# Toward an Understanding of the Returns to Cognitive Skills Across Cohorts *

Judith Hellerstein† Sai Luo‡ Sergio Urzúa§

August 23, 2023

**Abstract**

Recent research concludes that wage returns to cognitive skills have declined in the U.S. We reassess this finding. Using Yitzhaki (1996) decomposition methods, we document the impact of shifts in the distributions of pre-labor market cognitive skills for white men and women across two cohorts. These shifts explain the declining estimated returns to cognitive skills, especially for men. Measurement error does not seem to be driving this conclusion. Grappling with pre-labor market skill distributions is necessary for capturing the dynamics of returns to cognitive skills. This may prove especially important in the future given evolving pandemic-induced changes in skill development.

*We thank Katharine Abraham, Dan Black, Melissa Kearney, Richard Murphy, Nolan Pope, and seminar participants at Chicago Harris, George Washington University, SUNY-Buffalo, UIUC, University of Cambridge, George Mason University, and SED 2023 annual meeting for valuable comments. We are grateful to Rosella Gardecki, Mark Loewenstein, Randy Olsen, Donna Rothstein, and other researchers at the Bureau of Labor Statistics and the Center for Human Resource Research (CHRR) at the Ohio State University for their help with the data. Errors are our own.

†University of Maryland. E-mail: hellerst@umd.edu.

‡Shanghai University of Finance and Economics. E-mail: luosaiecon@gmail.com.

§University of Maryland. E-mail: surzua@umd.edu.

# 1 Introduction

Recent dramatic declines in standardized test scores in the United States have drawn new attention to the importance of tracking changes in the distribution of skills across cohorts (NAEP 2023). However, this is not the first time skill distributions have changed across cohorts in recent decades (although it may well prove the most dramatic). Altonji et al. (2012) details the changing distribution of skills across two earlier cohorts of youth in the U.S. represented in the 1979 National Longitudinal Survey of Youth (NLSY–79) and the 1997 National Longitudinal Survey of Youth (NLSY–97). Taking as given this dynamic, several studies have concluded that the wage returns to cognitive skills declined for young workers in the U.S. over the past 40 years (Castex et al. 2014; Deming 2017; Ashworth et al. 2021). The accumulation of skills is at least partly an endogenous response to their labor market returns (Card 1999), and labor market returns can be endogenous to the distribution of skills (Acemoglu 2002). Thus, it is important to simultaneously characterize (and study) the changing distributions of skills and the evolution of their labor market returns across cohorts.

In this paper, we re-examine whether the labor market returns to cognitive skill have declined across cohorts, emphasizing the relationship between the estimation of these returns and the changing measured distributions. We use the measure of cognitive skill from Altonji et al. (2012), which is derived from the reported scores on the Armed Forces Qualifying Test (AFQT) for the NLSY–79 and NLSY–97 samples. The AFQT scores across the two cohorts in the NLSY data are not directly comparable, so Altonji et al. (2012) concorded the scores across cohorts as well as adjusted them for the age of the test takers. Because the resulting concorded scores are not raw AFQT test scores, we refer to them as "adjusted" AFQT scores, or AAFQT.

Our analysis focuses on the changes in the measured distributions of and returns to cognitive skill *within* groups. Specifically, and because Blacks and Hispanics have experienced marked changes over time in access to dimensions of the U.S. economy and constitute significantly smaller samples in the NLSY, we study the samples of white non-Hispanic men and white non-Hispanic women.[1] As in Altonji et al. (2012), we document that the average AAFQT scores increased slightly over time for both white men and women. However, we focus more on the fact that the *distribution* of scores widened and became more left-skewed. In particular, for both

1. For brevity, in the paper we often refer to these groups as "men" and "women." Results for white Hispanics and Blacks are available in Appendix E.

groups, there is a thicker tail at the bottom of the AAFQT distribution in the younger cohort (NLSY–97), and a "hollowing-out" in the middle of the AAFQT distribution.

We then delve into the implications of these distributional changes in measured cognitive skills on estimates of the returns to cognitive skills, re-considering the recent findings of their decline in the U.S. across the two NLSY cohorts (Castex et al. 2014; Deming 2017; Ashworth et al. 2021). We estimate the univariate linear (OLS) relationship between the log of wages and AAFQT scores. We show that the wage return to AAFQT declined for white men, confirming previous studies (the result for white women is less clear). We then replicate previous results demonstrating that the observed distributions of AAFQT scores for white men and women grew increasingly left-skewed across the cohorts.

We show that this increased left-skewness is central to the finding that the wage returns to cognitive skills have declined for white men, but does not markedly affect the wage returns for white women. We do this by implementing Yitzhaki (1996) decompositions, tracing out the relationship between the measured distribution of cognitive skill (AAFQT) and the estimation of its wage return in typical log-linear (OLS) wage regressions. Then, by comparing Yitzhaki decompositions across the two cohorts, we show that the construction of the OLS estimate in the younger cohort (NLSY–97) places much higher weight relative to the NLSY–79 on the wages of individuals with low levels of cognitive skills. This empirical fact has a marked influence on changes in the estimated wage returns across cohorts for men but less on the estimate for women.

We then conduct exercises to generate counterfactual estimated wage returns. We do this by maintaining the reported wages in each cohort in the NLSY data, but adjusting the distributions of AAFQT scores to be the same across cohorts. This is done by a reweighting that arises directly from the Yitzhaki decomposition. We show that when the AAFQT distribution is fixed in each cohort to be that of the older cohort (NLSY–79), the estimated counterfactual returns to cognitive skills for men are unchanged across the two cohorts. This starkly contrasts with the falling returns generated by the OLS estimates. For women, on the other hand, reweighting increases the estimated (counterfactual) decline in the wage return relative to OLS. However, the absolute magnitude of the change for women is much less dramatic than for men.

We conclude by briefly assessing the possibility that measurement error in the AAFQT scores could be driving the results. Unfortunately, there is no easy way to correct or account for this source of bias, but we argue that it is unlikely to be the main driver of these findings.

# 2 Data and Related Research

## 2.1 NLSY and the Measure of Cognitive Skills

The National Longitudinal Surveys of Youth (NLSYs) have shaped our understanding of the U.S. labor market over the past 40 years. A large body of research has examined the experiences of the NLSY–79, a nationally representative sample of the cohort of American youth aged 14–22 when first surveyed in 1979. A growing body of more recent research has focused on the NLSY–97, a younger cohort aged 12 to 16 when first surveyed in 1997. The individuals in the NLSY–97 cohort are now old enough to draw comparisons between their labor market experiences in early adulthood and those of the NLSY–79. Comparing experiences and outcomes across these two cohorts is helping to explain and understand the evolution of the U.S. labor market . Plans are underway for a new NLSY cohort[2], which, if fielded as intended, will include a cohort of youth markedly affected during early years of school by the Covid-19 pandemic. Measuring the skill distributions of this new cohort and eventually studying their subsequent adult labor market outcomes will be critical, including via cross-cohort comparisons. This is especially true given early evidence from 13-year-olds that standardized test score declines already apparent in 2020 accelerated in 2023, alongside increased variance.[3]

As a measure of cognitive skills, researchers have utilized the AFQT scores of the NLSY respondents (e.g. Neal et al. 1996; Heckman et al. 2006; Urzúa 2008). These scores are based on specific sections of the Armed Services Vocational Aptitude Battery (ASVAB). Respondents in both NLSY cohorts took versions of the ASVAB, although the raw scores are not directly comparable across cohorts due to changes in test administration.[4]

As discussed above, Altonji et al. (2012) concord and adjust the AFQT scores across the two cohorts, creating what we refer to as AAFQT scores. Based on a number of assumptions, they conclude that the skill distribution widened between the NLSY–79 and NLSY–97, and that this likely has important implications for wage inequality.[5] When Altonji et al. wrote their paper, the NLSY–97 cohort was too young to have realized wage outcomes, so conclusions about wage

2. https://www.bls.gov/nls/nlsy26.htm accessed July 14, 2023

3. https://www.nationsreportcard.gov/highlights/ltt/2023/ accessed July 14, 2023.

4. See Appendix B for more details about the ASVAB and the AFQT score.

5. There are two fundamental differences in the test format and administration across the two cohorts. First, while the NLSY–79 respondents took a paper-based test, the NLSY–97 respondents took a computer-based test designed using Item Response Theory (IRT), so not all respondents answered all questions. Second, the NLSY–79 respondents were ages 15–23 when they took the test, while the NLSY–97 respondents were 12–18.

inequality remained speculative. Thus, this paper aims to enhance the existing work in this area.

### 2.2 Existing Findings of Wage Returns to Cognitive Skills

Since Altonji et al. (2012), two other influential papers have used realized wages for both NLSY cohorts to study how the early career wage returns to skills have changed across cohorts. Castex et al. (2014) focus on estimating changes in the skill price of AAFQT. While Deming (2017)'s primary interest is in changes in the price of measures of social skills, he also estimates changes in the skill price of AAFQT. Both papers estimate conventional linear hedonic wage functions where skill prices are constant across the skill distribution. Despite differences in specific choices of sample construction and model specifications, they both find that the wage returns to AAFQT have declined across cohorts.[6]

We first re-examine the findings of Castex et al. (2014) and Deming (2017). To this end, we estimate log-linear wage equations of the form:

$$\ln W_i^c \quad = \quad \alpha^c + \beta^c AAFQT_i^c + \epsilon_i^c, \tag{1}$$

where $W_i^c$ denotes the (log) average hourly wage of individual $i$ from cohort $c$ (NLSY–79 or NLSY–97) observed between the ages of 25 and 39, and $\epsilon_i^c$ is the associated error term.[7]

Figure 1 plots the relationship between (average) log wages and AAFQT scores. It also displays the OLS results of estimating equation (1). Panel A shows the results for white men. The estimated (log) wage return to an additional AAFQT point falls from 0.677 for the NLSY–79 cohort to 0.464 for the NLSY–97 cohort, a large and statistically significant drop of 0.212.

6. Ashworth et al. (2021) estimates a dynamic structural model using data from the two NLSY cohorts. Consistent with the previous findings, they conclude that returns to unobserved cognitive ability (measured in a factor model) have declined across cohorts. Weinberger (2014) compares two samples of 12th-graders from 1972 and 1992 seven years after graduation and concludes that the returns to math score *increased* across those cohorts. The samples' characteristics and the test score's nature might explain this distinctive change.

7. We estimate and report results from weighted least squares regressions, using the BLS custom sample weights. With some abuse of terminology, we refer to these regressions throughout the paper as "OLS" so as not to confuse sample weights with Yitzhaki weights discussed below, which are our key set of weights. We multiply $\ln W_i^c$ by 100 in our Figures and Appendices for ease of display. In addition, we consider univariate regressions. This facilitates the exposition of our Yitzhaki decomposition exercises, frees us from complicated issues related to concording other covariates (e.g. social skills) across NLSY cohorts, and does not impose functional form assumptions on the relationship between the covariates, AAFQT, and log wages. Conceptually, to add covariates linearly, one can first residualize both $\ln W$ and $AAFQT$ with covariates, and then apply the Yitzhaki decomposition to the residualized $\widetilde{\ln W}$ and $\widetilde{AAFQT}$.

For white women, as shown in Panel B, the estimated return to an additional AAFQT point falls from 0.830 for the NLSY–79 cohort to 0.789 for the NLSY–97 cohort. The drop is not significant and its magnitude is much smaller than that of men.[8]

In the next section, we present evidence of how the measured distribution of cognitive skills (AAFQT) has changed across cohorts. As discussed in Section 4, this is critical for understanding the evolution of the OLS estimates displayed in Figure 1.

## 3 Distributional Changes in Cognitive Skills

It is well known that the wage structure widened in the U.S. labor market over the decades encompassing the early adulthood of the NLSY cohorts (e.g. Katz et al. 1999; Card et al. 2002; Autor et al. 2008), and employment grew rapidly not only at the highest-skill jobs but also at the lowest-skill jobs (Autor et al. 2006; Autor et al. 2013). Much less discussed, at least in the context of labor market outcomes, is how the underlying skill distribution–using detailed skill measures other than education–changed over time.

Altonji et al. (2012) is an important exception. Though their focus is a composite skill index rather than a specific skill measure, the authors note the changing distribution of AAFQT scores and document a widening distribution of their composite skill index across the two NLSY cohorts. Figure 2 replicates their AAFQT result for white men and women. We also report corresponding distributional statistics. The first finding, as noted by Altonji et al., is that both the mean and median of AAFQT scores are slightly higher in the younger (NLSY–97) cohort for both groups.

Perhaps more strikingly, the skewness of the AAFQT distribution is much more pronounced for the younger cohort (0.81 for white men, 0.79 for white women) than for the older cohort (0.63 for white men, 0.56 for white women). Consistent with this, the kurtosis of the distributions also increased across cohorts for white men and women. Statistical tests comparing skewness, kurtosis, and the overall distributions of AAFQT scores all reject nulls of no differences across cohorts (See Appendix A Tables A.1 and A.2.).

For white men, the increasing mass of people with very low scores and the overall increase (though not large in magnitude) in the mean and median across the cohorts together create

8. Results including covariates, reported in Appendix A Table A.3, tell largely the same story as the univariate regressions.

"hollowing out" in the low-to-median range of the AAFQT distribution. There is also a hollowing out for white women, but it is driven more by gains in AAFQT scores for individuals in the 10th to 50th percentile. To our knowledge, these distributional changes in cognitive skills have received very little attention.[9] In the next section, we connect these changes to the estimated returns to cognitive skills. To do this, we hearken back to Yitzhaki (1996).

# 4 The Yitzhaki Decomposition

In order to understand the underlying mechanisms behind the estimated declines in the returns to AAFQT scores, we implement the Yitzhaki (1996) decomposition. Consider a generalization of the (log) linear wage equation (1):

$$Y = E[Y|H, C] + \epsilon = \alpha(C) + \beta(C)H + \epsilon,$$

where $C$ denotes a given cohort (that has, e.g., cohort-specific skill-neutral technology) and $H$ is human capital.

Let $B_C(h) = E(Y|C, H = h)$ be the regression curve and $b_C(h)$ be its slope, i.e. the unit treatment effect evaluated at $h$ for a given $C$. The Ordinary Least Squares (OLS) estimate of the linear relationship between $Y$ and $H$ can be expressed as:

$$\beta_C^{OLS} = \int_h w_C(h)\, b_C(h)\, dh. \tag{2}$$

Therefore, $\beta_C^{OLS}$ can be decomposed into unit treatment effects $b_C(h)$ and how they are weighted by $w_C(h)$.

In order to highlight the specific role of skewness in the weights, we rearrange Yitzhaki's original formulation and write the weights as:

$$w_C(h) = \left[\frac{F_{C,H}(h)\Big(1 - F_{C,H}(h)\Big)}{\sigma^2_{C,H}}\right] \Big(E_C(H|H > h) - E_C(H|H \leq h)\Big), \tag{3}$$

9. The basic patterns of distributional change in AAFQT, including the increased left-skewness and kurtosis, are not caused by changing correlation with the covariates such as measured of non-cognitive and social skills Deming (2017) and education. See Appendix A Figure A.1 for the distribution of AAFQT scores residualized by covariates.

where $F_{C,H}(h)$ is the cumulative density function of $H$ evaluated at $h$, $\sigma^2_{C,H}$ is the associated variance, and $E_C(\cdot)$ denotes the expectation. The weights, $w_C(h)$, are non-negative and solely depend on the distribution of $H$ given $C$, not on the distribution of $Y$. The role of $Y$ in the construction of the OLS estimates comes only through the unit treatment effects $b_C(h)$.

The first term in brackets in expression (3) reaches its peak when $F_{C,H}(h) = 0.5$, i.e., at the center of the distribution of $H$. Its contribution to the weights is fairly intuitive. But understanding the second expression in (3) is equally important. In particular, larger differences in the conditional expectations on either side of a given $h$ contribute more to the OLS weights. So this dispersion is essential for driving OLS estimates. In particular, left(right)-skewed distributions tend to put higher (lower) weight on $h$'s toward the bottom of the distribution. Thus, when unit treatment effects $b_C(h)$ differ across the distribution of $H$, the weighting scheme of the Yitzhaki decomposition plays a key role in the OLS estimate $\beta_C^{OLS}$. If the unit treatment effects are constant, the weights do not matter in practice.

In most empirical analyses, $H$ is discrete; in our application, $H$ is the AAFQT score. Therefore, we use the discrete version of expression 2 to implement the decomposition. Specifically, and dropping $C$ from the notation for parsimony, we first rank observations in increasing order of $H$, so $h_1 < h_2 < \cdots < h_n$, where $n$ denotes the number of distinctive realizations of $H$. Let $N_i$ be the number of duplicate observations for $h_i$ and let $N$ be the sum of all observations: $N = N_1 + \cdots + N_n$.[10] Then, let $\Delta h_i = h_{i+1} - h_i$ and $b_i = \Delta \bar{y}_i / \Delta h_i$. Thus, we can think of $b_i$ as the pairwise slope or estimated unit treatment effect and the OLS estimator can be expressed as:

$$\beta_{OLS} = \sum_{i=1}^{n-1} w_i b_i, \quad \text{with} \quad \sum_{i=1}^{n-1} w_i = 1 \ \text{ and } \ w_i \geq 0 \quad \forall i,$$

where the discrete weights $w_i$ can be written analogously to the continuous weights $w_h$:

$$w_i = \frac{1}{\sigma_h^2} \frac{\sum_{j=1}^{i} N_j}{N} \frac{\sum_{j=i+1}^{n} N_j}{N} \left( \frac{\sum_{j=i+1}^{n} N_j h_j}{\sum_{j=i+1}^{n} N_j} - \frac{\sum_{j=1}^{i} N_j h_j}{\sum_{j=1}^{i} N_j} \right) \Delta h_i. \tag{4}$$

Finally, we decompose the OLS estimate obtained from equation (1) in each NLSY cohort

10. When weighted least squares (WLS) is utilized instead of OLS, it is straightforward to extend the Yitzhaki decomposition. See Appendix C for details.

as:

$$\beta_{OLS}^{79} = \sum_{i=1}^{n-1} w_i^{79} b_i^{79} \qquad \text{and} \qquad \beta_{OLS}^{97} = \sum_{i=1}^{n-1} w_i^{97} b_i^{97} \tag{5}$$

These expressions indicate that one can examine the changing OLS returns to AAFQT scores across the NLSY cohorts by examining whether the change is mechanically driven by changing Yitzhaki weights, changing pairwise slopes, or both. Moreover, because the Yitzhaki weights are only a function of AAFQT scores, we can specifically examine how much the changing distribution of AAFQT scores between the two cohorts affects the construction of the OLS estimates.

## 5 Understanding the Wage Returns to Cognitive Skills

Figure 3 plots the Yitzhaki weights by gender for each NLSY cohort.[11] Given the similarities in the AAFQT distributions for white men and women, the weights – in particular, the changes in the weights across cohorts – are also alike between these groups.

Low AAFQT scores receive more weight for the NLSY–97 than for the NLSY–79. Looking back at Figure 2 and noting equation (3), it is the increasing left-skewness of the AAFQT distribution in the NLSY–97 that yields larger weights on low AAFQT scores for the NLSY–97 than for the NLSY–79. This is true for AAFQT scores up to around 140, about the 15th–20th percentile. Beyond this region, for AAFQT scores up to about the 75th percentiles, weights are higher for the NLSY–79. The weights are similar across the cohorts for the top quartile or so of AAFQT scores.

Examining how the Yitzhaki weights differ across cohorts does not alone explain why, mechanically, the estimated OLS returns to AAFQT are lower in the NLSY–97 relative to the NLSY–29. As is clear from equation (5), studying how the Yitzhaki weights work together with the pairwise slopes is critical for understanding this result. Figure 4 again depicts the (smoothed) Yitzhaki weights both each cohort, this time overlayed with smoothed pairwise slopes (the $b_i$'s).[12] While the weights look similar between white men and women, the slopes

11. To make the graphs more readable, we collapse AAFQT scores into bins containing three consecutive AAFQT points. We also overlay the binned weights with local linear regression curves estimated using Locally Weighted Scatterplot Smoothing (LOWESS) with a tricube weighting function and a bandwidth of 0.1 unless otherwise noted.

12. A bandwidth of 0.3 is used for smoothing the pairwise slopes.

reveal distinctive patterns.

For the sample of white men (top panel of Figure 4), the gradient of the relationship between AAFQT and log wages remains upward-sloping and relatively constant through much of the AAFQT distribution for the NLSY–79 cohort (except perhaps at the very top and very bottom–places where local linear regression performs less well). This leads to what appears to be a positive and essentially linear relationship between AAFQT and log wages for white men in the NLSY–79.

The gradient for white men in the NLSY–97 cohort is less constant; in particular, it has flat spots at various points in the distribution, especially for AAFQT scores in the 110 to 140 range (approximately the 5th–20th percentile). This region of AAFQT scores also displays large weights in the NLSY–97, implying that here the lower OLS estimate of the wage return to AAFQT is driven primarily by these flat spots in the local linear regression.

By contrast, the slopes for white women displayed in the bottom panel of Figure 4 are characterized by a constant gradient across much of the AAFQT distribution for both cohorts. Moreover, for much of the AAFQT distribution, the gradients between the two cohorts look quite similar. One exception emerges for AAFQT scores between 140 to 160 points, where the gradient is flat or slightly negative for the NLSY–97. Another exception is at the top of the AAFQT distribution (above the 75th percentile): the gradient for the NLSY–97 flattens out while the gradient for the NLSY–79 increases. This nonlinear pattern differs from that of white men in the NLSY–97 cohort. As discussed in the following subsection, this result is crucial in explaining the distinction between white women and men in the counterfactual OLS estimates.

To better understand how the different parts of the AAFQT distribution contribute to the OLS estimates of the wage returns to AAFQT, Figure 5 displays the progressive sum of the Yitzhaki decomposition from equation (5), starting with the lowest AAFQT score until the entire sum is calculated (producing the OLS estimate).[13]

The top left panel presents the results for white men. The contributions of the lowest AAFQT scores to the OLS estimates are not markedly different across cohorts. However, the progressive

13. For each AAFQT score $h_k$ from $h_1$ to $h_{n-1}$, we calculate and graph for each cohort $c$ the cumulative sum of the Yitzhaki decomposition:

$$\sum_{i=1}^{k} w_i^c b_i^c, \quad k = 1, \ldots, n-1$$

We graph the progressive sum by three-point bins and use a bandwidth of 0.3 for smoothing.

sums from the Yitzhaki decomposition begin to permanently diverge at an AAFQT score of around 110 (5th percentile), at which point the Yitzhaki sum for the NLSY–79 cohort rises quickly and continuously until it reaches its final level of 0.68. In contrast, the Yitzhaki sum for the NLSY–97 stays low until around an AAFQT score of 150 (25th percentile). It then rises quickly, only to fall again in the 160 to 180 points range before recovering and reaching its final level of 0.46. This is (not surprisingly) consistent with Figure 4, where the flat slopes for the NLSY–97 play a large role in depressing the NLSY–97 OLS estimate relative to the NLSY–79 estimate.

The bottom left panel of Figure 5 reports the results for white women. The Yitzhaki sums of the two cohorts move almost in tandem and remain close to each other for much of the AAFQT distribution, with two noticeable exceptions. First, the Yitzhaki sum for the NLSY–97 cohort stops rising when reaching around 150 points but then quickly catches up. Second, at an AAFQT score of around 195 points (75th percentile), the Yitzhaki sums of the two cohorts are still very close; then they diverge again. The sum for the NLSY–79 cohort continues to rise, eventually reaching its OLS estimate level of 0.83. In contrast, for the NLSY–97 cohort, it stays largely constant after the 195 points, before reaching its final level of 0.79. This, again, is consistent with Figure 4, where the gradients of the local linear regression diverge between the two cohorts at the top of the AAFQT distribution.

### 5.1 Counterfactual OLS Estimates

Given the different OLS wage returns to AAFQT *and* the changing AAFQT distributions between the NLSY–79 and NLSY–97 cohorts, we ask the following counterfactual question: Would the OLS returns to AAFQT have changed between the two cohorts if the distribution of AAFQT had not changed (holding wages at their observed values)? Or, equivalently: Would the OLS returns to AAFQT have changed if the Yitzhaki weights had been held fixed across the cohorts but the observed pairwise slopes had still been realized? To the best of our knowledge, this is the first time the Yitzhaki decomposition has been used to consider this kind of counterfactual comparison.

We answer this question by decomposing the observed difference between $\beta_{OLS}^{79}$ and $\beta_{OLS}^{97}$

as:

$$\begin{aligned}\beta_{OLS}^{79} - \beta_{OLS}^{97} &= \Big(\beta_{OLS}^{79} - \beta_{OLS}^{97}|w^{79}\Big) + \Big(\beta_{OLS}^{97}|w^{79} - \beta_{OLS}^{97}\Big), \\ &= \sum_i w_i^{79}(b_i^{79} - b_i^{97}) \; + \; \sum_i (w_i^{79} - w_i^{97})b_i^{97}. \end{aligned} \tag{6}$$

The first term in equation 6 is the counterfactual difference in the OLS estimates, holding fixed the AAFQT distribution (and the corresponding Yitzhaki weights) at the NLSY–79 level.[14]

We first present the counterfactual estimates for white men. Using expression (6), the OLS decline of 0.21 points for white men (as reported in Figure 1) can be decomposed as:

$$\begin{aligned}\beta_{OLS}^{79} - \beta_{OLS}^{97} &= \Big(0.677 - 0.689\Big) + \Big(0.689 - 0.464\Big) \\ &= -0.012 + 0.225 \end{aligned} \tag{7}$$

The first term is the counterfactual change in returns holding the Yitzhaki weights at NLSY–79 levels, a counterfactual that finds for white men that the returns to AAFQT scores stayed basically unchanged across the cohorts (an *increase* of 0.012 points). This finding highlights the critical role of the changing composition of AAFQT scores in the narrative that there has been a decline in the return to cognitive skills. When we use the weighting structure generated by the AAFQT distribution in the older cohort (NLSY–79) to calculate the (linear) returns to cognitive skills, we find no evidence that the returns have declined for white men.

We gain further insight into the mechanics behind this by again graphing cumulative contributions, this time comparing the counterfactual cumulative contributions for NLSY–97 (0.689) to the NLSY–79 OLS estimate (0.689). This is graphed in the top right panel of Figure 5. In contrast to the top left panel of OLS results, here, the counterfactual NLSY–97 sum exceeds the actual NLSY–79 OLS sum for a good portion of AAFQT scores below the median, as the larger NLSY–79 weights pull up the cumulative sum enough (relative to the NLSY–97 weights) to overtake the overall NLSY–79 sum. The two sums in the top right panel end up converging

14. An alternative decomposition is:

$$\begin{aligned}\beta_{OLS}^{79} - \beta_{OLS}^{97} &= \Big(\beta_{OLS}^{79} - \beta_{OLS}^{79}|w^{97}\Big) + \Big(\beta_{OLS}^{79}|w^{97} - \beta_{OLS}^{97}\Big) \\ &= \sum_i (w_i^{79} - w_i^{97})b_i^{79} \; + \; \sum_i w_i^{97}(b_i^{79} - b_i^{97}) \end{aligned}$$

The second term is the counterfactual difference in the OLS estimates, holding fixed the AAFQT distribution at the NLSY–97 level.

again above around the median AAFQT score, and increase in tandem thereafter.

The counterfactual estimates for white women are in stark contrast to those of white men. For this groups, we decompose the OLS decline (of 0.04 points), again holding the Yitzhaki weights at NLSY–79 levels:

$$\begin{aligned} \beta_{OLS}^{79} - \beta_{OLS}^{97} &= \Big(0.830 - 0.702\Big) + \Big(0.702 - 0.789\Big) \\ &= 0.128 - 0.087 \end{aligned} \tag{8}$$

The counterfactual estimate, the first term, indicates that the wage return to AAFQT scores went down for white women by 0.128 points. This counterfactual estimate for white women, like that for men, deviates from the actual OLS result, but unlike for men, both the OLS and counterfactual estimates show declines across the cohorts. The bottom right panel of Figure 5 shows that the cumulative sum for the counterfactual NLSY–97 estimate only falls below the NLSY–79 OLS cumulative sum in the top quarter or so of the AAFQT distribution; otherwise the two curves are very similar. More generally, the absolute magnitude of the overall difference between the actual NLSY–97 OLS estimate and the counterfactual estimate is much smaller for women than for men. This is entirely consistent with the results in Figure 4 and the left panels of Figure 5, where the observed differences for women across the cohorts are much less stark than for men.

## 6 The Measurement of Cognitive Skills

The interpretation of our findings hinges on the assumption that the changing AAFQT scores reflect true changes in cognitive skill. In other words, until now we have not considered the potential impact of measurement error in AAFQT scores.[15] In this section, we briefly consider three related questions: (1) whether AAFQT is a mismeasured proxy for cognitive skill; (2) if so, at what stage was much of the measurement error introduced?; (3) and whether it is possible to easily correct for measurement error.

First, the divergence and increased (left) skewness of AAFQT distribution in the NLSY–97 relative to that of the older cohort is not being driven by one specific section of the ASVAB that

15. There has long been concern about measurement error in test scores. See Griliches et al. (1972) for an early treatment.

Altonji et al. (2012) used to create AAFQT. Instead, as we show in Appendix A Figure A.2, it appears to some degree in all four parts of the AAFQT. No individual section of the AAFQT appears anomalous.

Second, the change does not seem to be a direct artifact of the concordance of the different test formats across the two ASVAB administrations. Segall (1997) documents the concordance process. We discuss the details of this process in Appendix B. While there are some anomalous aspects to the scores in the NLSY–97 cohort, Appendix B explains that they existed in the scores before any concordance was performed.

All told, to the extent that there may be measurement error in the AAFQT scores in the NLSY–97 that drive changes across cohorts, it seems to be present in all the original AFQT test score results and is not a function of adjustments made to concord the test scores across the cohorts.

For any measurement error in AAFQT scores that does exist, one could still potentially correct for it in the estimate of the returns to cognitive skill. Motivated by measurement error concerns, Castex and Dechter (2014) estimate two-staged least squares regressions in some of their analyses, using SAT scores as an instrument for AAFQT scores. However, AAFQT scores in the NLSY–97 are derived from "Item Response Theory" (IRT) models, and, as pointed out in Schofield (2014)) and Jacob et al. (2016), measurement error is non-classical for IRT-based test scores.[16] Thus, simple IV does not work. One alternative approach is the "mixed effects structural equations" method in Junker et al. (2012). But implementing this requires data on the responses of individual test-takers to each question in the ASVAB and these are unavailable for the NLSY–97.

In the end, we do not have enough information to reach a definitive conclusion about whether the changing distribution of AFQT scores–and thus, AAFQT scores–is subject to measurement error. We also do not have an obvious method to correct for measurement error. But there are two pieces of evidence that suggest that measurement error alone cannot be driving estimates of changing cognitive returns across the cohorts.

First, to the extent that AAFQT scores are mismeasured, this measurement error should affect both white men and women in the NLSY samples. Indeed, this could explain the consistent

16. In Appendix B Figure B.2 we plot the standard errors of the estimated IRT scores. The errors seem to be oddly large for low scores in the NLSY–97, suggesting some underlying issues with the IRT model used for the NLSY–97. We thank Dan Black for pointing this out.

shifts in the AAFQT distributions across cohorts for both men and women. But if measurement error is present in the same form for men and women, there is no clear reason that it should differentially affect the changes in estimated returns to cognitive skill across men and women that we observe.

Second, if measurement error in the construction of AFQT is driving the changing AFQT (and AAFQT) distributions across cohorts, it should be unique to AFQT tests. But this is not the case. We generate additional evidence from two widely-used longitudinal data sets from the National Center for Education Statistics (NCES), the National Education Longitudinal Study of 1988 (NELS:88) and the Educational Longitudinal Study of 2002 (ELS:02). The cohorts in these data are different than those in the NLSY cohorts–the NELS:88 sample is about 7–11 years older than the NLSY–97 and the ELS:02 sample is about 1-5 years younger than the NLSY–97. But interestingly, when we plot 12th-grade math scores for these two samples, we also find increasing left-skewness and kurtosis for white men and for white women across the cohorts.[17] This provides additional suggestive evidence that the changing AAFQT score distributions across cohorts are not driven by something anomalous in the NLSY data, and in particular not driven solely by measurement error in AAFQT scores.

# 7 Conclusion

Understanding the labor market returns to education and skills has been critically important in empirical labor economics (Becker 1964). This process is facilitated by a growing availability of datasets, such as the NLSY, that contain measures of both labor market outcomes and different dimensions of skills.

Using the (in our view, under-appreciated) Yitzhaki decomposition in various ways, we demonstrate how one can understand the mechanics behind changing estimated returns to AAFQT scores across two NLSY cohorts of white men and women. Of particular substantive empirical importance, we show that for white men in the NLSY, the estimated decline in the return to AAFQT scores critically depends on changes in the distribution of AAFQT scores between the two cohorts and in particular on the widening (and increased left-skewness) of the

17. Unlike for the NLSY cohorts, we find no evidence that wage returns to math scores decreased from the NELS:88 to the ELS:02 for men or women. This is consistent with Weinberger (2014) who compares NELS:88 to an earlier NCES dataset. See Appendix D. This is additional evidence of the fragility of the narrative of declines in returns to cognitive skill

AAFQT distribution in the NLSY–97.

To the extent that returns to cognitive skills did decline for the white men in our sample, the Yitzhaki weights and the accompanying unit slopes show that they did so only at the lower part of the AAFQT distribution. In contrast, the returns to AAFQT for white women are relatively constant across the cohorts, so changes in the AAFQT distribution for these women did not affect the estimated returns much. Assuming that the changes in the AAFQT distribution reflect real changes in cognitive skill, and assuming the resulting estimates of wage returns are also mostly correct[18], it is natural to contemplate the underlying economic reasons behind them. This is difficult. From a theoretical perspective, any economic model that seeks to rationalize the results for changing wage returns to cognitive skill should not focus solely on the determinants of wage returns (such as changing technology on the demand side or changing labor market experience on the supply side) but also should address how and why the distributions of cognitive skill are affected by underlying economic drivers of skill investment. But practically, the NLSY samples are small–there are between 1400 and 2200 people in each of the four subgroups in our analysis. This severely limits possibilities for using the NLSY to test theories that endogenize both heterogeneous skill investment and labor market returns.

18. We emphasize, however, that our discussion throughout the paper and in Appendix D suggests that the results on the declining returns to cognitive skill should be treated with some skepticism.

## References


Acemoglu, Daron. 2002. "Technical Change, Inequality, and the Labor Market." *Journal of Economic Literature.*

Altonji, Joseph, Prashant Bharadwaj, and Fabian Lange. 2009. *Constructing AFQT Scores that are Comparable Across the NLSY79 and the NLSY97.*

———. 2012. "Changes in the Characteristics of American Youth: Implications for Adult Outcomes." *Journal of Labor Economics.*

Ashworth, Jared, V. Joseph Hotz, Arnaud Maurel, and Tyler Ransom. 2021. "Changes across Cohorts in Wage Returns to Schooling and Early Work Experiences." *Journal of Labor Economics.*

Autor, David, and David Dorn. 2013. "The growth of low-skill service jobs and the polarization of the US labor market." *American Economic Review.*

Autor, David, Lawrence Katz, and Melissa Kearney. 2006. "The polarization of the US labor market." *American Economic Review.*

———. 2008. "Trends in US wage inequality: Revising the revisionists." *The Review of Economics and Statistics.*

Becker, Gary. 1964. *Human capital: A theoretical and empirical analysis with special reference to education.* University of Chicago press.

Bureau of Labor Statistics. 1992. *NLSY79 Profiles of American Youth: Addendum to Attachment 106.* Technical report.

———. *Administration of the CAT-ASVAB (NLSY97).* https://www.nlsinfo.org/content/cohorts/nlsy97/topical-guide/education/administration-cat-asvab-0. Accessed: 2020-06-23.

———. *Aptitude, Achievement Intelligence Scores (NLSY79).* https://www.nlsinfo.org/content/cohorts/nlsy79/topical-guide/education/aptitude-achievement-intelligence-scores. Accessed: 2020-06-23.

Card, David. 1999. "The causal effect of education on earnings." *Handbook of Labor Economics.*

Card, David, and John E DiNardo. 2002. "Skill-biased technological change and rising wage inequality: Some problems and puzzles." *Journal of Labor Economics.*

Castex, Gonzalo, and Kogan Dechter. 2014. "The Changing Roles of Education and Ability in Wage Determination." *Journal of Labor Economics.*

Defense Manpower Data Center. 2006. *ASVAB Technical Bulletin.* Technical report.

Deming, David. 2017. "The Growing Importance of Social Skills in the Labor Market." *The Quarterly Journal of Economics.*

Griliches, Zvi, and William Mason. 1972. "Education, Income, and Ability." *Journal of Political Economy.*

Heckman, James, Jora Stixrud, and Sergio Urzúa. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics.*

Ing, Pamela, Carole Lunney, and Randall Olsen. 2012. "Reanalysis of the 1980 AFQT Data from the NLSY79." *Center for Human Resource Research Working Paper.*

Jacob, Brian, and Jesse Rothstein. 2016. "The Measurement of Student Ability in Modern Assessment Systems." *Journal of Economic Perspectives.*

Junker, Brian, Lynne Schofield, and Lowell Taylor. 2012. "The use of cognitive ability measures as explanatory variables in regression analysis." *IZA Journal of Labor Economics.*

Katz, Lawrence, and David Autor. 1999. "Changes in the wage structure and earnings inequality." In *Handbook of Labor Economics,* vol. 3. Elsevier.

NAEP. 2023. *U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), various years, 1971–2023 Long-Term Trend (LTT) Reading and Mathematics Assessments.* Technical report.

Neal, Derek, and William Johnson. 1996. "The role of premarket factors in black-white wage differences." *Journal of Political Economy.*

Prada, Maria F, and Sergio Urzúa. 2017. "One Size Does Not Fit All: Multiple Dimensions of Ability, College Attendance, and Earnings." *Journal of Labor Economics.*

Quester, Aline, and Robert Shuford. 2017. *Population Representation in the Military Services: Fiscal Year 2015 Summary Report.* Technical report.

Sands, William, Brian Waters, and James McBride, eds. 1997. *Computerized Adaptive Testing: From Inquiry to Operation.* American Psychological Association.

Schofield, Lynne. 2014. "Measurement error in the AFQT in the NLSY79." *Economics Letters.*

Segall, Daniel. 1997. "Equating the CAT-ASVAB." In *Computerized Adaptive Testing: From Inquiry to Operation.* American Psychological Association.

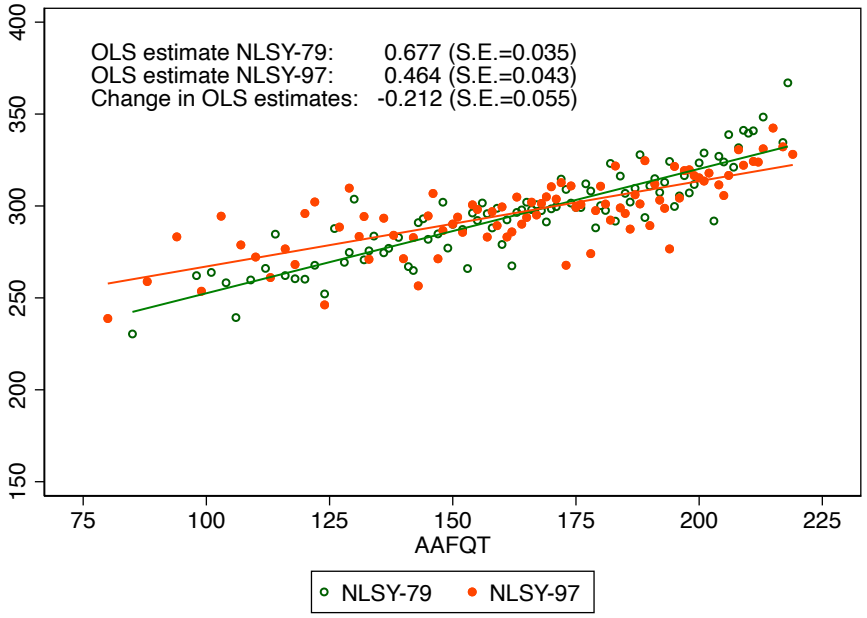
Urzúa, Sergio. 2008. "Racial Labor Market Gaps The Role of Abilities and Schooling Choices." *Journal of Human Resources.*

Weinberger, Catherine. 2014. "The Increasing Complementarity between Cognitive and Social Skills." *Review of Economics and Statistics.*

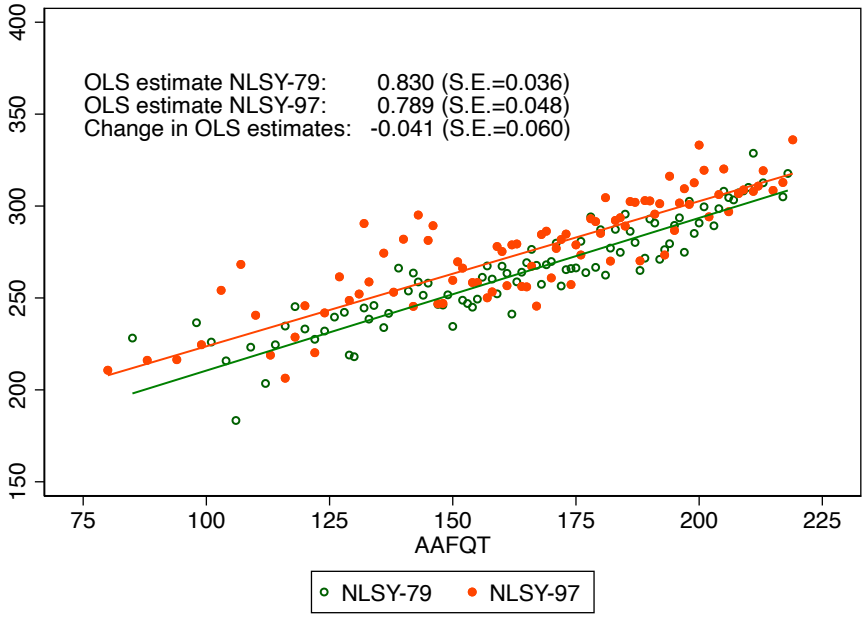
Yitzhaki, Shlomo. 1996. "On Using Linear Regressions in Welfare Economics." *Journal of Business & Economic Statistics.*

Figure 1: OLS Estimates of the Wage Returns to AAFQT

(a) White Non-Hispanic Men



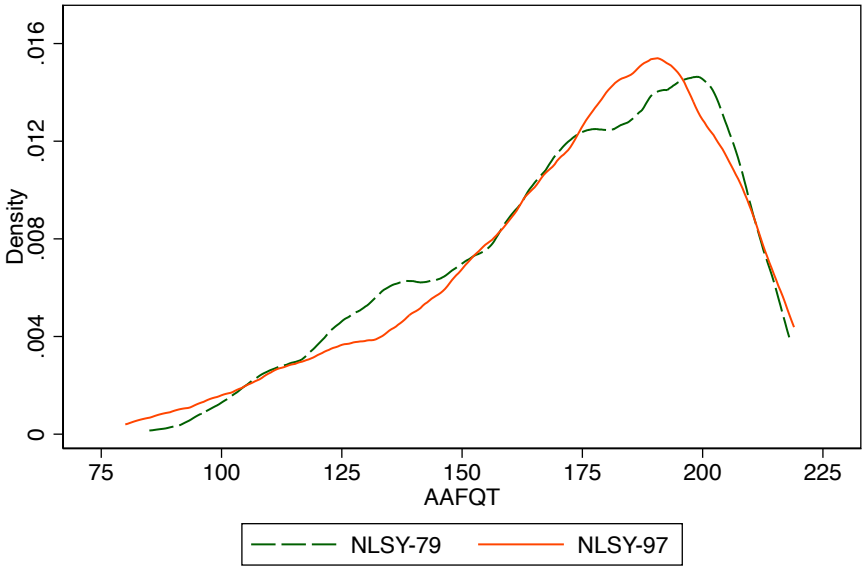

(b) White Non-Hispanic Women




*Note*: In each figure, we plot the average log wages against each value of AAFQT scores separately for the two cohorts. The wage observations are from ages 25–39. AAFQT scores are concorded by Altonji et al. (2012). The OLS estimates and the fitted lines are based on the univariate regression in Equation (1). See Appendix A Table A.3 for the full regression results without and with covariates. We multiply log wages by 100 for ease of display. BLS custom sample weights are used.

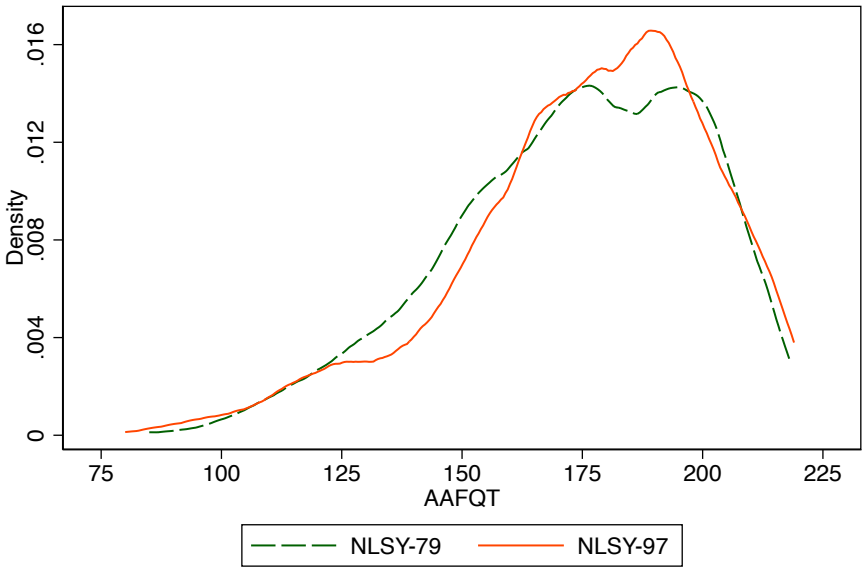
Figure 2: Adjusted AFQT Distribution for White Non-Hispanic Men and Women

(a) White Non-Hispanic Men




NLSY-79: Mean = 172.7, SD = 29.0, Skewness = 0.62, p1 = 104, p5 = 118, p10 = 130, p50 = 178
NLSY-97: Mean = 173.4, SD = 29.8, Skewness = 0.81, p1 = 94, p5 = 113, p10 = 129, p50 = 179

(b) White Non-Hispanic Women

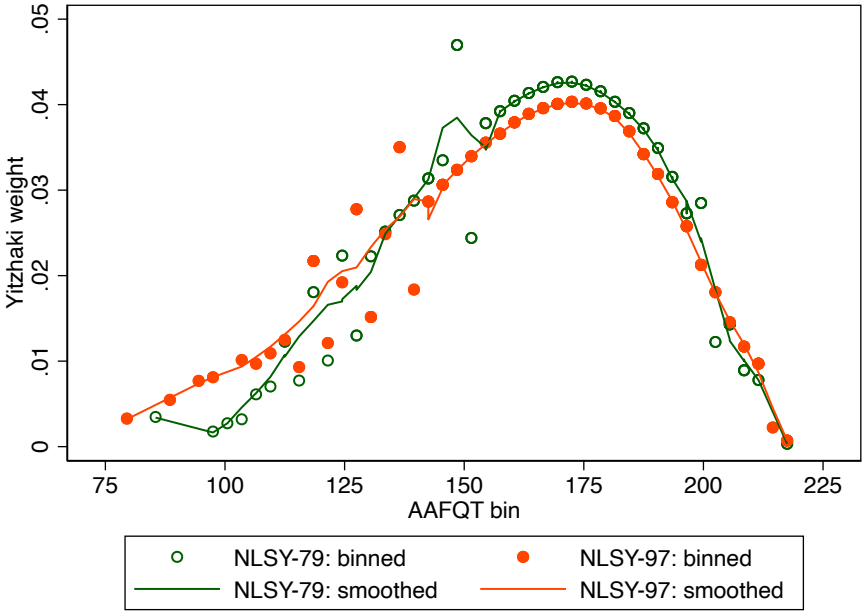



NLSY-79: Mean = 173.3, SD = 26.0, Skewness = 0.55, p1 = 106, p5 = 124, p10 = 137, p50 = 176
NLSY-97: Mean = 175.4, SD = 26.2, Skewness = 0.77, p1 = 103, p5 = 122, p10 = 140, p50 = 179

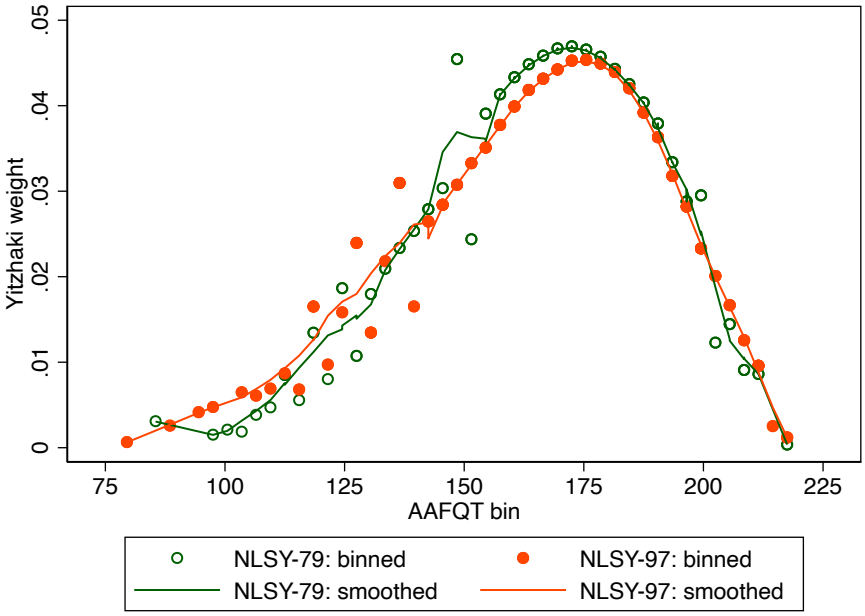
*Note*: In each figure, we plot the density of AAFQT scores separately for the two cohorts. AAFQT scores are concorded by Altonji et al. (2012). At the bottom of each figure, we present distribution statistics for each cohort. See Appendix A Tables A.1 and A.2 for a full list of distribution statistics and tests for cross-cohort changes. BLS custom sample weights are used.

Figure 3: Yitzhaki Weights

(a) White Non-Hispanic Men



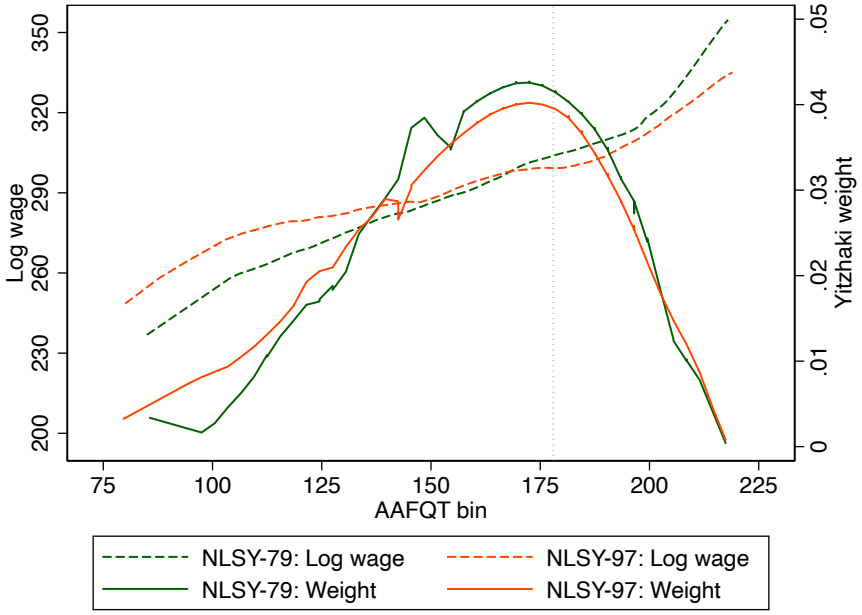

(b) White Non-Hispanic Women



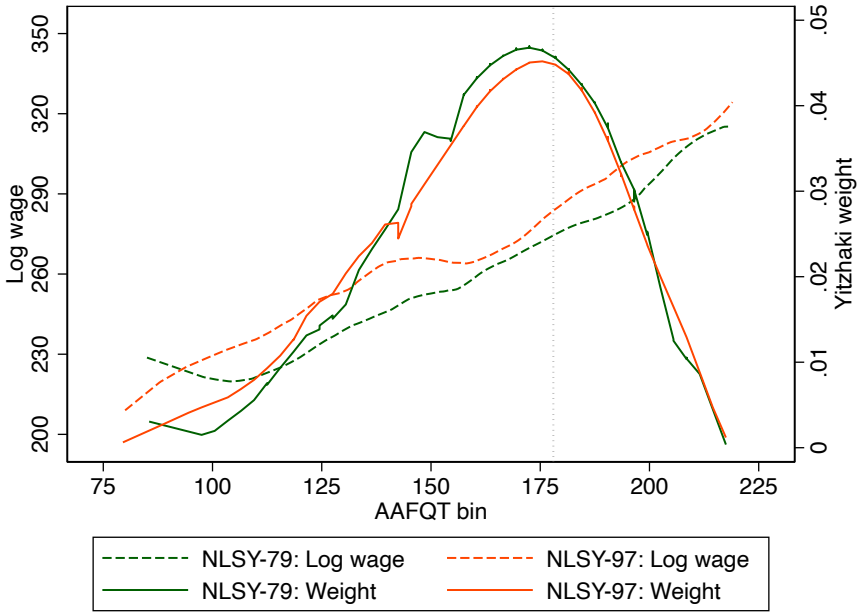

*Note*: In each figure, we plot the Yitzhaki weights using the formula in equation (4), separately for the two cohorts. The binned scatterplots are created by summing up Yitzhaki weights in each bin, which contains three consecutive AAFQT points. The smoothed curve is created using Locally Weighted Scatterplot Smoothing (LOWESS) with a tricube weighting function and a bandwidth of 0.1. See Appendix C for how BLS custom sample weights are incorporated in the Yitzhaki decomposition.

Figure 4: Smoothed Yitzhaki Weights and Local Linear Regression

(a) White Non-Hispanic Men



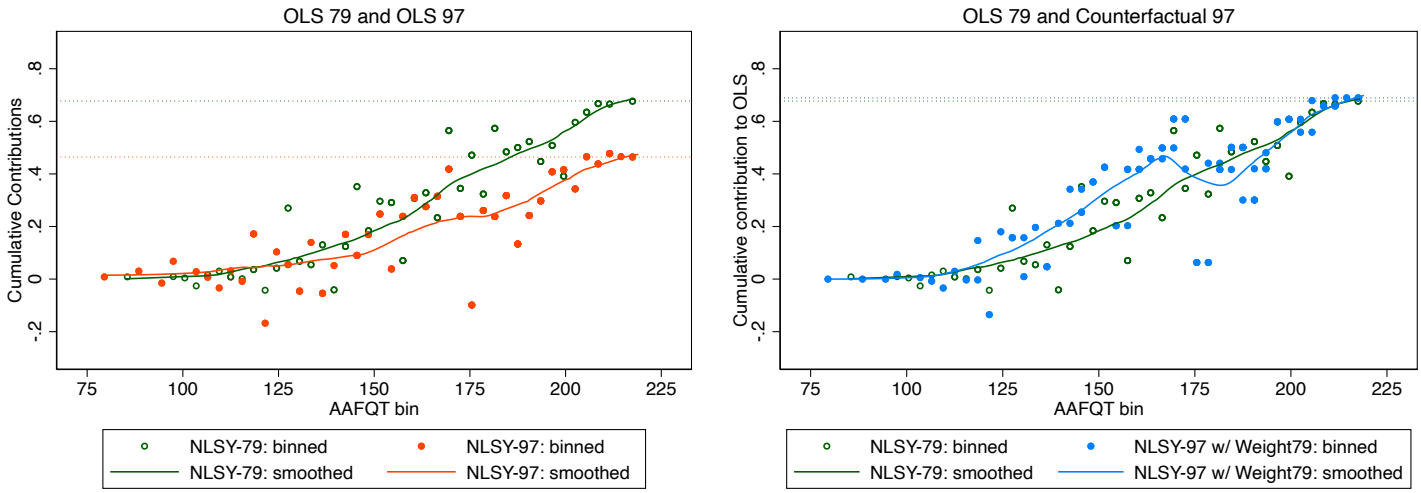

(b) White Non-Hispanic Women



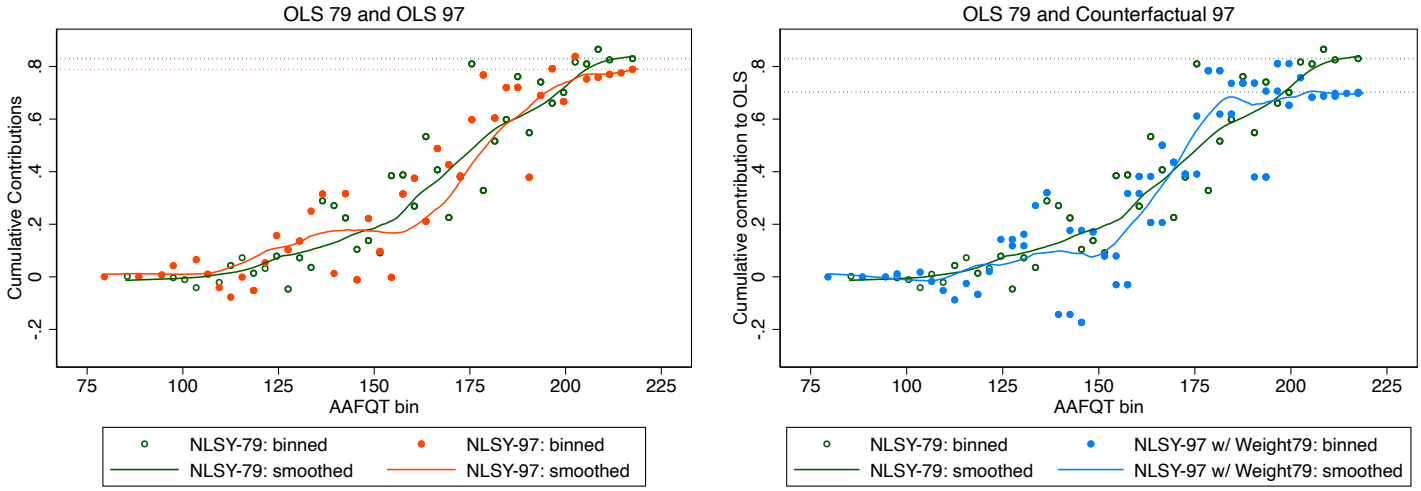

*Note*: In each figure, we overlay the smoothed Yitzhaki weights (right axis) with smoothed pairwise slopes (left axis), separately for the two cohorts. The smoothed curves are created using Locally Weighted Scatterplot Smoothing (LOWESS) with a tricube weighting function. To make the graphs more readable, a bandwidth of 0.1 and 0.3 is used for smoothing the weights and the slopes respectively. See Appendix C for how BLS custom sample weights are incorporated in the Yitzhaki decomposition.

Figure 5: Cumulative Contributions to Actual & Counterfactual OLS Estimates

(a) White Non-Hispanic Men




(b) White Non-Hispanic Women




*Note*: In each figure, we graph the progressive sum of the Yitzhaki decomposition from equation (5), starting with the lowest AAFQT score until the entire sum is calculated (producing the OLS estimate). The top left and bottom left panels plot the progressive sum for the actual OLS estimates of the two cohorts. The top right and bottom right panels plot the counterfactual estimate for NLSY–97 (using NLSY–97 weights) with the actual NLSY–79 OLS estimate. We graph the progressive sum by three-point bins and use a bandwidth of 0.3 for smoothing. See Appendix C for how BLS custom sample weights are incorporated in the Yitzhaki decomposition.

# Appendices

## A Supplemental Tables and Figures using AAFQT scores and subsections

Table A.1: Distribution Statistics of AAFQT score

| | White Men | | White Women | |
|---|---|---|---|---|
| | NLSY–79 | NLSY–97 | NLSY–79 | NLSY–97 |
| Mean | 172.7 | 173.4 | 173.3 | 175.4 |
| S.D. | 29.0 | 29.8 | 26.0 | 26.2 |
| Skewness | 0.62 | 0.81 | 0.55 | 0.77 |
| Kurtosis | 2.57 | 3.12 | 2.79 | 3.39 |
| p1 | 104 | 94 | 106 | 103 |
| p5 | 118 | 113 | 124 | 122 |
| p10 | 130 | 129 | 137 | 140 |
| p25 | 154 | 155 | 156 | 161 |
| p50 | 178 | 179 | 176 | 179 |
| p75 | 197 | 196 | 194 | 194 |
| p90 | 207 | 208 | 205 | 206 |
| p95 | 210 | 213 | 210 | 212 |
| p99 | 217 | 217 | 217 | 217 |
| Test of Equal Distribution | $p < 0.01$ | | $p < 0.01$ | |

*Note*: Distribution statistics of AAFQT scores (concorded by Altonji et al. (2012)) are presented for the two NLSY cohorts, and for white non-Hispanic men and white non-Hispanic women. BLS custom sample weights are used. For the test of equal distributions between NLSY–79 and NLSY–97, we report p-values of the chi-squared test.

Table A.2: Bootstrap Results, White Non-Hispanic Men and Women

| | Observed difference NLSY97 – NLSY79 | Bootstrap std. err. | z | p-value | Normal based [95% conf. | interval] |
|---|---|---|---|---|---|---|
| **White Non-Hispanic Men** | | | | | | |
| Mean | 0.77 | 0.85 | 0.91 | 0.36 | -0.89 | 2.44 |
| S.D. | 0.82 | 0.57 | 1.44 | 0.15 | -0.30 | 1.94 |
| Skewness | 0.19 | 0.05 | 3.82 | **0.00** | 0.09 | 0.29 |
| Kurtosis | 0.55 | 0.13 | 4.25 | **0.00** | 0.30 | 0.80 |
| p1 | -10 | 3.30 | -3.03 | **0.00** | -16.46 | -3.54 |
| p5 | -5 | 2.34 | -2.14 | **0.03** | -9.58 | -0.42 |
| p10 | -1 | 1.91 | -0.52 | 0.60 | -4.74 | 2.74 |
| p25 | 1 | 1.86 | 0.54 | 0.59 | -2.66 | 4.66 |
| p50 | 1 | 1.19 | 0.84 | 0.40 | -1.33 | 3.33 |
| p75 | -1 | 0.98 | -1.02 | 0.31 | -2.92 | 0.92 |
| p90 | 1 | 1.06 | 0.94 | 0.35 | -1.08 | 3.08 |
| p95 | 3 | 0.56 | 5.32 | **0.00** | 1.89 | 4.11 |
| p99 | 0 | 0.48 | 0.00 | 1.00 | -0.95 | 0.95 |
| **White Non-Hispanic Women** | | | | | | |
| Mean | 2.09 | 0.75 | 2.79 | **0.01** | 0.62 | 3.56 |
| S.D. | 0.16 | 0.57 | 0.28 | 0.78 | -0.95 | 1.27 |
| Skewness | 0.22 | 0.06 | 3.78 | **0.00** | 0.11 | 0.34 |
| Kurtosis | 0.59 | 0.16 | 3.73 | **0.00** | 0.28 | 0.90 |
| p1 | -3 | 3.40 | -0.88 | 0.78 | -0.95 | 1.27 |
| p5 | -2 | 2.59 | -0.77 | 0.44 | -9.67 | 3.67 |
| p10 | 3 | 2.54 | 1.18 | 0.24 | -1.97 | 7.97 |
| p25 | 5 | 1.26 | 3.98 | **0.00** | 2.54 | 7.46 |
| p50 | 3 | 1.04 | 2.88 | **0.00** | 0.96 | 5.04 |
| p75 | 0 | 1.06 | 0.00 | 1.00 | -2.08 | 2.08 |
| p90 | 1 | 1.17 | 0.85 | 0.39 | -1.29 | 3.29 |
| p95 | 2 | 0.82 | 2.43 | **0.02** | 0.38 | 3.62 |
| p99 | 0 | 0.97 | 0.00 | 1.00 | -1.90 | 1.90 |

*Note*: We bootstrap 2,000 times to construct standard errors, p-values, and confidence intervals for the cross-cohort difference in distribution statistics of AAFQT scores (concorded by Altonji et al. (2012)). We present the results separately for white non-Hispanic men and white non-Hispanic women. BLS custom sample weights are used.

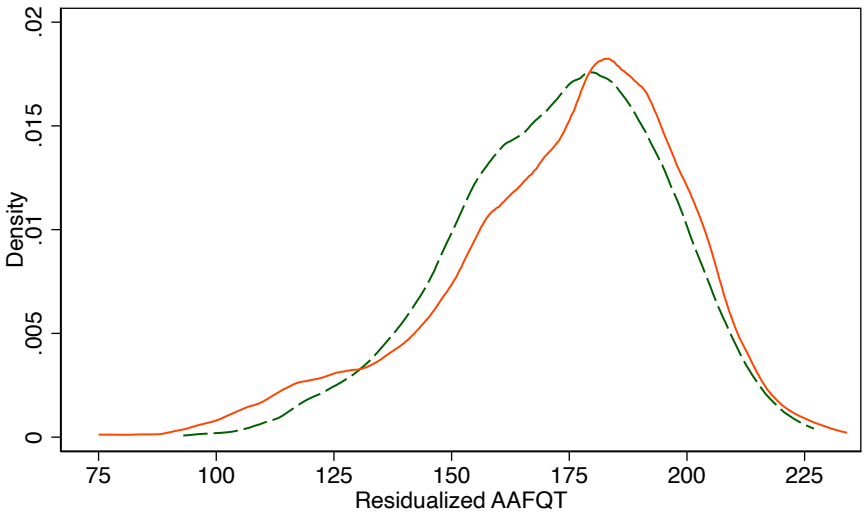
Table A.3: OLS estimates for White Men and Women

| | White Men | | White Women | |
|---|---|---|---|---|
| | NLSY–79 | NLSY–97 | NLSY–79 | NLSY–97 |
| **Panel A: Univariate Regression** | | | | |
| AAFQT | 0.677*** | 0.463*** | 0.830*** | 0.789*** |
| | [0.035] | [0.043] | [0.036] | [0.048] |
| Change from NLSY–79 | | –0.212*** | | –0.041 |
| | | [0.055] | | [0.060] |
| Obs | 2099 | 1584 | 2191 | 1488 |
| **Panel B: Control for Non-cognitive & Social Skills** | | | | |
| AAFQT | 0.605*** | 0.452*** | 0.761*** | 0.768*** |
| | [0.036] | [0.043] | [0.038] | [0.047] |
| Change from NLSY–79 | | –0.153*** | | 0.007 |
| | | [0.056] | | [0.061] |
| Obs | 2099 | 1584 | 2191 | 1488 |
| **Panel C: Control for Education** | | | | |
| AAFQT | 0.404*** | 0.161*** | 0.437*** | 0.315*** |
| | [0.042] | [0.051] | [0.044] | [0.054] |
| Change from NLSY–79 | | –0.243*** | | –0.122* |
| | | [0.067] | | [0.069] |
| Obs | 2099 | 1584 | 2191 | 1488 |
| **Panel D: Control for Non-cognitive & Social Skills & Education** | | | | |
| AAFQT | 0.363*** | 0.174*** | 0.406*** | 0.321*** |
| | [0.043] | [0.050] | [0.044] | [0.053] |
| Change from NLSY–79 | | –0.188*** | | –0.085 |
| | | [0.066] | | [0.069] |
| Obs | 2099 | 1584 | 2191 | 1488 |

*Note*: We present OLS estimates of the wage returns to AAFQT scores (concorded by Altonji et al. (2012)). Panel A presents results for the univariate regression in equation (1). Panel B controls for measures of non-cognitive skills and social skills (created by Deming (2017)). Panel C controls for the highest grade completed. Panel D controls for both measures of non-cognitive and social skills, and the highest grade completed. We present results separately for NLSY–79 and NLSY–97, and for white non-Hispanic men and white non-Hispanic women. We also present the estimated change in the OLS estimates across cohorts. BLS custom sample weights are used. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

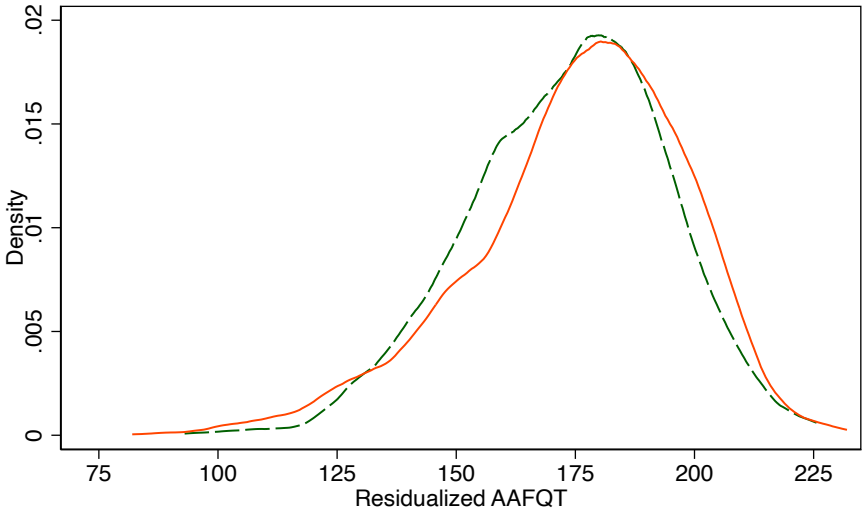
Figure A.1: Residualized AAFQT Distribution for White Non-Hispanic Men and Women

(a) White Non-Hispanic Men




NLSY-79: Mean = 172.6, SD = 22.2, Skewness = 0.37, Kurtosis: 2.90, p1 = 116, p5 = 133, p50 = 175
NLSY-97: Mean = 173.4, SD = 25.4, Skewness = 0.75, Kurtosis: 3.47, p1 = 104, p5 = 122, p50 = 178

(b) White Non-Hispanic Women

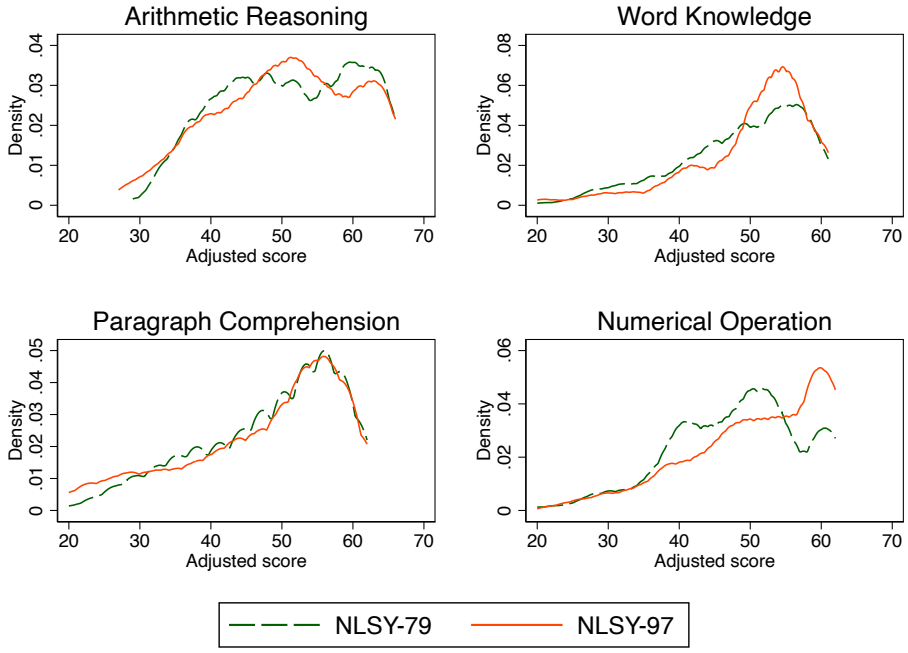



NLSY-79: Mean = 173.3, SD = 20.6, Skewness = 0.32, Kurtosis: 3.01, p1 = 124, p5 = 138, p50 = 175
NLSY-97: Mean = 175.4, SD = 22.6, Skewness = 0.66, Kurtosis: 3.49, p1 = 110, p5 = 132, p50 = 178

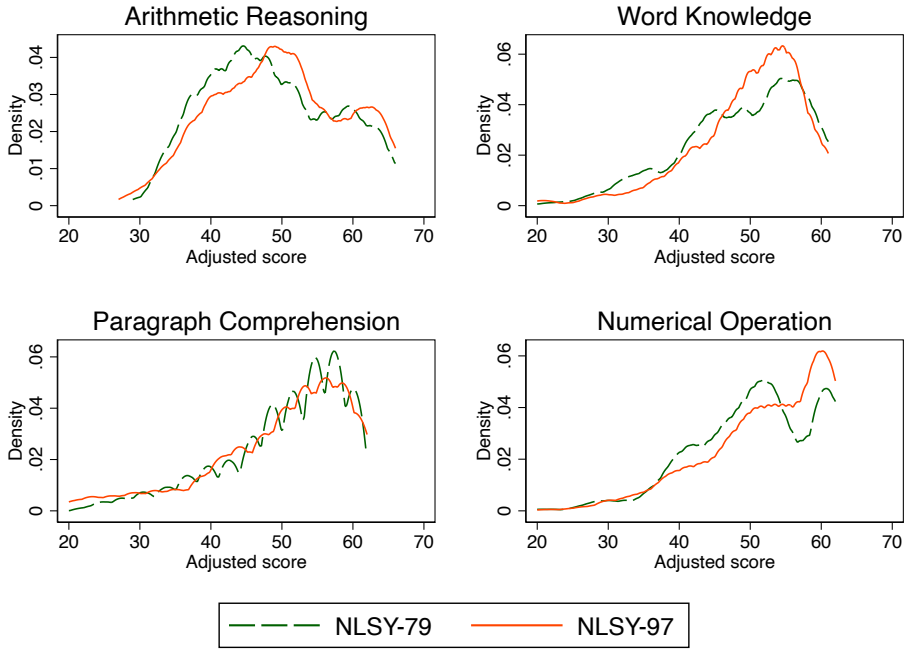
*Note*: In each figure, we plot the density of residualized AAFQT scores separately for the two cohorts. AAFQT scores (Altonji et al. 2012) are residualized by measures of non-cognitive and social skills (Deming 2017), and the highest grade completed. At the bottom of each figure, we present distribution statistics for each cohort. BLS custom sample weights are used.

Figure A.2: Adjusted ASVAB subsection scores for White Non-Hispanic Men and Women

(a) White Non-Hispanic Men




(b) White Non-Hispanic Women




*Note*: In each figure, we plot the density of four ASVAB subsection scores, separately for the two cohorts. The four subsections (Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, and Numerical Operation) are used to create the AFQT score. We adjust each subsection score following Altonji et al. (2012). See Appendix B for a discussion of other versions of the AFQT score. BLS custom sample weights are used.

# B Notes on Armed Forces Qualification Test (AFQT)

This appendix describes the background and essential details of the Armed Forces Qualification Test (AFQT) score. This is a collection of information from different sources: a manuscript by Altonji et al. (2009), a technical bulletin by Defense Manpower Data Center (2006) which includes several chapters from Sands et al. (1997), annual reports on population representation in the military services (e.g. Quester et al. 2017), and the introduction on the NLSY website (Bureau of Labor Statistics; Bureau of Labor Statistics 1992; Bureau of Labor Statistics). The AFQT score is constructed based on multiple sections of the Armed Services Vocational Aptitude Battery (ASVAB), a set of tests developed by the Department of Defense (DOD) for screening military enlistees and assigning them to military occupations. Economists have long been using the AFQT score, as well as other tests in the ASVAB, to measure skills and abilities (Neal and Johnson, 1996; Heckman et al. 2006; Altonji et al., 2012; Prada et al. 2017). This is facilitated by the data of the NLSY–79 and the NLSY–97, as survey respondents took the ASVAB test.

### History of the ASVAB and the NLSY

The ASVAB was first introduced in 1968 and has undergone several adjustments and revisions since. One important adjustment has been to update the norms of the ASVAB (Defense Manpower Data Center 2006). In practice, the military sets a goal of selecting only applicants who rank higher than X% of American youth in the national distribution of ability and skill. Different military branches have different qualification cutoffs, and many of them now use a cutoff of 30%–40% for applicants with a high school diploma. Recruiters therefore need to know how the Xth percentile youth in the national population scores on the ASVAB in order to compare military applicants to this benchmark. To ensure that contemporary applicants are always compared to an appropriate benchmark, the benchmark must be updated over time.[19]

In 1979, after questioning the appropriateness of using the World War II reference population as the benchmark, the DOD and Congress decided to let the NLSY–79 respondents take the ASVAB, and the DOD used their scores as the new benchmark for military enlistees. The NLSY–79 served as a natural group to benchmark the ASVAB because it is a nationally representative sample of the cohort of Americans born 1957–1964. The respondents took the ASVAB in the

19. For example, in 2015, the military services typically do not accept applicants who score in the bottom 30th percentile in the national AFQT distribution. In addition, DOD requires that at least 60 percent of new enlistees score at the 50th percentile or higher in the national AFQT distribution (Quester et al. 2017).

summer and fall of 1980, following the standard ASVAB procedures. This study of benchmarking the ASVAB using the NLSY–79 is called "Profile of American Youth (PAY–80)."

A major revision of the ASVAB occurred when it shifted from a paper-based test to a computer-based test. The military started to implement the computer-based ASVAB on a large scale in 1996–1997, after about two decades of research and evaluation. The NLSY–97 respondents took the computer-based test, while the NLSY–79 respondents took the paper-based test.

In the paper-based test, all respondents received the same set of questions. In the computer-based test, the next questions that respondents received depend on their answers to previous questions. For example, if a respondent answered a question correctly, then the next question becomes more difficult. This adaptive feature of the computer-based test means that different respondents can receive different sets of questions and with different orderings. The raw count of correct answers is therefore no longer directly comparable across respondents. Instead, item response theory (IRT) models are used to construct estimates of ability and skill (also called "thetas") for each respondent of the computer-based ASVAB. These IRT estimates are supposed to be comparable across respondents.[20]

Due to the test format change, the military needed a new benchmark for the computer-based ASVAB. As the NLSY–97 respondents were 12-17 when first interviewed in 1997, and some were deemed too young for the purpose of benchmarking military enlistees, two other nationally representative samples were identified to complete the computer-based ASVAB during the NLSY–97 screening process. The first sample, the Student Testing Program (STP), consisted of students who expected to be in grades 10–12 in the fall of 1997. Included were many respondents who also participated in the NLSY–97, as well as youth who refused to participate in or were not eligible for the NLSY–97. The second sample, the Enlistment Testing Program (ETP), was a nationally representative sample of youth aged 18–23 as of June 1997. The ASVAB performance of respondents in these two samples (again, which includes some NLSY–97 respondents) was then used to benchmark the computer-based ASVAB for the military.

20. Two sections, numerical operations and coding speed, in the computer-based ASVAB are administered in a non-adaptive format (that is, everyone answers the same questions in the same order). The scores of these two sections are therefore not "thetas" estimated from IRT. However, the two sections are still done on computers, so the scores are not directly comparable to the scores of the same sections but in paper format.

### Concordance of different formats of ASVAB

A practical issue coming from ASVAB's format change is how to concord the paper-based and computer-based test scores. This is of significant importance for the military because, ideally, the selection criteria into the Armed Forces should be held broadly consistent before and after the test format change. This is also extremely important for researchers because otherwise the AFQT score and the ASVAB subsection scores, as measures of skills and abilities, are not comparable between the NLSY–79 and the NLSY–97 cohorts (Altonji et al. 2012).

Daniel Segall, a researcher at the DOD specializing in psychometrics, developed a mapping between the paper-based and computer-based ASVAB scores (Segall 1997). He drew a sample of military applicants in two rounds, in 1988 (N=8,040) and from 1990 to 1992 (N=10,379). In each round, one-third of the participants were randomly assigned to take the paper-based ASVAB, and the other two–thirds took the computer-based ASVAB. Using the test performance of these military applicants, Segall created a mapping to link each computer-based ASVAB component score to a paper-based ASVAB component score. Since the computer-based ASVAB scores ("thetas" estimated from IRT models) are continuous and the paper-based ASVAB scores (counts of correct answers) are discrete by construction, Segall applied certain smoothing and grouping to the score distributions in the mapping procedure. For further technical details, see Segall (1997).

In their efforts to concord the AFQT score between the NLSY–79 and the NLSY–97, Altonji et al. (2012) relied heavily on Segall's mapping. Since the mapping is not publicly available, the authors sent the computer-based ASVAB subsection IRT scores in the NLSY–97 to Segall, who mapped the scores into paper-based scores so that they could directly be compared to the scores of the NLSY–79 respondents. With the scores from Segall in hand, the authors adjusted for one more important difference between the two NLSY cohorts: test-taking ages. The NLSY–79 respondents were around ages 15–23 and the NLSY–97 respondents were around ages 12–18 when they took the ASVAB. On average, ASVAB performance improves as people age, so it is critical to address the differential test-taking ages both within and across cohorts.

To construct the mapping across ages, the authors exploited the fact that both cohorts have a nontrivial share of respondents taking the ASVAB at age 16. Under the (somewhat strong) assumption that a person's *ranking* in the AFQT score distribution does not vary with age, the authors mapped a person at age X (which is not 16) to the score distribution of age 16 by their

ranking in the score distribution of age X. For example, if a youth in the NLSY–79 took the test at age 20 and ranked the 25th percentile within the AFQT score distribution of age 20, the youth will be mapped to have the 25th percentile score of the age-16 distribution in the NLSY–79. This relies on the assumption that whoever at the 25th percentile in the score distribution at age 16 will remain at the 25th percentile at age 20. Whether this rank-invariant assumption holds remains to be analyzed and tested. More details can be found in Altonji et al. (2009).[21]

In Figures B.1, we graph the distribution of the original IRT-based scores (called "thetas") for three of the four different ASVAB sections for the NLSY–97 cohort. As a reminder, these scores are original to the NLSY–97 data and were not further processed to concord to the 1979 cohort. We also graph the IRT-based scores for the NLSY–79 cohort that were constructed after-the-fact from the original paper-based tests by researchers from the Ohio State University.[22]

The divergence and increased skewness of scores in the NLSY–97 IRT-based scores relative to NLSY–79 scores are visible in different sections (especially Word Knowledge and Paragraph Comprehension) of Figures B.1, suggesting that changes in the AAFQT scores across the cohorts do not seem to be a function of the concordance that was done to create AAFQT-equivalent scores for the 1997 cohort. That said, it is not obvious whether the "thetas" are directly comparable between the NLSY–79 and the NLSY–97, for at least three reasons. First, we do not know if the IRT models and estimation methods used for the two cohorts are the same. Second, even if the models and methods are the same, the raw data imported to the models may still not be comparable due to the different test formats used. Third, there is a strong hint that something is amiss in the IRT scores for the NLSY–97.

Figure B.2 plots the standard errors of the estimated IRT scores ("thetas") for different ASVAB sections. As pointed out in past studies (Schofield 2014; Jacob et al. 2016), "thetas" in IRT models are more precisely estimated for the middle of the distribution, leading to a non-classical measurement error structure with larger errors at the tails. This particular measurement error issue is a feature of the IRT, generally, and not just for NLSY datasets (Jacob and Rothstein, 2016). The error structure of the thetas in the NLSY–79 is generally symmetric, and the standard errors are minimized toward the middle of the distribution, exactly as expected from the IRT model. What is odd is that in the NLSY–97, the distribution of the standard

21. The adjusted AFQT score created by Altonji et al. is what we referred to as the AAFQT score in the main text.

22. See Ing et al. (2012) for details. IRT-based scores are not available for the numerical operations section of ASVAB in the NLSY–79.

errors is not symmetric nor minimized around the middle of the theta distribution.[23]

### Different versions of AFQT score

The ASVAB has multiple sections. The AFQT score is a sum of scores from four ASVAB sections. By picking scores from different sections, two versions of the AFQT score have been constructed and used. The AFQT–80, probably the most widely used AFQT score, is the summation of arithmetic reasoning (AR), numerical operations (NO), paragraph comprehension (PC), and word knowledge (WK). The formula is AFQT–80 = AR + 0.5*NO + PC + WK.

In 1989, according to the NLS website, it was realized that the numerical operations section had some design inconsistencies that resulted in unreliable scores (Bureau of Labor Statistics). The DOD decided to replace numerical operations with math knowledge (MK) in the construction of the AFQT score. The new score is called AFQT–89. The formula is AFQT–89 = AR + MK + 2*VE. Verbal composite (VE) can be seen as a weighted average of PC and WK with unequal weights. WK receives a higher weight because there are more questions in the WK section.

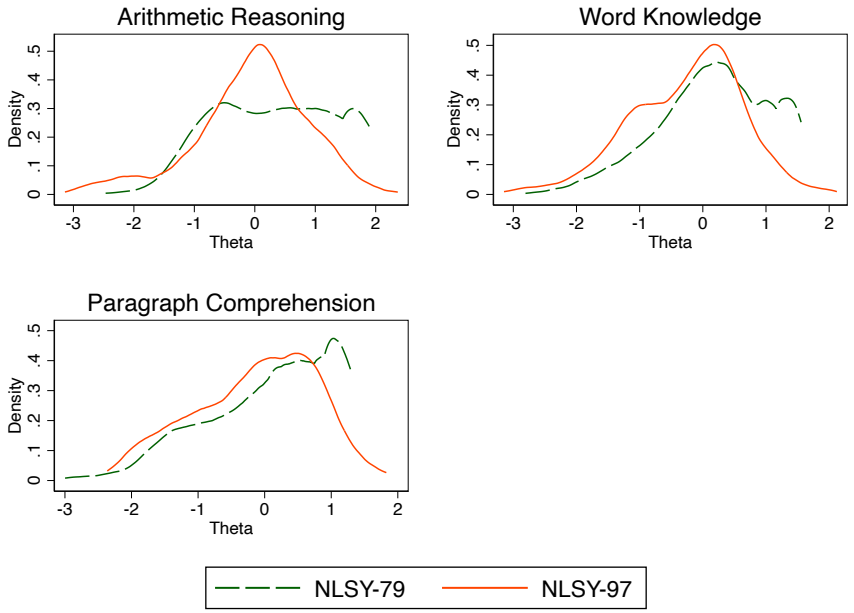
Different studies have used different versions of the AFQT score. Neal and Johnson (1996) used the AFQT–89 in the published version of their paper, and noted that results are similar using the AFQT–80. Altonji et al. (2012) used the AFQT–80 and created the adjusted score that is supposed to be comparable between the NLSY–79 and the NLSY–97. More recent studies have been using their adjusted AFQT–80 score (Castex et al. 2014; Deming 2017). Although Altonji et al. (2012) only did the adjustment for the AFQT–80, their method can be applied to the AFQT–89 and/or the ASVAB subsection scores.

In the paper, we use the AFQT–80 score in order to be able to use the AAFQT scores across the two cohorts, consistent with Altonji et al. (2012), Castex et al. (2014), and Deming (2017), who also compare the NLSY–79 with the NLSY–97.

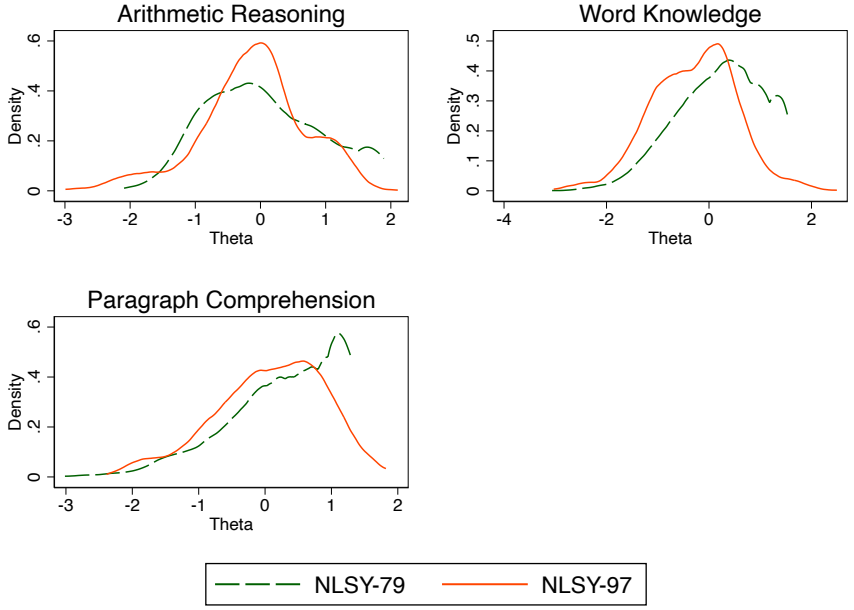
23. We thank Dan Black for pointing this out to us.

Figure B.1: IRT-based ASVAB subsection scores for White Non-Hispanic Men and Women

(a) White Non-Hispanic Men



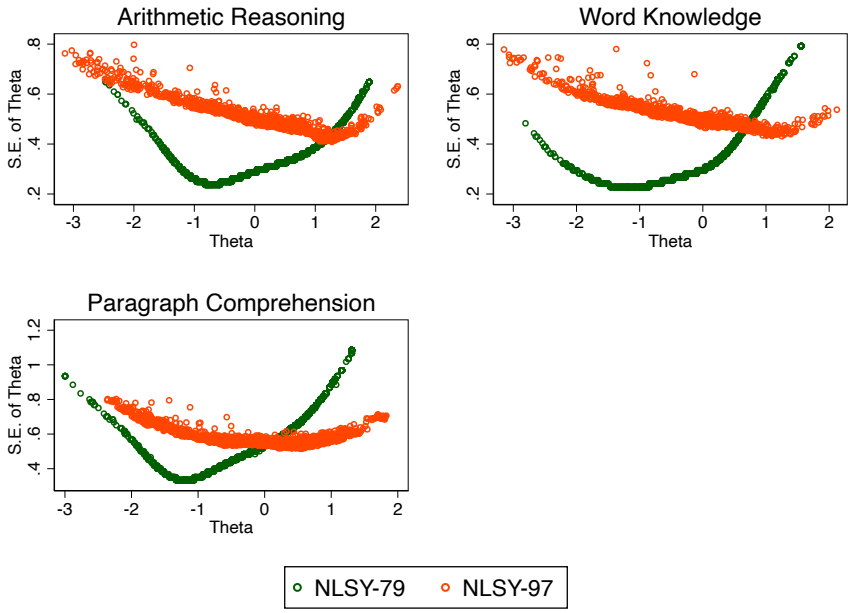

(b) White Non-Hispanic Women



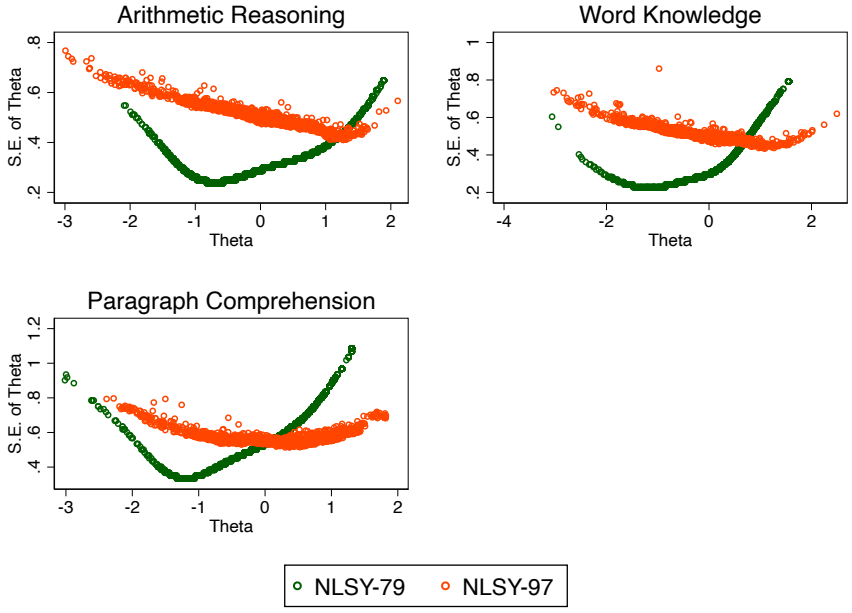

*Note*: In each figure, we plot the density of IRT-based ASVAB subsection scores ("thetas"), separately for the two cohorts. The NLSY–97 scores are original to the NLSY–97 data and not further processed to concord to the 1979 cohort. The NLSY–79 scores are constructed after-the-fact from the original paper-based tests by researchers from the Ohio State University (Ing et al. 2012). The IRT-based score for the Numerical Operation subsection is not available for the NLSY–79. BLS custom sample weights are used.

Figure B.2: Standard Errors of IRT-based ASVAB subsection scores for White Non-Hispanic Men and Women

(a) White Non-Hispanic Men




(b) White Non-Hispanic Women




*Note*: In each figure, we plot the standard errors of the estimated IRT scores ("thetas") for different ASVAB sections, separately for the two cohorts. The NLSY–97 scores are original to the NLSY–97 data and not further processed to concord to the 1979 cohort. The NLSY–79 scores are constructed after-the-fact from the original paper-based tests by researchers from the Ohio State University (Ing et al. 2012). The IRT-based score for the Numerical Operation subsection is not available for the NLSY–79. BLS custom sample weights are used.

# C Yitzhaki Decomposition with Weights

For simplicity, Yitzhaki's decomposition formula (Proposition 1 in Yitzhaki 1996) assumes that each value of X has only one observation. In practice, each value of X can be linked to multiple observations in the data. As suggested by Yitzhaki (1996), all observations with the same $X$ should be aggregated, leading to a *grouped* dataset in which the outcome $Y$ is averaged within each value of $X$. In a univariate model, we can recap the original OLS estimate by using the grouped data and weighting the grouped regression with group size. In addition, each observation in the data can represent multiple observations in the population. It is sometimes more appropriate to use Weighted Least Squares (WLS) with sample weights than OLS (Solon, Haider, and Wooldridge 2015). In this appendix, we extend Yitzhaki's formula to allow for these two types of weights.

Following Yitzhaki's notation, let $y_i$ and $x_i$ $(i = 1, \ldots, n)$ be observations and ranked in the increasing order of X. An important simplification that Yitzhaki makes is that $\Delta x_i = x_{i+1} - x_i > 0$, i.e., each value of X has only one observation. Here we extend Yitzhaki's set-up and allow there to be duplicate observations. Let there be $N_i$ duplicate observations for $(x_i, y_i)$. Let $b_i = \Delta y_i / \Delta x_i$ be the slope of two adjacent values of $X$.

Like Yitzhaki (1996), we are interested in decomposing the point estimate. Given this, the two types of weights mentioned above are both equivalent to adding duplicate observations. The distinction between the two cases is the construction of $y_i$. In the first case (without sample weights), $y_i$ is the average of all $Y$ linked to $x_i$. In the second case (with sample weights, i.e. WLS), $y_i$ is the weighted average of all $Y$ linked to $x_i$.

With duplicate observations, the sample covariance of $Y$ and $X$ can be expressed as:

$$\begin{aligned} \text{cov}(y, x) &= \tfrac{1}{2n(n-1)} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} N_i N_j \left(x_i - x_j\right)\left(y_i - y_j\right) \\ &= \tfrac{1}{n(n-1)} \textstyle\sum_{i=2}^{n} \sum_{j=1}^{i-1} N_i N_j \left(x_i - x_j\right)\left(y_i - y_j\right) \end{aligned}$$

Note that when there are no duplicate observations ($N_i = 1, \text{for all} i$), the expression becomes $\text{cov}(y, x) = \frac{1}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(x_i - x_j\right)\left(y_i - y_j\right)$, which is what Yitzhaki presents in Proposition 1 (Yitzhaki 1996).

Like Yitzhaki, we substitute $(x_i - x_j) = \Delta x_i + \Delta x_{i+1} + \cdots + \Delta x_{j-1}$ and $(y_i - y_j) = b_i \Delta x_i +$

$b_{i+1}\Delta x_{i+1} + \cdots + b_{j-1}\Delta x_{j-1}$. After collecting like terms, we get:

$$\operatorname{cov}(y,x) = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \left\{ \sum_{j=i}^{n-1} (N_1 + \cdots N_i)(N_{j+1} + \cdots + N_n)\Delta x_j + \sum_{j=1}^{i-1} (N_{i+1} + \cdots + N_n)(N_1 + \cdots N_j)\Delta x_j \right\} \Delta x_i b_i$$

Again, when there are no duplicate observations, the expression simplifies to $\operatorname{cov}(y,x) = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \left\{ \sum_{j=i}^{n-1} i(n-j)\Delta x_j + \sum_{j=1}^{i-1} j(n-i)\Delta x_j \right\} \Delta x_i b_i$, as in Yitzhaki (1996).

Similarly, we can get the expression for $\operatorname{cov}(x,x)$:

$$\operatorname{cov}(\mathrm{x},\mathrm{x}) = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \left\{ \sum_{j=i}^{n-1} (N_1 + \cdots N_i)(N_{j+1} + \cdots + N_n)\Delta x_j + \sum_{j=1}^{i-1} (N_{i+1} + \cdots + N_n)(N_1 + \cdots N_j)\Delta x_j \right\} \Delta x_i$$

We can then write down the OLS/WLS estimator as a weighted average of $b_i$:

$$\widetilde{b_{OLS/WLS}} = \frac{\operatorname{cov}(y,x)}{\operatorname{cov}(x,x)} = w_i b_i, \qquad \text{where } \sum_{i=1}^{n-1} w_i = 1$$

where the weight $w_i$ is:

$$w_i = \frac{\left\{ \sum_{j=i}^{n-1} (N_1 + \cdots N_i)(N_{j+1} + \cdots + N_n)\Delta x_j + \sum_{j=1}^{i-1} (N_{i+1} + \cdots + N_n)(N_1 + \cdots N_j)\Delta x_j \right\} \Delta x_i}{\sum_{k=1}^{n-1} \left\{ \sum_{j=i}^{n-1} (N_1 + \cdots N_k)(N_{j+1} + \cdots + N_n)\Delta x_j + \sum_{j=1}^{k-1} (N_{k+1} + \cdots + N_n)(N_1 + \cdots N_j)\Delta x_j \right\} \Delta x_k}$$

The numerator of $w_i$ can be written equivalently in a more intuitive expression:

$$\left( \sum_{j=1}^{i} N_j \right) \left( \sum_{j=i+1}^{n} N_j \right) \cdot \left( \frac{\sum_{j=i+1}^{n} N_j x_j}{\sum_{j=i+1}^{n} N_j} - \frac{\sum_{j=1}^{i} N_j x_j}{\sum_{j=1}^{i} N_j} \right) \cdot \Delta x_i$$

As a comparison, the continuous version of the weighting function $w(x)$ is:

$$w(x) = \frac{F_X(x) \cdot (1 - F_X(x))}{\sigma_X^2} \{ E(X \mid X > x) - E(X \mid X \leq x) \}$$

The first term in the discrete weighting function $\left(\sum_{j=1}^{i} N_j\right)\left(\sum_{j=i+1}^{n} N_j\right)$ matches with $F_X(x) \cdot (1 - F_X(x))$ in the continuous weighting function. The second term in the discrete weighting function $\frac{\Sigma_{j=i+1}^{n} N_j x_j}{\sum_{j=i+1}^{n} N_j} - \frac{\Sigma_{j=1}^{l} N_j x_j}{\sum_{j=1}^{i} N_j}$ matches with $E(X \mid X > x) - E(X \mid X \leq x)$ in the continuous weighting function. Compared to the case with no duplicate observations, here both the cumulative density and the conditional expected value are expressed in a weighted form.

# D Evidence from NELS

In the absence of any way to formally correct for potential measurement errors in AAFQT scores in the NLSYs, in this appendix, we provide additional evidence from other data sources. We use two nationally representative longitudinal data sets from the National Center for Education Statistics (NCES), the National Education Longitudinal Study of 1988 (NELS:88) and the Educational Longitudinal Study of 2002 (ELS:02). The NELS:88 cohort was first surveyed in 1988 as 8th graders and the ELS:02 cohort was first surveyed in 2002 as 10th graders. The NELS:88 cohort is 7–11 years older than the NLSY–97, and the ELS:02 cohort is 1–5 years younger than the NLSY–97.

While the NCES datasets do not contain AAFQT scores, we use respondents' math test scores from the senior year of high school as a measure of cognitive skills, and the (log of) hourly earnings eight years after high school as the wage measure.[24] We choose the NELS:88 and the ELS:02 for a cross-cohort comparison because the tests in ELS:02 are adapted from NELS:88 and they share many test items by design. Based on these shared test items and using the IRT method, ELS:02 constructs a NELS-equated math score that basically tells how many questions an ELS:02 student would have answered correctly had they taken the NELS:88 test.[25]

Table D.2 compares the 12th-grade math score distributions between the two NCES cohorts, separately for white men and white women. From NELS:88 to ELS:02, the mean of math scores has gone up while the distribution has become more left-skewed. As can be more clearly seen in Figure D.1, there is also a hollowing out in the low-to-middle part of the math score distribution, for both white men and white women. All of these patterns are also present in the evolution of the AAFQT score distributions across the two cohorts in the NLSY, even though the cohorts in the NLSY and NCES do not overlap.

That said, while there are important similarities between the math score distribution of the NELSs and the AAFQT score distribution of the NLSYs, there are also differences. In particular, we do not find an increasing mass of low math scorers in the ELS:02 as in the NLSY–97. It is unclear, however, how much of this difference is due to different samples, different test formats, or different test score construction processes.

24. Weinberger (2014) uses weekly earnings as the main labor market outcome measure. We use hourly earnings to stay consistent with our baseline analysis of NLSY.

25. The IRT-based math score in NCES can be subject to measurement errors. But there is nothing to suggest that measurement issues with IRT in the NCES would lead to exactly the same pattern of measurement error as AAFQT scores in the NLSY–97.

We examine how changes in the distribution of math scores across the two cohorts contribute to changing OLS estimates of wage returns to math scores. To do this, in Table D.1, we first present the OLS estimates of univariate regressions of log hourly earnings on math scores, separately for white men and white women.

The wage returns to math scores are positive and significant for both cohorts and for both men and women. Notably, the change in the OLS estimate across the two NCES cohorts is statistically insignificant. If anything, there is an increase in the OLS estimate, which is consistent with what Weinberger (2014) has documented by comparing NELS:88 with an older cohort.

We then perform the Yithaki decomposition on the baseline OLS estimates. Figure D.2 plots the Yithaki weights together with the smoothed local linear regressions and Figure D.3 plots the cumulative contributions to OLS estimates, separately for the two NCES cohorts and for white men and white women. First, the weights shift to the right across cohorts, but the shift mainly happens for high math scorers as compared to low math scorers (especially for white men). As we discussed in deriving the Yitzhaki decomposition, this is solely a mechanical result of changes in the math distributions.

Second, the pairwise slopes show more nonlinearities for the sample of white men than women, with flat regions in the mid-low range of the math score distribution. This is true for both NCES cohorts, and is similar to what we found for AAFQT scores in the NLSY–97 (but not in the NLSY-79). In contrast, the slopes seem broadly positive and linear for the sample of white women in both NELS:88 and ELS:02, as with both cohorts of NLSY–97 data. Given that the birth years of the two NCES cohorts span those of the NLSY–97, it is perhaps not surprising that the wage returns to measured cognitive ability more closely match those in the NLSY–97 than those in the older NLSY cohort.

Regardless of the differences between the two data sources in the samples and in the measures of cognitive abilities used, we find broad similarities in the evolution of cognitive test scores across cohorts. We also find wage returns to measured cognitive ability in the NELS that are consistent with the younger (and more similarly aged) NLSY–97. We read this additional evidence as supporting the view that measurement error alone in AAFQT scores cannot be driving changes in the test score distributions across NLSY cohorts, or associated changes in wage returns.

Table D.1: Returns to 12th-grade Math Score, NELS:88 and ELS:02

| | White Men | | White Women | |
|---|---|---|---|---|
| | NELS | ELS | NELS | ELS |
| Math | 0.445*** | 0.538*** | 0.946*** | 1.146*** |
| | [0.107] | [0.132] | [0.106] | [0.134] |
| Change from NELS | | 0.092 | | 0.200 |
| | | [0.170] | | [0.171] |
| Obs | 2474 | 2558 | 2461 | 2824 |

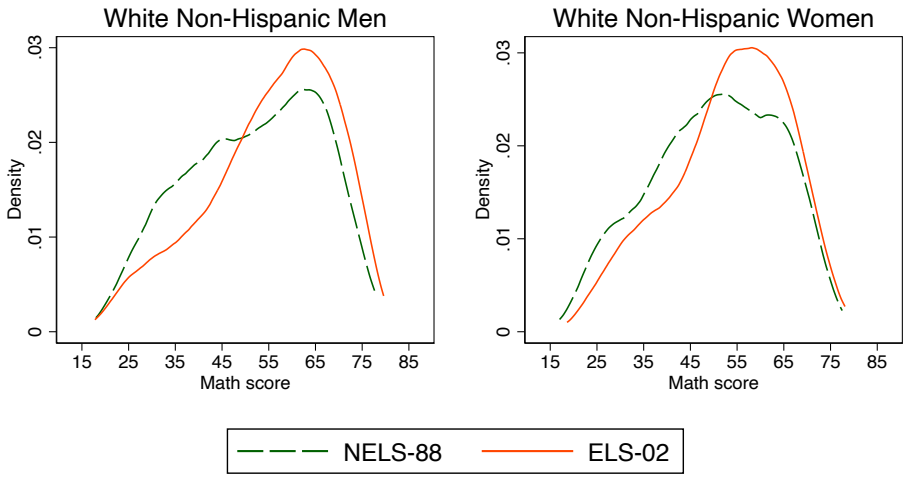
*Note*: We present OLS estimates of the wage returns to 12th-grade math scores in a univariate regression. We present results separately for NELS:88 and ELS:02, and for white non-Hispanic men and white non-Hispanic women. We also present the estimated change in the OLS estimates across cohorts. Sample weights are used. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table D.2: Distribution Statistics of 12th-grade Math Score in NELS:88 and ELS:02

| | White Men | | White Women | |
|---|---|---|---|---|
| | NELS | ELS | NELS | ELS |
| Mean | 51.6 | 55.4 | 50.1 | 53.1 |
| S.D. | 14.1 | 13.6 | 13.6 | 12.6 |
| Skewness | 0.25 | 0.59 | 0.24 | 0.46 |
| Kurtosis | 2.07 | 2.60 | 2.19 | 2.49 |
| p1 | 22.1 | 22 | 21.5 | 23.2 |
| p5 | 27.5 | 28.7 | 26 | 29.3 |
| p10 | 31.4 | 34.7 | 30.2 | 34.1 |
| p25 | 40.9 | 46.9 | 40.3 | 45 |
| p50 | 53.2 | 57.6 | 51 | 54.7 |
| p75 | 63.7 | 65.9 | 61.3 | 62.9 |
| p90 | 69.5 | 71.7 | 67.5 | 68.5 |
| p95 | 72.2 | 73.8 | 70.3 | 71 |
| p99 | 76 | 77 | 74.4 | 75.1 |

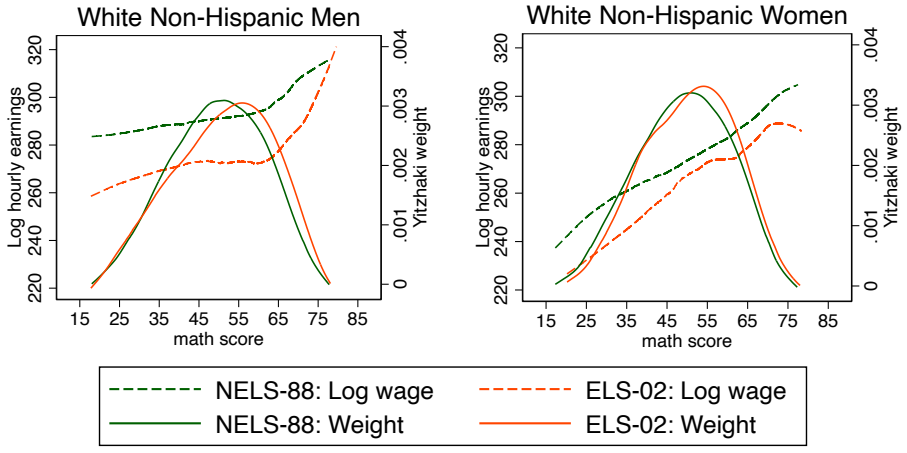
*Note*: We present the distribution statistics for 12th-grade math scores separately for NELS:88 and ELS:02, and for white non-Hispanic men and white non-Hispanic women. Sample weights are used.

Figure D.1: Distribution of Math Score, NELS and ELS



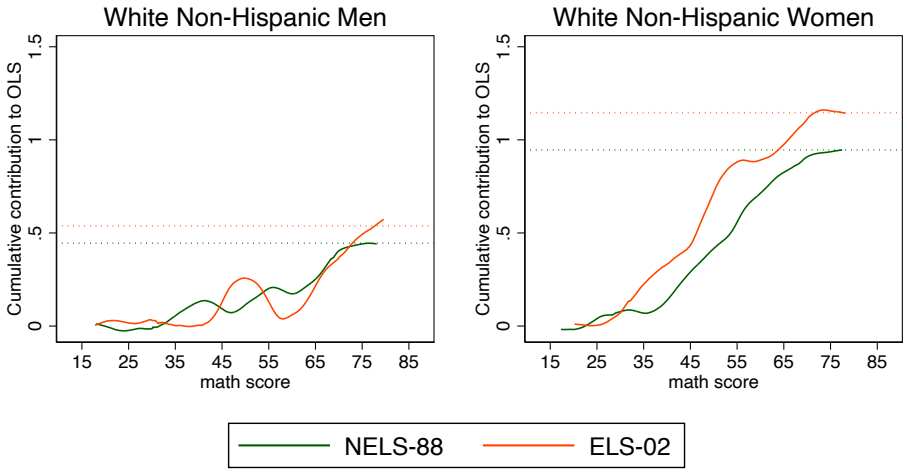

*Note*: In each figure, we plot the density of 12th-grade math scores separately for the two cohorts. Sample weights are used.

Figure D.2: Smoothed Yitzhaki Weights and Slopes, NELS and ELS




*Note*: In each figure, we overlay the smoothed Yitzhaki weights (right axis) with the smoothed pairwise slopes (left axis), separately for the two cohorts. The smoothed curves are estimated using Locally Weighted Scatterplot Smoothing (LOWESS) with a tricube weighting function. Sample weights are used.

Figure D.3: Cumulative Contribution to OLS, NELS and ELS




*Note*: In each figure, we graph the progressive sum of the Yitzhaki decomposition, starting with the lowest math score until the entire sum is calculated (producing the OLS estimate). We use a bandwidth of 0.3 for smoothing. Sample weights are incorporated in the Yitzhaki decomposition.

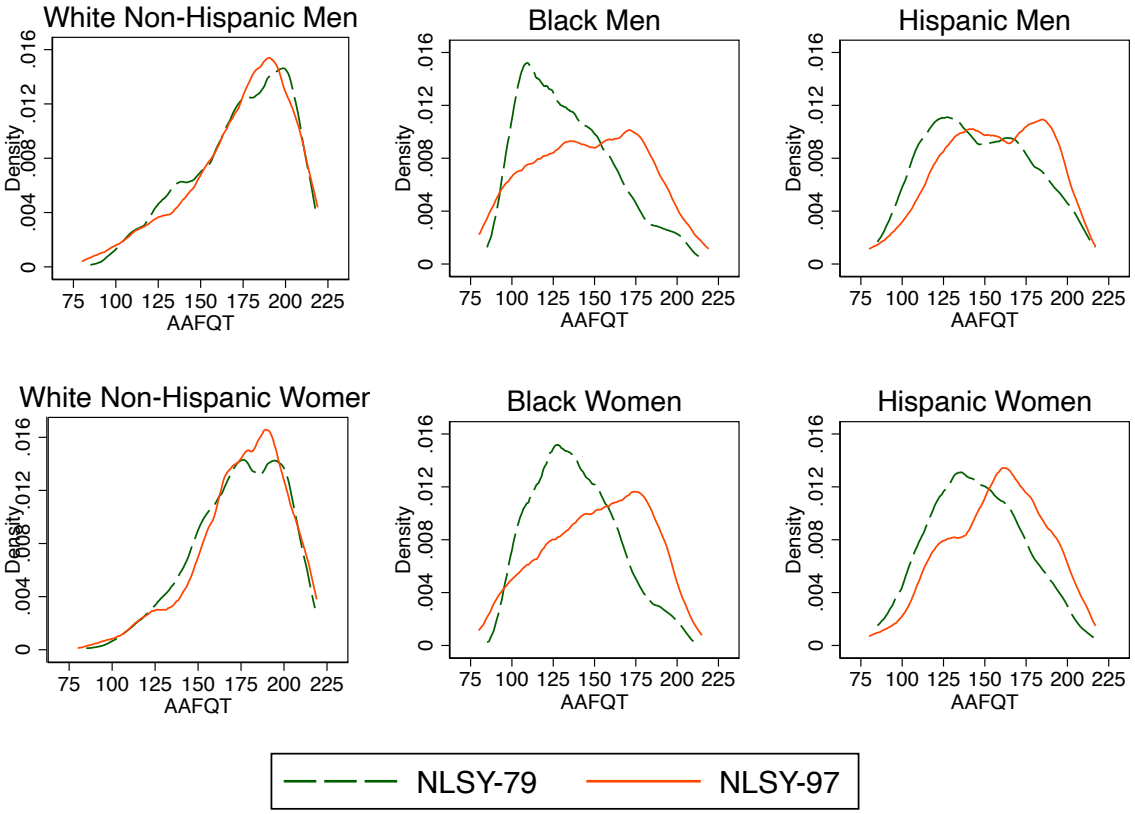
# E Tables and Figures for the Black and Hispanic Samples

Table E.1: OLS Estimates: By Gender and By Race and Ethnicity

| | **White** | | **Black** | | **Hispanic** | |
|---|---|---|---|---|---|---|
| | Men | Women | Men | Women | Men | Women |
| **Panel A**: Univariate Regression | | | | | | |
| AAFQT | 0.677*** | 0.830*** | 0.672*** | 0.962*** | 0.576*** | 0.823*** |
| | [0.035] | [0.036] | [0.051] | [0.044] | [0.060] | [0.054] |
| AAFQT * NLSY–97 | -0.212*** | -0.0407 | -0.141 | -0.176** | -0.228** | -0.257*** |
| | [0.055] | [0.060] | [0.086] | [0.072] | [0.094] | [0.097] |
| Obs | 3683 | 3679 | 1978 | 2156 | 1338 | 1373 |
| **Panel B**: Control for Social and Non-cognitive Skills | | | | | | |
| AAFQT | 0.605*** | 0.761*** | 0.592*** | 0.911*** | 0.460*** | 0.676*** |
| | [0.036] | [0.038] | [0.054] | [0.046] | [0.062] | [0.057] |
| AAFQT * NLSY–97 | -0.153*** | 0.00687 | -0.0812 | -0.134* | -0.117 | -0.104 |
| | [0.056] | [0.061] | [0.089] | [0.077] | [0.094] | [0.098] |
| Obs | 3683 | 3679 | 1978 | 2156 | 1338 | 1373 |

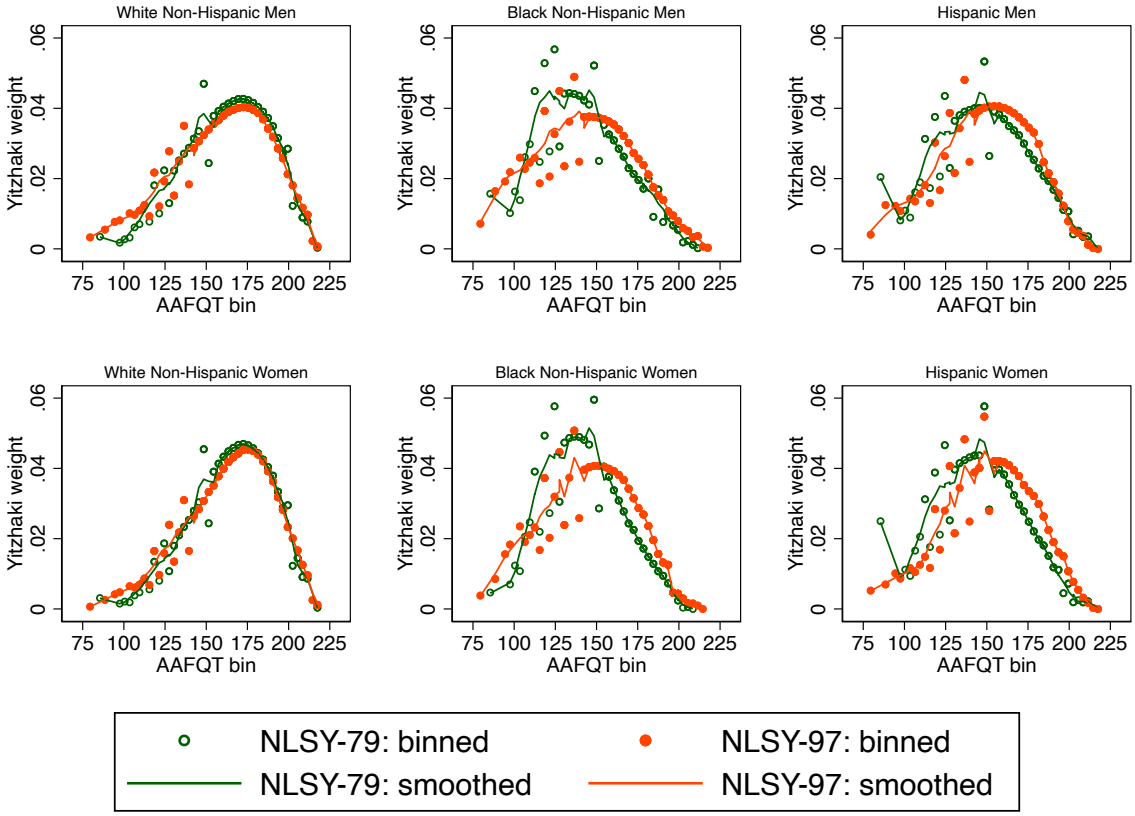
*Note*: We present OLS estimates of the wage returns to AAFQT scores, by gender and by race and ethnicity. Panel A presents results for the regression of log wages on the AAFQT score, a dummy variable for the NLSY–97 cohort, and their interaction term. Panel B further controls for measures of non-cognitive skills and social skills (created by Deming (2017)) and their interactions with the NLSY–97 dummy. Sample weights are used. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure E.1: Adjusted AFQT Distribution By Gender and By Race and Ethnicity



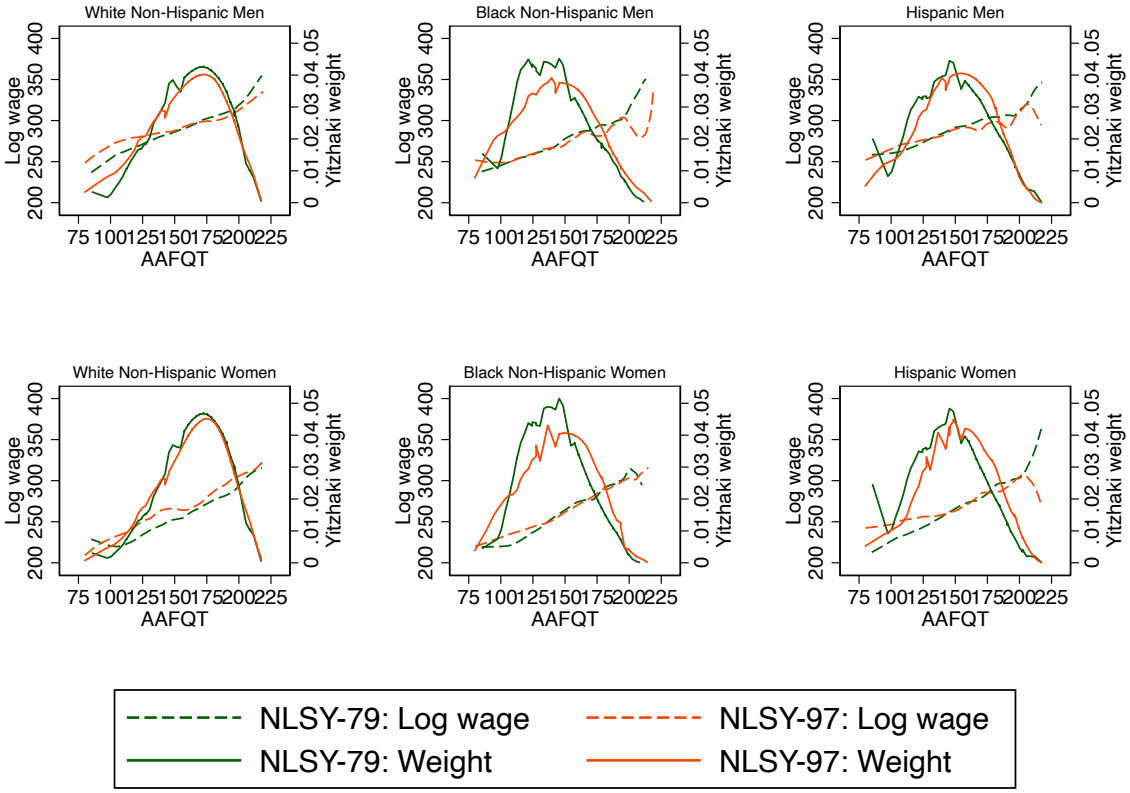

*Note*: We plot the density of AAFQT scores by gender and by race and ethnicity. AAFQT scores are concorded by Altonji et al. (2012). BLS custom sample weights are used.

Figure E.2: Smoothed Yitzhaki Weights By Gender and By Race and Ethnicity



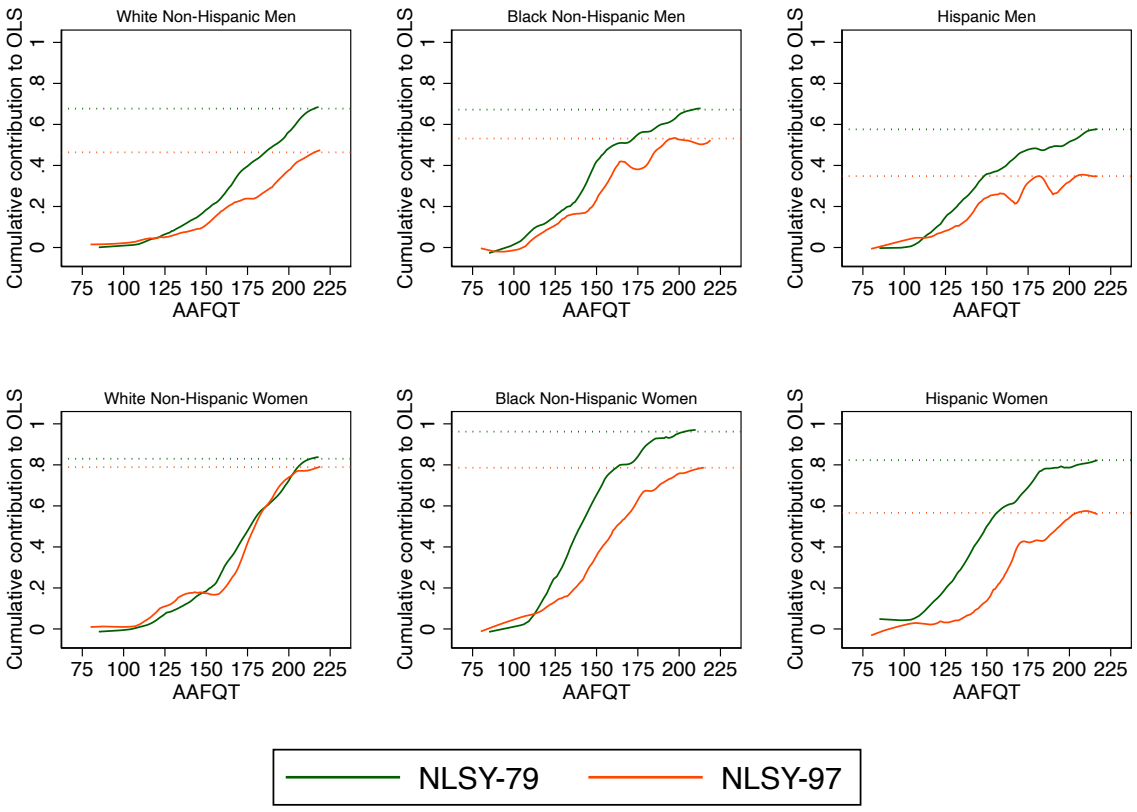

*Note*: We plot the Yitzhaki weights by gender and by race and ethnicity. The binned scatterplots are created by summing up Yitzhaki weights in each bin, which contains three consecutive AAFQT points. The smoothed curve is created using Locally Weighted Scatterplot Smoothing (LOWESS) with a tricube weighting function and a bandwidth of 0.1. BLS custom sample weights are incorporated in the Yitzhaki decomposition.

Figure E.3: Smoothed Yitzhaki Weights and Local Linear Regression By Gender and By Race and Ethnicity




*Note*: We overlay the smoothed Yitzhaki weights (right axis) with smoothed pairwise slopes (left axis), by gender and by race and ethnicity. The smoothed curves are created using Locally Weighted Scatterplot Smoothing (LOWESS) with a tricube weighting function. BLS custom sample weights are incorporated in the Yitzhaki decomposition.

Figure E.4: Cumulative Contributions to OLS Estimates By Gender and By Race and Ethnicity




*Note*: We graph the progressive sum of the Yitzhaki decomposition, by gender and by race and ethnicity. We use a bandwidth of 0.3 for smoothing. BLS custom sample weights are incorporated in the Yitzhaki decomposition.