The Variance of Achievement Increases During Childhood*

Eric Nielsen
Federal Reserve Board

July 18, 2023

Abstract

Economic analyses typically standardize test scores to have a unit variance within each grade or age. However, the variance in economically relevant achievement may not be constant across grades/ages, so a "standard deviation of achievement" may not have a constant meaning. This paper constructs economically interpretable, cardinal test scales by estimating the relationship between individual test items and outcomes such as school completion and early-career labor market earnings. The standard deviation of achievement according to these new scales increases by 50-400% between kindergarten and eighth grade. Standardizing these new test scales separately by grade completely obscures notable increases during childhood in the white-black achievement gap. Similarly, a reanalysis of Dahl and Lochner (2012) reveals that family income has much larger effects on achievement for older children and that this heterogeneity is more than completely obscured when scores are converted to standard deviation units by grade prior to analysis. Overall, the large increases in the variance of achievement documented here call into question many common analyses using age- or grade-standardized test scores.

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

1 Introduction

Large literatures in economics and social science use achievement test scores and other psychometric measures as outcomes, often with the interpretation that these measures are proxies for human capital. Because different test scales have different units, an almost universal practice in these literatures is to standardize scores to have a mean of zero and a standard

*Nathan Ausubel and Victoria Chbane provided outstanding research assistance. The analysis and conclusions set forth here are those of the authors and do not indicate concurrence by other members of the research staff, the Board of Governors, or the Federal Reserve System. Please do not cite or circulate. Contact: eric.r.nielsen@frb.gov

deviation (sd) of one and to then report estimates in the resulting "standard deviation units." Usually, this standardization is conducted separately by some combination of age, grade, and year. The assumption implicit in this approach is that a "standard deviation of achievement" is a well-defined concept that has a stable meaning across the groups being analyzed.

In this paper, I argue that the standardization of achievement scores to have a unit variance within a grade/age is not benign: the true, economically relevant standard deviation of achievement is much larger for older children.[1] Thus, a standard deviation corresponds to a much greater quantity of achievement in later grades. Empirical designs that pool sd-unit estimates across grades in effect blend estimates with very different true magnitudes.

To estimate the standard deviation of economically-relevant achievement in each grade, I follow Nielsen (2019) and estimate item-level models that relate individual test items (questions) to long-run, economically interpretable outcomes such as college completion, wage rates, and total labor market earnings in early adulthood. I then use these estimated models to predict outcomes for each test-taker based on their vector of item responses. The "item-anchored" scales thus aggregate the individual items in proportion to their utility in predicting outcomes, and they are in economically interpretable, outcome units. Importantly, because the amount of measurement error in these item-anchored test scores might differ systematically by grade, I estimate the grade-specific item-anchored scale reliabilities and adjust the cross-grade standard deviation comparisons accordingly.

As pointed out in prior literature going back to at least Stevens (1946), the application of standard statistical techniques such as regression, mean differences, etc. to standard psychometric scores introduces several key conceptual concerns. First, psychometrically derived test scores, which I will refer to as "given" scores, are not cardinal measures of achievement *in the contexts in which economists and social scientists typically use them.* That is, a fixed-magnitude change in a given scale will generally map non-linearly to changes in economically relevant and interpretable outcomes (Cunha et al., 2021). Second, the given scales are based on psychometric methods which aggregate vectors of item responses into scalars in ways that may be totally unrelated to the real-world, economic value of the skills covered by the items (Nielsen, 2019).[2] To be clear, these observations are not an indictment of psychometrics as a field – the problem lies with social scientists using psychometric measures in contexts where they were not designed to be used.

The use of item-anchored achievement scales solves both of these problems. The item anchored scales are cardinally interpretable: each unit change in an item-anchored score

[1] The meaning of a standard deviation might also differ over time, although the analysis in this paper does not speak to this question.

[2] See also Nielsen (2022); Bond and Lang (2018); Schroeder and Yitzhaki (2017); Cawley et al. (1999); Cunha et al. (2010); Jacob and Rothstein (2016); Lord (1975), among many others.

corresponds to a fixed increase the the predicted value of a cardinally interpretable outcome. Moreover, the item-anchored scales aggregate the individual item response data in a way that places greater weight on items that are relatively more predictive of the anchor (outcome).

I construct item-anchored achievement scales for children in pre-kindergarten through eighth grade using the National Longitudinal Survey of Youth 1979 Child and Young Adult Data (CNLSY). I use item-level data taken from the Peabody Individual Achievement Test (PIAT) and various long-run economic outcomes such as highest grade completed, college completion, and early adult (age 25) wages and total income. I also consider as a short-run outcome the standardized PIAT score at the next sitting of the test, which are administered every two years in the CNLSY. My baseline anchor models assume a grade-specific, linear relationship between the PIAT items and the outcome, residualized by survey year, race/sex indicators, and their interactions. To reduce the large number of potential parameters in these models and to avoid concerns related to over-fitting, I estimate the anchor models via lasso regression. Alternative, more flexible model specifications and different dimension-reduction techniques yield very similar estimates.

Within a given grade, the observed variance in the item-anchored scores reflects both the true variance of item-anchored achievement as well as grade-specific measurement error. If the amount of measurement error differs notably by grade, then its presence will bias naive estimates comparing the estimated variance of achievement in different grades. Fortunately, the method developed in Nielsen (2019) allows one to calculate both an item-anchored achievement scale and its reliability. I therefore adopt that paper's "split-half" approach in which the PIAT items are divided into two comparable groups each of which is then used to estimate a separate item-anchored scale. I use one of these scales as my baseline item-anchored achievement measure and the other to construct an instrumental variable which allows me to estimate the relevant reliability.

Depending on the subject matter of the test and the economic outcome used as the anchor, I estimate that the standard deviation of achievement in grades 7-8 is 50-400% larger than the standard deviation in pre-k and kindergarten. The scales anchored to log income and log wages at age 25 display the largest increases, while the school completion anchored increases are more modest (though still large), at around 50%. Scores anchored to the next-observed standardized PIAT scores also show increases of around 50%.

The use of item-level data is critical for these results. In an alternative anchoring analysis in which I follow Bond and Lang (2018) by anchoring the given PIAT scale scores flexibly to the same outcomes, the variances do not show clear trends for some anchors and show generally smaller increases for others.

The large increases in the standard deviations of item-anchored achievement scales has

implications for the measurement of trends in achievement inequality. For example, the mean white-black differences in item-anchored achievement show no clear trends when these scores are standardized separately by grade. However, when the scores are kept in their economically interpretable, non-standardized units, they generally show large increases in white-black achievement inequality. While the precise increases in percentage terms differ by anchor, the eighth grade white-black item-anchored gaps are generally about twice as large as the corresponding pre-k gaps. Moreover, because these gaps are in directly interpretable units, this analysis shows both that the level of the white-black achievement gap and its increase are economically very significant. The grade-8 item-anchored achievement gaps correspond to gaps of 15-20% in labor market income and wages at age 25, 0.14 in the probability of completing college, and about 0.6 grades completed.

These very large standard deviation increases between early and late elementary school more generally suggest the analyses which standardize scores by age or grade might either obscure or falsely generate age heterogeneity. Many empirical analyses using sd-unit test scores document larger effects/estimates for younger children. Additionally, research using such scores also commonly documents intervention "fade-out" whereby initially large causal effects diminish as the treated children age. The much larger spread in achievement for older children implies that both of these empirical regularities would arise as statistical artifacts from by-grade/age standardization.

I demonstrate that standardization can indeed create and obscure age heterogeneities through a reanalysis of Dahl and Lochner (2012), a paper which uses an IV strategy to argue that household income has a significant positive effect on standardized PIAT scores in the CNLSY data. Using Dahl and Lochner (2012)'s data and method, there is some weak evidence that the effect of income on achievement is larger for children under 12 years old. Using the item-anchored test scales standardized by grade, I likewise find some evidence of larger effects for this younger age group. By contrast, the non-standardized estimates, which are in economically interpretable "outcome units" tell a completely different story – the estimated effects are typically much larger for older children. Household income in fact has larger effects on achievement in older children, but, because the variance of achievement is also greater for older children, the sd-unit estimates obscure this fact.

The rest of the paper is organized as follows. Section 2 reviews the literatures on test scaling, anchoring, and intervention fade out. Section 3 discusses the CNLSY item and outcome data. Section 4 presents the empirical framework for measuring grade-level achievement variance using item-outcome relationships. Section 5 presents the headline empirical results on the growth in the spread in achievement from pre-kindergarten through eighth grade, while Section 6 shows that this growth in variance has significant consequences for estimated trends

in white-black achievement inequality. Section 7 shows that the growth in achievement also significantly alters estimated age heterogeneity in causal effects estimated using achievement test scores. Section 8 concludes. Appendix A collects additional empirical results.

2 Literature

This paper is not the first to note that psychometric scales are not cardinally interpretable, particularly for economic applications.[3] While this research has tended not to focus on higher-order moments, the non-cardinality of psychometric skills renders statistics such as the standard deviation uninterpretable (Stevens, 1946). In prior work, I have assessed the robustness of male-female comparisons of achievement variance (Domicolo and Nielsen, 2022), finding that these comparisons are quite sensitive to order-preserving rescalings of achievement test scales.

The anchoring methodology used in this paper builds on the foundations laid in Bond and Lang (2018) and Nielsen (2019, 2022).[4] In particular, both the method I use to anchor scores at the item level and the method I use to correct for possibly differential measurement error by grade are identical to those used in Nielsen (2019, 2022). In turn, those papers adapt the conceptual framework and empirical methodology from Bond and Lang (2018) to the item-anchoring case. Item-anchoring is also related to Bettinger et al. (2013), which finds that different ACT subtests are not equally useful at predicting college performance. Anchoring different AFQT subtests to log wages, Cawley et al. (1999) find substantial non-linearities that differ by subtest and age.

This paper contributes also to the literature on intervention fade out – the very common occurrence that measured effects from an educational intervention decrease in magnitude over time, in some cases disappearing completely. Bailey et al. (2020) reviews this literature and argues that estimated fade out is unlikely to reflect solely a statistical artifact. However, that claim relies on evidence using given, vertically-scaled achievement tests, not scales anchored at the item level to economically interpretable outcomes.

A number of papers within economics have explored the "statistical artifact" explanation for fade out. Lang (2010) argues that the fade out of teacher effects in later grades might be due to test score standardization. Cascio and Staiger (2012) tests this idea with a model of educational production that anchors final-grade, given test scores on college completion, finding that the variance of knowledge increases 37%-56% between kindergarten and the end

[3] See Lord (1975); Stevens (1946); Ballou (2009); Jacob and Rothstein (2016); Schroeder and Yitzhaki (2017); Bond and Lang (2018); Nielsen (2023), among many others.

[4] Anchoring is a popular method to handle the non-cardinality of test scores. Notable examples include Cunha et al. (2010), Cunha and Heckman (2008), Chetty et al. (2014), Jackson (2018), Cawley et al. (1997), among many others.

of high school. This estimated increase is quite a bit less than what I find over a shorter range of grades and thus can only account for a modest amount of (1) the larger estimated impacts (in sd units) of interventions in early grades and (2) the more rapid fade out of these same interventions.[5] Wan et al. (2021) show that plausible, order-preserving rescalings of given test scores can eliminate or reverse estimated fade out in a well-known RCT of an early-childhood mathematics intervention. Outside of economics, Hill et al. (2008) notes that typical annual growth rates, measured in standard-deviation units, are three times larger in kindergarten than in middle school, implying that a treatment effect of Δ sd in kindergarten would be the same as an effect of $\Delta/3$ sd in middle school when expressed in "months of school" units.

3 CNLSY Item and Outcome Data

This paper uses the National Longitudinal Survey of Youth 1979 Child and Young Adult Data (CNLSY). The CNLSY follows the children of women from the National Longitudinal Survey of Youth 1979 (NLSY79), a survey that itself tracks a nationally-representative cohort who were in their mid-teens in 1980. Thus, the children in the CNLSY were mostly born between the mid-1980s and 2000. The CNLSY contains detailed demographic data, a wide array of psychometric measures, and various longer-run outcomes including school completion and labor market outcomes.

The anchored achievement measures I construct use psychometric data culled from sittings of the PIAT math and reading recognition exams which were administered to CNLSY respondents between the ages of 5 and 14 in every survey wave. These exams have a number of features that make them uniquely well-suited to item anchoring. First, they contain the necessary item-response data. Second, it is feasible to pool the test data across both of these dimensions because the specific test items are the same across all survey waves and respondent ages.[6] Finally, the given PIAT scale scores are widely used and studied, so there are many papers in prior literature to which estimates using the item-anchored scales can be compared.

While children in all grades face the same test questions, the administration of the PIAT math and reading exams depends on the grade of the test-taker. Within each PIAT subject test, the questions are arrayed in order of increasing difficulty. For math, each test admin-

[5]The two methods differ in many ways. The model in Cascio and Staiger (2012) assumes that knowledge accumulation follows an AR(1) process with innovations whose variances either grow or shrink linearly across grades. These and other parameters are identified off of across-grade test-test and test-outcome correlations. Compared to my method, Cascio and Staiger (2012) places more parametric structure on the process of knowledge accumulation, considers only one outcome anchor, and does not utilize item-level data.

[6]This constancy is not stated explicitly in the CNLSY documentation but was confirmed in correspondence with BLS staff.

istration starts with a "basal" item that depends on the child's grade and some initial item responses, with older children generally starting at a higher-numbered (and therefore more difficult) item. The child then progresses through the test, answering progressively more difficult items, until a "ceiling" is reached where the child answers a sufficiently high share of items incorrectly, at which point the exam stops.[7] The process for reading recognition is similar, except that the basal item depends on the child's basal math item instead of her grade in school. Because of this dependence, I combine the math and reading recognition items together in my baseline measure of achievement. I also report results estimated on the math items alone.[8]

Table 1: Summary Statistics

	Mean	Std. Dev.	N
Male	0.51	0.50	9,222
Hispanic	0.08	0.27	9,222
Black	0.16	0.37	9,222
Birth Year	1987.18	6.46	9,222
Grade	3.60	1.38	9,222
College	0.28	0.45	6,694
Highest Grade Completed	13.41	2.12	6,168
Log Income at 25	10.13	0.97	5,440
Log Wage at 25	2.23	0.92	4,985

Using the longitudinal dimension of the CNLSY, I construct a number of economically relevant, long-run outcomes for each survey respondent. For school completion, I construct an estimate of the highest grade completed as well as an indicator equal to one if the child completed college. For labor market outcomes, I estimate the total wage earnings at age 25 as well as the hourly wage rate at age 25.[9] Table 1 presents summary statistics for our key

[7] In detail, the starting item depends on the child's grade. If the child answers the starting item incorrectly, the test moves back to the prior grade's starting item. This process is repeated iteratively until the child answers the starting item correctly or until item 1 is reached. From the resulting starting item, the student answers 5 consecutive items. If these are answered correctly, the exam proceeds. If not, items are next asked in reverse order (getting less difficult) until 5 consecutive items are answered correctly or until item 1 is reached. The final item in this sequence of 5 is the "basal" item. The test then proceeds from the basal until the ceiling is reached. In the 1986 and 1996-2014 survey waves, the ceiling is the last item in the first set of 7 consecutive items where 5 of the responses are incorrect. For 1988-1994, the ceiling is the last of 5 consecutive items answered incorrectly. For each sitting of the exam, the basal and ceiling items are noted and all item responses between them are recorded.

[8] Though I do not show them both for the sake of brevity and because of the challenge in imputing item responses below the basal item, the results using the reading items alone with the same imputation rules likewise find large increases in item-anchored variance.

[9] Total wage earnings at age 25 is constructed by adding together total income for each reported job in the CNLSY, replacing the reported income intervals with the midpoints of the intervals. A smoothed outcome for total wage earnings is then calculated to account for missing data points at age 25 by averaging the total wage earnings across ages 24, 25 and 26. The hourly wage rate at age 25 is constructed by dividing weekly earnings by hours worked in a week. Weekly earnings and hours worked in a week are constructed similarly

variables of analysis other than the PIAT test items.

4 Constructing Item-Anchored Achievement Scales

4.1 Conceptual Framework

This paper follows the empirical framework developed in Nielsen (2019), which itself modifies the framework in Bond and Lang (2018). Consider a survey participant i in grade g in CNLSY survey year t. Let S_i be some economically-relevant outcome for i, such as college completion or later-life earnings. Further, let $D_{i,g} = [d_{i,1,g}, \ldots, d_{i,M,g}]$ be the vector of item response indicators for some achievement test consisting of M dichotomous items: $d_{i,m,g} = 1$ if i answers question m correctly and zero otherwise. Finally, let $X_{i,g}$ be some other observable characteristics of the individual, such as their race, sex, and survey wave.

The achievement of i in grade g is *defined* as

$$A_{i,g} = \mathbb{E}[S_i | D_{i,g}, X_{i,g}]. \tag{1}$$

The actual outcome S_i can then be written as $S_i = A_{i,g} + \eta_{i,g}$ where $\mathbb{E}[A_{i,g}\eta_{i,g}] = 0$ by construction. $A_{i,g}$ is not observed, but $S_{i,g}$ and $D_{i,g}$ are. To estimate $A_{i,g}$, I therefore assume that $\mathbb{E}[S_i|D_{i,g}, X_{i,g}] = f(D_{i,g}, X_{i,g})$ for some known function f. Data on S_i and $(D_{i,g}, X_{i,g})$ can then be used to estimate $\hat{A}_{i,g} = \hat{f}(D_{i,g}, X_{i,g})$.[10]

This paper is concerned with $SD(A_{i,g})$. A naive estimate of this quantity is just the sample standard deviation of the anchored scores. However, this will be an overestimate of the true standard deviation because the anchored scales are estimated with error.

As demonstrated in Nielsen (2019), it is possible to estimate the reliability of the anchored test scales using a "split items" approach. The basic idea of the approach is to divide the test items into two disjoint groups, call them group (1) and group (2). Each group is used to create separate anchored scales, the group (1)-scale and the group (2)-scale. The group (1)-scale then serves as the scale from which I measure the standard deviation of achievement, while the group (2)-scale is used to construct an instrument which allows me to estimate the group (1)-scale's reliability.

The details of the method are as follows. For an estimated item-anchored achievement measure $\hat{A}_g$, we have $\sigma^2_{\hat{A}_g} = \sigma^2_{A_g} + \sigma^2_{\nu_g}$, where $\sigma^2_{A_g}$ is the true variance of anchored achievement in grade g and $\sigma^2_{\nu_g}$ is the measurement error variance. Measurement error in this setting can

to how total wage earnings are constructed.

[10] Imposing a function form is only necessary because the CNLSY sample sizes in each grade are small relative to the number of items. Given enough data, these expectations could be estimated totally non-parametrically using simple sample averages for each possible realization of $(D_{i,g}, X_{i,g})$.

come from two possible sources: estimation error in $\hat{\mathbb{E}}[S_i|D_{i,g}, X_{i,g}]$ and mis-specification of the form of $\mathbb{E}[S_i|D_{i,g}, X_{i,g}]$. In the empirical work, I select quite flexible specifications for the form of this expectation and I additionally show that yet-more flexible specifications produce quantitatively similar estimates. Thus, the maintained assumption throughout the paper is that estimation error is the only source of measurement error ν.

Consider the infeasible regression

$$\hat{A}_{i,g} = \kappa_g + \gamma_g A_{i,g} + \epsilon_{i,g}. \tag{2}$$

The probability limit of $\hat{\gamma}_g$ from this regression is $R_g \equiv \sigma^2_{A_g}/(\sigma^2_{A_g} + \sigma^2_{\nu_g})$, the reliability of the estimated anchored scale. Because $(\sigma^2_{A_g} + \sigma^2_{\nu_g})$ is directly estimable from data, recovering an estimate of R_g would allow one to estimate $\sigma^2_{\nu_g}$. This regression is infeasible of course because $A_{i,g}$ is not observed. However, S_i is a noisy proxy for $A_{i,g}$: $S_i = A_{i,g} + \eta_{i,g}$, where $\eta_{i,g}$ is simply defined as the component of S_i on predictable from the anchor data (test items and demographics). Thus, consider the feasible regression

$$\hat{A}_{i,g} = \tilde{\kappa}_g + \tilde{\gamma}_g S_g + \tilde{\epsilon}_{i,g} \tag{3}$$

The probability limit of $\hat{\tilde{\gamma}}_g$ will be attenuated towards zero by the factor $\sigma^2_{A_g}/(\sigma^2_{A_g} + \sigma^2_{\eta_g})$. This errors-in-variables problem is solvable with an instrument for S_i – a variable $Z_{i,g}$ correlated with $A_{i,g}$ and uncorrelated with $\eta_{i,g}$.

Let $\hat{A}^{(1)}_{i,g}$ and $\hat{A}^{(2)}_{i,g}$ be the item-anchored scales estimated on the group-(1) and group-(2) items, respectively. Taking $\hat{A}^{(1)}_{i,g}$ as the anchored scores of interest, we want to run an IV regression to estimate equation (3). An instrument for S_i in this case can be constructed using the group-(2) anchored scale scores by taking the average S among test-takers who are not i but who have the same (or similar) values on the group-(2) anchored scale.[11] That is, the instrument $Z^{(1)}_{i,g}$ is defined by

$$Z^{(1)}_{i,g} = \frac{\sum_{j \neq i} S_{j,g} \times \mathbb{I}(\hat{A}^{(2)}_{j,g} = \hat{A}^{(2)}_{i,g})}{\sum_{j \neq i} \mathbb{I}(\hat{A}^{(2)}_{j,g} = \hat{A}^{(2)}_{i,g})}. \tag{4}$$

$Z_{i,g}$ satisfies the exogeneity requirement thanks to the leave-one-out construction. Relevance is satisfied because the $(\hat{A}^{(2)}_{j,g} = \hat{A}^{(2)}_{i,g})$ condition guarantees under the maintained assumptions that $A_{i,g} \approx A_{j,g}$. To summarize, the steps of the correction procedure are:

1. Estimate the item group-(1) and group-(2) anchored test scales.
2. Construct $\{Z^{(1)}_{i,g}\}$ according to equation (5).

[11] Naturally, the labels (1) and (2) are interchangeable here – once you have two distinct scales, either one can be used as the base and the other to construct the instruments.

3. Estimate equation (3), using $Z_{i,g}^{(1)}$ as an instrument for S_i, yielding $\hat{\tilde{\gamma}}_g^{(1)}$.

4. Estimate $\hat{\sigma}_{A_g}$ using $\sqrt{Var(\hat{A}_g^{(1)}) \times \hat{\tilde{\gamma}}_g^{(1)}}$.

5. Supposing further that $A_{i,g} \sim N(\bar{A}_g, \sigma_{A_g}^2)$ and $\nu_{i,g} \sim N(0, \sigma_{\nu_g}^2)$ and letting $\hat{q}_g^{(1)}(p)$ be the estimated p^{th} percentile of $\hat{A}_g^{(1)}$, estimate the corrected $p-(1-p)$ gap as[12]

$$\frac{\left(\hat{q}_g^{(1)}(p) - \hat{q}_g^{(1)}(1-p)\right)\sqrt{Var(\hat{A}_g^{(1)}) \times \hat{\tilde{\gamma}}_g^{(1)}}}{SD(\hat{A}_g^{(1)})}.$$

Given-Score Anchoring

Anchoring at the item-level is my preferred alternative to using psychometric scales in their native units for the two key reasons outlines in the Introduction: lack of cardinality and the (possibly) economically arbitrary aggregation of test items. Nonetheless, it is also possible to construct anchored scales using the given psychometric scores instead of the raw test items. In particular, I simply adopt directly the approach taken in Bond and Lang (2018). This method, which served as the inspiration for the item-anchoring approach, is quite similar.

1. Define achievement as $\theta_{i,g} = \mathbb{E}[S_{i,g}|T_{i,g}, X_{i,g}]$, where $T_{i,g}$ is a test score in its native (or linearly rescaled) units. Let $\hat{\theta}_{i,g}$ be the estimated achievement, calculated as the average outcome S_g of all individuals with observed test scores equal to (or close to, depending on the data) $T_{i,g}$.

2. Just as in the item-anchored case, the infeasible regression of $\hat{\theta}_{i,g}$ on $\theta_{i,g}$ would recover the reliability of $\hat{\theta}_{i,g}$ which is necessary to properly adjust the sample standard deviation of $\hat{\theta}_{i,g}$.

3. Instead of this infeasible regression, run an instrumental variables regression of $\hat{\theta}_{i,g}$ on $S_{i,g}$ using instruments $\xi_{i,g}$. These instruments are constructed similarly to the $Z_{i,g}$ in the item-anchored method, but instead of using one half of the current-grade items to construct the alternate anchored scale, one instead uses the given anchored scores for the observed grade immediately prior to g. Letting $l(g)$ denote this last observed grade (with $l(g) = g-2$ in most cases),

$$\xi_{i,g} = \frac{\sum_{j \neq i} S_{j,l(g)} \times \mathbb{I}(\hat{\theta}_{j,l(g)} = \hat{\theta}_{i,l(g)})}{\sum_{j \neq i} \mathbb{I}(\hat{\theta}_{j,l(g)} = \hat{\theta}_{i,l(g)})}. \tag{5}$$

[12] Under fairly general conditions, ν_g will be approximately normal when the form of $\mathbb{E}[S_i|D_{i,g}, X_{i,g}]$ is correctly specified. I will also assume normality to correct the estimated white-black achievement gaps in Section 6.

4.2 Implementation

Applying the method outlined in the previous section to the CNLSY data requires a number of decisions and adjustments. This section describes these implementation details.

Correcting for measurement error in the anchored scores requires the construction of two anchored tests scales using disjoint subsets of the test items. I create these groups simply by taking the even- and odd-numbered test items. Aside from its simplicity, this approach has some advantages stemming from the particular structure of the PIAT exam. Recall from Section 3 that the PIAT exam starts at an easy "basal" question that depends on the age of the child. The child then works up the test item list in order of increasing difficulty. The exam stops when the child starts getting most of the questions wrong (see footnote 7 for details). This exam structure implies that test items that have similar numbers also have similar difficulty. Thus, dividing the test items by whether their items numbers are even or odd ensures that the items used for each anchored scale run the gamut of difficulty.

Depending on the age of their mothers, children in the CNLSY are observed in a range of survey years for a given grade. In order to account for cohort effects and for demographic characteristics in a consistent way across different model-selection techniques, I first residualize the outcomes S using the survey year interacted with the grade group and race/-sex indicators and their interactions interacted with the grade group. I then estimate the anchored scales by relating the residuals of this first stage to the vectors of item indicators.

The staggered entry of children into the CNLSY sample means that outcomes are sometimes not observable for a given individual simply because they are not old enough. Thus, I estimate each anchor model on the subset of the full sample for whom the relevant outcome is at least potentially observable.[13]

The method requires the selection of a particular functional form for $\mathbb{E}[S_i|D_{i,g}, X_{i,g}]$. I select a flexible, linear functional form so as to avoid imposing more structure on this expectation than necessary. In the interest of sample size, I pool the data into pairs of grades: pre-k/k, grades 1-2, grades 3-4, grades 5-6, and grades 7-8. I find similar results when I estimate these models separately by grade, but in some cases the estimates are more volatile, particularly the reliability estimates coming through the IV procedure. Finally, I suppose that the (residualized) outcome is a linear function of the dichotomous item responses interacted with the test-taker's grade. That is, letting $S_{i,p}^{(r)}$ denote the residualized outcome for student i in grade group p consisting of grades g and $g+1$, I suppose for item group (1)

$$\tilde{S}_{i,p}^{(r)} = D_{i,p}^{(1)} W_g \mathbb{I}(i \in g) + D_{i,p}^{(1)} W_{g+1} \mathbb{I}(i \in g+1). \tag{6}$$

[13] I define an observation as "highest grade completed feasible" and "college feasible" if the reported year of birth is earlier than 1993. The log wage at 25 and log income at 25 feasibility conditions are obvious – the individual needs to be at least 25 in the most recent CNLSY wave.

I construct the instruments $Z_{i,g}^{(1)}$ using $j \neq i$ where $\hat{A}_{j,g}^{(2)}$ is in the same 2-percentile bin (50 total) as i.

The number of parameters in this model is quite large, and thus I employ lasso regression with the penalty parameter selected via cross-validation to select which covariates enter the final anchor relationship. Other approaches such as elastic net regression and even ordinary least squares regression yield very similar estimates. I also experiment with estimating more flexible specifications. In particular, I modify equation (6) to allow for all possible two-way interactions between items. Such interactions might occur if, for example, an economic outcome requires each of a number of different skills assessed by different test items. Empirically, however, I find very little difference in either the anchored scales or in the estimated variances from these richer specifications.

Given-Score Anchoring

To implement the given-score anchoring scheme outlined in the previous section, I make two approximations. First, I anchor the age-standardized piat math, reading, and combined achievement scores divided into percentile buckets. That is, I divide the given achievement distribution for grade g into 100 equally-sized bins. The anchored scale for i in grade g is then defined as the average S_g for all j in the same percentile bucket as i. Second, as in the item-anchored scheme described above, I construct the instruments $\xi_{i,g}$ using $j \neq i$ where $\hat{\theta}_{j,g-1}$ is in the same 2-percentile bin (50 total) as i.

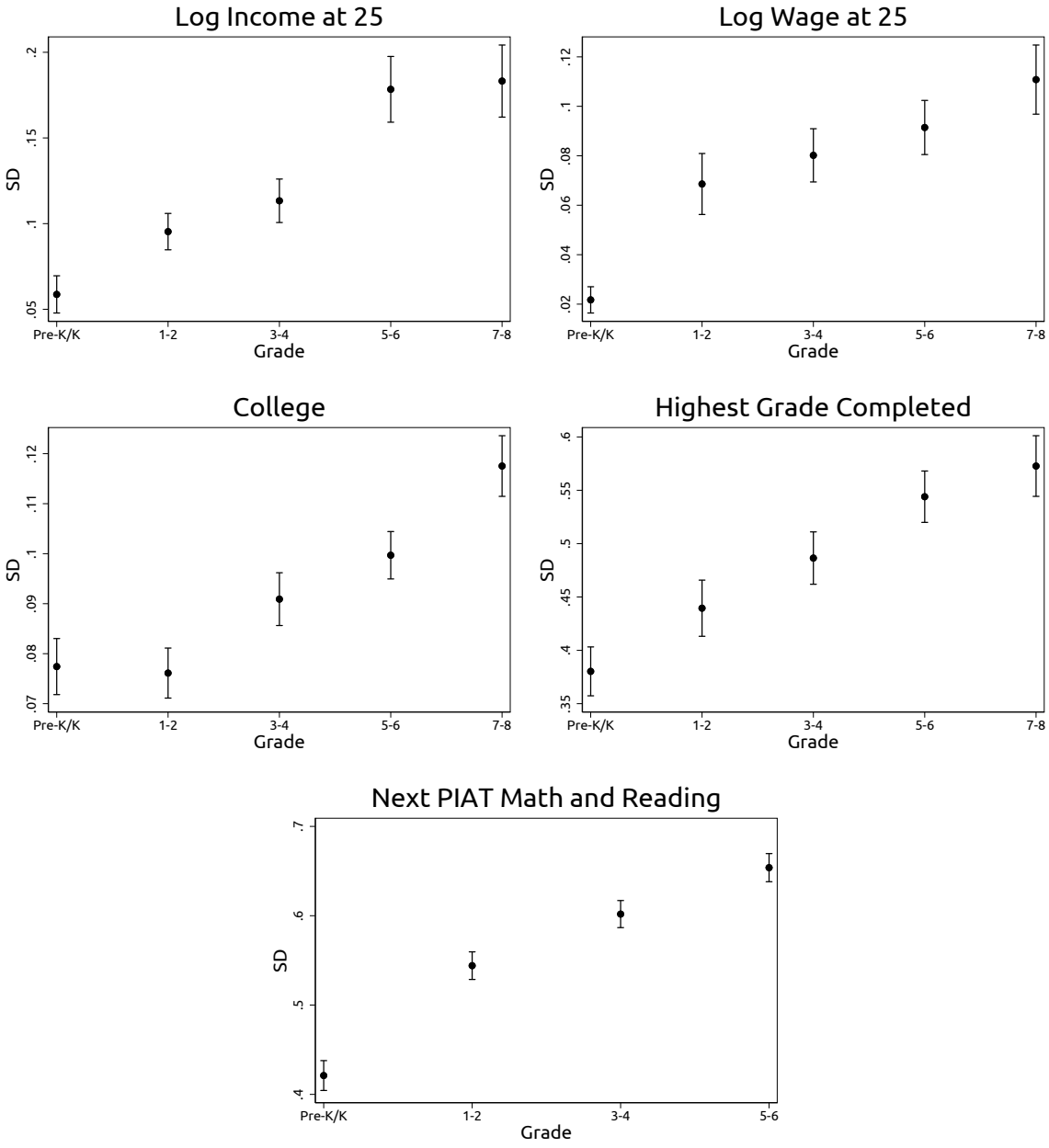
5 The Spread of Achievement Through Grade 8

I now present evidence that the spread in item-anchored achievement, measured using the standard deviation as well as the gap between the 90^{th} versus 10^{th} percentiles, increases notably between pre-kindergarten and eighth grade.

Figure 1 shows the evolution of item-anchored scales using the math and reading items combined for a range of different outcomes. The top left panel shows that the standard deviation of scales anchored to log income at age 25 roughly triples from around 0.06 log points to 0.18 log points. Moreover, the panel suggests some non-linearity, with comparatively modest increases in the standard deviation through grades 1/2 followed by sharp increases through grades 5/6 and a more gradual increase from grades 5/6 through grades 7/8. The top right panel shows that the standard deviations of the scales anchored at the item level to log wages at age 25 likewise increase notably from around 0.02 log points to 0.11 log points, with more rapid increases at the youngest and oldest grades. The college-anchored standard deviations, shown in the middle left panel, also display a non-linear pattern. Indeed, the estimated standard deviation actually decreases between pre-k/k and grades 1/2, albeit not by a statistically significant amount, before increasing sharply and statistically significantly

thereafter. The standard deviations of the scales anchored to the highest grade completed, shown in the middle right panel, increase roughly linearly through grades 5/6, with a smaller increase between grades 5/6 and grades 7/8. The net change in standard deviation for both of these school completion anchors is quite large: +52% for college completion and +51% for the highest grade completed.

Figure 1: The Standard Deviation of Item-Anchored Math and Reading Achievement



Note: Estimated standard deviations from anchor models following the form in equation (6). 95% confidence intervals based on 1,000 bootstrap iterations holding the anchored scales using the even and odd items fixed.

In contrast to the long-run, non-psychometric outcomes discussed so far, the bottom

panel of Figure 1 anchors to the age-standardized, combined PIAT math and reading scores from the next survey wave, a gap of two years.[14] The idea is to anchor to an outcome that may be of direct interest to educators or to social scientists unconvinced by my arguments against the interpretability of psychometric scales in their native units. Indeed, one of the primary uses of standardized achievement test scores is to mark students' progress through school. As with the more temporally distant, economic anchors shown in the other panels, the next PIAT item-anchored standard deviations increase from just above 0.4 in pre-k/k to around 0.65 in grades 5/6, a 55% increase. There is again evidence of non-linearity, with a much larger increase between pre-k/k and grades 1/2 than between the higher grades, which show a roughly linear trend.

Figure 4 in the Appendix repeats the analysis using just the PIAT math items. Overall, these results are very similar to the combined results in Figure 1. The most notable difference is that the next PIAT item anchored standard deviations show a less dramatic, though still statistically significant, increase from about 0.41 to 0.49.

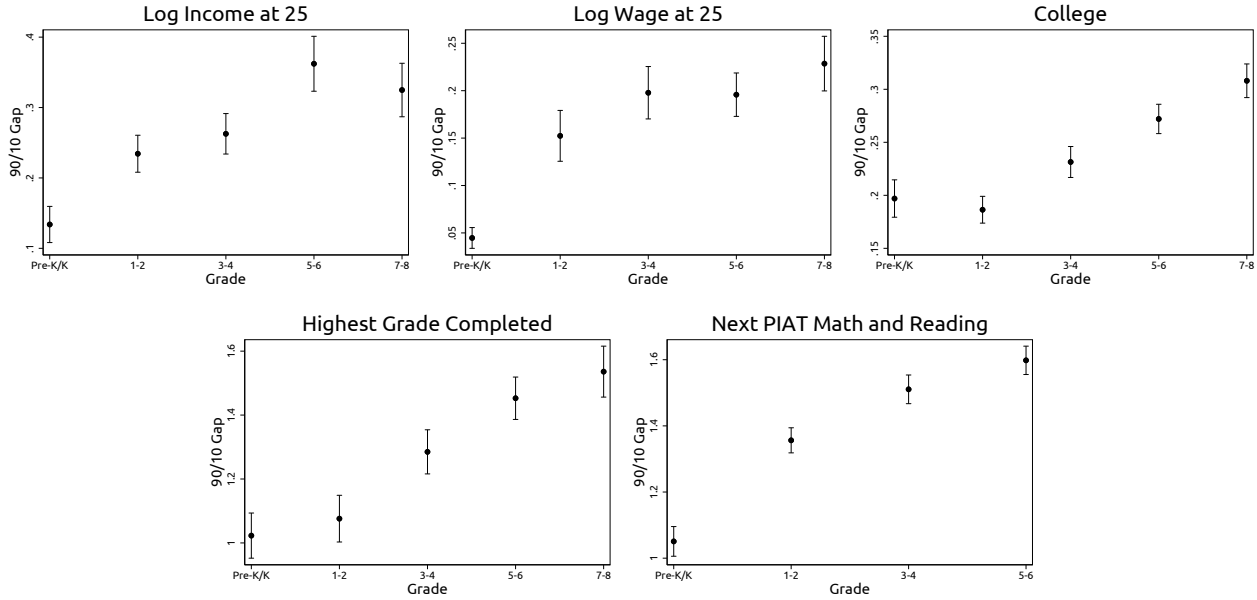
The standard deviation results provide strong evidence that the spread in achievement increases as students progress through school. Additionally, these results speak directly to the common practice of standardizing test scores to have a unit standard deviation in each grade. However, the standard deviation does not clearly show what is going on in the extremes of the achievement distribution. Therefore, Figure 2 presents the estimated, measurement-error-corrected differences between the 90^{th} and 10^{th} percentiles of the various item-anchored test scales. As with the standard deviation estimates in Figure 5, the 90/10 gaps show large and generally quite steady increases between pre-k/k and grades 7/8. The only case where steady increases are not apparent is between grades 5/6 and 7/8 when log labor income or wage rates at 25 are used as the anchor – the estimated change for both outcomes is slightly negative although not statistically significant. Nonetheless, the net increases in the 90/10 gaps are all enormously positive, with the percentage increases ranging between 50% to nearly 400%.[15]

These spread estimates depend on two estimated components: the unadjusted distribution of item-anchored scores and the estimated reliability of the item-anchored scales. The smaller is the estimated reliability, the smaller will be the corrected estimate relative to the raw estimate. Thus, an increase in the estimated spread during childhood could come from a combination of two sources: increasing raw spreads and/or increasing estimated reliabilities.

[14] Because of this two year gap, I do not show anchored estimates for the grade 7/8 group, because very few observations in these grades have subsequent PIAT scores.

[15] Figure 5 in the Appendix shows similar results for the gap between the 99^{th} and 1^{st} percentiles. The 99/1 log income at 25 gap shows increases between all grades, although the college gap does not.

Figure 2: The 90/10 Gap in Item-Anchored Math and Reading Achievement



Note: Estimated differences between the 90th and 10th percentiles of item-anchored score distributions from anchor models following the form in equation (6). 95% confidence intervals based on 1,000 bootstrap iterations holding the anchored scales using the even and odd items fixed.

Figure 6 in the Appendix shows that the estimated reliabilities do tend to be higher in later grades, although this is not uniformly the case. For instance, while the log income at 25 reliabilities display a clear upward trend, the highest grade completed reliabilities do not. Higher scale reliabilities at older ages is intuitive and is consistent with psychometric findings in other settings. Nonetheless these estimated reliability differences are too small in magnitude to account for most of the overall spread increases reported in Figures 1 and 2. For instance, the log income at 25 reliability increases from about 0.55 to 0.81 from pre-k to eighth grade, implying that the reliability adjustment factor increases from about 0.74 to 0.9.[16] Were the unadjusted standard deviations constant, the reliability adjustment alone would then yield an estimated adjusted increase of 21%, which is much less than the roughly 200% adjusted increase reported in Figure 1. The headline result that the spread in achievement increases substantially during childhood does not depend in a quantitatively important way on the reliability adjustment.

Finally, Figure 3 presents the standard deviations of the given-anchored scores, estimated following the approach in Bond and Lang (2018). Unlike the item-anchored scales, the given-anchored scales present less clear evidence of standard deviation increases through childhood.[17] Indeed, the log income at age 25 results show no clear pattern for grades 3/4

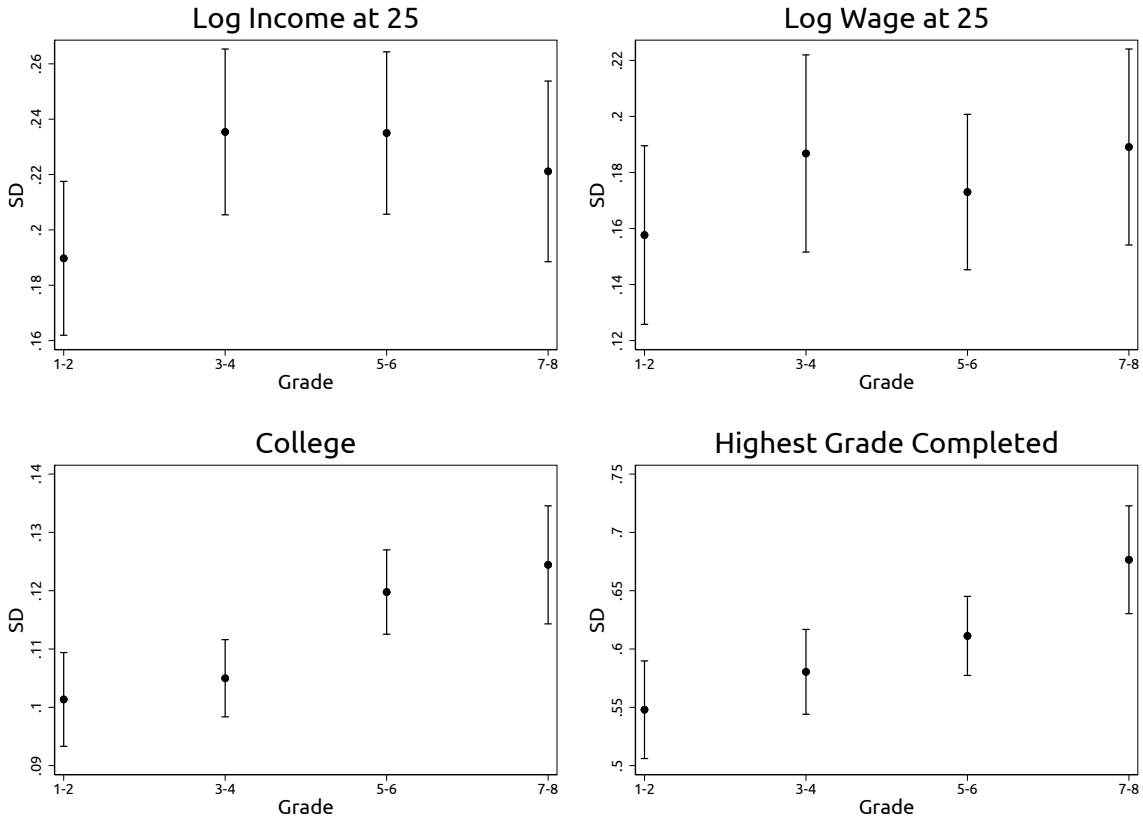
[16]Recall that the adjustment factor is the square root of the reliability estimate.

[17]Because the measurement error adjustment in this case uses the prior-wave given anchored scales to construct the instrument, I can only compute measurement-error corrected standard deviations for grades

and above and only modest, weak evidence of a lower standard deviation in grades 1/2. The college and highest grade completed estimates, by contrast, do show clear increases. While these school-completion point estimates are generally similar to their item-anchored counterparts, the standard deviations are uniformly less precisely estimated.

A comparison of Figure 3 with Figures 1 highlights the importance of considering the aggregation of test items in the construction of achievement scales. In prior work, I showed that item- versus given-anchored scales can disagree on mean achievement gaps (Nielsen, 2019) and male-female variance comparisons (Nielsen, 2022). The results presented here show that these two anchoring methods can also disagree on the evolution of the spread in achievement through childhood.

Figure 3: The Standard Deviation of Given-Anchored Achievement



Note: Estimated standard deviations from anchor models using given scores, following the method in Bond and Lang (2018) and outlined in Section 4. 95% confidence intervals based on 1,000 bootstrap iterations holding the anchored scales using the even and odd items fixed.

1/2 - 7/8.

6 Standardization and Achievement Gap Trends

Section 5 shows that the variance in item-anchored achievement increases dramatically between pre-k and eighth grade. In this section, I demonstrate that these increases have important consequences for the measurement of trends in achievement inequality.

Table 2 shows the net changes in the mean item-anchored white - black achievement gaps between pre-k/k and eighth grade (or sixth grade in the case of the "next observed PIAT" anchor). For each set of items (math and reading combined or just math alone) and each anchor, two gap-change estimates are shown: those using the item-anchored scores in their native "anchor" units and those using item-anchored scores that have been standardized to have a unit variance separately by grade.

As pointed out in Bond and Lang (2018) and Nielsen (2019), measurement error biases towards zero raw mean differences in anchored scores. I thus follow these papers by adjusting the raw mean differences by the grade-specific reliabilities estimated as in Section 4. In particular, assuming normality, I estimate the white-black item-anchored achievement gap in grade g as

$$\frac{1}{\hat{\bar{\gamma}}_g^{(1)}} \times \left(\frac{\sum_i \hat{A}_{i,g}^{(1)} \mathbb{I}(i \in \text{white})}{\sum_i \mathbb{I}(i \in \text{white})} - \frac{\sum_i \hat{A}_{i,g}^{(1)} \mathbb{I}(i \in \text{black})}{\sum_i \mathbb{I}(i \in \text{black})} \right). \tag{7}$$

The unit variance estimates correspond conceptually to the typical presentation in prior literature (e.g. Fryer Jr and Levitt (2004)) others). However, because the true variance of item-anchored achievement is larger in later grades, this by-grade standardization will shade toward zero mean gaps in later grades comparatively more than in earlier grades, thus biasing down the estimated trend in the achievement gap.

This downward bias is evident in Table 2.[18] The anchor-unit gap changes are uniformly positive and are highly statistically significant in almost all cases. These estimates imply that white/black achievement inequality increases very substantially during childhood. By contrast, the sd-unit gap changes are only sometimes statistically distinguishable from zero and always represent much smaller percentage changes from the pre-k/k baseline than the corresponding anchor-unit changes. Moreover, some of the sd-unit estimates are actually negative, indicating a decrease in the white/black achievement inequality.

In addition to the methodological point regarding the bias introduced by standardizing already-cardinal test scores, the results in Table 2 are substantively interesting in their own right. White/black achievement inequality is notably greater at older grades/ages for

[18] See also Figure 7 in the Appendix. The changes in male-female item anchored achievement inequality, not shown for brevity, similarly show smaller-magnitude changes when scores are standardized separately by grade. The item-anchored scales generally find that women have higher mean achievement than men in pre-k/k and that this advantage erodes steadily in higher grades, typically vanishing completely by eighth grade.

every anchor outcome. For example, the item-predicted white-black log earnings at 25 gap increases by roughly 0.09 between pre-k/k and grades 7/8, while the item-predicted college completion gap grows by 0.06 probability units. These estimates thus suggest both that the cumulative disparities in endowments and investments lead to large achievement gaps by the start of school (i.e. the pre-k/k gaps are already very large in outcome units) and that these disparities continue to grow rapidly during the first 8-9 years of formal schooling. While these estimates do not provide guidance as to which social processes are generating these widening achievement gaps, they do suggest that there may be substantial scope to arrest/reduce racial achievement disparities in elementary and middle school.

Table 2: Changes in the White-Black Item-Anchored Achievement Gaps

Math & Reading	Log Income at 25 8^{th}-Pre-K	Log Wage at 25 8^{th}-Pre-K	College 8^{th}-Pre-K	Highest Grade 8^{th}-Pre-K	Next PIAT 6^{th}-Pre-K
Anchor	0.089***	0.089***	0.061***	0.272***	0.292***
	(0.021)	(0.026)	(0.013)	(0.049)	(0.031)
Change	131%	869%	69%	78%	110%
By-Grade SD	-0.301	0.781*	0.130	0.166	0.222***
	(0.328)	(0.467)	(0.166)	(0.131)	(0.066)
Change	-26%	90%	11%	18%	35%
Math					
Anchor	0.033	0.101***	0.016	0.150***	0.077***
	(0.031)	(0.035)	(0.016)	(0.043)	(0.029)
Change	25%	103%	15%	33%	18%
By-Grade SD	-0.771	-0.202	-0.211	−0.358***	-0.000
	(0.547)	(0.745)	(0.230)	(0.134)	(0.075)
Change	-43%	-10%	-14%	-25%	-0%

Note: The "Anchor" estimates report the net change in the item-anchored scales over the grades indicated, while the "By-Grade-SD" estimates report the net changes using scores that have been standardized to have a unit variance by grade. Bootstrapped standard errors based on 1,000 bootstrap iterations shown in parentheses. The "Change" estimates report the estimated gap changes as a percentage of the pre-k/k gaps. The math & reading log wage at 25 anchored gap change percentage increase is very large because the pre-k/k gap, at around 0.01, is quite close to 0. * reflects 0.1 significance, ** reflects 0.05 significance, and *** reflects 0.01 significance.

7 Standardization and Causal Effects

Most papers estimating causal effects on test scores do so for test scores reported in by-grade or age sd units. However, if effects are estimated on children with widely varying ages, or if effects are estimated separately on younger and older children and then compared, the common use of sd-unit scores may not be benign. If the true variance in achievement is

greater at older ages, then the standardized effects will systematically understate effect sizes at older ages relative to younger ages.

This section demonstrates using estimates from a well-cited and well-executed paper that these concerns are not merely theoretical. I consider estimated age heterogeneity from Dahl and Lochner (2012), an influential paper that estimates the causal impact of household income on children's test scores using the CNLSY. That paper finds significant effects of income on test scores, with some evidence that the effects are larger for younger children.

I use the replication files from Dahl and Lochner (2012) to replicate their sample and method, which identifies the causal effects of household income on children's test scores via an instrumental variables strategy.[19] The instrument leverages expansions in the earned income tax credit (EITC) to construct a plausibly exogenous source of variation in household income.[20] The outcomes studied in Dahl and Lochner (2012) are standardized PIAT math and reading measures scores. Because Dahl and Lochner (2012) uses the same data as I do, it is then straightforward to swap into the replication files the various item-anchored achievement measures inplace of the standardized PIAT measures. That is, I can keep everything about the analysis constant except the way that the PIAT data are used to construct achievement measures.

Table 3 presents math estimates based on the method from Dahl and Lochner (2012) for two samples: children less than 12 years old and children 12 years old or older.[21] Columns (2) and (4) present results in standard deviation units, where the standardization is carried out separately by grade. Columns (1) and (3), by contrast, present results for various item-anchored scales in their native "anchor" units. The sd-unit results tell largely the same story whether the outcomes are the PIAT scores (used in Dahl and Lochner (2012)) or the item-anchored scores. The estimated impact of $1,000 of additional family income is around 0.03-0.1 sd in achievement, with the effects typically larger for the under-12 sample, although

[19]The only difference is that I do not have fine geographic information available in my sample, as this requires one to apply for secure data from the NLS.

[20]In detail, using Dahl and Lochner (2012)'s notation, the outcome equation estimated is

$$\Delta y_{ia} = x_i'\alpha + \Delta w_{ia}'\beta + \Delta I_{ia}\delta_0 + \Phi(P_{i,a-1}) + \eta_{ia},$$

where Δy_{ia} is student i's achievement change at age a, I_{ia} is i's family income, $\Phi(P_{i,a-1})$ is a flexible function of lagged pre-tax family income, and x_i and w_{ia} are fixed and time-varying student characteristics. The causal effect of interest, δ_0, is then estimated using predicted changes in EITC income as a function of lagged pre-tax income as an instrument. Letting $\chi_a^s(P)$ denote the amount of EITC income accruing to a youth whose family is on EITC schedule s (which can vary with family structure) with pre-tax family income P, the instrument is given by

$$\Delta\chi_a^{IV}(P_{i,a-1}) \equiv \chi_a^{s_{i,a-1}}(\hat{\mathbb{E}}[P_{i,a}|P_{i,a-1}]) - \chi_{a-1}^{s_{i,a-1}}(P_{i,a-1}).$$

[21]Data difficulties not withstanding (see Section 3), the results using reading items, not shown, tell qualitatively the same story.

the differences by age are mostly not statistically significant.

The non-standardized results in columns (1) and (3) paint quite a different picture than the standardized results. The item-anchored results are small and never statistically significant for the under-12 subsample. By contrast, the corresponding results for the 12-and-older subsample are often much larger and are usually statistically significant. For instance, the college-anchored math scale estimate for the 12+ group, at 0.027, is nearly three and half times larger than the corresponding estimate for the under-12 group. The log(income) and log(wage) anchored estimates are likewise more than three times larger for older students, while the next PIAT estimates are 27% larger. Only the college item-anchored math estimates are comparable in magnitude across the two age groups, although only the older group's estimate is statistically significant.

Table 3: The Effect of Income on Math Achievement in Younger and Older Children

	< 12 years		≥ 12 years	
	(1)	(2)	(3)	(4)
Units	anchor	sd	anchor	sd
PIAT (Dahl and Lochner (2012))		0.107		0.037
		(0.073)		(0.028)
College	0.008	0.121	0.027*	0.079*
	(0.008)	(0.087)	(0.014)	(0.04)
High School	0.002	0.099	0.024*	0.078
	(0.006)	(0.088)	(0.014)	(0.05)
Highest Grade Completed	0.055	0.144	0.055**	0.096**
	(0.039)	(0.099)	(0.026)	(0.044)
Log(income)	0.008	0.099	0.034*	0.116**
	(0.01)	(0.091)	(0.018)	(0.057)
Log(wage)	0.01	0.164	0.034**	0.112**
	(0.009)	(0.115)	(0.016)	(0.055)
Next PIAT $(t+2)$	0.038	0.077	0.048*	0.065*
	(0.035)	(0.07)	(0.026)	(0.035)

Note: All estimates use the specification from Table 6 of Dahl and Lochner (2012) but do not use the confidential geocoded data from that paper. * reflects 0.1 significance and ** reflects 0.05 significance.

The item-anchored estimates imply that income does in fact have a significant effect on childhood skills, but only for older children. These estimates are all economically interpretable because the item-anchored scales are. They imply that an additional $1,000 of household income for kids aged 12 and over increases predicted college completion by about 0.03, highest grade completed by 0.06 years, income and wages at age 25 by about 3.5% each, and PIAT scores two years in the future by 0.05 sd.

Taken together, the results in Table 3 highlight the importance of item-anchoring and

the importance of not standardizing scores. Without anchoring, there is no direct way to assess whether the test score effects identified in Dahl and Lochner (2012) are economically meaningful. Moreover, standardizing completely reverses the true age heterogeneity in the effects – the standardized scores suggest modestly larger effects for younger children, while the cardinally interpretable item-anchored effects are much larger for older children.

8 Discussion and Conclusion

In this paper, I showed that the variances of economically motivated, cardinal measures of achievement increase dramatically between pre-kindergarten and eighth grade. I constructed such measures by estimating models that relate the full vector of item responses to long-run, economic outcomes. The standard deviation of these item-anchored scales increase 50-400% between pre-k and eighth grade. I then showed that these increases have important consequences both for the estimation of trends in achievement quality and in the estimation of heterogeneous treatment effects by child grade/age.

These results build on prior evidence demonstrating that the naive use of psychometrically-derived achievement measures in economic applications might yield misleading or biased results. The standard practice of reporting results in sd units may not be benign, particularly if the data involve students of very different ages.

These results suggest that, where possible, researchers should either not pool results across ages/grades or should do so only for achievement measures that are in the same cardinally interpretable units. An interesting question for future research would be to assess which types of skills are generating the widening variance documented here. Answering this question would require more information on the content of the items than is available in the CNLSY data used in this study. Additionally, it would be interesting to investigate concurrently the functional form of the best-fit item-anchoring models. That is, do certain types of skills seems to interact? Answering such questions would likely require substantially larger sample sizes than are available in the CNLSY.

References

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., and Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest*, 21(2):55–97.

Ballou, D. (2009). Test scaling and value-added measurement. *Education finance and Policy*, 4(4):351–383.

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.

Bond, T. and Lang, K. (2018). The Black–White Education Scaled Test-Score Gap in Grades K-7. *Journal of Human Resources*, 53(4):891–917.

Cascio, E. U. and Staiger, D. O. (2012). Knowledge, tests, and fadeout in educational interventions. Technical report, National Bureau of Economic Research.

Cawley, J., Conneely, K., Heckman, J., and Vytlacil, E. (1997). *Cognitive Ability, Wages, and Meritocracy*, pages 179–192. Springer New York, New York, NY.

Cawley, J., Heckman, J., and Vytlacil, E. (1999). On policies to reward the value added by educators. *The Review of Economics and Statistics*, 81(4):720–727.

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.

Cunha, F., Heckman, J. J., and Schennach, S. (2010). Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica*, 78:883–931.

Cunha, F., Nielsen, E., and Williams, B. (2021). The econometrics of early childhood human capital and investments. *Annual Review of Economics*, 13(1):487–513.

Dahl, G. B. and Lochner, L. (2012). The impact of family income on child achievement: Evidence from the earned income tax credit. *American Economic Review*, 102(5):1927–1956.

Domicolo, C. and Nielsen, E. (2022). Male–female achievement variance comparisons are not robust. *Economics Letters*, 220:110853.

Fryer Jr, R. G. and Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *Review of economics and statistics*, 86(2):447–464.

Hill, C. J., Bloom, H. S., Black, A. R., and Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child development perspectives*, 2(3):172–177.

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.

Lang, K. (2010). Measurement matters: Perspectives on education policy from an economist and school board member. *Journal of Economic Perspectives*, 24(3):167–182.

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

Nielsen, E. (2019). Test Questions, Economic Outcomes, and Inequality. *Finance and Economics Discussion Series 2019-013, Federal Reserve Board*.

Nielsen, E. (2022). Is the greater variability in achievement for males a psychometric artifact? *Working Paper*.

Nielsen, E. (2023). The income-achievement gap and adult outcome inequality. *Journal of Human Resources*.

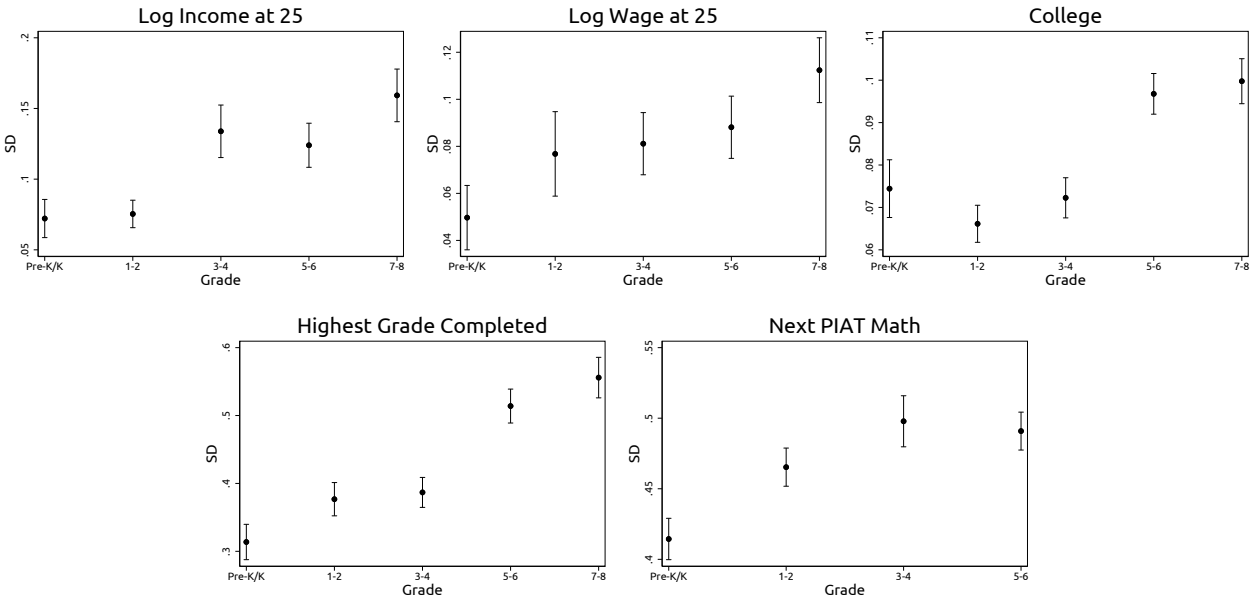
Schroeder, C. and Yitzhaki, S. (2017). Revisiting the Evidence for Cardinal Treatment of Ordinal Variables. *European Economic Review*, 92:337 – 358.

Stevens, S. (1946). On the Theory of Scales of Measurement. *Science*, 103:677–680.

Wan, S., Bond, T. N., Lang, K., Clements, D. H., Sarama, J., and Bailey, D. H. (2021). Is intervention fadeout a scaling artefact? *Economics of Education Review*, 82:102090.

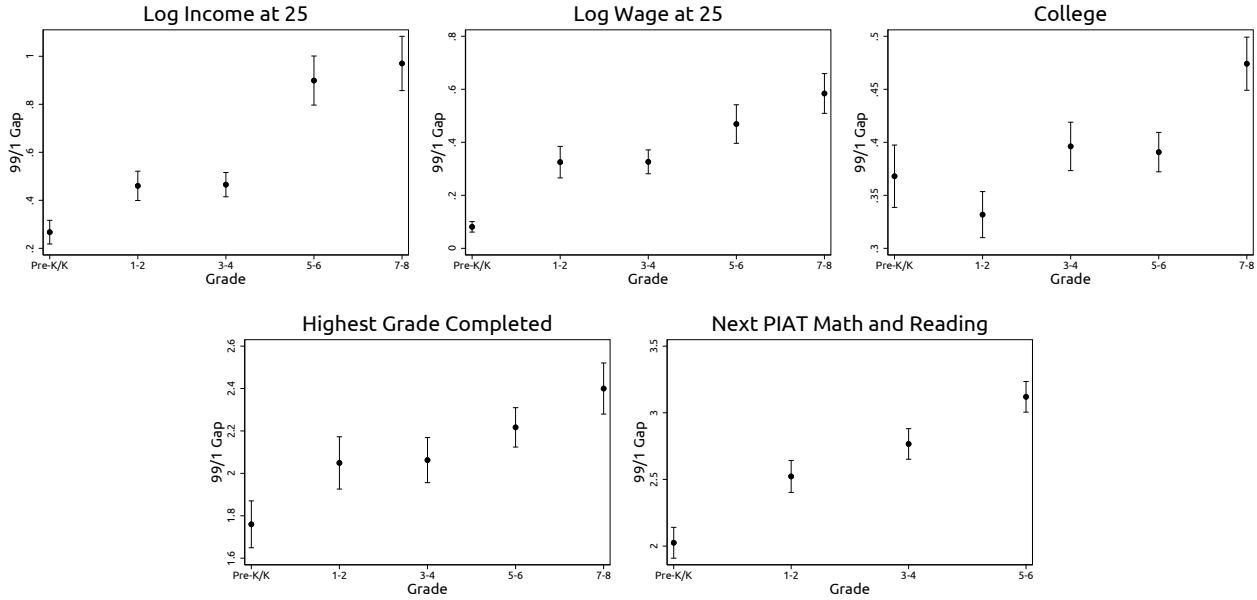
A Supplementary Results

Figure 4: The Standard Deviation of Item-Anchored Math Achievement



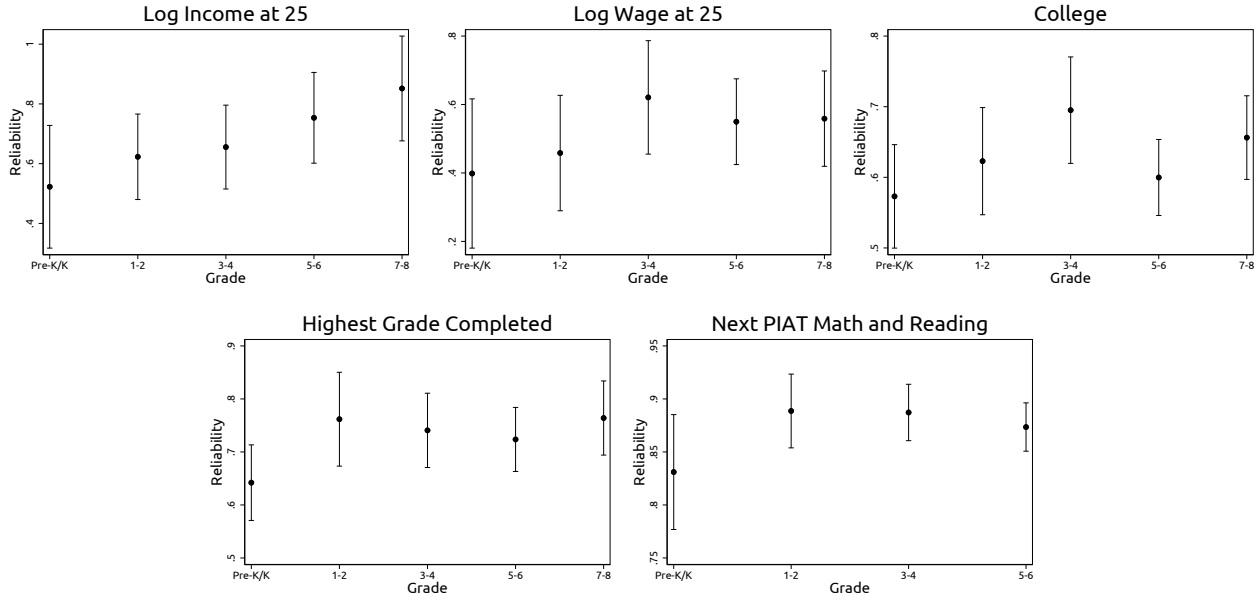
Note: Estimated standard deviations from anchor models following the form in equation (6). 95% confidence intervals based on 1,000 bootstrap iterations holding the anchored scales using the even and odd items fixed.

Figure 5: The 99/1 Gap in Item-Anchored Reading and Math Achievement



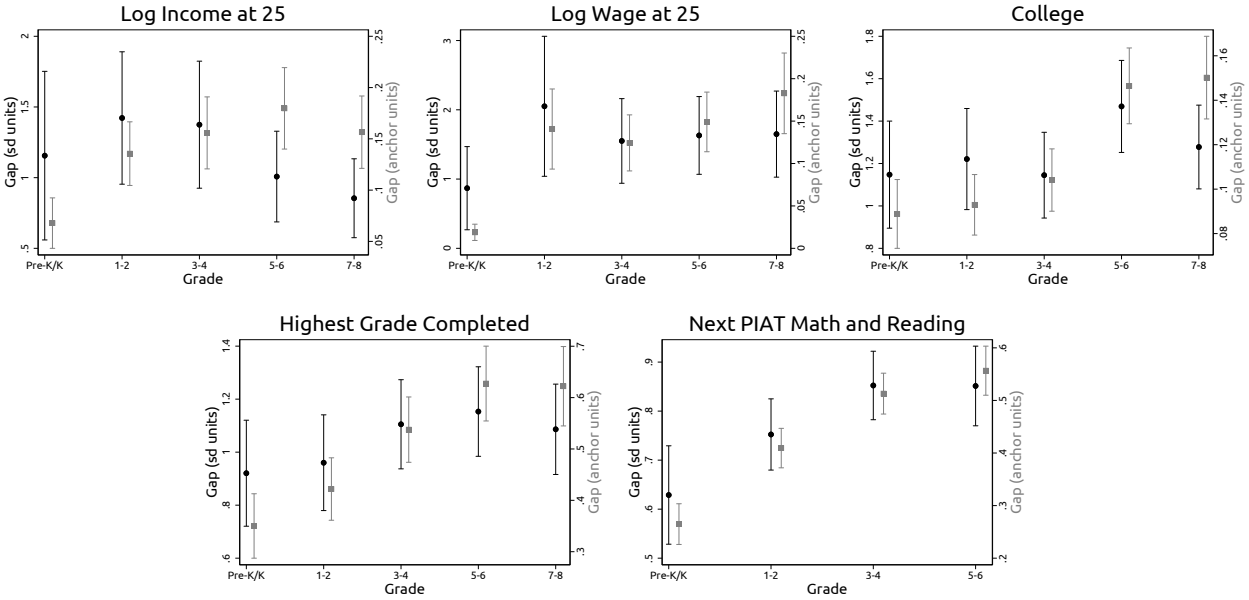
Note: Estimated differences between the 99$^{\text{th}}$ and 1$^{\text{st}}$ percentiles of item-anchored score distributions from anchor models following the form in equation (6). 95% confidence intervals based on 1,000 bootstrap iterations holding the anchored scales using the even and odd items fixed.

Figure 6: The Reliability of Item-Anchored Reading and Math Achievement



Note: Reliabilities estimated using the approach described in 4 using anchor models following the form in equation (6). 95% confidence intervals based on 1,000 bootstrap iterations holding the anchored scales using the even and odd items fixed.

Figure 7: Math & Reading Item-Anchored Black-White Achievement Gaps



Note: The left-axis scale (black) shows the mean white-black achievement gaps using item-anchored scores standardized to have a unit variance by grade. The right-axis scale (gray) shows the same gaps using non-standardized scores in anchor (outcome) units. 95% confidence intervals based on 1,000 bootstrap iterations holding the anchored scales using the even and odd items fixed. Estimates are based on anchor models of the form in equation (6) and adjust for measurement error using the method in Nielsen (2019).