Combining Experimental and Observational Data to Estimate Treatment Effects

Susan Athey, Raj Chetty, & Guido Imbens

IAAE Invited Session

January 5th, 2021, ASSA Meetings

How can we systematically exploit experimental data to assist in answering questions that they cannot directly answer?

- Methods for doing so will make experiments more valuable by extending the value beyond the narrow questions they were intended for.

- Methods for doing so will make observational studies more credible by grounding them using experimental data.

What is average effect of small class size on 8th grade test scores?

	Project STAR (Experimental)		New York (Observational)	
Grade	3rd	8th	3rd	8th
Mean Controls	0.011	?	0.157	0.155
(regular class)	(0.015)	?	(0.001)	(0.001)
Mean Treated	0.212	?	0.019	0.037
(small class)	(0.025)	?	(0.001)	(0.002)
Difference	0.201	?	−0.138	−0.118
	(0.029)	?	(0.002)	(0.002)

Starting Point

- -0.118 is not credible estimate for causal effect of small class size on 8th grade scores

- Reason is that -0.138 (estimate for third grade for New York) is so different from 0.201 (estimate for third grade from experiment).

- Observational Sample has low internal validity

Question

How do we use the experimental data to improve the observational estimate?

Simple and naive ()difference-in-differences) approach: take observational study estimate for 8th grade and subtract estimated bias for 3rd grade score:

$$\widehat{\tau}^{8,NY}_{\text{adj}} \stackrel{?}{=} \widehat{\tau}^{8,NY} - \left(\widehat{\tau}^{3,NY} - \widehat{\tau}^{3,PS}\right) = -0.118 - \left(-0.138 - 0.201\right)$$

Two Problems

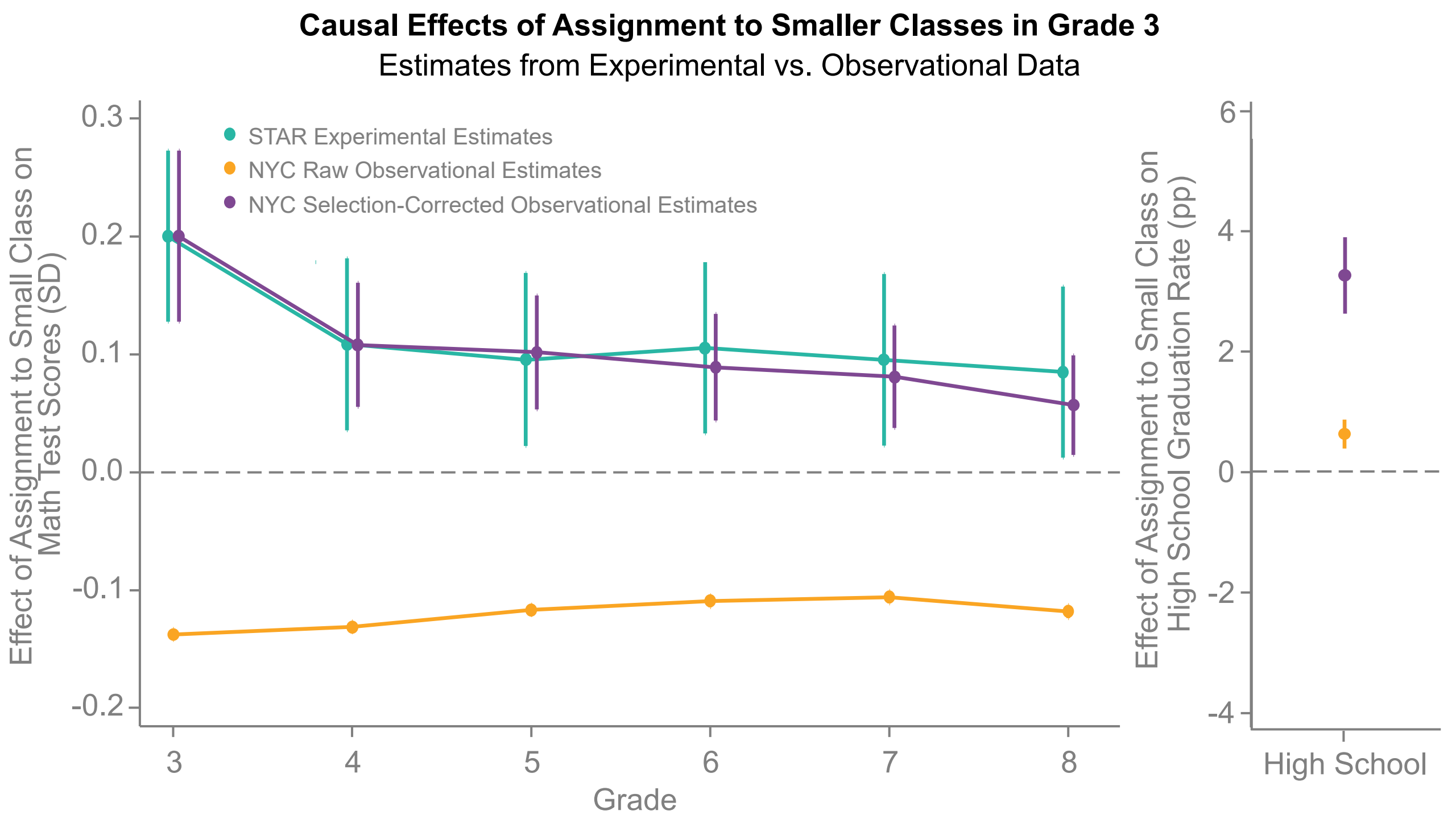
- This can only work if the primary and secondary outcome are measured on the same scale.

- Not clear what to do with multiple secondary outcomes.

Here: general approach to adjustment that allows for different scale and multiple secondary outcomes.

Preview of Results

	Project STAR (Experimental)		New York (Observational)	
Grade	3rd	8th	3rd	8th
Mean Controls (regular class)	0.011 (0.015)		0.157 (0.001)	0.155 (0.001)
Mean Treated (small class)	0.212 (0.025)		0.019 (0.001)	0.037 (0.002)
Difference	0.201	(0.085)	−0.138	−0.118
Adjusted			**0.201**	**0.057**


Causal Effects of Assignment to Smaller Classes in Grade 3
Estimates from Experimental vs. Observational Data
STAR Experimental Estimates
NYC Raw Observational Estimates
NYC Selection-Corrected Observational Estimates
Effect of Assignment to Small Class on Math Test Scores (SD)
0.3
0.2
0.1
0.0
-0.1
-0.2
3
4
5
6
7
8
Grade
Effect of Assignment to Small Class on High School Graduation Rate (pp)
6
4
2
0
-2
-4
High School

Set Up in Current Paper

Observation Scheme: ✓ is observed, ? is missing

Units	Sample G_i	Treatm. W_i	Primary Outc. Y_i^8	Secondary Outc. Y_i^3	Pretreat Var X_i
1 to N_{PS}	PS	✓	?	✓	✓
$N_{\text{PS}}+1$ to N	NY	✓	✓	✓	✓

Setting also considered in Roseman et al (2018, 2020) and Kallus and Mao (2020).

Notation

$G_i \in \{\text{PS}, \text{NY}\}$ Sample/Group Indicator, Project Star or New York

$W_i \in \{0, 1\}$ Treatment Indicator, Regular Class or Small Class

$Y_i^8(0), Y_i^8(1)$ Potential Outcomes for Primary Outcome, 8th grade test score, $Y_i^8 = Y_i^8(W_i)$ Realized Value

$Y_i^3(0), Y_i^3(1)$ Potential Outcomes for Secondary Outcome, 3rd grade test score, $Y_i^3 = Y_i^3(W_i)$ Realized Value

$\tau = \mathbb{E}[Y_i^8(1) - Y_i^8(0) | G_i = \text{NY}]$ Estimand: average effect of treatment on primary outcome in observational study.

2. Maintained Assumptions

Assumption 1 (External Validity of the Observational Study) *The observational sample is a random sample of the population of interest.*

(definitional)

Assumption 2 (Internal Validity of the Experimental Sample) *For* $w = 0, 1$,

$$W_i \perp\!\!\!\perp \left(Y_i^8(w), Y_i^3(w)\right) \mid G_i = \text{PS}.$$

(satisfied by design)

Assumption 3 (Conditional External Validity) *The experimental study has conditional external validity if*

$$G_i \perp\!\!\!\perp \left(Y_i^8(0), Y_i^8(1), Y_i^3(0), Y_i^3(1)\right).$$

Strong assumption: why is -0.138 (estimate on 3rd grade for New York) different from 0.201 (Project Star estimate)?

- Because of unobserved confounders in New York sample,

- **NOT** because New York population is different from Project Star Population (possibly after adjusting for covariates).

• We do **NOT** want to assume unconfoundedness in the observational sample:

Assumption 4 (Unconfoundedness in the Observational Sample)
For $w = 0, 1$,

$$W_i \perp\!\!\!\perp \left(Y_i^3(w), Y_i^8(w)\right) \Big| G_i = \text{NY}$$

(That would solve problems, but lead to testable restrictions that are easily rejected with the current data.)

3. Latent Unconfoundedness

Instead we assume:

Assumption 5 (Latent Unconfoundedness)

$$W_i \perp\!\!\!\perp Y_i^8(w) \mid Y_i^3(w), G_i = \mathsf{NY} \quad \forall w.$$

Tricky assumption: we cannot use this on its own:

$$W_i \perp\!\!\!\perp Y_i^8(0) \mid Y_i^3(0), G_i = \text{NY}$$

We can estimate

$$\mathbb{E}\Big[Y_i^8(0) \mid Y_i^3(0), W_i = 0, G_i = \text{NY}\Big]$$

but we cannot average this over the marginal dist of $Y_i^3(0)$ because to see $Y_i^3(0)$ we need to condition on $W_i = 0$.

But, from the experimental sample we can estimate the distribution of $Y_i^3(0)$ and that allows us to do the averaging.

Main Result

Theorem 1 *Suppose that Assumptions 1-3 and 5 hold, so that the experimental study is unconfounded and has conditional external validity, and the observational study has latent unconfoundedness.*

Then the average effect of the treatment on the primary outcome in the observational study is point-identified.

4. Estimation Strategy: Control Function Approach

• Estimate relation (cumulative distribution function) between secondary outcome and treatment and pre-treatment variables in experimental sample.

$$F_{Y^3|W,G}(y|w, \text{PS}) = \text{pr}\left(Y_i^3 \leq y \middle| W_i = w, G_i = \text{PS}\right)$$

• Evaluate this distribution function (this is the **control function**):

$$\eta_i = F_{Y^3|W,G}\left(Y_i^3 \middle| W_i, \text{PS}\right)$$

• By construction:

$$\eta_i | G_i = \text{PS} \sim U[0, 1]$$

If assignment is unconfounded in the observational sample, then

$$\eta_i | G_i = \text{PS} \sim U[0, 1]$$

Deviations from uniformity of distribution of η_i captures violations of unconfoundedness.

Latent Unconfoundedness implies that it captures **all** violations of unconfoundedness

Intuition

η_i measures where individual i is in the distribution of $Y_i(W_i)$ in experimental sample.

Suppose that the average value of η_i in the observational study is larger than 0.5 (evidence of selection). Then η_i must at least be correlated with unobserved confounder.

Thus: we should compare treated and control individuals with the same value of η_i.

Latent unconfoundedness

$$W_i \perp\!\!\!\perp Y_i^8(w) \mid Y_i^3(w), G_i = \text{NY}$$

(in combination with maintained assumptions) implies that:

$$W_i \perp\!\!\!\perp Y_i^8(w) \mid \eta_i, G_i = \text{NY}$$

Thus:

- Estimate the average effect of the treatment on the primary outcome adjusting for the control variable.

6. Conclusion

• Principled way to combine experimental data to adjust flawed estimates based on observational data alone.

• Just one case where we leverage strengths of observational and experimental data.

References

Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely (No. w26463). National Bureau of Economic Research.

Athey, S., Chetty, R., & Imbens, G. (2020). Combining Experimental and Observational Data to Estimate Treatment Effects on Long Term Outcomes. arXiv preprint arXiv:2006.09676.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological

research: Conceptual, strategic, and statistical considerations. Journal of personality and social psychology, 51(6), 1173.

Kallus, N., & Mao, X. (2020). On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. arXiv preprint arXiv:2003.12408.

Kallus, N., Puli, A. M., & Shalit, U. (2018). Removing hidden confounding by experimental grounding. In Advances in neural information processing systems (pp. 10888-10897).

Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). John Wiley & Sons.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. Statistics in medicine, 8(4), 431-440.

Rosenman, E., Owen, A. B., Baiocchi, M., & Banack, H. (2018). Propensity Score Methods for Merging Observational and Experimental Datasets. arXiv preprint arXiv:1804.07863.

Rosenman, E., Basse, G., Owen, A., & Baiocchi, M. (2020). Combining Observational and Experimental Datasets Using Shrinkage Estimators. arXiv preprint arXiv:2002.06708.

VanderWeele, T. (2015). Explanation in causal inference: methods for mediation and interaction. Oxford University Press.