International Statistical Institute

WORLD STATISTICS CONGRESS 2025 THE HAGUE

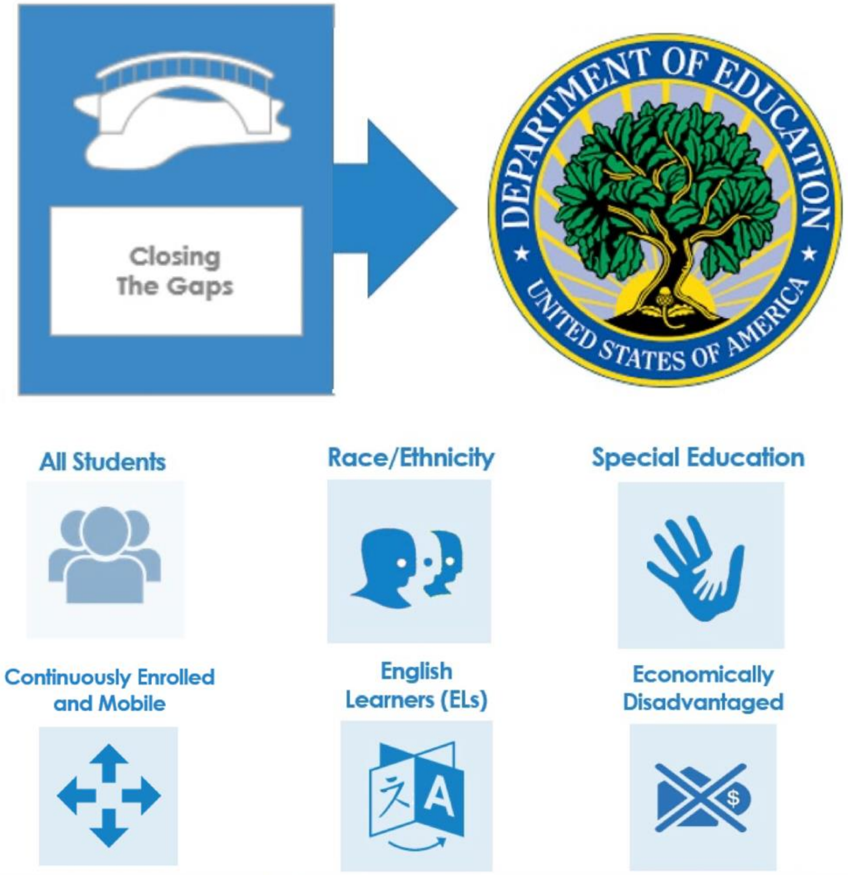
Propensity scores for coarsened data due to small-cell suppression of subgroup covariates: The case of school matching in a typical U.S. state

Joshua D. Wasserman, Michael R. Elliott, Ben B. Hansen
University of Michigan

Tuesday 7 October, 2 p.m. – 3:40 p.m.

Public test score data and school accountability in the United States

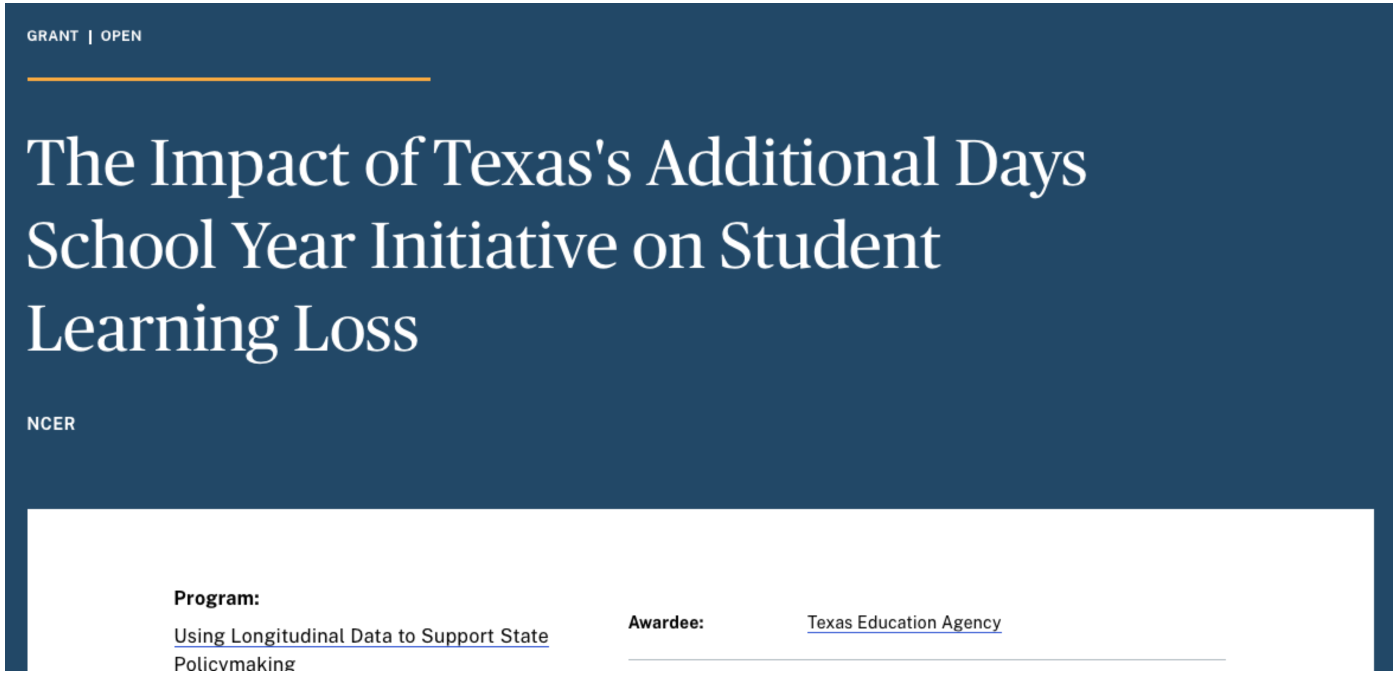
- United States mandates public release of annual summaries of state standardized test scores
- Schools in need of ``**targeted support and improvement**'' if **any of these demographic groups** lag their peers (Every Student Succeeds Act, 2015):
  - Economically disadvantaged students
  - Students from major racial or ethnic groups
  - Children with disabilities
  - English learners
- Identified schools must make plans to close the gaps with approval from their local education agency (LEA)



Source: Texas Education Agency. *Understanding the Closing the Gaps Domain*. Link.

Texas's Additional Days School Year initiative (ADSY)

- Achievement gaps grow over the summer despite their stasis during the school year (Heyns, 1978, Cooper et al., 1996)
- ADSY attempts to combat this slide by funding additional school days
- LEAs decide which schools should add days
- How to estimate the effects of ADSY on standardized test scores?


GRANT | OPEN

The Impact of Texas's Additional Days School Year Initiative on Student Learning Loss

NCER

Program: Using Longitudinal Data to Support State Policymaking

Awardee: Texas Education Agency

Source: Institute of Education Sciences. *The impact of Texas's additional days school year initiative on student learning loss.* Link.

A standard propensity score analysis

- Notation:
  - *Ti* = 1 if school *i* (*i* = 1, …, n) is in the intervention group, 0 if in the control group
  - *Wijk* = average score on test *j* (*j* = 1, …, J) among students in subgroup *k* (k = 1, …, K) at school *i*
  - ***Wi*** = (Wi_{11}, …, Wi_{JK})
  - ***Zi*** = additional confounders with treatment
- An evaluation of ADSY could use publicly available data to estimate the following propensity score model:

$$\log\left(\frac{P(T_i = 1|\mathbf{W}_i, \mathbf{Z}_i)}{1 - P(T_i = 1|\mathbf{W}_i, \mathbf{Z}_i)}\right) = \beta_0 + \mathbf{W}_i\beta_W + \mathbf{Z}_i\beta_Z \qquad (1)$$

The relationship between ``obtained'' and ``true'' test scores

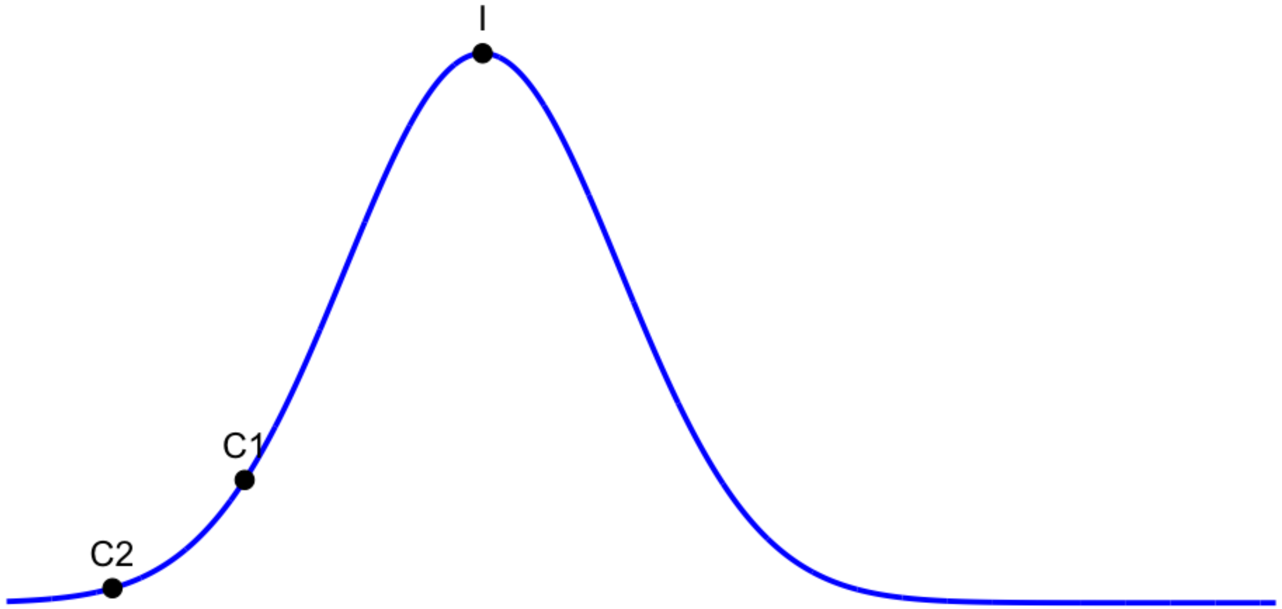
- Classical test theory models the "obtained" score *Wijk* by a simple additive measurement error model:

$$W_{ijk} = X_{ijk} + \epsilon_{ijk} \qquad (2)$$

- *Xijk* is the average ``true'' score for students in a subgroup
- ``True'' score: average score if the student took the test infinitely many times
- The ``error'' represents random test-day fluctuations due to students arriving late, sleeping poorly the night before, randomly guessing correctly, etc.
- **Balancing *Wijk* does not balance *Xijk*** (De Gil et al., 2015; Raykov, 2012)

Why estimate propensity scores conditional on ``true” test scores?

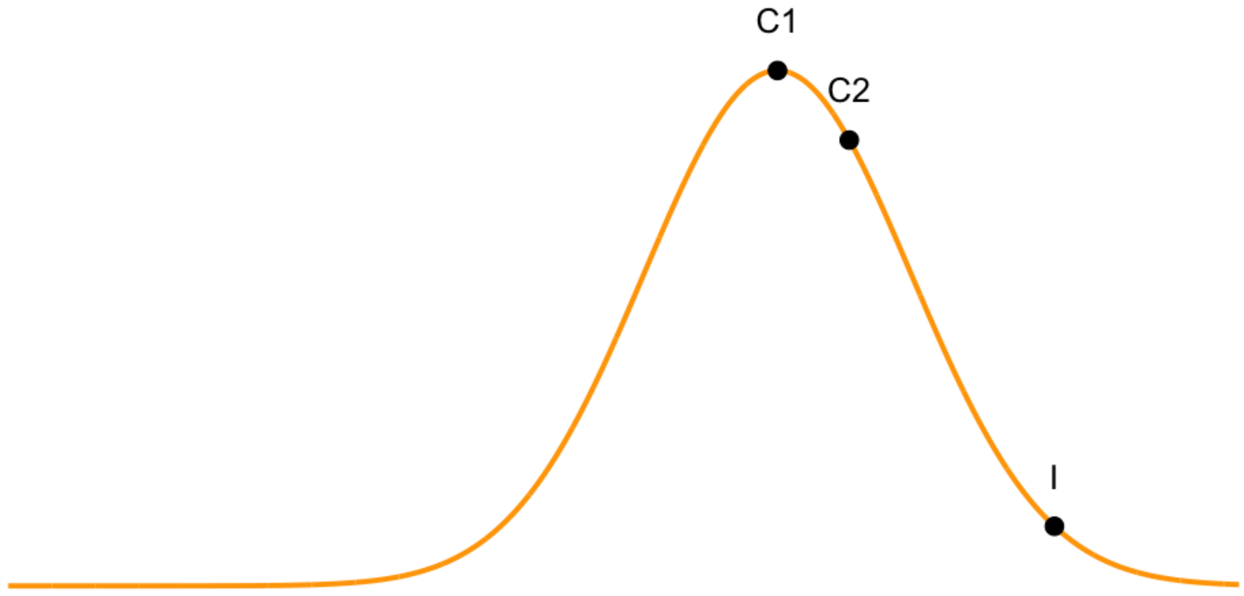
1. Schools with outlying W’s may be easier to match



One intervention and two controls with the same values of *Xijk*, but different values of *Wijk*. Curve in blue is the distribution of *Wijk* | *Xijk*.

Why estimate propensity scores conditional on ``true” test scores?

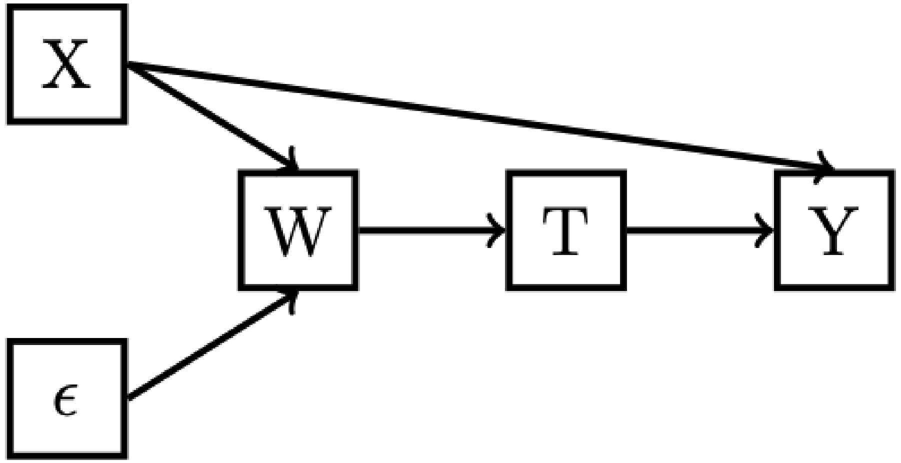
1. Schools with outlying W’s may be easier to match



One intervention and two controls with the same values of *Xijk*, but different values of *Wijk*. Curve in yellow is the distribution of *Wijk* | *Xijk*.

Why estimate propensity scores conditional on ``true" test scores?

2. Obtained scores predict intervention, but true scores predict outcomes



- Rubin (1997) notes the propensity score considers strong and weak predictors of outcomes equally
- PS estimates constructed with the goal of modeling assignment may balance predictors of intervention assignment more closely than prognostic variables

Why estimate propensity scores conditional on ``true" test scores?

3. States withhold small subgroups' obtained scores from public datasets to protect student privacy

- Propensity score estimates conditioned on obtained scores are precluded for any school with withheld scores
- At right: minimum number of students for reporting average subgroup score by state (source: Jacob et al., 2014)

State[a]	Minimum number[b]
Nebraska	10
Nevada	10
New Hampshire	10
New Jersey	30
New Mexico	10
New York	5
North Carolina	5
North Dakota	*ns*
Ohio	10
Oklahoma	5
Oregon	*ns*
Pennsylvania	40
Rhode Island	10
South Carolina	10
South Dakota	10
Tennessee	10
Texas	5

Estimating a propensity score that conditions on true test scores when they are measured with error

- Assume ***Wi*** is a ``strong'' surrogate (Lockwood and McCaffrey, 2016) in that:

$$\mathbf{X}_i \perp T_i \mid \mathbf{W}_i,\ \mathbf{Z}_i$$

- Additionally, assume:

$$\mathbf{W}_i \mid \mathbf{X}_i,\ \mathbf{Z}_i \sim \mathcal{N}(\mathbf{X}_i,\ \Sigma_i)$$

- Then P(Ti = 1 | ***Xi***, ***Zi***) is:

$$P(T_i = 1 | \mathbf{X}_i, \mathbf{Z}_i) = \int \frac{1}{1 + \exp(-\beta_0 - w\beta_W - \mathbf{Z}_i\beta_Z)} (\beta_W^T \Sigma_i \beta_W)^{-1} \phi(\frac{w\beta_W - \mathbf{X}_i\beta_W}{\beta_W^T \Sigma_i \beta_W}) dG(w\beta_W) \qquad (3)$$

- With ***Wi*** a strong surrogate *(*and the logistic regression in (1) being correct), P(Ti = 1 | ***Wi***, ***Xi***, ***Zi***) = P(Ti = 1 | ***Wi***, ***Zi***), so β's may be estimated from observed data via ML

Estimating true scores to use in propensity score estimates

- Estimate true scores using fitted values from the following hierarchical linear model (HLM):

$$W_{ijk} = \mathbf{Z}_i\gamma + \delta_{ij}^{(c)} + \delta_{ijk}^{(s)} + \epsilon_{ijk} \qquad (4)$$

- $\delta_{ij}^{(c)}$ is a random school-level effect; $\delta_{ijk}^{(s)}$ is a random subgroup-within-school effect
- $\delta_{ijk}^{(s)}$ is identified using an external estimate of the measurement error covariance matrix

Estimating true scores to use in propensity score estimates

- Agencies publish estimates of the conditional SD of the measurement error given a true score (CSEM)
- Notation:
  - σ_{ijk} = the CSEM associated with X_{ijk}
  - m_{ijk} = number of test-takers in the subgroup
- Estimate Σ_i prior to fitting the HLM by:

$$\Sigma_i = \begin{pmatrix} \frac{\sigma^2_{i11}}{m_{i11}} & 0 & \dots & 0 \\ 0 & \frac{\sigma^2_{i12}}{m_{i12}} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \frac{\sigma^2_{iJK}}{m_{iJK}} \end{pmatrix}$$

Table B.3.1. Spring 2024 STAAR Grades 3–5 Mathematics Conditional Standard Error of Measurement for Scale Scores

	Grade 3		Grade 4		Grade 5	
Raw	SS	CSEM	SS	CSEM	SS	CSEM
0	860		910		1000	
1	934	133	1025	133	1087	133
2	1031	96	1121	96	1182	96
3	1089	80	1179	79	1240	79
4	1132	70	1221	70	1283	70
5	1166	64	1255	63	1317	63
6	1195	59	1284	59	1345	59
7	1221	56	1309	55	1370	55
8	1244	53	1331	53	1392	52
9	1265	51	1351	50	1412	50
10	1284	50	1370	49	1431	48
11	1303	48	1388	47	1448	47
12	1321	48	1405	46	1464	46
13	1338	47	1421	45	1480	45
14	1354	46	1437	45	1495	44

Source: Texas Education Agency. *Technical Digest 2023-2024*. Link.

Results from a simulation study

- Compare three sets of PS estimates for full matching within calipers:
  - Proposed maximum likelihood (ML) estimates
  - Regression calibration (RC) estimates
  - Estimates from the ``Naive'' logistic regression on the error-prone subgroup scores
- Assess:
  - % of intervention schools retained in matching
  - Balance on ***Xi*** from matching
  - Bias and RMSE of intervention effect estimates

Results from a simulation study

Subgroup Sizes	Caliper Width	ML	RC	Naive
All large	0.5	11	12	16
	0.7	9	10	13
	1	7	8	10
2 moderate, 2 small	0.5	4	8	28
	0.7	3	7	25
	1	2	5	21

Table 1. Average % of intervention schools for which full matching within calipers fails to find a match. Notes. Caliper widths reported on the logit scale.

Results from a simulation study

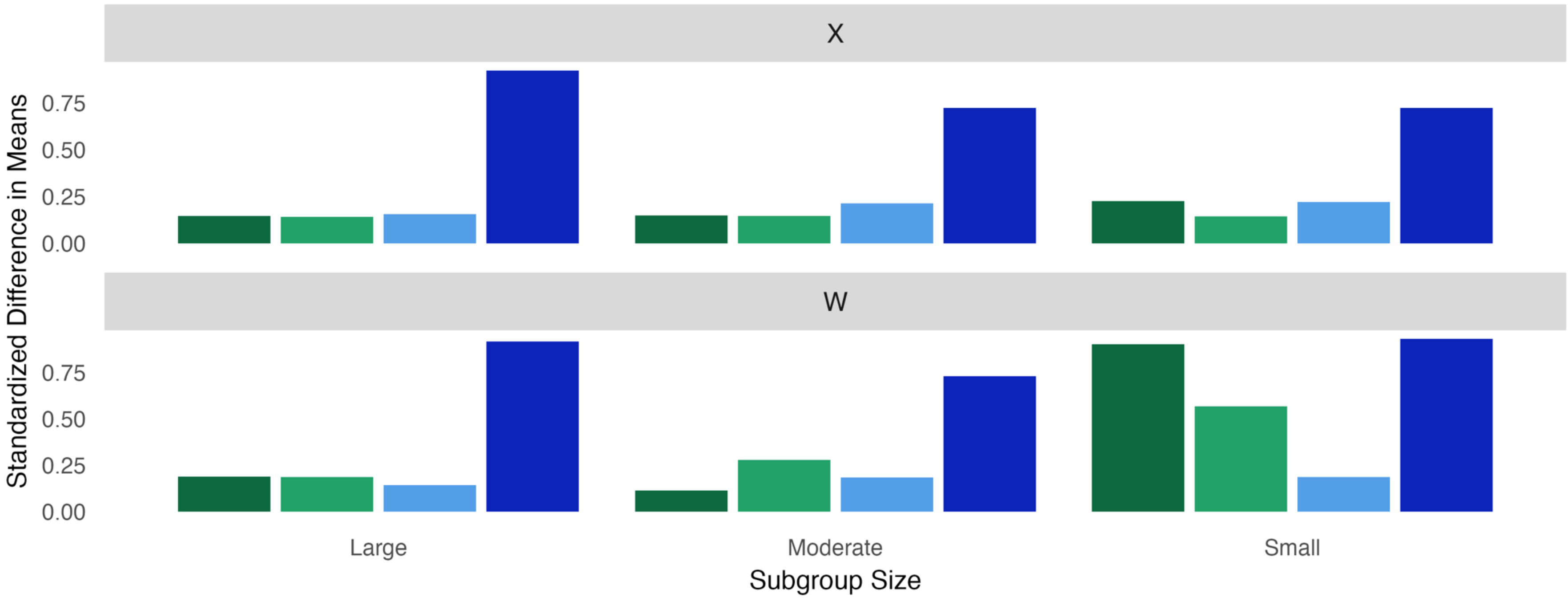


Figure 1. Standardized difference in means of ***Xi*** and ***Wi*** between intervention and control groups. Sample matched on ML PS estimates in dark green; on RC estimates in light green; on naive estimates in sky blue; and without matching in dark blue.

Results from a simulation study

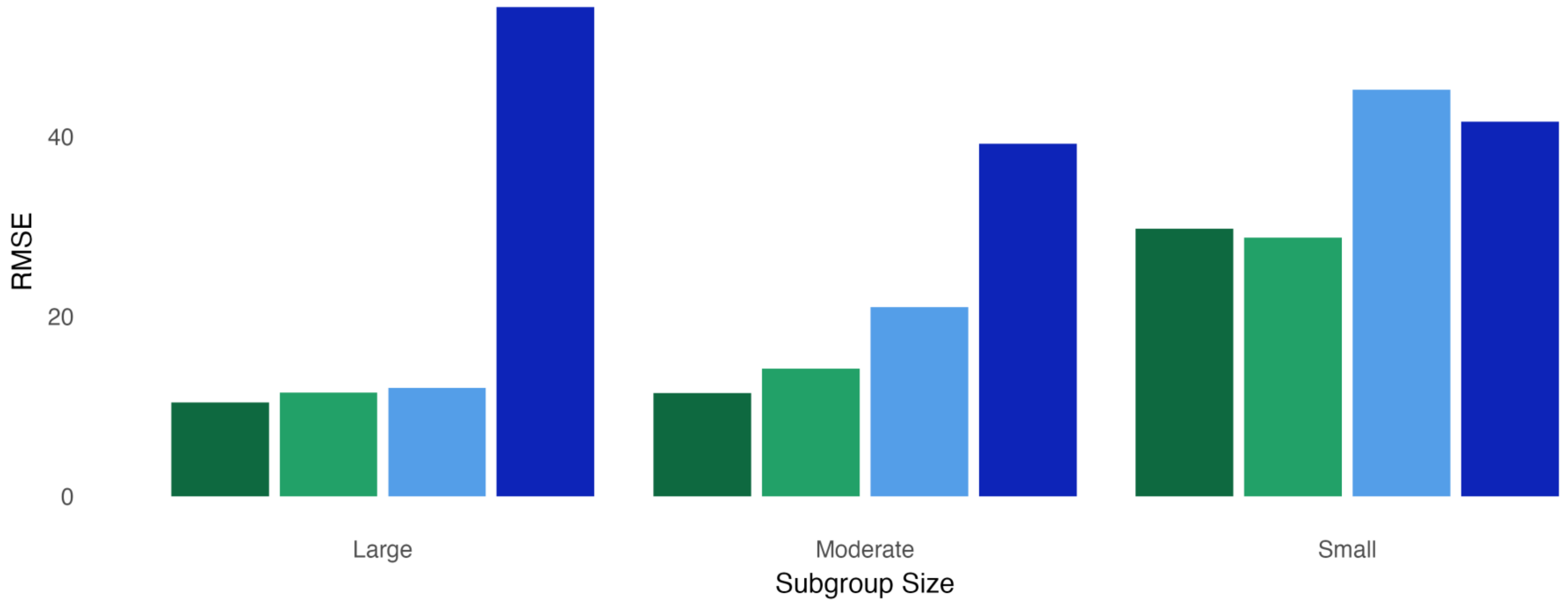


Figure 2. RMSE in matching estimators. Estimator from matching on ML PS estimates in dark green; on RC estimates in light green; on naive estimates in sky blue; and without matching in dark blue.

Using PS estimates to form a matched comparison group to ADSY schools

- 125 public schools in the 2020-2021 added days prior to spring 2021 testing
- Form a matched comparison group of schools with similar average true scores for students in subgroups of different sizes
    - Median school that year enrolled 233 Hispanic/Latino and 27 Black or African-American students
- Challenges:
    1. No testing in spring 2020 due to COVID-19 shutdown
        - At least one—sometimes two—pretests needed to remove selection bias (Hallberg et al., 2018)
        - Use what's available (spring 2019 testing) as an exercise in matching on proposed PS estimates
    2. TEA withholds average test scores from publicly available data for subgroups with <5 students
        - When ***Wi*** is not observed, ML and RC still produce PS estimates; naive approach does not
        - Apply ML and RC to publicly available data as is
        - Apply naive method to dataset that uses restricted—access student-level data to impute missing averages, giving it an oracle-type boost

Using PS estimates to form a matched comparison group to ADSY schools

Subgroup	Grade	Subject	ML	RC	Naive/ Complete	Unmatched
Hispanic or Latino	3	M	1.17	0.11	1.95	20.13
Black or African-American	3	M	4.91	16.26	4.05	9.10
Hispanic or Latino	3	R	0.06	2.94	0.69	2.94
Black or African-American	3	R	6.04	19.29	9.11	2.33
Hispanic or Latino	4	M	3.88	1.28	5.55	4.07
Black or African-American	4	M	3.40	1.10	1.08	9.00
Hispanic or Latino	4	R	1.18	0.73	3.18	3.37
Black or African-American	4	R	7.11	3.55	3.18	12.89
Hispanic or Latino	5	M	5.44	1.96	0.75	16.25
Black or African-American	5	M	0.71	0.84	2.79	2.02
Hispanic or Latino	5	R	4.67	4.60	7.64	11.59
Black or African-American	5	R	0.85	2.78	7.05	2.68
Overall average			**3.28**	4.62	3.92	8.03

Table 3. Standardized differences in ADSY and matched control means of $\widehat{\mathbf{X}}_i = (\widehat{X}_{i11}, \ldots, \widehat{X}_{iJK})$.

The take-home consideration when estimating propensity scores with public test score data:

Drown out the noise!

References

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achivement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*(3), 227–268. https://doi.org/10.3102/00346543066003227

De Gil, P. R., Bellara, A. P., Lanehart, R. E., Lee, R. S., Kim, E. S., & Kromrey, J. D. (2015). How do propensity score methods measure up in the presence of measurement error? A Monte Carlo study. *Multivariate Behavioral Research, 50*(5), 520–532. https://doi.org/10.1080/00273171.2015.1022643

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). https://congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Heyns, B. (1978). *Summer Learning and the Effects of Schooling*. Academic Press.

Hallberg, K., Cook, T.D., Steiner, P.M., Clark, M.H. (2018). Pretest measures of the study outcome and the elimination of selection bias: Evidence from three within study comparisons. *Prevention Science, 19*(3), 274—283. https://doi.org/10.1007/s11121-016-0732-6

References, cont.

Jacob, R. T., Goddard, R. D., & Kim, E. S. (2014). Assessing the use of aggregate data in the evaluation of school-based interventions: Implications for evaluation research and state policy regarding public-use data. *Educational Evaluation and Policy Analysis, 36*(1), 44–66. https://doi.org/10.3102/0162373713485814

Lockwood, J. R., & McCaffrey, D. F. (2016). Matching and weighting with functions of error-prone covariates for causal inference. *Journal of the American Statistical Association, 111*(516), 1831–1839. https://doi.org/10.1080/01621459.2015.1122601

Monahan, J. F., & Stefanski, L. A. (1992). Normal scale mixture approximations to F*(z) and computation of the logistic-normal integral. In N. Balakrishnan (Ed.), *Handbook of the Logistic Distribution* (1st ed., pp. 529–540). Marcel Dekker, Inc.

Raykov, T. (2012). Propensity score analysis with fallible covariates: A note on a latent variable modeling approach. *Educational and Psychological Measurement, 72*(5), 715–733. https://doi.org/10.1177/0013164412440999

References, cont.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*(8_Part_2), 757–763. https://doi.org/10.7326/0003-4819-127-8_Part_2-199710151-00064

Estimating a propensity score that conditions on true test scores when they are measured with error

Proposal A: Regression calibration (RC)

- Procedure:
  1. Separately for each test, fit the following hierarchical linear model:

$$W_{ijk} = \mathbf{Z}_i\gamma + \delta_{ij}^{(c)} + \delta_{ijk}^{(s)} + \epsilon_{ijk} \qquad (3)$$

    - $\delta ij^{(c)}$ is a random school-level effect; $\delta ijk^{(s)}$ is a random subgroup-within-school effect

  2. Obtain fitted values $\widehat{\mathbf{X}}_i = (\widehat{X}_{i11}, \ldots, \widehat{X}_{iJK})$
  3. Fit the logistic regression:

$$\log\left(\frac{P(T_i = 1|\widehat{\mathbf{X}}_i, \mathbf{Z}_i)}{1 - P(T_i = 1|\widehat{\mathbf{X}}_i, \mathbf{Z}_i)}\right) = \beta_0 + \widehat{\mathbf{X}}_i\beta_{\widehat{X}} + \mathbf{Z}_i\beta_Z \qquad (4)$$

- This closely approximates the regression on ***Xi*** and ***Zi*** (Carroll et al., 2006), but the theory justifying RC assumes ***Wi*** is a \`\`weak'' surrogate (Lockwood and McCaffrey, 2016). Symbolically:

$$\mathbf{W}_i \perp T_i \mid \mathbf{X}_i,\ \mathbf{Z}_i$$

Estimating a propensity score that conditions on true test scores when they are measured with error

Proposal B: Maximum likelihood (ML)

- Procedure:
  1. Fit the hierarchical linear models from step 1 of RC
  2. Obtain the fitted values from step 2 of RC
  3. Fit the logistic regression:

$$\log\left(\frac{P(T_i = 1|\mathbf{W}_i, \mathbf{Z}_i)}{1 - P(T_i = 1|\mathbf{W}_i, \mathbf{Z}_i)}\right) = \beta_0 + \mathbf{W}_i\beta_W + \mathbf{Z}_i\beta_Z \qquad (6)$$

  4. Use the fitted values and estimated coefficients from steps 2 and 3 in the approximation to (3) proposed in Monahan and Stefanski (1992):

$$P(T_i|\mathbf{X}_i, \mathbf{Z}_i) \approx \sum_{t=1}^{3} p_t \cdot \Phi\left(\frac{s_t(\hat{\beta}_0 + \widehat{\mathbf{X}}_i\hat{\beta}_W + \mathbf{Z}_i\hat{\beta}_Z)}{\sqrt{1 + s_t^2\hat{\beta}_W^T\widehat{\Sigma}_i\hat{\beta}_W}}\right) \qquad (7)$$

Results from a simulation study

- Data generation [1 assessment (J = 1), 4 mutually exclusive student subgroups (K = 4)]:
  1. Simulate ***Zi***'s representing student body demographic summaries
  2. Simulate Gaussian $\delta ij^{(c)}$'s and $\delta ijk^{(s)}$'s; mean = 0, SD's = $\boldsymbol{\tau}^{(c)}$ and $\boldsymbol{\tau}^{(s)}$, respectively
  3. Form $Xijk = \boldsymbol{Zi}\gamma + \delta ij^{(c)} + \delta ijk^{(s)}$
  4. Simulate *mijk*'s
     - Setting 1: all *mijk* simulated from a distribution for a large subgroup; diagonal elements of Σi are small
     - Setting 2: 2 *mijk* simulated from a distribution a moderate subgroup, 2 simulated from a distribution for a small subgroup
  5. Simulate Gaussian *Wijk*'s conditional on *Xijk;* mean = *Xijk*, SD = σ
  6. Simulate *Ti* as Bernoulli with P(*Ti* = 1 | ***Wi***, ***Zi***) logistic in ***Wi*** and ***Zi***
  7. Simulate Gaussian *Yijk*'s conditional on *Xijk;* mean and SD the same as for *Wijk* | *Xijk*

Using PS estimates to form a matched comparison group to ADSY schools

Subgroup	Subject	PS Estimator	Estimate (SE)	T-stat.	Adj. p-value
White	M	ML	12.14 (8.25)	1.47	0.45
		RC	14.19 (8.65)	1.64	0.34
		Naive	12.67 (8.24)	1.54	0.40
Asian	M	ML	42.78*** (11.55)	3.71	$8.58 \cdot 10^{-4}$
		RC	47.30** (13.56)	3.49	$1.96 \cdot 10^{-3}$
		Naive	43.92** (12.38)	3.55	$1.56 \cdot 10^{-3}$
White	R	ML	6.16 (7.09)	0.87	0.85
		RC	8.08 (6.50)	1.24	0.61
		Naive	2.10 (7.04)	0.30	1.00
Asian	R	ML	41.38** (12.75)	3.24	$4.73 \cdot 10^{-3}$
		RC	42.22** (13.94)	3.03	$9.82 \cdot 10^{-3}$
		Naive	41.08** (13.49)	3.05	$9.27 \cdot 10^{-3}$

Table 4. Effect estimates from a placebo test. Notes. Effects measured using average pretest scores for subgroups whose scores were not included in PS estimation. P-values have been adjusted using a max-T correction.