Test Items, Outcomes, and Achievement Gaps

Eric Nielsen*

October 15, 2018

Abstract

Standard psychometric methods aggregate test items into achievement scores without considering how items relate to economic outcomes. This paper constructs alternative achievement measures using the estimated relationship between individual test items and economic outcomes such as school completion and labor market earnings. Item-anchored scales rank students differently than standard psychometric scales, and these ranking differences have important implications for estimated achievement gaps by race, gender, and parental income. Typically, though not always, item-anchored achievement gaps are substantially larger than gaps calculated using standard psychometric scores. Black/white and high-/low-income achievement gaps are generally about 0.2-0.6 standard deviations larger (about 20-60%) when test items are related to labor market outcomes and 0.06-0.2 standard deviations larger (6-20%) when high school and college completion are used. Test items can fully explain black-white earnings differences, but can explain only half of the earnings difference between youth from high- versus low-income households. Finally, conditional on item-anchored math scores, item-anchored reading scores have significantly positive relationships with various economic outcomes in contrast to psychometric scales where the reading relationships are zero or even negative conditional on math. Keywords: human capital, educational outcomes, achievement gaps, measurement error. JEL Codes: JEL Codes: I.24, I.26, C.2.

1 Introduction

Education and human capital are fundamental to understanding economic outcomes. Group differences in human capital play a key role in economists' explanations for group differences in labor market success, health, and many other outcomes. A key question,

*Bryce Turner and Danielle Nemschoff provided excellent research assistance. This paper benifited from discussions with Ralf Meisenzahl and seminar participants at the Federal Reserve Board. The views and opinions expressed in this paper are solely those of the author and do not reflect those of the Board of Governors or the Federal Reserve System. Contact: eric.r.nielsen@frb.gov

then, is to what extent skill differences drive observed inequalities. Because human capital is not directly observable, economists often turn to achievement test scores as proxies. Test scores correlate with economic outcomes across a variety of contexts, justifying their use as measures of human capital and lending support for policies focused on their improvement. Indeed, the equation of test scores with human capital is so widespread within economics that researchers often treat such scores as they would any other economically interpretable, desirable outcome.[1]

However, achievement tests measure human capital imperfectly because the skills emphasized by the test may not be the skills which correspond most directly to human capital. Ultimately, every test scale may be thought of as a method for combining individual test items into a single index. The problem is that psychometric methods combine items without reference to the *economic* importance of the skills the items measure. This is not a failing of the test itself; the problem lies in applying the test to a purpose – measuring human capital – for which it was not designed. Standard achievement scales may emphasize items that are not predictive of economic success while de-emphasizing items that are predictive. In turn, this could bias achievement comparisons if different groups perform differentially better or worse on the items that are related to later success.

Consider the following two question test as an example. Question 1 correlates strongly with labor market outcomes, while question 2 is uninformative. If the psychometric procedure weights these two questions equally, the test will have three possible scores: none correct, one correct, or both correct. This example illustrates two important implications of using test scores to assess labor market outcomes. First, a test can have too many scores – only question 1 is relevant for understanding labor market outcomes so only 2 scores (question 1 right or wrong) are needed. Second, irrelevant questions can obfuscate useful information. Question 2 is superfluous and its inclusion in the test hides the relationship

[1]There are many, many examples of this approach in economics. For instance, Hanushek and Rivkin (2012, 2009), Hoxby (2000), Reardon (2011), Clotfelter et al. (2009) all use test scores in this way. This list is not meant to be exhaustive this use of test scores is the rule rather than the exception in economics.

between question 1 and labor market success. Some individuals with the middle score (one right, one wrong) will do well because question 1 is predictive, and some will do less well because question 2 is not. This will look like unexplained variation in outcomes conditional on test scores, but this variation would be explained by a better targeted scale that used only the predictive item.[2]

This paper studies the alignment between achievement scales and human capital. It does this using recently available data on individual Armed Forces Qualifying Test (AFQT) items from the National Longitudinal Survey of Youth 1979 (NLSY79) to assess which items are most relevant for predicting long-run economic outcomes such as school completion and labor income. In particular, the paper constructs new, "item-anchored" scores by weighting test items in proportion to how strongly they correlate with these outcomes. Based on these item-anchored scores, the paper documents that many achievement gaps by race, gender, and socioeconomic status are considerably larger than those so far reported in the literature.

This exercise is only possible because of the item-level data and the long-run outcomes available in the NLSY79. Most research on achievement gaps would not be able to employ a similar approach, as both features of the data are unusual. Even where long-run outcomes have been used to anchor scores (Cunha and Heckman, 2008; Bond and Lang, 2018), item level data have not been available or have not been used.[3] The NLSY79 item-level data is used by Schofield (2014) to argue that the measurement error in the IRT-derived AFQT scores is non-normal.

Achievement has no natural units, so in some sense the most fundamental way to compare different test scales is through their induced rank ordering of students. Even if the cardinal properties of two scales are different, they might nonetheless rank students in the same way. However, this is not the case in the NLSY79; the item-anchored scales

[2]Please see Appendix B for a formal treatment of this example.

[3]Polachek et al. (2015) take a different approach by estimating human capital production function parameters for each survey respondent in the NLSY79. They then relate these "human capital" ability measures to standard measures of cognitive and noncognitive skills.

often rank students very differently than the given (psychometrically-derived) scales. It is not uncommon for an individual student's ranking to differ by 10-20 percentile points between these alternative scales. Some students do well on items that are predictive of later success but poorly on uninformative items, resulting in a low given score, while for other students the situation is reversed.

The item-anchored scales also yield very different achievement gaps once measurement error is properly handled. Measurement error is a subtle issue in this setting – simply taking the mean difference in the item-anchored scales will result in an achievement gap which is biased towards zero. I adapt the empirical approach in Bond and Lang (2018) to adjust for this bias. My innovation is to leverage the item-level data to estimate multiple independent item-anchored scores, which can then be used to undo the effect of shrinkage. It is this method which allows me to estimate the relevant reliabilities of the item-anchored scales.

The given test-score gaps in the NLSY79 are in line with what has been reported in the literature (Neal and Johnson (1996); Reardon (2011); Downey and Yuan (2005), and many others), while the item-anchored gaps are typically much larger. For instance, I estimate that black-white and high-/low-income math and reading gaps are all around 1 standard deviation (sd) using given scores. By contrast, the high-/low-income gaps are 0.2-0.4 sd larger and the black/white gaps 0.2-0.6 sd larger when I anchor on later-life earnings. Anchoring to high school completion leads to more modest, though still sizable, increases on the order of 0.06-0.20 sd. Item-anchoring does not increase all of the gaps, however – the college anchored black/white math gap is a full 0.20 sd smaller than the gap calculated using given math scores.

There are two reasons why the item-anchored achievement gap estimates differ from those calculated in the standard way using given test scores. First, the item-anchored test reliabilities (the signal-to-noise ratios for the tests) are often much lower than what is typically assumed in the literature – the anchoring-relevant reliability need bear little

relation to the psychometrically-derived reliability typically reported. Second, some groups do particularly well answering items that are particularly predictive of economic success. Some of these items are not emphasized by the psychometric scoring system, so that the given test scores effectively "hide" the group's success.

The item-anchored gaps can be directly compared to the actual outcome gaps. Consistent with prior literature using given test scores (Lang and Manove, 2011), test items predict larger black/white gaps in school completion than are actually observed – black youth complete more schooling than can be predicted by their achievement alone. However, the item-anchored and actual black white earnings and wage gaps are almost identical, a result reminiscent of Neal and Johnson (1996). Differences in school quality are one explanation for these results, although Lang and Manove (2011) argue against this interpretation. Importantly however, my estimates rule out labor market discrimination as an alternative mechanism because the earnings and wage gaps are estimated only on white men. By contrast, the item-anchored scales dramatically under-predict both school completion and adult earnings gaps by parental income.

Finally, the item-anchored scores help resolve the "reading puzzle" – the documented phenomenon in which reading scores, though positively correlated with income and other outcomes, are weakly or even negatively correlated with outcomes conditional on math scores (Sanders, 2016; Kinsler and Pavan, 2015; Arcidiacono, 2004). Joint regressions of item-anchored reading and math scores on outcomes (including outcomes not used in the anchoring regressions) suggest a sizable role for reading conditional on math, in contrast to what regressions using given scores find in the NLSY79. 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 combined as they are in standard psychometric models. Nonetheless, the item-anchored reading coefficients are still only about half as large as the math coefficients for most outcomes and most regression specifications.

By leveraging new item-level data and the long-term outcomes available in the NLSY79,

I am able to advance the literature on achievement gaps by demographic groups. My results suggest that the standard approach to measuring achievement differences using psychometric scores may provide biased estimates of *human capital* differences. Policy interventions that seek to raise test scores may be focusing their energies in the wrong place – some improvements in observed test scores may not translate to improvements in actual economic outcomes, and causal effects estimated on test scores may mask heterogeneous impacts by race, gender, and socioeconomic status. In turn, this suggests that economists and policy makers would do well to focus on the construction of achievement tests that are more closely aligned with the economic outcomes of interest.

The rest of the paper is organized as follows: Section 2 discusses the test item and outcome data in the NLSY79, while Section 3 presents preliminary evidence that these items correlate quite differently with different outcomes. Section 4 discusses the general empirical and conceptual framework. Sections 5 - 6 present the main empirical estimates, while Section 7 address the relationship between item anchoring and the reading puzzle. Section 8 concludes. Appendix A presents all tables and figures and Appendix B contains supplementary discussion and analysis.

2 Data

The NLSY79 is a high quality, nationally representative survey that follows a sample of roughly 12,500 individuals aged 14-22 in 1979 through to the present. Each round of the survey collects extensive information on educational and labor market outcomes. These data allow me to construct school completion and lifetime earnings variables. Additionally, survey respondents took a battery of achievement tests, the Armed Services Vocational Aptitude Battery (ASVAB), in the base year of the survey. Importantly, item-level response data are available for these tests. I now describe each of these pieces, test items and later-life outcomes, in greater detail.

The ASVAB, which was administered at the start of the survey, consists of a series of subject and skill specific tests which collectively are used by the United States military in making enlistment and personnel decisions. The Armed Forces Qualifying Test (AFQT), a measure of math and reading achievement widely studied within economics, is based on a number of the components of the ASVAB. I make use of the individual item response data for the math and reading components that go into the AFQT.[4] Although each survey respondent in the base year of the NLSY79 should have taken the ASVAB, in practice about 1,500 individuals do not have and valid item-level data. Throughout, I use only the subsample of respondents for whom at least some item-level data is available.[5]

I use longitudinal data in the NLSY79 to construct school completion and labor earnings variables. Please refer to Nielsen (2015b) for more details on how I construct these measures. For school completion, I use the highest grade completed reported at any point in the first 15 years of the survey.[6] I define "high school" as an indicator of 12 or more grades of schooling completed and "college" as an indicator 16 or more grades completed.

The first earnings variable I study is a measure of wages at age 30 (wage_30). For each survey round, I compute wages by dividing total reported labor income by total reported hours worked. I then estimate wage_30 by taking the average annual wage for the three survey rounds closest to each individual's age-30 round. I average to smooth out transitory wage/earnings fluctuations. Although I can observe wages at younger ages, I restrict myself to age 30 because almost all schooling is completed by this age. Additionally, this

[4]The math items come from the arithmetic reasoning (30 items) and mathematics knowledge (25 items) ASVAB components, while the reading items come from the paragraph comprehension (15 items) and word knowledge (35 items) components. The constituent ASVAB components defining the AFQT changed after the administration of the test in the NLSY79. I use the current definition of math and reading, rather than the definition which held in 1980.

[5]I require only that the items used be non-missing. Since some respondents have item data for some ASVAB components and not for others, in practice this means that the samples used across different ASVAB components are slightly different. In cases where the items are not entirely missing, 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 NLSY. The assumption I make therefore is that leaving a question blank and getting the question incorrect amount to the same thing.

[6]In some cases, the highest grade completed reported by the respondent actually falls between successive rounds of the survey. My fill-in rule assumes that the higher value is correct.

age falls shortly after the typical "crossing point" past which more educated adults earn more on average than less educated adults.

The second earnings variable I study is the present discounted value of lifetime labor income (pdv_labor). The construction of pdv_labor is complicated because missing data, the choice of labor force participation, and various survey limitations all become first-order problems. As with the school completion variables, I follow Nielsen (2015b) in constructing my measure. First, to deal with missing labor income, I adopt an extreme, "pessimistic" imputation rule which assigns to each missing labor income the minimum labor income observed for the individual over the life of the survey. This pessimistic rule is not meant to be realistic; rather, it will tend to compress the earnings distribution.[7] Second, measuring the labor income budget set requires one to take a stand on selection into and out of employment. I assume that unemployment is involuntary, so that full income = observed income. I make this assumption because the alternative, that wage×full-time hours is the correct measure, is largely driven by estimated wages, which is the other outcome I study. Third, and finally, the NLSY79 survey respondents are only in their late 40's in the most recent survey round; I therefore need to make some assumptions on their earnings from the present until retirement. I assume that each individual's earnings growth after the last survey follows the education-specific growth rates from a pseudo-panel of male earnings constructed from the 2005 American Community Survey. Additionally, I assume each respondent retires 20 years after the most recent survey round, when the respondents are in their mid-late sixties.[8]

Table 1 in Appendix A presents the summary statistics of the main variables used in the analysis. The college completion rate in the NLSY79 is about 23%, while the high

[7] In Nielsen (2015b) I also study "optimistic" imputations which assign the maximum observed income, rather than the minimum. Though different in levels, these two imputation rules produce income measures which are quite correlated with each other. I therefore stick with pessimistic imputation to simplify the discussion.

[8] 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 outlined above. I use a 5% discount rate throughout.

school completion rate is 89%. The average wage earned at age 30 is about $19.50 with a standard deviation of $11.25. The pdv_labor variable has an average of $435,000; changing the imputation rule and the assumption on labor supply would increase this average substantially.[9] Roughly 9% of the sample is missing each of the ASVAB components.

3 Item-Outcome Correlations

Before delving into the anchoring analysis, I first present some simple evidence that test items differ widely in how strongly they predict school completion and labor market success. These large differences open the door for item-anchored scales to have non-trivial effects on achievement calculations.

Figure 1 shows the distributions of the estimated coefficients and R^2's for bivariate regressions of each test item on college and high school completion. The left panels show that each item, taken in isolation, is positively correlated with school completion. However, there is quite a range in the estimated coefficients – the point estimates range from 0.07 to over 0.30. Likewise, the right panels show that items differ widely in how much outcome variation they can explain. Interestingly, the distributions suggest that math items tend to be more predictive of college completion, while reading items are more predictive of high school completion. Finally, the bottom two panels plot the distributions of the differences of the coefficient and R^2 estimates for each item across the two school completion outcomes. The wide spread in these distributions implies that the items that are highly predictive of one type of school completion are not necessarily predictive of the other. Figure 2, which repeats the analysis using the labor income outcomes in logs, tells a similar story: items differ widely in how strongly they predict labor market success.

Similarly, Figure 3 presents the math and reading item coefficient distributions in multivariate regressions in which all items (both math and reading) are included in the

[9] For instance, optimistic imputation + fully chosen labor supply results in a pdv_labor measure with an average of $1.2 million and a standard deviation of over $2 million.

right hand side. Most of the item coefficients are again positive (although some are negative) and the distributions are again quite spread out. As before, the items that are most predictive for one outcome are not generally the items that are most predictive for other outcomes.

4 Conceptual Framework

This section presents the framework that I use to analyze test items and economic outcomes. For ease of exposition, I refrain here from discussing the techniques I employ to handle measurement error in the calculation of item-anchored achievement gaps. That analysis is presented in Section 6.

Let $j \in \{1,\ldots,M\}$ index a sample of test-taking students drawn independently from some population, and let S_j denote the economic outcome (school completion, wages, etc.) of interest for student j. All other observable characteristics of the student (race, gender, family background, etc.) are denoted by X_j.

Students take an achievement test with N dichotomous items. Let $\mathbf{D}_j$ denote the full vector of item responses from student j: $\mathbb{D}_j = [D_{1,j},\ldots,D_{N,j}]$ where $D_{i,j} = 1$ if j gets question i correct, and 0 otherwise. These items are combined using some standard psychometric framework to produce a standardized (mean 0, standard deviation 1) test score z_j. In the NLSY79 data used in this paper, the z_j will be based on a three parameter logistic IRT model. However, this detail is not important. What matters is that the scores are constructed from the $\mathbf{D}_j$ without reference to S_j. Often, I will refer to the z_j scale as the "given" scale or the "given z scores."

It is the given z scores, constructed by testing agencies and other data providers, that are almost always treated by economists as direct measures of human capital rather than as estimated proxies. As described in the introduction, this introduces two distinct problems, both of which will be remedied by anchoring at the item level.

First, the units of z_j are not meaningful. Achievement does not have natural units, so without reference to some external outcome it is not possible to determine whether a given score represents a lot of achievement or relatively little. This means that changes in the given scale will also not generally be informative and statistics calculated using the given scale will generally be biased. In symbols, for outcome S_j, ignoring measurement error, we might have for some increasing function ψ the relation $S_j = \psi(z_j)$. Instead of using the given z-scores directly, it would therefore be preferable to use the alternative scale $\psi(z)$ as it is already in interpretable units. Interpretability is not the only reason to prefer $\psi(z)$ to z, however. When ψ is nonlinear, as it often appears to be empirically, statistics calculated using the two scales may disagree dramatically – they may even differ in sign (Nielsen, 2015a; Schroeder and Yitzhaki, 2017; Bond and Lang, 2013).

Second, given z scores represent a particular choice about how to map each of the 2^N possible sequences of item responses to achievement. Since this map will generally be chosen without reference to economically interpretable outcomes, it is possible that scoring procedure will obscure useful information about the relationship between test items and outcomes.

I propose a framework which overcomes both of these conceptual problems. As in Bond and Lang (2018), I guarantee cardinal interpretability by defining achievement A_j as the expected value of S_j:

$$S_j \equiv A_j + \eta_j, \ \mathbb{E}[\eta_j] = \mathbb{E}[\eta_j A_j] = 0. \tag{1}$$

Note that the η_j term is orthogonal by construction. Because only S_j is observed for each student j, A_j must be estimated. Rather than estimating A_j directly from z_j, I instead propose to estimate it directly from the item-level responses: $\hat{A}_j = \hat{\mathbb{E}}[S_j|\mathbf{D}_j]$. In other words, I allow test items to enter directly into the anchoring relationship, rather than only through the given score. In particular, for some flexible function f, I suppose that

$$S_j = f(\mathbf{D}_j) + \varepsilon_j. \tag{2}$$

I then use estimates of f to construct outcome-denominated achievement scores. Provided that $\hat{f}(\mathbf{D}_j)$ approximates $\mathbb{E}[S_j|\mathbf{D}_j]$ sufficiently well, this approach will produce achievement measures which are responsive to the relationship between individual test items and outcomes.

I sometimes condition on X_j as well as $\mathbf{D}_j$ in defining the anchoring relationship $\hat{A}_j = \hat{\mathbb{E}}[S_j|\mathbf{D}_j, X_j]$. In particular, I use only white men to estimate the anchoring relationships for labor income outcomes. Doing this avoids selection and interpretation difficulties stemming from the higher labor force non-participation rates for female and black respondents.

4.1 Linear Regression and Probit Scales

Because $\mathbf{D}_j$ can take on many possible values for tests with even moderate numbers of items, it is necessary in empirical work to place restrictions on the class of functions considered for f.[10] This section takes the simplest case, linear regression (or probit, as appropriate) with no interaction terms across items. These simple models produce anchored scales which, at the mean, look similar to the scales one gets by anchoring the z-scores directly. Although there is no *a priori* reason to rule out the possibility that a given item is only valuable in concert with another item or combination of items, I find that allowing for such interactions produces qualitatively similar anchored scales once dimension reduction techniques are employed to limit the number of parameters estimated.

Therefore, I suppose that the relationship between S_j and $\mathbf{D}_j$ is linear. When S_j is dichotomous, I will sometimes suppose that S_j and $\mathbf{D}_j$ are related through the probit

[10] Of course, in general $\mathbb{E}[S_j|\mathbf{D}_j]$ is non-parametrically identified by the population averages of S in each "bucket" defined by $\mathbf{D}$.

function. That is, I suppose

$$S_j = \mathbf{D}_j'\mathbf{W}_r + \varepsilon_j, \text{ or } S_j = \Phi(\mathbf{D}_j'\mathbf{W}_p + \varepsilon_j). \tag{3}$$

Let $\hat{A}_j^{(r)}$ and $\hat{A}_j^{(p)}$ denote the regression and probit-anchored scores (predicted values) for student j.

It is important to emphasize that the item coefficient vectors $\mathbf{W}_r$ and $\mathbf{W}_p$ in Equation 3 should not be interpreted as structural parameters; the estimated coefficient on an individual item is not indicative of any causal relationship between that item and the outcome S. Rather, the goal is simply to estimate $\mathbb{E}[S_j|\mathbf{D}_j]$ flexibly. Individual elements of $\hat{\mathbf{W}}_r$ and $\hat{\mathbf{W}}_p$ are therefore allowed to be negative even though it may not be plausible that true causal effect of any of these items should be less than zero. The lack of a structural interpretation of the item coefficients also implies that they cannot be used to identify which particular items are valuable or not valuable. However, I care only about the predicted values of S in my application, properly adjusted for measurement and estimation error.

5 Empirical Results – Rank Stability

I first assess how the item-anchored test scales estimated using either linear or probit regressions compare to the given test scales. I next compare the item-anchored scales to "z-anchored scales" based on models which estimate models relating given z scores to outcomes.

This exercise yields two main insights. First, the item-anchored scales rank students very differently than the given scales. There is a wide range of item-anchored scores associated with each given score; there are typically individuals with very different predicted outcomes based on their item responses who nonetheless have the same given score. Second, economic outcomes are not simply linear transformations of the given scores. A

fixed change in the given score corresponds to a large change in outcomes in some regions and a small change in others.

Figure 4 plots school completion anchored scales estimated using either the individual test items or the given z scores against the given z scores. The dotted line in each panel is the 45-degree line – the given test scale plotted against itself. Because the item-anchored scores are not simple functions of the given z-scores, I plot the average item anchored score conditional on the given score.

The schooling-anchored scales using z scores and individual test items are quite similar to each other and quite different than the given z scale. The college anchored math and reading scales are convex; differences in achievement at the bottom of the given scale do not translate to differences in college completion, while differences at the top do. The situation is reversed for the high school anchored scales. Improvements at the bottom ends of the given achievement scales translate very strongly changes in high school completion, while improvements in the top ends are not very valuable. These results are intuitive. Low-achievement youth are not likely to be on the margin for attending/completing college, so the college anchored scales should be flat for such students. At the same time, these students are comparatively likely to be on the high school completion margin, explaining the steep anchored relationship for below-mean z-scores. Interestingly, the high school anchored scales are much more concave for math than reading.

The achievement scales depicted in Figure 4 are nonlinear functions of observed scores. This means that standard statistics, such as regression coefficients and mean differences, computed using observed scores will be biased if school completion is the outcome of interest. However, Figure 4 does not make clear why studying test items individually matters. To understand the difference between item-anchored scales and z-anchored scales, it is necessary to move beyond the conditional means of the item-anchored distributions.

Figure 5 plots the conditional means along with the interquartile and middle 90 percent

ranges for the item-anchored school completion scales.[11] This figure demonstrates that there is a whole distribution of item-anchored scores for each given score. Individuals whose item responses led to the same given score might have very different predicted school completion based on which particular items they got wrong and right. For college completion (left panels), this variation is greater at the top end of the observed score distribution. For instance, among students 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. 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.

The range of item-anchored scores depicted in Figure 5 implies that the item-anchored scores will rank students differently than the given scores. This is not true for the z-anchored scales, as these are just monotone transformations of the given scores. In other words, the item anchored scores do not just disagree with the given scores about how valuable achievement is (as the anchored z-scores do), they disagree fundamentally about which students are performing well and which students are not.

The relatively wide 90^{th} percentile ranges depicted in Figure 5 suggest that the ranking differences between the given scores and the item-anchored scores might be quite substantial. Indeed, Figure 6, which plots the absolute value of the difference in percentiles according to different scales, shows that it is not uncommon for the percentile ranking of a student to differ by 0.1 - 0.2 or more between the given z scale and the school completion anchored scale.[12] Interestingly, the ranking differences between the anchored scales and the given scales are similar in distribution to the differences between the anchored scales and the scales anchored on the other school completion variable. There are substantial fractions of students who rank quite differently in math or reading achievement depending

[11] In detail, I first divide the NLSY79 sample into ventiles based on the given test scores. I then plot the means and percentiles of the ventile-specific item anchored distributions on the y-axis.

[12] The figure plots only the distribution of the absolute percentile differences because the distribution of percentile differences is symmetric by construction.

on whether one is interested in college or high school completion.

Figures 7 - 9 repeat the analysis for the log wage and log pdv_labor scales. Compared to the school-anchored scales, Figure 7 shows that the item-anchored and z-anchored scales using log labor income are more similar to the given z scale but less similar to each other. Of course, this implies that the dollar outcome scales, rather than the log dollar outcome scales, would be nonlinear relative to the given scales. The biggest differences are seen for pdv_labor; we will see in the next section that this measure also typically yields the largest (in sd units) achievement gap estimates. As before, Figure 8 shows that there is substantial variation in the item-anchored scores associated with each observed (given) score. Unlike with school completion, the spread of these conditional distributions appears to be fairly constant across the range of observed scores, although the distributions are a bit more spread out in reading for lower given scores. Finally, Figure 9 shows that the log pdv_labor-anchored scales display slightly more rank shuffling (relative to the given scores) than the school completion-anchored scales.

6 Empirical Results – Achievement Gaps

I now turn to the measurement of achievement gaps using item-anchored test scales. Section 5 showed that relating economic outcomes to individual test items results in test scales that are quite different in terms of how they rank students and how much emphasis they place on different parts of the given test score distribution. Both of these differences raise the possibility that the item-anchored scales will yield different mean achievement gaps if some groups perform particularly well or poorly on items that are particularly predictive of outcomes. Empirically, I find that most item-anchored gaps are larger than the given gaps, but some are smaller.

6.1 Mean Achievement Gaps, Shrinkage, and Measurement Error

Estimating achievement gaps using the item-anchored scores introduces several issues related to shrinkage and measurement error. I adapt the framework from Bond and Lang (2018) to handle these issues. The strategy of using instruments to recover the relevant "shrinkage term" (to be explained below) is from their work. My innovation lies in leveraging the item-level data to construct the necessary instruments.

Let h and l denote two groups of students whose achievement we want to compare. Our goal is to estimate $\Delta A_{h,l} \equiv \bar{A}_h - \bar{A}_l$, where $\bar{A}_g$ is the average achievement of group g. Each individual's achievement is measured with error: $\hat{A}_j = A_j + \nu_j$ where ν_j is error that comes from estimation and specification error. If $A \sim N(\bar{A}, \sigma_A^2)$ and if $\nu_j \sim N(0, \sigma_\nu^2)$ iid in the population,

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

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

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

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

Given some way to consistently estimate $R_{A,\nu}$, one could use Equation 5 to recover a consistent estimate of $\Delta A_{h,l}$. A biased estimator of $R_{A,\nu}$ is the regression of $\hat{A}_j$ on S_j

$$\hat{A}_j = \kappa + \gamma S_j + \varepsilon_j. \tag{6}$$

The OLS estimate of γ is biased because S_j is a noisy measure of A_j: $\text{plim}\hat{\gamma}_{OLS} = R_{A,\nu}R_{A,\eta} < R_{A,\nu}$. This is a classic errors-in-variables problem, so with the appropriate instrument an unbiased estimate of γ can be recovered. In general, what is needed are at least two noisy measures of S_j, where the noise is uncorrelated across measures.

The item-level construction allows one to construct many such noisy measures 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 3 can be estimated separately on each group to produce anchored scores $\hat{A}_j^{(1)}$ and $\hat{A}_j^{(2)}$. Each of these scores is a noisy measure of A_j. Now consider estimating Equation 6 using $\hat{A}_j^{(1)}$. An instrument for S_j in this equation is the average S among students who are not j but who nevertheless have the same value of $\hat{A}_j^{(2)}$. That is, an instrument using item group (1) as the base is

$$\zeta_j^{(1)} = N_j^{-1} \sum_{j' \neq j:\ \hat{A}_j^{(2)} = \hat{A}_{j'}^{(2)}} S_{j'} \tag{7}$$

This instrument is relevant because achievement is persistent – test-takers should do similarly well on tests with similar items. The exogeneity condition is satisfied thanks to the leave-one-out construction.

There are several things to note about this method. First, groups (1) and (2) can be interchanged in the above construction. Using group (1) to construct either the scale ($\hat{A}_j$) or the the instruments (ζ_j) should yield a consistent estimate of $\Delta A_{h,l}$. Second, there are many different ways to partition the items into two groups, and each partition yields two valid estimators by interchanging the roles of the groups. Third, there is no reason to restrict the partition to two groups. With a partition of the items into K disjoint groups, $K-1$ instruments could be constructed as above and then combined using GMM. For

now, I restrict the analysis to only 2 groups with equal numbers of test items where I assign odd-numbered items to group (1) and even numbered items to group (2). To the extent that the ASVAB component tests organize items by content, this procedure ensures that items from each content area are included in both groups (1) and (2). Given this approach, it does not appear to matter very much which of the two groups is used to construct item-anchored test scale and which is used to construct the instrument. A more comprehensive treatment of the possibilities suggested by the above discussion is left for future work.

A technical point concerning the construction of the instrument $\zeta_j^{(1)}$ is the selection of other students $\{j'\text{s}\}$ such that $\hat{A}_j^{(2)} = \hat{A}_{j'}^{(2)}$. With 25 or more items in each group, it will frequently be the case that no other youth will have the same item-anchored score in group (2) as youth j. Therefore, I divide the sorted $\hat{A}_j^{(2)}$s into 200 equally sized bins and estimate $\zeta_j^{(1)}$ using the j' in the same bin as j. Putting this all together, my approach consists of the following steps:

1. Divide the items into groups (1) and (2) such that each group has roughly half of the total items.
2. Estimate $\hat{A}_j^{(1)}$ and $\hat{A}_j^{(2)}$ using Equation 3. Let $\Delta\hat{A}_{h,l}^{(1)}$ be the raw (unadjusted) mean achievement gap estimated using the $\hat{A}_j^{(1)}$.
3. Construct the $\zeta_j^{(1)}$ as in Equation 7. Estimate $\hat{\gamma}^{(1)}$ from Equation 6 using instrumental variables regression.
4. Estimate the *h*-*l* achievement gap using $\frac{\Delta\hat{A}_{h,l}^{(1)}}{\hat{\gamma}^{(1)}}$.

The final 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. My baseline tables treat these factors as known, consistent with work that uses psychometrically-derived reliability estimates and consistent with the anchored estimates I produce which use such reliabilities. In other words, I construct standard errors that

account only for the sampling variation in $\Delta\hat{A}^{(1)}_{h,l}$. Therefore, I also produce standard error estimates which take the sampling distribution of $1/\hat{\gamma}^{(1)}$ into account, either through the bootstrap or through an asymptotic approximation. These methods yield larger standard errors, but overall they do not change any of the important empirical conclusions of the paper. Table 2 demonstrates that the asymptotic and bootstrapped standard errors are generally about 25-50% larger than naive standard errors that do not adjust for the variation in $1/\hat{\gamma}^{(1)}$.

6.2 School Completion Gaps

I first present results comparing achievement gaps estimated using given z scores to those using scores anchored at the item level to various school completion measures. Table 3 presents three sets of gaps using probit models: white/black, male/female, and high-/low-income. I present the item-anchored gaps in standard deviation units for comparability to the z scores and in outcome (school completion probability) units for economic interpretability.

The item-anchored white/black achievement gaps are quite different than the given z-score gaps. The given gaps for both math and reading are around 1 sd. The high school anchored gaps are about 0.13-0.19 sd larger, while the difference is even larger for the college anchored reading gap, at 0.22 sd. By contrast, the college anchored math gap is 0.19 sd *smaller* than the given gap. In sum, black/white math achievement inequality is almost 20% smaller in standard deviation terms when test items are anchored to college completion, while the other anchored gaps are 15-20% larger. This suggests that 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 math items that are predictive of high school completion.

Turning to male/female differences, the given scores suggest that males have a modest

0.17 sd advantage in math and a smaller 0.10 sd deficit in reading. Anchoring to school completion dramatically lowers the male advantage in math; the college-anchored gap, at 0.13 sd, is 24% lower and the high school-anchored gap, at 0.05 sd, is 71% lower. Similarly, anchoring to high school completion shrinks the female advantage in reading by 20% to 0.08 sd, while anchoring to college completion removes the female advantage entirely – the gap falls to 0. Achievement scales which emphasize items associated with school completion reduce, or in some cases remove, apparent male-female achievement differences.

Anchoring to school completion has less dramatic effect on the income-achievement gap, defined here as the mean difference between the top and bottom household income quintiles. The high school gap is only 6% (0.06 sd) larger in math and about 9% (0.08 sd) larger in reading. The college results show the same divergence between math and reading achievement seen above; the college anchored math gap is about 5% (0.05 sd) smaller than the corresponding z gap, while the anchored reading gap is 31% (0.28 sd) larger.

Table 3 also presents the estimated item-anchored reliabilities. The math reliabilities are 0.83 (high school) and 0.88 (college); these are both quite close to the 0.85 reliability reported in the NLSY. By contrast, the high school reading reliability, at 0.87, is larger than the reported value of 0.81, while the college reading reliability is much smaller, at 0.74. A consistent finding is that the outcome-relevant reliabilities are often quite different from the reliabilities reported by the NLSY and also quite different for the same test items across different outcomes. These reliabilities are mostly larger than what Bond and Lang (2018) find using a similar procedure where the anchoring outcome is high school completion and prior-year lagged test scores are used to construct the instruments necessary to adjust for shrinkage. 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 Bond and Lang as measurement error but not by my "within year" IV procedure. Additionally, the assessments studied by Bond and Lang are different than the

AFQT and the students are much younger, both of which could independently explain their lower reliability estimates.

The second column of Table 3 shows gaps calculated by anchoring the given scores directly to outcomes. Each gap is adjusted by the NLS-reported test reliability. Generally, these anchored gaps are larger than the item-anchored gaps, sometimes substantially so, and they are never substantially smaller. These differences are driven both by differences in the estimated test reliabilities and differences in the "raw" (unadjusted) gaps. One extreme example is the college-anchored male/female math gap where the z-anchored gap is the same as the raw gap, at 0.1 sd, while the item-anchored gap is 0.

Because the anchored scores are in school-completion units, it is possible to directly compare group differences in predicted school completion given the item-response data with the actually observed group differences in school completion. The predicted black-/white school completion gaps are uniformly larger than the actual school completion gaps. For instance, math items predict a college completion gap of 0.2, while the actual gap is only 0.13. Black youth are more successful at completing high school and college than their math and reading test results would predict. One possible explanation for this finding is that black youth may on average be attending lower quality schools where the probability of graduating is higher for a given level of true achievement. Turning to male/female gaps, the predicted gaps are usually small and positive in favor of men, while the actual gaps uniformly favor women slightly. That is, purely on the basis of test performance, men would be expected to complete both high school and college at slightly greater rates than women, but in fact women have higher completion rates than men. Whereas test items over-predict black/white gaps, they under-predict achievement gaps by parental income. In each case, the predicted gap is about 0.05-0.06 less than the actual gap. Economically advantaged youth have an even larger school completion advantage than what would be predicted on the basis of their already high level of academic achievement.

6.3 Labor Income and Wage Anchored Gaps

I now repeat the analysis of the previous section using test scales anchored to log wages at age 30, ln(wage_30), and log lifetime labor wealth, ln(pdv_labor). Table 4 presents my preferred estimates which use only white males to construct the anchored scales. These are my preferred estimates because white men have the greatest labor force attachment in the NLSY79 data so selection plays less of a role in the estimates. Moreover, to the extent that discrimination and other barriers for women and racial minorities are operative in the labor market, the item-anchored achievement scales estimated using only white males will be more interpretable. Achievement gaps estimated using these white-male scales answer the question: "If test items correlated to outcomes for everyone as they do for white men, what would be the achievement gap between these two groups?" Anchored scales estimated on the full NLSY79 sample will blend the relationships between skills and outcomes operative for different demographic subgroups. Nonetheless, I report full-sample estimates in Table 5 for completeness.

The baseline results presented in Table 4 imply that given scores underestimate achievement gaps by race, gender, and parental income. The item-anchored white/black gaps are uniformly much larger than the given gaps. The ln(wage_30) gaps for math and reading are 20-25% (0.18-0.22 sd) larger, while the ln(pdv_labor) gaps are roughly 40-55% (0.42-0.63 sd) larger. The already sizable achievement gap between white and black youth is even larger when one considers the items that are predictive of subsequent labor market success. Turning to male/female gaps, the item-anchored advantages for men in math are much larger than the z gaps (+41% or 0.07 sd for ln(wage_30) and +65% or 0.12 sd for ln(pdv_labor)) while the female advantage in reading stays roughly flat. Finally, the item-anchored high-/low-income gaps are uniformly much larger (23-33% or 0.20-0.37 sd) than the given gaps.

The estimated reliabilities are again informative. The ln(wage_30) reliability for reading is 0.84, slightly higher than the value reported in the NLSY. However, the other

item-anchored scales have much lower reliabilities, ranging from 0.64 to 0.074. In particular, the ln(pdv_labor) reliabilities are both quite low relative to the psychometrically derived values reported in the NLSY. This again highlights the fact that the economically relevant amount of measurement error in test scores may be greater than what is captured by standard methods.

The comparison between the item-predicted and actual outcome gaps again reveals some interesting patterns. The differences between the predicted and actual black/white gaps are negligible; the earnings differences between black and white NLSY79 respondents are almost exactly consistent with their measured achievement differences. This is not the case for the male/female gap; the test items predict small advantages for men in math and even smaller advantages for women in reading even though the actual gaps are quite large in favor of men. The actual ln(wage_30) gap is 0.22, while for ln(pdv_labor) the gap is even larger, at 0.44. These gaps are more then 4 times as large as what the male advantage in math would predict. (The larger actual gap for ln(pdv_labor) is consistent with the fact that women are less likely to be engaged in full time employment, implying that wage differences understate overall earnings differences.) Finally, the predicted high-/low-income gaps, though sizable, are only about half as large as the actual gaps. As with education, the realized advantages for youth from high-income households are greater than what can be predicted on the basis of their measured achievement alone. This result stands in stark contrast to the gaps by race which are predicted almost perfectly by test scores alone.

The black-white wage gaps reported in Table 4 are consistent with Neal and Johnson (1996). These authors find that adding AFQT to log wage regressions reduces the black-white gap for men by about two-thirds and actually pushes the black-white gap for women modestly positive. In order to compare my estimates to theirs, I need to estimate wage gaps separately by gender. Table6 estimates black/white and high-/low-income gaps on the male-only subset of my data. The item-anchored scores explain between 75%-90% of

the observed black/white wage and labor income gaps – less than my full sample results but more than Neal and Johnson (1996). By contrast, share of the high-/low-income gap that can be explained actually falls slightly. Table 7 repeats the analysis for the female-only subsample. Again, the share of the high-/low-income gap that can be explained is quite a bit lower. Interestingly, and in contrast to Neal and Johnson (1996), I find that test items over-predict the black/white earnings gap among women: for both wage income and full labor income, I predict gaps that are about 40% larger than observed.

Using the full sample, rather than just white men, to estimate the anchored scales yields broadly similar results. Table 5 shows that while the differences between the given gaps and the anchored gaps are typically (though not uniformly) smaller, the qualitative conclusions are largely unchanged. The one exception is the male/female reading gap. The item-anchored and given gaps in reading were very similar to each other using white male prices. Using the full sample instead to estimate the anchoring relationship yields a large positive gap for men in reading. That is, the item-anchored scales estimated on the full sample suggest that men have a sizable advantage in both math and reading, while the given scores and the item-anchored scores using just white male skill prices find a significant reading advantage for women and a significant math advantage for men. This is intuitive – women are half the sample and earn far less than men. The item-anchored scale estimated on the full sample will thus tend to pick out particularly "male" items in order to predict earnings, making it appear as though men have a sizable advantage in both reading and math.

7 The "Reading Puzzle" and Item-Anchoring

Prior research (Sanders, 2016; Kinsler and Pavan, 2015) has noted that in multivariate regressions of outcomes on math and reading scores, the coefficients on reading are often much smaller than the coefficients on math. In some cases, the reading coefficients are

even significantly negative. This section demonstrates that using item-anchored scores can resolve this puzzle. Item-anchored math and reading scores both have large and significant coefficient estimates in the types of multivariate regressions that typically give rise to the the reading puzzle. Nonetheless, the coefficient estimates on item-anchored reading achievement are still typically smaller than those on math.

Table 8 presents the regression coefficients on math and reading for a number of model specifications when ln(wage_30) is the outcome variable. The table presents different specifications which controls or not for education and parental income. The estimated coefficients are large and significantly positive for math but small and insignificant for reading when the given scores are used (odd-numbered columns).[13] A one standard deviation increase in the given math score is associated with a 0.1-0.18 increase in log wages at age 30, while a similar increase in the given reading score corresponds to a -0.01 to 0.03 change in log wages. Switching from the given z scores to the wage_30 item-anchored scores barely shifts the math coefficient estimates while dramatically increasing the reading coefficient estimates. The item-anchored scores imply that a one standard deviation in reading skill corresponds to a 0.05-0.09 increase in log wages. The anchored reading estimates are typically a bit more than half as large as the math estimates.

It is important to emphasize that these results are not tautological – the anchored math and reading scales are constructed independently of each other. Nonetheless, the scales are constructed to be maximally predictive of ln(wage_30). This may make their predictive significance less surprising. However, Table 9 shows that using item scales anchored to outcomes other than ln(wage_30) largely yields the same set of results: the estimated item-anchored reading coefficients are significantly above 0 and much closer to the corresponding math estimates.

The dependent variable in every model presented in Table 9 is ln(wage_30). Columns

[13]The smallest difference is between columns (7) and (8) which run the models on the full NLSY79 rather than the white male subsample. As discussed in section 6 wage data is often missing for women and minorities limiting the interpretability of these estimates.

(2) and (5) show that using scales anchored to other wage measures barely changes the math and reading estimates whether or not additional controls are included. Using high school anchored scores (columns (4) and (8)) yields math and reading coefficients that are significantly greater than 0 and very close to each other. The reading coefficients using college anchored scores (columns (3) and (7)) are comparatively closer to 0.[14] Nonetheless, even in this case, the reading estimates are still significantly above the corresponding estimates using the given scores, and the gap between the reading and math estimates is also much smaller.

8 Discussion and Conclusion

This paper demonstrates the value of using item-level data in the economic analysis of achievement. Alternative test scales based on anchoring item level data to outcomes rank students very differently than given scales and have important implications for measures achievement gaps by race, gender, and parental income. Generally, these gaps are 15-50% larger using scales anchored at the item level to school completion or labor income outcomes. For example, while black/white earnings inequality can be fully explained by differences in these item-anchored scales, only about half of the gap in adult earnings between youth from high- and low-income households can be explained.

In addition to these empirical results, the paper develops a method for using item-level data to estimate test reliability. This is an important contribution because the relevant measure of reliability depends on the anchoring outcome and need not be related to the reliability reported by the test designers. While other work (Bond and Lang, 2018) has shown how to estimate the anchored reliabilities using lagged test scores in panel data, panel data is not always available. Moreover, panel methods will classify skills that are predictive of outcomes but which are not tested in consecutive years as measurement error,

[14]The smaller reading estimates using college anchored scores may reflect the weaker relationship between reading items and college completion previously documented.

whereas the method developed here will not. Nonetheless, there is still more work to be done methodologically, as the split of items into "test scale" and "instrument" uses was carried out in an ad hoc, though intuitive, manner. Efficiency gains from an improved procedure are likely. Additionally, the method allows for the direct estimation of the relevant measure of test reliability. The direct estimation of reliability has two significant advantages. First, the the anchoring-relevant reliability may be quite different than the reliability calculated using psychometric methods. Second, the reliability is estimated, so its simultaneous estimation allows for the construction of standard errors which take this estimation into account.

There are a number of potentially interesting extensions to the work presented here. First, one could carry out the anchoring analysis using the items from various noncognitive/behavioral assays such as the Rotter Locus of Control and the Rosenberg Self Esteem scales that are included in the NLSY79. Second, one could investigate which test items are driving the differences between the given and anchored scales. Unfortunately, the content of the ASVAB items is not publicly available. The best one can do, then, is to connect the items to the IRT item-level parameters, which are reported. Finally, one could estimate item-anchored scores in the CNLSY, which has the children of the women in the NLSY79. Although the achievement tests in the NLSY79 and CNLSY are different, one might nonetheless be able to estimate a new measure of intergenerational persistence in human capital.

Overall, the results in this paper suggest that economists would do well to consider more closely the alignment between the assessments they are using and the economic outcomes that they are ultimately interested in. Although test scores are strongly related to outcomes, they are not designed with these outcomes in mind. Achievement test scores can thus obscure important relationships between skills and outcomes with significant effects on measured achievement inequality.

References

Arcidiacono, P. (2004). Ability sorting and the returns to college major. *Journal of Econometrics*, 121(1):343 – 375. Higher education (Annals issue).

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)*.

Clotfelter, C., Ladd, H., and Vigdor, J. (2009). The Academic Achievement Gap in Grades 3-8. *The Review of Economics and Statistics*, 91:398–419.

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.

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.

Hanushek, E. and Rivkin, S. (2012). The Distribution of Teacher Quality and Implications for Policy. *Annual Review of Economics*, 4:131–57.

Hanushek, E. A. and Rivkin, S. G. (2009). Harming the Best: How Schools Affect the Black-White Achievement Gap. *Journal of Policy Analysis and Management*, 28(3):366–393.

Hoxby, C. (2000). The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *Quarterly Journal of Economics*, 115(4):1239–1285.

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.

Lord, F. (1975). The 'Ability' Scale in Item Characteristics Curve Theory. *Psychometrika*, 40:205–217.

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, Board of Governors of the Federal Reserve System*.

Nielsen, E. (2015b). The Income-Achievement Gap and Adult Outcome Inequality. *Finance and Economics Discussion Series, Board of Governors of the Federal Reserve System (U.S.)*, 041.

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.

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(Supplement C):337 – 358.

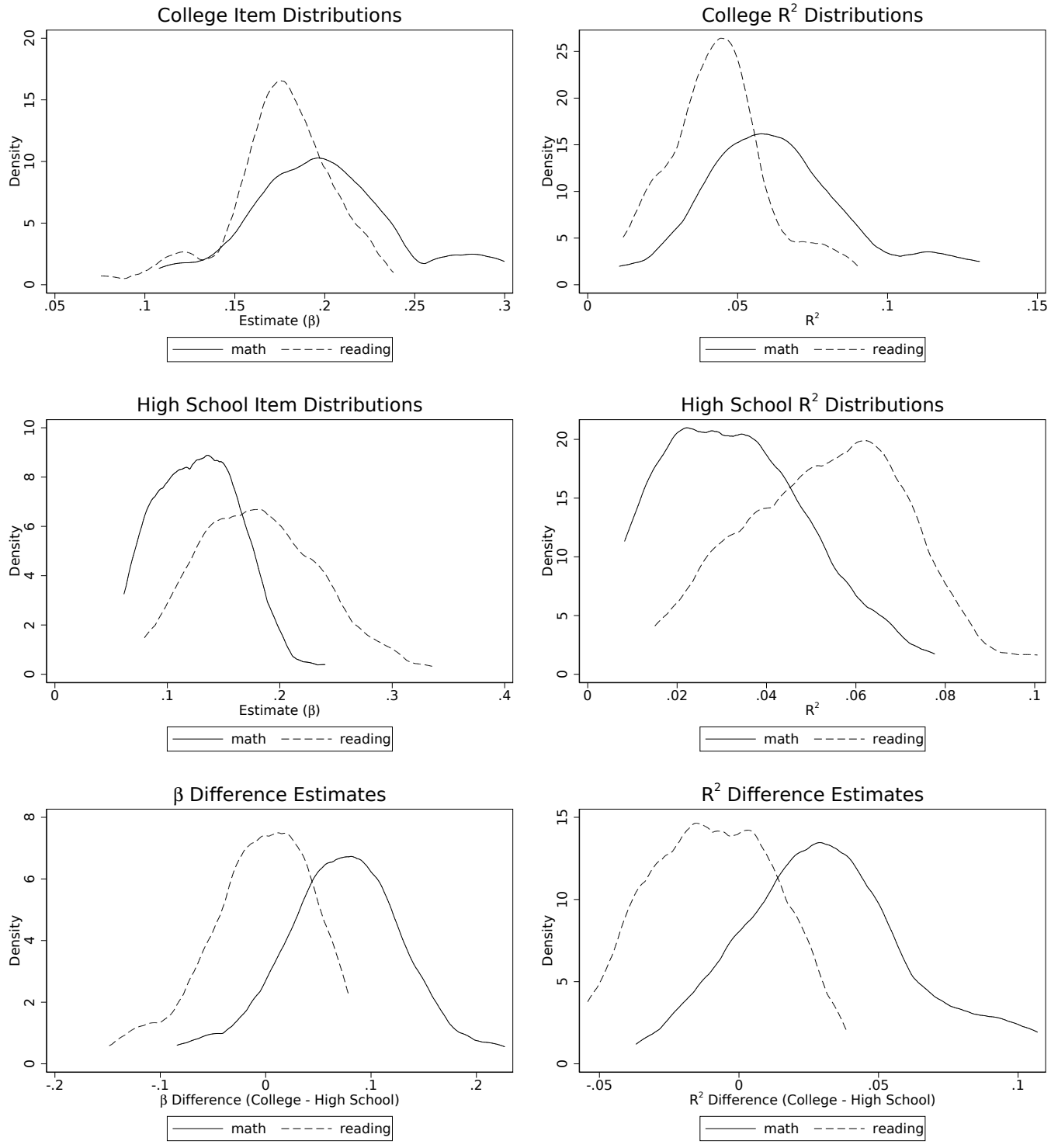
A Appendix

Table 1: Summary Statistics

Variable	Mean	Std. Dev.	N
base year age	17.77	2.33	12686
male	0.51	0.5	12686
black	0.14	0.35	12686
base year hh income ($1,000)	70.12	46.33	12249
high school	0.89	0.31	12686
college	0.23	0.42	12686
highest grade completed	13.26	2.48	12686
pdv_labor ($1,000)	434.85	332.99	12686
wage_30	19.43	11.27	8667
math	99.07	18.95	11878
reading	98.32	19.32	11878
afqt	147.85	27.16	11878
ar missing	0.09	0.29	12686
wk missing	0.09	0.29	12686
pc missing	0.09	0.29	12686
mk missing	0.09	0.29	12686

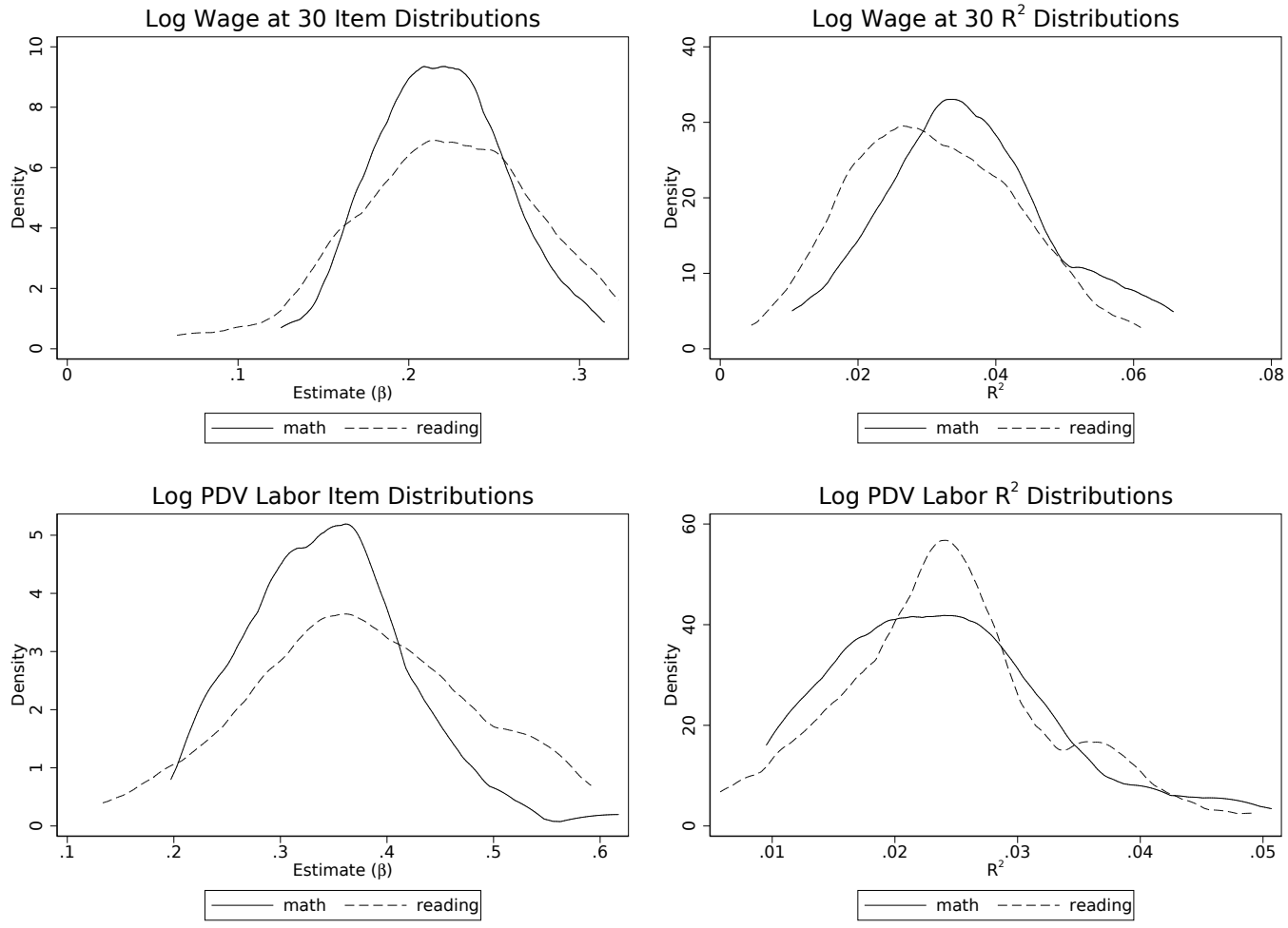
Note: All statistics use base-year sampling weights. Dollar values in 2017-constant dollars deflated using the CPI-U. A discount rate of 5% is used to construct pdv_labor.

Figure 1: School Completion Item-by-Item Regression Coefficient and R^2 Distributions



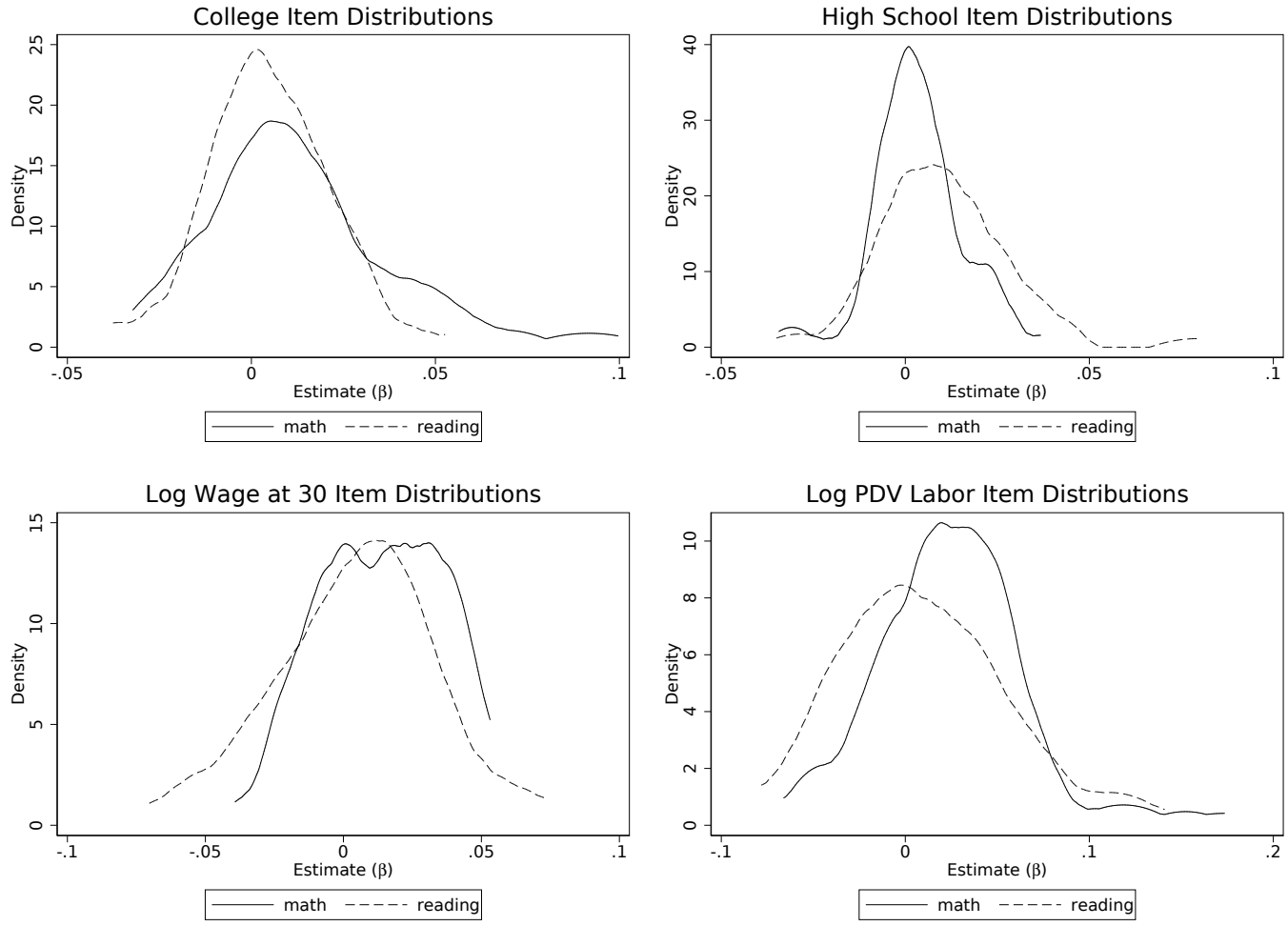
Note: Panels depict kernel densities across test items (i) for regressions of the form $y_j = \alpha_i + \beta_i D_{i,j} + \varepsilon_{i,j}$ where y_j is a school completion indicator (high school or college) for youth j.

Figure 2: Labor Income Item-by-Item Regression Coefficient and R^2 distributions



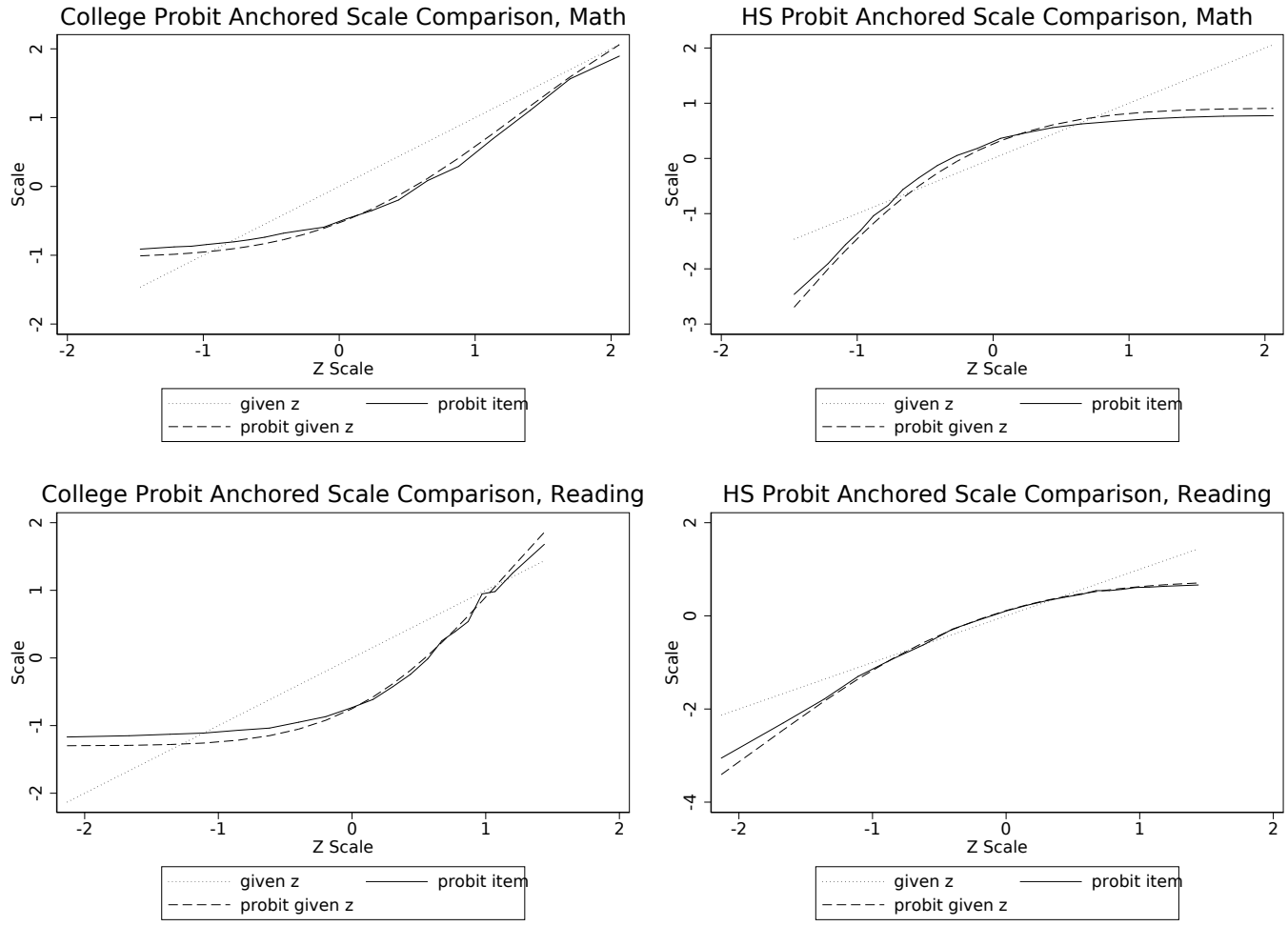
Note: Panels depict kernel densities across test items (i) for regressions of the form $y_j = \alpha_i + \beta_i D_{i,j} + \varepsilon_{i,j}$ where y_j is either the log wages at age 30 or the log present discounted value of labor income for youth j.

Figure 3: Item Regression Coefficient, Full Combined Regressions



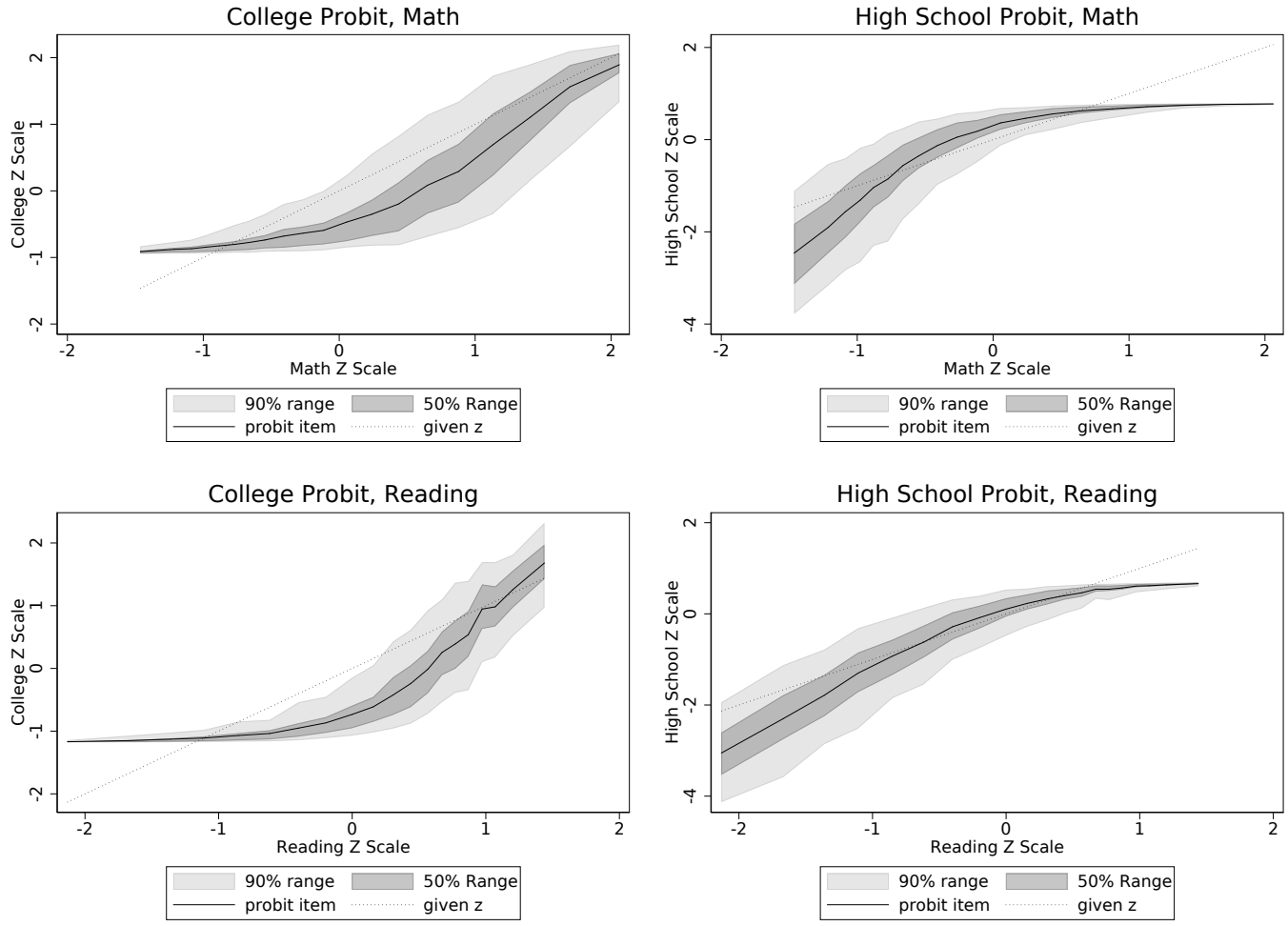
Note: Panels depict kernel densities across test items (i) for regressions of the form $y_j = \alpha + \beta_1 D_{1,j} + \ldots + \beta_N D_{N,j} + \varepsilon_{i,j}$ for various indicated outcomes y_j.

Figure 4: Schooling Completion Mean Scales for Math and Reading



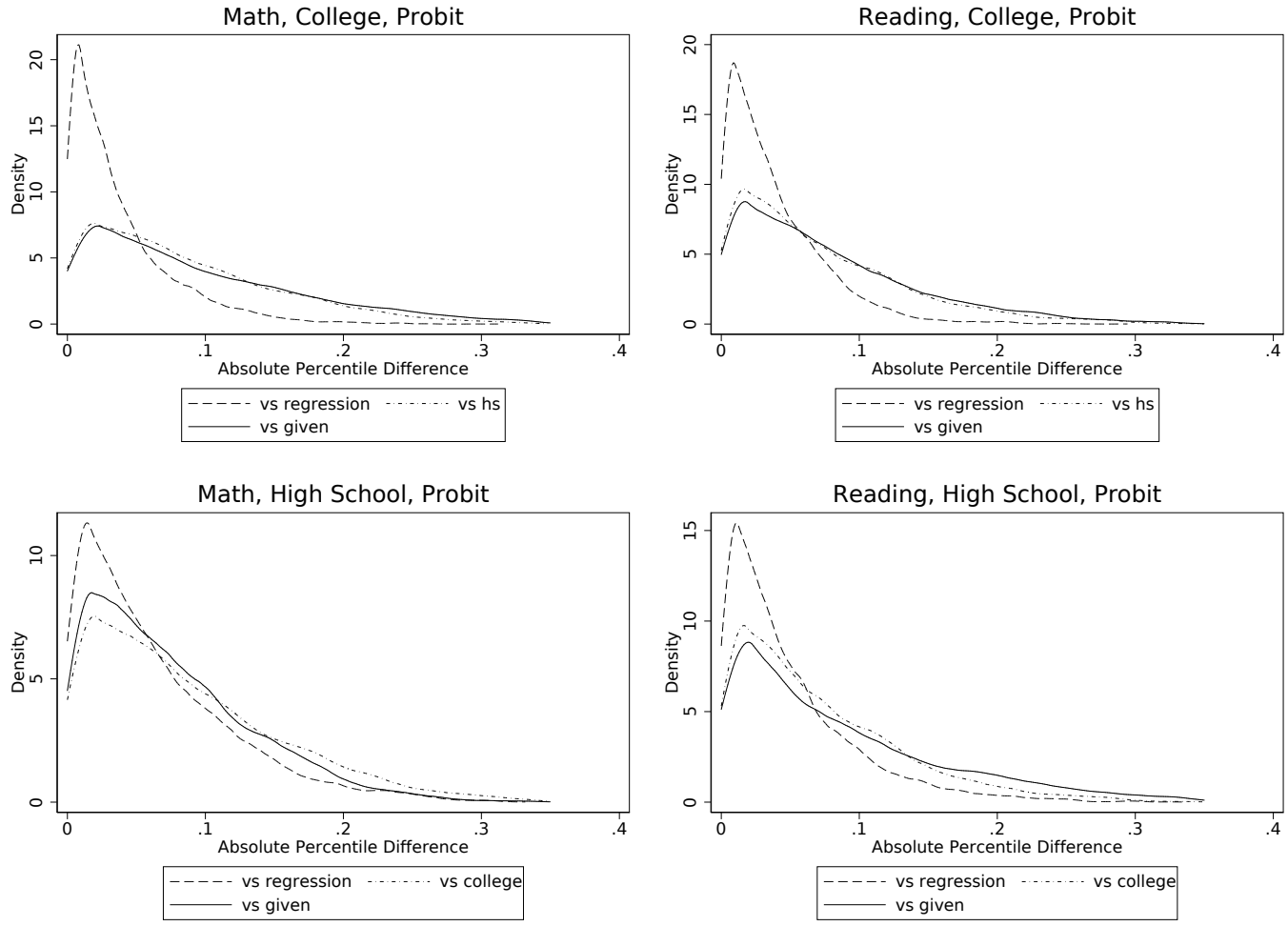
Note: The "probit given z" scales are constructed using simple probits of the age-adjusted z scores on either high school or college indicators. The "probit item" scales are based on linear, item-level probit regressions. The mean predicted values of the item-level regressions are plotted for each ventile of the given z distribution.

Figure 5: School Completion Scales for Math and Reading



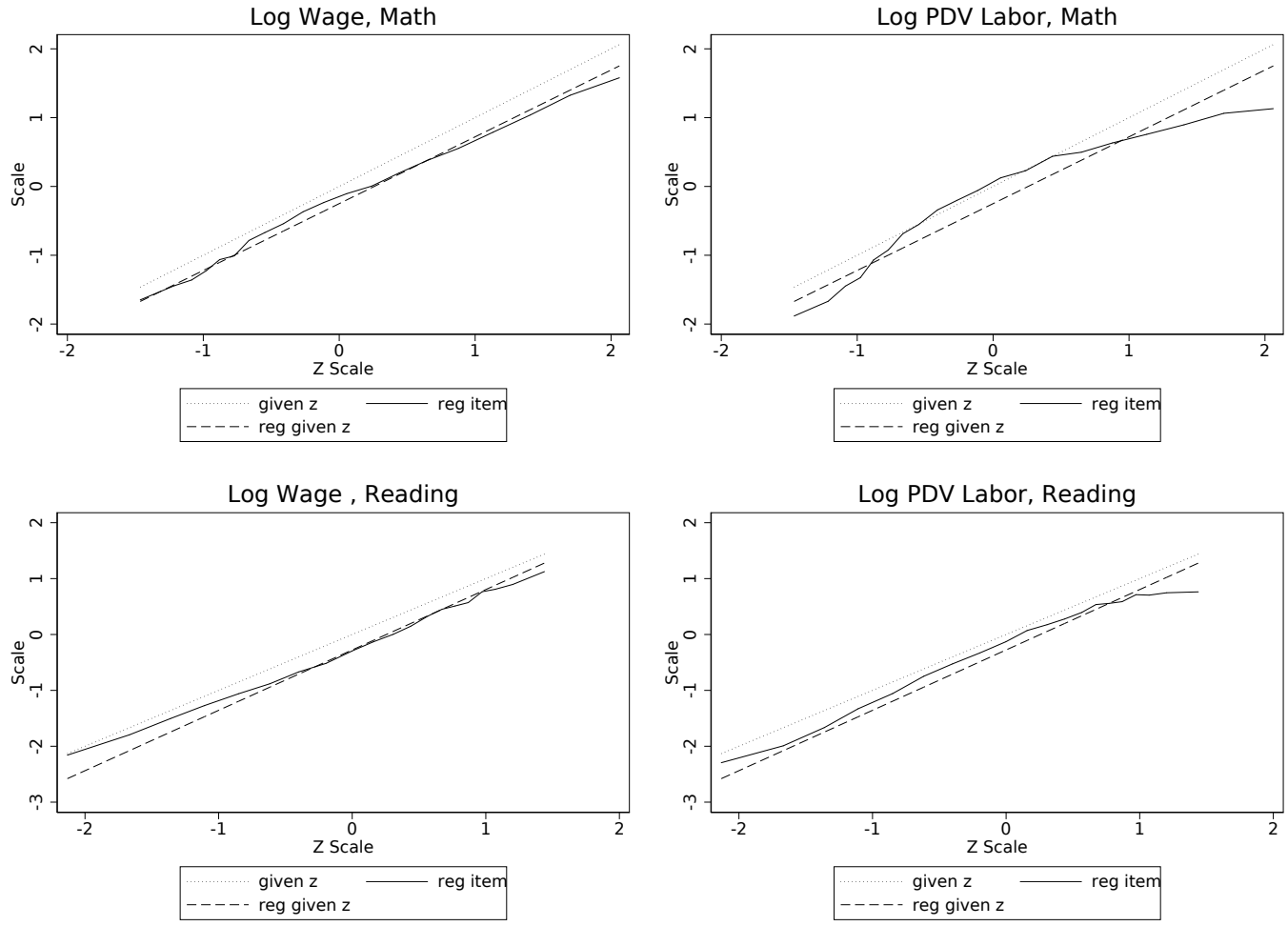
Note: The "probit item" scales are based on linear, item-level probit regressions. The mean predicted values of the item-level regressions are plotted for each ventile of the given *z* distribution, along with the middle 50% and middle 90% range.

Figure 6: Percentile Differences, Schooling Scales



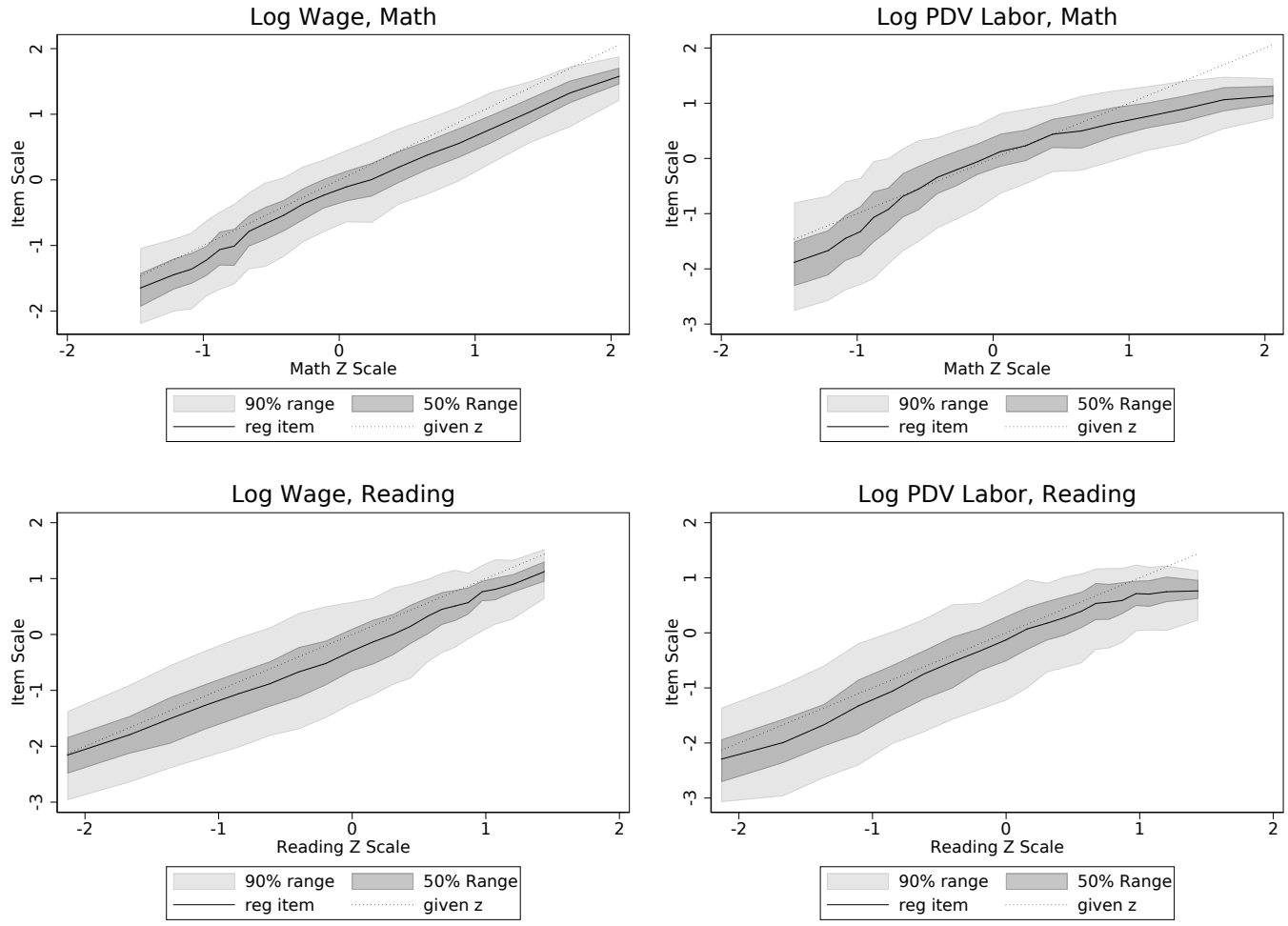
Note: Each panel plots the distribution of the percentile differences between the titular probit item-anchored scale and (1) the regression item-anchored scale ("vs regression"), (2) the given scale ("vs given"), and (3) the probit anchored scale using the other schooling outcome ("vs hs" or "vs college").

Figure 7: Labor Income Scale for Math and Reading



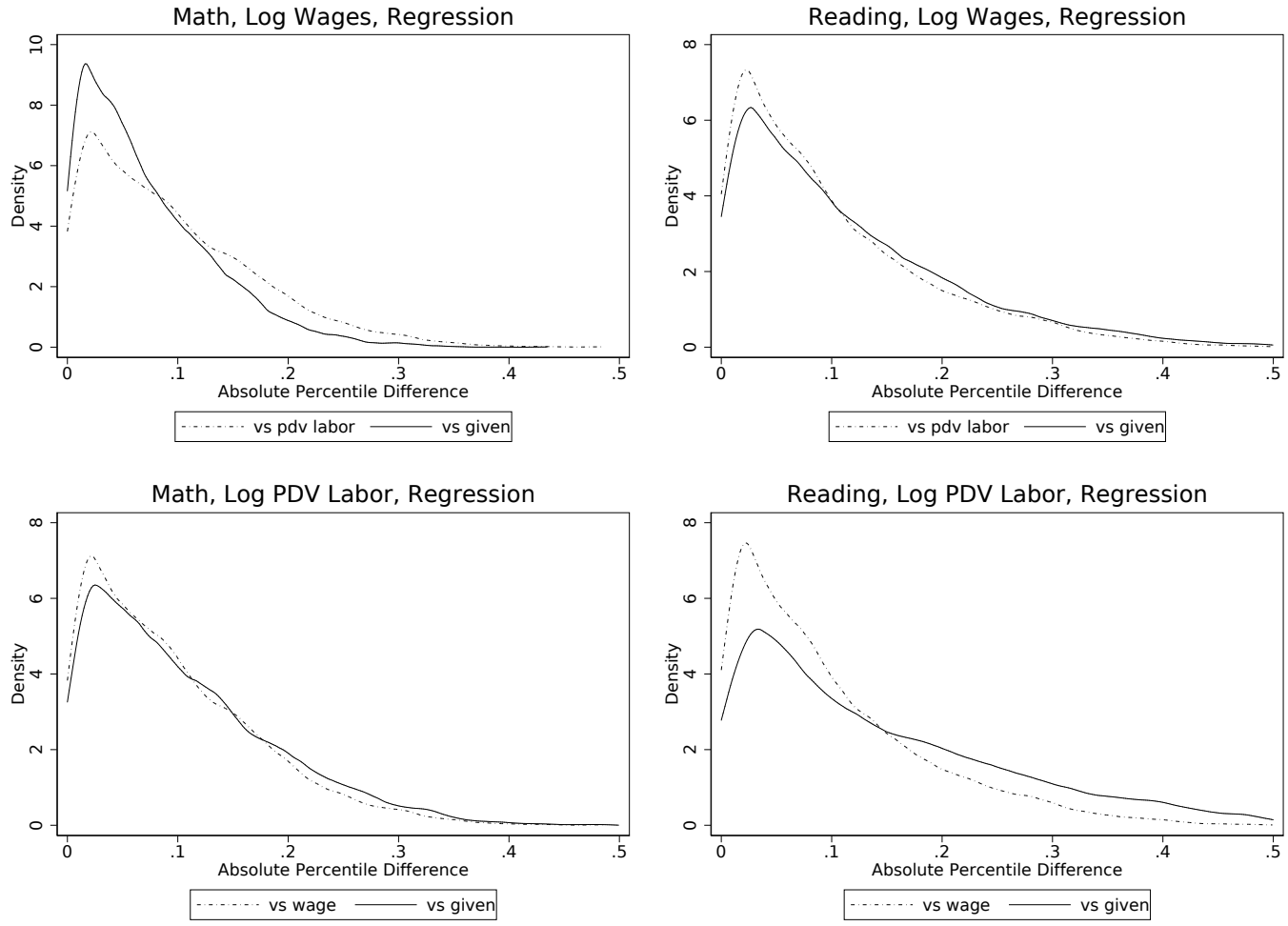
Note: The "reg given z" scales are constructed using simple regressions of the age-adjusted z scores on the logarithms of the pdv_labor variables. The "reg item" scales are based on linear, item-level probit regressions (in log space). The mean predicted values of the item-level regressions are plotted for each ventile of the given *z* distribution.

Figure 8: Labor Income Scales for Math and Reading



Note: The "reg item" scales are based on linear, item-level probit regressions (in log space). The mean predicted values of the item-level regressions are plotted for each ventile of the given z distribution, along with the middle 50% and middle 90% range.

Figure 9: Percentile Differences, Labor Income Scales



Note: Top panels plot kernel density estimates of the individual-level percentile differences between the log wage anchored scales and either the log(pdv_labor) scales or the given scales. Bottom panels repeat the analysis for log(pdv_labor).

Table 2: Naive, Asymptotic, and Bootstrapped Standard Errors

Test	outcome	gap	naive	asymp	boot
Math	hs	white/black	0.03	0.04	0.06
Reading	hs	white/black	0.03	0.04	0.05
Math	college	white/black	0.03	0.03	0.03
Reading	college	white/black	0.04	0.04	0.04
Math	log labor	white/back	0.03	0.06	0.06
Reading	log labor	white/black	0.04	0.06	0.07
Math	log wages	white/black	0.03	0.05	0.06
Reading	log wages	white/black	0.03	0.06	0.07
Math	hs	male/female	0.02	0.02	0.03
Reading	hs	male/female	0.02	0.02	0.03
Math	college	male/female	0.02	0.02	0.03
Reading	college	male/female	0.03	0.03	0.04
Math	log labor	male/female	0.02	0.03	0.04
Reading	log labor	male/female	0.02	0.03	0.03
Math	log wages	male/female	0.02	0.02	0.04
Reading	log wages	male/female	0.02	0.02	0.03
Math	hs	high/low	0.03	0.04	0.06
Reading	hs	high/low	0.03	0.04	0.05
Math	college	high/low	0.03	0.03	0.05
Reading	college	high/low	0.04	0.04	0.06
Math	log labor	high/low	0.03	0.06	0.07
Reading	log labor	high/low	0.03	0.06	0.07
Math	log wages	high/low	0.03	0.05	0.07
Reading	log wages	high/low	0.03	0.06	0.07

Note: The "naive" standard errors are calculated without accounting for the sampling variation in $1/\hat{\gamma}^{(1)}$. The "asymp" standard errors account for this sampling variability using the delta method. The "boot" standard errors are based on a normal approximation using 250 bootstrapped estimates where the instruments use $\hat{A}^{(2)}$ sorted into 20 (rather than 200) equinumerous bins.

Table 3: NLSY79 Item-Anchored School Completion Gaps, Probit

White/Black	z	noitem z	item z	predicted	actual	item R
math, college	0.99	0.98	0.80	0.20	0.13	0.88
	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)	.
reading, college	1.06	1.21	1.28	0.25	0.13	0.74
	(0.02)	(0.03)	(0.04)	(0.01)	(0.01)	.
math, hs	0.99	1.20	1.18	0.15	0.06	0.83
	(0.03)	(0.03)	(0.03)	(0.00)	(0.01)	.
reading, hs	1.06	1.39	1.19	0.16	0.06	0.87
	(0.02)	(0.03)	(0.03)	(0.00)	(0.01)	
Male/Female						
math, college	0.17	0.23	0.13	0.03	-0.01	0.88
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	.
reading, college	-0.10	-0.10	0.00	0.00	-0.01	0.74
	(0.02)	(0.02)	(0.03)	(0.01)	(0.01)	.
math, hs	0.17	0.12	0.05	0.01	-0.03	0.83
	(0.02)	(0.02)	(0.02)	(0.00)	(0.01)	.
reading, hs	-0.10	-0.15	-0.08	-0.01	-0.03	0.87
	(0.02)	(0.02)	(0.02)	(0.00)	(0.01)	.
High/Low Income						
math, college	0.99	1.06	0.94	0.23	0.29	0.88
	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)	.
reading, college	0.91	1.16	1.19	0.23	0.29	0.74
	(0.03)	(0.03)	(0.04)	(0.01)	(0.01)	.
math, hs	0.99	1.07	1.05	0.13	0.18	0.83
	(0.03)	(0.03)	(0.03)	(0.00)	(0.01)	.
reading, hs	0.91	1.10	0.99	0.13	0.18	0.87
	(0.03)	(0.04)	(0.03)	(0.00)	(0.01)	.

Note: All scales constructed using age indicators in addition to item-level indicators. Instruments use the outcome scale divided into 200 equinumerous bins. Standard errors treat the estimated reliabilities as known with certainty.

Table 4: NLSY79 Item-Anchored Log Labor Earnings Gaps, Regression, White Male Prices

White/Black	z	noitem z	item z	predicted	actual	item R
math, wage	0.99	1.13	1.22	0.24	0.24	0.74
	(0.03)	(0.03)	(0.04)	(0.01)	(0.02)	.
reading, wage	1.06	1.41	1.24	0.23	0.24	0.84
	(0.02)	(0.03)	(0.03)	(0.01)	(0.02)	.
math, pdv_labor	0.99	1.13	1.62	0.48	0.44	0.64
	(0.03)	(0.03)	(0.04)	(0.01)	(0.03)	.
reading, pdv_labor	1.06	1.41	1.48	0.42	0.44	0.70
	(0.02)	(0.03)	(0.04)	(0.01)	(0.03)	.
Male/Female						
math, wage	0.17	0.20	0.24	0.05	0.22	0.74
	(0.02)	(0.02)	(0.03)	(0.01)	(0.01)	.
reading, wage	-0.10	-0.14	-0.10	-0.02	0.22	0.84
	(0.02)	(0.02)	(0.02)	(0.00)	(0.01)	.
math, pdv_labor	0.17	0.20	0.29	0.09	0.45	0.64
	(0.02)	(0.02)	(0.03)	(0.01)	(0.02)	.
reading, pdv_labor	-0.10	-0.14	-0.09	-0.02	0.45	0.70
	(0.02)	(0.02)	(0.03)	(0.01)	(0.02)	.
High/Low Income						
math, wage	0.99	1.13	1.20	0.24	0.46	0.74
	(0.03)	(0.03)	(0.04)	(0.01)	(0.02)	.
reading, wage	0.91	1.22	1.11	0.20	0.46	0.84
	(0.03)	(0.03)	(0.03)	(0.01)	(0.02)	.
math, pdv_labor	0.99	1.13	1.36	0.40	0.86	0.64
	(0.03)	(0.03)	(0.04)	(0.01)	(0.03)	.
reading, pdv_labor	0.91	1.22	1.21	0.35	0.86	0.70
	(0.03)	(0.03)	(0.04)	(0.01)	(0.03)	.

Note: All scales constructed using age indicators in addition to item-level indicators. Instruments use the outcome scale divided into 200 equinumerous bins. Standard errors treat the estimated reliabilities as known with certainty.

Table 5: NLSY79 Item-Anchored Log Labor Earnings Gaps, Regression

White/Black	z	noitem z	item z	predicted	actual	item R
math, wage	0.99	1.13	1.18	0.26	0.24	0.84
	(0.03)	(0.03)	(0.03)	(0.01)	(0.02)	.
reading, wage	1.06	1.41	1.34	0.27	0.24	0.81
	(0.02)	(0.03)	(0.03)	(0.01)	(0.02)	.
math, pdv_labor	0.99	1.13	1.23	0.46	0.44	0.86
	(0.03)	(0.03)	(0.03)	(0.01)	(0.03)	.
reading, pdv_labor	1.06	1.41	1.24	0.43	0.44	0.90
	(0.02)	(0.03)	(0.03)	(0.01)	(0.03)	.
Male/Female						
math, wage	0.17	0.20	0.24	0.05	0.22	0.84
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	.
reading, wage	-0.10	-0.14	0.17	0.03	0.22	0.81
	(0.02)	(0.02)	(0.02)	(0.00)	(0.01)	.
math, pdv_labor	0.17	0.20	0.24	0.09	0.45	0.86
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	.
reading, pdv_labor	-0.10	-0.14	0.15	0.05	0.45	0.90
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	.
High/Low Income						
math, wage	0.99	1.13	1.13	0.25	0.46	0.84
	(0.03)	(0.03)	(0.03)	(0.01)	(0.02)	.
reading, wage	0.91	1.22	1.21	0.24	0.46	0.81
	(0.03)	(0.03)	(0.03)	(0.01)	(0.02)	.
math, pdv_labor	0.99	1.13	1.04	0.39	0.86	0.86
	(0.03)	(0.03)	(0.03)	(0.01)	(0.03)	.
reading, pdv_labor	0.91	1.22	1.01	0.35	0.86	0.90
	(0.03)	(0.03)	(0.03)	(0.01)	(0.03)	.

Note: All scales constructed using age indicators in addition to item-level indicators. Instruments use the outcome scale divided into 200 equinumerous bins. Standard errors treat the estimated reliabilities as known with certainty.

Table 6: NLSY79 Item-Anchored Log Labor Earnings Gaps, Regression, White Male Prices, Male Sample

White/Black	z	noitem z	item z	predicted	actual	item R
math, wage	1.04	1.19	1.27	0.25	0.30	0.74
	(0.04)	(0.05)	(0.05)	(0.01)	(0.03)	.
reading, wage	1.07	1.42	1.24	0.23	0.30	0.84
	(0.04)	(0.05)	(0.05)	(0.01)	(0.03)	.
math, pdv_labor	1.04	1.19	1.66	0.49	0.55	0.64
	(0.04)	(0.05)	(0.06)	(0.02)	(0.04)	.
reading, pdv_labor	1.07	1.42	1.50	0.43	0.55	0.70
	(0.04)	(0.05)	(0.06)	(0.02)	(0.04)	.
High/Low Income						
math, wage	0.96	1.10	1.14	0.22	0.48	0.74
	(0.04)	(0.05)	(0.05)	(0.01)	(0.03)	.
reading, wage	0.90	1.21	1.14	0.21	0.48	0.84
	(0.04)	(0.05)	(0.05)	(0.01)	(0.03)	.
math, pdv_labor	0.96	1.10	1.32	0.39	0.92	0.64
	(0.04)	(0.05)	(0.06)	(0.02)	(0.04)	.
reading, pdv_labor	0.90	1.21	1.29	0.37	0.92	0.70
	(0.04)	(0.05)	(0.06)	(0.02)	(0.04)	.

Note: All scales constructed using age indicators in addition to item-level indicators. Instruments use the outcome scale divided into 200 equinumerous bins. Standard errors treat the estimated reliabilities as known with certainty.

Table 7: NLSY79 Item-Anchored Log Labor Earnings Gaps, Regression, White Male Prices, Female Sample

White/Black	z	noitem z	item z	predicted	actual	item R
math, wage	0.93	1.06	1.16	0.23	0.17	0.74
	(0.04)	(0.04)	(0.04)	(0.01)	(0.03)	.
reading, wage	1.06	1.41	1.24	0.23	0.17	0.84
	(0.03)	(0.04)	(0.04)	(0.01)	(0.03)	.
math, pdv_labor	0.93	1.06	1.56	0.46	0.33	0.64
	(0.04)	(0.04)	(0.06)	(0.02)	(0.04)	.
reading, pdv_labor	1.06	1.41	1.46	0.42	0.33	0.70
	(0.03)	(0.04)	(0.05)	(0.01)	(0.04)	.
High/Low Income						
math, wage	1.01	1.15	1.23	0.24	0.42	0.74
	(0.04)	(0.04)	(0.05)	(0.01)	(0.03)	.
reading, wage	0.93	1.24	1.10	0.20	0.42	0.84
	(0.03)	(0.04)	(0.05)	(0.01)	(0.03)	.
math, pdv_labor	1.01	1.15	1.37	0.41	0.77	0.64
	(0.04)	(0.04)	(0.06)	(0.01)	(0.04)	.
reading, pdv_labor	0.93	1.24	1.13	0.32	0.77	0.70
	(0.03)	(0.05)	(0.06)	(0.02)	(0.04)	.

Note: All scales constructed using age indicators in addition to item-level indicators. Instruments use the outcome scale divided into 200 equinumerous bins. Standard errors treat the estimated reliabilities as known with certainty.

Table 8: Reading Puzzle Regressions – Wages at Age 30

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	z	item z	z	item z	z	item z	z	item z
math	0.17***	0.15***	0.11***	0.12***	0.10***	0.11***	0.09***	0.10***
	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
read	0.03	0.09***	-0.00	0.07***	-0.01	0.06***	0.03**	0.05***
	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
education	no	no	yes	yes	yes	yes	yes	yes
parental income	no	no	no	no	yes	yes	yes	yes
white male only	yes	yes	yes	yes	yes	yes	no	no
Observations	2363	2262	2363	2262	2289	2187	8012	7776
Adjusted R^2	0.115	0.156	0.138	0.166	0.169	0.194	0.211	0.219

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Table shows the estimated coefficients on math and reading for regression of the form $\ln(wage) = \alpha + \beta_1 math + \beta_2 read + \gamma X + \varepsilon$, where X denotes education, race, and parental income controls (or not, as indicated). Column labels correspond to either the given-z scores or the wage anchored scores.

Table 9: Reading Puzzle Regressions – Wages at Age 30, Alternative Scales

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	ln_w30	ln_p14	college	high school	ln_w30	ln_p14	college	high school
math	0.15***	0.14***	0.13***	0.11***	0.11***	0.09***	0.07***	0.05**
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)
read	0.09***	0.08***	0.05***	0.09***	0.06***	0.05***	0.03*	0.04**
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.02)
education	no	no	no	no	yes	yes	yes	yes
parental income	no	no	no	no	yes	yes	yes	yes
white male only	yes	yes	yes	yes	yes	yes	yes	yes
Observations	2262	2262	2262	2262	2187	2187	2187	2187
Adjusted R^2	0.156	0.128	0.109	0.080	0.194	0.180	0.163	0.161

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Table shows the estimated coefficients on math and reading for regression of the form $\ln(wage) = \alpha + \beta_1 math + \beta_2 read + \gamma X + \varepsilon$, where X denotes education, race, and parental income controls (or not, as indicated). Column labels correspond to different anchoring outcomes for the math and reading scales.

B Additional Analysis and Discussion

B.1 Item Response Theory and Economic Outcomes

This appendix formalizes some of the discussion in the introduction (Section 1) regarding standard psychometric models and economic outcomes. Throughout, I will consider the canonical 3PL IRT model; the arguments using other psychometric models are similar. In the 3PL model, the probability that a test-taker with ability θ answers question i correctly is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}},$$

where (c_i, b_i, a_i) are item specific parameters that govern how ability θ translates to correct answers (the item characteristic curve or ICC).[15]

This set-up shows the general difficulty with using IRT scales in economics research. The scale of θ is *defined* as the scale such that $P_i(\theta)$ has the above functional form. There is no economic reason why this scale should be interesting. This scale is generically not unique – Lord (1975) showed that there exist infinitely many alternative scales that can fit any set of item-response data equally well, provided that the appropriate changes are made to the functional form of the ICC curve. More importantly, any two test items that have the same item-level parameters (c_i, b_i, a_i) will contribute equally to the estimation of θ. But the skills measured by two such equivalent questions might differ dramatically in their importance.

Consider the following example, which expands on the discussion in Section 1. Consider a two question exam, where each item measures a unique skill. Let w denote the outcome (e.g. wages) that we care about, and suppose that skill 1 (measured by test item 1) leads to $w_1 > 0$, while skill 2 is worthless ($w_2 = 0$). To make the math as simple as possible, suppose that guessing is not possible, so that $c_i = 0$, and that each test item has infinite discrimination for its skill. In other words, the test-taker gets item i correct if and only if she possesses skill i. Finally, suppose the skills are independent of each other and that each is possessed by half of the population.

The assumption that the skills are equally prevalent and independent implies that the 3PL model will assign the same item parameters to both questions. Moreover, this implies that there will be 3 possible values of θ: θ_L = both items wrong, θ_M = one item correct and one item incorrect, and θ_H = both items correct. In order for θ to be distributed symmetrically with mean 0 and standard deviation 1, we must have

$$(\theta_L, \theta_M, \theta_H) = (-\sqrt{2}, 0, \sqrt{2}).$$

How informative is the θ scale? It is straightforward to calculate that $\mathbb{E}[w|\theta]$ is linear

[15] The parameter c_i is the guessing probability: as $\theta \to -\infty$, $P_i(\theta) \to c_i$. Similarly, a_i, the "discrimination," is the maximum slope of the ICC. Finally, b_i is the "difficulty," or the point in θ space at which the slope of a_i is realized.

in θ with a slope that depends on w_1:

$$\mathbb{E}[w|\theta] = \left(\frac{w_1}{2\sqrt{2}}\right)\theta.$$

Note that the slope in the above relationship is simply equal to the covariance of w and θ. Similarly, we can calculate $var(w|\theta)$, the variance of w conditional on θ:

$$var(w|\theta_L) = var(w|\theta_H) = 0, \;\; var(w|\theta_m) = \left(\frac{w_1}{\sqrt{2}}\right)^2$$

There is no unexplained variation in w when scores are low or high, because the low score guarantees the valuable skill s_1 is not possessed, while the high score guarantees that it is. There is still residual variance for middle scores because some of the middle score test-takers have the valuable skill, while others do not.

The conditional expectations and variances above are the best an analyst interested in outcome w could do if she were restricted to using the IRT-derived θ scale. However, clearly she could do better using item-level data. In particular, given data on w and each test item, she could figure out that item one yields w_1 always and item 2 is independent of w. Armed with this information, she could construct a new achievement scale, $\tilde{\theta}$ that using only item 1. This achievement scale would have two values: $\tilde{\theta}_L$ if item 1 is wrong, $\tilde{\theta}_H$ if item 1 is right. Of course, since the units of w are interpretable by hypothesis, a convenient to use would be $\tilde{\theta}_L = 0$ and $\tilde{\theta}_H = w_1$. However, to facilitate the comparison with θ, let us make $\tilde{\theta}$ a z-score by setting

$$(\tilde{\theta}_L, \tilde{\theta}_H) = (-1, 1).$$

It is straightforward to see that this scale is more informative than the old scale. First, the slope of the conditional expectation function is steeper:

$$\mathbb{E}[w|\tilde{\theta}] = \left(\frac{w_1}{2}\right)\tilde{\theta}.$$

Similarly, because $\tilde{\theta}$ perfectly reveals w, the residual variance is always 0:

$$var(w|\tilde{\theta}_L) = var(w|\tilde{\theta}_H) = 0.$$