American Economic Review: Papers & Proceedings 2015, 105(5): 502–507
http://dx.doi.org/10.1257/aer.p20151025

Evaluating Econometric Evaluations of Post-Secondary Aid[†]

By Josh Angrist, David Autor, Sally Hudson, and Amanda Pallais*

The question of whether and how financial aid affects college enrollment remains central in discussions of higher education policy. Most econometric investigations of this question identify causal effects using non-experimental strategies such as covariate conditioning, differences-in-differences panel methods, and regression discontinuity (RD) designs. The resulting empirical analyses have produced a wide range of estimates, perhaps reflecting the diversity of the models and assumptions used in this work (see research surveyed in Deming and Dynarski 2009).

In an effort to produce credible and robust estimates of the causal effects of aid on post-secondary outcomes, we've worked with the Susan Thompson Buffett Foundation (STBF) to conduct a randomized evaluation of STBF's longstanding scholarship program for Nebraska high school seniors. Findings to date are detailed in our working paper Angrist et al. (2014).[1] Our focus here is methodological: in the spirit of LaLonde's (1986) pioneering comparison of job training effects from randomized and non-experimental analyses, and the recent Wing and Cook (2013) within-study evaluation of a RD design, we compare our experimental results with covariate-controlled estimates from a pre-experimental cohort and with RD estimates from the experimental sample. The results show covariates do little to mitigate selection bias, but RD estimates that exploit institutional idiosyncrasies in the award process come close to an appropriately-defined experimental benchmark. On the other hand, the RD estimates are sensitive to controls for the running variable.

I. The Buffett Scholarship

The STBF-eligible applicant pool contains Nebraska-resident high school seniors and graduates of in-state high schools who have not yet attended college. Awardees are selected on the basis of financial need, college readiness, and a review of personal statements and reference letters. These award winners, called Buffett Scholars, receive grants worth full tuition and fees that can be used at any Nebraska public college. Grants are renewable for up to five years, so that students attending the most-expensive in-state public college can receive more than $60,000 in total. Buffett Scholars who attend one of the three University of Nebraska campuses also participate in Learning Communities (LCs), an academic services intervention.

Our experiment has randomly awarded more than 2,000 scholarships since 2012. Prior to the experiment, applicants were chosen by individual reviewers without a formal ranking procedure. Starting in 2012, reviewers scored applicants using a common rubric. The highest scoring applicants (301 out of 1,430 eligible applicants in 2012) were guaranteed awards, while the lowest-scoring (127) were removed from consideration. The remainder (1,003) were subject to random assignment, with award rates varying by students' intended colleges,

* Angrist: Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 and NBER (e-mail: angrist@mit.edu); Autor: Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 and NBER (e-mail: dautor@mit.edu); Hudson: Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 (e-mail: slhudson@mit.edu); Pallais: Harvard University, Cambridge, MA 02138 and NBER (e-mail: apallais@fas.harvard.edu). We thank Sydnee Caldwell and Olivia Kim for superb research assistance and the staff of the Susan Thompson Buffett Foundation (STBF) for their expert input. We acknowledge financial support from STBF and the MIT SEII seed fund. The views expressed here are those of the authors alone and do not necessarily reflect those of the institutions or funders involved with this work.

† Go to http://dx.doi.org/10.1257/aer.p20151025 to visit the article page for additional materials and author disclosure statement(s).

[1] Briefly, these results show that scholarship offers increased total financial aid received substantially. This in turn generated modest gains in initial enrollment, with a marked treatment-induced shift from two- to four-year campuses, an effect that increases in the second post-program year.

which we call strata. The analysis reported here compares covariate-controlled estimates from the 2011 non-experimental cohort to the 2012 experimental results. In addition, we use the 2012 threshold for guaranteed awards to construct RD estimates from the experimental data.[2]

II. Econometric Methods

Let D indicate aid offers, and let Y_1 and Y_0 denote potential outcomes in treated and untreated states, respectively. The observed outcome is determined by

$$y = Y_0 + (Y_1 - Y_0)D.$$

We wish to estimate the average treatment effect (ATE), defined as

$$\delta \equiv E[Y_1 - Y_0].$$

Random assignment within strata (that is, conditional random assignment (CRA)) implies that potential outcomes in treated and untreated states are mean-independent of treatment conditional on stratum, denoted by x:

$$(1) \qquad E[Y_j|D,x] = E[Y_j|x];\ j = 0,1.$$

Given CRA, simple treatment-control contrasts within strata,

$$\Delta_x = E[y|x,D = 1] - E[y|x,D = 0],$$

provide unbiased estimates of the strata-specific average treatment effects, $\delta_x \equiv E[Y_1 - Y_0|x]$. Averaging these conditional contrasts across strata generates a matching estimand for ATE:

$$\delta = E[\delta_x] = E[\Delta_x].$$

We're also interested in regression estimates computed by fitting

$$(2) \qquad y = \alpha_x + \rho D + \eta,$$

where α_x is a fixed-effect for each stratum and η is the regression error. Angrist (1998) shows that the regression estimand in such models is also an average of conditional treatment effects, specifically

$$\rho = E\left\{\frac{\lambda_x(1-\lambda_x)}{E[\lambda_x(1-\lambda_x)]}\delta_x\right\},$$

where $\lambda_x \equiv E[D|x]$ is the conditional probability of treatment within strata. Ordinary least squares (OLS) therefore estimates a variance-weighted average of strata-specific average causal effects.

A natural starting point for identification without random assignment is a generic conditional independence assumption (CIA). The CIA swaps a vector of controls, denoted $\mathbf{X}$, for x in the CRA, equation (1). The CIA is a strong assumption, not guaranteed to hold in non-randomized studies. We use the CIA to compute non-experimental estimates of treatment effects in the pre-experimental cohort in three ways: OLS, swapping α_X for α_x in equation (2); propensity score weighting as in Hirano, Imbens, and Ridder (2003); and a hybrid regression/reweighting procedure suggested by Kline (2011).

The propensity score weighting (HIR) estimator builds on the fact that the CIA implies

$$E[Y_1|\mathbf{X}] = E\left[\frac{yD}{\lambda_X}|\,\mathbf{X}\right]$$

$$E[Y_0|\mathbf{X}] = E\left[\frac{y(1-D)}{1-\lambda_X}|\,\mathbf{X}\right],$$

where $\lambda_X \equiv E[D|\mathbf{X}]$ is the conditional-on-covariates probability of treatment. Bringing these expressions inside a single expectation and over a common denominator, the average treatment effect is

$$\delta = E\left\{\frac{y[D-\lambda_X]}{\lambda_X[1-\lambda_X]}\right\}.$$

Our estimates based on this formula use logit to model λ_X.

Kline's estimator begins with linear models for conditional means:

$$E[Y_0|\mathbf{X}] = E[y|\mathbf{X},D = 0] = \mathbf{X}'\beta_0$$

$$E[Y_1|\mathbf{X}] = E[y|\mathbf{X},D = 1] = \mathbf{X}'\beta_1.$$

[2] Scholarship application reviewers were unaware of the thresholds for random assignment when scoring applications.

Table 1—Descriptive Statistics

	Experimental sample		Observational sample		RD sample	
	Control mean (1)	Treatment − control (2)	Control mean (3)	Treatment − control (4)	Control mean (5)	Treatment − control (6)
Female	0.621	0.009	0.695	0.107***	0.647	0.039
		(0.031)		(0.030)		(0.040)
White	0.673	0.031	0.562	−0.139***	0.481	−0.140***
		(0.028)		(0.029)		(0.040)
EFC ($)	3,337	−52	1,826	−1,245***	1,832	−1,248***
		(235)		(180)		(266)
Income ($)	50,738	339	40,012	−12,605***	39,055	−8,784***
		(2,170)		(2,850)		(2,155)
GPA	3.44	0.005	3.49	0.056 **	3.53	0.061*
		(0.026)		(0.026)		(0.032)
Sample size	504	1,003	593	1,052	265	624

Notes: The experimental sample contains the 2012 treatment and control groups. The observational sample is the 2011 pre-experimental cohort. The RD sample contains applicants from the 2012 control group and guaranteed-award group who scored within four points of the guaranteed award cutoff. The running variable, (the reviewer score), ranges from 11 to 26.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

These imply

$$\delta = (\beta_1 - \beta_0)' E[\mathbf{X}].$$

Finally, we compute sharp RD estimates in the 2012 data using the STBF reviewer score, s, as our running variable. These come from a version of equation (2) estimated in the sample of applicants who scored within four points of the cutoff used for guaranteed funding. The treatment here is a dummy variable, Z, indicating reviewer scores above the cutoff. The sample to the left of the cutoff is limited to those in the randomly-assigned control group.

III. Data and Results

Random assignment of scholarship offers balanced the characteristics of treatment and control applicants in the 2012 experimental sample, as the first two columns of Table 1 show. By contrast, award winners in the pre-experimental (2011) cohort had higher high school GPAs and lower family incomes, on average, than those not awarded scholarships. This is consistent with a review process that favors both merit and financial need. We see broadly similar differences on either side of the threshold for guaranteed awards in 2012, though the income and gender differentials are smaller. These cross-threshold covariate gaps, shown in columns 5 and 6 of Table 1, compare experimental controls who scored no lower than four points below the threshold with students who were guaranteed awards and scored at most four points above the threshold. Total points ranged from 11 to 26.

Randomly-assigned Buffett scholarships increased four-year sophomore enrollment by a precisely-estimated 14.4 percentage points, an impressive effect given that only 64 percent of controls were enrolled in four-year colleges in the second follow-up year. This finding—our first experimental benchmark—appears in column 1 of Table 2. This column also shows that the experimental estimates are essentially invariant to the choice of estimator used for covariate adjustment, an expected consequence of random assignment.

Given the substantial differences between 2011 treated and control applicants, it's not surprising that treatment-control comparisons in this sample miss the experimental benchmark. Averaging across strata, the estimated treatment effect is just 0.091 in the 2011 sample, an estimate reported in column 2 of Table 2. OLS with strata fixed effects produces estimates

Table 2—Effects on Four-Year College Enrollment in Year Two

	Experimental sample (1)	Observational sample (2)	Experimental RD sample (3)	RD sample (4)
Control mean	0.639	0.708	0.685	0.685
Raw difference	0.142***	0.086***	0.107***	0.044
	(0.028)	(0.027)	(0.033)	(0.037)
Panel A. Strata-adjusted estimates				
Matching	0.144***	0.091***	0.116***	0.096***
	(0.024)	(0.023)	(0.027)	(0.031)
OLS	0.144***	0.091***	0.116***	0.099***
	(0.024)	(0.022)	(0.027)	(0.032)
Panel B. Esitmates with selection controls				
OLS	0.143***	0.094***	0.120***	0.107***
	(0.023)	(0.022)	(0.026)	(0.032)
OLS with r.v. controls				0.024
				(0.064)
HIR	0.143***	0.097*	0.119*	
	(0.054)	(0.058)	(0.063)	
Kline	0.143***	0.092***	0.119 ***	
	(0.022)	(0.021)	(0.025)	
Sample size	1,003	1,052	715	624

Notes: Samples for columns 1, 2, and 4 are defined in Table 1. The sample for column 3 includes applicants in the experimental sample who scored within four points of the guaranteed award cutoff. Estimates in panel B are from models that include linear controls for GPA, EFC, imputed family income, and dummies for gender and nonwhite race. HIR standard errors are bootstrapped with 1,000 replications.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

indistinguishable from this matching estimate. The apparent negative selection bias in the observational estimates presumably reflects STBF's focus on measures of need that are negatively correlated with post-secondary outcomes.

Many of the variables used to screen applicants appear in our data, so perhaps selection bias can be eliminated using statistical controls. As panel B of column 2 shows, however, controlling for covariates, whether by regression, propensity score matching, or Kline-reweighting, boosts the observational estimates only slightly.[3]

[3] In results not reported here, we find that reweighting to produce estimates of the treatment on the treated or the treatment on the untreated leaves the observational estimates largely unchanged and still well below the experimental benchmark. As in Angrist and Rokkanen (2012), HIR and linear reweighting estimates are similar, but the HIR estimates are much less precise.

Why does controlling for covariates move the observational estimates so little? Although covariates in the observational sample are highly imbalanced, and we would expect those listed in Table 1 to predict college enrollment, it turns out that they explain too little of the outcome variation in our data to matter much as controls. As Pischke and Schwandt (2014) note, when the covariates at hand are noisy or imperfect proxies for strong predictors of outcomes, the addition of even highly imbalanced controls can have little impact on estimated treatment effects.

A. *Within-Study RD*

Our RD analysis uses only applicants that scored within four points of the cutoff for guaranteed awards. The relevant experimental benchmark therefore compares randomized treatment and control observations that fall in

this bandwidth. The strata-adjusted experimental estimate in this high-scoring subsample, shown in column 3 of Table 2, is 0.116. The gap between this estimate and the full-sample result in column 1 is likely due to the negative merit gradient in treatment impacts documented in our working paper. Specifically, applicants who appear best-prepared for college gained the least from STBF awards.

Comparing treatment and control observations in the RD bandwidth produces a strata-adjusted (OLS) estimate of 0.099. This estimate, reported in column 4 of Table 2, is noticeably closer to its experimental benchmark than is the raw difference in column 4. Adding the full set of controls produces an OLS estimate (0.107) that closes the gap even further.

These OLS estimates come from regression models that omit the running variable, in this case, the STBF reviewer score, *s*. This omission is justified by the assumption that applicants near the treatment threshold are similar enough to eliminate secular running variable effects on outcomes. Moreover, as in Angrist and Rokkanen (2012), the inclusion of other covariates may mitigate the need for running variable controls. The STBF reviewer score is known to be a function of covariates like GPA, so conditional on these variables, scores may be as good as randomly assigned.[4]

Do running variable controls matter? In practice, we're handicapped here by the coarseness of the running variable, which offers only four points of support upon which to base implicit extrapolation of average potential outcomes across the award threshold. Indeed, as column 4 of Table 2 shows, inclusion of a linear running variable control interacted with treatment (that is, *s* and *sZ*) generates an imprecisely estimated treatment effect close to zero. This imprecision and sensitivity to running variable controls emerges in spite of the fact that the coefficients on *s* and *sZ* aren't significantly different from zero.

[4]This argument can't be applied to the 2011 sample because the 2011 review process did not rely on a clearly defined scoring rubric.

IV. Discussion

In the absence of random assignment, institutional knowledge opens the door to credible quasi-experimental research designs. This knowledge may come in the form of information on the covariates that determine treatment assignment, as in Dehejia and Wahba's (1999) influential re-examination of LaLonde (1986). In our application, covariates strongly related to treatment assignment are too weakly related to outcomes to eliminate selection bias. Other institutional features turn out to be more valuable: an RD estimate exploiting an award threshold replicates experimental findings for applicants near the cutoff (though not the overall treatment effect). A key weakness here, however, is the coarseness of the running variable. The addition of controls for the running variable and its interaction with treatment status generates an imprecise RD estimate of zero.

REFERENCES

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66 (2): 249–88.

Angrist, Joshua, David Autor, Sally Hudson, and Amanda Pallais. 2014. "Leveling Up: Early Results from a Randomized Evaluation of Post-Secondary Aid." National Bureau of Economic Research Working Paper 20800.

Angrist, Joshua, and Miikka Rokkanen. 2012. "Wanna Get Away? RD Identification Away from the Cutoff." National Bureau of Economic Research Working Paper 18662.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448): 1053–62.

Deming, David, and Susan Dynarski. 2009. "Into College, Out of Poverty? Policies to Increase the Postsecondary Attainment of the Poor." National Bureau of Economic Research Working Paper 15387.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71 (4): 1161–89.

Kline, Patrick. 2011. "Oaxaca-Blinder as a Reweighting Estimator." *American Economic Review* 101 (3): 532–37.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (4): 604–20.

Pischke, Jörn-Steffen, and Hannes Schwandt. 2014. "Poorly Measured Confounders are More Useful on the Left Than on the Right." http://econ.lse.ac.uk/staff/spischke/ec533/C_var_note.pdf (accessed January 7, 2015).

Wing, Coady, and Thomas D. Cook. 2013. "Strengthening the Regression Discontinuity Design Using Additional Design Elements: A Within-Study Comparison." *Journal of Policy Analysis and Management* 32 (4): 853–77.