# Test Questions, Economic Outcomes, and Inequality

Eric Nielsen

Federal Reserve Board

Spring 2020

# Disclaimer

Disclaimer: The views and opinions expressed in this presentation are solely those of the author and should not be interpreted as reflecting the official policy or position of the Board of Governors or the Federal Reserve System.

Danielle Nemschoff and Bryce Turner provided excellent research assistance.

# Measuring human capital is important.

**human capital is fundamental to understanding economic outcomes and inequality**

- labor market success, health, family formation, etc.

**how should human capital be measured?**

- not directly observable
- economic outcomes slow to be realized

**test scores are therefore frequently used as proxies**

- correlate with earnings, health, and many other outcomes
- readily available and easy-to-use
- achievement gaps/trends, value-added, policy effects, etc.

# Test scales are not interval measures of human capital.

**does a 1 sd change “mean” the same thing everywhere?**

- depends on the context and outcome of interest
  - improvements at bottom valuable for hs, at top for college

**standard statistics biased in presence of non-linearities**

- Bond and Lang (forthcoming), Bond and Lang (2013), Nielsen (wp), Schroeder and Yitzhaki (2017)
- many estimates very sensitive to small shifts in scale

# Test scales aggregate items without economics.

**test scales aggregate test items into a single index**

- e.g. "score = percent correct" treats all tests with the same number of right answers equivalently

**aggregation does not consider economic outcomes**

- skills emphasized by the test may differ from the skills associated with labor market success

**test scales may obscure real human capital differences**

- some groups may do better on "outcome relevant" items
- could conversely falsely identify human capital differences

**most analysis takes scale construction as a given**

# Goals of this paper:

1. **construct meaningful test scales by relating individual test items to economic outcomes**
   - solves both the aggregation and interval-scale problems
   - high school and college completion, wages, lifetime earnings

2. **compare rankings of item-anchored and standard scales**

3. **estimate achievement gaps by race, gender, and parental income**
   - IV methods to handle measurement error/shrinkage
   - estimate economically-relevant test reliability

# Overview of main results:

1. **item-anchored and given scales rank students differently**
   - ± 20 percentile point shifts not uncommon
2. **item-anchoring dramatically alters achievement gaps**
   - item-anchored gaps generally 0.1-0.5 sd larger
3. **item-anchored scores fully predict some outcome gaps**
   - black/white earnings gaps fully predicted
   - black/white employment gaps mostly predicted
   - high-/low-income gaps roughly twice as large as predicted
4. **item-anchoring resolves the “reading puzzle”**
   - reading scores jointly significant with math in wage regressions

# This work contributes to several literatures.

**test scores and cardinality**

- Bond and Lang (2013), Schroeder and Yitzhaki (2017)

**anchoring to later-life outcomes**

- Cunha, Heckman and Schennach (2010), Polacheck (2015), Bond and Lang (2018), many others

**reading puzzle**

- Sanders (wp), Kinsler and Pavan (2015), Arcidiacono (2004)

**achievement gaps by race and income**

- Neal and Johnson (1996), Fryer and Levitt (2004), Lang and Manove (2011), Ritter and Tayor (2011), Reardon (2011), many others

# NLSY79 item-level data

**National Longitudinal Survey of Youth (1979)**

- nationally representative, longitudinal survey
- 11,406 youth aged 14-22 in 1979 in my analysis sample

**Armed Forces Qualifying Test (AFQT)**

- math = math knowledge + arithmetic reasoning (55 items)
- reading = passage comp + word knowledge (50 items)
- blank items treated as incorrect

**reported scores estimated using 3PL IRT model**

- use age-standardized ($z$ scores) for math and reading
- about 9% of sample missing achievement measures

# NLSY79 outcomes data

**college and high-school completion**

- highest grade completed through 1994

**average wages at age 30**

- average three rounds nearest age 30
- missing frequently, especially for women and minorities

**present discounted value of lifetime labor income**

- pessimistic imputation of missing labor income
- extrapolate to retirement using age/education profiles from ACS

# Notation and Framework

**$M$ individuals indexed by $i$ take a test with $N$ binary questions indexed by $j$**

- $d_{i,j} = 1$ if $i$ gets $j$ correct, 0 o.w.
- $D_i = [d_{i,1}, \ldots, d_{i,N}]$, individual $i$'s vector of item responses
- $S_i =$ economic outcome of interest for $i$ (e.g. earnings)

**goal: estimate achievement $A_i$, defined by $\mathbb{E}[S_i|D_i]$**

$$S_i = A_i + \eta_i, \ \mathbb{E}[\eta_i A_i] = 0$$

**construct $\hat{A}_i$ by estimating for some $f$:**

$$\hat{A}_i \equiv \hat{S}_i = \hat{f}(D_i)$$

# OLS/probit anchoring

**start with linear regression (or probit) model**

$$S_i = D_i' W + \varepsilon_i, \text{ or } S_i = \Phi(D_i' W + \varepsilon_i)$$

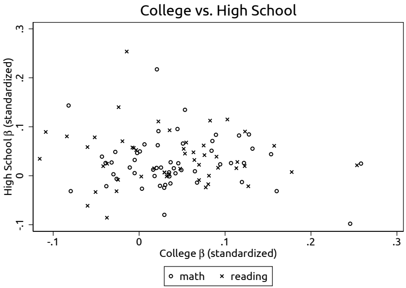
**assumes no interactions between test items**

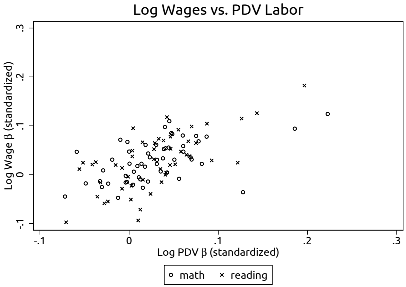
- models with interactions produce similar scales
- do not need to know *which* items are predictive
- more work needed on dimension reduction

**elements of $W$ are not structural parameters**

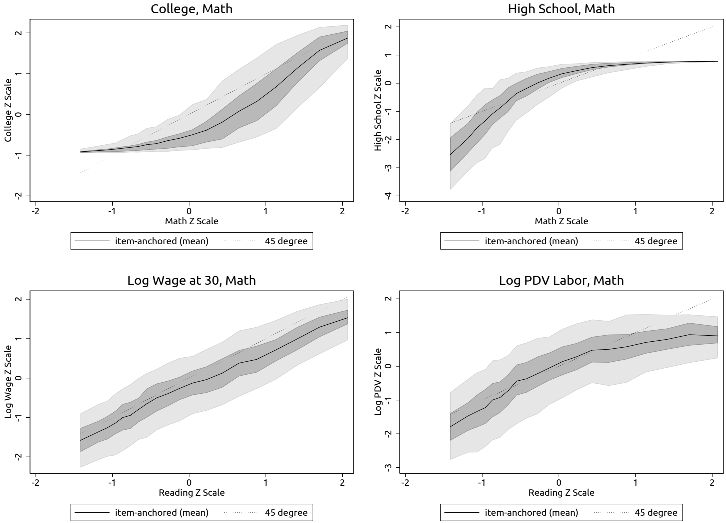
- just trying to flexibly estimate $\mathbb{E}[S_i|D_i]$
- no causality here, just anchoring

# Weights differ across items and outcomes.





# Item-anchored scores differ from given scores.



▸ reading ▸ percentile differences

# Interpretation of the item-anchored scores.

**achievement simple defined as $\mathbb{E}[S|D]$**

- inherently multidimensional: different outcomes $\implies$ different achievement measures
- anything correlated with both items and outcomes will contribute to achievement
- e.g. private schools teach Homer → Homer items predict income, even if Homer is useless

**is the "Homer problem" a problem?**

- potentially a problem for all research linking test scores to outcomes
- AFQT items selected and validated to test broad skills free of cultural bias
  - AFQT does not ask about Homer...

# The item weights are not just picking up confounders.

**similar baseline results estimating scales on different demographic subgroups**

- baseline labor outcome scales estimated using only white men
- adding demographic controls makes little difference

**at the item-weight level:**

- $\hat{W}(S)_{j,g}$ = item $j$'s weight for $S$, estimated on group $g$
- rarely or never reject $\hat{W}(S)_{j,g} = \hat{W}(S)_{j,g'}$
- often reject $\hat{W}(S)_{j,g} = \hat{W}(S')_{j,g}$

▸ schooling table

# Achievement gaps

**how do item-anchored gaps compare to given gaps?**

- some groups may do well on “outcome-relevant” items

**measurement error becomes important**

- no longer care simply about rank order

**“raw” gaps biased toward 0 due to measurement error**

- adapt methods from Bond and Lang (2018)
- disjoint sets items to construct independent measurements
- difficult to handle with observed scores (no way to generate “new” measurements)

# Anchoring and measurement error

**suppose that $A \sim N(\bar{A}, \sigma_A^2)$ in the population**

- normality for simplicity
- linear approximation if normality fails

**goal: estimate $\Delta A_{H,L} \equiv \bar{A}_H - \bar{A}_L$**

- differences by race, gender, parental income, etc.

**problem: anchored scores estimated with error**

$$\hat{A}_i = A_i + \nu_i$$

**naive averages of $\hat{A}_i$ yield downward-biased estimates**

# Anchoring and measurement error

**suppose** $\nu_i \sim N(0, \sigma_\nu^2)$

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

$\implies$ $\bar{\hat{A}}_H - \bar{\hat{A}}_L$ **is biased towards 0**

$$\text{plim}(\bar{\hat{A}}_H - \bar{\hat{A}}_L) = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_\nu^2}(\Delta A_{H,L})$$

**bias depends on the amount of signal in** $\hat{A}$

- $\sigma_\nu^2$ has nothing to do with psychometric reliability

# Anchoring and measurement error

**problem: need to estimate** $R_{A,\nu} \equiv \frac{\sigma_A^2}{\sigma_A^2+\sigma_\nu^2}$

**if** $A_i$ **were observed could regress** $\hat{A}_i$ **on** $A_i$

- plim$\hat{\beta} = R_{A,\nu}$
- but $A_i$ is not known (that's the whole problem!)

**solution: use** $S_i$ **in place of** $A_i$

$$\hat{A}_i = \kappa + \gamma S_i + \varepsilon_i$$

- but $S_i = A_i + \eta_j \implies plim(\hat{\gamma}) < R_{A,\nu}$
- errors-in-variables problem biases $\gamma$

# Instrumenting to remove attenuation bias

**instrument by creating two separate anchored scales**

- partition test items into disjoint groups (1) and (2)
- estimate $\hat{A}_i^{(1)}$ and $\hat{A}_i^{(2)}$ separately on these groups

**consider** $\hat{A}_i^{(1)} = \kappa_1 + \gamma_1 S_i + \varepsilon_i$

$$\begin{aligned} z_i^{(1)} &= N_i^{-1} \sum_{i' \neq i:\ \hat{A}_i^{(2)} = \hat{A}_{i'}^{(2)}} S_{i'} \\ plim(\hat{\gamma}_1^{IV}) &= R_{A,\nu} \end{aligned}$$

**leave-one-out instrument**

# Instrumenting to remove attenuation bias

**$\hat{A}_i^{(1)}$ and $\hat{A}_i^{(2)}$ can be used interchangeably**

- $z_i^{(1)}$ and $z_i^{(2)}$ both yield consistent estimates
- can compare and test for equality
- in practice, results are very similar either way

**generically, no $i'$ such that $\hat{A}_i^{(2)} = \hat{A}_{i'}^{(2)}$ exactly**

- divide $\hat{A}^{(2)}$ into many percentile buckets
- use the average outcome within each bucket

**very many ways to create groups (1) and (2)**

- equal numbers of items for now
- could create more than 2 groups

# Inference

**calculations typically treat reliabilities as known**

- reliability errors are seldom available

**item-anchored scales and reliabilities are estimated**

- bootstrapped standard errors are 50-150% larger
- (almost) all gap comparisons remain highly significant
- "naive" standard errors similar to given score errors

**show non-bootstrapped standard errors today**

- want standard errors comparable (in construction) to standard calculations using given scores
- accounting for reliability estimation another advantage of item-anchoring
- conclusions not changed using larger, bootstrapped errors

# Item-anchored black/white gaps

| White/Black | $z$ | item $z$ | predicted | actual | item $R$ |
|---|---|---|---|---|---|
| math, college | 0.98 | 0.81 | 0.20 | 0.13 | 0.87 |
| | (0.03) | (0.03) | (0.01) | (0.01) | . |
| reading, college | 1.05 | 1.28 | 0.25 | 0.13 | 0.74 |
| | (0.02) | (0.04) | (0.01) | (0.01) | . |
| math, wage | 0.98 | 1.21 | 0.25 | 0.24 | 0.75 |
| | (0.03) | (0.04) | (0.01) | (0.02) | . |
| reading, wage | 1.05 | 1.20 | 0.23 | 0.24 | 0.87 |
| | (0.02) | (0.03) | (0.01) | (0.02) | . |
| math, pdv | 0.98 | 1.49 | 0.46 | 0.45 | 0.69 |
| | (0.03) | (0.04) | (0.01) | (0.03) | . |
| reading, pdv | 1.05 | 1.34 | 0.41 | 0.45 | 0.75 |
| | (0.02) | (0.04) | (0.01) | (0.03) | . |

# Item-anchored high/low income gaps

| High/Low Income | $z$ | item $z$ | predicted | actual | item $R$ |
|---|---|---|---|---|---|
| math, college | 0.99 | 0.94 | 0.23 | 0.29 | 0.87 |
| | (0.03) | (0.03) | (0.01) | (0.01) | . |
| reading, college | 0.90 | 1.20 | 0.23 | 0.29 | 0.74 |
| | (0.03) | (0.04) | (0.01) | (0.01) | . |
| math, wage | 0.99 | 1.19 | 0.24 | 0.46 | 0.75 |
| | (0.03) | (0.04) | (0.01) | (0.02) | . |
| reading, wage | 0.90 | 1.09 | 0.20 | 0.46 | 0.87 |
| | (0.03) | (0.03) | (0.01) | (0.02) | . |
| math, pdv | 0.99 | 1.18 | 0.36 | 0.82 | 0.69 |
| | (0.03) | (0.04) | (0.01) | (0.03) | . |
| reading, pdv | 0.90 | 1.06 | 0.32 | 0.82 | 0.75 |
| | (0.03) | (0.04) | (0.01) | (0.03) | . |

# Item-anchored male/female gaps

| Male/Female | $z$ | item $z$ | predicted | actual | item $R$ |
|---|---|---|---|---|---|
| math, college | 0.18 | 0.13 | 0.03 | -0.00 | 0.87 |
| | (0.02) | (0.02) | (0.01) | (0.01) | . |
| reading, college | -0.11 | 0.01 | 0.00 | -0.00 | 0.74 |
| | (0.02) | (0.03) | (0.01) | (0.01) | . |
| math, wage | 0.18 | 0.24 | 0.05 | 0.22 | 0.75 |
| | (0.02) | (0.03) | (0.01) | (0.01) | . |
| reading, wage | -0.11 | -0.10 | -0.01 | 0.22 | 0.87 |
| | (0.02) | (0.02) | (0.00) | (0.01) | . |
| math, pdv | 0.18 | 0.25 | 0.07 | 0.47 | 0.69 |
| | (0.02) | (0.03) | (0.01) | (0.02) | . |
| reading, pdv | -0.11 | -0.08 | -0.03 | 0.47 | 0.75 |
| | (0.02) | (0.03) | (0.01) | (0.02) | . |

# Median regression can address wage selection.

**wage data frequently not available, selection likely an issue**

- missing 17% for white men, 25% or more for other groups

**let $\tilde{S}_i$ be the latent wage ($S_i = \tilde{S}_i$ if working)**

- suppose $\tilde{S}_i = D_i' W + \varepsilon_i$ with $\varepsilon$ iid and median$(\varepsilon) =$ mean$(\varepsilon)$
- median$[\tilde{S}_i | D_i] =$ mean$[\tilde{S}_i | D_i]$

**“fill in” missing wages with 0’s and identify means using median regression**

- assumes negative selection: $\tilde{S}_i <$ median$[\tilde{S} | D_i]$ for $S_i = 0$
- otherwise same procedure as before

# Item-anchored wage gaps, median regression

| white/black | naive $z$ | item-anchored | predicted | actual | item $R$ |
|---|---|---|---|---|---|
| math | 0.98 | 1.48 | 0.34 | 0.24 | 0.64 |
| | (0.03) | (0.04) | (0.01) | (0.02) | . |
| reading | 1.05 | 1.27 | 0.25 | 0.24 | 0.78 |
| | (0.02) | (0.04) | (0.01) | (0.02) | . |
| male/female | | | | | |
| math | 0.18 | 0.31 | 0.05 | 0.22 | 0.64 |
| | (0.02) | (0.03) | (0.01) | (0.01) | . |
| reading | -0.11 | -0.18 | -0.05 | 0.22 | 0.78 |
| | (0.02) | (0.03) | (0.01) | (0.01) | . |
| high/low | | | | | |
| math | 0.99 | 1.44 | 0.30 | 0.46 | 0.64 |
| | (0.03) | (0.04) | (0.01) | (0.02) | . |
| reading | 0.90 | 1.12 | 0.21 | 0.46 | 0.78 |
| | (0.03) | (0.04) | (0.01) | (0.02) | . |

# Item-anchored scores explain racial employment gaps.

- Prior research finds that test scores do not explain black/white employment gaps (Ritter and Taylor 2011)
- Item-anchoring to Ritter and Taylor's outcomes:
    - items predict 70% of the black/white gap in cumulative unemployment through 2004 for men
    - 78% of the cumulative weeks not working through 2004

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

# Item-anchoring helps resolve the “reading puzzle.”

Table: Reading Puzzle Regressions – Wages at Age 30

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | given | given-anchored | item-anchored | given | given-anchored | item-anchored |
| math | 0.17*** | 0.17*** | 0.16*** | 0.10*** | 0.11*** | 0.11*** |
| | (0.02) | (0.02) | (0.01) | (0.02) | (0.02) | (0.02) |
| reading | 0.03 | 0.02 | 0.09*** | -0.01 | -0.01 | 0.06*** |
| | (0.02) | (0.02) | (0.01) | (0.02) | (0.02) | (0.02) |
| education | no | no | no | yes | yes | yes |
| parental income | no | no | no | yes | yes | yes |
| white male only | yes | yes | yes | yes | yes | yes |
| Observations | 2,306 | 2,306 | 2,217 | 2,232 | 2,232 | 2,142 |
| Adjusted $R^2$ | 0.12 | 0.12 | 0.16 | 0.17 | 0.17 | 0.20 |

**similar results using scales item-anchored to different outcomes (schooling, etc.)** ▸ other anchors

# Wrapping up

**achievement tests misaligned with human capital**

- item-anchoring changes ranks and achievement gaps
- LATEs and other causal effects may be mis-identified

**achievement inequality is worse than test scores indicate**

- black/white and high/low income gaps anchored to labor market outcomes are much larger than naive gaps

**many future avenues of research**

- optimal instrument construction
- noncognitive skills
- what makes an item predictive?
- anchored changes away from the mean

# Thank you!

# Appendix: What unites the predictive items?

**do not know the content of the items**

**do know the IRT parameters for each item: discrimination ($\alpha_j$), difficulty ($\beta_j$), guessability ($\gamma_j$)**

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

**regress $\hat{W}_j$ on $\alpha_j$, $\beta_j$, and $\gamma_j$**

- labor market outcomes, hs completion: high discrimination, low difficulty, and low guessing probability
- college completion: high discrimination, high difficulty, low guessing
- most variation not explained ($R^2 \approx 0.1 - 0.4$)

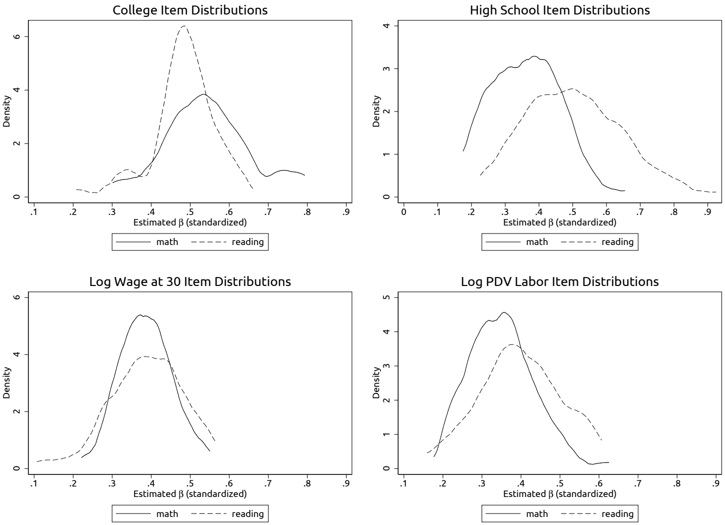
▸ regression table

# Appendix: IRT Parameter Regression

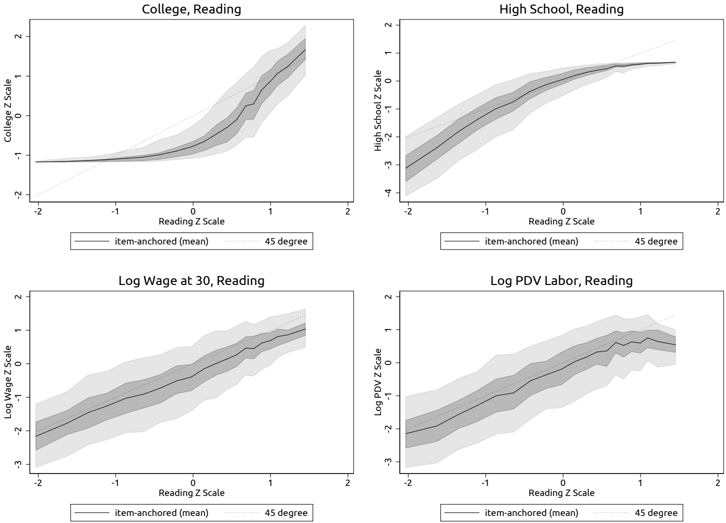
| | (1)<br>wage item | (2)<br>wage full | (3)<br>pdv item | (4)<br>pdv full | (5)<br>hs item | (6)<br>hs full | (7)<br>col item | (8)<br>col full |
|---|---|---|---|---|---|---|---|---|
| discrimination | 0.04*** | 0.00 | 0.05*** | -0.00 | 0.02*** | -0.00 | 0.03*** | -0.00 |
| | (0.00) | (0.00) | (0.01) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) |
| difficulty | -0.03*** | 0.00 | -0.12*** | -0.02*** | -0.06*** | -0.01*** | 0.01** | 0.02*** |
| | (0.01) | (0.00) | (0.01) | (0.01) | (0.00) | (0.00) | (0.00) | (0.00) |
| guessing | -0.08* | -0.02 | -0.07 | 0.00 | -0.00 | 0.01 | -0.10** | -0.01 |
| | (0.05) | (0.03) | (0.09) | (0.06) | (0.03) | (0.02) | (0.04) | (0.03) |
| Obs. | 105 | 105 | 105 | 105 | 105 | 105 | 105 | 105 |
| Adj. $R^2$ | 0.372 | -0.022 | 0.566 | 0.086 | 0.738 | 0.317 | 0.482 | 0.205 |

▸ back

# Item correlations with outcomes vary widely.



# Appendix: Item-anchored vs Given, Reading.


College, Reading

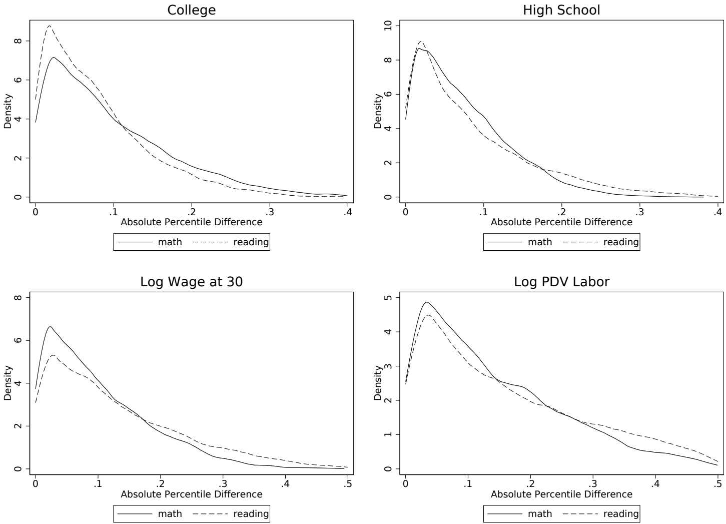
High School, Reading

Log Wage at 30, Reading

Log PDV Labor, Reading

▸ back

# Percentile ranks differ between item-anchored and given scales.



back

# Appendix: Item-by-Item Hypothesis Tests, Share Rejected

| | $\hat{W}(\text{test})_{j,g} = \hat{W}(\text{test})_{j,g'}$ | | $\hat{W}(\text{school})_{j,g} = \hat{W}(\text{school})_{j,g'}$ | | $\hat{W}(\text{test})_{j,g} = \hat{W}(\text{school})_{j,g}$ | |
|---|---|---|---|---|---|---|
| male/female | no mc | Bernoulli | no mc | Bernoulli | no mc | Bernoulli |
| math, hs | 0.15 | 0.00 | 0.13 | 0.00 | 0.51 | 0.22 |
| reading, hs | 0.18 | 0.04 | 0.14 | 0.02 | 0.44 | 0.26 |
| math col | 0.15 | 0.00 | 0.13 | 0.02 | 0.45 | 0.13 |
| reading col | 0.18 | 0.04 | 0.08 | 0.00 | 0.58 | 0.12 |
| black/white | no mc | Bernoulli | no mc | Bernoulli | no mc | Bernoulli |
| math, hs | 0.05 | 0.02 | 0.07 | 0.00 | 0.56 | 0.36 |
| reading, hs | 0.10 | 0.00 | 0.08 | 0.00 | 0.50 | 0.28 |
| math, col | 0.05 | 0.02 | 0.07 | 0.00 | 0.42 | 0.15 |
| reading, col | 0.10 | 0.00 | 0.06 | 0.02 | 0.46 | 0.14 |
| high-/low-income | no mc | Bernoulli | no mc | Bernoulli | no mc | Bernoulli |
| math, hs | 0.20 | 0.05 | 0.07 | 0.02 | 0.58 | 0.18 |
| reading, hs | 0.18 | 0.00 | 0.18 | 0.04 | 0.50 | 0.32 |
| math, col | 0.20 | 0.05 | 0.16 | 0.00 | 0.58 | 0.22 |
| reading, col | 0.18 | 0.00 | 0.14 | 0.04 | 0.66 | 0.28 |

▸ back

# Appendix: reading puzzle with other anchors

| | (1)<br>ln_w30 item | (2)<br>college item | (3)<br>college given | (4)<br>high school item | (5)<br>high school given |
|---|---|---|---|---|---|
| math | 0.11*** | 0.07*** | 0.10*** | 0.05** | 0.08*** |
| | (0.02) | (0.02) | (0.02) | (0.02) | (0.02) |
| reading | 0.06*** | 0.03* | -0.02 | 0.04** | 0.01 |
| | (0.02) | (0.02) | (0.02) | (0.02) | (0.02) |
| education | yes | yes | yes | yes | yes |
| parental income | yes | yes | yes | yes | yes |
| white male only | yes | yes | yes | yes | yes |
| Observations | 2,142 | 2,142 | 2,232 | 2,142 | 2,232 |
| Adjusted $R^2$ | 0.20 | 0.16 | 0.17 | 0.16 | 0.16 |

▸ back