The Variance of Achievement Increases During Childhood

Eric Nielsen

Federal Reserve Board

Disclaimer: The views and opinions expressed in this presentation are solely those of the authors and should not be interpreted as reflecting the official policy or position of the Board of Governors or the Federal Reserve System.
Nathan Ausubel and Victoria Chbane provided outstanding research assistance.

Motivation I: Item-Anchoring

Psychometric scales are not cardinal *for economic applications.*

- economic outcomes are nonlinear in test scores
- cardinality necessary to meaningfully compare means, variances

Anchoring given scores is an imperfect solution.

- $A = \mathbb{E}[S|T]$ for some cardinal outcome S and test score T
- rescale scores to be cardinally interpretable

Problem: how and why are the given scales constructed?

- by educators, psychometricians
- mastery of a curriculum, etc.

Scores not designed to serve as a proxy for human capital!

Motivation I: Item Anchoring

Any test score is based on some method of aggregating questions.

- M binary questions $\rightarrow 2^M$ possible vectors of item responses
- many, many choices about how to aggregate
- IRT, % correct, etc. represent particular choices

Psychometric scores aggregate items in a *non-economic* way.

- items not weighted by economic usefulness
- test designers have different objectives

Solution: anchor at the item level.

- $A = \mathbb{E}[S|D]$ where $D =$ full vector of item responses

Motivation I: Item Anchoring – A Simple Example

A test is 1/2 trig items and 1/2 statistics items, equally weighted.

- stats is useful in the labor market, trig is not[1]

"Eric" knows trig but not stats. "Jesse" knows stats but not trig.[2]

- Jesse and Eric score equally on the exam
- Jesse earns more than Eric

Standard analysis or given-score anchoring:

- Earnings difference unexplained by achievement

But this is just because we are measuring achievement strangely.

- downweight trig, upweight stats $\Longrightarrow$
- Jesse scores higher than Eric and out earns him

[1] Sorry Mrs. Hamilton...

[2] Names chosen randomly.

Motivation I: Item Anchoring Matters

Nielsen (2019)

- white-black gaps in wage- and lifetime income-anchored achievement equal observed gaps
- white-black differences in employment predictable from items
- item-anchored scores resolve the "reading puzzle"

Nielsen (2023)

- Males do not consistently have greater variance on item-anchored test scales

Bruhn et al. (2023)

- teacher vam, fade out, variation in student achievement
- presentation in about 30 minutes...

Motivation II: Are SD-Unit Scores Meaningful?

Achievement tests commonly scaled to have a unit variance by grade/age.

- achievement gaps, causal effects, etc. reported in "sd units"

The variance of economically-relevant skills may not be constant across grades/ages.

- the range of skills/tasks at older ages is much greater than at younger ages
- sd-units may not have a fixed meaning

Achievement gaps, causal effects, etc. reported in sd-units might erroneously mask or create heterogeneities by age.

- point also applies to percentile-unit scores

This Paper: Achievement Variance at Different Ages

1. Cardinal achievement measures via item-anchoring.
   - aggregation based on item-outcome relationships
   - split-half IV correction for measurement error
   - lasso to handle large number of items

2. Variance in achievement by grade, pre-k through 8^{th} grade.
   - 90/10 and 99/1 gaps as well

3. Assess importance of standardization for:
   - the evolution of the white-black achievement gap
   - the causal effect of income on achievement

Preview of Results

The standard deviation of achievement increases *a lot* during childhood.

- 50% to 400% depending on the anchor

This result depends on the use of item-level data.

- given-anchoring yields smaller/null increases in variance

By-grade standardization totally obscures:

- large increases in white-black achievement inequality
- larger causal effects of income on *older* children

Standardization and ignoring item-level data are not innocuous.

Contributions to Several Literatures

Non-interpretable variance of achievement – Lang (2010); Cascio and Staiger (2012); Stevens (1946), Nielsen (2023b)

Test scores and cardinality – Bond and Lang (2013); Lord (1975); Nielsen (2023a); Domicolo and Nielsen (2022); Cawley et al. (1999); Bettinger et al. (2013)...

Anchoring – Nielsen (2019); Bond and Lang (2018); Heckman, Cunha, Schennach (2010); Polacheck et al. (2015)...

Intervention fade-out – Bailey et al. (2020); Wan et al. (2021); Hill et al. (2008)

Any literature that uses sd-unit test scores.

CNLSY Item-Level Data

PIAT math and reading exams.

- age 5-14 respondents in every CNLSY wave
- item content fixed across survey waves
- 84 math, 84 reading items

Items asked in order of increasing difficulty.

- "basal" item depends on age and several "trial" questions
- exam stops when most items answered incorrectly

Fill-in rule:

- items below basal = correct
- items above final question = incorrect

Conceptual Framework and Method

Individuals i in grade g take a test with binary items indexed by j

- $d_{i,j,g} = 1$ if i gets j correct, 0 o.w.
- $D_{i,g} = [d_{i,1,g}, \ldots, d_{i,N_g,g}]$ is i's vector of item responses
- $S_i =$ economic outcome of interest for i (e.g. earnings)
- $X_{i,g} =$ other controls (e.g. survey wave, age, etc.)

Goal: estimate achievement $A_{i,g}$, defined by $\mathbb{E}[S_i|D_{i,g}, X_{i,g}]$

$$S_i = A_{i,g} + \eta_{i,g}, \ \mathbb{E}[\eta_{i,g} A_{i,g}] = 0$$

Construct $\hat{A}_{i,g}$ by estimating for some f:

$$\hat{A}_{i,g} \equiv \hat{S}_i = \hat{f}(D_{i,g}, X_{i,g})$$

Conceptual Framework and Method

Interested in estimating statistics like $\sigma^2_{A_g}$.

- $\hat{A}_{i,g}$ is estimated with error: $\sigma^2_{\hat{A}_g} = \sigma^2_{A_g} + \sigma^2_{\nu_g}$
- regression of $\hat{A}_{i,g}$ on $A_{i,g}$ estimates $R_g \equiv \sigma^2_{A_g}/(\sigma^2_{A_g} + \sigma^2_{\nu_g})$

Feasible regression of $\hat{A}_{i,g}$ on S_i downward biased for R_g.

- $S_i = A_{i,g} + \eta_{i,g}$
- need an instrument for S_i

Split-half IV approach:

- partition test items into disjoint (1) and (2)
- estimate $\hat{A}^{(1)}_{i,g}$ and $\hat{A}^{(2)}_{i,g}$ separately on these groups
- $Z^{(1)}_{i,g}$ = average S_j among $j \neq i$ where $\hat{A}^{(2)}_{i,g} = \hat{A}^{(2)}_{j,g}$

Empirical Implementation

Residualize S_i with year, race, sex, and interactions $\rightarrow \tilde{S}_i$.

- alternative residualizations yield very similar results.

By grade, separate lasso regressions of $\tilde{S}_i$ on odd and even item indicators.

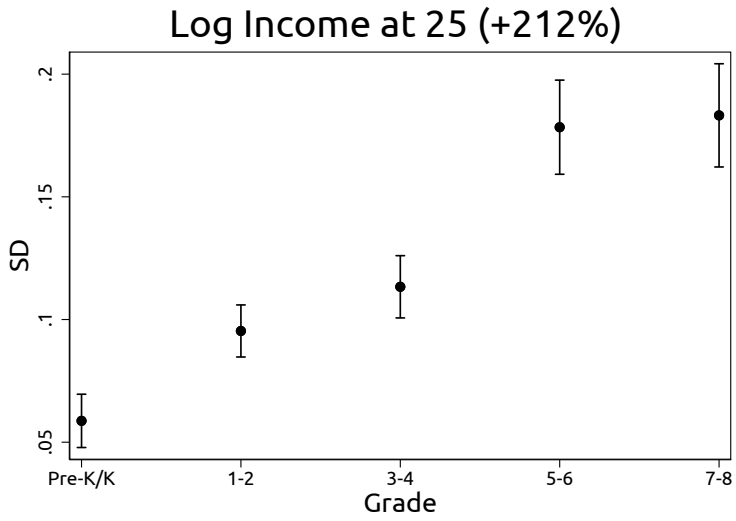
- yields $\hat{A}^{(1)}_{i,g}$ and $\hat{A}^{(2)}_{i,g}$

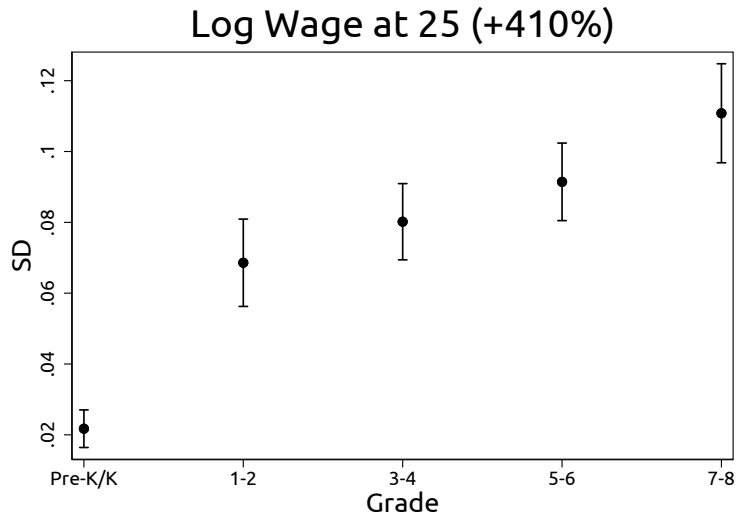
Regress $\hat{A}^{(1)}_{i,g}$ on $\tilde{S}_i$, instrumenting with $Z^{(1)}_{i,g}$.

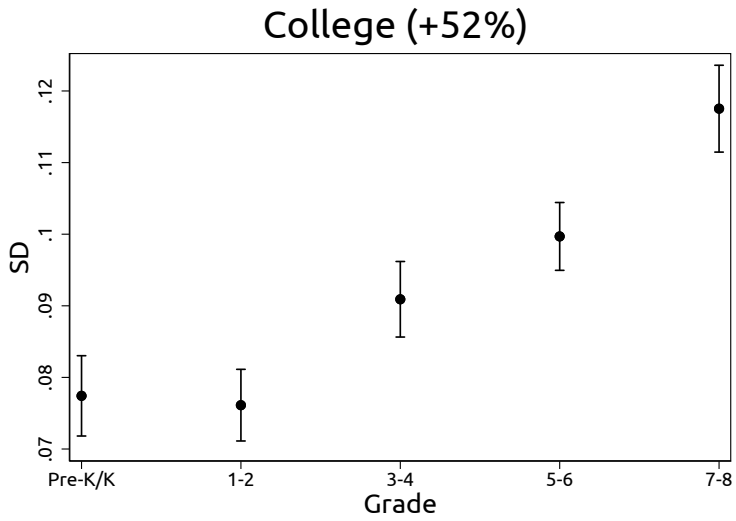
- yields $\hat{\tilde{\gamma}}^{(1)}_g$

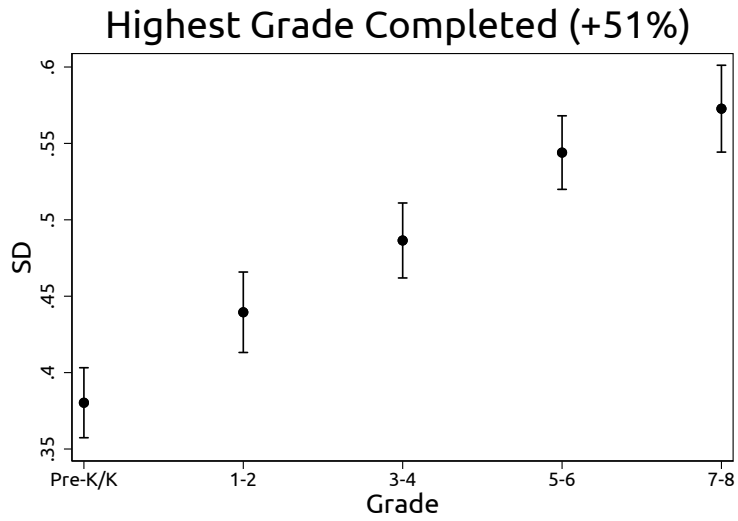
Estimate $\hat{\sigma}_{A_g}$ using $\sqrt{Var(\hat{A}^{(1)}_g) \times \hat{\tilde{\gamma}}^{(1)}_g}$.

The SD of Achievement Through Childhood

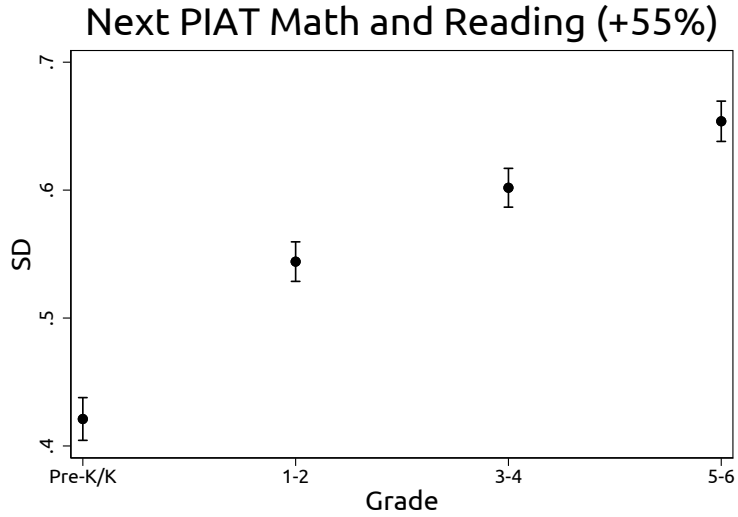








The SD of Achievement Through Childhood



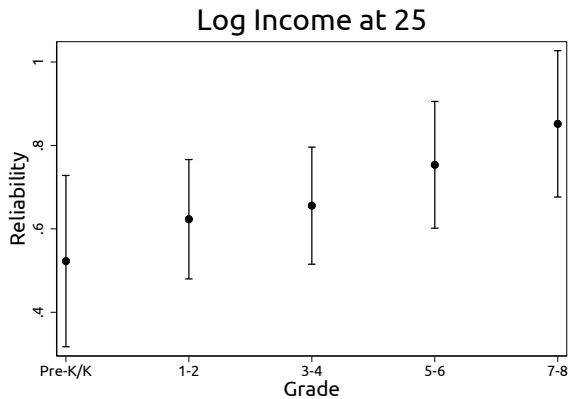
Lasso typically selects more items in higher grades.

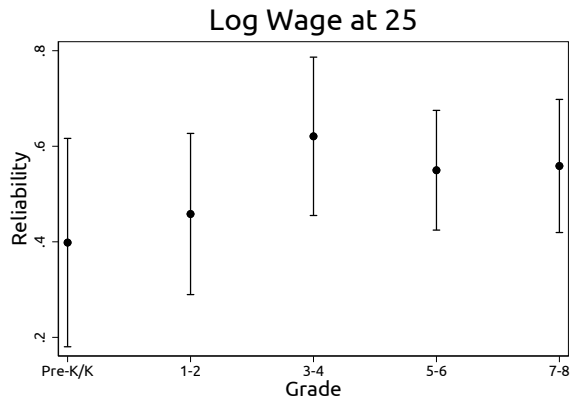
- about 20-25 in pre-k/k to 40-50 in grade 8
- not always – highest grade completed

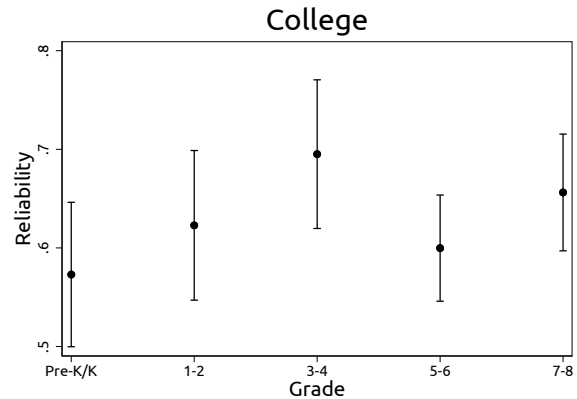
Alternative models yield qualitatively similar results.

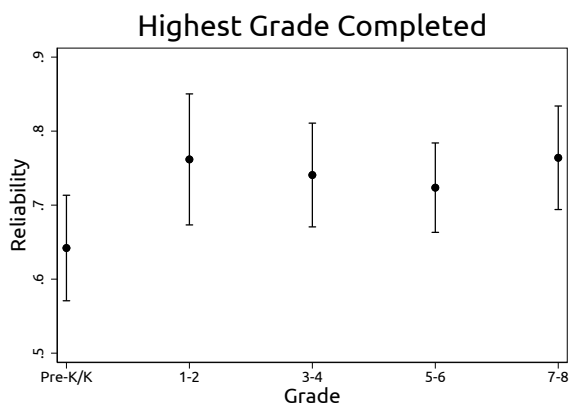
- lasso with 2-way item interactions, OLS, elasticnet, alternative lasso set-ups

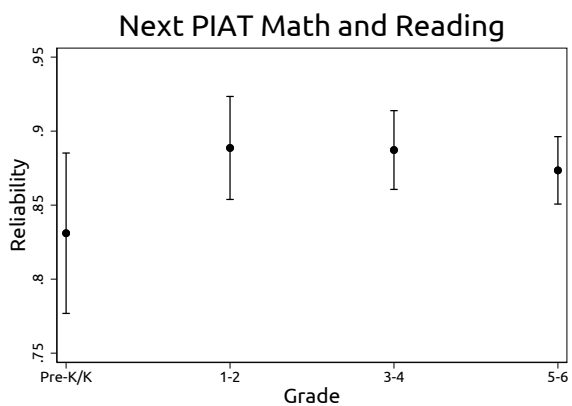
The Reliability of Item-Anchored Achievement



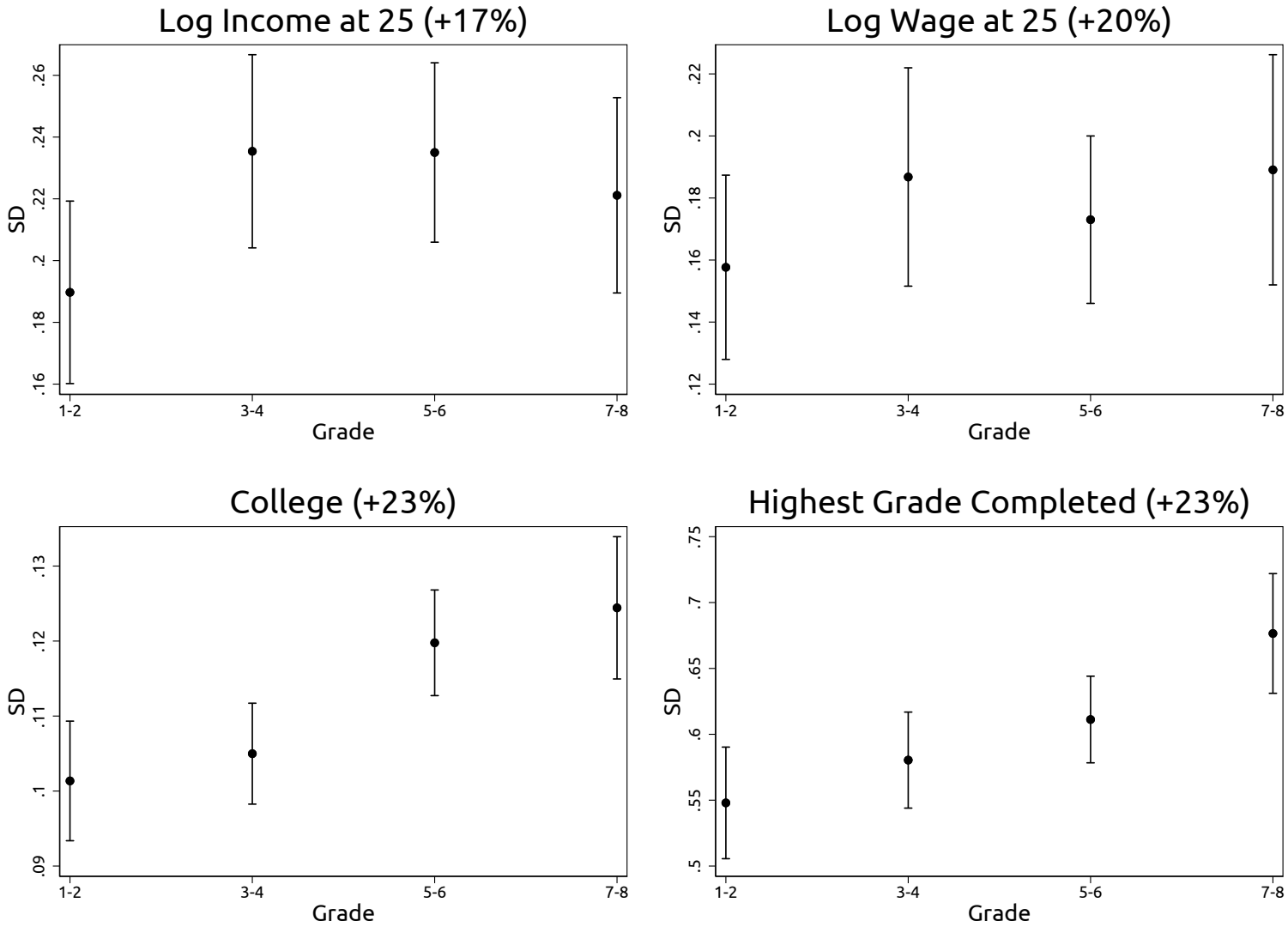








The SD of Given-Anchored Achievement



Implications for Young/Old Estimates

Growing variance $\Longrightarrow$ standardization shrinks later-age estimates

Compared to pre-k/k, same-size estimates in grades 7/8 will be

- 32% as large for log income at 25
- 20% as large for log wage at 25
- about 65% as large for highest grade, college, and next PIAT

These declines are similar in magnitude to:

- causal effect estimates on older versus younger children – e.g., Kane, Rockoff, and Staiger (2008); Dee and Jacob (2011)
- estimates of effect fade-out – e.g., Krueger and Whitmore (2001)

The Black-White Math & Reading Achievement Gap

	Log Income	Log Wage	College	Highest Grade	Next PIAT
Anchor	0.09^{***}	0.09^{***}	0.06^{***}	0.27^{***}	0.29^{***}
	(0.02)	(0.03)	(0.01)	(0.04)	(0.03)
Change	131%	869%	69%	78%	110%
By-Grade SD	-0.30	0.78^{*}	0.13	0.17	0.22^{***}
	(0.33)	(0.47)	(0.17)	(0.13)	(0.07)
Change	-26%	90%	11%	18%	35%

Anchor units – significant increases in achievement inequality.

SD units – mixed significance, smaller percentage changes.

Effect of Income on Achievement – Dahl and Lochner 2012

	sd units			anchor units		
	< 12	≥ 12	Δ_{sd}	< 12	≥ 12	Δ_{anchor}
PIAT	0.11	0.04	0.07			
	(0.07)	(0.03)	(0.08)			
College	0.12	0.08*	0.04	0.01	0.03*	-0.02
	(0.09)	(0.04)	(0.10)	(0.01)	(0.01)	(0.02)
Highest Grade	0.14	0.10**	0.05	0.06	0.06**	0.00
	(0.1)	(0.04)	(0.11)	(0.04)	(0.03)	(0.05)
Log(income)	0.10	0.12**	-0.02	0.01	0.03*	-0.03*
	(0.10)	(0.06)	(0.11)	(0.01)	(0.02)	(0.02)
Log(wage)	0.16	0.11**	0.05	0.01	0.03**	-0.02*
	(0.12)	(0.06)	(0.13)	(0.01)	(0.02)	(0.02)
Next PIAT	0.08	0.07*	0.01	0.04	0.05*	-0.01
	(0.07)	(0.04)	(0.08)	(0.04)	(0.03)	(0.04)

Anchor units – income has much larger effects on older children

SD units – income might have larger effects on younger children

Conclusion

The sd of achievement is much larger for older children.

Converting scores to sd units is not benign:

- white-black achievement gap trends
- causal effects of income

The use if item-anchored scores is critical.

Future research:

- what (if anything) unites predictive items?
- how do items/skills interact?
- more data and methodological exploration needed

Thank you!

Eric Nielsen
Federal Reserve Board
eric.r.nielsen at frb dot gov
https://sites.google.com/site/ericnielsenecon/home