NBER WORKING PAPER SERIES

TOWARD AN UNDERSTANDING OF THE RETURNS TO COGNITIVE SKILL ACROSS COHORTS

Judith K. Hellerstein
Sai Luo
Sergio Urzúa

Working Paper 32229
http://www.nber.org/papers/w32229

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2024, Revised July 2026

We thank Katharine Abraham, Dan Black, Melissa Kearney, Richard Murphy, Nolan Pope, and seminar participants at the University of Chicago, George Washington University, SUNY-Buffalo, UIUC, University of Cambridge, George Mason University, SED 2023 annual meeting, and the World Bank 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Toward an Understanding of the Returns to Cognitive Skill Across Cohorts
Judith K. Hellerstein, Sai Luo, and Sergio Urzúa
NBER Working Paper No. 32229
March 2024, Revised July 2026
JEL No. J24, J31

ABSTRACT

Recent research concludes that wage returns to cognitive skill have declined in the U.S. We reassess this finding. Using decomposition methods, we document the pivotal role played by dynamic shifts in the distributions of pre-labor market cognitive skill. These shifts explain the declining estimated returns, especially for white men. We discard measurement error as an explanation. Although often overlooked, grappling with changing pre-labor market skill distributions is necessary for capturing the evolution of their labor market returns. This may prove especially important in the future, given continuing changes in skill development in recent youth cohorts.

Judith K. Hellerstein
University of Maryland, College Park
Department of Economics
and NBER
hellerst@econ.umd.edu

Sergio Urzúa
University of Maryland, College Park
Department of Economics
and NBER
urzua@econ.umd.edu

Sai Luo
Shanghai University of Finance
and Economics
luosaiecon@gmail.com

1 Introduction

Large pandemic-related 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 (National Center for Education Statistics, 2024b,a). However, this is not the first time skill distributions have changed across cohorts in recent decades. 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 its 1997 version (NLSY–97). Taking this trend as given, several studies have concluded that the wage returns to cognitive skill declined for young workers in the U.S. over the past 40 years (Castex and Dechter, 2014; Deming, 2017; Ashworth et al., 2021). Yet, the accumulation of skills is at least partly an endogenous response to their anticipated labor market returns (Card, 1999). Thus, it is important to simultaneously 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 these returns and the changing measured distributions of skill. We follow previous studies that consider the log-linear relationship between wages and cognitive skill. We use the cognitive skill measure from Altonji et al. (2012), which is derived from the reported scores on the Armed Forces Qualification Test (AFQT) for the NLSY–79 and NLSY–97 samples. These scores across these two cohorts are not directly comparable, so Altonji et al. (2012) concorded them across samples and adjusted them for the age of the test takers. Because the resulting scores are not raw AFQT test scores, we refer to them as "adjusted" AFQT scores, or AAFQT.

Our analysis focuses on the evolution of the returns to cognitive skill *within* two 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 both groups' average AAFQT scores increased slightly over time. However, our primary focus lies in the widening and left-skewing of the distribution across the cohorts. In particular, we document a thicker tail at the bottom of

[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 Online Appendix Section 6.

the AAFQT distribution for men and women in the younger cohort (NLSY–97), accompanied by a "hollowing-out" in its low-to-middle range. Importantly, our analysis extends beyond this setting. We demonstrate similar patterns using data from two other longitudinal surveys administered by the National Center for Education Statistics (NCES). This, along with other indirect evidence presented, suggests that the evolving distribution of AAFQT scores captures a complex shift in cognitive skill across cohorts.

After documenting the evolution of the skill distribution, we shed light on the potential economic forces driving it. To this end, we introduce a model of endogenous skill investment and fixed cost consistent with the basic patterns of rising labor market returns and substantial parental education expenditures that characterized the 1980s and early 1990s (when NLSY–97 parents were investing in children). Within this framework, we simulate cognitive skill distributions under different parameterizations. The resulting distribution under higher labor market returns displays a higher average skill level, a hollowing out below the mean, and a longer left tail. These patterns mimic those observed in reality.

We then examine how changes in the measured cognitive skill distribution affect estimates of the labor market returns to this skill. Specifically, we reevaluate recent findings regarding declining returns in the U.S. by using Armed Forces Qualification Test (AFQT) scores derived from the National Longitudinal Survey of Youth (NLSY) cohort. We estimate the univariate linear relationship between the logarithm of wages and AAFQT scores. Our analysis shows that the wage return to AAFQT has declined for white men, which confirms previous studies; however, the results for white women are less definitive.

To illustrate how changes in the distribution of human capital can affect these estimates, we utilize the decompositions developed by Yitzhaki (1996). This framework connects the distribution of measured skills to the resulting ordinary least squares (OLS) wage returns. We begin with model-generated data for demonstration and then return to the NLSY data to explore the connection between AAFQT scores and observed wages.

Our findings reveal that the shifting distribution of AAFQT scores across the NLSY cohorts, particularly the increased left skewness, is crucial for understanding the changes in estimated wage returns to cognitive skill. Specifically, we find that the OLS estimates in the younger cohort (NLSY–97) give much greater weight to the wages of individuals with lower cognitive skill levels compared to the NLSY–79 cohort. This empirical observation significantly affects changes in the estimated wage returns across cohorts for men, while its impact on women is less pronounced. Our findings complement recent evidence linking the rise in skill dispersion to the growth of residual wage inequality observed since the late 1980s (Lochner et al., 2025).

We conduct simulation exercises to further explore the previous results and generate counterfactual estimated wage returns. We do this by maintaining the reported wages in each NLSY cohort but adjusting the underlying distributions of AAFQT scores to be the same across them. The structural weighting scheme of the Yitzhaki decomposition enables this strategy. The findings confirm the importance of accounting for the skill's evolution: we show that when the AAFQT distribution is fixed (mirroring that of the older cohort, NLSY–79), the estimated counterfactual returns to cognitive skill for men remain unchanged across the two cohorts. This starkly contrasts with the falling returns generated by the OLS estimates. For women, on the other hand, reweighting amplifies 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.

The fact that changing estimated wage returns to cognitive skill across the NLSY cohorts is so intrinsically linked (especially for men) to the changing skill distribution is a novel result. It suggests that closer attention needs to be paid to changes in the entire cognitive skill distributions of recent cohorts of youth. It also underscores the need to carefully interpret the estimated returns to cognitive skill as these cohorts enter adulthood and gain experience in the labor market.[2]

Several observations motivate our focus on the interconnection between pre-labor market skills and their future returns. First, previous evidence indicates that returns to cognitive skill declined for children of the NLSY–97 cohort relative to children of the NLSY–79 (e.g., Castex and Dechter, 2014; Deming, 2017). However, early school investments made by NLSY–97 parents were made in the 80s and early 90s, when real wages for those at the bottom of the wage distribution and the less educated were falling and when the returns to education were

[2] In addition to the well-known importance of cognitive skill in the labor market that is our focus, there is increasing recognition of the importance of socio-emotional and social skills (e.g., Deming, 2017; Edin et al., 2022). That said, there has been much less consensus among economists and psychologists on measuring those skills. As a result of this, consistent measures of these skills across cohorts and datasets are elusive. For example, no common questions were asked, and no tests in the two NLSY cohorts served as consistent socio-emotional or social skill measures. While we conduct a detailed analysis of cognitive skill and its wage returns in our analysis, it is worth keeping in mind that we still do not know much about the evolution of skills, endogenous skill investment, and labor market returns.

increasing (e.g., Juhn et al., 1993; Autor et al., 2008). Therefore, parents raising children in the 80s and early 90s might reasonably have expected that the returns to skill for their children upon adulthood would have been higher than those of the NLSY–79. We show that increasing returns can generate the divergence in the skill distribution that fits the pattern in the NLSY data.

Second, because of the rising dispersion in real wages in the 80s and early 90s, children in the NLSY–97 cohort whose parents earned at the bottom of the wage distribution earned lower real wages than similarly situated parents of the NLSY–79 cohort. This suggests an increasing role of credit constraints over this period, another issue our conceptual framework explores. Moreover, during the relevant period, there is no evidence (of which we are aware) that the cost of raising a child decreased in real terms.[3] These observations suggest that the real cost of raising a child increased between the two NLSY cohorts for families at the bottom of the income distribution. We show this can explain the changing patterns observed in the skill distribution, particularly the hollowing out below the mean.[4]

The paper is organized as follows. Section 2 describes our sources of information, while Section 3 documents the distributional changes in cognitive skill. Section 4 introduces our conceptual framework, which connects with the Yitzhaki decomposition described in Section 5. Section 6 presents our main findings, and Section 7 concludes.

2 Data and Related Research

2.1 NLSY and the Measure of Cognitive Skill

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

[3] The evidence does vary in terms of how much the cost of raising a child did increase, and data availability is an issue. One example is Laughlin (2013), for information starting in 1985.

[4] Two additional facts are worth noting. First, there was a sharp decrease in the fraction of two-parent families during this period. For single-parent families, the investment cost in children relative to total family income is mechanically higher than in two-earner families. In the NLSY samples, the share of white non-Hispanic men living in a two-parent family at about age 14 decreases from 85% in the older cohort (NLSY–79) to 62% in the younger cohort (NLSY–97). Our empirical results are robust to conditioning on family structure. Second, there is at least some evidence (e.g., Lochner and Monge-Naranjo, 2011) that credit constraints were more binding on human capital investment starting in the 1980's. This alone would also have caused a larger share of the population to be unable to overcome the costs of human capital investments.

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,[5] 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 and increased variance already apparent in 2020 accelerated in 2023.[6]

As a measure of cognitive skill, researchers have utilized the AFQT scores of the NLSY respondents (e.g., Neal and Johnson, 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 the ASVAB, but the raw scores are not directly comparable. 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. Specifically, Altonji et al. (2012) grapple carefully with two fundamental differences in the ASVAB 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. Altonji et al. (2012) use a concordance between the two test formats (see Segall (1997) for details) to generate test scores on a comparable scale across cohorts. Second, the NLSY–79 respondents were ages 15–23 when they took the test, while the NLSY–97 respondents were 12–18. Because AFQT test scores generally rise with age, an age adjustment is necessary to compare scores across ages.

Altonji et al. (2012) exploit the fact that in both cohorts, a non-trivial number of test-takers were 16 years old. Assuming that test-score rank does not change with age, the test scores of 16-year-olds are ranked by percentile, and these scores are then applied to percentiles of test takers at each age other than 16. In Online Appendix Section 1, we provide more information about the ASVAB, the construction of the AFQT, changes in test administration, the concordance

[5] `https://www.bls.gov/nls/nlsy26.htm`, accessed February 28, 2024
[6] `https://www.nationsreportcard.gov/highlights/ltt/2023/`, accessed February 28, 2024.

across NLSY cohorts, and additional references.

Based on a number of reasonable assumptions, Altonji et al. (2012) conclude from an analysis of AAFQT scores that the skill distribution widened between the NLSY–79 and NLSY–97, and that this likely has important implications for wage inequality. When Altonji et al. wrote their paper, the NLSY–97 cohort was too young to have realized wage outcomes, so conclusions about wage inequality remained speculative. Thus, this paper aims to enhance the existing work in this area.

2.2 Existing Findings of Wage Returns to Cognitive Skill

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 and Dechter (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 skill, 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.[7]

We first re-examine the findings of Castex and Dechter (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 is a measure of the wage of an individual i from cohort c (NLSY–79 or NLSY–97) and ϵ_i^c is the associated error term. We use two samples of White Non-Hispanic individuals from the NLSY–79 and NLSY–97. Basic summary statistics for these samples can be found in Appendix Table A.1.[8,9] To operationalize W_i, we compute the average of all hourly wages observed for

[7] 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 scores' different natures might explain this distinctive difference.

[8] 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.

[9] 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 skill) across NLSY cohorts,

an individual between the ages of 25–39. We then take the natural logarithm, and multiply by 100.[10]

Figure 1 plots the relationship between log wages and AAFQT scores. It also displays the OLS results of estimating equation (1) using data from each cohort. Panel A shows the results for white men. The estimated 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 log points. For white women, as shown in Panel B, the estimated return to an additional AAFQT point falls from 0.830 (NLSY–79) to 0.789 (NLSY–97). The drop is not significant, and its magnitude is much smaller than men's. In Table 1 we provide OLS results for multiple versions of (1) where covariates are also included. These results tell largely the same story as the univariate regressions, as expected given the point estimates reported in Castex and Dechter (2014) and Deming (2017).

As discussed in Section 5, characterizing the evolution of the cognitive skill distribution (AAFQT) is critical for understanding the changes in the OLS estimates displayed in Figure 1. Therefore, we examine this evolution across cohorts next.

3 Distributional Changes in Cognitive Skill

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 and Autor, 1999; Card and DiNardo, 2002; Autor et al., 2008). Employment grew rapidly not only at the highest-skill jobs but also at the lowest-skill jobs (Autor et al., 2006; Autor and Dorn, 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 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 $\widetilde{AAFQT}$.

[10] The correlation between age and AAFQT scores in the NLSY samples is negligible. Nonetheless, in Online Appendix Section 7, we examine the role of age in shaping wage differentials. First, we assess how the OLS regression results (as in Table 1) are affected by age-related changes in wages across cohorts. Consistent with previous research (e.g., Luo, 2023), early-career wage profiles differ between the two cohorts: wages increase more rapidly with age in the NLSY–79 than in the NLSY–97. The return to AAFQT also rises modestly with age in both samples but is statistically significant only for women. Second, we report results using wage observations restricted to ages 31—33. The main results remain robust under this restriction.

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. 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 is much more pronounced for the younger cohort (-0.81 for white men, -0.77 for white women) than for the older cohort (-0.62 for white men, -0.55 for white women). Consistent with this, the kurtosis 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 Tables 2 and 3.

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 "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 AAFQT scores for individuals between the 10th and 50th percentiles. To our knowledge, these distributional changes in cognitive skill have received very little attention. Moreover, they are not caused by changing correlations with other covariates, such as socio-emotional and social skills and years of education.[11]

Of course, any conclusions about changing cognitive skill (or changing returns) must rest on the assumption that these changes are not merely artifacts of measurement error.[12] Given the underlying complexities associated with constructing AAFQT scores, this is a serious concern.

In principle, one can correct for measurement error in estimates of the returns to cognitive skill. Indeed, motivated by these 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 and Rothstein (2016), measurement error is non-classical for IRT-based test scores.[13] Thus, conventional instrumental variables methods are not valid. One alternative approach is the "mixed effects structural equations" method in Junker et al. (2012). However, implementing this requires data on the responses of individual test-takers to each question in the ASVAB, and these are unavailable for

[11]See Appendix Figure A.1 for the distribution of AAFQT scores residualized by covariates.

[12]Concerns about measurement error in test scores have persisted for decades. See Griliches and Mason (1972) for an early discussion.

[13]In Online Appendix Figure OA.2, we plot the standard errors of the estimated IRT scores. The errors seem oddly large for low scores in the NLSY–97, suggesting some underlying issues with the IRT model. We thank Dan Black for pointing this out.

the NLSY–97. As a result, available methods for measurement error correction prove inadequate in this context. That said, three pieces of available evidence suggest that measurement error might not be a major concern in our setting.

First, the divergence and increased left-skewness of the AAFQT distribution in the NLSY–97, compared to that of the older cohort, is not driven by any one specific ASVAB section that Altonji et al. (2012) used to construct the AAFQT. Instead, as we show in Appendix Figure A.2, it appears to some degree in all components of the AAFQT.

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. While there are some anomalous aspects to the scores in the NLSY–97 cohort, Online Appendix Section 1 explains that they existed in the scores before any concordance was performed.

Third, and probably most convincingly, if measurement error in the construction of the AAFQT is driving the changing distributions across cohorts, it should be unique to AAFQT tests. But this is not the case. In the next two sub-sections, we present additional evidence from two sources of data collected by the National Center for Education Statistics (NCES).

3.1 Evidence from NELS and ELS

The National Education Longitudinal Study of 1988 (NELS:88) and the Educational Longitudinal Study of 2002 (ELS:02) are nationally representative longitudinal NCES data sets. 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. As a benchmark, note that 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.

The NCES datasets do not contain AAFQT scores, so we use respondents' math test scores from the senior year of high school as a measure of cognitive skill, following Weinberger (2014).[14] The ELS:02 tests are adapted from NELS:88, and they share many test items by design. Based on these shared test items and using IRT, ELS:02 contains an NELS:88-equated math score.[15]

[14] We find similar distributional changes in arithmetic reasoning scores, the subsection of ASVAB that directly measures mathematical skill and is one of the four components of AAFQT, across the two NLSY cohorts. We also find similar distributional changes in 10th-grade math scores between NELS and ELS. See Tables OA.1 and OA.2 in Online Appendix Section 1.

[15] The IRT-based math scores in NCES data can be subject to measurement error too. But there is nothing to suggest that measurement issues in NCES data would lead to the same measurement error pattern as AAFQT scores in the NLSY.

Table 4 compares the 12th-grade math score distributions between the two NCES cohorts, separately for white men (A) and white women (B). From the NELS:88 to ELS:02 cohorts, the average math score rose while the distribution became more left-skewed. As can be seen in Figure 3, there is also a hollowing out in the low-to-middle part of the math score distribution, for both groups. These patterns are present in the evolution of the AAFQT score distributions across the two NLSY cohorts, even though the NLSY and NELS/ELS cohorts do not perfectly overlap.[16]

3.2 Additional evidence from NAEP

The long-term trend (LTT) assessments from the National Assessment of Education Progress (NAEP) are designed and administered by NCES to measure the educational progress of successive cohorts of American students over a long period. In this sub-section, we focus on the math scores of 13-year-olds, the age in LTT NAEP close to that of AFQT takers in the NLSY, which is also available for more recent years.[17] We present published distributional statistics from 1978-2023 for the full sample of non-Hispanic whites in Table 5; detailed public statistics are unavailable separately for white men and women.[18] The first notable trend is a steady improvement until 2012 in math performance both in the mean and across the distribution. In addition, the standard deviation stayed relatively stable at around 30, consistent with the information on scores at various percentiles. However, because NCES does not publish the full distribution of scores, one cannot know with certainty whether there was a divergence in scores at the tails, something that may be especially important for scores below the 10th percentile.

From 2012 to 2020, trends in LTT NAEP math scores reversed. The mean math score of 13-year-olds declined from 293 to 291. Perhaps more strikingly, the change extended beyond the mean and fell differentially for students at different parts of the distribution. In particular, the math score declined by 7 points (from 253 to 246) for students at the 10th percentile, while increasing by 2 points for students at the 75th percentile and by 3 points for students at the

[16] While there are important similarities across the two sources of test scores, there are also differences. In particular, we do not find an increasing mass of very low math scorers across the NELS/ELS cohorts, in contrast to the NLSY.

[17] NAEP's LTT math scores are also available in some years for 9-year-olds and 17-year-olds. Data on LTT NAEP scores of 17-year-olds are unavailable after 2012.

[18] The LTT assessments were updated in 2004. Both the original and revised assessment formats were administered that year. See National Center for Education Statistics (2005) for a discussion of this design change.

90th percentile. The standard deviation increased substantially from 31 to 36.[19] Post the onset of the Covid-19 pandemic, the distributional changes in math scores have somewhat accelerated, as the mean score continued to decline from 2020 to 2023, and the decline was greater for the bottom of the score distribution. Specifically, as shown in Table 5, the 10th percentile score declined by 10 points (from 246 to 236), while the 75th percentile declined by 5 points and the 90th percentile score declined by 6 points.[20]

The published distributional change of LTT NAEP math scores since 2012 resembles in many ways the divergence of AAFQT scores between NLSY–79 and NLSY–97. Although NAEP's data do not allow us to look at statistics such as skewness and kurtosis, the increasing mass of students at the lower end of the math score distribution and the increasing standard deviation both suggest that there has been a divergence of math scores since 2012 among 13-year-olds.

In summary, the previous two subsections highlight consistent trends in the evolution of cognitive test scores across the NLSY and other data sources. Together, this bolsters confidence that the observed shifts in returns to cognitive skill from the NLSY reflect a real and general trend. Next, we introduce a model that illustrates an underlying economic mechanism capable of driving these distributional changes in measured cognitive skill. Consistent with evidence that cohort-based changes in AAFQT scores appear across all relevant ASVAB components (see Appendix Figure A.2), we assume that a single economic parameter influences these shifts.

[19] The 2020 LTT NAEP for 13-year-olds was conducted between October and December of 2019. The distributional changes in math scores from 2012 to 2020 are therefore free of the influence of Covid-19 disruptions.

[20] There are two other notable sets of test score trends representing more recent cohorts that also clearly demonstrate the same kinds of distributional changes. The first is the main NAEP test, commonly known as "The Nation's Report Card." See https://nces.ed.gov/nationsreportcard accessed January 3, 2025. Main NAEP has only been collected since 1990, but as with LTT NAEP scores, the standard deviation of main NAEP math scores has increased over the last decade or so. Additionally, for the seven years prior to the Covid pandemic, the data are consistent with a decline in math scores in the lower half of the distribution but no decline in scores at the 75th and 90th percentiles. Scores for 2024 (the most recent year available) indicate that the math test-score declines since Covid have continued for the bottom of the distribution (10th and 25th percentiles). In contrast, the pandemic-related test score declines for the 75th and 90th percentiles of the distribution have ceased, and test scores in 2024 at the 75th and 90th percentiles were both higher than the last administration in 2022. The second is the U.S. results from TIMSS, "Trends in International Mathematics and Science Study," which began in 1995. TIMSS math scores for U.S. 8th graders increased for those in the 90th percentile between 2011 and 2019 but decreased for those at the 10th percentile. See https://nces.ed.gov/pubs2022/2022041.pdf accessed November 23, 2024. Recently reported TIMSS results for 2023 (the most recent year of data available) document declines since 2019 for U.S. 8th graders across the math test score distribution (National Center for Education Statistics (2024a)).

4 A Model of Endogenous Skill with Fixed Investment Costs

A simple model of endogenous skill formation can illustrate the economic mechanisms leading to a change in the distribution of skill arising from a change in the wage returns to the same skill. Expanding on the literature showing that skill investment responds endogenously to technology (Heckman et al., 1998), we analyze the critical role investment costs play in shaping the trajectory of skill development, especially when it contains a fixed component. To this end, we consider a simple adaptation of the "q" theory of investment (Tobin and Brainard, 1976) to our setting. Our goal is not to fully fit the model to the data, but rather to point out in a simple framework that a simultaneous change in the marginal productivity of human capital *and* in the evolution of investment costs affects the population distribution of cognitive skill.[21]

Let H_t denote the level of human capital at time t, where H_0 represents the initial human capital endowment. Let I_t be human capital investment ($I_t = (dH_t/dt)/H_t$), $q(I_t)$ the cost of adding human capital, r the marginal productivity of human capital (in units of output), and ρ the discount rate. The agent maximizes the present value of future net flows of output:

$$Max\ _{I_t} \int_0^\infty (r - q(I_t)) H_t e^{-\rho t} dt$$

subject to a cost function $q(I)$.

Uzawa (1969) shows that when the cost function is convex ($q'(I) > 0, q''(I) > 0, q(0) = 0, q'(0) = 1$), the optimal I^* is uniquely determined. Consider a modified version where the cost function $q(I_t)$ has an initial fixed cost and is then convex. In this case, there are two possible optimal investment levels I^*: a high investment level determined by $\frac{r-q(I^*)}{\rho-I^*} = q'(I^*)$, and a low investment level (at a corner solution). We adopt this cost function (with fixed cost) which allows for heterogeneous optimal investment decisions.

Who ends up with the low-level optimal investment? If some individuals (or their families) are credit-constrained and affected by high fixed costs, they may optimally choose $I^* = 0$ (a corner solution). On the other hand, for a given discount rate, the high-level optimal investment would increase as the marginal productivity of human capital rises, for example, due to technological changes as described in Acemoglu (2002).

[21] Because our empirical application focuses on a uni-dimensional measure of skill, we use the terms "cognitive skill", "skill", and "human capital" interchangeably.

4.1 The impact of r on I^*

We now simulate how the optimal investment level, I^*, changes with the rate of return, r. We consider two different regimes. Both assume that the initial distribution of human capital, H_0, follows a (left-skewed) beta distribution $B(5, 1.5)$ and that $\rho = 0.1$. In Regime 1, $r = 0.2$ and I_H^* is 5%. In Regime 2, $r = 0.21$ and the high optimum I_H^* is 5.3%.

As shown below, an increase in r generates the divergence in the skill distribution that matches the pattern in the NLSY data. This triggers our interest in the marginal product of human capital. To examine the impact of r, we focus on how investments depend on an increasing share of the population unable to overcome the fixed cost.[22]

To operationalize the implications of an increasing population share constrained by the fixed cost happening simultaneously to an increase in r, we assume that in Regime 1 described above, the constraint is not binding so that everyone reaches the high optimum ($I_H^* = 5\%$). But under Regime 2, we assume that the constraint is binding for 5% of the population. This group ends up with the low optimum ($I_L^* = 0$), while the rest of the population reaches the high optimum ($I_H^* = 5.3\%$).

To further assess the implications of fixed costs, we consider two sub-cases of Regime 2. In Regime 2(a), we let the simulation be agnostic so that the fixed cost affects a random 5% of the population. This case mimics the situation where children across the baseline cognitive skill spectrum could all face barriers to skill investment. In Regime 2(b), we assume that the fixed cost constrains people in the bottom 5% of the initial human capital (H_0) distribution, which mimics the situation where children from disadvantaged backgrounds (who may have lower baseline cognitive skill accumulation) are more likely to face constraints in making skill investments.

Figure 4 depicts the resulting distribution of the stock of human capital, H_1 under the different regimes.[23] The distribution of human capital in both Regimes 2(a) and 2(b) is more divergent and left-skewed compared to Regime 1. This divergence is propelled in both cases by the 5% of the population making low human capital investments due to the fixed cost, coupled

[22] This is motivated by the evidence suggesting that the real cost of raising a child increased between the two NLSY cohorts for families at the bottom of the income distribution. This would have caused a larger share of the population to be unable to overcome the fixed cost of investment–and similarly for any variable costs. For simplicity, we do not formally model credit constraints in our simulation, but an increasing role for credit constraints would produce similar results.

[23] We draw 10,000 observations of initial human capital H_0 from the $B(5, 1.5)$ distribution and calculate H_1 (human capital after ten investment periods) for each observation under each of the investment regimes.

with the higher return to human capital that causes unconstrained individuals in both versions of Regime 2 to invest at a higher level than under Regime 1. The contrast between regimes is more pronounced for Regime 2(b) when the simulation assumes that the bottom 5% of the H_0 distribution is the population's share constrained to the low investment optimum.

This pattern of divergent human capital between regimes is similar to what we observe across NLSY cohorts. This demonstrates how the distribution of human capital can endogenously evolve in response to changing (expected) returns to human capital coupled with fixed investment costs.

5 The Yitzhaki Decomposition

To describe how changing measured distributions of human capital is related to changing estimated linear returns to human capital, we implement a Yitzhaki decomposition (Yitzhaki, 1996). 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 and H is human capital. The Ordinary Least Squares (OLS) estimate of the linear relationship between Y and H can be expressed as:

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

where $b_C(h)$ is the slope of $E(Y|C, H = h)$ evaluated at h for a given C. Therefore, β_C^{OLS} can be decomposed into unit treatment effects, $b_C(h)$, and associated weights, $w_C(h)$. To highlight the specific role of skewness in the weights, we rearrange Yitzhaki's original formulation and write the weights as:

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

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. As a result, 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 (2) reaches its peak when $F_{C,H}(h) = 0.5$, which corresponds to the center of the distribution of H. Its contribution to the weights is intuitive. However, understanding the second expression in (2) is equally important. Specifically, larger differences in the conditional expectations on either side of a given h contribute more to the OLS weights, making this dispersion critical for driving the point estimates. In particular, left-skewed distributions tend to assign higher weight to h values toward the bottom of the distribution, whereas right-skewed distributions assign lower weight. Therefore, when unit treatment effects $b_C(h)$ vary across the distribution of H, the weighting scheme used in the Yitzhaki decomposition plays a crucial role in computing β_C^{OLS}. In contrast, when the unit treatment effects are constant, the weights become irrelevant in practice.

In both the empirical analyses and in the empirical simulation we conduct below, realizations of H are discrete. 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: $N = N_1 + \cdots + N_n$.[24] Then, let $\Delta h_i = h_{i+1} - h_i$, $\Delta \bar{y_i} = y_{\bar{i+1}} - \bar{y_i}$ ($\bar{y_i}$ is the average outcome for individuals reporting $h(i)$), 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:

$$w_i = \frac{1}{\sigma_h^2} \left(\frac{\sum_{j=1}^{i} N_j}{N} \right) \left(\frac{\sum_{j=i+1}^{n} N_j}{N} \right) \left(\frac{\sum_{j=i+1}^{n} N_j h_j}{\sum_{j=i+1}^{n} N_j} - \frac{\sum_{j=1}^{i} N_j h_j}{\sum_{j=1}^{i} N_j} \right) \Delta h_i. \tag{3}$$

Finally, consider two cohorts C_1 and C_2, each investing in human capital under potentially different returns to human capital (due to, say, different technology) and potentially different fixed investment costs. We decompose the OLS estimate obtained from equation (1) for each

[24] When weighted least squares (WLS) is utilized instead of OLS, it is straightforward to extend the Yitzhaki decomposition. See Online Appendix Section 3 for details.

cohort as:

$$\beta_{OLS}^{C_1} = \sum_{i=1}^{n-1} w_i^{C_1} b_i^{C_1} \qquad \text{and} \qquad \beta_{OLS}^{C_2} = \sum_{i=1}^{n-1} w_i^{C_2} b_i^{C_2}. \tag{4}$$

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

5.1 Implementing The Yitzhaki Decomposition on Simulated Data

To demonstrate how the Yitzhaki decomposition works in practice, we return to the simulation in Section 4, where regimes can alternatively be thought of as cohorts (if the regimes occur at different times). We project the stock of human capital H_1 in each of the simulations' regimes onto wages by assuming that wages for each individual i are $W_i = \alpha \times e^{r \times H_{1i}} \times \epsilon_i$, where ϵ_i is assumed to be distributed according to a normal distribution ($N(0, 0.01)$), and we generate the resulting empirical distribution of log wages for the simulated data. We then replicate the analysis under each regime. We start with the analytical formula of Equation (2).

The difference between the regimes in the human capital investment environments simultaneously impacts both the Yitzhaki weights and the slopes. Figure 5 shows that more weight is allocated to the lower part of the H_1 distribution in Regimes 2(a) and 2(b) compared to Regime 1. The slopes are just the rates of return, r, across regimes.[25]

The slopes and weights across regimes can then be combined via the Yitzhaki decomposition to produce the different OLS estimates. As expected, the estimated β^{OLS} is 0.2 in Regime 1 and 0.21 in Regimes 2(a) and 2(b). In the real world, however, the data are always coarser, H is measured in discrete units, and sample sizes are often small for each skill level. Thus, even if the functional form of the underlying relationship between skill and wage is truly (log) linear, the estimated unit slopes using real data will not be constant. To see this, and as a segue to using the NLSY data, we draw a random sample of 2,000 observations from our simulated dataset. We

[25] The difference across regimes in H_1 is even more pronounced for Regime 2(b), as there is missing mass in Regime 2(b) because the lowest 5% of H_0 who are constrained by the fixed cost to make no further human capital investments fall behind the rest of the human capital distribution.

bin the values of H_1 so that there are approximately 120 equal-sized bins.[26] We then implement the Yitzhaki decomposition on these data using the discrete formula of Equation (3).

It is evident that the simulated weights and pairwise slopes obtained in this case (smaller sample and discretized data) exhibit more noise compared to the continuous, original simulated sample. As illustrated in Figure 6, although the simulated weights are still governed by the analytical weighting function defined in Equation (2) and generally follow the pattern shown in Figure 5, there is significantly more variation in the Yitzhaki weight associated with each binned value of H compared to the continuous version. And the variation (and noise) in the data becomes even more pronounced in the pairwise slopes, as seen in Figure 7. While the slopes in Regime 2 are, on average, higher than those in Regime 1, approximately 30% of the sample displays a reversed order, making the slopes in Figure 7 less distinguishable between the two regimes. This reflects a more realistic scenario involving small sample sizes and discrete values of H, similar to the real-world NLSY data.

6 The Evolution of Wage Returns to Cognitive Skill

We now turn to the more important exercise of implementing the Yitzhaki decomposition on the actual data for the two NLSY cohorts.[27] Figure 8 plots the Yitzhaki weights by gender for each cohort. To make the graphs more readable, the dots in the Figure are Yitzhaki weights averaged across bins containing three consecutive AAFQT scores. We also overlay these binned weights with local linear regression curves.[28] Given the similarities in the AAFQT distributions for white men (a) and women (b) as seen in Figure 2, the weights – and in particular, the weight changes across cohorts – are also alike between these groups.

For both groups, due to the increasing left-skewness of the AAFQT distribution, low AAFQT scores receive more weight for the NLSY–97 than for the NLSY–79. This is true for AAFQT

[26] We choose the specific sample size and bin numbers to replicate the actual NLSY sample size and the number of unique AAFQT values.

[27] As previously noted, we maintain the framework of previous studies of considering the linear relationship between the log of wages and AAFQT. We show in Online Appendix Section 5 that the Yitzhaki decomposition does allow for consideration of nonlinear relationships, as long as the returns are parameterized as constant within discrete ranges of AAFQT (e.g. above/below median or quartiles) of AAFQT. This is done via a double residualization process. In Online Appendix Section 5, we estimate OLS regressions and conduct the Yitzhaki decompositions allowing for differential returns to cognitive skill above and below the median. The main findings of these results are consistent with our overall conclusion from considering linear returns–one should be skeptical of concluding that returns to AAFQT have declined across cohorts.

[28] We use Locally Weighted Scatterplot Smoothing (LOWESS) with a tricube weighting function and a bandwidth of 0.1.

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–79. As is apparent from equation (4), studying how the Yitzhaki weights work together with the pairwise slopes is critical for understanding this result. Recall that the pairwise slope of the Yitzhaki decomposition is $b_i = \Delta \bar{y}_i / \Delta h_i$. In Figure 9, we depict smoothed versions of the $\bar{y}_i$'s for each cohort (bandwidth of 0.1) and we overlay the graph with the (smoothed) Yitzhaki weights that we previously generated in Figure 8.[29]

Interesting patterns emerge from our analysis. First, for the sample of white men (top panel of Figure 9), the gradient of the wage curve for the NLSY–79 does essentially appear to be steeper through much of the distribution of AAFQT scores.

Second, although the weights look similar between white men and women (bottom panel of Figure 9), the slopes of their wage curves diverge, with the slopes for women in general being flatter than for men. In addition, differences in the slopes of wages for white women across the two cohorts are characterized by more variability than for men, and with less clear patterns between the cohorts.[30] One notable feature for women is that the gradient of the wage curve becomes flatter or negative for a fairly wide region of AAFQT scores between 140 to 160 points, especially for the NLSY97 cohort. This pattern is not apparent for white men, and 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, the left panels of Figure 10 displays the progressive sum of the Yitzhaki decomposition from equation (4), starting with the lowest AAFQT

[29] Smoothing of the unit slopes is necessary for depicting the wage curve because the actual pairwise slopes, as shown in Online Appendix Figure OA.1, are very noisy. This happens because pairwise slopes represent differences in average log wages between adjacent AAFQT scores, differences that are exacerbated by the small sample sizes at any given AAFQT score. Even with the smoothing we do in Figure 9, the wage curves are somewhat noisy and contain small regions of AAFQT scores with negative slopes. Of course, if we were to increase the bandwidth in Figure 9 when smoothing the wages, these "wiggles" in the wage curve would become less prominent (and obviously eventually disappear). Nonetheless, even with only the modest smoothing of wages in the Figure, clear patterns emerge.

[30] The pairwise slopes themselves can be found in Online Appendix Figure OA.1.

score until the entire sum is calculated (producing the OLS estimate).[31] We graph the progressive sum by three-point bins, and also present smoothed versions of them in each panel of the figure.

The top left panel presents the results for white men. The dots on the graph are the progressive sums themselves. The contributions of the lowest AAFQT scores to the OLS estimates are not markedly different across cohorts. However, the progressive sums from the Yitzhaki decomposition begin to permanently diverge in the smoothed version at an AAFQT score of around 110 (5th percentile), at which point the Yitzhaki sum for the NLSY–79 cohort rises quickly until it reaches its final level of 0.68. In contrast, the cumulative smoothed sum for the NLSY–97 rises much more slowly for much of the AAFQT distribution. This is consistent with Figure 9, where the flatter overall slopes for the NLSY–97 play a large role in depressing the NLSY–97 OLS estimate relative to the NLSY–79 estimate. Notably, even for the unsmoothed versions, the Yitzhaki sum for the NLSY–97 (the orange dots) is consistently below that of the NLSY–79 (the green dots) starting at an AAFQT score that is below the median AAFQT score of 176. This highlights the large role of lower AAFQT scorers in driving the final OLS estimates.

The bottom left panel of Figure 10 reports the results for white women. The Yitzhaki sums of the two cohorts move largely in tandem and remain close to each other for much of the AAFQT distribution, with two exceptions. First, the Yitzhaki sum for the NLSY–97 cohort is below that of the NLSY–79 until around 140 points but then overtakes it. 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 9, where the gradients of the local linear regression diverge between the two cohorts at the top of the AAFQT distribution.[32]

[31] More precisely, 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:

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

[32] Online Appendix Section 4 contains results and discussion for the Yitzhaki decomposition applied to the NELS/ELS data described in Section 3.1. Unlike with AAFQT scores across the cohorts, but more consistent with Weinberger (2014), the point estimates of the OLS returns to cognitive skill in NELS/ELS (as represented by 12th-grade math scores) are larger in the younger cohorts than the older cohorts for both men and women, but they are not statistically significantly different across the cohorts.

6.1 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 NLSY 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 β^{79}_{OLS} and β^{97}_{OLS} as:

$$\begin{aligned}
\beta^{79}_{OLS} - \beta^{97}_{OLS} &= \left(\beta^{79}_{OLS} - \beta^{97}_{OLS}|w^{79}\right) + \left(\beta^{97}_{OLS}|w^{79} - \beta^{97}_{OLS}\right), \\
&= \sum_{i=1}^{n-1} w^{79}_i (b^{79}_i - b^{97}_i) \; + \; \sum_{i=1}^{n-1} (w^{79}_i - w^{97}_i) b^{97}_i. \qquad (5)
\end{aligned}$$

The first term in equation (5) is the counterfactual difference in the OLS estimates, holding fixed the AAFQT distribution (and the corresponding Yitzhaki weights) at the NLSY–79 level.[33]

We first present the counterfactual estimates for white men. Using expression (5), the OLS decline of 0.21 points for white men (as reported in Figure 1) can be decomposed as:

$$\begin{aligned}
\beta^{79}_{OLS} - \beta^{97}_{OLS} &= \Big(0.677 - 0.689\Big) + \Big(0.689 - 0.464\Big), \qquad (6) \\
&= -0.012 + 0.225.
\end{aligned}$$

Thus, holding the Yitzhaki weights at NLSY–79 levels, the returns to AAFQT scores stay basically unchanged across cohorts (an *increase* of 0.012 points). In other words, when we use the weighting structure generated by the AAFQT distribution in the older cohort (NLSY–79)

[33] An alternative decomposition is:

$$\begin{aligned}
\beta^{79}_{OLS} - \beta^{97}_{OLS} &= \left(\beta^{79}_{OLS} - \beta^{79}_{OLS}|w^{97}\right) + \left(\beta^{79}_{OLS}|w^{97} - \beta^{97}_{OLS}\right), \\
&= \sum_i (w^{79}_i - w^{97}_i) b^{79}_i \; + \; \sum_i w^{97}_i (b^{79}_i - b^{97}_i),
\end{aligned}$$

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

to calculate the (linear) returns to cognitive skill, we find no evidence that the returns have declined for white men. 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 skill.

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 (0.677). This is graphed in the top right panel of Figure 10. 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 again above around the median AAFQT score, and increase in tandem thereafter.[34]

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

$$\begin{aligned}\beta^{79}_{OLS} - \beta^{97}_{OLS} &= \Big(0.830 - 0.702\Big) + \Big(0.702 - 0.789\Big), \qquad (7)\\ &= 0.128 - 0.087.\end{aligned}$$

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. Still, unlike for men, both the OLS and counterfactual estimates show declines across the cohorts. The bottom right panel of Figure 10 shows that the cumulative sum for the counterfactual NLSY–97 estimates 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 consistent with the finding that the observed differences for women across the cohorts are much less stark than for men.

[34]The negative slope in the counterfactual NLSY–97 sum at AAFQT scores between around 170 and 180 is due to two very low binned outliers. Using even a slightly larger bandwidth – 0.08 or 0.05 instead of 0.03 – largely eliminates it. See Online Appendix Section 8 for details.

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.

It is natural to contemplate the underlying economic reasons behind concurrent changes in cognitive skill distribution and its implications on wage returns across cohorts. From a theoretical perspective, any economic model that seeks to rationalize changing wage returns should not focus solely on their conventional determinants (such as changing technology on the demand side or changing labor market experience on the supply side) but should also address how and why the distributions of cognitive skill are affected by underlying economic drivers of the human capital investment process. We illustrate this point by introducing a model of endogenous skill investment. Within its framework, we show that the cognitive skill distribution (as observed in NLSY) can result from changing wage returns arising from an external factor like changing technology and the structure of investment costs.[35]

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 linear return to AAFQT scores critically depends on changes in the distribution of AAFQT scores between the two NLSY cohorts and, in particular, on the widening (and increased left-skewness) of the AAFQT distribution in the NLSY–97.

To the extent that returns to cognitive skill 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. There also appear to be no changes in the returns to cognitive skill (as measured by math test scores) for either men or women in the NELS/ELS

[35] 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. Administrative data matching students' standardized test scores to later-life labor market outcomes is a more promising data source for future research.

samples.

Given all the evidence, we are skeptical of simple narratives of declining returns to cognitive skills across U.S. youth cohorts. Moreover, close attention should be paid to the dramatically changing skill distribution of recent cohorts of American youth, with increased emphasis on moments beyond the mean and variance. In a few years, as these young cohorts enter adulthood, it will be critical to link the changing skill distribution to subsequent adult labor market returns.

References

Acemoglu, D. (2002). Technical change, inequality, and the labor market. *Journal of Economic Literature 40*(1), 7–72.

Altonji, J. G., P. Bharadwaj, and F. Lange (2012). Changes in the Characteristics of American Youth: Implications for Adult Outcomes. *Journal of Labor Economics 30*(4), 783–828.

Ashworth, J., V. J. Hotz, A. Maurel, and T. Ransom (2021). Changes across cohorts in wage returns to schooling and early work experiences. *Journal of Labor Economics 39*(2), 413–447.

Autor, D. H. and D. Dorn (2013). The growth of low-skill service jobs and the polarization of the us labor market. *American Economic Review 103*(5), 1553–1597.

Autor, D. H., L. F. Katz, and M. S. Kearney (2006). The polarization of the us labor market. *American Economic Review 96*(2), 189–194.

Autor, D. H., L. F. Katz, and M. S. Kearney (2008). Trends in us wage inequality: Revising the revisionists. *The Review of Economics and Statistics 90*(2), 300–323.

Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis with special reference to education*. Chicago: University of Chicago Press.

Card, D. (1999). The causal effect of education on earnings. In O. C. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3, pp. 1801–1863. Elsevier.

Card, D. and J. E. DiNardo (2002). Skill-biased technological change and rising wage inequality: Some problems and puzzles. *Journal of Labor Economics 20*(4), 733–783.

Castex, G. and E. K. Dechter (2014). The Changing Roles of Education and Ability in Wage Determination. *Journal of Labor Economics 32*(4), 685–710.

Deming, D. J. (2017). The Growing Importance of Social Skills in the Labor Market. *The Quarterly Journal of Economics 132*(4), 1593–1640.

Edin, P.-A., P. Fredriksson, M. Nybom, and B. Ockert (2022). The Rising Return to Non-Cognitive Skill. *American Economic Journal: Applied Economics 14*(2), 115–153.

Griliches, Z. and W. M. Mason (1972). Education, Income, and Ability. *Journal of Political Economy 80*(3, Part 2), S74–S103.

Heckman, J. J., L. Lochner, and C. Taber (1998). Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents. *Review of Economic Dynamics 1*(1), 1–58.

Heckman, J. J., J. Stixrud, and S. Urzúa (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics 24*(3), 411–482.

Jacob, B. A. and J. Rothstein (2016). The Measurement of Student Ability in Modern Assessment Systems. *Journal of Economic Perspectives 30*(3), 85–108.

Juhn, C., K. M. Murphy, and B. Pierce (1993). Wage Inequality and the Rise in Returns to Skill. *Journal of Political Economy 101*(3), 410–442.

Junker, B. W., L. S. Schofield, and L. J. Taylor (2012). The use of cognitive ability measures as explanatory variables in regression analysis. *IZA Journal of Labor Economics 1*(1), 4.

Katz, L. F. and D. H. Autor (1999). Changes in the wage structure and earnings inequality. In O. C. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3, pp. 1463–1555. Elsevier.

Laughlin, L. L. (2013). Who's minding the kids? child care arrangements. Technical Report P70-135, U.S. Census Bureau.

Lochner, L. J. and A. Monge-Naranjo (2011). The nature of credit constraints and human capital. *American Economic Review 101*(6), 2487–2529.

Lochner, L. J., Y. Park, and Y. Shin (2025). The evolution of unobserved skill returns in the u.s.: A new approach using panel data. Working Paper w33407, National Bureau of Economic Research.

Luo, S. (2023). Racial gaps in the early careers of two cohorts of american men. `https://ssrn.com/abstract=4556291`. SSRN Electronic Journal.

National Center for Education Statistics (2005). NAEP 2004 Trends in Academic Progress: Three Decades of Student Performance in Reading and Mathematics. Technical Report NCES 2005-463, U.S. Department of Education, Institute of Education Sciences.

National Center for Education Statistics (2024a). TIMSS 2023 U.S. Highlights Web Report (NCES 2024-184). Technical report, U.S. Department of Education, National Center for Education Statistics.

National Center for Education Statistics (2024b). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), various years, 1971–2024 Long-Term Trend (LTT) Reading and Mathematics Assessments. Technical report, U.S. Department of Education, Institute of Education Sciences.

Neal, D. and W. R. Johnson (1996). The role of premarket factors in black-white wage differences. *Journal of Political Economy 104*(5), 869–895.

Schofield, L. S. (2014). Measurement error in the AFQT in the NLSY79. *Economics Letters 123*(2), 193–196.

Segall, D. O. (1997). Equating the CAT-ASVAB. In W. A. Sands, B. K. Waters, and J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation*, pp. 117–130. Washington, D.C.: American Psychological Association.

Tobin, J. and W. Brainard (1976). Asset markets and the cost of capital. Discussion Paper 427, Cowles Foundation for Research in Economics.

Urzúa, S. (2008). Racial Labor Market Gaps The Role of Abilities and Schooling Choices. *Journal of Human Resources 43*(4), 919–971.

Uzawa, H. (1969). Time preference and the penrose effect in a two-class model of economic growth. *Journal of Political Economy 77*(4, Part 2), 628–652.

Weinberger, C. (2014). The Increasing Complementarity between Cognitive and Social Skills. *Review of Economics and Statistics 96*(5), 849–861.

Yitzhaki, S. (1996). On Using Linear Regressions in Welfare Economics. *Journal of Business & Economic Statistics 14*(4), 405–411.

Table 1: 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***	0.464***	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***	0.165***	0.437***	0.322***
	[0.042]	[0.051]	[0.044]	[0.053]
Change from NLSY–79		–0.239***		–0.116*
		[0.067]		[0.068]
Obs.	2099	1584	2191	1488
Panel D: Control for Non-cognitive & Social Skills & Education				
AAFQT	0.363***	0.177***	0.406***	0.327***
	[0.043]	[0.050]	[0.044]	[0.052]
Change from NLSY–79		–0.185***		–0.079
		[0.066]		[0.068]
Obs.	2099	1584	2191	1488

Note: We present OLS estimates of the wage returns to AAFQT scores (concorded by Altonji et al. (2012)). The dependent variable is the log of average hourly wages, which we then multiply by 100 for ease of display. 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Distribution Statistics of AAFQT scores, NLYS–79 and NLSY–97

	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 associated with the chi-squared distribution.

Table 3: Statistical Tests for Equal Moments across Cohorts: NLSY–79 vs. NLSY–97, 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 4: Distribution Statistics of 12th-grade Math Score in NELS:88 and ELS:02

	A. White Men		B. White Women	
	NELS	ELS	NELS	ELS
Mean	51.8	56.0	51.2	54.2
S.D.	13.8	13.2	13.0	12.1
Skewness	-0.27	-0.62	-0.29	-0.52
Kurtosis	2.11	2.71	2.29	2.63
p1	22.1	22.6	22.2	25.3
p5	28	29.6	27.6	31.4
p10	32.1	36.5	32.4	35.9
p25	41.2	48.1	41.8	46.7
p50	53.4	58.3	52.3	55.8
p75	63.5	66.3	61.8	63.4
p90	69.2	71.8	67.7	68.6
p95	71.9	74.1	70.4	71.2
p99	75.9	77	74.4	75.3

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.

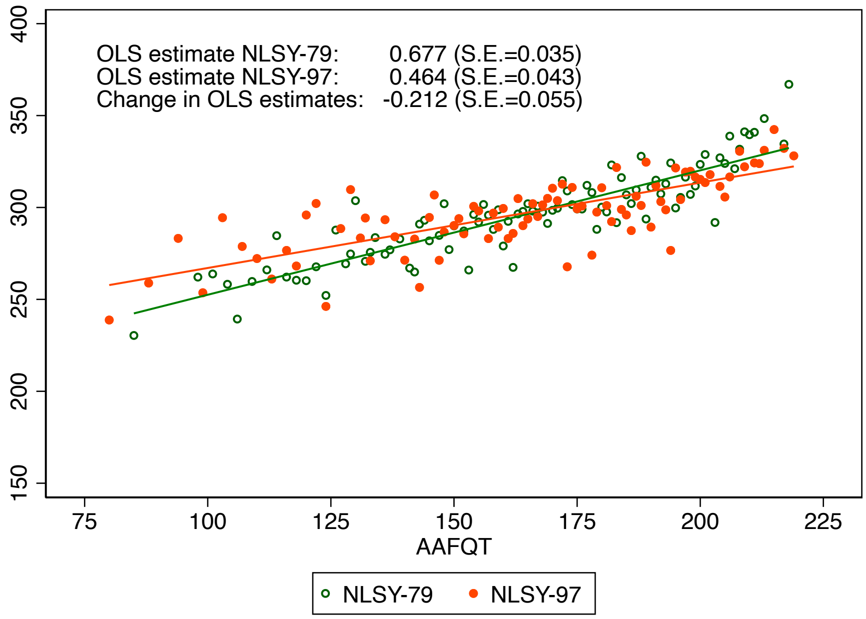
Table 5: Distribution Statistics of 13-year-old Math Score in NAEP Long-term Trend Assessment, Non-Hispanic Whites

Year	Mean	p10	p25	p50	p75	p90	S.D.
1978*	272	226	248	272	296	317	36
1982*	274	234	254	275	296	314	31
1986*	274	236	254	273	293	312	29
1990*	276	239	257	277	296	313	29
1992*	279	242	260	279	298	315	28
1994*	281	243	262	282	300	318	30
1996*	281	245	262	281	300	318	29
1999*	283	245	263	283	303	322	30
2004*	288	251	270	290	309	326	31
2004	287	249	269	289	308	326	31
2008	290	251	272	292	311	328	32
2012	293	253	273	294	313	331	31
2020	291	246	270	293	315	334	36
2023	285	236	262	287	310	328	37

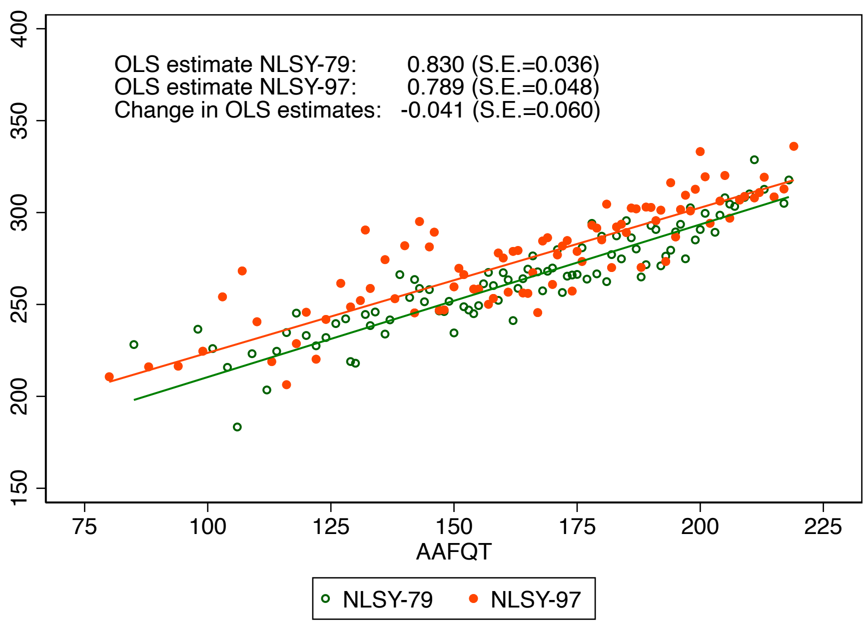
Note: The math scores are from NAEP long-term trend assessments. In 2004, changes were made to the assessment design. * represents the old assessment format. Source: https://nces.ed.gov/nationsreportcard/ltt/ accessed February 28, 2024.

Figure 1: OLS Estimates of the Wage Returns to AAFQT

(A) White Non-Hispanic Men



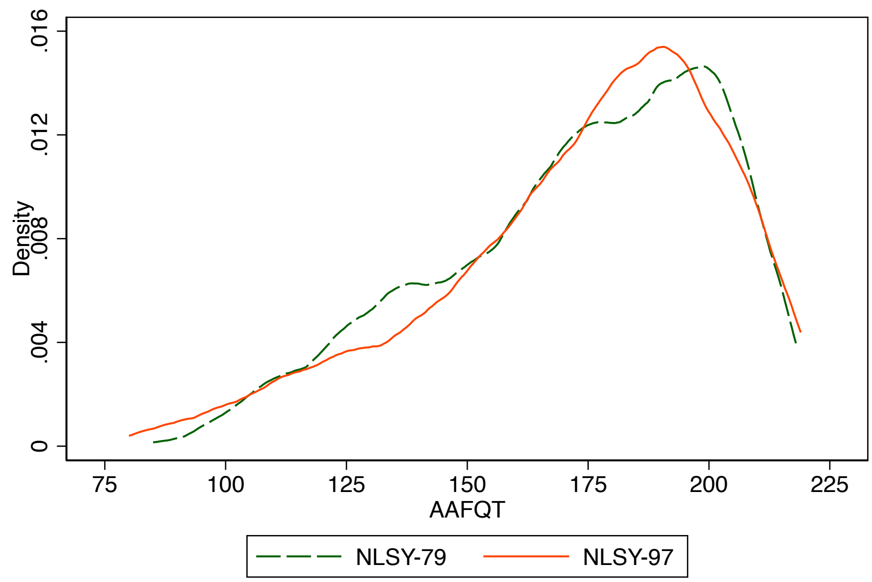
(B) White Non-Hispanic Women



Note: In each figure, we plot 100 times 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 Table 1 for the full regression results without and with covariates. BLS custom sample weights are used.

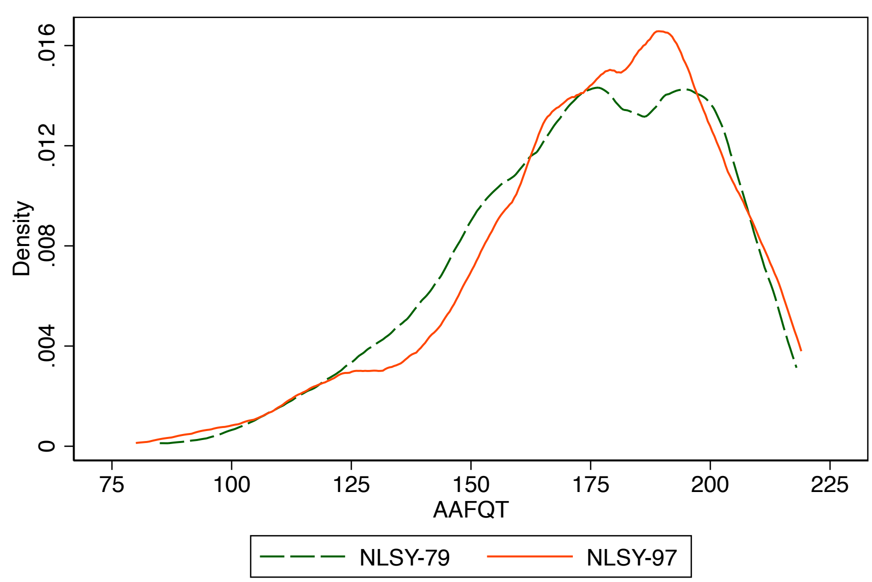
Figure 2: Adjusted AFQT Distribution for White Non-Hispanic Men and Women

(A) White Non-Hispanic Men



NLSY-79: Mean = 172.7, SD = 29, Skewness = -.62, p1 = 104, p5 = 118, p50 = 178
NLSY-97: Mean = 173.4, SD = 29.8, Skewness = -.8100000000000001, p1 = 94, p5 = 113, p50 =

(B) White Non-Hispanic Women

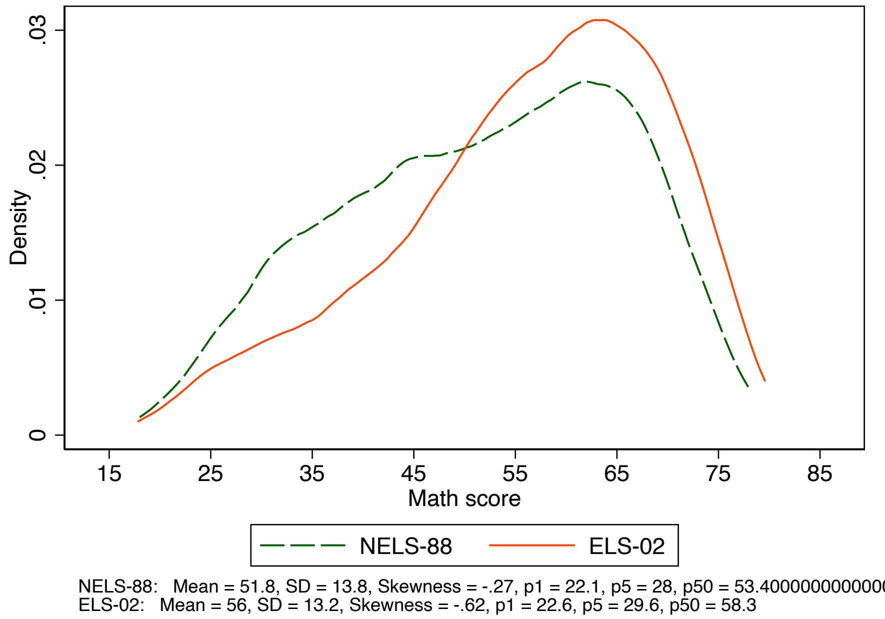


NLSY-79: Mean = 173.3, SD = 26, Skewness = -.55, p1 = 106, p5 = 124, p50 = 176
NLSY-97: Mean = 175.4, SD = 26.2, Skewness = -.77, p1 = 103, p5 = 122, p50 = 179

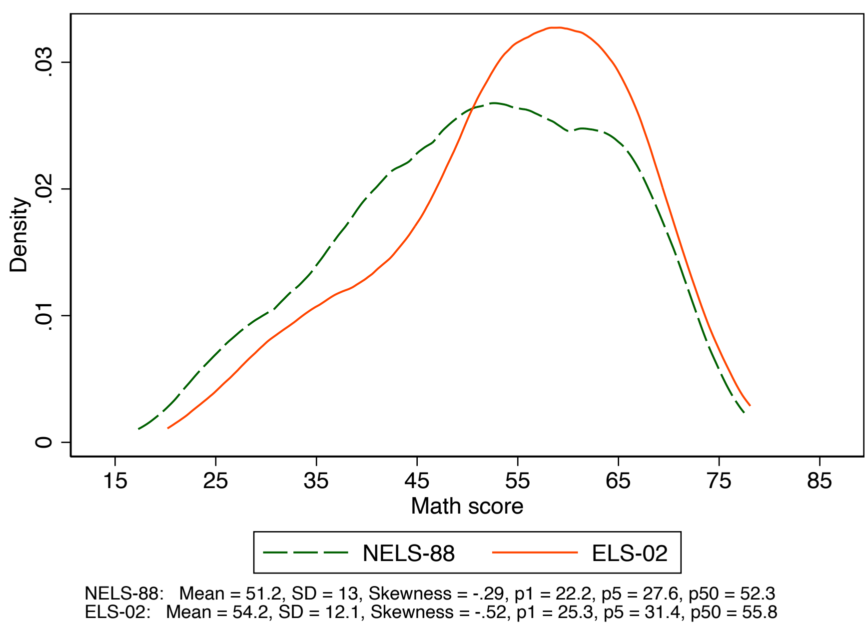
Note: In each figure, we plot the density of AAFQT scores separately for the two cohorts. AAFQT scores are concorded by Altonji et al. (2012). At the bottom of each figure, we present distribution statistics for each cohort. See Tables 2 and 3 for a full list of distribution statistics and tests for cross-cohort changes. BLS custom sample weights are used.

Figure 3: Distribution of 12th-grade Math Score, NELS and ELS

(A) White Non-Hispanic Men

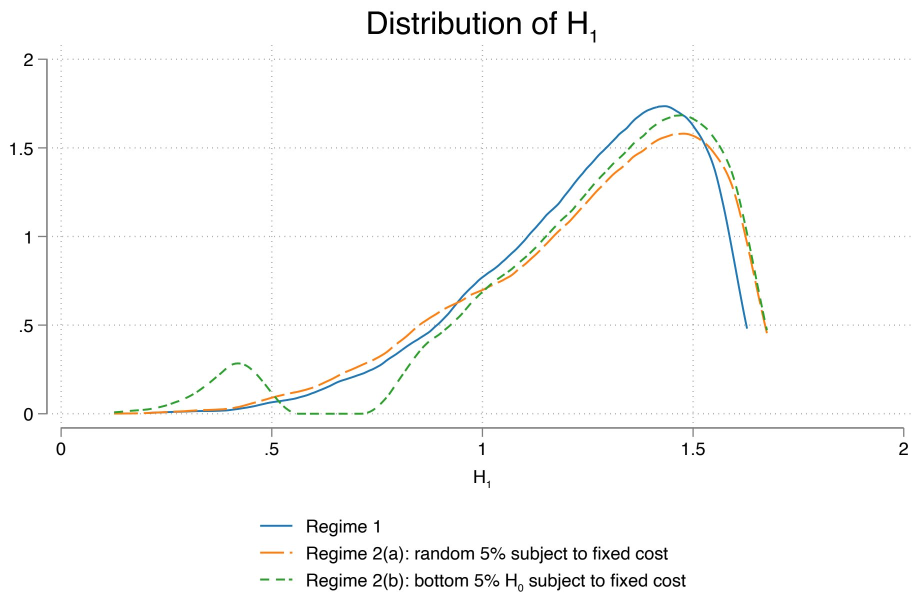


(B) White Non-Hispanic Women



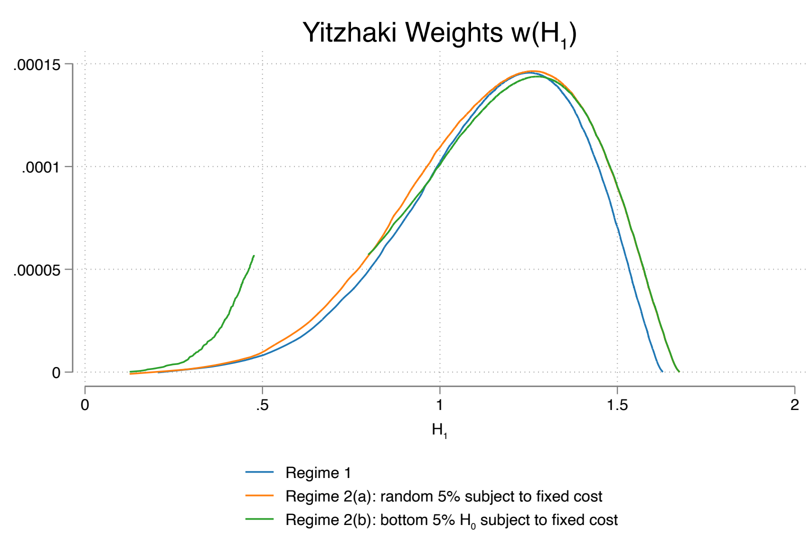
Note: In each figure, we plot the density of 12th-grade math scores separately for the two cohorts. NELS-equated math scores are used. At the bottom of each figure, we present distribution statistics for each cohort. See Tables 4 for a full list of distribution statistics. Sample weights are used.

Figure 4: Simulated distributions of H_1 before (regime 1) and after (regime 2) the change in productivity



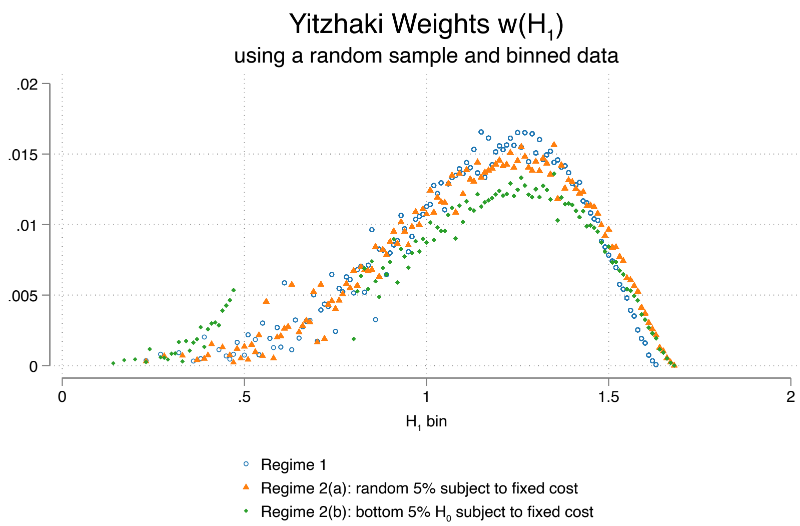
Note: The figure plots the simulated distribution of the stock of human capital, H_1, under different regimes. We draw 10,000 observations of initial human capital H_0 and calculate H_1 under each regime. In Regime 1, $r = 0.2$ and there is no binding constraint of investment cost. In Regime 2(a) and 2(b), $r = 0.21$ and 5% of the population faces a binding fixed investment cost constraint. In Regime 2(a), a random 5% of the population faces the constraint, and in Regime 2(b), the bottom 5% of H_0 distribution faces the constraint.

Figure 5: Simulated Yitzhaki weights $w(H_1)$ before (regime 1) and after (regime 2) the change in productivity



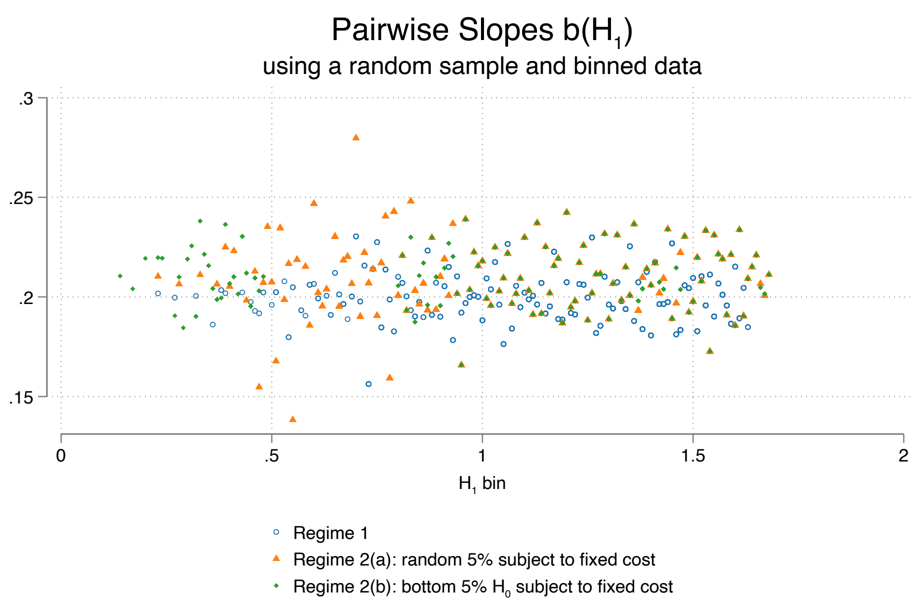
Note: The figure plots the simulated Yitzhaki weights, $w(H_1)$, under different regimes. We draw 10,000 observations of initial human capital H_0 and calculate H_1 under each regime. The Yitzhaki weights are then calculated using equation (2). See the notes to Figure 4 and Section 4 for the definition of each regime.

Figure 6: Simulated Yitzhaki weights $w(H_1)$ before (regime 1) and after (regime 2) the change in productivity, using a random sample and binned data



Note: The figure plots the simulated Yitzhaki weights, $w(H_1)$, using a 20% random sample of our full simulated dataset ($N = 10,000$). We bin the values of H_1 so that there are approximately 120 equal-sized bins, replicating the actual NLSY sample size and the number of unique AAFQT values. We then implement the Yitzhaki decomposition using equation (3). See notes to Figure 4 and Section 4 for the definition of each regime.

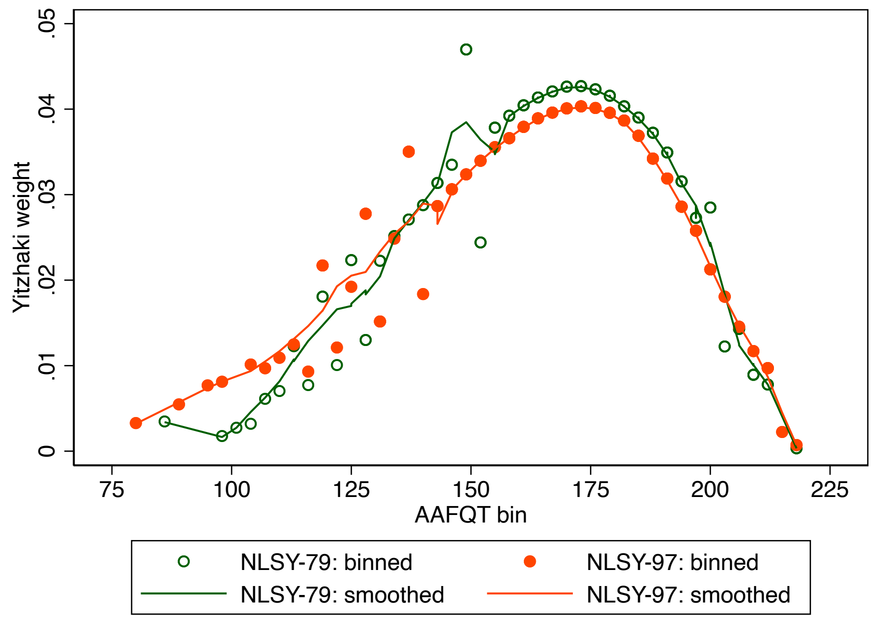
Figure 7: Simulated Yitzhaki slopes $b(H_1)$ before (regime 1) and after (regime 2) the change in productivity, using a random sample and binned data



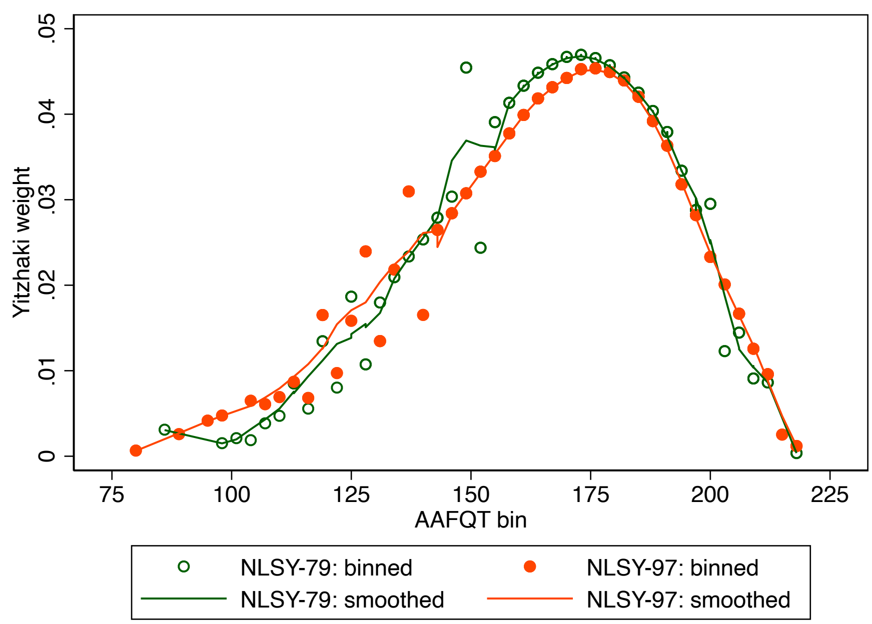
Note: The figure plots the simulated Yitzhaki slopes $b(H_1)$ using a 20% random sample of our full simulated dataset ($N = 10,000$). We bin the values of H_1 so that there are approximately 120 equal-sized bins, replicating the actual NLSY sample size and the number of unique AAFQT values. See notes to Figure 4 and Section 4 for the definition of each regime.

Figure 8: Yitzhaki Weights

(A) White Non-Hispanic Men



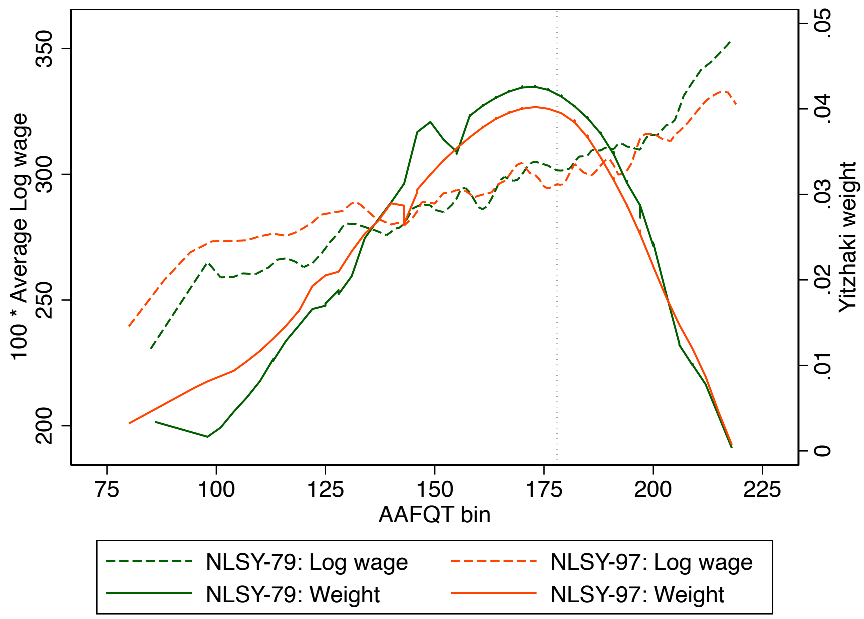
(B) White Non-Hispanic Women



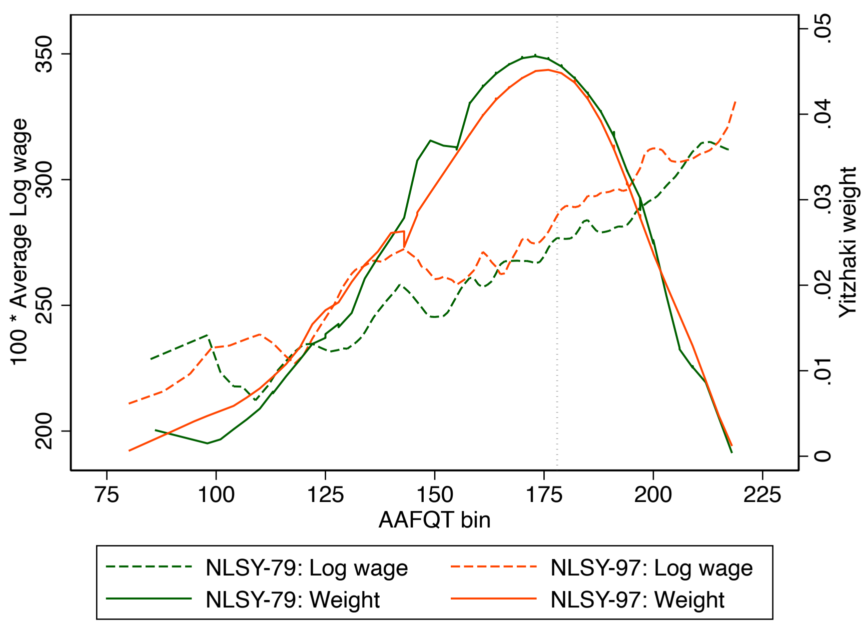
Note: In each figure, we plot the Yitzhaki weights using the formula in equation (3), separately for the two NLSY cohorts. The binned scatterplots are created by summing up Yitzhaki weights in each bin, which contains three consecutive AAFQT points. The smoothed curve is created using Locally Weighted Scatterplot Smoothing (LOWESS) with a tricube weighting function and a bandwidth of 0.1.

Figure 9: Smoothed Yitzhaki Weights and Local Linear Regression

(A) White Non-Hispanic Men



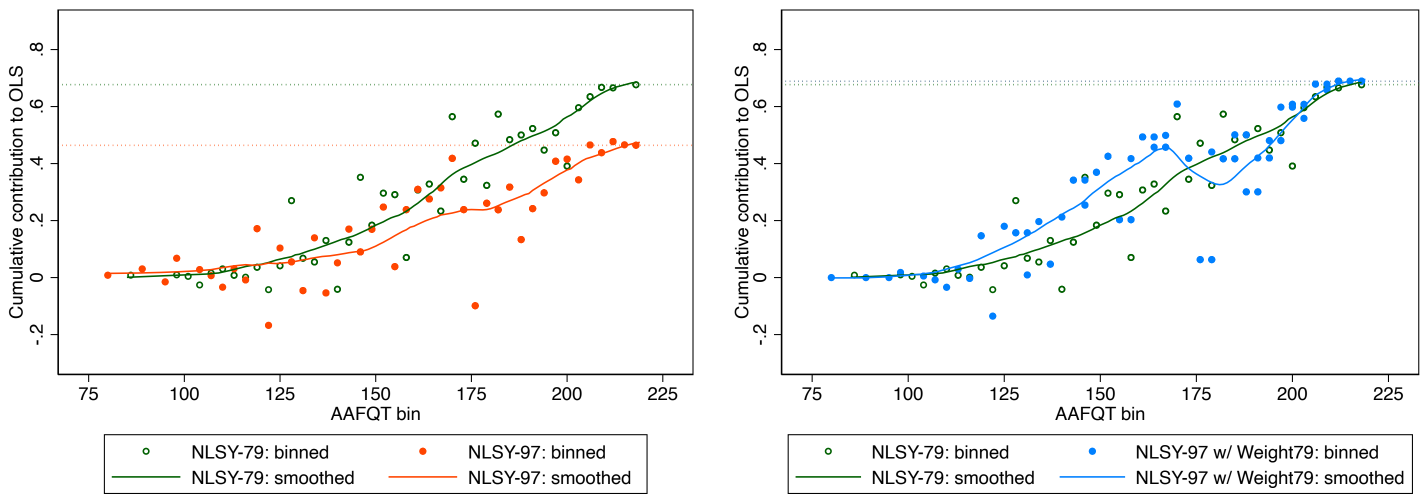
(B) White Non-Hispanic Women



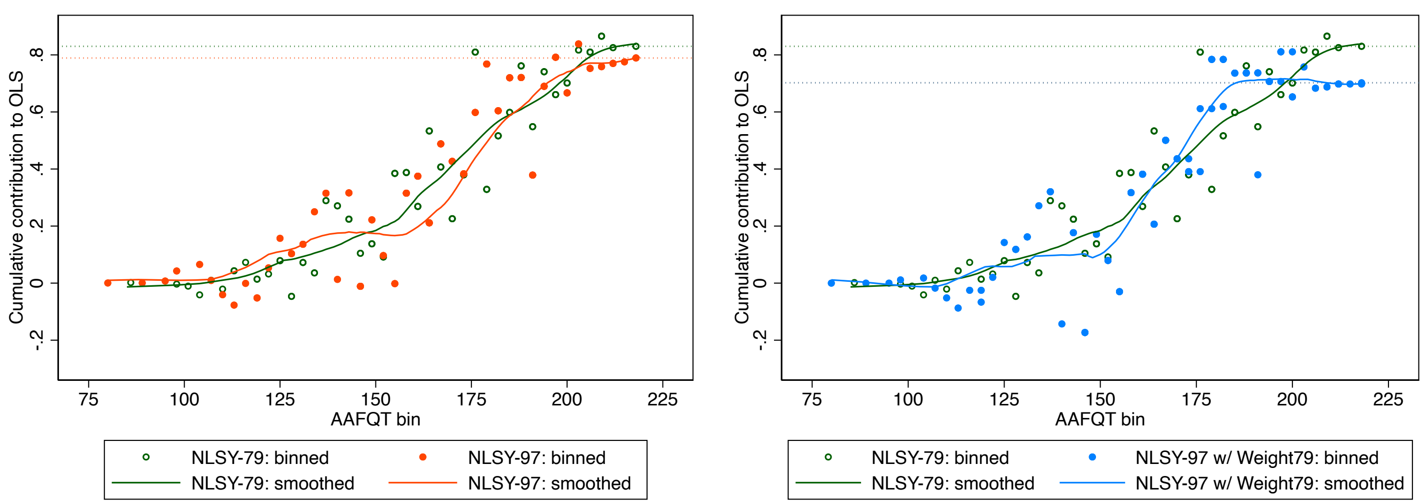
Note: In each figure, we overlay the smoothed Yitzhaki weights (right axis) with smoothed average log wages (left axis), separately for the two NLSY cohorts. We multiply average log wages by 100 for ease of display. The smoothed curves are created using Locally Weighted Scatterplot Smoothing (LOWESS) with a tricube weighting function. A bandwidth of 0.1 is used for smoothing both the weights and the wages.

Figure 10: Cumulative Contributions to Actual & Counterfactual OLS Estimates

(A) White Non-Hispanic Men



(B) White Non-Hispanic Women

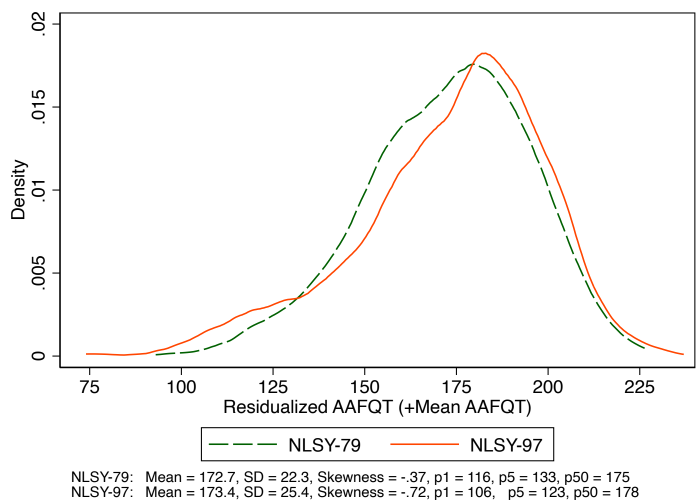


Note: In each figure, we graph the progressive sum of the Yitzhaki decomposition from equation (4), starting with the lowest AAFQT score until the entire sum is calculated (producing the OLS estimate). The top left and bottom left panels plot the progressive sum for the actual OLS estimates of the two cohorts. The top right and bottom right panels plot the counterfactual estimate for NLSY–97 (using NLSY–79 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.

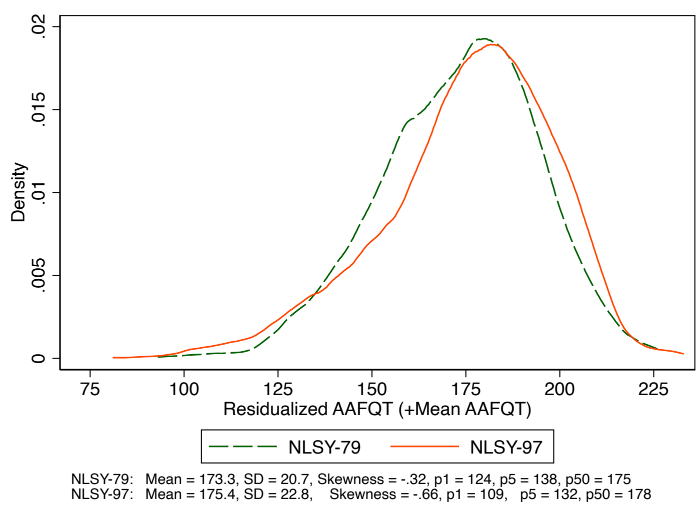
Appendices

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

(A) White Non-Hispanic Men



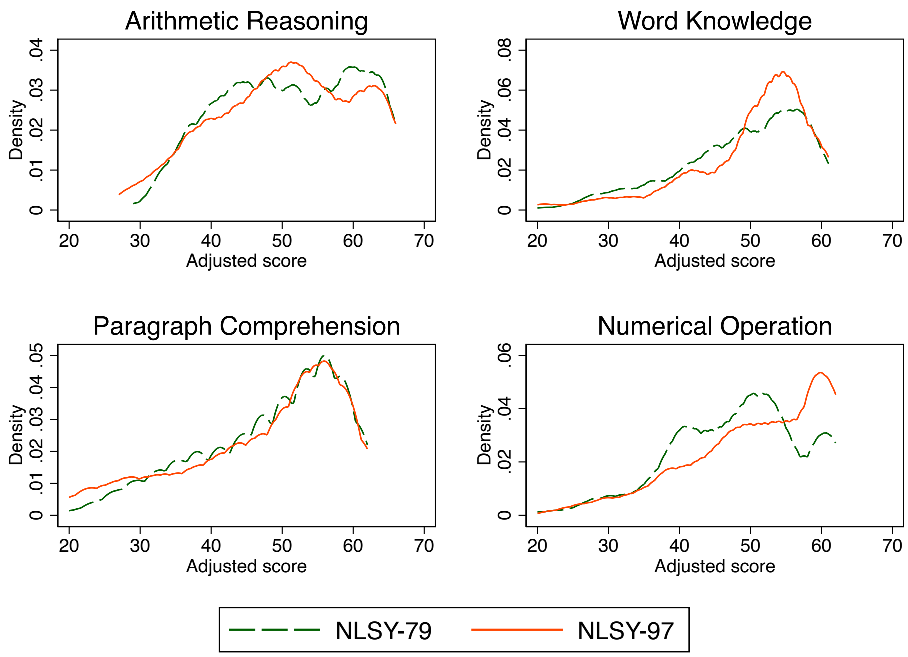
(B) White Non-Hispanic Women



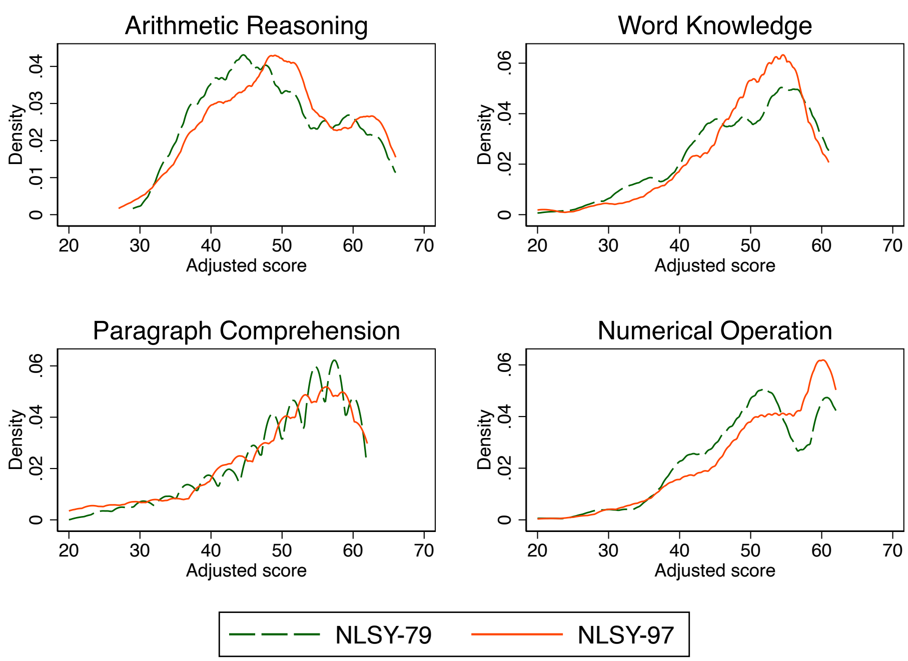
Note: In each figure, we plot the density of residualized AAFQT scores separately for the two NLSY cohorts. AAFQT scores (Altonji et al., 2012) are residualized by measures of non-cognitive and social skills (Deming, 2017), and the highest grade completed. At the bottom of each figure, we present distribution 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 NLSY 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 Online Appendix Section 1 for a discussion of other versions of the AFQT score. BLS custom sample weights are used.

Table A.1: Summary Statistics

	White Men		White Women	
	NLSY–79	NLSY–97	NLSY–79	NLSY–97
Average Age	30.8	30.1	30.8	30.1
High School Graduate (incl. GED)	0.42	0.30	0.40	0.21
Some College	0.21	0.24	0.24	0.23
Bachelor and Above	0.31	0.41	0.32	0.52
AAFQT (Altonji et al.)	172.7	173.4	173.3	175.4
Mother's Highest Grade Completed	12.1	13.4	12.0	13.4
Living with Both Parents as Child	0.81	0.62	0.80	0.58
Observations	2099	1584	2191	1488

Note: BLS custom sample weights are used. Average age is calculated with the years of non-missing wage observations over ages 25–39. Family structure (living with both parents) is measured at age 14 for the NLSY–79 and at the first wave (ages 12–16) for the NLSY–97.

Online Appendix for "Toward an Understanding of the Returns to Cognitive Skill Across Cohorts"

1 Notes on the Armed Forces Qualification Test (AFQT)

The Armed Forces Qualification Test (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 parts of 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 NLSY–79 and the NLSY–97, as survey respondents took the ASVAB. There are two existing versions of the AFQT: AFQT–80 and AFQT–89. The former, AFQT–80 scores, form the basis of the AAFQT scores we use for our analysis, as in Altonji et al. (2012), Castex and Dechter (2014), and Deming (2017). AFQT–80 is the summation of scores on four sections of the ASVAB: arithmetic reasoning (AR), numerical operations (NO), paragraph comprehension (PC), and word knowledge (WK). The formula is AFQT–80 = AR + 0.5*NO + PC + WK.

Additional background information on the Armed Forces Qualification Test (AFQT) can be found in Altonji et al. (2009), Defense Manpower Data Center (2006), Sands et al. (1997), annual reports on population representation in the military services (e.g. Quester and Shuford (2017)), and the BLS website.[1] We draw from these sources in some of the discussion below.

The ASVAB shifted from a paper-based to a computer-based test starting at 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. NLSY–79 respondents were paid $50 for taking the ASVAB, and NLSY–97 respondents were paid $75. In the paper-based test, all respondents received the same set of questions. The computer-based test is adaptive, so that a correct answer a respondent gives to one question leads to a more difficult subsequent question, whereas the opposite is true for a wrong answer. This adaptive feature means that different test-takers can receive different sets of questions and with different orderings. The raw count of correct answers is therefore no longer directly comparable across test-takers. Instead, item response theory (IRT) models are used to construct estimates of ability and skill (also called "thetas") for each test-taker of the computer-based ASVAB. These IRT

[1] For example, *Administration of the CAT-ASVAB (NLSY97)* available at `https://www.nlsinfo.org/content/cohorts/nlsy97/topical-guide/education/administration-cat-asvab-0` (Accessed: 2020-06-23), and *Aptitude, Achievement & Intelligence Scores (NLSY79)* available at `https://www.nlsinfo.org/content/cohorts/nlsy79/topical-guide/education/aptitude-achievement-intelligence-scores` (Accessed: 2020-06-23)

estimates are supposed to be comparable across test-takers.[2]

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.

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

[2] 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 paper version of the same sections.

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 is 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).[3]

In Figures OA.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.[4]

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

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

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

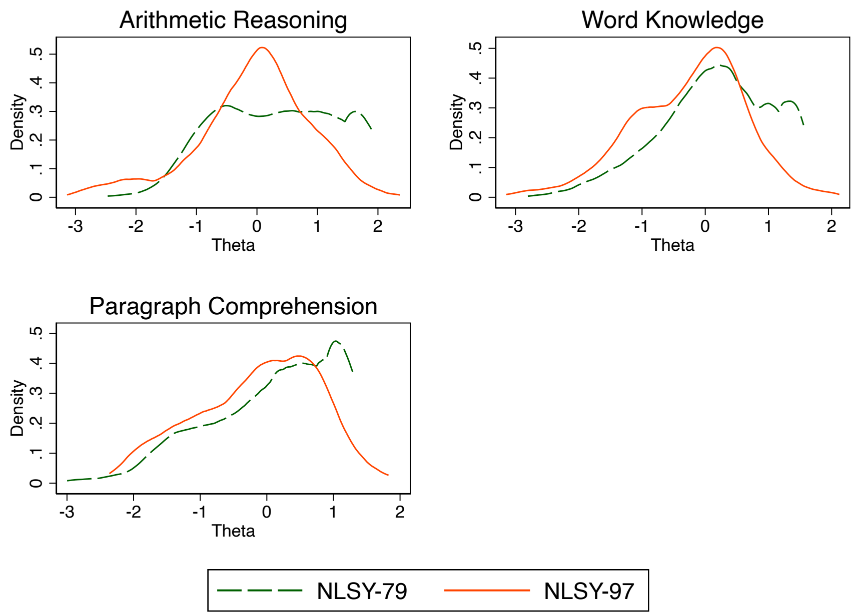
Comprehension) of Figures OA.1, suggesting that changes in the AAFQT scores across the cohorts do not seem to be a function of the concordance that was done to create AAFQT-equivalent scores for the 1997 cohort. That said, it is not obvious whether the "thetas" are directly comparable between the NLSY–79 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 OA.2 plots the standard errors of the estimated IRT scores ("thetas") for different ASVAB sections. As pointed out in past studies (Schofield, 2014; Jacob and Rothstein, 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 errors is not symmetric nor minimized around the middle of the theta distribution.[5]

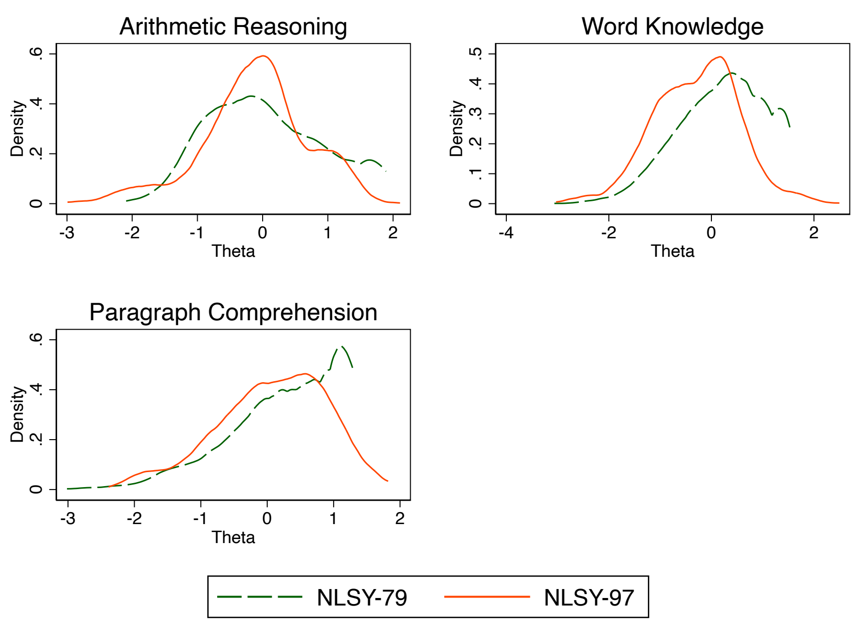
[5]We thank Dan Black for pointing this out to us.

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

(A) White Non-Hispanic Men



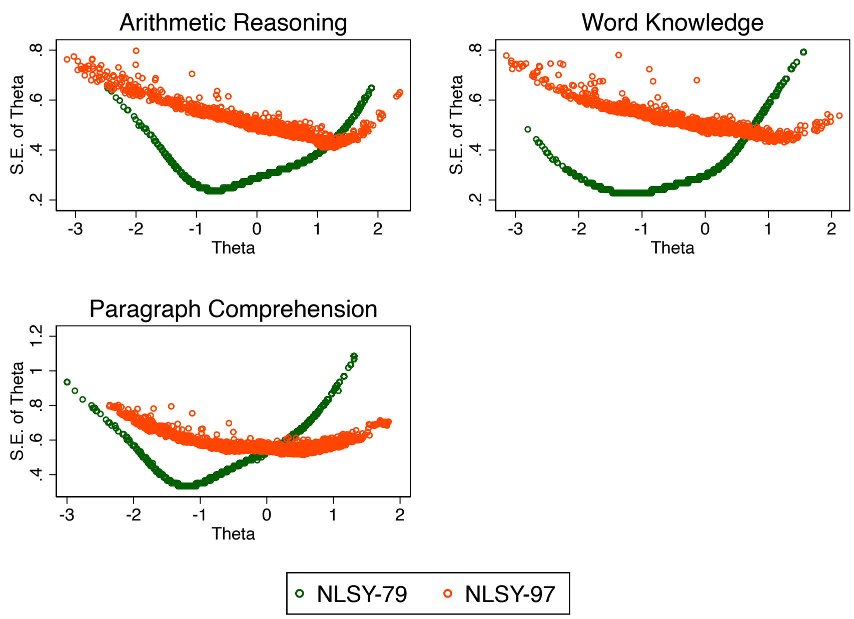
(B) White Non-Hispanic Women



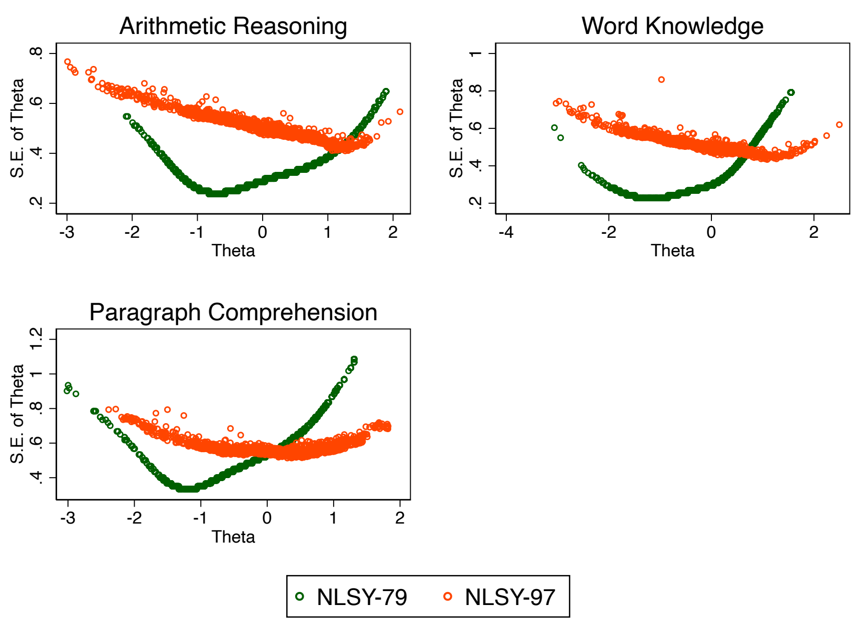
Note: In each figure, we plot the density of IRT-based ASVAB subsection scores ("thetas"), separately for the two cohorts. The NLSY–97 scores are original to the NLSY–97 data and not further processed to concord to the 1979 cohort. The NLSY–79 scores are constructed after-the-fact from the original paper-based 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 OA.2: Standard Errors of IRT-based ASVAB subsection scores for White Non-Hispanic Men and Women

(A) White Non-Hispanic Men



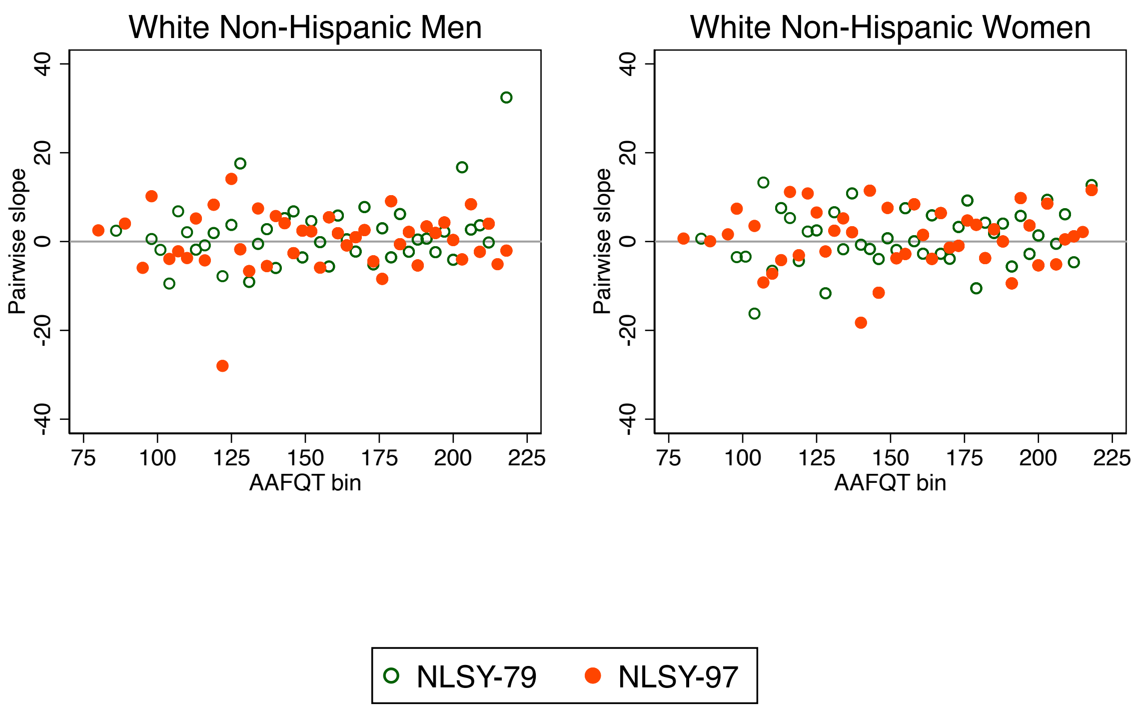
(B) White Non-Hispanic Women



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.

2 Supplemental Figures and Tables

Figure OA.3: Pairwise Slopes from Yitzhaki Decomposition



Note: In each figure, we plot the pairwise slopes ("betas") associated with each AAFQT bin, separately for white non-Hispanic men and women. Each bin contains three consecutive AAFQT points, and the slopes are averaged within each bin.

Table OA.1: Distribution Statistics of Arithmetic Reasoning score

	White Men		White Women	
	NLSY–79	NLSY–97	NLSY–79	NLSY–97
Mean	51.2	50.6	48.5	49.6
S.D.	9.5	10.0	9.0	9.2
Skewness	-0.14	-0.27	0.21	-0.03
Kurtosis	1.91	2.20	2.08	2.21
p1	32	28	32	30
p5	35	34	35	35
p10	38	36	38	38
p25	43	43	42	43
p50	51	51	47	50
p75	59	59	56	57
p90	64	64	62	62
p95	65	65	64	64
p99	66	66	66	66
Test of Equal Distribution	$p < 0.01$		$p < 0.01$	

Note: Distribution statistics of the ASVAB's Arithmetic Reasoning subsection scores (concorded by the authors following 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 OA.2: Distribution Statistics of 12th- and 10th-grade Math Score in NELS and ELS, Non-Hispanic Whites

	12th Grade				10th Grade			
	White Men		White Women		White Men		White Women	
	NELS	ELS	NELS	ELS	NELS	ELS	NELS	ELS
Mean	51.81	56.05	51.18	54.24	46.67	50.57	46.62	49.31
S.D.	13.82	13.23	13.03	12.07	13.53	12.62	12.82	11.59
Skewness	-0.27	-0.62	-0.29	-0.52	-0.17	-0.41	-0.22	-0.39
Kurtosis	2.11	2.71	2.29	2.63	1.99	2.55	2.16	2.50
p1	22.1	22.6	22.2	25.3	19.9	20.1	19.6	22.0
p5	28.0	29.6	27.6	31.4	24.1	27.4	24.0	28.0
p10	32.1	36.5	32.4	35.9	27.6	32.5	27.8	32.7
p25	41.2	48.1	41.8	46.7	35.6	41.8	37.3	41.4
p50	53.4	58.3	52.3	55.8	47.9	52.1	47.6	50.7
p75	63.5	66.3	61.8	63.4	57.6	60.1	56.8	58.3
p90	69.2	71.8	67.7	68.6	64.3	65.7	63.3	63.4
p95	71.9	74.1	70.4	71.2	67.1	68.9	66.0	66.2
p99	75.9	77.0	74.4	75.3	70.7	74 .6	70.0	70.8

Note: We present the distribution statistics of 12th-grade and 10th-grade math scores separately for NELS:88 and ELS:02, and for white non-Hispanic men and white non-Hispanic women. We use the NELS-equated math scores created by NCES. Sample weights are used.

3 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 recover the original OLS estimate by using the grouped data and weighting the grouped regression by 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 rather 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:

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

Note that when there are no duplicate observations ($N_i = 1$, for all i), the expression becomes $\operatorname{cov}(y,x) = \frac{1}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(x_i - x_j\right)\left(y_i - y_j\right)$, 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:

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

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 $\operatorname{var}(x)$:

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

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

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

and

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

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.

4 The Yitzhaki Decomposition Using NELS/ELS Data

This appendix provides evidence on the returns to cognitive skill using the NELS/ELS NCES data. Specifically, we examine how changes in the distribution of 12th-grade math scores across the NELS/ELS cohorts contribute to changing OLS estimates of wage returns to math scores.

In Table OA.3, we present the OLS estimates of univariate regressions of log hourly earnings on math scores, separately for white men and white women. The hourly earnings are measured as year 1999 for the NELS cohort and as year 2011 for the ELS cohort, which are seven years after high school graduation for both cohorts. We convert the earnings data to the 2013 price, and for ease of display, we multiply log hourly earnings by 100. As described in the paper, we use the math scores of 12th-graders.

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, consistent with what Weinberger (2014) documented in comparing NELS:88 with an older cohort.

We then perform the Yithaki decomposition on the baseline OLS estimates. Figure OA.4 plots the Yithaki weights together with the smoothed local linear regressions and Figure OA.5 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 (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 smoothed log hourly earnings 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 of the smoothed log hourly earnings 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.

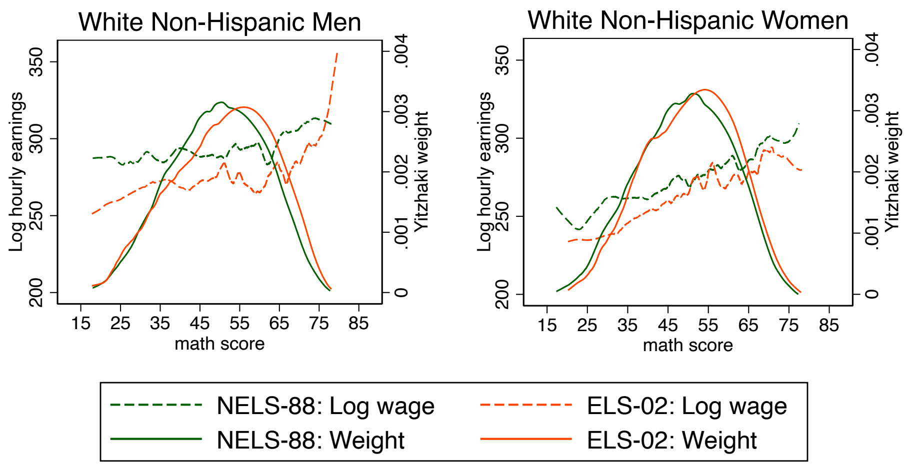
Regardless of the differences between the NLSY and the NELS/ELS in the samples and in the measures of cognitive abilities used, we find broad similarities in the evolution of cognitive test scores across cohorts. We also find wage returns to measured cognitive ability in the NELS that are consistent with the younger (and more similarly aged) NLSY–97. These results provide additional evidence that the changes in the distribution of AAFQT scores reflect real changes in cognitive skill, and that changes in the estimated returns to AAFQT are real and nuanced across the skill distribution.

Table OA.3: 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

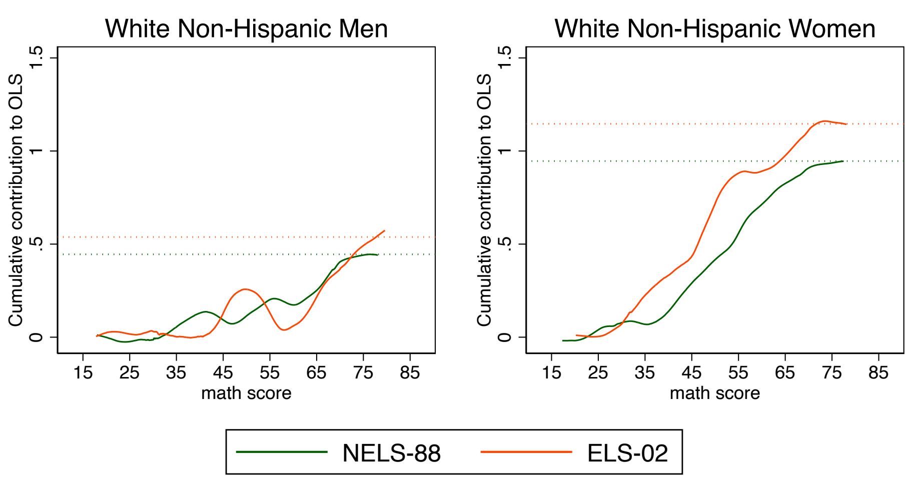
Note: We present OLS estimates of the wage returns to 12th-grade math scores in a univariate regression. The dependent variable is log hourly earnings measured seven years after high school graduation, which we multiply by 100 for the ease of display. We present results separately for NELS:88 and ELS:02, 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure OA.4: Smoothed Yitzhaki Weights and Slopes, NELS and ELS



Note: In each figure, we overlay the smoothed Yitzhaki weights (right axis) with the smoothed log hourly earnings (left axis), separately for the two cohorts. We multiply log hourly earnings by 100 for ease of display. We use math scores from 12th-graders. The smoothed curves are estimated using Locally Weighted Scatterplot Smoothing (LOWESS) with a tricube weighting function. A bandwidth of 0.1 is used for smoothing both the weights and the earnings. Sample weights are used.

Figure OA.5: 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.

5 Yitzhaki Decomposition under Non-linear Returns

In this appendix we describe an empirical strategy to implement the Yitzhaki decomposition under non-linear returns. Consider the following regression specification, which allows returns to differ above and below the median AAFQT score:

$$y_i = \beta_0 + \beta_1[A_i \times L_i] + \beta_2[A_i \times H_i] + \beta_3 H_i + \epsilon_i,$$

where A_i denotes the AAFQT score, $L_i = 1$ if the individual's score is below the median, and $H_i = 1$ if the score is above the median. Coefficients β_1 and β_2 capture the returns to AAFQT for individuals below and above the median, respectively. Table OA.4 reports the corresponding OLS point estimates. The point estimates of the returns are lower below the median than above the median, across both cohorts and both for men and for women. However, the differences in the returns across cohorts are statistically significant only for men and only below the median AAFQT score. This is all consistent with the results from our full sample Yitzhaki decompositions of linear returns.

To implement the Yitzhaki decompositions, note that β_1 and β_2 from the previous equation can be recovered using two double-residual regressions. First, consider the estimation of β_1. We residualize both y and $A \times L$ using $A \times H$ and H as controls. Then, we estimate:

$$\widetilde{y_i} = \theta_0 + \beta_1[\widetilde{A_i \times L_i}] + e_i \tag{1}$$

where

$$\begin{aligned}\widetilde{y_i} &= y_i - (\hat{\alpha_0} + \hat{\alpha_2}[A_i \times H_i] + \hat{\alpha_3} H_i) \\ &= \begin{cases} y_i - \hat{\alpha_0} & \text{if } L_i = 1, \\ y_i - (\hat{\alpha_0} + \hat{\alpha_3}) - \hat{\alpha_2} A_i & \text{if } H_i = 1 \end{cases}\end{aligned}$$

and

$$
\begin{aligned}
\widetilde{A_i \times L_i} &= [A_i \times L_i] - (\hat{\gamma_0} + \hat{\gamma_2}[A_i \times H_i] + \hat{\gamma_3} H_i) \\
&= \begin{cases} A_i - \hat{\gamma_0} & \text{if } L_i = 1, \\ 0 & \text{if } H_i = 1. \end{cases}
\end{aligned}
$$

where we use the fact that $\hat{\gamma_2} = 0$ and $\hat{\gamma_0} + \hat{\gamma_3} = 0$.

Similarly, we can recover β_2 from:

$$
\widetilde{\widetilde{y_i}} = \theta_0 + \beta_2[\widetilde{A_i \times H_i}] + u_i \tag{2}
$$

where

$$
\begin{aligned}
\widetilde{\widetilde{y_i}} &= y_i - (\hat{\lambda_0} + \hat{\lambda_1}[A_i \times L_i] + \hat{\lambda_3} H_i) \\
&= \begin{cases} y_i - \hat{\lambda_0} - \hat{\lambda_1} A_i & \text{if } L_i = 1 \\ y_i - \hat{\lambda_0} - \hat{\lambda_3} & \text{if } H_i = 1 \end{cases}
\end{aligned}
$$

and

$$
\begin{aligned}
\widetilde{A_i \times H_i} &= [A_i \times H_i] - (\hat{\delta_0} + \hat{\delta_1}[A_i \times L_i] + \hat{\delta_3} H_i) \\
&= \begin{cases} 0 & \text{if } L_i = 1 \\ A_i - \hat{\delta_3} & \text{if } H_i = 1 \end{cases}
\end{aligned}
$$

where $\hat{\delta_0} = \hat{\delta_1} = 0$.

From expressions (1) and (2) we can implement the Yitzhaki decompositions. Figure OA.6 below plots the cumulative contributions to the OLS estimates of β_1 and β_2, for white men (Panel a) and white women (Panel b). The OLS estimates from Table OA.4 are reported using the horizontal dotted lines.

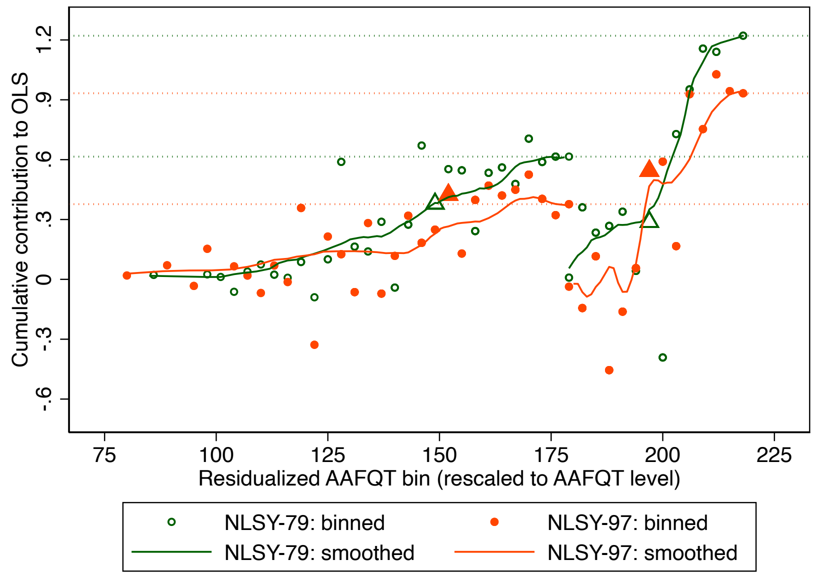
Unlike when computing OLS estimates via minimizing the sum of squared residuals, here, all observations contribute to the cumulative sums. More specifically, the above median AAFQT observations do contribute to the calculation of the below median slopes in the Yitzhaki decomposition. To see this, note that for the derivation the estimate of β_1 via the Yitzhaki decomposition applied to equation (1), the entire distribution of AAFQT scores, including those that are above the median prior to residualization, is used to construct Yitzhaki weights. Similarly, Yitzhaki slopes are constructed for the entire residualized distribution of log(wages). Mechanically, the observations with above median AAFQT scores are all stacked at $\widetilde{A_i} = 0$, and the average of their $\widetilde{y_i}$ across all above median observations is 0.

Table OA.4: OLS Results for non-linear returns to the AAFQT

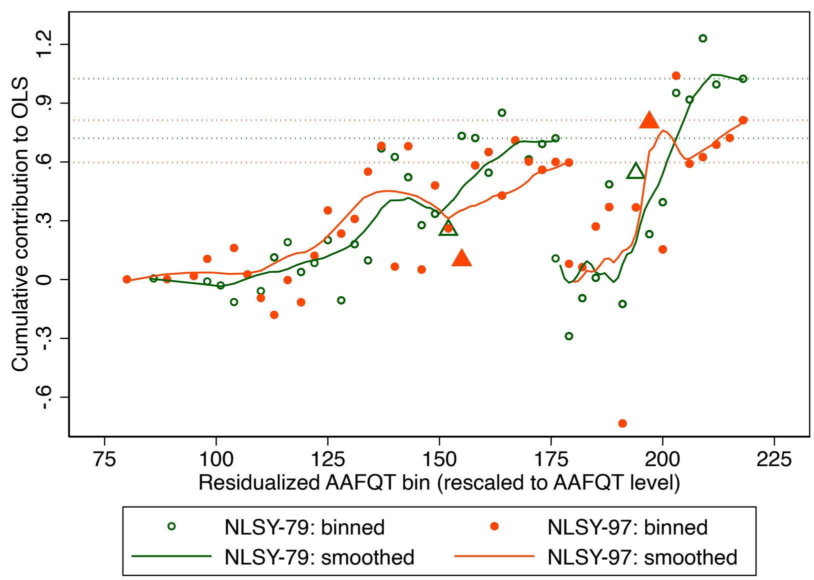
	White Men		White Women	
	NLSY–79	NLSY–97	NLSY–79	NLSY–97
AAFQT * Below Median	0.615***	0.377***	0.721***	0.597***
	[0.0606]	[0.0718]	[0.0680]	[0.0806]
AAFQT * Above Median	1.221***	0.933***	1.025***	0.813***
	[0.132]	[0.183]	[0.145]	[0.185]
Above Median	-117.5***	-105.5***	-53.84*	-30.66
	[27.40]	[37.62]	[30.35]	[38.27]
Constant	194.6***	233.8***	143.8***	172.7***
	[9.241]	[10.92]	[10.36]	[12.24]
Change from NLSY–79 for Below Median		-0.238**		-0.123
		[0.0939]		[0.105]
Change from NLSY–79 for Above Median		-0.289		-0.211
		[0.226]		[0.236]
Obs	2099	1584	2191	1488

Figure OA.6: Cumulative Contributions to OLS Estimates, Below and Above the Median

(A) White Non-Hispanic Men



(B) White Non-Hispanic Women



Note: To rescale $\widetilde{A_i \times L_i}$ and $\widetilde{A_i \times H_i}$, so that we can plot them on the same x-axis and in the same range as the original AAFQT distribution, we add the below-median AAFQT sample mean to $\widetilde{A_i \times L_i}$ and the above-median AAFQT sample mean to $\widetilde{A_i \times H_i}$. For calculating the estimate of β_1, the cumulative sum is 0 at the lowest below-median AAFQT score and rises to the estimated β_1 at the median of AAFQT. For calculating β_2, the cumulative sum begins anew at 0 at the median of AAFQT and rises to the estimate of β_2 when it reaches the top of the AAFQT distribution. Bins that include observations where $\widetilde{A_i \times L_i} = 0$ when calculating the cumulative sums for the below-median slopes, and again for bins that include $\widetilde{A_i \times H_i} = 0$ for calculating the above-median slopes are marked separately with triangle symbols.

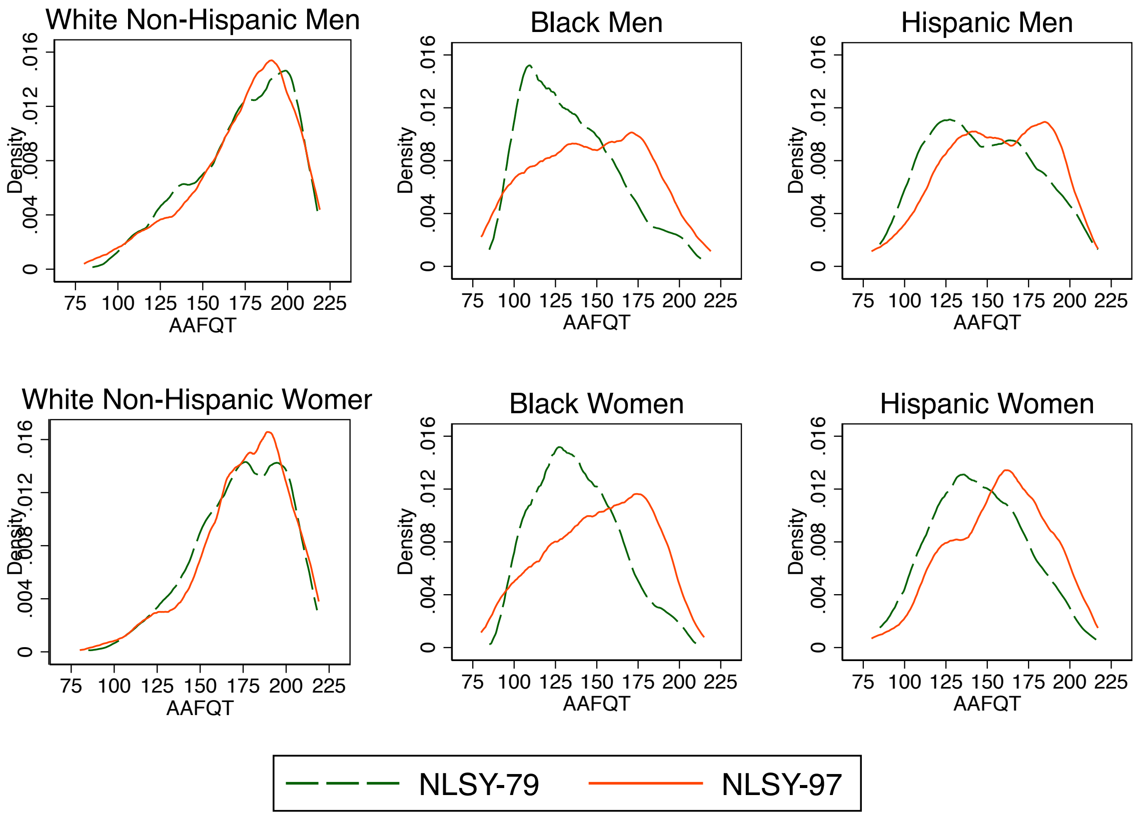
6 Tables and Figures for the Black and Hispanic Samples

Table OA.5: 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

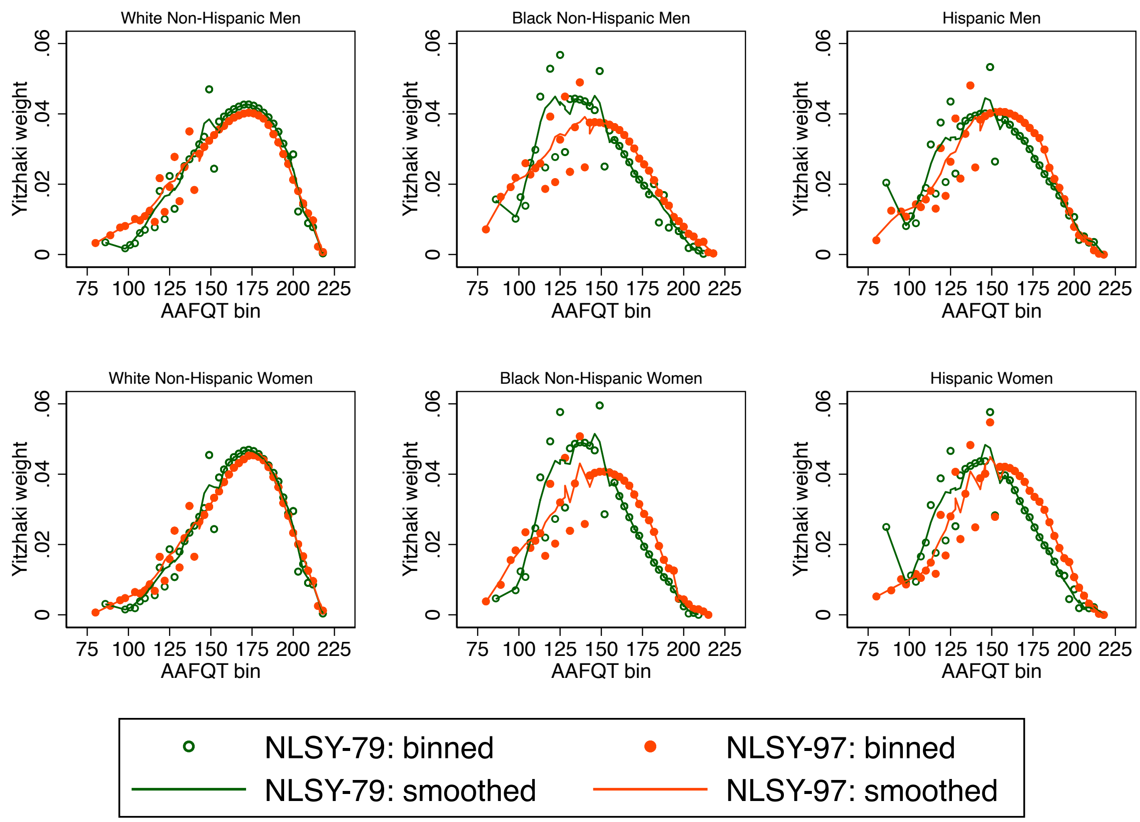
Note: We present OLS estimates of the wage returns to AAFQT scores, by gender and by race and ethnicity. The dependent variable is the log of average hourly wages, which we then multiply by 100. Panel A presents results for the regression of log wages on the AAFQT score, a dummy variable for the NLSY–97 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 OA.7: Adjusted AFQT Distribution By Gender and By Race and Ethnicity



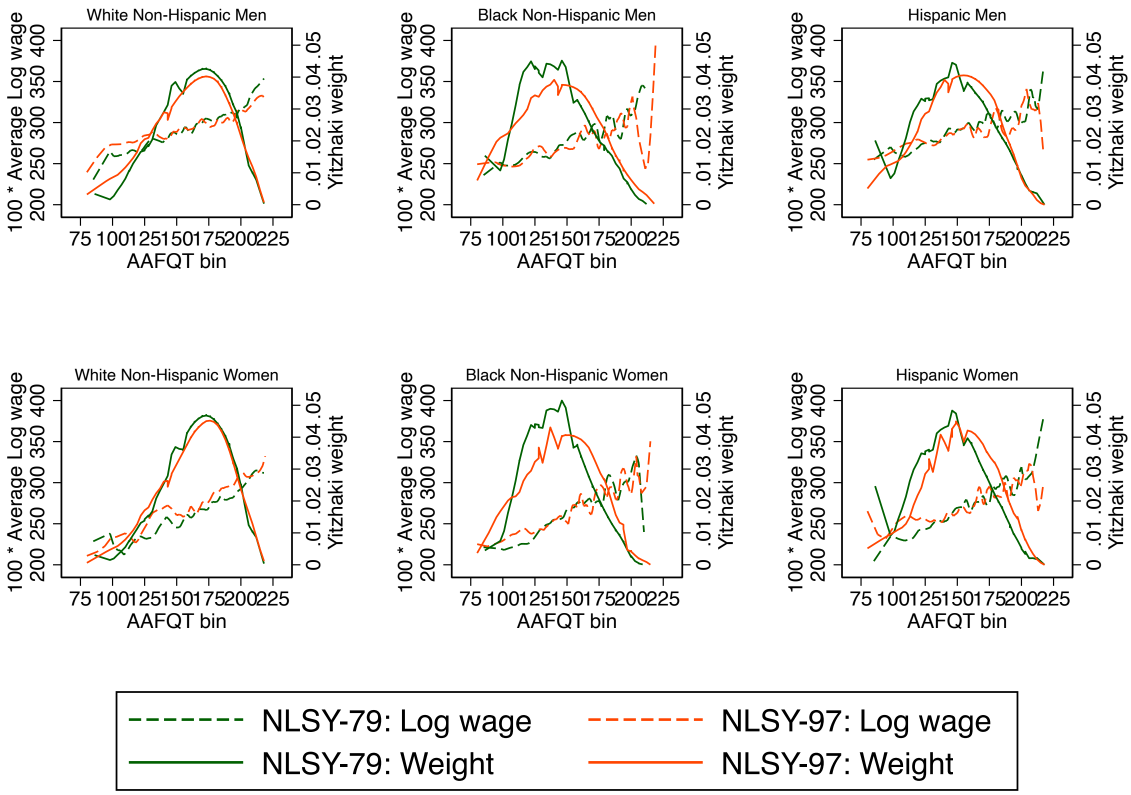
Note: We plot the density of AAFQT scores by gender and by race and ethnicity. AAFQT scores are concorded by Altonji et al. (2012). BLS custom sample weights are used.

Figure OA.8: Smoothed Yitzhaki Weights By Gender and By Race and Ethnicity



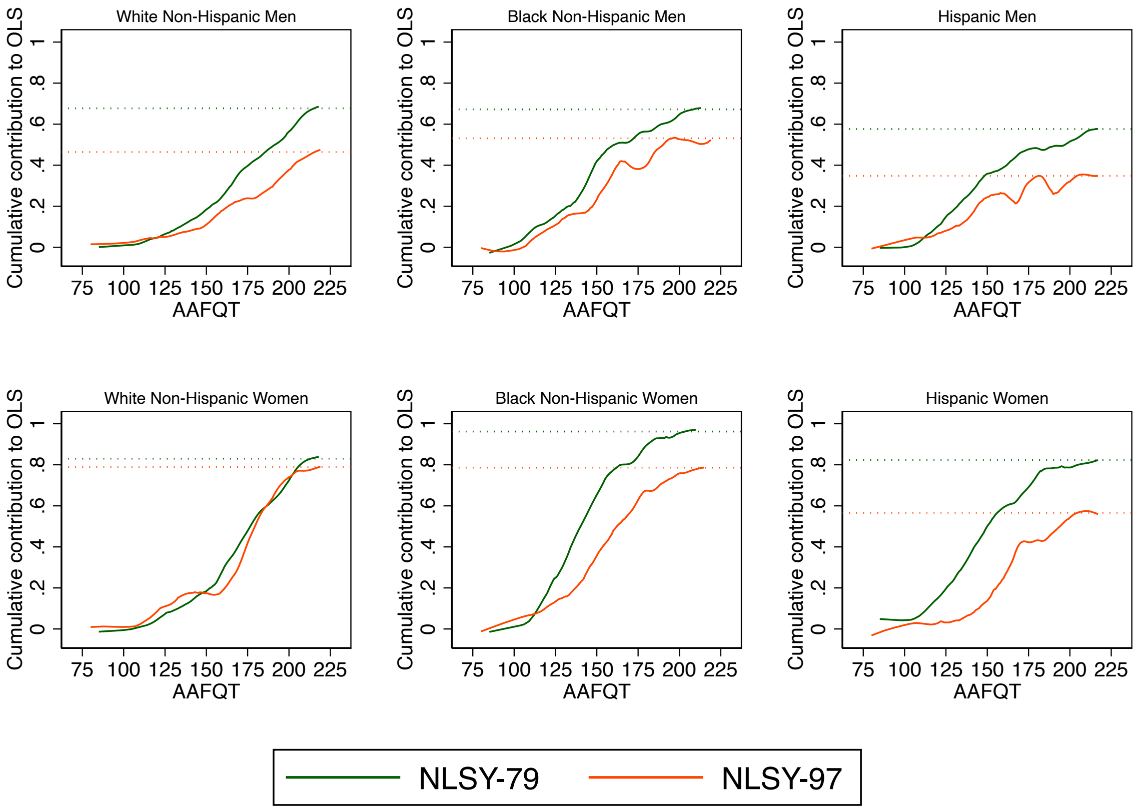
Note: We plot the Yitzhaki weights by gender and by race and ethnicity. The binned scatterplots are created by summing up Yitzhaki weights in each bin, which contains three consecutive AAFQT points. The smoothed curve is created using Locally Weighted Scatterplot Smoothing (LOWESS) with a tricube weighting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

Figure OA.9: 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 log wages (left axis), by gender and by race and ethnicity. We multiply log wages by 100 for ease of display. The smoothed curves are created using Locally Weighted Scatterplot Smoothing (LOWESS) with a tricube weighting function. A bandwidth of 0.1 is used for smoothing both the weights and the wages. BLS custom sample weights are incorporated in the Yitzhaki decomposition.

Figure OA.10: 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.

7 Robustness Check: The Role of Age

This appendix examines the role of age in potentially driving wage differentials across cohorts. First, Table OA.6 shows that the NLSY–97 cohort is slightly younger, about 0.7 years on average.

Table OA.7 reports the results from OLS regressions for white men (column 1) and white women (column 2). Panel A includes no controls; Panel B includes a control for the average age of the individual across the observed wage observations. A comparison of Panels A and B indicates that controlling for average age has little impact on any of the estimated returns to AAFQT. This is not surprising, as the correlation between average age and AAFQT is only 0.003 for the NLSY–79 and 0.073 for the NLSY–97. The interaction term between average age and AAFQT is positive for both cohorts and marginally significant for only women (Panel C). The cross-cohort change in the coefficient on the interaction term is insignificant for both men and women.

To further assess the influence of age, Table OA.8 reports results where the range of ages used to form the average hourly wage is restricted to wage observations for ages of 31–33 only. For comparison, we first present OLS estimates using wage observations over 25–39 (our baseline). The point estimates from the restricted age range are very similar to our baseline estimates and further confirm that differences in the ages for which wages are observed across the two cohort in our full samples of observations do not drive the baseline results.

Narrowing down the prime age window: To examine the robustness of our counterfactual exercises with respect to age, we report the results of implementing the decomposition in equation (5) using average wages for individuals aged 31–33. The results are as follows:

$$\begin{aligned}
\text{White men, 31-33}: \beta^{79}_{OLS} - \beta^{97}_{OLS} &= \Big(0.696 - 0.641\Big) + \Big(0.641 - 0.527\Big) = 0.055 + 0.114, \\
\text{White women, 31-33}: \beta^{79}_{OLS} - \beta^{97}_{OLS} &= \Big(0.818 - 0.643\Big) + \Big(0.643 - 0.881\Big) = 0.175 - 0.238.
\end{aligned}$$

The counterfactual results holding NLSY–79 weights fixed continue to support the baseline finding that changes in group composition account for a substantial share of the decline in OLS returns for white men, while amplifying the decline for white women.

Table OA.6: Age distribution in each NLSY sample

sample	Mean	SD	Min	p10	p25	p50	p75	p90	Max
NLSY–79	30.9	4.1	25	26	27	30	34	37	39
NLSY–97	30.2	4.0	25	25	27	29	33	36	39

Note: The sample used in this table includes each of the annual observations per individual that are averaged together for our wage analyses. Here, we pool white men and white women. BLS sample weights are used. The correlation between average age and AAFQT is 0.003 in the NLSY–79 and 0.073 in the NLSY–97.

Table OA.7: OLS Regressions with Age Control & Interaction

	(1) White Men	(2) White Women
Panel A: No Control		
AAFQT	0.677***	0.830***
	[0.0346]	[0.0363]
AAFQT * NLSY–97	-0.212***	-0.0407
	[0.0554]	[0.0597]
Obs	3683	3679
Panel B: Control for Age		
AAFQT	0.671***	0.825***
	[0.0344]	[0.0367]
AAFQT * NLSY–97	-0.226***	-0.0467
	[0.0551]	[0.0599]
Obs	3683	3679
Panel C: Interacted with Age		
AAFQT	-0.965	-0.316
	[1.004]	[0.613]
AAFQT * Age	0.0531	0.0369*
	[0.0326]	[0.0197]
AAFQT * NLSY–97	0.720	-0.641
	[1.432]	[1.066]
AAFQT * Age * NLSY–97	-0.0302	0.0208
	[0.0469]	[0.0350]
Obs	3683	3679

Note: Panel A uses our baseline specification without covariates. Panel B controls for average age, averaged across survey waves, and allows its coefficient to vary by cohort. Panel C further includes the interaction between AAFQT and average age, and also allows its coefficient to vary by cohort. Robust standard errors are reported in brackets. BLS sample weights are applied. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table OA.8: OLS Regressions with Average Wage over Different Age Ranges

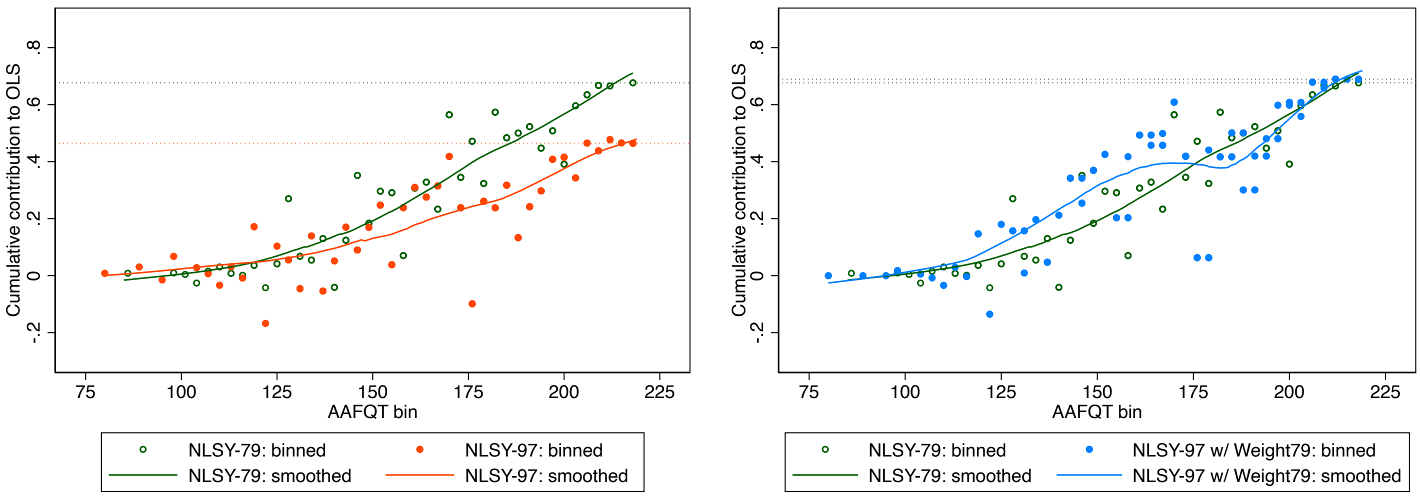
	(1) White Men	(2) White Women
Panel A: 25–39 (Baseline)		
AAFQT	0.677***	0.830***
	[0.0346]	[0.0363]
AAFQT * NLSY–97	-0.212***	-0.0407
	[0.0554]	[0.0597]
Obs.	3683	3679
Panel B: 31–33		
AAFQT	0.696***	0.818***
	[0.0416]	[0.0534]
AAFQT * NLSY–97	-0.169***	0.0629
	[0.0635]	[0.0812]
Obs.	3157	2926

Note: Panel A uses our baseline wage measure, averaging wage observations over ages 25–39. Panel B averages wage observations over ages 31–33. Robust standard errors in brackets. BLS sample weights are used. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

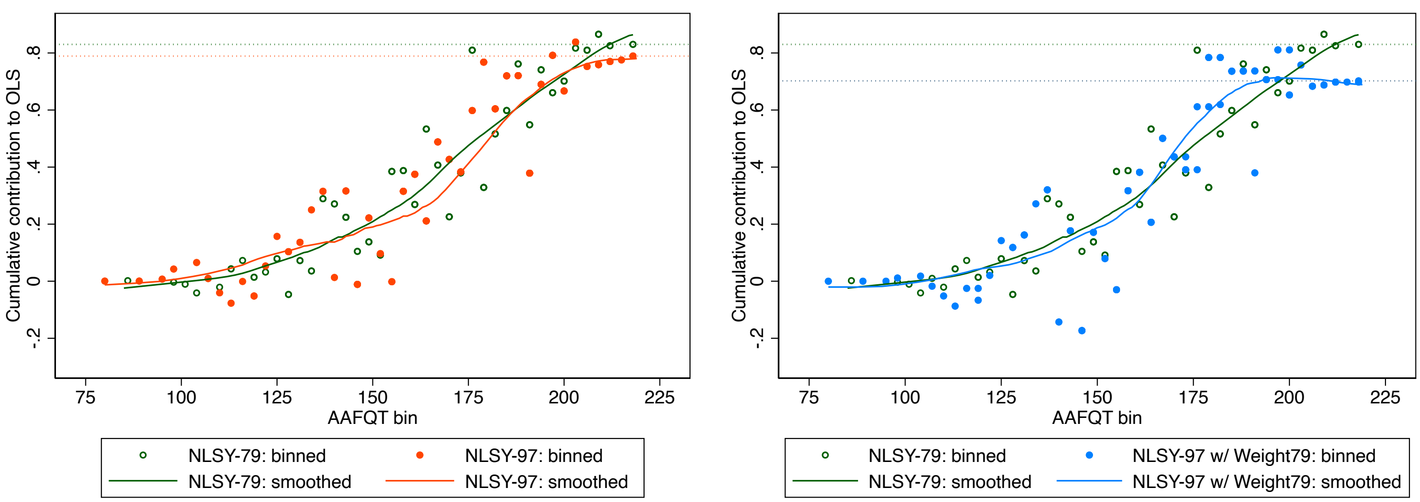
8 Robustness Check: Bandwidth Selection

Figure OA.11: Cumulative Contributions to Actual & Counterfactual OLS Estimates: Bandwidth of 0.5.

(A) White Non-Hispanic Men



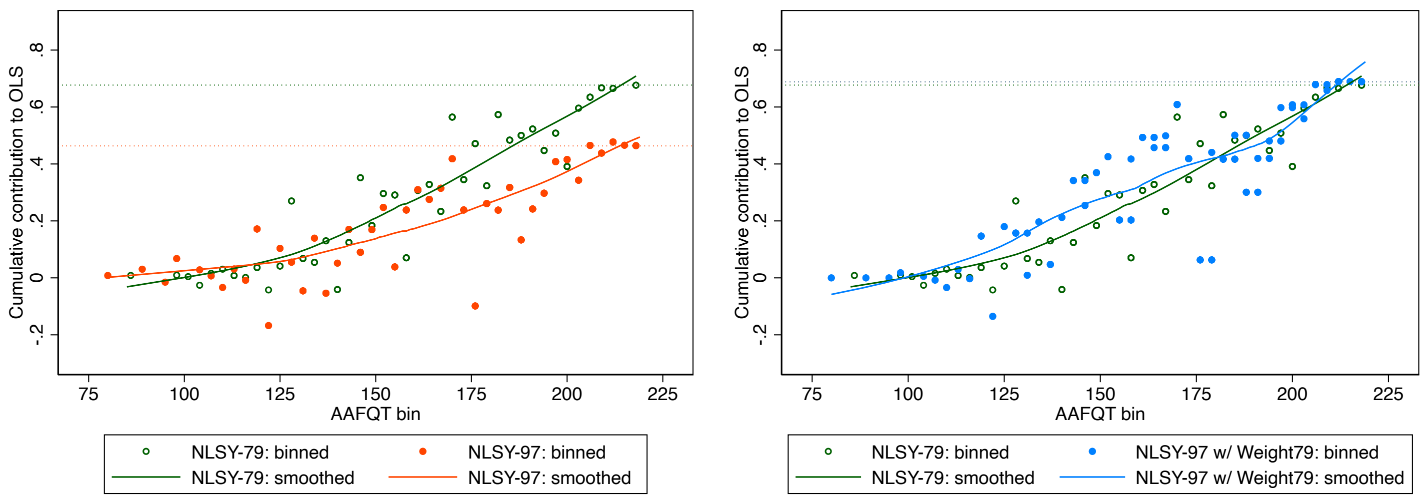
(B) White Non-Hispanic Women



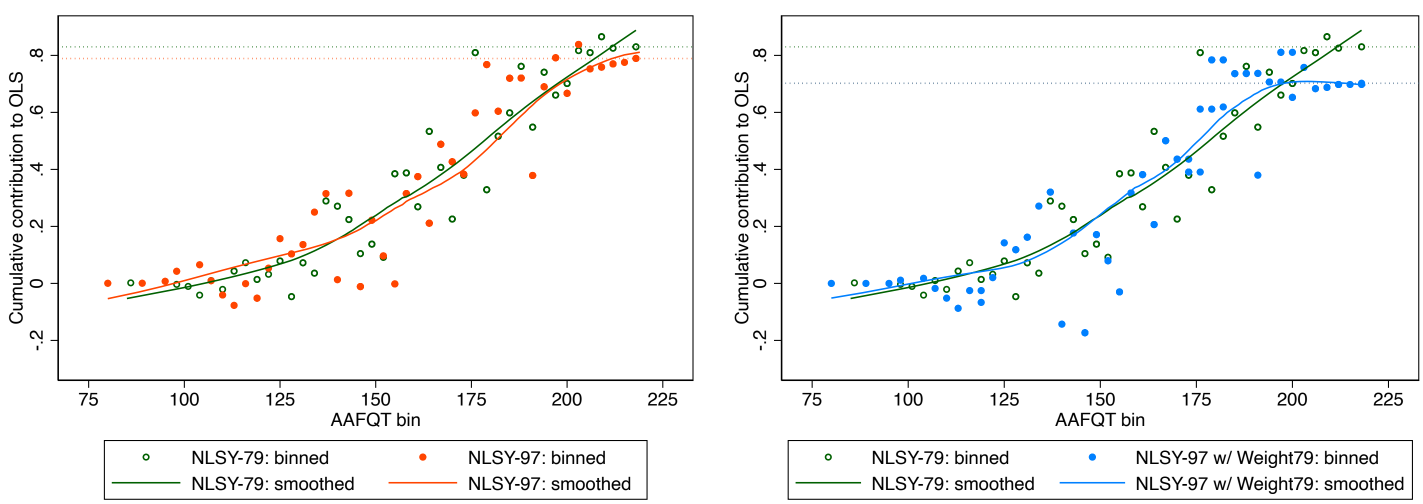
Note: In each figure, we graph the progressive sum of the Yitzhaki decomposition, starting with the lowest AAFQT score until the entire sum is calculated (producing the OLS estimate). The top left and bottom left panels plot the progressive sum for the actual OLS estimates of the two cohorts. The top right and bottom right panels plot the counterfactual estimate for NLSY–97 (using NLSY–79 weights) with the actual NLSY–79 OLS estimate. We graph the progressive sum by three-point bins and use a bandwidth of 0.5 for smoothing.

Figure OA.12: Cumulative Contributions to Actual & Counterfactual OLS Estimates: Bandwidth of 0.8.

(A) White Non-Hispanic Men



(B) White Non-Hispanic Women



Note: In each figure, we graph the progressive sum of the Yitzhaki decomposition, 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–79 weights) with the actual NLSY–79 OLS estimate. We graph the progressive sum by three-point bins and use a bandwidth of 0.8 for smoothing.

References

Altonji, J. G., P. Bharadwaj, and F. Lange (2009). *Constructing AFQT Scores that are Comparable Across the NLSY79 and the NLSY97.* Unpublished manuscript.

Altonji, J. G., P. Bharadwaj, and F. Lange (2012). Changes in the Characteristics of American Youth: Implications for Adult Outcomes. *Journal of Labor Economics 30*(4), 783–828.

Castex, G. and E. K. Dechter (2014). The Changing Roles of Education and Ability in Wage Determination. *Journal of Labor Economics 32*(4), 685–710.

Defense Manpower Data Center (2006). ASVAB Technical Bulletin. Technical report, Department of Defense.

Deming, D. J. (2017). The Growing Importance of Social Skills in the Labor Market. *The Quarterly Journal of Economics 132*(4), 1593–1640.

Heckman, J. J., J. Stixrud, and S. Urzúa (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics 24*(3), 411–482.

Ing, P., C. Lunney, and R. Olsen (2012). Reanalysis of the 1980 afqt data from the nlsy79. Technical report, The Ohio State University.

Jacob, B. A. and J. Rothstein (2016). The Measurement of Student Ability in Modern Assessment Systems. *Journal of Economic Perspectives 30*(3), 85–108.

Neal, D. and W. R. Johnson (1996). The role of premarket factors in black-white wage differences. *Journal of Political Economy 104*(5), 869–895.

Prada, M. F., G. Rucci, and S. Urzúa (2017). One Size Does Not Fit All: Multiple Dimensions of Ability, College Attendance, and Earnings. *Journal of Labor Economics 35*(S1), S339–S384.

Quester, A. and R. Shuford (2017). Population Representation in the Military Services: Fiscal Year 2015 Summary Report. Technical report, Department of Defense, Arlington, VA.

Sands, W. A., B. K. Waters, and J. R. McBride (Eds.) (1997). *Computerized Adaptive Testing: From Inquiry to Operation.* Washington, D.C.: American Psychological Association.

Schofield, L. S. (2014). Measurement error in the AFQT in the NLSY79. *Economics Letters 123*(2), 193–196.

Segall, D. O. (1997). Equating the CAT-ASVAB. In W. A. Sands, B. K. Waters, and J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation*, pp. 117–130. Washington, D.C.: American Psychological Association.