Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random to Nonrandom Assignment.

William R. Shadish
University of California, Merced

The Problem

- Randomized Experiments Yield Unbiased and Consistent Effect Estimates
- But they are not always feasible or ethical
- Under what circumstances can nonrandomized experiments yield accurate estimates?

Nonrandomized Experiments

- A central hypothesis about the use of nonrandomized experiments is that their results can well-approximate results from randomized experiments
  - especially when the results of the nonrandomized experiment are appropriately adjusted by, for example, selection bias modeling or propensity score analysis.
  - I take the goal of such adjustments to be: ***to estimate what the effect would have been if the nonrandomly assigned participants had instead been randomly assigned to the same conditions and assessed on the same outcome measures***.
  - The latter is a counterfactual that cannot actually be observed
  - So how is it possible to study whether these adjustments work?

Kinds of Empirical Comparisons of Randomized to Nonrandomized Experiments

- In general, there have been three different ways to study this question:
  - Computer Simulations
  - Single Study Comparisons
  - Meta-Analytic Comparisons
- Such studies have not consistently supported the effectiveness of adjustments.
- However, these methods all provide a poor test of the adjustments, each for different reasons:

Computer Simulations

- Important method of learning about the issue, especially with
  - Increased computer power
  - User-friendly simulation programs (GAUSS)
- However, this method is of limited use because
  - The nature of selection bias in nonrandomized experiments is unknown, but
  - To use computer simulations one must program selection bias, so it is known.
  - Any results are always subject to the question of whether they have really modeled selection bias.

Single Study Comparisons

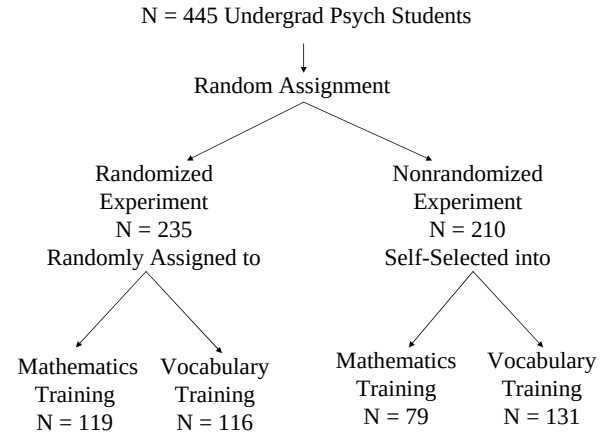
- Long history: E.g., Lalonde in the 1980s
- Widely used today: Bloom et al.; Glazerman et al
  - Start with a randomized experiment
  - Then find a nonrandomized control
  - Substitute the nonrandomized control for the randomized control.
  - Try to adjust the QE answer to see if you get the same answer
- Tends to conclude that QE cannot be made to approximate the estimates from RE (e.g., Glazerman et al.)
- Fatal Flaw: The nonrandomized control differs from the randomized control in more ways than just assignment method
- Thus it cannot provide a good test of the counterfactual: ***if the nonrandomly assigned participants had instead been randomly assigned to the same conditions and assessed on the same outcome measures***.

Meta-Analytic Comparisons

- Find large numbers of randomized and nonrandomized experiments on the same question, and compare average effect sizes (e.g., Lipsey and Wilson)
- Takes advantage of diversity over many studies to explore the role of covariates that are confounded with assignment method (e.g., different kinds of control groups; e.g., Shadish & Ragsdale).
- Sometimes yields somewhat more optimistic conclusions
- Problems:
  - One can never know for certain that one knows and adequately measures all those confounds.
  - No access to individual data to use some adjustment methods (e.g., propensity score analysis)

A New Approach: A Laboratory Analogue

- One way to control (on expectation) for such confounds is to randomize them—i.e., to randomly assign participants to being in a randomized or nonrandomized experiment in which they are otherwise treated identically.
- This also gives access to individual data so adjustments to quasi-experimental results can be tried.
- The closest of any method to testing the right question: ***if the nonrandomly assigned participants had instead been randomly assigned to the same conditions and assessed on the same outcome measures***.
- Here is the design as we implemented it:



More on the Design

- All participants pretested on a host of covariates
- Chose math and vocab training because
  - Good analogue to educational interventions
  - Relevant to college students
  - Easy to control effect size with item difficulty
  - Math phobias cause plausible selection bias
- All participants treated together without knowledge of the different conditions.
- All participants posttested on both math and vocab outcomes.

Unadjusted Results:
Effects of Math Training on Math Outcome

	Math Tng Mean	Vocab Tng Mean	Mean Diff	Absolute Bias
Unadjusted Randomized Experiment	11.35	7.16	4.19	
Unadjusted Quasi-Experiment	12.38	7.37	5.01	.82

Conclusions:

1. The effect of math training on math scores was larger when participants could self-select into math training.
2. The 4.19 point effect (out of 18 possible points) in the randomized experiment was overestimated by 19.6% (.82 points) in the nonrandomized experiment

Unadjusted Results:
Effects of Vocab Training on Vocab Outcome

	Vocab Tng Mean	Math Tng Mean	Mean Diff	Absolute Bias
Unadjusted Randomized Experiment	16.19	8.08	8.11	
Unadjusted Quasi-Experiment	16.75	7.75	9.00	.89

Conclusions:

1. The effect of vocab training on vocab scores was larger (9 of 30 points) when participants could self-select into vocab training.
2. The 8.11 point effect (out of 30 possible points) in the randomized experiment was overestimated by 11% (.89 points) in the nonrandomized experiment.

Adjustments to Quasi-Experiments

- It is no surprise that randomized and nonrandomized experiments might yield different answers.
- Can we adjust the answers?
  - Propensity Scores
  - ANCOVA
  - Structural Equation Modeling

Propensity Scores

- The conditional probability of being in the treatment or comparison group given available predictors of group membership.
- The propensity score reduces all the information in the predictors to one number.
  - This can make it easier to do matching or stratifying when there are multiple matching variables available.

Estimation of Propensity Scores in Our Data Set

- Used SPSS (MVA) to impute missing data in the covariates (EM method)
- Used stepwise logistic regression with subsequent forced entry of variables out of balance
  - For example: Math and vocabulary proxy pretests, ACT, GPA, measures of previous exposure to math courses, math anxiety, Demographics
  - But also "Big 5" personality traits (extraversion, emotional stability, agreeableness, intellect, and conscientiousness)

(Y^0, Y^1)

Two Criteria

- Balance: After PS Stratification, are T and C balanced on pretest covariates?
  - I.e., mimic a randomized experiment
  - Necessary but not sufficient because of hidden bias
- Strong Ignorability
  - If we identify and measure all covariates that are related to both treatment *Z* and potential outcomes, treatment assignment is "strongly ignorable" given **X**.
  - There is no test for this, so strong ignorability is frequently assumed without thorough justification

Balance: Rubin 2001 Criteria

Table 3. Rubin's (2001) Balance Criteria Before and After Propensity Score Stratification

Analysis	Propensity Score B	Propensity Score R	Number of Covariates with Variance Ratio ≤1/2	>1/2 and ≤4/5	>4/5 and ≤5/4	>5/4 and ≤2	>2
Before Any Adjustment	-1.13	1.51	0	2	17	6	0
After Stratification on Propensity Scores Constructed from All Covariates	-0.03	0.93	0	1	22	2	0

B = the standardized difference in the mean propensity score in the two groups (B) should be near zero.
R = the ratio of the variance of the propensity score in the two groups (R) should be near one, and
Ratios = The ratio of the variances of the covariates after adjusting for the propensity score must be close to one

Mathematics Outcome

	Mean Difference (standard error)	Absolute Bias (Δ)	Percent Bias Reduction (PBR)	R^2
Covariate-Adjusted Randomized Experiment	4.01 (.35)	.00		.58
Unadjusted Quasi-Experiment	5.01 (.55)	1.00		.28
PS Stratification	3.72 (.57)	.29	71%	.29
Plus Covariates	3.74 (.42)	.27	73%	.66
PS Linear ANCOVA	3.64 (.46)	.37	63%	.34
Plus Covariates	3.65 (.42)	.36	64%	.64
PS Nonlinear ANCOVA	3.60 (.44)	.41	59%	.34
Plus Covariates	3.67 (.42)	.34	66%	.63
PS Weighting	3.67 (.71)	.34	66%	.16
Plus Covariates	3.71 (.40)	.30	70%	.66

• Bias reduction in Math Outcome is 59-73%.

• No adjustment method stood out as best.

• Adding covariates reduces standard error nontrivially.

Vocabulary Outcome

	Mean Difference (standard error)	Absolute Bias (Δ)	Percent Bias Reduction	R^2
Covariate-Adjusted Randomized Experiment	8.25 (.37)			.71
Unadjusted Quasi-Experiment	9.00 (.51)	.75		.60
PS Stratification	8.15 (.62)	.11	86%	.55
Plus Covariates	8.11 (.52)	.15	80%	.76
PS Linear ANCOVA	8.07 (.49)	.18	76%	.62
Plus Covariates	8.07 (.47)	.18	76%	.76
PS Nonlinear ANCOVA	8.03 (.50)	.21	72%	.63
Plus Covariates	8.03 (.48)	.22	70%	.77
PS Weighting	8.22 (.66)	.03	96%	.54
Plus Covariates	8.19 (.51)	.07	91%	.76

• Bias reduction in Vocab Outcome is 70-96%

• No adjustment method stood out as best.

Predictors of Convenience

- We had a rich set of covariates.
- Bad practice: We also tested the effectiveness of propensity score adjustments based only on predictors of convenience (sex, age, ethnicity, marital status)
- We got good balance (but we will see that is misleading—hence balance is nec not suff):

Balance for Predictors of Convenience

Table 3. Rubin's (2001) Balance Criteria Before and After Propensity Score Stratification

Analysis	Propensity Score B	Propensity Score R	Number of Covariates with Variance Ratio ≤1/2	>1/2 and ≤4/5	>4/5 and ≤5/4	>5/4 and ≤2	>2
Before Any Adjustment							
	-1.13	1.51	0	2	17	6	0
After Stratification on Propensity Scores Constructed from All Covariates							
	-0.03	0.93	0	1	22	2	0
After Stratification on Propensity Scores Constructed from Predictors of Convenience Balance Tested only on the 5 Predictors of Convenience							
	-0.01	1.10	0	0	5	0	0
After Stratification on Propensity Scores Constructed from Predictors of Convenience Balance Tested on All 25 Covariates							
	-0.01	1.10	0	2	16	7	0

• Notice balance on these four covariates is pretty good (row three), as is balance on all 25 covariates (row four).

Mathematics Outcome

	Mean Difference (standard error)	Absolute Bias (Δ)	Percent Bias Reduction (PBR)	R^2
Covariate-Adjusted Randomized Experiment	4.01 (.35)	.00		.58
Unadjusted Quasi-Experiment	5.01 (.55)	1.00		.28
PS Stratification	3.72 (.57)	.29	71%	.29
Plus Covariates	3.74 (.42)	.27	73%	.66
PS Linear ANCOVA	3.64 (.46)	.37	63%	.34
Plus Covariates	3.65 (.42)	.36	64%	.64
PS Nonlinear ANCOVA	3.60 (.44)	.41	59%	.34
Plus Covariates	3.67 (.42)	.34	66%	.63
PS Weighting	3.67 (.71)	.34	66%	.16
Plus Covariates	3.71 (.40)	.30	70%	.66
PS Stratification with Predictors of Convenience	4.84 (.51)	.83	17%	.28
Plus Covariates	5.06 (.51)	1.05	-5%[a]	.35

Vocabulary Outcome

	Mean Difference (standard error)	Absolute Bias (Δ)	Percent Bias Reduction	R^2
Covariate-Adjusted Randomized Experiment	8.25 (.37)			.71
Unadjusted Quasi-Experiment	9.00 (.51)	.75		.60
PS Stratification	8.15 (.62)	.11	86%	.55
Plus Covariates	8.11 (.52)	.15	80%	.76
PS Linear ANCOVA	8.07 (.49)	.18	76%	.62
Plus Covariates	8.07 (.47)	.18	76%	.76
PS Nonlinear ANCOVA	8.03 (.50)	.21	72%	.63
Plus Covariates	8.03 (.48)	.22	70%	.77
PS Weighting	8.22 (.66)	.03	96%	.54
Plus Covariates	8.19 (.51)	.07	91%	.76
PS Stratification with Predictors of Convenience	8.77 (.48)	.52	30%	.62
Plus Covariates	8.68 (.47)	.43	43%	.65

Balance Redux

- All propensity score studies I know assume if they got balance, that is sufficient.
- But our results clearly show that is not the case.
- Balance is a necessary but not sufficient condition

Exploring Strong Ignorability

- There is no test for it, but
- The key is having the covariates that predict treatment condition and outcome
- We have been playing with the data to see how much difference it makes to have more and better covariates.
- Consider the following correlations between our covariates and both treatment and outcome.

	Treat.	Vocabulary			Mathematics		
Covariate set	Z	Y^t_{vocab}	Y^c_{vocab}	Y_{vocab}	Y^t_{math}	Y^c_{math}	Y_{math}
dem*	0.22	0.41	0.49	0.38	0.48	0.35	0.35
pre	0.24	0.60	0.49	0.47	0.50	0.47	0.45
aca	0.07	0.58	0.45	0.37	0.63	0.57	0.50
top	0.43	0.33	0.41	0.46	0.45	0.48	0.54
psy	0.18	0.41	0.44	0.32	0.36	0.24	0.24
dem+pre*	0.28	0.64	0.60	0.51	0.61	0.57	0.54
dem+aca	0.24	0.64	0.58	0.48	0.68	0.61	0.57
dem+top	0.44	0.48	0.59	0.53	0.62	0.59	0.61
dem+psy	0.28	0.63	0.59	0.51	0.60	0.49	0.41
pre+top	0.44	0.63	0.57	0.58	0.59	0.59	0.62
pre+aca	0.26	0.64	0.54	0.50	0.68	0.62	0.60
pre+psy	0.30	0.63	0.59	0.51	0.60	0.49	0.49
dem+pre+top	0.44	0.66	0.66	0.59	0.67	0.67	0.66
dem+pre+aca*	0.30	0.68	0.63	0.54	0.71	0.65	0.63
dem+pre+aca+top*	0.45	0.45	0.70	0.68	0.62	0.75	0.73
dem+pre+aca+top+psy*	0.47	0.47	0.72	0.71	0.63	0.79	0.74

Outcome prediction generally good, but treatment pred more variable

Adjustments

- Now consider how well these different sets of covariates reduce bias in Vocabulary Outcome (Results were similar for Math Outcome):

Percent Bias Remaining

Vocabulary	PS-Stratification			PS-ANCOVA			PS-Weighting		
		s.e.			s.e.			s.e.	
Adjusted Randomized Experiment									
Unadjusted Quasi-Experiment									
Adjusted Quasi-Experiments									
dem*	8.60	0.47	46	8.69	0.49	58	8.68	0.47	57
pre	8.57	0.43	43	8.56	0.44	41	8.47	0.43	30
aca	8.78	0.43	70	8.69	0.45	59	8.69	0.44	58
top	8.53	0.49	38	8.36	0.54	14	8.44	0.49	25
psy	8.78	0.47	70	8.77	0.48	69	8.72	0.47	62
dem+pre*	8.54	0.42	39	8.47	0.44	29	8.41	0.41	21
dem+aca	8.55	0.42	39	8.49	0.43	32	8.51	0.42	35
dem+top	8.43	0.46	24	8.38	0.51	17	8.43	0.46	24
dem+psy	8.52	0.45	35	8.48	0.48	30	8.55	0.44	40
pre+top	8.20	0.42	-7	8.19	0.48	-8	8.32	0.43	9
pre+aca	8.48	0.42	31	8.37	0.42	16	8.27	0.42	2
pre+psy	8.41	0.40	21	8.45	0.45	27	8.32	0.42	9
dem+pre+top	8.29	0.42	6	8.26	0.46	1	8.33	0.42	10
dem+pre+aca*	8.24	0.40	-2	8.28	0.42	4	8.21	0.40	-5
dem+pre+aca+top*	8.20	0.40	-7	8.02	0.44	-31	8.20	0.41	-8
dem+pre+aca+top+psy*	8.14	0.39	-15	8.06	0.45	-25	8.11	0.39	-19

Note the relationship between having variables that predict treatment and bias reduction. More clear in the next table of correlations:

Correlations

		PSSV	PSAV	PSWV	ANCOVAV	PSSM	PSAM	PSWM	ANCOVAM
auc	Pearson Correlation	-.734	-.818	-.760	-.770	-.882	-.870	-.887	-.879
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
	N	19	19	19	19	19	19	19	19
cort	Pearson Correlation	-.758	-.768	-.794	-.810	-.855	-.849	-.837	-.843
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000
	N	19	19	19	19	19	19	19	19
corv	Pearson Correlation	-.844	-.789	-.804	-.804	-.604	-.494	-.551	-.418
	Sig. (2-tailed)	.000	.000	.000	.000	.006	.031	.014	.075
	N	19	19	19	19	19	19	19	19
corm	Pearson Correlation	-.788	-.834	-.737	-.760	-.675	-.557	-.614	-.511
	Sig. (2-tailed)	.000	.000	.000	.000	.002	.013	.005	.025
	N	19	19	19	19	19	19	19	19

This table shows that the higher the correlation of the predictor set with treatment or outcome (the rows), the higher bias reduction no matter what method is used (the columns).

Conversely, the next table shows that balance is essentially unrelated to bias reduction:

Correlations

		PSSV	PSAV	PSWV	ANCOVAV	PSSM	PSAM	PSWM	ANCOVAM
B	Pearson Correlation	-.137	-.149	-.129	.106	-.168	-.249	-.272	-.315
	Sig. (2-tailed)	.576	.543	.598	.665	.491	.303	.260	.190
	N	19	19	19	19	19	19	19	19
R	Pearson Correlation	-.192	-.247	-.169	-.053	-.070	.028	-.065	.025
	Sig. (2-tailed)	.430	.308	.489	.831	.775	.909	.792	.920
	N	19	19	19	19	19	19	19	19
BPCT	Pearson Correlation	-.281	-.161	-.359	-.407	-.164	-.112	-.094	-.058
	Sig. (2-tailed)	.244	.511	.131	.084	.503	.647	.701	.813
	N	19	19	19	19	19	19	19	19
RPCT	Pearson Correlation	.164	.232	.222	.186	.362	.436	.380	.514
	Sig. (2-tailed)	.501	.340	.361	.447	.128	.062	.108	.024
	N	19	19	19	19	19	19	19	19

Here the rows are Rubin's (2001) balance metrics, and the columns are bias reduction as in the previous table.

Observations

- Balance is unrelated to bias reduction
- Predicting treatment or predicting outcome are strongly related to bias reduction.
- Lesson: You really do need a good set of covariates to get bias reduction.

ANCOVA

- To simplify, I didn't go over the ordinary OLS ANCOVA results, but they did as well as the more complicated propensity score methods.
- For example, look at the last row of the next two tables:

Mathematics Outcome

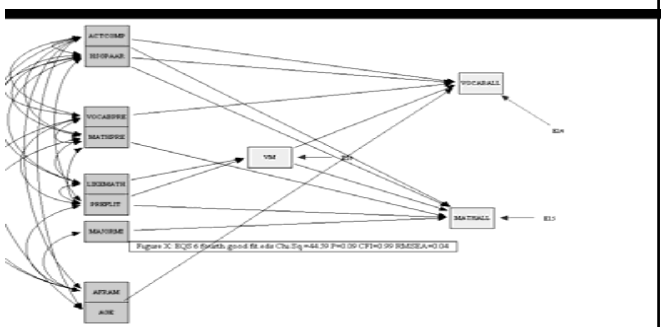
	Mean Difference (standard error)	Absolute Bias (Δ)	Percent Bias Reduction (PBR)	R^2
Covariate-Adjusted Randomized Experiment	4.01 (.35)	.00		.58
Unadjusted Quasi-Experiment	5.01 (.55)	1.00		.28
PS Stratification	3.72 (.57)	.29	71%	.29
Plus Covariates	3.74 (.42)	.27	73%	.66
PS Linear ANCOVA	3.64 (.46)	.37	63%	.34
Plus Covariates	3.65 (.42)	.36	64%	.64
PS Nonlinear ANCOVA	3.60 (.44)	.41	59%	.34
Plus Covariates	3.67 (.42)	.34	66%	.63
PS Weighting	3.67 (.71)	.34	66%	.16
Plus Covariates	3.71 (.40)	.30	70%	.66
PS Stratification with Predictors of Convenience	4.84 (.51)	.83	17%	.28
Plus Covariates	5.06 (.51)	1.05	-5%[a]	.35
ANCOVA Using Observed Covariates	3.85 (.44)	.16	84%	.63

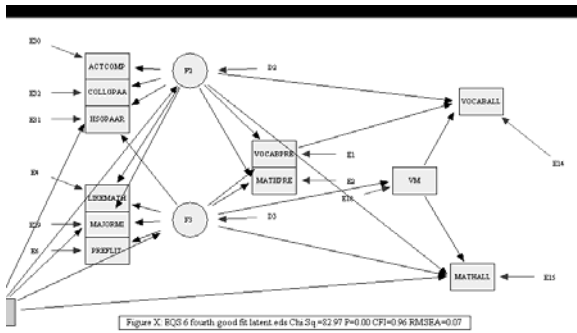
Vocabulary Outcome

	Mean Difference (standard error)	Absolute Bias (Δ)	Percent Bias Reduction	R^2
Covariate-Adjusted Randomized Experiment	8.25 (.37)			.71
Unadjusted Quasi-Experiment	9.00 (.51)	.75		.60
PS Stratification	8.15 (.62)	.11	86%	.55
Plus Covariates	8.11 (.52)	.15	80%	.76
PS Linear ANCOVA	8.07 (.49)	.18	76%	.62
Plus Covariates	8.07 (.47)	.18	76%	.76
PS Nonlinear ANCOVA	8.03 (.50)	.21	72%	.63
Plus Covariates	8.03 (.48)	.22	70%	.77
PS Weighting	8.22 (.66)	.03	96%	.54
Plus Covariates	8.19 (.51)	.07	91%	.76
PS Stratification with Predictors of Convenience	8.77 (.48)	.52	30%	.62
Plus Covariates	8.68 (.47)	.43	43%	.65
ANCOVA Using Observed Covariates	8.21 (.43)	.05	94%	.76

Structural Equation Models as Adjustments

- If ordinary ANCOVA did well, perhaps SEM would do well too.
- After all, it can do more complex models than ordinary ANCOVA:





Randomized Results		Math Effect	Vocab Effect
		4.01	8.25
Observed Variable Models			
Model	CFI	Math Effect	Vocab Effect
First good fit	.987	4.37	8.48
Second good fit	.971	3.83	8.19
Third good fit	.980	3.96	8.38
Fourth good fit	.990	3.96	8.43
Fifth good fit	.978	3.91	8.32
Observed Mediational Models			
Fourth good fit	.983	3.96	8.43
Latent Variable Models			
Model	CFI	Math Effect	Vocab Effect
Fourth good fit	.961	3.69	8.49

Discussion

- These analyses are encouraging that nonrandomized experiments might yield results similar to randomized experiments if
  - Both balance
  - And strong ignorability are met
- It doesn't seem to matter much which analytic method is used.

Discussion

- This laboratory analogue is an improvement over past methods for studying this question
  - Though it has clear generalizability questions
- We are currently replicating, and also doing a study randomly assigning to a RE and an RDD (preliminary results are encouraging).
- One little piece in the renaissance of social experimentation.

The End

Acknowledgement: M.H. Clark (SIU), Peter Steiner (NU)