the highest grade completed observed in any wave up to and including t. Occasionally, the highest grade completed for a respondent will decrease between one survey and the next. These data are difficult to interpret; my fill-in rule assumes that the lower value is incorrect. I only use the grade-completion variables up to 14 years after the start of the survey, as this is as far out as I can go in the NLSY97. Very few people change their education status after age 30 in the NLSY79, so this restriction should have little effect on my estimates.

Appendix C. Asymptotic Bias Simulation Procedure For Cliff's δ

I use the unadjusted test-score reliabilities drawn from Table 10 in all of the simulations. For each test s, I estimate $(\hat{\mu}_{s,t,G}, \hat{\sigma}_{s,t,G})$ for $G \in \{H, L\}$ and $t \in \{1979, 1997\}$ using the sample means and standard deviations. I draw random pseudo-samples of test scores of size N, for N large, from each distribution $N(\hat{\mu}_{s,t,G}, \hat{\sigma}_{s,t,G})$. I use these pseudosamples to estimate $\tilde{\delta}$. Then, if $R_{s,t}$ is the reliability of assessment s in year t, I generate "noiseless" pseudosamples drawn from $N(\hat{\mu}_{s,t,G}, \sqrt{R_{s,t}}\hat{\sigma}_{s,t,G})$ and use these pseudosamples to compute δ. I compute δ and $\hat{\delta}$ for each achievement test for the whole range of possible reliabilities reported in Reardon[31] and report the extrema of this procedure.

To simulate the bias stemming from income measurement error, I suppose the true distribution of income is lognormal with mean μ_t and variance σ_t^2 in both surveys t. I also assume that observed log income $\tilde{m}_{i,t}$ is equal to true log income plus a normally distributed classical measurement error $\eta_{i,t}$ with variance $\sigma_{\eta,t}$. Finally, I suppose that observed standardized test scores are linear in true log income: $s_{i,t} = a_t + B_t m_{i,t} + \varepsilon_{i,t}$, $\mathbb{E}[\varepsilon_{i,t}|m_{i,t}] = 0$, $\varepsilon_{i,t} \sim N(0, \sigma_{\varepsilon,t})$. Under these assumptions, a linear regression of test scores on observed log income will recover an asymptotically biased estimate of B_t:

$plim\ \hat{B}_t^{ols} = B_t R_t = B_t \left(\frac{\sigma_t^2}{\sigma_t^2+\sigma_{\eta,t}^2}\right)$. If R^2 is the true share of variance explained, then the asymptotic share explained in the noisy regression is $\tilde{R}^2 = R_t R^2$. These facts imply that the following procedure will provide valid approximations for δ and $\tilde{\delta}$:

(1) Estimate $\hat{\tilde{\mu}}_t$ and $\hat{\tilde{\sigma}}_t$ from the sample means and standard deviations of the log income distributions. Then, for some large $N \in \mathbb{N}$, draw a sample $\{m_{i,t}\}$ of size N from $N(\hat{\tilde{\mu}}_t, R_m \hat{\tilde{\sigma}}_t^2)$. These $\{m_{i,t}\}$ are the "clean" income values.

(2) Run a linear regression of $s_{i,t}$ on the observed log income values. Using the $\tilde{R}^2$ from this regression, simulate a population of errors $\{\varepsilon_{i,t}\}$ of size N by drawing a random sample from $N\left(0, \frac{1-\tilde{R}^2}{R_m}\right)$. Then, for each $m_{i,t}$ in the created sample from step 1, simulate a test score via $s_{i,t} = R_m \hat{B}_t^{ols} m_{i,t} + \varepsilon_{i,t}$.

(3) Create a virtual population of noisy incomes $\{\tilde{m}_{i,t}^v\}$ of size N via $\tilde{m}_{i,t}^v = m_{i,t} + \eta_{i,t}$, where the $\eta_{i,t}$ are iid draws from $N(0, (1-R_m)\hat{\tilde{\sigma}}_t^2)$. Repeat these steps for the other survey.

(4) For the clean data, calculate $\hat{\delta}$ from the scores that correspond to incomes in the top 20 percent and bottom 20 percent of the true income distributions for years t and $t+1$. For the noisy data, calculate $\hat{\tilde{\delta}}$ analogously using the noisy income distribution and compute $\text{Bias}(R_m) = \frac{\hat{\delta}-\hat{\tilde{\delta}}}{\hat{\delta}}$.

Appendix D. Anchoring Methods

I use only the "raw" (not age-standardized) test scores for the regression-based anchoring analysis described in Section 6. For each achievement test, I use the NLSY79 to estimate equation 5. My baseline specifications are linear in s. Higher-order polynomials produce similar gap/change estimates but are much harder to adjust for test-score measurement error. Various plausible specifications of equation 5 produce similar gap/change estimates.

The basic idea is to use the estimated coefficients from equation 5 to convert test scores into y-denominated units. In particular, if $\hat{y}_{i,G,t}$ is the predicted value of y for survey respondent i in racial/gender group G in year t, then the anchored average achievement for G in year t is estimated by the empirical average of the $\{\hat{y}_{i,G,t}\}$'s. These anchored averages can then be used to estimate any gap/change of interest. In this framework, the regression estimate of β_1 determines how strongly changes in test scores are reflected in changes in y. If test scores are estimated noisily (and substantial psychometric work suggests that they are), the OLS estimate of β_1 will be attenuated. I therefore manually adjust the OLS estimates of β_1 to reflect plausible guesses about the noisiness of the s. The NLS reports reliability estimates for each component AFQT test. I back out the implied measurement error variances using the observed test-score variances and the reliability estimates. I then compute "regression-adjusted" reliabilities using only the components of s that are orthogonal to the other regressors. Finally, I scale up the $\hat{\beta}_1$'s by the inverses

of these reliabilities prior to estimating the $\hat{y}$'s. Adjusting for measurement error in this way typically increases the gap-change estimates by about 20-30%. Table 10 shows the regression-adjusted test-score reliabilities I use.

Implementing the quantile regression-based procedure outlined in Section 6 requires that I estimate $K_{79}(y|s)$ for each $s \in [\underline{s}, \bar{s}]$ and the marginal test-score distributions $F_{t,G}$ for each $t \in \{79, 97\}$ and $G \in \{H, L\}$. I estimate the marginal test-score distributions using a smoothed kernel density estimator. I estimate $K_{79,t}(y|s)$ in two steps. First, I estimate polynomial quantile regressions of the form $y^{(\tau)} = \alpha^{(\tau)} + \beta_1^{(\tau)} s + \beta_2^{(\tau)} s^2 + \ldots \beta_n^{(\tau)} s^n$ for each $\tau \in \{\tau_1, \ldots, \tau_M\}$, where $0 < \tau_i < \tau_{i+1} < 1,\ 1 \leq i \leq M-1$. My baseline estimates use linear quantile regressions ($n = 1$), as my measurement error adjustment procedure (described below) only works well in this case.

Test-score measurement error is a thorny issue in this setting. In general, errors-in-variables will bias the quantile regression coefficients, and, by extension, the corresponding gap/change estimates. There is no straightforward, readily available method for handling measurement error in polynomial quantile regressions. I adjust for test-score measurement error by setting $n = 1$ and multiplying each estimated quantile regression coefficient by R_s^{-1}, where R_s^{-1} is the inverse of the corresponding test reliability in Table 10. To my knowledge, there is no general theorem supporting this procedure. Nonetheless, simulation exercises suggest that this procedure will be approximately valid when the latent test-score and measurement error distributions are symmetric and when the reliability of the observed scores is in the range reported by the NLS. The approximation is most accurate when $\tau \approx 0.5$; the adjustment will overcompensate for measurement error bias when τ is close to one and will under compensate when τ is close to zero. Adjusting for measurement error in this way has a relatively large effect for some of the quantile gap-change estimates. I do not adjust the cubic quantile regression-based estimates reported in the online appendix. I note only that test-score measurement error will generally have the effect of muting the estimated gradient between test scores and outcomes. Since the estimated reliabilities of the AFQT assessments are similar in the two NLSY surveys, this implies that my baseline cubic estimates should be viewed as conservative.

I use the estimated quantile regressions to estimate $\tilde{Q}_{79}(u|s)$, the quantile function of y conditional on s. Since the quantile regressions do not guarantee that the resulting $\tilde{Q}_{79}(u|s)$ is monotone in u, I estimate $\hat{K}_{79}(y|s)$ by $\int_0^1 \mathbb{I}\left(\tilde{Q}_{79}(u|s) \leq y\right) du$. Even if $\tilde{Q}_{79}(u|s)$ is not monotone, $\hat{K}_{79}(y|s) = \int_0^1 \mathbb{I}\left(\tilde{Q}_{79}(u|s) \leq y\right) du$ will be. Finally, I use cubic b-splines to smooth out the stepwise function defined by the above integral. The derivative of this smoothed cdf yields a smooth estimate of the conditional density of y given s, $\hat{k}_{79}(y|s)$.

My empirical work allows the estimated relationship between s and y to depend on student characteristics. In particular, I estimate $K_{79}(y|s, x)$ for a student with characteristics $x \in \{\{\text{Black}, \text{White}\} \times \{\text{Male}, \text{Female}\}\}$ via two methods. My baseline specification

estimates quantile regressions subsetting on the relevant race/sex category prior to estimation. That is, I estimate $y^{(\tau)} = \alpha^{(\tau)} + \beta_{x,1}^{(\tau)} s + \beta_{x,2}^{(\tau)} s^2 + \ldots + \beta_{x,n}^{(\tau)} s^n + \gamma_x^{(\tau)}(\text{age dummies})$ for $x \in \{\{\text{Black}, \text{White}\} \times \{\text{Male}, \text{Female}\}\}$. I also produce estimates, presented in the online appendix, in which I run the quantile regressions on the full data set but include dummies for x. The estimates produced using these alternate specifications are quite similar to those I report in this paper.