Should Value-Added Models Weight All Students Equally?

Case Tatro

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

October 24, 2025

Disclaimer

The views expressed are those of the author and do not necessarily reflect the positions or policies of the North Carolina Department of Public Instruction.

Value-Added (VA) Models: Basics

- School systems needed a way to evaluate teachers.
- Economists came up with statistical models to estimate a teacher's impact on student outcomes.
  - Common outcomes include math and reading standardized test scores.
  - Other outcomes include absences, disciplinary incidents, etc.
- VA models are a popular way to evaluate teachers.
  - In 2023, 30 states used test-score VA measures as part of teacher evaluations (National Council on Teacher Quality, 2024).

Value-Added (VA) Models: Construction

1. Step 1: Use a regression model to predict students' outcomes using their predetermined characteristics, the most important of which is a student's lagged outcome
2. Step 2: Find the difference between a student's actual and predicted outcome (**residual**).
3. Step 3: Calculate the simple average of the residual for all students within a classroom.
4. (Optional) Step 4: Apply a shrinkage correction.

Value-Added (VA) Models: State of Literature

- **Pros**:
  1. VA models do not reward teachers for having students who would have scored highly on standardized tests regardless (e.g., Kane and Staiger, 2008; Hanushek and Rivkin, 2010; Bacher-Hicks, Kane, and Staiger, 2014; Chetty, Friedman, and Rockoff, 2014a).
     - Approximately forecast unbiased.
  2. Teachers estimated to be good at raising test scores are also estimated to be good at promoting students' long-term outcomes. (e.g., Chetty, Friedman, and Rockoff, 2014b; Lavy and Megalokonomou, 2024).
     - School systems can evaluate teachers on long-run impacts without waiting for students' long-term outcomes to become realized.

Value-Added (VA) Models: State of Literature

- **Cons**:
  1. Test-score VA measures, while predictive, explain very little of the variation in a teacher's long-run impact (e.g., Deming, 2009; Chetty et al., 2011; Chamberlain, 2013).
     - e.g. Chetty et al. (2011) finds test score impacts explain 20% of a teacher's long-run impacts.
     - Including a teacher's impacts on students' non-cognitive outcome closes some of this gap.
  2. Teachers have different test-score impacts on different types of students (e.g., Lavy, Paserman, and Schlosser, 2012; Condie, Lefgren, and Sims, 2014; Fox, 2016; Delgado, 2020; Aucejo et al., 2022; Gershenson et al., 2022; Graham et al., 2023).
     - Including students with different levels of baseline achievement (e.g., Biasi, Fu, and Stromme, 2021; Eastmond et al., 2024).

This Paper

- **Question**: Would a weighted average of student residuals improve the accuracy of a teacher's impacts on test-score in predicting a teacher's impacts on long-run outcomes?
- **Setting**: Elementary school (grades 3-5) teachers in North Carolina.
- **Data**: North Carolina Education Research Data Center (NCERDC).
- **Method**: Estimate weights across students such that I minimize the prediction error of a teacher's long-run impacts using test-score residuals.
- **Takeaway**: A weighted average does a better job than an unweighted average at predicting a teacher's long-run impacts.

Is Equally-Weighting Students Reasonable?

1. Equal-weighting implies that raising test scores from high to higher is just as important as raising test scores from below-basic to basic.
   - On the Armed Forces Qualification Test (AFQT), answering easier questions correctly is more predictive of individuals' long-run outcomes (Nielsen, 2019).
2. Equal-weighting seems unlikely to be the best way to estimate a teacher's long-run impacts.
   - Consider high school graduation as the long-run outcome of interest.
   - Perhaps the students most at risk of not graduating are lower-achieving students.
   - If so, a more informative test-score VA model might be one that places a higher weight on a teacher's impacts on short-run outcomes for low-achieving students.

Research Questions

1. What are the optimal weights for predicting a teacher's long-run impacts using a teacher's impacts on test scores?
2. How much more predictive is a weighted VA measure compared to a conventional VA measure?
3. Do these optimal weights represent true differences in teacher effects or an efficient use of a small sample?

Preview of Findings

1. The highest-achieving students (based on students' baseline achievement) receive the highest weight.
   - This holds even when predicting long-run outcomes for the lowest-achieving students.
2. A weighted VA improves the predictive power of a teacher's long-run impacts by about 10%.
3. The highest-achieving students receive the highest weight for two reasons.
   1. The highest-achieving students have less-noisy residuals (small-sample efficiency).
   2. A teacher's true impacts on the highest-achieving students reflect general aspects of teaching that are especially important for promoting high school graduation (true effects).

NCERDC Data

- **Years**: 1998-2011.
- **Grades**: 4-5.
- **Outcomes**: Math and reading standardized test scores, high school graduation in North Carolina, absences, suspensions.
- **Characteristics**: Race, gender, economic disadvantaged status, lagged standardized math and reading test scores.

Summary Stats

Constructing VA: Step 1

$$\tilde{Y}_{i,j,s,t} = \alpha + \gamma \tilde{Y}_{i,t-1} + \tilde{\mathbb{X}}_i \beta + \eta_{i,j,s,t} \quad (1)$$

- Variables indexed by outcome s for student i in classroom $j(i)$ in year t.
- $\sim$ indicates variable demeaned at the classroom level.
- Tilde indicates demeaned at the classroom-year level.
- $Y_{i,j,s,t}$: Student outcome in year t.
- $Y_{i,t-1}$: Student outcomes in past year (cubic).
- $\mathbb{X}_i$: Vector of characteristics (gender, ethnicity, economic disadvantaged status, etc.).

Constructing VA: Step 2

$$\nu_{i,j,s,t} = Y_{i,j,s,t} - \hat{Y}_{i,j,s,t}$$
$$\hat{Y}_{i,j,s,t} = \hat{\alpha} + \hat{\gamma} Y_{i,t-1} + \mathbf{X}_i \hat{\beta}$$
$$\epsilon_{i,j,s,t} = v_{i,j,s,t} - \frac{1}{N} \sum_{i=1}^{N} \nu_{i,j,s,t}$$

- $\hat{Y}_{i,j,s,t}$: Student's predicted value using Step 1 coefficients and student characteristics.
- $\nu_{i,j,s,t}$: Difference between a student's actual and predicted outcome (**residual**).
- $\epsilon_{i,j,s,t}$: Recentered residual.

Background: Constructing VA Measures (Steps 3 & 4)

$$\hat{VA}_{j,s,t} = \sum_{i=1}^{N_j} \epsilon_{i,j,s,t}$$

$$VA'_{j,s,t} = \Omega^* * \hat{VA}_{j,s,t}$$

- $\hat{VA}^k_{j,s,t}$: Teacher j's unadjusted test-score VA for subject s in year t.
- $VA'_{j,s,t}$: Teacher j's shrunken test-score VA for subject s in year t.
  - I prefer the method used in Mulhern & Opper (2023) to shrink these VA estimates.
  - Incorporates information about other outcomes when shrinking unadjusted value-added measures. Example

How to Weight Students in a VA Model?

Goal: Estimate weights on a teacher's impact on student test scores that maximize the predictive power of a teacher's long-run impacts.

- The weights are based on lagged test scores.
- Unweighted average (conventional VA) is a special case of my weighted VA measure.
- I use high school graduation as my long-run outcome.

Weighted VA: Defining VA^*

$$VA^*_{j,s,t} = \frac{1}{W} \sum_{i,k} \beta_k \mathbf{1}\{i \in k\} \mathbf{1}\{i \in j\} \epsilon_{i,j,s,t}$$

$$W = \sum_{i} \sum_{k} \beta_k \mathbf{1}\{i \in k\} \mathbf{1}\{i \in j\}$$

- $\epsilon_{i,j,s,t}$: the re-centered test-score residual in subject s for student i in teacher j's class in year t.
- Student i is grouped into bin k based on student i's lagged test score.
- β_k represents the coefficient on each bin k.
- Use 5 bins, or quintiles of lagged student achievement within a school, grade, and year.

Weighted VA: Toy Example

- Suppose that what a school district cares about is a teacher's impact on high school graduation.
- Also suppose a student's probability of graduating is given by a probit model of their lagged math or reading standardized test score.
- Under such a scenario, the lowest-achieving students are most at risk of not graduating high school.
- Therefore, we would expect the **optimal way to weight a teacher's test-score impacts on students is to place the highest weight on the lowest-achieving students.**

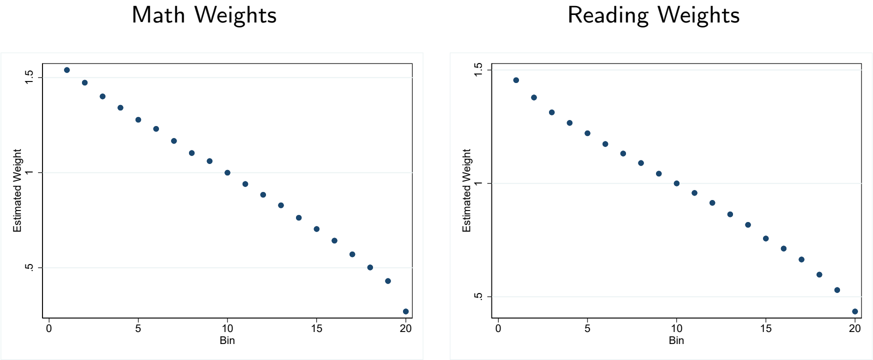
Toy Example: Econometrics

$$Pr(Graduated_{i,j,t}|\tilde{Y}_{i,k,s,t-1}) = \Phi(\delta_0 + \delta_1 \tilde{Y}_{i,k,s,t-1}$$
$$+\delta_2 \tilde{Y}^2_{i,k,s,t-1} + \delta_3 \tilde{Y}^3_{i,k,s,t-1} + \mathbf{X}_i \Delta + \omega_{i,j,s,t})$$

- $Y_{i,t-1}$: Vector of student outcomes in past year (cubic).
- $\mathbf{X_i}$: Vector of student demographic information
- $\omega_{i,j,s,t}$: Residual clustered at the classroom level.
- **Suppose**
  1. A teacher's VA in each bin is independent of VA in every other bin.
  2. A teacher's VA for a student's particular bin is the only factor that affects a student's graduation probability.
- Then, test-score effects are proportional to how at-risk a student is of not graduating high school.

Toy Example: Estimated Weights

Math and Reading Weighting Functions



Weighted VA: Econometric Strategy

1. Divide lagged test scores for math or reading into 5 quintiles within a grade and year.
2. Within each class, classify students into their respective bin.
3. Estimate weights on each bin to maximize the predictive power of this weighted VA on a teacher's leave-one-year-out (LOYO) high school graduation VA.
    - Normalize weights such that the weight on the median bin is 1.
    - Weights above 1 indicate students who receive a higher weight.
    - Weights below 1 indicate students who receive a lower weight.

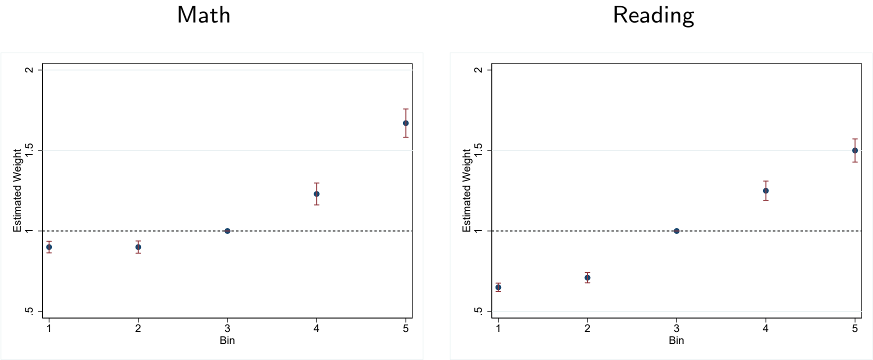
Weighted VA: Estimation

$$\min_{(\beta_k)s} \left[\tilde{VA}^{grad}_{j,-t} - \beta_0 - VA^*_{j,s,t} \right]^2$$

- $\tilde{VA}^{grad}_{j,-t}$: Teacher j's LOYO high school graduation VA for year t (Jackson, 2018). Full Details
- $VA^*_{j,s,t}$: Teacher j's weighted VA measure using student residuals for subject s in year t.
- Estimate using non-linear least squares.

Weighted VA: Initial Estimates

Initial Bin-Weight Estimates for Math and Reading



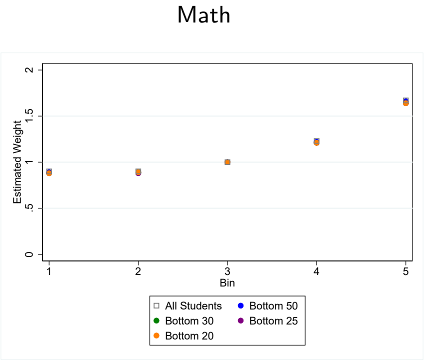
Weighted VA: Weights for Low HS Grad VA

$$\min_{(\beta_k)s} \left[LowGrad\tilde{VA}_{j,b,-t} - \beta_0 - VA^*_{j,s,t} \right]^2$$

- $LowGrad\tilde{VA}^{grad}_{j,b,-t}$: Teacher j's LOYO high school graduation VA measured only for students in the bottom b percentile of the lagged achievement distribution.
- $VA^*_{j,s,t}$: Teacher j's weighted VA measure using student residuals for subject s in year t.

Weighted VA: Weights by HS Grad VA Definition

Estimated Bin-Weights by Subject



Weighted VA: How Much More Predictive?

- **Question**: How much more predictive is a weighted VA measure compared to a conventional VA measure?
- **Method**: Regress a teacher's high school graduation VA on a teacher's weighted versus unweighted test-score VA.
- Compare R^2 values.
- **Findings**: Weighted VA increases the percentage of explained variation by $\sim$ 20% for math and $\sim$ 8% for reading (in the baseline).

Predictive Power: Baseline

Predictive Power of Weighted vs Unweighted VA: Baseline

	Math		Reading	
	(1)	(2)	(3)	(4)
Conventional (Unweighted) VA	0.000566***		0.000967***	
	(0.000114)		(0.000127)	
Weighted VA		0.000581***		0.000984***
		(0.000106)		(0.000123)
Observations	84,704	84,704	84,678	94,678
R^2	0.0010	0.0012	0.0017	0.0019
% Increase in Explained Variation	–	19.56	–	8.15

Notes: $p < 0.001^{***}$, $p < 0.05^{**}$, $p < 0.1^{*}$.

Predictive Power: Non-Cognitive VA Controls

Predictive Power of Weighted vs Unweighted VA: Non-Cognitive

	Math		Reading	
	(1)	(2)	(3)	(4)
Conventional (Unweighted) VA	0.000473***		0.000958***	
	(0.000133)		(0.000147)	
Weighted VA		0.000507***		0.000986***
		(0.000124)		(0.000143)
Suspension VA	0.0112***	0.0112**	0.0112***	0.0111***
	(0.00421)	(0.00421)	(0.00420)	(0.00420)
Behavioral Index VA	0.00392*	0.00385*	0.00366	0.00362*
	(0.00208)	(0.00208)	(0.00207)	(0.00207)
Observations	67,595	67,595	67,571	67,571
R^2	0.0024	0.0026	0.0033	0.0035
% Increase in Explained Variation	–	8.11	–	4.90

Notes: $p < 0.001^{***}$, $p < 0.05^{**}$, $p < 0.1^{*}$.

Predictive Power: Joint Math and Reading

Predictive Power of Weighted vs Unweighted VA: Joint Estimation

	Baseline		**Non-Cognitive**	
	(1)	(2)	(3)	(4)
Conventional Math VA	0.000143		0.000100	
	(0.000147)		(0.000146)	
Conventional Reading VA	0.000879***		0.000826***	
	(0.000158)		(0.000157)	
Weighted Math VA		0.000175		0.000135
		(0.000128)		(0.000128)
Weighted Reading VA		0.000827***		0.000779***
		(0.000138)		(0.000123)
Suspension VA			0.0111***	0.0111**
			(0.00420)	(0.00420)
Behavioral Index VA			0.00366*	0.00359*
			(0.00207)	(0.00208)
Observations	67,549	67,549	67,549	67,549
R^2	0.0017	0.0019	0.0031	0.0033
% Increase in Explained Variation	–	13.27	–	6.36

Notes: $p < 0.001^{***}, p < 0.05^{**}, p < 0.1^{*}$.

Why This Pattern of Weights

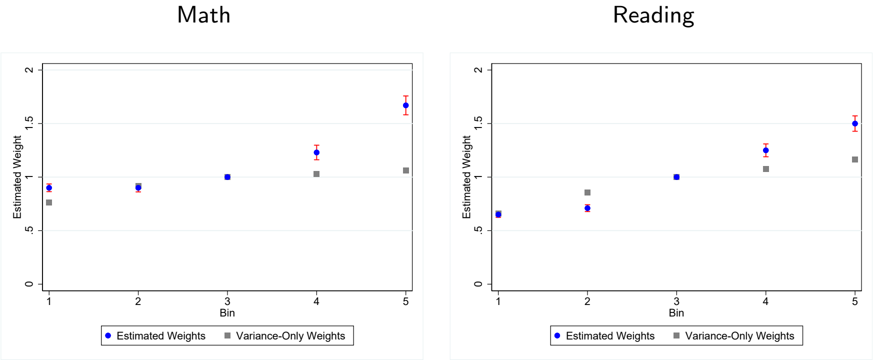
1. Differences in noisiness of student residuals across bins.
   - The bins which receive the highest weight are bins in which student test score residuals are the least noisy.
2. A teacher's impact on particular students is more informative about a teacher's impacts on other students.
   - The bins which receive the highest weight are bins whose true effects are most correlated with the true effects of other bins.
3. Attribute differences in weights not explained by (1) or (2) as a true-effects story.
   - Bin-specific VA may weight general aspects of teaching differently.
   - The bins which receive the highest weights are bins for which the bin-specific VA more heavily weights aspects of teaching especially important for promoting high school graduation.

Evaluating Explanation (1)

- Compare estimated weights to theoretical weights.
- Theoretical weights: Inverse of relative variance of student residuals.

Weighted VA: Comparing Estimated to Variance Weights

Estimated and Variance-based Bin-Weights by Subject



Disentangling (1) & (2) versus (3)

- Suppose small-sample efficiency completely explained the pattern of weights.
- If true, I should see similar weights when I use a teacher's impact on current test scores to predict a teacher's impact on test scores for the next cohort of students.

Weighted VA: Subsequent Cohort Bin Weight Estimates

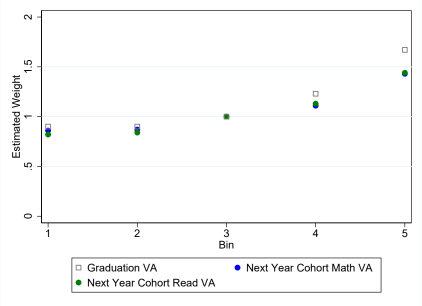
$$\min_{(\beta_k)s} \left[VA^s_{j,t+1} - \beta_0 - VA^*_{j,s,t} \right]^2$$

- $VA^s_{j,t+1}$: Teacher j's VA for math and reading scores using students in teacher j's class in the next year $(t+1)$.
- $VA^*_{j,s,t}$: Teacher j's weighted VA measure using student residuals for subject s in year t.

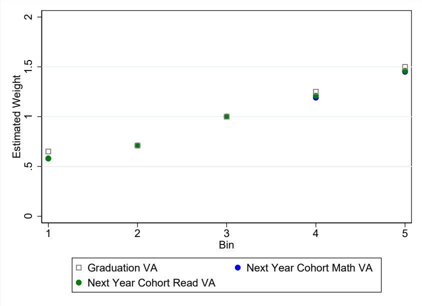
Weighted VA: Teacher VA on Subsequent Cohort

Estimated Bin-Weights by Subject

Math



Reading



Other Robustness Checks

- I observe a similar pattern of weights when looking at smaller and larger classes separately. Class Size Results
- Weights are similar when I swap the years used to estimate a teacher's high school graduation VA and impact on test scores. Reverse Out-of-Sample Results
- Weights are similar when I predict LOYO test-score impacts in subsequent grades. Future Test Score Results
- Weights are similar when I predict LOYO bin-specific math and reading VA. Bin-Specific VA Results

Conclusion

1. Value-added models become more accurate in predicting a teacher's long-run impacts when weighted.
2. The highest-achieving students receive the highest weight, even when using short-run impacts for all students to predict long-run outcomes for the lowest-achieving students.
3. This optimal pattern of weights is due to a combination of factors.
    1. Small-sample efficiency.
    2. True-differences in teacher quality.

Thank You

Questions, comments, feedback? ctatro1@binghamton.edu

NCERDC Data: Summary Stats

Summary Statistics of Student Data

Variable	Mean	SD	Min	Max
Female	0.510	(0.500)		
Black	0.235	(0.424)		
Hispanic	0.053	(0.225)		
White	0.483	(0.500)		
Asian	0.0159	(0.125)		
Economically Disadvantaged	0.838	(0.368)		
Student With Disabilities	0.157	(0.364)		
Academically Gifted	0.0282	(0.166)		
English Language Learner	0.125	(0.111)		
Ever Suspended	0.36	(0.48)		
Graduated High School	0.805	(0.396)		
-Ln(1+Absences)	-0.0973	(0.437)	-4.95	0
-Days Suspended	-0.0880	(1.314)	-447	0
Classroom Size	22.869	(3.73)	10	35
Student-Year Observations		2,587,625		
Students		1,633,504		

Back

Summary of Value-Added Measures by Dimensionality (Grades 3-5) Pooled Years

Group (1)	Subject (2)	Unadjusted Std. Dev (3)	1D Std. Dev (4)	2D Std. Dev (5)	2D Teachers (6)
Whole-Class (Typical VA)	Math	0.209	0.108	0.149	45,221
	Reading	0.155	0.0557	0.0711	45,221
Top 50%	Math	0.216	0.0851	0.132	35,763
	Reading	0.157	0.0282	0.0371	35,763
Bottom 50%	Math	0.230	0.102	0.146	41,209
	Reading	0.191	0.0541	0.0622	41,211
Top 30%	Math	0.224	0.0834	0.122	30,245
	Reading	0.169	0.0254	0.0298	30,246
Bottom 30%	Math	0.246	0.0969	0.140	38,988
	Reading	0.222	0.0529	0.0590	39,005
Top 25%	Math	0.230	0.0839	0.120	29,156
	Reading	0.177	0.0252	0.0296	29,156
Bottom 25%	Math	0.253	0.0948	0.135	38,026
	Reading	0.234	0.0529	0.0558	38,032

Back

Intuitive Example

- Suppose we are estimating adjusted measures for a teacher j's math and reading value-added in year t.
- Let's suppose our unadjusted value-added measures indicate teacher j's value-added in year t is 1 for math and 0 for reading.
- Suppose through our estimation we obtain a shrinkage matrix for teacher j of

$$\begin{bmatrix} 0.5 & 0.25 \\ 0.1 & 0.6 \end{bmatrix}.$$

- Our shrunken value-added estimate for teacher j would be:
  - $\tilde{VA}_{math} = 0.5 * \hat{VA}_{math} + 0.25 * \hat{VA}_{read} = 0.5$
  - $\tilde{VA}_{read} = 0.1 * \hat{VA}_{math} + 0.6 * \hat{VA}_{read} = 0.25$

Back

Verifying Distributional VA

- Estimate shrunken VA for each teacher for math and reading following Mulhern & Opper (2023).
- Separate VA for top and bottom performing students.
- Use 3 different splits to classify top/bottom students.
  - Top/Bottom 50% (Eastmond et al., 2024).
  - Top/Bottom 30%.
  - Top/Bottom 25%.
- I classify students based on their lagged test score in a given subject compared to the lagged test scores of all students in a given subject within a school and grade.

VA Estimates

How Correlated Are These VA Measures?

- **Goal**: Estimate correlation in latent VA for top and bottom performing students within a split.
- **Method**: Define and minimize a log-likelihood function using MLE to estimate:
  1. Latent VA for top students.
  2. Latent VA for bottom students.
  3. Variances of latent VA for top and bottom students.
  4. Covariance between latent VA measures.
  5. **Correlation between latent VA measures**

How Correlated Are These VA Measures: MLE

$$LLF \equiv -v'\Sigma v$$

$$v = \begin{bmatrix} \hat{VA}^{top}_{j,s,t} \\ \hat{VA}^{bot}_{j,s,t} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma^2_{VA_{top}} & \sigma_{VA_{top}VA_{bot}} \\ \sigma_{VA_{top}VA_{bot}} & \sigma^2_{VA_{bot}} \end{bmatrix}$$

- v: Matrix of estimated (unadjusted) value-added for top and bottom students.
- Σ: Estimated matrix of the latent VA variance-covariance matrix.
- Observed variance in VA measures is a combination of latent variance (signal) and sampling error (noise).

More Details

How Correlated Are These VA Measures: Results

Estimated Latent Correlation Among Top/Bottom Value-Added

Split	Subject	Baseline Correlation (1)	Baseline Std. Error (2)	Classroom Moments Correlation (3)	Classroom Moments Std. Error (4)
Top/Bottom 50%	Math	0.836	0.00227	0.811	0.0024
	Reading	0.682	0.00431	0.645	0.00436
Top/Bottom 30%	Math	0.590	0.00463	0.590	0.00470
	Reading	0.439	0.00672	0.313	0.00743
Top/Bottom 25%	Math	0.509	0.00531	0.531	0.00524
	Reading	0.362	0.00722	0.212	0.00795

How Correlated Are These VA Measures: MLE

$$LLF \equiv -v'\Sigma v$$

$$= \frac{1}{\det \Sigma}\left(\sigma^2_{VA_{top}}(\hat{VA}^{top}_{j,s,t})^2 + \sigma^2_{VA_{bot}}(\hat{VA}^{bot}_{j,s,t})^2 + 2\sigma_{VA_{top}VA_{bot}} * \hat{VA}^{top}_{j,s,t} * \hat{VA}^{bot}_{j,s,t}\right)$$

$$\det \Sigma = \sigma^2_{VA_{top}} \cdot \sigma^2_{VA_{bot}} - (\sigma_{VA_{top}VA_{bot}})^2$$

- Assume observed variance in VA measures is signal plus noise.
- For example:
- $\hat{\sigma}^{top}_{j,s,t} = \exp(2 * \sigma_{VA_{top}}) + \eta^2_{top}$

Back

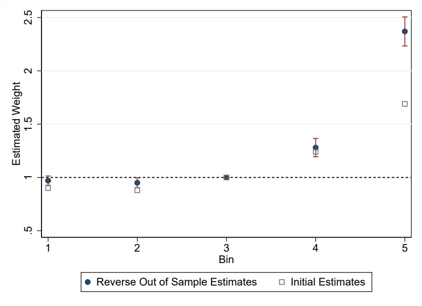
Weighted VA: Out of Sample Details (Jackson, 2018)

- **Issue**: Including students in both test-score VA and non-cognitive VA induces mechanical correlation between the two VA measures.
  - Students with unusually high/low test-score residuals generally also have unusually high/low non-cognitive residuals.
- **Fix**: Estimate pooled VA excluding the cohort used in the test-score residuals.
  - Calculate VA for all years besides year t.
  - Denote this "Out of Sample" or "Leave-year-out" VA $VA_{j,s,-t}$

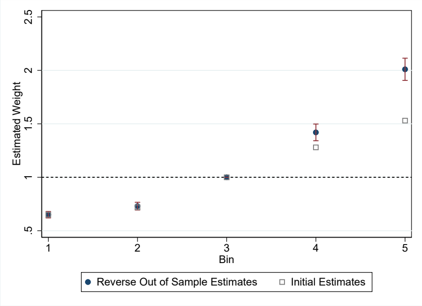
Back

Reverse Out of Sample Bin-Weights by Subject

Math



Reading

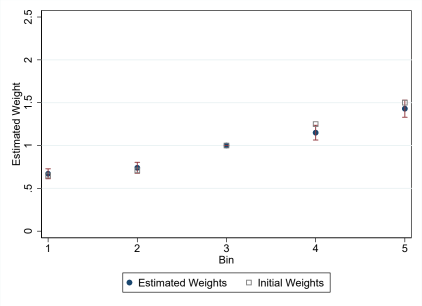


Back

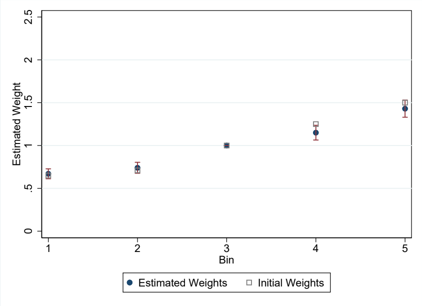
Weights and Relative Variance: Grades 3-5 Reading 5 Bins by Class Size

Reading Weight Estimates and Relative Variance by Class Size

Classes 10-19 Students



20-35 Students

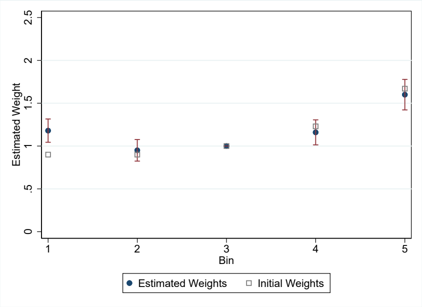


Back

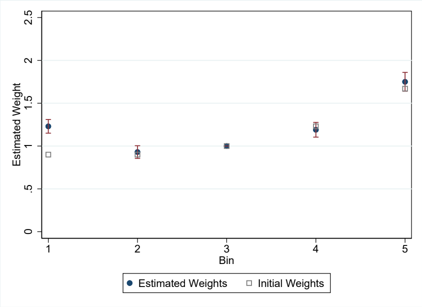
Weights and Relative Variance: Grades 3-5 Math 5 Bins by Class Size

Math Weight Estimates and Relative Variance by Class Size

Classes 10-19 Students



20-35 Students

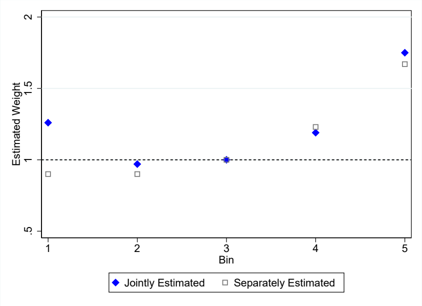


Back

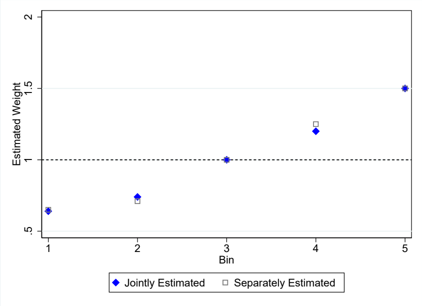
Weighted VA: Replace or Not Replace Missings

Initial Bin-Weight Estimates Replace or No Replace Missing

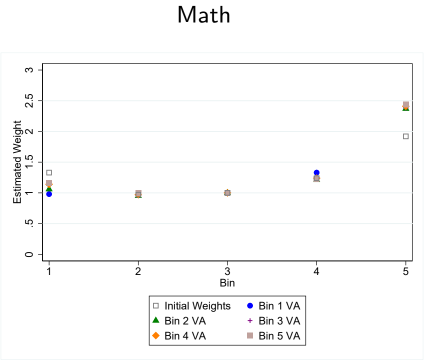
Math

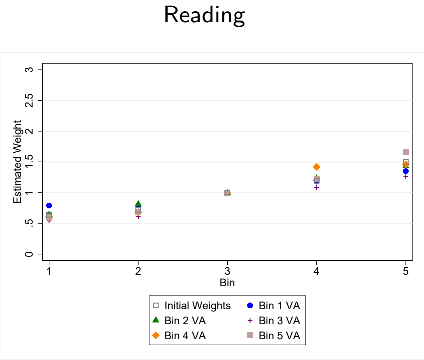


Reading

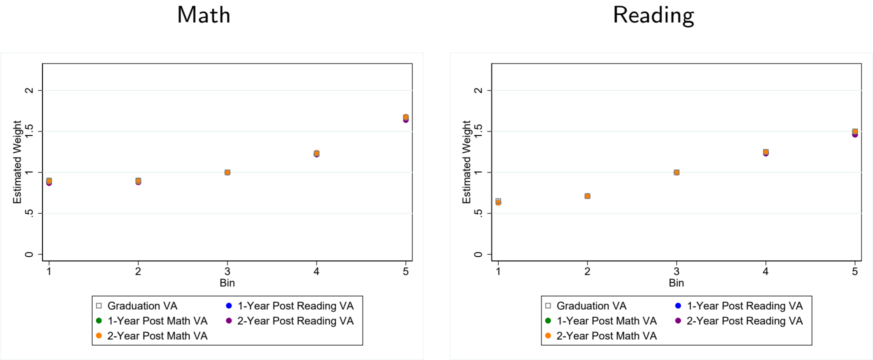


Estimated HS Grad Bin-Weights by Subject





Estimated Bin-Weights by Subject



Weighted VA: Bin-Specific VA

Estimated Weights by Subject