Test Questions, Economic Outcomes, and Inequality

Eric Nielsen*

Federal Reserve Board

This draft: April 2020

Abstract

Standard achievement scales aggregate test questions without considering their relationship to economic outcomes. This paper constructs new achievement scales by estimating the relationships between individual test questions and outcomes such as school completion, wages, and lifetime earnings. These new scales rank individuals differently than standard scales, yielding achievement gaps by race, gender, and household income that are 0.1-0.5 standard deviations larger in most cases. Test questions almost fully predict observed white-black differences in labor market outcomes, while predicting roughly half of these gaps by household income and almost none by gender.

Keywords: human capital, inequality, achievement gaps, measurement error
JEL Codes: I.24, I.26, J.24, C.2

*Danielle Nemschoff and Bryce Turner provided excellent research assistance. This paper benefited from discussions with Ralf Meisenzahl, Steve Laufer, Neil Bhutta, Jesse Bruhn, Paul Peterson, and seminar participants at the Federal Reserve Board and the NBER Education Meeting. The views and opinions expressed in this paper are solely those of the author and do not reflect those of the Federal Reserve Board or the Federal Reserve System. Contact: eric.r.nielsen@frb.gov.

1 Introduction

Human capital is fundamental to understanding labor market success, educational attainment, and many other economic outcomes. Group differences in human capital therefore play a key role in common explanations for inter-group inequalities along many dimensions. However, human capital is not directly observable, so researchers and policy makers often turn to achievement test scores calculated using standard psychometric methods. These scores correlate with economic outcomes across a variety of contexts, motivating their use as proxies. However, I argue in this paper that such scores do not measure human capital well because the skills that are rewarded by psychometric scoring systems do not correspond closely to the skills that predict economic success. I argue further that improved test scales based on the relationships between individual test questions and economic outcomes paint a very different picture of human capital differences across individuals and across groups defined by race, gender, and household income.

Every test scale can be thought of as a method for aggregating individual test items (questions) into a scalar index. Standard psychometric methods combine items without reference to the economic importance of the skills the items measure. Such methods may yield biased estimates of human capital inequality if different individuals or groups perform differentially better or worse on items that are systematically related to later success. To be clear, this is not a failing of psychometric methods per se – the problem lies in using them for a purpose for which they were not designed.

In this paper, I construct new achievement scales that more accurately measure human capital. Using recently available data on individual Armed Forces Qualifying Test (AFQT) items from the National Longitudinal Survey of Youth 1979 (NLSY79), I assess which items are most relevant for predicting high school and college completion, wages at age 30, and total lifetime labor earnings. I construct "item-anchored" scores by weighting the individual test items in proportion to how strongly they correlate with these outcomes, conditional on

the other item responses. In other words, I assign to each item a "skill price" based on that item's ability to predict an outcome, net of the other items. I then use these skill prices to compare human capital (achievement) differences across individuals and across different demographic groups. This use of item data is novel – even where outcomes have been used to anchor scores, item data have not been available or have not been used.[1]

Achievement does not have natural units, so the most fundamental way to compare different achievement scales is through their induced rankings of individuals. Even if they disagree cardinally, two scales might nonetheless rank test-takers identically. This is not the case in my application – the item-anchored scales rank youth very differently than the NLSY79-given scales, with ranking differences of 10-20 percentile points common. Some youth do well on items that are predictive of later success but poorly on uninformative items, resulting in a low given score and a high item-anchored score, while for others, the situation is reversed. The item-anchored and given scales disagree fundamentally about which test-takers are doing well and which are doing poorly.

I next turn to the estimation of achievement gaps by race, gender, and household income using the item-anchored scales. This requires that I use the item data in an additional, novel way to assess measurement error in the item-anchored scales. Adjusting for measurement error is important for calculating mean achievement gaps – simply taking the "naive" difference in the mean item-anchored scores between two groups will result in estimates that are biased towards zero. I adapt the empirical approach in Bond and Lang (2018), who use lagged scores and instrumental variables methods to undo this downward bias. I follow an analogous approach, but instead of lagged scores, I split the items into disjoint groups which I use to estimate two different item-anchored scales. One of these two scales then serves as my baseline measure of achievement, while the other is used to construct the instruments necessary to correct for the bias created by measurement error.

[1]See Cunha and Heckman (2008); Bond and Lang (2018); Chetty et al. (2014); Jackson (2018); Polachek et al. (2015), among many others. The empirical exercise in this paper is related to Bettinger et al. (2013) who show that only some of the ACT subtests are useful for predicting college performance and retention. Schofield (2014) is one of the few studies to make use of the item data in the NLSY79.

Achievement gaps calculated using the item-anchored scales are typically much larger than gaps calculated using the NLSY79-given scales. For instance, I estimate that white/black and high-/low-income (top and bottom household income quintiles) math and reading gaps are all around 1 standard deviation (sd) using the NLSY79-given scales, roughly in line with prior literature.[2] By contrast, these gaps are 0.2-0.5 sd larger when I anchor at the item level on wages at age 30 or lifetime earnings. Anchoring to high school completion leads to more modest increases in these gaps of 0.06-0.20 sd. Item-anchoring does not increase all gaps, however – the college item-anchored white/black math gap is a full 0.2 sd smaller than the given gap. The true extent of achievement inequality along many dimensions is masked by standard achievement scales.

The item-anchored gaps can be directly compared to the actual outcome gaps because they are in the same units. Test items predict larger white/black gaps in school completion than are actually observed, consistent with prior literature using standard psychometric test scales (Lang and Manove, 2011). By contrast, the item-anchored and actual white/black lifetime earnings and wage gaps are very similar – the sizable racial differences in these labor market outcomes can be almost fully explained by differential item responses. This result strengthens the headline conclusion in Neal and Johnson (1996). I both explain a greater share of the early-adult wage gap, and I additionally explain the gap in lifetime earnings, an outcome not studied in that paper.[3] The predictability of the white/black lifetime earnings gap suggests that racial differences in employment may also be predictable from measured achievement, in contrast to prior literature. Indeed, I find that applying my method to the outcomes studied in Ritter and Taylor (2011) yields item-anchored scales that predict most of the observed white/black male employment gap. Notably, I obtain all of these labor market results using only white men to construct the item-anchored scales, so that my findings do not depend on differences by gender and race in how items correlate with outcomes.

[2]See Neal and Johnson (1996); Reardon (2011); Downey and Yuan (2005), among many others.

[3]To be clear, Neal and Johnson (1996) could not have studied lifetime earnings because the NLSY79 respondents were too young to credibly estimate this outcome when their paper was written.

The item-anchored scales do less well at explaining outcome differences between youth from high- and low-income households. The item-anchored scales modestly under-predict school completion differences by household income while predicting half or less of the observed differences in wages and lifetime earnings. Similarly, the item-anchored scales do not predict well the observed differences in outcomes by gender – the predicted gaps in school completion are generally larger than what is observed, while the predicted gaps in wages and lifetime earnings are much smaller.

Finally, I show that the item-anchored scores resolve the "reading puzzle" – the phenomenon in which reading scores, though positively correlated with wages in isolation, are weakly or even negatively correlated with wages conditional on math scores.[4] In contrast to regressions using given scores, joint regressions of wages on item-anchored scores suggest a sizable role for both reading and math. Reading skills do seem to have explanatory power above and beyond their correlation with math skills, but this is not visible when reading items are aggregated using standard psychometric methods.

The conceptual framework and empirical results in this paper have implications for many research agendas in economics and social science. Psychometric scales are in widespread use across many disciplines, and often there is a conceptual wedge between what these scales measure and how they are used. This paper demonstrates that alternative measures better-aligned with a particular research application can yield notably different, and more credible, results. As there is nothing special about the applications studied here that should make them particularly sensitive to the way in which the items are used, these results suggest that key findings from many other literatures may also depend critically on the economically arbitrary aggregation of test items into achievement scales.

Beyond these cautionary conclusions, the paper provides a constructive way forward. It shows how to leverage item-response data to construct achievement scales better aligned with the outcome of interest and demonstrates how to use these same data to correct for

[4]See Sanders (2016); Kinsler and Pavan (2015); Arcidiacono (2004).

biases stemming from measurement error. Overall, the work presented here suggests that we should, where possible, devote more effort to better aligning our achievement measures with the outcomes we are truly interested in. This could take the form of weighting items based on how they relate to economic outcomes. It could also involve a more judicious selection of the items that go into an achievement scale in the first place.

The rest of the paper is organized as follows. Section 2 discusses the test item and outcome data in the NLSY79. Section 3 discusses the general empirical and conceptual framework, while Section 4 presents preliminary evidence that these items correlate quite differently with different outcomes. Sections 5 and 6 present the main empirical estimates, while Section 7 addresses the reading puzzle. Section 8 concludes. Section 9 at the end contains all of the tables and figures referenced in the paper.

2 NLSY79 Item and Outcome Data

The NLSY79 is a nationally representative survey that follows a cohort of youth aged 14-22 in 1979 through to the present. Each round of the survey collects extensive information on educational and labor market outcomes. Additionally, in the base year of the survey, respondents took the Armed Services Vocational Aptitude Battery (ASVAB), a widely used achievement test.[5] Item response data that code each item as correct or incorrect for each respondent are available for several of the ASVAB component tests.

The ASVAB consists of a series of subject- and skill-specific tests covering both academic and applied subjects. I make use of the item response data from the math and reading components of the ASVAB that together comprise the oft-studied Armed Forces Qualifying Test (AFQT). The given scores for these component tests are estimated using the three parameter logistic (3PL) model from item response theory (IRT). The math items come from the arithmetic reasoning (30 items) and mathematics knowledge (25 items) tests, while the reading items come from the paragraph comprehension (15 items) and word knowledge

[5]The ASVAB is primarily used by the United States military in making enlistment and personnel decisions. I therefore drop the subsample of the initial sampling frame that was designed to be representative of the military population. However, my results change very little if this group is included.

(35 items) tests.[6]

The math and reading items in the NLSY79 are particularly well-suited to my anchoring exercise. These were low-stakes and likely unfamiliar tests for most of the the NLSY79 sample. The item responses therefore are unlikely to reflect differential coaching/practice by race and socioeconomic background, as might be a concern with more widely-used or higher-stakes assessments. Moreover, the NLSY79 respondents were on the cusp of adulthood when the tests were administered. As such, the item responses can be viewed as summary measures of the skills these youth took into adulthood.

I use longitudinal data in the NLSY79 to construct various school completion and labor earnings outcomes.[7] For school completion, I use the highest grade completed reported at any point in the first 15 years of the survey, with high school defined as 12 or more grades completed and college as 16 or more completed. The first labor market outcome I construct is the average hourly wage at age 30 (wage_30), which I estimate by dividing total annual labor income by total hours worked for each survey round. I then average over the three rounds closest to each individual's age-30 round to smooth out transitory fluctuations. I focus on respondents at age 30 in order to study early career success at an age when formal education is mostly completed. The second labor market outcome I construct is the present discounted value of lifetime labor income (pdv_labor). The construction of pdv_labor is complicated because I do not observe labor income in all years for all survey respondents, either due to missing data, unemployment, or labor-force non-participation. To deal with such missing observations, I adopt a "pessimistic" imputation rule which assigns to each missing value the minimum labor income observed for the individual over the life of the survey. This rule is not meant to be realistic; rather, it will tend to compress the earnings distribution. It is also sensitive to the extensive margin of labor force participation, in contrast to wage_30, as well as to alternative imputation approaches that use more realistic fill-in rules.[8]

[6] I use the current definition of the math and reading subtests of the AFQT, rather than the definition which held in 1980.

[7] Please refer to Nielsen (2015b) for more details.

[8] In Nielsen (2015b), I study this pessimistic measure along with others which adopt alternative imputation

Table 1 presents the summary statistics of the main variables used in the analysis. Roughly 9% of the sample is missing every ASVAB component – I drop these roughly 1,500 individuals from the anchoring analysis.[9] My final analysis sample consists of 11,406 youth.[10]

3 Conceptual Framework

This section presents a framework to analyze test items and economic outcomes. For ease of exposition, I defer the discussion of the techniques I employ to handle measurement error in the estimation of item-anchored achievement gaps until Section 6.

Let $i \in \{1, \ldots, N\}$ index a sample of test-taking individuals drawn independently from some population. Denote by S_i the economic outcome of interest for individual i and by X_i all other non-test observables. The achievement test consists of M dichotomous items indexed by j. Let D_i denote the vector of item responses for individual i: $D_i = [d_{i,1}, \ldots, d_{i,M}]$ where $d_{i,j} = 1$ if i gets item j correct, and 0 otherwise. These items are combined using some framework (IRT in the NLSY79) to produce a standardized (mean 0, standard deviation 1) test score z_i. It is these standardized scores that are typically treated as direct measures of human capital rather than as estimated proxies. Such treatment introduces two distinct problems, both of which are remedied by item-level anchoring.

First, achievement has no natural scale – there is no way to determine whether a given score represents a lot of achievement or relatively little. Anchoring scores by estimating the

rules. Though different in levels, the resulting measures using different fill-in rules are highly correlated with each other. The pessimistic imputations are not pessimistic in the sense of maximizing the measured achievement gaps. Rather, the descriptor comes from the assumption that latent wages are very low in survey rounds where wage income is not observed. To project labor earnings to retirement, I assume that each individual's earnings growth after the latest survey round follows the education-specific growth rates from a pseudo-panel of male earnings constructed from the 2005 American Community Survey and that everyone exits the labor force 20 years after the latest survey round. A final complication is that the NLSY79 moved to a biennial format after 1994. I impute labor earnings for the odd-numbered years using linear interpolation after applying the pessimistic imputation rule to the survey years. I discount future income exponentially at a 5% annual rate.

[9] These individuals are also all missing their IRT-estimated ASVAB component scores. In additional cases where the items are not entirely missing, so that an IRT-estimated score is present, I set the missing items to 0, corresponding to "incorrect." For individuals who took the assessment (so that not all items are missing), blank (unanswered) items are coded as missing by the NLSY79. The assumption I make therefore is that leaving a question blank and getting the question incorrect are equivalent.

[10] I use the full range of ages in my analysis to get as large a sample as possible. However, restricting the sample to youth less than 18 years old at the survey start yields qualitatively similar empirical results.

relationship between z_i and S_i solves this indeterminacy by rescaling so that test scores are in directly interpretable "outcome" units. Interpretability is not the only reason to prefer the anchored scale, however. If the transformation from the given scale to the anchored scale is nonlinear, as is often the case empirically, statistics calculated using the two scales may disagree dramatically – they may even differ in sign.[11]

Second, and a key insight in this paper, the given scores represent a particular choice about how to map each of the 2^M possible sequences of item responses to a scalar measure of achievement. Psychometric methods select this map without reference to economic outcomes, so the resulting scores may obscure useful information about economically relevant skills contained in the item responses. Anchoring the given scores to outcomes does not address this concern, as the information loss occurs during the construction of the scores themselves.

I propose a framework that overcomes both of these conceptual challenges. Similarly to Bond and Lang (2018), I guarantee interpretability by defining achievement A_i as the expected value of the outcome S_i conditional on the item responses D_i and possibly on additional controls X_i:

$$A_i \equiv \mathbb{E}[S_i|D_i, X_i] \tag{1}$$

This implies $S_i = A_i + \eta_i$, where $\mathbb{E}[A_i\eta_i] = 0$ by construction.

This framework is somewhat different from the standard used widely in social science. Because achievement is simply defined as the expected outcome conditional on the item responses, achievement estimates will differ by definition when different outcomes are considered. Achievement on this view is inherently multidimensional and context dependent. This framework matches the typical conception of human capital, which is generally taken to have multiple dimensions (health, job-specific skills, etc.). Additionally, the differences between my framework and the psychometric standard are smaller than they initially appear. Both approaches collapse item-response vectors down to a scalar measure of achievement. The approaches simply differ in how much signal they take from a given item – item-anchored scales emphasize items that predict outcomes, while psychometric scales do not.

[11] See Nielsen (2015a); Schroeder and Yitzhaki (2017); Bond and Lang (2013).

Because only S_i is observed, A_i must be estimated. I thus suppose that $S_i = f(D_i, X_i) + \varepsilon_i$ for some function f. I then use estimates of f to construct outcome-denominated achievement scores. These item-anchored scores, given by $\hat{A}_i = \hat{f}(D_i, X_i) = \hat{\mathbb{E}}[S_i|D_i, X_i]$, are responsive to the relationship between individual test items and outcomes. Note that I include X_i in the anchoring relationship. In practice however, other than age indicators and a constant, I typically allow demographics to enter only through the selection of which subsample of the data I use to estimate the anchored scale.

It is necessary empirically to place restrictions on the class of functions considered for f because D_i can take on many possible values with even a moderate number of items. I consider only linear regression and linear probit models in this paper. Linearity effectively assumes that there are no interactions between items in the anchoring relationship. Although there is no a priori reason to rule out interactions, I find that allowing for them produces anchored scales quite similar to what I report here.[12] The item-anchored scales I estimate are therefore based on linear regression and probit models of the form[13]

$$S_i = D_i W + X_i \Gamma + \varepsilon_i, \text{ or } S_i = \Phi(D_i W + X_i \Gamma + \varepsilon_i). \tag{2}$$

The resulting estimated item coefficient vectors $\hat{W}$ should not be interpreted causally. Rather, the goal is simply to estimate $A_i = \mathbb{E}[S_i|D_i, X_i]$. Indeed, the only reason to assume any parametric model at all is because my sample size is not large enough for non-parametric anchoring. Given a large-enough sample, the item-anchored scale could be estimated as the sample average of S for each possible combination of item responses and controls. For this reason, my analysis focuses on the individual ranks and group-level achievement gaps implied by the item-anchored scales, not on the specific anchored item weights.

[12] Specifically, I estimate models that allow for two-way interactions, and I employ various regularization techniques to keep the number of estimated parameters reasonable relative to the sample size.

[13] Another concern might the overfitting. To address this, I estimate leave-one-out anchored scales, where the predicted outcome for a given individual is based on an anchoring regression which excludes that individual. The resulting item-anchored scales are very similar to my baseline scales – the correlations are typically well above 0.99, and the corresponding achievement gaps are likewise quite similar. Given this similarity, and given that conceptually I want to "grade" everyone in the same way (so that two test-takers with the same item-responses always receive the same score), I do not pursue leave-one-out estimates further.

The given math and reading scores in the NLSY79 are estimated using the workhorse three parameter logistic (3PL) IRT model. I review here this widely-used IRT model to make clear the differences between my approach and the standard approach taken in psychometrics to using item-level data. The 3PL model supposes that a test-taker with achievement θ_i answers item j with probability (i.e., has item response function)

$$P(d_{i,j} = 1|\theta_i, \alpha_j, \beta_j, \gamma_j) = \gamma_j + \frac{1 - \gamma_j}{1 + e^{-\alpha_j(\theta - \beta_j)}}.$$

The parameters $(\alpha_j, \beta_j, \gamma_j)$ characterize the item. The discrimination parameter α_j gives the maximum slope of the item response function – the higher is α_j, the more sharply does the item distinguish between individuals with similar levels of achievement. The item difficulty parameter β_j gives the location of this maximum slope – difficult (high-β) items distinguish between high-θ individuals, while less difficult items distinguish between lower-θ individuals. Finally, γ_j gives the guessing probability – the probability that a test-taker with minimal achievement answers the item correctly. Importantly, these IRT parameters are not mechanically connected to the skills covered by the item and are not estimated using data on economic outcomes. Two items that have the same estimated parameters will contribute equally in the determination of θ, regardless of how well they predict outcomes.

Although the framework and estimation clearly differ between IRT and item-anchoring, the ultimate goals of the two approaches are analogous. Item-anchoring seeks a scale that is interval with respect to some economic outcome. It constructs this scale using information on the joint distribution of item responses and the outcome. IRT likewise seeks a scale that is interval in the log-odds of a correct response to each test question, and it constructs this scale using the distribution of item responses.

4 Item-Level Analysis

Before delving into the analysis of the item-anchored scales, I first study the relationship between individual test items and various economic outcomes. I present evidence that test items differ widely in how strongly they predict different outcomes. Moreover, the items that

are particularly predictive of one outcome may not be predictive of other outcomes. Finally, the psychometric parameters that characterize each item's contribution to the NLSY79-given scale are only modestly related to that item's ability to predict a given outcome. Taken together, these results demonstrate that item-anchoring and psychometric methods make use of the item response data in very different ways, raising the possibility that they will disagree about individual and group achievement comparisons.

4.1 Item-Outcome Regression Coefficients

Figure 1 shows the distributions of the estimated coefficients for bivariate regressions of indicators for each test item on school completion and labor market outcomes, where each outcome is converted to standard-deviation units for comparability. The panels of Figure 1 show that every item is positively associated with every outcome. However, there is quite a range in the estimated coefficients. For instance, using standardized college completion as the outcome yields math item coefficient estimates that range from about 0.3 sd to 0.8 sd. For each outcome, one can easily reject zero for most item coefficients, and one can also easily reject that all of the coefficients are equal. Comparing the distributions, a relatively greater number of math items than reading items are strongly associated with college completion. The situation for high school completion and lifetime earnings is reversed – reading items seem to be particularly predictive of these outcomes.

Items that are predictive of one outcome are often not predictive of other outcomes. This is made clear in Figure 2, which plots the item-level coefficients for different outcomes against each other. The top left panel shows that the college completion item coefficients are not very correlated with the high school completion item coefficients.[14] Even math items with college coefficient estimates in the right tail have high school coefficients no higher or lower than the average. By contrast, the top right panel shows that the ln(wage_30) and ln(pdv_labor) item coefficients are positively correlated – the items that predict lifetime earnings also, on average, predict wages at age 30.[15] Even here there is variation, with some items strongly

[14]The correlations are 0.11 for math and 0.37 for reading.

[15]The correlations are 0.83 for math and 0.93 for reading. To some degree, these positive associations

predictive of only one labor market outcome. The bottom two panels show that regressions that include all items simultaneously yield similar conclusions.

4.2 Item-Outcome Correlations and IRT Parameters

A natural question is whether the highly predictive items have any discernible commonalities. My ability to investigate this question is limited – the only item-level information available in the NLSY79 comes from the estimated 3PL IRT parameters used in the construction of the given scales. I assess how these IRT parameters are related to outcomes by estimating models which relate item-level outcome regression coefficients to the IRT parameters. Specifically, I estimate regressions of the form[16]

$$\hat{W}_j = \delta_0 + \delta_1 \underbrace{\text{discrimination}_j}_{\alpha_j} + \delta_2 \underbrace{\text{difficulty}_j}_{\beta_j} + \delta_3 \underbrace{\text{guessing}_j}_{\gamma_j} + \varepsilon_j, \tag{3}$$

where $\hat{W}_j$ is the first-stage regression coefficient from a model relating an indicator for test item j to an outcome (school completion, wage at 30, etc.) standardized to have a mean of zero and standard deviation of one. I pool math and reading items in these regressions – the results are similar but less precise if I consider math and reading separately. The relatively small number of items (105 in total) limits somewhat the depth of analysis. Despite this, the estimates of equation 3, presented in Table 2, are suggestive and largely intuitive.

Table 2 shows that the IRT parameters can partially predict which test items will be strongly correlated with economic outcomes. In particular, columns (1) and (3) show that items that have high discrimination, low difficulty, and low guessing probability tend to be more strongly associated with labor market outcomes. Column (5) shows similar results for high school completion. Column (7), by contrast, finds that items with high difficulty tend to be more strongly associated with college completion, with item discrimination and guessing probability entering as before. The even-numbered columns repeat the analysis for

are mechanical, as wages at age 30 enter directly into the estimation of pdv_labor. However, much of the variation in ln(pdv_labor) is not driven by ln(wage_30) – the correlation between the two is only 0.57 and bivariate regressions show that ln(wage_30) can explain roughly 33% of the variation in ln(pdv_labor).

[16]Please see Section 3 for definitions of these parameters in the context of the 3PL model. I transcribed the IRT parameters manually from the NLSY79 codebook.

coefficients estimated in joint regressions of each outcome on all of the items simultaneously. Very few significant links are apparent in these "full" regressions other than item difficulty, which is negatively related to item predictiveness for most outcomes other than college completion, where the association is positive.

The results for item discrimination and guessing probability are quite intuitive. A highly discriminating item operates like a threshold – test-takers with achievement above some cutoff are very likely to get the item correct, while those below are very unlikely. Such an item serves as a clear, discrete measure of the skill being tested and should therefore be correlated with an outcome if the tested skill is itself associated with the outcome. Conversely, easy-to-guess items do not distinguish clearly between high achievers, who got the question correct because they knew the answer, and low achievers, who happened to guess well. Such items should therefore have comparatively weak correlations with outcomes, which is what we find.

Less intuitive is the finding that less difficult items are more predictive of earnings and wages. One possible explanation for this result is that more difficult items may be more purely academic and may thus simply not measure economically relevant skills. The positive relationship between item difficulty and college completion is consistent with this hypothesis – difficult items predict advanced school completion but not earnings. Alternatively, easier items might measure general, foundational skills that are widely applicable in many settings.

4.3 Item-Outcome Correlations: Skills or Confounders?

The weights used in the construction of the item-anchored scales are not causal. As I argued in Section 3, the purpose of the anchoring regressions is to construct reasonable approximations of $\mathbb{E}[S_i|D_i, X_i]$, not to determine the causal effect of getting a given item correct on economic outcomes. The responses to a particular item will therefore affect the anchored scales if that item is correlated with outcomes for any reason, including reasons that have little to do with human capital. For example, an item that is heavily correlated with family income will be given a lot of weight simply because family income predicts most

outcomes, even if the skill assessed by the item is itself useless.[17] Without causal estimates of the skills measured by the different test items, I cannot rule out that "confounded" items like this are an important contributor to my empirical results.[18]

However, a number of observations mitigate this concern. First, even if some items are correlated with outcomes through confounders, the empirical question answered by the method is still important in its own right: how much of the observed gap in outcomes between different groups can be predicted on the basis of test items recorded in youth? Second, the items in the NLSY79 version of the ASVAB went through extensive vetting for content, applicability, difficulty, and absence of cultural bias (Bock and Mislevy, 1981). Third, I find very similar gap estimates using scales anchored on different demographic groups, suggesting that confounders correlated with demographics are playing at most a modest role.[19] Fourth, the item-anchored scales do a poor job predicting gender differences in labor market outcomes. Since the large labor market differences between men and women are likely driven primarily by non-skill factors (work disruptions due to children, occupational sorting, discrimination, etc.), these poor predictions serve as a reality check for the method.

Fifth and finally, a direct comparison of the item weights across scales estimated on different demographic subsets of the data suggests that the item weights are more reflective of skills than confounders. Table 3 compares the anchored weights for each item estimated on different demographic subsamples of the data (men, high-income, etc.) for each economic outcome. In almost all cases, one cannot reject, even at very low confidence levels, that an item's weight for a given anchor is the same across the different demographic groups. This suggests that the items really do correspond to outcome-relevant skills and that the item weights are more plausibly viewed as skill prices rather than confounded correlations.

Table 3 also compares the item weights across different outcomes holding the demographic

[17]Note that restricting the estimation to white men, as I do in my preferred specifications using labor market outcomes as anchors, will not generally solve this problem and could even make it worse.

[18]It is worth noting that this problem is not unique to item anchoring; it arises also in any analysis linking psychometrically derived achievement measures to outcomes (e.g., high IRT math scores might predict earnings spuriously through some confounder).

[19]Compare the labor estimates in Tables 7 - 9 and the school completion estimates in Tables 5 and 6.

group used for anchoring fixed. These comparisons suggest that anchoring to economic outcomes makes a substantive difference for which items are emphasized – one can typically reject the null that a given item's weights are equal (in sd terms) across different outcomes. Notably, this result holds when comparing weights estimated using economic outcomes to weights estimated using the given psychometric scores themselves as the anchors. The differences between the item-anchored and the given scales do not arise simply through estimation error but also through systematic differences in how the items are weighted.

5 Empirical Results – Scale Comparisons

I now compare the item-anchored scales to the NLSY79-given scales. This comparison yields two important results. First, economic outcomes are not linearly related to the given scores. Second, the item-anchored scales rank individuals very differently than the given scales. Both of these results suggest that there is scope for the item-anchored scales to disagree with the given scales about group differences in achievement.

Figure 3 plots the item-anchored scores against the given scores. The x-axis for each panel plots the given scores in standard deviation units. The y-axis plots the mean (solid black line) as well as the middle-50% and middle-90% ranges of the item-anchored scores, also in standard deviation units, for each ventile of the given score distribution. The first thing to note in Figure 3 is that the conditional means of the item-anchored scales have nonlinear relationships with the given scales, particularly for school completion.[20] As noted in previous literature, these nonlinearities mean that standard estimators (mean differences, OLS, etc.) using the given scales may produce severely biased estimates.[21]

The nonlinearities in the school completion scales are intuitive. The college item-anchored math and reading scales have convex relationships to the given scales; differences in achievement at the bottom ends of the given scales do not translate to differences in college completion, while differences at the top do. The situation is reversed for high school – improvements

[20]The non-linearities evident in the school completion panels are not simply an artifact of the probit models used to estimate the item-anchored scales – linear models (not shown) produce similar patterns.

[21]See Bond and Lang (2013); Schroeder and Yitzhaki (2017); Jacob and Rothstein (2016); Nielsen (2015a).

at the bottom ends of the given scales translate strongly to changes in high school completion, while improvements at the top are not very valuable. Low-achievement youth are not likely to be on the margin for completing college, so the college-anchored scales are flat for them. At the same time, these individuals are more likely to be on the margin for completing high school, explaining the steep anchored relationship for below-average given scores.

The item-anchored scales are closer to being linearly related to the given scales for ln(wage_30). This means that the relationship between observed scores and expected wages is convex – improvements at the top of the given math and reading scales yield out-sized gains in wages. The ln(pdv_labor) scales show some concavity – the conditional means are roughly linear for lower given scores before flattening at the top end.

The middle 50% and 90% ranges in each panel of Figure 3 show that there is a fairly wide distribution of item-anchored scores associated with each given score ventile. Individuals whose item responses led to a similar given score can have very different predicted outcomes based on their particular item responses. For college completion, this variation is greater at the top of the observed score distribution. For instance, among youth with given math scores about 1 sd above the mean, the middle 90% of the item-anchored scores cover almost 2 sd on the item-anchored scale, while for those 1 sd below the mean, the corresponding range is only about 0.2 sd. Some apparently high-performing youth are actually predicted to have very low rates of college completion, while others are even more likely to finish than their high given scores would indicate. Low-performing youth according to the given scale, however, are always predicted to have low rates of college completion. The pattern for high school is reversed – there is a lot of variation in the item-anchored scores at the bottom of the given scale and very little variation at the top. In contrast, the ranges of the ln(wage_30) and ln(pdv_labor) item-anchored scales appear to be fairly constant across the given score distributions. The range of item-anchored scores in each given score ventile again tends to be quite large. For example, the range of log(pdv_labor) scores for youth at the mean of the given reading distribution is about 2 sd.

The 50% and 90% ranges depicted in Figure 3 imply that the item-anchored scales rank test-takers very differently than the given scales. Indeed, Figure 4, which plots the differences in percentile ranks according to the item-anchored and given scales, shows that it is not uncommon for the ranking of a given test-taker to differ by 10 to 20 percentile points. Notably, the labor outcome scales display more rank shuffling relative to the given scale than the school completion scales. This is intuitive, as math and reading items should be more closely aligned to academic performance than to success in the labor market.

The item-anchored scores do not just disagree with the given scores about how valuable achievement is, they disagree fundamentally about which individuals are performing well or poorly. It is worth noting that such a disagreement could not happen if I instead anchored the given scores directly to outcomes, as has been done in prior literature. Such a procedure will generally yield some (non-linear) monotone translation from the given scale to the anchored scale, so that all rank orders are preserved.

6 Empirical Results – Achievement Gaps

In this section, I turn to the measurement of achievement gaps using the item-anchored test scales. I show how to leverage the item response data to estimate the amount of measurement error in these scales. That is, I use the item response data in two ways: first to estimate the item-anchored scales and second to estimate the reliability (signal-to-noise ratio) of these scales. Adjusting for measurement error, I find that the item-anchored scales typically imply much greater achievement inequality than the NLSY79-given scales. Moreover, I find that the item-anchored scales can explain significant fractions of the observed differences in school completion and labor market outcomes by race and household income.

6.1 Achievement Gaps and Measurement Error

The item-anchored test scores are estimated with error, which implies that achievement gaps estimated using the unadjusted difference in the mean item-anchored scores across groups will be biased towards zero. I adapt the approach developed in Bond and Lang (2018) to

handle this problem. The basic idea is to construct two independent anchored scales that can be used in an instrumental variables setting to undo the bias generated by the measurement error. The strategy of using instruments to recover the relevant "shrinkage term" (to be explained below) is from Bond and Lang (2018), who use lagged scores to construct the instruments. My innovation lies in leveraging the item response data, rather than lagged scores, to construct the needed instruments.

Let h and l denote two groups whose achievement we wish to compare. The goal is to estimate $\Delta A_{h,l} \equiv \bar{A}_h - \bar{A}_l$, where $\bar{A}_g$ is the population average achievement of group g. Each individual's achievement is measured with error: $\hat{A}_i = A_i + \nu_i$. Here, ν_i is error that in principle can come from estimation error in the anchoring relationship (i.e., $\hat{W} \neq W$) or from miss-specification in the anchoring relationship (i.e., $\mathbb{E}[S|D, X]$ is not linear). However, my maintained assumption is that the population anchoring relationship truly is linear, so that ν_i comes only from estimation error.[22] Then, if $A \sim N(\bar{A}, \sigma_A^2)$ and if $\nu_i \sim N(0, \sigma_\nu^2)$ iid in the population,[23]

$$\mathbb{E}[S_i|\hat{A}_i] = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_\nu^2}\hat{A}_i + \frac{\sigma_\nu^2}{\sigma_A^2 + \sigma_\nu^2}\bar{A}. \tag{4}$$

Equation (4) says that the expected outcome of individual i conditional on the item-anchored achievement $\hat{A}_i$ is "shrunk" towards the population mean achievement $\bar{A}$. This is intuitive – tests are noisy, so the best guess about an individual's true achievement gives weight to both the realized score and the population expected value, with the observed score given more weight the less noisy it is.

A naive estimator for $\Delta A_{h,l}$ is the difference in mean item-anchored scores. However, equation (4) shows that this estimator will be biased towards 0. Letting $\hat{A}_h - \hat{A}_l$ denote the sample difference in the mean item-anchored scores,

[22] Section 3 argues that the estimated scales under linearity are quite similar to the estimated scales which allow for item-by-item interactions.

[23] Under the assumption that $\mathbb{E}[S|D, X]$ is linear, v_i will be approximately normal under standard assumptions. See Bond and Lang (2018) for an extensive discussion and demonstration of the robustness of the general approach outlined here to violations in the normality assumption on A_i.

$$\text{plim}(\hat{A}_h - \hat{A}_l) = R_{A,\nu} \times \Delta A_{h,l}, \quad R_{A,\nu} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_\nu^2}. \tag{5}$$

The intuition for equation (5) is that, for each test-taker, measurement error implies that the best guess of achievement should be shaded to the population average. However, for a group mean, this measurement error is immaterial, so the shading towards 0 is not needed.

Given a consistent estimator of $R_{A,\nu}$, one could use equation (5) to recover a consistent estimate of $\Delta A_{h,l}$. A biased estimator of $R_{A,\nu}$ is given by γ in the regression

$$\hat{A}_i = \kappa + \gamma S_i + \epsilon_i. \tag{6}$$

The OLS estimate of γ is biased toward 0 because S_i is also a noisy measure of A_i. This classic errors-in-variables problem is solvable with an appropriate instrument for S_i – a variable ζ_i that is correlated with A_i and uncorrelated with η_i.

The item response data allows for the construction of many such instruments simply by estimating different item-anchored scales using disjoint subsets of the test items. For example, if the items are partitioned into two groups (1) and (2), equation (2) can be estimated separately on each group to produce anchored scores $\hat{A}_i^{(1)}$ and $\hat{A}_i^{(2)}$. Each of these scores is a noisy measure of A_i. Now consider estimating equation (6) using $\hat{A}_i^{(1)}$. An instrument for S_i in this equation is the average S among test-takers who are not i but who nevertheless have the same value of $\hat{A}_i^{(2)}$. That is, an instrument using item group (1) for the main scale is

$$\zeta_i^{(1)} = N_i^{-1} \sum_{i' \neq i : \hat{A}_i^{(2)} = \hat{A}_{i'}^{(2)}} S_{i'}, \tag{7}$$

where N_i is the number observations not equal to i satisfying $\hat{A}_i^{(2)} = \hat{A}_{i'}^{(2)}$. This instrument is relevant because the condition $\hat{A}_i^{(2)} = \hat{A}_{i'}^{(2)}$ guarantees that $A_{i'} = A_i$ for all i' used to construct $\zeta_i^{(1)}$. The exogeneity condition is satisfied thanks to the leave-one-out construction.[24]

[24] A technical point concerning the construction of the instruments is the selection of the other test-takers $\{i'\}$ such that $\hat{A}_i^{(2)} = \hat{A}_{i'}^{(2)}$. With about 25 items in each group, it will frequently be the case that no one else will have the exact same $\hat{A}^{(2)}$ as individual i. Therefore, I divide the sorted $\hat{A}^{(2)}$ into 100 equally sized bins and estimate $\zeta_i^{(1)}$ using the observations in the same bin as i. The resulting instruments still satisfy relevance because the $\{i'\}$ used to construct ζ_i will have achievement close to A_i on average. Using 20 or 200 bins produces very similar results. While I do not report the first stage results for brevity, the instruments are uniformly very strong, with F-stats above 400 in all cases. Additionally, it does not appear to matter

Although there are many different ways to partition the items into two disjoint groups, I restrict the analysis here to two groups with equal numbers of items where I assign odd-numbered items to group (1) and even numbered items to group (2). To the extent that the math and reading tests organize items by content, this procedure helps ensure that items from each content area are included in both groups.

Putting everything together, my method consists of the following steps:

1. For each achievement test (math and reading), divide the items into groups (1) and (2) such that each group has half of the total items.
2. Estimate $\hat{A}^{(1)}$ and $\hat{A}^{(2)}$ separately using groups (1) and (2) via equation (2). Let $\Delta\hat{A}^{(1)}_{h,l}$ be the raw (unadjusted) achievement gap estimated using $\hat{A}^{(1)}$.
3. Construct the instruments $\{\zeta^{(1)}_i\}$ as in equation (7). Estimate $\hat{\gamma}^{(1)}$ from equation (6) using these instruments.
4. Estimate the achievement gap using $\Delta\hat{A}^{(1)}_{h,l}/\hat{\gamma}^{(1)}$.[25]

One econometric question is how to construct the standard errors for the anchored achievement gaps. The issue is whether to treat the inflation factors $1/\hat{\gamma}^{(1)}$ as estimated or known. The main results treat these factors as known, consistent with work that uses psychometrically-derived reliability estimates. I also generate bootstrapped standard errors which take the sampling distribution of $1/\hat{\gamma}^{(1)}$ into account. Table 4 demonstrates that the bootstrapped standard errors are 25-50% larger than the baseline standard errors which treat the reliabilities as known. Nonetheless, using bootstrapped standard errors would not change any of the important empirical conclusions of the paper.

6.2 School Completion Item-Anchored Gaps

Table 5 shows school completion item-anchored gaps for three demographic comparisons: white/black, male/female, and high-/low-income. I present the item-anchored gaps in standard-deviation units for comparability to the NLSY79-given gaps and in outcome (school completion probability) units for economic interpretability.

very much in my setting which of the two groups is used to construct the item-anchored test scale and which is used to construct the instrument.

[25]Note that the reliability adjustments estimated using the group-2 items are applied to scales anchoring using the group 1 items only, not the full set of items.

The item-anchored white/black gaps are quite different than the given math and reading gaps, which are both around 1 sd. The high school item-anchored gaps and the college item-anchored reading gap are all about 0.2 sd larger than their given-scale counterparts. By contrast, the college item-anchored math gap is about 0.2 sd smaller than the given gap. Black youth do comparatively well on math items that are particularly predictive of college completion, although the anchored gap is still quite large in absolute terms. At the same time, black youth perform relatively poorly on reading items that are predictive of high school or college completion and on math items that are predictive of high school completion.

Turning to gender differences, the given scores imply that males have a 0.18 sd advantage in math and a 0.11 sd deficit in reading, with both estimates roughly consistent with prior literature.[26] Anchoring to school completion at the item level lowers the male advantage in math – the college gap is 0.13 sd, and the high school gap is only 0.05 sd. Similarly, item-anchoring to high school completion shrinks the female advantage in reading to 0.08 sd, while anchoring to college completion removes the female advantage entirely. Achievement scales which emphasize items associated with school completion reduce, or in some cases remove, apparent male-female achievement differences.

Item-anchoring to school completion also changes the estimated achievement gaps between youth from high- and low-income households. Specifically, Table 5 shows the given and item-anchored gaps for youth in the top and bottom quintiles of the base-year household income distribution. The given math and reading gaps are both about 1 sd, while the high school item-anchored gaps are roughly 0.1 sd larger. The college item-anchored gaps again show a divergence between math and reading, with the math gap a touch (0.05 sd) smaller and the reading gap substantially (0.3 sd) larger than the corresponding given gaps.

Table 5 also presents the estimated reliabilities for the item-anchored scales. The math reliabilities are 0.81 for high school and 0.87 for college; these are both quite close to the 0.85 reliability reported in the NLSY79. The high school reading reliability, at 0.86, is larger

[26]See Fryer and Levitt (2010); Dee (2007); Le and Nguyen (2018); Downey and Yuan (2005).

than the NLSY79 value of 0.81, while the college reading reliability is smaller, at 0.74. A consistent finding is that the item-anchored reliabilities are different from, and often smaller than, the reliabilities reported by the NLSY79 and also quite different for the same tests across different anchoring outcomes.[27]

The second column of Table 5 shows that anchoring at the item-level, as compared to anchoring using the given scores, is critical for generating my empirical results. The column shows gaps calculated by anchoring the given scores directly to outcomes, adjusted using the NLSY79-reported test reliabilities.[28] These given-anchored gaps are generally larger than the item-anchored gaps, sometimes substantially so, and they are never much smaller. The differences between the item- and given-anchored results are driven both by differences in the estimated test reliabilities and differences in the unadjusted gaps. One extreme example is the college male/female math gap, where the given-anchored estimate is the same as the raw gap, at 0.1 sd, while the item-anchored estimate is zero. This same general pattern repeats for the labor market outcomes – the given- and item-anchored gaps are often quite different from each other. These results highlight that simply anchoring the given scores to outcomes, thereby allowing for a non-linear relationship between the given scale and the economically relevant scale, is not enough – the item-anchored scales are fundamentally different.

Finally, I compare group differences in predicted versus actual school completion. The predicted white/black gaps are uniformly much larger than the actual gaps. Math and reading items predict college gaps of 0.20 and 0.25, respectively, both much larger than the still-sizeable 0.13 gap observed in the data. One possible explanation for this finding is that black youth may on average be attending lower quality schools where the probability

[27]Because the item-anchored scales use to construct the gaps use only half of the items for each test, with the other half used to construct the instruments, the reliability comparisons to the NLSY79 are not quite apples-to-apples. The item-anchored reliabilities are mostly larger than what Bond and Lang (2018) find anchoring to high school completion with prior-year lagged test scores used to adjust for measurement error. One possible explanation for this difference is that any skill that is predictive of outcomes within a given year that is not predictive across years will be viewed by their framework as measurement error but not by my within-year procedure. Additionally, the assessments studied by Bond and Lang (2018) are different and the respondents are much younger, either of which could explain their lower reliability estimates.

[28]The ad hoc use of the reported test reliabilities is not well-motivated, but does represent an effort to account for measurement error using typically-available psychometric data.

of graduating is higher for a given level of achievement. For gender, the predicted gaps are usually small and positive in favor of men, while the actual gaps slightly favor women. The exception is high school item-anchored reading, which correctly predicts an advantage for women. However, even this prediction is not very accurate – the predicted gap is only a quarter as large as the actual gap. Test items under-predict school completion gaps by household income by roughly 0.05-0.06. Youth from high-income households have an even larger advantage in school completion than what would be predicted on the basis of their item responses alone.

For robustness, Table 6 show alternative achievement gaps where the item-anchored scales are estimated on women or black respondents only. In both cases, the estimated gaps are quite close to the full-sample results in Table 5.[29] The similarity of these results supports my preferred interpretation of the anchored scales as largely measuring human capital rather than the spurious influences of confounders.

6.3 Wage and Lifetime Earnings Item-Anchored Gaps

I now repeat the analysis anchoring at the item level to ln(wage_30) and ln(pdv_labor). Table 7 presents my preferred estimates that use only white males to construct the anchored scales. White males have greater labor force attachment in the NLSY79, so selection plays less of a role in the estimates. Moreover, to the extent that discrimination and other barriers are operative in the labor market, item-anchored achievement scales estimated using only white males will be more interpretable. The resulting achievement gaps answer the question: "If items predicted outcomes for everyone as they do for white men, what would be the achievement gap between these two groups?"[30]

The white/black estimates in Table 7 are striking. Both the ln(wage_30) and ln(pdv_labor) item-anchored scales show greater achievement inequality by race than the NLSY79-given scales, which both show achievement gaps of around 1 sd. The item-anchored white/black

[29]Estimates using scales anchored on men and white respondents only, omitted for brevity, are naturally also quite close to the full-sample estimates.

[30]For completeness, Table 9 shows that using the full sample to estimate the anchored scales yields qualitatively similar results in almost all cases.

gaps are 0.15-0.23 sd larger for ln(wage_30) and 0.29-0.51 sd larger for ln(pdv_labor). Moreover, the predicted white/black gaps for both of these outcomes are almost exactly equal to the observed gaps – the large differences in mean wages and lifetime earnings by race can be almost perfectly predicted solely by differential item response patterns.

The white/black results are consistent with Neal and Johnson (1996), who find that adding AFQT scores to log wage regressions reduces the white/black gap for men by about two-thirds in the NLSY79. In order to compare my estimates to theirs, Table 8 estimates race and income gaps on the male-only subset of my data. The item-anchored scores explain between 77%-84% of the observed white/black wage and labor income gaps for men – less than what I found in the full sample but more than Neal and Johnson. It is worth emphasizing again that my results apply to white/black gaps in both wages in early adulthood and total lifetime earnings, whereas Neal and Johnson (1996) have results only for early-adult wages due to the young age of the NLSY79 respondents when that paper was written.

Turning to differences by household income, the item-anchored scales again show more achievement inequality than the given scales, with the item-anchored high-/low-income gaps about 0.16-0.20 sd larger than the given gaps, which are both around 1 sd. However, in contrast with the white/black results, the math and reading items predict log wage and lifetime earnings gaps that are roughly half as large as the gaps that are actually observed. This echos the findings for school completion – skill differences appear to be an important part of the story for understanding adult inequality between high- and low-income youth, but other factors must also be playing a very significant role.

Test items do a poor job of predicting the sizeable gender gaps in labor market outcomes. Using the ln(wage_30) and ln(pdv_labor) item-anchored scales instead of the given scales increases the male advantage in math by about 0.07 sd off a base of 0.18 sd. By contrast, item-anchoring to labor outcomes affects the female advantage in reading very little. These item-anchored gaps translate to modest predicted gender gaps in wages and earnings – the actual gap in ln(wage_30) is more than four times as large as what the male advantage

in item-anchored math would predict, while the ln(pdv_labor) gap is almost seven times as large. The differences are even more extreme for reading, as the item-anchored scales predict that women should slightly out-earn men.

6.4 Median Regression Wage Anchored Gaps

The baseline wage gaps use only white men with non-missing wages to estimate the item-anchored scales. Despite their relatively high labor force participation, wages are still missing for about 20% of the white men in my sample, raising concerns that selection may be driving the estimated achievement gaps. In this section, I present alternative estimates that use median regression to estimate the ln(wage_30) item-anchored scales using the full sample of white men, not just those with non-missing wages. The key assumption underlying the approach is that, conditional on the item responses, the latent unobserved wages are below the population median. With this assumption, I can identify the conditional medians, and, under further assumptions, the conditional means of the latent wage distributions.

Let $\tilde{S}$ be the latent (possibly unobserved) wage, and suppose that $\text{median}[\tilde{S}|D] = \mathbb{E}[\tilde{S}|D]$ for all vectors of item responses D.[31] Thus, if I can identify the median latent wage conditional on D, then I can also identify the mean latent wage conditional on D and proceed as before to construct the item-anchored achievement gaps. Two conditions are sufficient to identify the conditional medians. First, the unobserved latent wages must be below the median conditional on D – selection into observed wage income must always be positive. Second, the conditional medians should be linear in D.[32] Under these two conditions, the median can be identified by creating a new outcome $\breve{S}$ equal to S when the outcome is non-missing and $\min\{S_i\}$ otherwise and running a median regression of $\breve{S}$ on D.

Table 10 presents the ln(wage_30) item-anchored gaps estimated on the white male sample using median regression. As before, the item-anchored gaps are typically substantially larger

[31] As an example, the conditional medians will equal the conditional means if $\tilde{S} = DW + \varepsilon$ with ε drawn from any mean-0, symmetric distribution.

[32] As with the OLS-anchored scales, linearity is only needed for computational convenience. Given enough data, the conditional medians could be non-parametrically identified provided that fewer than half of the observations are missing wage income for each D.

than the given gaps. Moreover, the median-anchored math gaps tend to be much larger than the regression-anchored gaps, with the white/black math gap ballooning from 1.21 sd to 1.48 sd and the high-/low-income gap increasing from 1.19 sd to 1.44 sd. These larger math gaps are partially due to the lower estimated reliability of the median-anchored math scale (0.64 vs 0.75). The median-anchored reading estimates, despite a similarly lower reliability, are generally much closer to the regression-anchored estimates. Overall, these results do not suggest that selection is driving my main findings, which are, if anything, conservative.

Turning to the actual versus predicted ln(wage_30) gaps using median regression, the reading item responses exactly predict the white/black gap and a bit less than half of the high-/low-income gap. These results are very similar to the baseline OLS-based estimates. By contrast, the median-regression white/black predicted wage gap is about 0.1 log points larger than the OLS-anchored gap, while the high/low-income gap is about 0.06 log points larger. Once again, selection does not appear to be a significant driver of my main results.

6.5 Are White/Black Gaps in Employment Predictable?

The finding that test items can explain a significant share of the observed white/black gap in lifetime earnings, an outcome that depends both on wages and on cumulative employment, suggests that white/black differences in employment might be predictable from item responses. This result would be notable given prior literature finding that these employment differences are substantially less well-explained by observables (test scores, completed education, local labor market conditions, etc.) than differences in wages.[33] Most directly relevant for my application, Ritter and Taylor (2011) argue that the headline result in Neal and Johnson (1996) does not carry over from wages to employment in the NLSY79. These authors find that while controlling for AFQT scores and other pre-market factors does reduce white/black differences in both unemployment and total time not working, the remaining (unexplained) gaps are still quite large.

To investigate further, I apply my item-anchoring and gap estimation methods directly

[33]See Ritter and Taylor (2011); Fairlie and Sundstrom (1999); Stratton (1993); Abowd and Killingsworth (1984).

to the employment outcomes studied in Ritter and Taylor (2011): cumulative weeks of unemployment and cumulative weeks not working through 2004.[34] This allows me to assess whether the predicted white/black gaps in employment outcomes according to the item-anchored scales match the observed gaps in these outcomes.

Table 11 presents the resulting item-anchored male white/black employment gaps, anchored using white men.[35] In sd units, the same pattern is evident as with the other item-anchored gaps: white/black achievement inequality for men is about 0.5-0.6 sd larger using the employment item-anchored scales than using the given AFQT scale.[36] Turning to the actual versus predicted gaps, the predicted white/black unemployment gap is about 70% as large as the actual gap (40 weeks cumulatively through 2004 versus 57 weeks), while the predicted not working gap is about 78% as large (114 weeks versus 145 weeks).[37] Notably, these percentages are quite similar to the corresponding percentages for the earnings outcomes studied in Section 6.3. Overall, I do not find evidence that white/black employment differences for men are substantially less predictable using pre-market observables than wage differences. However, this similarity is only evident using test scales specifically constructed at the item level to explain employment and wage outcomes.

7 The Reading Puzzle and Item-Anchoring

Prior research (Sanders, 2016; Kinsler and Pavan, 2015) has noted that in multivariate regressions of labor market outcomes on math and reading scores, the coefficients on reading are often much smaller than the coefficients on math and in some cases are even significantly negative. This is puzzling – reading and math are distinct, and reading should be valuable

[34] These variables are available through the replication package associated with Ritter and Taylor (2011). Although I could extend these employment variables to NLSY79 survey rounds past 2004, for ease of comparison I simply use the exact outcome variables used in their paper.

[35] The results using scales anchored using the full sample, which I omit for brevity, are extremely similar.

[36] I use the AFQT, which combines the math and reading scales, in this analysis for direct comparability to Ritter and Taylor (2011). The item-anchored scales are constructed using equation (2) in which D_i consists of the full vector of math and reading items for each test-taker i.

[37] The predicted gaps in Table 11 are negative because these outcomes, unlike the others studied to this point, are "bads." A negative gap in unemployment means that white men are predicted to suffer fewer weeks of unemployment than black men. The item-anchored scales are defined so that a higher score corresponds to lower predicted unemployment (or weeks not working).

economically. I demonstrate here that using item-anchored scores can resolve this puzzle. Item-anchored math and reading scores both have large, statistically significant coefficient estimates in the types of joint regressions that give rise to the reading puzzle using standard, psychometrically derived scores. This is true both when the items are anchored directly to labor market outcomes and when they are anchored to school completion.

I assess the effect of item anchoring on the reading puzzle by estimating for different definitions of math and reading (given, item-anchored) models of the form

$$ln(\text{wage\_30})_i = \alpha + \beta_1 \text{math}_i + \beta_2 \text{reading}_i + \text{controls}_i + \varepsilon_i. \quad (8)$$

Table 12 presents the math and reading regression coefficients for various specifications of equation (8) that include or exclude household income and highest grade completed dummies as additional controls. I estimate equation (8) only on white males for the same reasons that I anchor only on this group. Echoing prior literature, columns (1) and (4) show that the estimated coefficients using the NLSY79-given scores (in sd units) are large and statistically significant for math and small and insignificant for reading. Anchoring the given scales to ln(wage_30) does not resolve the puzzle – columns (2) and (5) show that given-anchored math is significantly associated with ln(wage_30), while given-anchored reading is not. Item-anchored scores, however, do resolve the reading puzzle. Columns (3) and (6) in Table 12 show that the item-anchored math coefficients are very similar to the given-scale coefficients, while the item-anchored reading coefficients are much larger, at 0.05-0.09 log points. Although these reading estimates are about half the size of the math estimates, they are still economically large and are statistically distinguishable from 0 at the 1% level.

These results are not tautological – the item-anchored math and reading scales are constructed independently of each other. It could have been the case that the component of reading orthogonal to math was uncorrelated with ln(wage_30). Nonetheless, the scales used in Table 12 are constructed to be maximally predictive of ln(wage_30). This may make the significance of item-anchored reading less surprising. However, Table 13 shows that the results are similar using school completion outcomes as the anchors. In particular, columns (2)

and (4) show that high school and college item-anchored scales result in significant, though smaller, math and reading estimates. Finally, columns (3) and (5) show that the reading puzzle persists if I instead anchor the given scores using probit models; simply allowing for non-linearity with given scores does not yield the same results as item anchoring.

The results in this section highlight the more general point that the item-anchored scales relate to each other, and to various outcomes, differently than the given scales. Any calculation using test scores as an outcome or as a control might yield strikingly different results with item-anchored scores used in place of the given scores. Findings such as the reading puzzle, which have been interpreted as saying something interesting about the relationship between math, reading, and economic outcomes, may in fact just reflect arbitrary, poorly-motivated choices about how to aggregate test items.

8 Discussion and Conclusion

In this paper, I argued that test scales anchored at the item level to economic outcomes are more credible measures of human capital than the psychometric scales widely used in social science. I showed that the choice of scale matters: item-anchored scales rank individuals differently than psychometric scales and yield gaps by race, gender, and household income that are often significantly larger. Moreover, item-anchored scales almost fully predict white/black gaps in wages and lifetime earnings and predict half of the gaps in these outcomes by household income. I found also that test items predict well white/black differences in employment, in contrast to prior literature. Finally, I showed that item-anchoring resolves the "reading puzzle" – item-anchored reading scores are positively associated with wages, conditional on math.

The results in this paper suggest that social scientists and policy makers would do well to consider more closely the alignment between the achievement scales they are using and the economic outcomes in which they are ultimately interested. Psychometric scales are not designed with economic outcomes in mind, and their use may bias our understanding of group

and individual differences in economically relevant skills. Many common analyses, including tracking student progress, measuring school and teacher effectiveness through value-added models, and estimating causal impacts of interventions/shocks on student achievement, may depend critically on economically arbitrary choices about how to aggregate test items.

References

Abowd, J. M. and Killingsworth, M. R. (1984). Do Minority/White Unemployment Differences Really Exist? *Journal of Business & Economic Statistics*, 2(1):64–72.

Arcidiacono, P. (2004). Ability Sorting and the Returns to College Major. *Journal of Econometrics*, 121(1):343 – 375. Higher Education (Annals Issue).

Bettinger, E. P., Evans, B. J., and Pope, D. G. (2013). Improving College Performance and Retention the Easy Way: Unpacking the ACT Exam. *American Economic Journal: Economic Policy*, 5(2):26–52.

Bock, R. and Mislevy, R. (1981). Data Quality Analysis of the Armed Services Vocational Aptitude Battery. Technical report, National Opinion Research Center.

Bond, T. and Lang, K. (2013). The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results. *Review of Economics and Statistics*, 95:1468–1479.

Bond, T. and Lang, K. (2018). The Black-White Education-Scaled Test-Score Gap in Grades K-7. *Journal of Human Resources (forthcoming)*.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9):2633–79.

Cunha, F. and Heckman, J. J. (2008). Formulating, Identifying, and Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Journal of Human Resources*, 43:738–782.

Dee, T. S. (2007). Teachers and the Gender Gaps in Student Achievement. *The Journal of Human Resources*, 42(3):528–554.

Downey, D. B. and Yuan, A. S. V. (2005). Sex Differences in School Performance During High School: Puzzling Patterns and Possible Explanations. *The Sociological Quarterly*, 46, 2.

Fairlie, R. W. and Sundstrom, W. A. (1999). The Emergence, Persistence, and Recent Widening of the Racial Unemployment Gap. *ILR Review*, 52(2):252–270.

Fryer, R. G. and Levitt, S. D. (2010). An Empirical Analysis of the Gender Gap in Mathematics. *American Economic Journal: Applied Economics*, 2, 2:201–240.

Jackson, K. C. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes. *Journal of Political Economy*, 126(5):2072–2107.

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

Kinsler, J. and Pavan, R. (2015). The Specificity of General Human Capital: Evidence from College Major Choice. *Journal of Labor Economics*, 33(4):933–972.

Lang, K. and Manove, M. (2011). Education and Labor Market Discrimination. *The American Economic Review*, 101(4):1467–1496.

Le, H. T. and Nguyen, H. T. (2018). The Evolution of the Gender Test Score Gap Through Seventh Grade: New Insights from Australia Using Unconditional Quantile Regression and Decomposition. *IZA Journal of Labor Economics*, 7(1):2.

Neal, D. A. and Johnson, W. R. (1996). The Role of Premarket Factors in Black-White Wage Differences. *The Journal of Political Economy*, 104:869–895.

Nielsen, E. (2015a). Achievement Gap Estimates and Deviations from Cardinal Comparability. *Finance and Economics Discussion Series, Federal Reserve Board*.

Nielsen, E. (2015b). The Income-Achievement Gap and Adult Outcome Inequality. *Finance and Economics Discussion Series, Federal Reserve Board*.

Polachek, S. W., Das, T., and Thamma-Apiroam, R. (2015). Micro- and Macroeconomic Implications of Heterogeneity in the Production of Human Capital. *Journal of Political Economy*, 123(6):1410–1455.

Reardon, S. (2011). *The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations*, chapter 5, pages 91–116. Russell Sage Foundation, New York.

Ritter, J. A. and Taylor, L. J. (2011). Racial Disparity in Unemployment. *The Review of Economics and Statistics*, 93(1):30–42.

Sanders, C. (2016). Reading Skills and Earnings: Why Does Doing Words Good Hurt Your Wages? *Working Paper*.

Schofield, L. S. (2014). Measurement Error in the AFQT in the NLSY79. *Economics Letters*, 123, 3:262–265.

Schroeder, C. and Yitzhaki, S. (2017). Revisiting the Evidence for Cardinal Treatment of Ordinal Variables. *European Economic Review*, 92:337 – 358.

Stratton, L. S. (1993). Racial Differences in Men's Unemployment. *ILR Review*, 46(3):451–463.

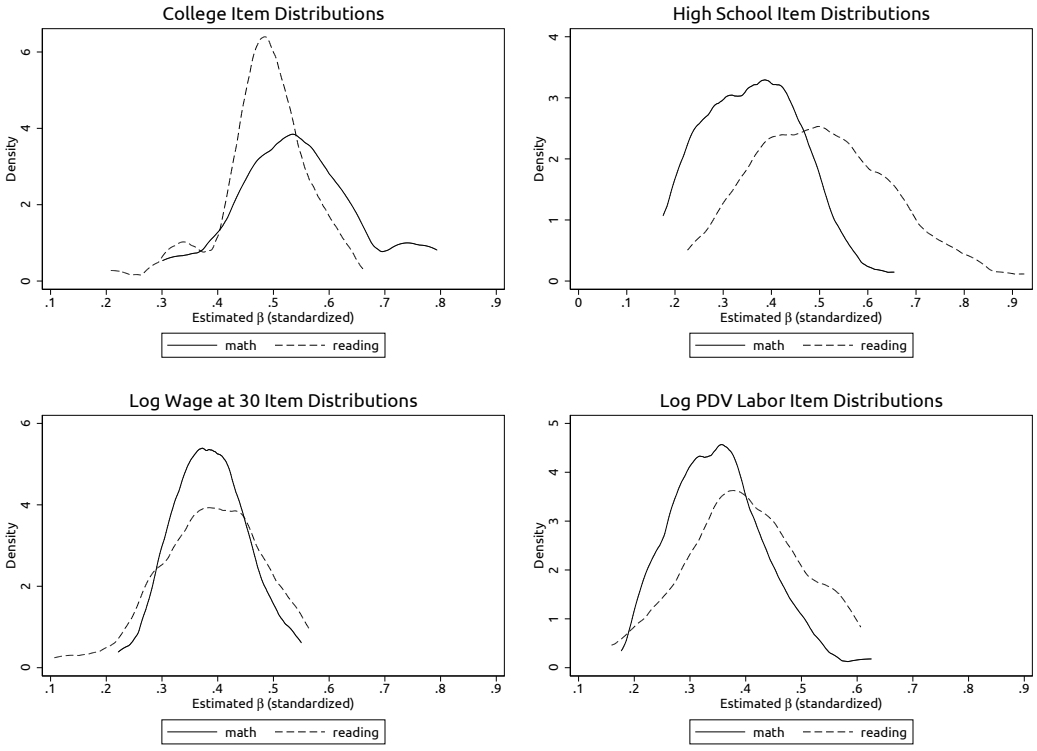
9 Tables and Figures

Table 1: Summary Statistics

Variable	Mean	Std. Dev.	N
age (base year)	17.73	2.32	11406
male	0.50	0.50	11406
black	0.14	0.34	11406
hh income (base year, $1,000)	70.84	46.4	10985
high school	0.89	0.32	11406
college	0.24	0.43	11406
highest grade completed	13.28	2.5	11406
pdv labor ($1,000)	440.88	332.2	11406
wage at 30	19.43	11.27	8521
math	99.05	18.99	10721
reading	98.23	19.38	10721
afqt	147.75	27.25	10721
ar missing	0.08	0.28	11406
wk missing	0.08	0.28	11406
pc missing	0.09	0.28	11406
mk missing	0.09	0.28	11406

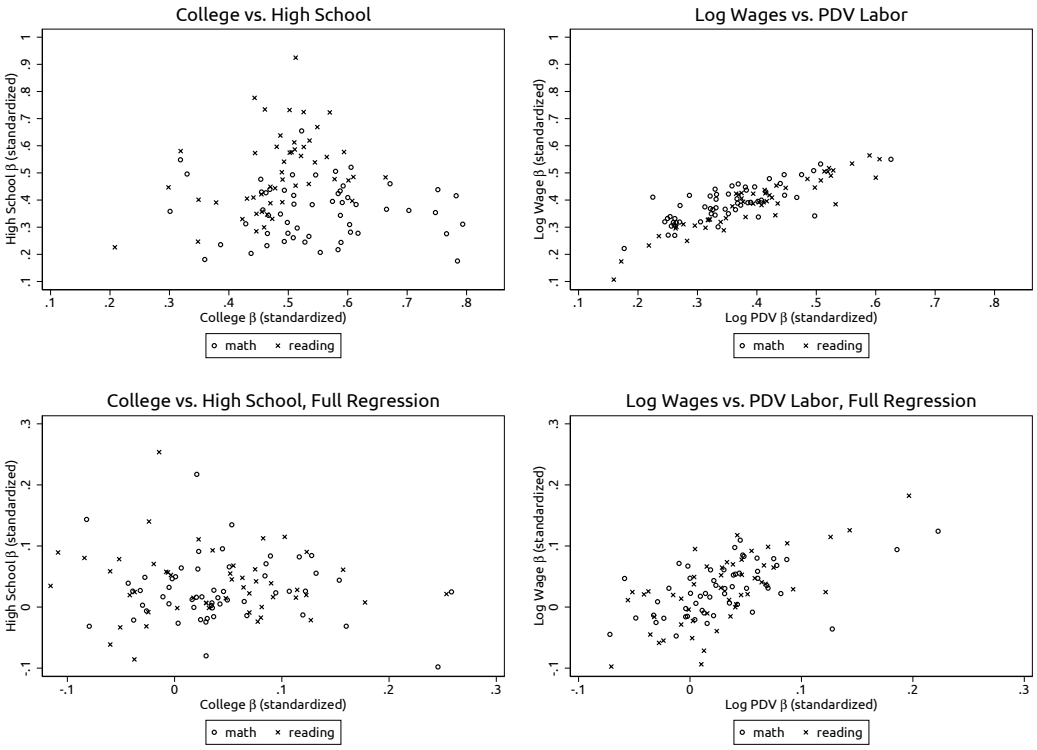
Notes: Dollar values in 2017-constant dollars deflated using the CPI-U.

Figure 1: Item-by-Item Regression Coefficient Distributions



Notes: Panels plot kernel densities across test items (j) of estimated regression coefficients ($\hat{W}_j$'s) from regressions of the form $y_i = \alpha_j + W_j d_{i,j} + \varepsilon_{i,j}$, where y_i is a standardized school completion indicator (high school or college) or labor market outcome (ln(wage_30) or ln(pdv_labor)).

Figure 2: Item-by-Item Regression Coefficient Comparisons



Notes: The top panels plot for each test item j pairs of estimated regression coefficients ($\hat{W}_j$'s) from regressions of the form $y_i = \alpha_j + W_j d_{i,j} + \varepsilon_{i,j}$, where y_i is a standardized (mean zero, standard deviation one) school completion indicator (high school or college) or labor market outcome (ln(wage_30) or ln(pdv_labor)). The bottom "Full Regression" panels repeat the analysis for regressions of the form $y_i = \alpha_j + \sum_j W_j d_{i,j} + \varepsilon_{i,j}$.

Table 2: Item Predictiveness and IRT Parameters

	(1) wage item	(2) wage full	(3) pdv item	(4) pdv full	(5) hs item	(6) hs full	(7) col item	(8) col full
discrimination	0.06***	0.00	0.05***	-0.00	0.05***	0.00	0.09***	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
difficulty	-0.05***	0.00	-0.11***	-0.02***	-0.17***	-0.04***	0.02*	0.05***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
guessing	-0.14*	-0.04	-0.06	0.00	-0.00	0.02	-0.24**	-0.03
	(0.08)	(0.06)	(0.08)	(0.06)	(0.08)	(0.05)	(0.10)	(0.07)
Obs.	105	105	105	105	105	105	105	105

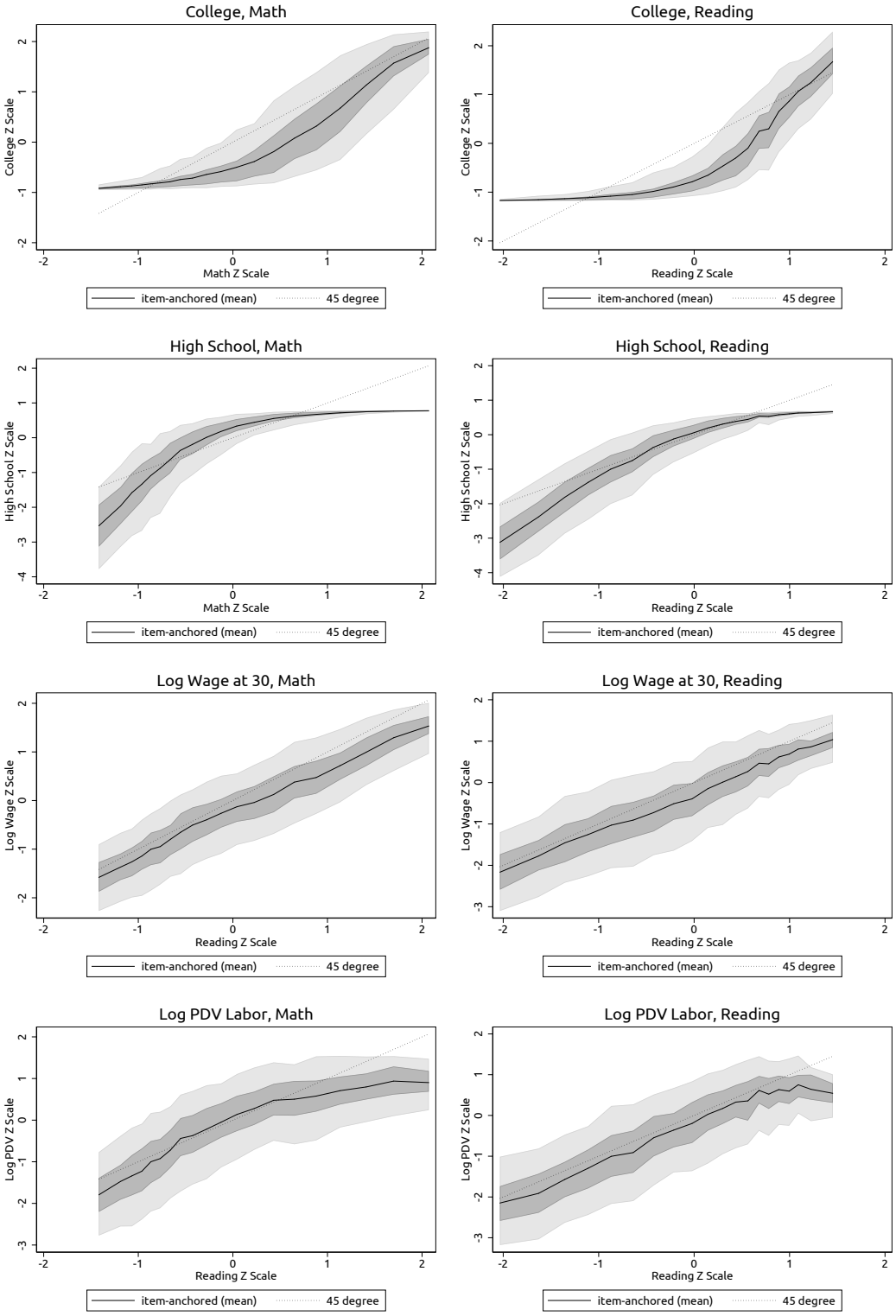
Notes: Each column represents a regression of the form $\hat{W}_j = \delta_0 + \delta_1 \text{discrimination}_j + \delta_2 \text{difficulty}_j + \delta_3 \text{guessing}_j + \varepsilon_j$, pooling math and reading together. $\hat{W}_j$ is the estimated regression coefficient for item j in a regression of an economic outcome in sd units on item(s) and discrimination_j, difficulty_j, and guessing_j are the irt-estimated discrimination, difficulty, and guessing probability for item j. The odd-numbered columns ("item") use $\hat{W}_j$'s estimated separately item-by-item, while the even-numbered columns ("full") use $\hat{W}_j$'s estimated jointly across all math or reading items. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 3: Item-by-Item Hypothesis Tests Share Rejected

Null tested:	Across-Group $\hat{W}(S)_{j,g} = \hat{W}(S)_{j,g'}$		Across-Anchor $\hat{W}(\text{given})_{j,g} = \hat{W}(S)_{j,g}$	
white/black	$\alpha = 0.05$	$\alpha = 0.50$	$\alpha = 0.05$	$\alpha = 0.50$
math, hs	0.00	0.02	0.36	0.45
reading, hs	0.00	0.00	0.28	0.38
math, college	0.00	0.02	0.15	0.27
reading, college	0.02	0.02	0.14	0.32
math, ln(wage_30)	0.00	0.00	0.13	0.27
reading, ln(wage_30)	0.00	0.00	0.26	0.32
math, ln(pdv_labor)	0.02	0.02	0.25	0.35
math, ln(pdv_labor)	0.00	0.04	0.26	0.36
male/female				
math, hs	0.00	0.00	0.22	0.35
reading, hs	0.02	0.08	0.26	0.38
math college	0.02	0.02	0.13	0.36
reading college	0.00	0.04	0.12	0.36
math, ln(wage_30)	0.00	0.02	0.09	0.20
reading, ln(wage_30)	0.02	0.12	0.18	0.22
math, ln(pdv_labor)	0.02	0.05	0.18	0.29
reading, ln(pdv_labor)	0.02	0.12	0.20	0.30
high-/low-inc.				
math, hs	0.02	0.04	0.18	0.29
reading, hs	0.04	0.12	0.32	0.38
math, college	0.00	0.09	0.22	0.40
reading, college	0.04	0.06	0.28	0.48
math, ln(wage_30)	0.00	0.04	0.15	0.25
reading, ln(wage_30)	0.00	0.02	0.26	0.36
math, ln(pdv_labor)	0.00	0.00	0.24	0.31
reading, ln(pdv_labor)	0.00	0.02	0.18	0.40

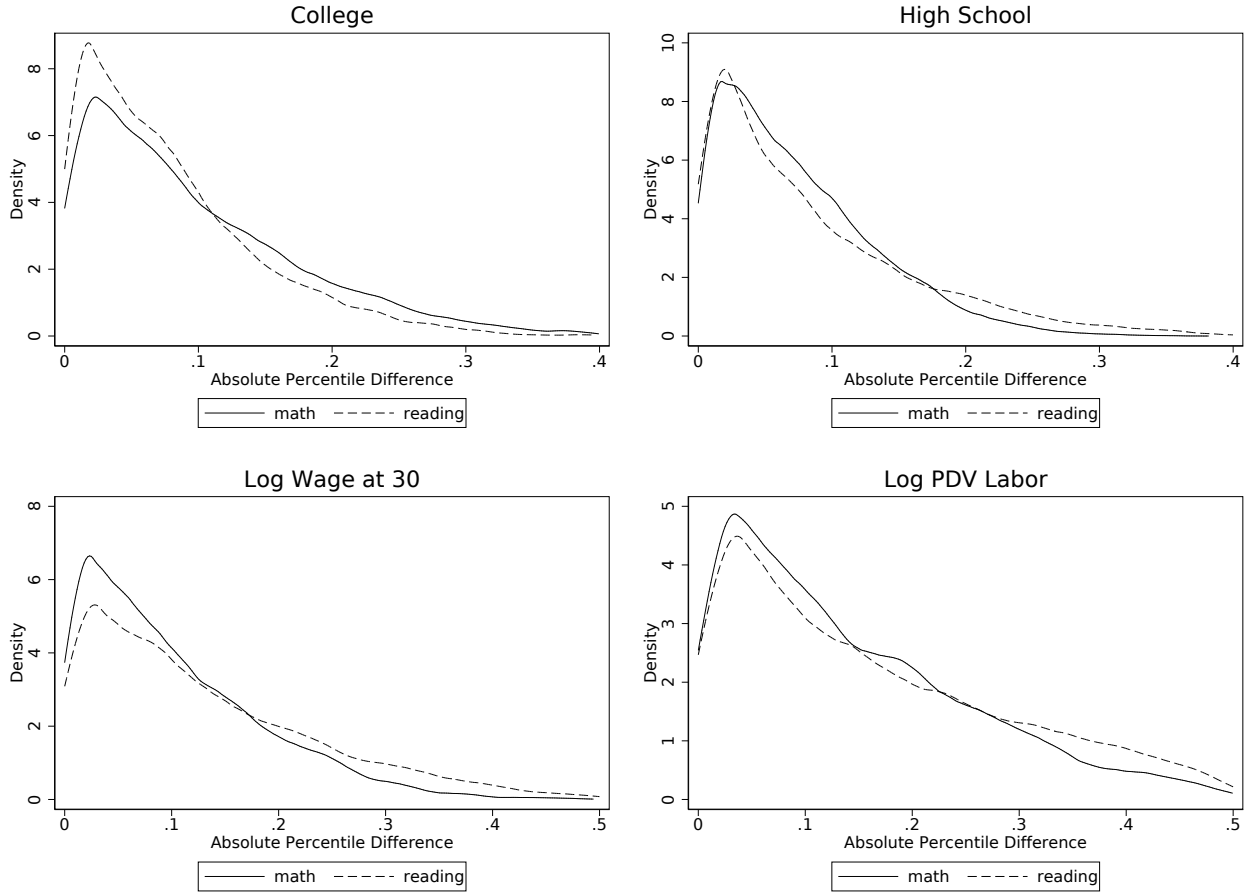
Notes: $\hat{W}(S)_{j,g}$ denotes the estimated coefficient (weight) for item j estimated on group g using outcome S. The columns show the share (across j) of various null hypotheses rejected at $\alpha = 0.05$ and $\alpha = 0.50$. The rejections use the Bonferroni correction to account for multiple testing, as each table entry corresponds to either 50 (reading) or 55 (math) separate hypothesis tests. In the Across-Anchor columns, group g corresponds to male, white, or high-income respondents, respectively.

Figure 3: The Relationship Between the Item-Anchored and NLSY79 Scales



Notes: Each panel plots the mean, along with the middle 50% and 90% range, of the item-anchored scores in each ventile of the NLSY79-given score distribution. The light dashed lines correspond to the 45-degree line of equality.

Figure 4: Percentile Differences – Item-Anchored and NLSY79 Scales



Notes: Each panel plots for math and reading a kernel density estimate of the distribution of the absolute value of the percentile differences between the item-anchored and NLSY79-given scores. If p is the percentile of a test-taker in the NLSY79-given score distribution and q is her percentile in the item-anchored distribution, the panels plot the density of $|p - q|$.

Table 4: Bootstrapped Standard Errors for Item-Anchored Achievement Gaps

White/Black		non-bootstrapped	bootstrapped
math	high school	0.03	0.05
reading	high school	0.03	0.05
math	college	0.03	0.03
reading	college	0.04	0.05
math	ln(pdv_labor)	0.03	0.04
reading	ln(pdv_labor)	0.04	0.05
math	ln(wage_30)	0.03	0.05
reading	ln(wage_30)	0.03	0.07
math	ln(wage_30) (median anchor)	0.03	0.04
reading	ln(wage_30) (median anchor)	0.03	0.05
Male/Female			
math	high school	0.02	0.03
reading	high school	0.02	0.03
math	college	0.02	0.03
reading	college	0.03	0.04
math	ln(pdv_labor)	0.02	0.03
reading	ln(pdv_labor)	0.02	0.04
math	ln(wage_30)	0.02	0.04
reading	ln(wage_30)	0.02	0.04
math	ln(wage_30) (median anchor)	0.02	0.03
reading	ln(wage_30) (median anchor)	0.02	0.03
High/Low			
math	high school	0.03	0.05
reading	high school	0.03	0.05
math	college	0.03	0.05
reading	college	0.04	0.06
math	ln(pdv_labor)	0.03	0.05
reading	ln(pdv_labor)	0.03	0.05
math	ln(wage_30)	0.03	0.05
reading	ln(wage_30)	0.03	0.06
math	ln(wage_30) (median anchor)	0.03	0.05
reading	ln(wage_30) (median anchor)	0.03	0.05

Notes: The non-bootstrapped standard errors are calculated without accounting for the sampling variation in the shrinkage adjustment factor $1/\hat{\gamma}^{(1)}$. The bootstrap standard errors are based on a normal approximation using 250 bootstrapped estimates, where the instruments use $\hat{A}^{(2)}$ sorted into 20 equinumerous bins.

Table 5: Item-Anchored School Completion Gaps

White/Black	given (z)	given-anchored (z)	item-anchored (z)	predicted	actual	item R
math, college	0.98	0.98	0.81	0.20	0.13	0.87
	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)	.
reading, college	1.05	1.22	1.28	0.25	0.13	0.74
	(0.02)	(0.04)	(0.04)	(0.01)	(0.01)	.
math, hs	0.98	1.21	1.21	0.15	0.06	0.81
	(0.03)	(0.03)	(0.03)	(0.00)	(0.01)	.
reading, hs	1.05	1.40	1.22	0.17	0.06	0.86
	(0.02)	(0.03)	(0.03)	(0.00)	(0.01)	.
Male/Female						
math, college	0.18	0.23	0.13	0.03	-0.00	0.87
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	.
reading, college	-0.11	-0.10	0.01	0.00	-0.00	0.74
	(0.02)	(0.02)	(0.03)	(0.01)	(0.01)	.
math, hs	0.18	0.12	0.05	0.01	-0.04	0.81
	(0.02)	(0.02)	(0.02)	(0.00)	(0.01)	.
reading, hs	-0.11	-0.16	-0.09	-0.01	-0.04	0.86
	(0.02)	(0.02)	(0.02)	(0.00)	(0.01)	.
High/Low Income						
math, college	0.99	1.07	0.94	0.23	0.29	0.87
	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)	.
reading, college	0.90	1.16	1.20	0.23	0.29	0.74
	(0.03)	(0.03)	(0.04)	(0.01)	(0.01)	.
math, hs	0.99	1.09	1.10	0.14	0.19	0.81
	(0.03)	(0.03)	(0.04)	(0.00)	(0.01)	.
reading, hs	0.90	1.13	1.03	0.14	0.19	0.86
	(0.03)	(0.04)	(0.03)	(0.00)	(0.01)	.

Notes: The first column shows gaps calculated using the NLSY79-given scales in sd units. The second column shows gaps calculated using given-anchored scales in sd units, and the third uses the item-anchored scales in sd units. The fourth column shows the predicted school completion gaps using the item-anchored scales, while the fifth column shows the actual gaps. The sixth column shows the estimated reliabilities of the item-anchored scales following the method outlined in Section 6. Estimates based on probit anchoring models that include age indicators in addition to item indicators. Instruments use the outcome scale divided into 100 equinumerous bins. Standard errors treat the estimated reliabilities as known with certainty. Please see Table 4 for bootstrapped alternatives that account for this uncertainty.

Table 6: Item-Anchored School Completion Gaps, Female and Black Samples

	Female Only		Black Only		
White/Black	item-anchored (z)	predicted	item-anchored (z)	predicted	actual
math, college	0.80	0.21	0.86	0.22	0.13
	(0.03)	0.01)	(0.03)	(0.01)	(0.01)
reading, college	1.35	0.26	1.22	0.24	0.13
	(0.04)	(0.01)	(0.04)	(0.01)	(0.01)
math, hs	1.32	0.15	1.56	0.15	0.06
	(0.04)	(0.00)	(0.04)	(0.00)	(0.01)
reading, hs	1.15	0.16	1.59	0.16	0.06
	(0.03)	(0.00)	(0.04)	0.00)	(0.01)
Male/Female					
math, college	0.14	0.04	0.07	0.02	-0.00
	(0.02)	(0.01)	(0.02)	0.01)	(0.01)
reading, college	0.05	0.01	-0.05	-0.01	-0.00
	(0.03)	(0.01)	(0.02)	(0.00)	(0.01)
math, hs	0.09	0.01	0.04	0.00	-0.04
	(0.03)	(0.00)	(0.03)	(0.00)	(0.01)
reading, hs	-0.06	-0.01	-0.22	-0.02	-0.04
	(0.02)	(0.00)	(0.03)	(0.00)	(0.01)
High/Low Income					
math, college	0.91	0.24	0.89	0.23	0.29
	(0.03)	(0.01)	(0.03)	(0.01)	(0.01)
reading, college	1.28	0.25	1.04	0.21	0.29
	(0.04)	(0.01)	(0.03)	(0.01)	0.01)
math, hs	1.15	0.13	1.36	0.13	0.19
	(0.04)	(0.00)	(0.05)	(0.00)	0.01)
reading, hs	0.97	0.13	1.32	0.13	0.19
	(0.03)	(0.00)	(0.05)	(0.01)	(0.01)

Notes: All anchored scales constructed using either the female-only or black-only subsamples of the data. The first and second columns show the item-anchored gaps estimated on the female-only subsample, while the third and fourth show the gaps estimated using the black-only subsample. The fifth column shows the actual gaps. Estimates based on probit anchoring models that include age indicators in addition to item indicators. Instruments use the outcome scale divided into 100 equinumerous bins. Standard errors treat the estimated reliabilities as known with certainty.

Table 7: Item-Anchored Log Labor Earnings Gaps

White/Black	given (z)	given-anchored (z)	item-anchored (z)	predicted	actual	item R
math, ln(wage_30)	0.98	1.13	1.21	0.25	0.24	0.75
	(0.03)	(0.03)	(0.04)	(0.01)	(0.02)	.
reading, ln(wage_30)	1.05	1.42	1.20	0.23	0.24	0.87
	(0.02)	(0.03)	(0.03)	(0.01)	(0.02)	.
math, ln(pdv_labor)	0.98	1.13	1.49	0.46	0.45	0.69
	(0.03)	(0.03)	(0.04)	(0.01)	(0.03)	.
reading, ln(pdv_labor)	1.05	1.42	1.34	0.41	0.45	0.75
	(0.02)	(0.03)	(0.04)	(0.01)	(0.03)	.
Male/Female						
math, ln(wage_30)	0.18	0.20	0.24	0.05	0.22	0.75
	(0.02)	(0.02)	(0.03)	(0.01)	(0.01)	.
reading, ln(wage_30)	-0.11	-0.14	-0.10	-0.01	0.22	0.87
	(0.02)	(0.02)	(0.02)	(0.00)	(0.01)	.
math, ln(pdv_labor)	0.18	0.20	0.25	0.07	0.47	0.69
	(0.02)	(0.02)	(0.03)	(0.01)	(0.02)	.
reading, ln(pdv_labor)	-0.11	-0.14	-0.08	-0.03	0.47	0.75
	(0.02)	(0.02)	(0.03)	(0.01)	(0.02)	.
High/Low Income						
math, ln(wage_30)	0.99	1.14	1.19	0.24	0.46	0.75
	(0.03)	(0.03)	(0.04)	(0.01)	(0.02)	.
reading, ln(wage_30)	0.90	1.23	1.09	0.20	0.46	0.87
	(0.03)	(0.04)	(0.03)	(0.01)	(0.02)	.
math, ln(pdv_labor)	0.99	1.14	1.18	0.36	0.82	0.69
	(0.03)	(0.03)	(0.04)	(0.01)	(0.03)	.
reading, ln(pdv_labor)	0.90	1.23	1.06	0.32	0.82	0.75
	(0.03)	(0.04)	(0.04)	(0.01)	(0.03)	.

Notes: All anchored scales estimated on the white male subsample of the data. The first column shows gaps calculated using the NLSY79-given scales in sd units. The second column shows gaps calculated using given-anchored scales in sd units, and the third uses the item-anchored scales in sd units. The fourth column shows the predicted ln(wage_30) and ln(pdv_labor) gaps using the item-anchored scales, while the fifth column shows the actual gaps. The sixth column shows the estimated reliabilities of the item-anchored scales following the method outlined in Section 6. Estimates based on regression anchoring models that include age indicators in addition to item indicators. Item-anchored scales constructed using white men only. Instruments use the outcome scale divided into 100 equinumerous bins. Standard errors treat the estimated reliabilities as known with certainty. Please see Table 4 for bootstrapped alternatives that account for this uncertainty.

Table 8: Item-Anchored Log Labor Earnings Gaps, Male Sample

White/Black	given (z)	given-anchored (z)	item-anchored (z)	predicted	actual	item R
math, ln(wage_30)	1.04	1.20	1.27	0.25	0.30	0.75
	(0.04)	(0.05)	(0.06)	(0.01)	(0.03)	.
reading, ln(wage_30)	1.06	1.45	1.21	0.23	0.30	0.87
	(0.04)	(0.05)	(0.05)	(0.01)	(0.03)	.
math, ln(pdv_labor)	1.04	1.20	1.54	0.47	0.56	0.69
	(0.04)	(0.05)	(0.06)	(0.02)	(0.04)	.
reading, ln(pdv_labor)	1.06	1.45	1.39	0.42	0.56	0.75
	(0.04)	(0.05)	(0.06)	(0.02)	(0.04)	.
High/Low Income						
math, ln(wage_30)	0.96	1.12	1.15	0.22	0.47	0.75
	(0.04)	(0.05)	(0.06)	(0.01)	(0.03)	.
reading, ln(wage_30)	0.93	1.26	1.15	0.21	0.47	0.87
	(0.04)	(0.05)	(0.05)	(0.01)	(0.03)	.
math, ln(pdv_labor)	0.96	1.12	1.20	0.36	0.85	0.69
	(0.04)	(0.05)	(0.06)	(0.02)	(0.04)	.
reading, ln(pdv_labor)	0.93	1.26	1.21	0.37	0.85	0.75
	(0.04)	(0.05)	(0.06)	(0.02)	(0.04)	.

Notes: The first column shows gaps calculated using the NLSY79-given scales in sd units. The second column shows gaps calculated using given-anchored scales in sd units, and the third uses the item-anchored scales in sd units. The fourth column shows the predicted ln(wage_30) gaps using the item-anchored scales, while the fifth column shows the actual gaps. The sixth column shows the estimated reliabilities of the item-anchored scales following the method outlined in Section 6. Estimates based on regression anchoring models that include age indicators in addition to item indicators. Item-anchored scales constructed using white men only. Instruments use the outcome scale divided into 100 equinumerous bins. Standard errors treat the estimated reliabilities as known with certainty.

Table 9: Item-Anchored Log Labor Earnings Gaps, Full Sample

White/Black	given (z)	given-anchored (z)	item-anchored (z)	predicted	actual	item R
math, ln(wage_30)	0.98	1.13	1.18	0.28	0.24	0.84
	(0.03)	(0.03)	(0.03)	(0.01)	(0.02)	.
reading, ln(wage_30)	1.05	1.42	1.33	0.28	0.24	0.82
	(0.02)	(0.03)	(0.03)	(0.01)	(0.02)	.
math, ln(pdv_labor)	0.98	1.13	1.22	0.46	0.45	0.86
	(0.03)	(0.03)	(0.03)	(0.01)	(0.03)	.
reading, ln(pdv_labor)	1.05	1.42	1.26	0.45	0.45	0.89
	(0.02)	(0.03)	(0.03)	(0.01)	(0.03)	.
Male/Female						
math, ln(wage_30)	0.18	0.20	0.25	0.06	0.22	0.84
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	.
reading, ln(wage_30)	-0.11	-0.14	0.17	0.04	0.22	0.82
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	.
math, ln(pdv_labor)	0.18	0.20	0.24	0.09	0.47	0.86
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	.
reading, ln(pdv_labor)	-0.11	-0.14	0.16	0.05	0.47	0.89
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	.
High/Low Income						
math, ln(wage_30)	0.99	1.14	1.13	0.26	0.46	0.84
	(0.03)	(0.03)	(0.03)	(0.01)	(0.02)	.
reading, ln(wage_30)	0.90	1.23	1.22	0.25	0.46	0.82
	(0.03)	(0.04)	(0.03)	(0.01)	(0.02)	.
math, ln(pdv_labor)	0.99	1.14	1.02	0.38	0.82	0.86
	(0.03)	(0.03)	(0.03)	(0.01)	(0.03)	.
reading, ln(pdv_labor)	0.90	1.23	1.03	0.36	0.82	0.89
	(0.03)	(0.04)	(0.03)	(0.01)	(0.03)	.

Notes: The first column shows gaps calculated using the NLSY79-given scales in sd units. The second column shows gaps calculated using given-anchored scales in sd units, and the third uses the item-anchored scales in sd units. The fourth column shows the predicted ln(wage_30) and ln(pdv_labor) gaps using the item-anchored scales, while the fifth column shows the actual gaps. The sixth column shows the estimated reliabilities of the item-anchored scales following the method outlined in Section 6. Estimates based on regression anchoring models that include age indicators in addition to item indicators. Item-anchored scales constructed using the full sample. Instruments use the outcome scale divided into 100 equinumerous bins. Standard errors treat the estimated reliabilities as known with certainty. Please see Table 4 for bootstrapped alternatives that account for this uncertainty.

Table 10: Item-Anchored Log Wage Gaps Using Median Regression

White/Black	given (z)	given-anchored (z)	item-anchored (z)	predicted	actual	item R
math	0.98	1.13	1.48	0.34	0.24	0.64
	(0.03)	(0.03)	(0.04)	(0.01)	(0.02)	.
reading	1.05	1.42	1.27	0.25	0.24	0.78
	(0.02)	(0.03)	(0.04)	(0.01)	(0.02)	.
Male/Female						
math	0.18	0.20	0.31	0.05	0.22	0.64
	(0.02)	(0.02)	(0.03)	(0.01)	(0.01)	.
reading	-0.11	-0.14	-0.18	-0.05	0.22	0.78
	(0.02)	(0.02)	(0.03)	(0.01)	(0.01)	.
High/Low Income						
math	0.99	1.14	1.44	0.30	0.46	0.64
	(0.03)	(0.03)	(0.04)	(0.01)	(0.02)	.
reading	0.90	1.23	1.12	0.21	0.46	0.78
	(0.03)	(0.04)	(0.04)	(0.01)	(0.02)	.

Notes: All anchored scales estimated on the white male subsample of the data. The first column shows gaps calculated using the NLSY79-given scales in sd units. The second column shows gaps calculated using given-anchored scales in sd units, and the third uses the item-anchored scales in sd units. The fourth column shows the predicted ln(wage_30) gaps using the item-anchored scales, while the fifth column shows the actual gaps. The sixth column shows the estimated reliabilities of the item-anchored scales following the method outlined in Section 6. Estimates based on median regression anchoring models as outlined in Section 6.4 that include age indicators in addition to item indicators. Instruments use the outcome scale divided into 100 equinumerous bins. Standard errors treat the estimated reliabilities as known with certainty. Please see Table 4 for bootstrapped alternatives that account for this uncertainty.

Table 11: Item-Anchored Employment Gaps – White vs. Black Males

	AFQT (z)	AFQT-anchored (z)	item-anchored (z)	predicted	actual	item R
unemployed	1.13	1.33	1.59	-40	-57	0.65
	(0.04)	(0.05)	(0.07)	(2)	(4)	.
not working	1.13	1.33	1.76	-114	-145	0.66
	(0.04)	(0.05)	(0.07)	(5)	(9)	.

Notes: All anchored scales estimated on the white male subsample of the data. The first column shows white/black achievement gaps for men calculated using the AFQT scale in sd units. The second column shows gaps calculated using AFQT-anchored scales in sd units, and the third uses the item-anchored scales in sd units. The fourth column shows the predicted gaps in cumulative weeks unemployed through 2004 and cumulative weeks not working through 2004 using the corresponding item-anchored scales, while the fifth column shows the actual gaps in these outcomes. The sixth column shows the estimated reliabilities of the item-anchored scales following the method outlined in Section 6. Estimates based on regression anchoring models that include age indicators in addition to item indicators. Instruments use the outcome scale divided into 100 equinumerous bins. Standard errors treat the estimated reliabilities as known with certainty.

Table 12: Reading Puzzle Regressions – Wages at Age 30

	(1) given	(2) given-anchored	(3) item-anchored	(4) given	(5) given-anchored	(6) item-anchored
math	0.17***	0.17***	0.16***	0.10***	0.11***	0.11***
	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)
reading	0.03	0.02	0.09***	-0.01	-0.01	0.06***
	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)
education	no	no	no	yes	yes	yes
parental income	no	no	no	yes	yes	yes
white male only	yes	yes	yes	yes	yes	yes
Observations	2,306	2,306	2,217	2,232	2,232	2,142

Notes: Table shows the estimated coefficients on math and reading for regression of the form $\ln(\text{wage\_30}) = \alpha + \beta_1 \text{math} + \beta_2 \text{reading} + \gamma X + \varepsilon$, where X denotes education and household income controls (or not, as indicated). All columns show estimates only for white men. Column labels correspond to the math and reading test scores used (given, given-anchored, or item-anchored). The given-anchored columns use cubics in the given scores to construct the anchoring relationships. All regressions use test scores in sd units. Standard errors based on 1,000 bootstrap iterations. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 13: Reading Puzzle Regressions – Wages at Age 30, Alternative Scales

	(1) ln(wage_30) item	(2) college item	(3) college given	(4) high school item	(5) high school given
math	0.11***	0.07***	0.10***	0.05**	0.08***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
reading	0.06***	0.03*	-0.02	0.04**	0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
education	yes	yes	yes	yes	yes
parental income	yes	yes	yes	yes	yes
white male only	yes	yes	yes	yes	yes
Observations	2,142	2,142	2,232	2,142	2,232

Notes: Table shows the estimated coefficients on math and reading for regression of the form $\ln(\text{wage\_30}) = \alpha + \beta_1 \text{math} + \beta_2 \text{reading} + \gamma X + \varepsilon$, where X denotes education and household income controls (or not, as indicated). All columns show estimates only for white men. Column labels correspond to the different anchoring outcomes for the math and reading scales. The given-anchored columns use cubics in the given test scores to construct the anchoring relationships. All regressions use test scores in sd units. Standard errors based on 1,000 bootstrap iterations. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.