

When Experiments and Observational Studies give comparable Causal Estimates:

Thomas D. Cook

NYU, 2008

We probably all agree that:

- Experiments are best for causal inference
- Experiments are not always possible
- Observational studies can be unbiased if RD, IV, or if selection process is completely known
- But RD excepted, it is hard to know when these conditions hold in actual applications.
- This the root of our dilemma

We probably also agree that:

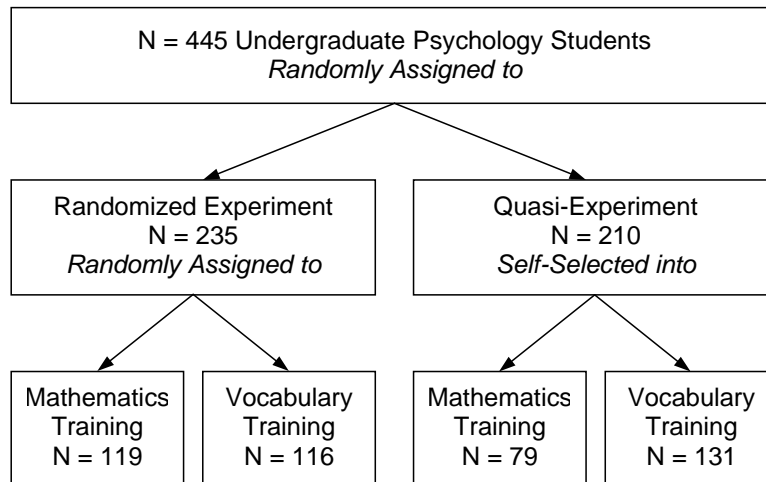
- One criterion for judging the quality of proposed observational studies is theoretical, embedded in logic and algebra-- e.g, formal proofs for RDD and IV
- Another is empirical. Does a particular kind of observational study “often” reproduce the same results as experiments
- This second criterion is the focus here today

Purposes

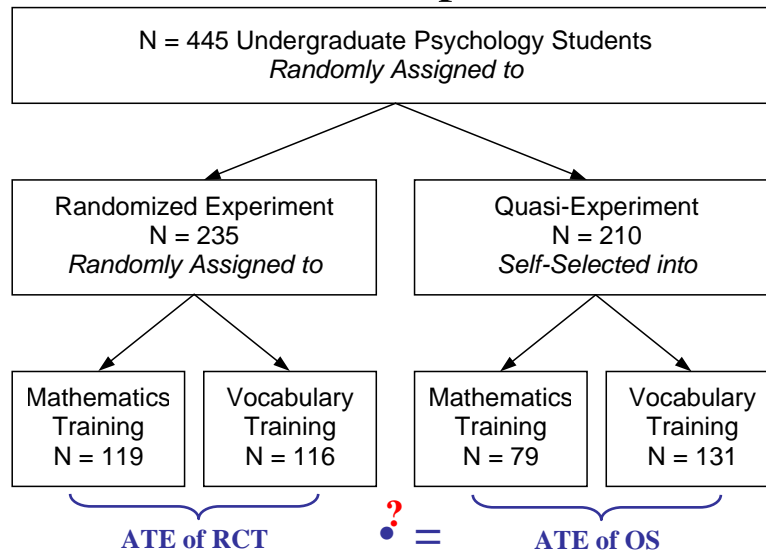
- Describe a study by Shadish, Clark & Steiner (JASA, 2008), contending that an adjusted quasi-experiment gives same result as an experiment
- Ask if it can be replicated (Pohl, Steiner, Eisenmann, Soellner & Cook)
- Explain why Shadish et al got the results they did (Steiner, Cook, Shadish & Clark; and Steiner & Cook)
- Probe the generality of these results across non-laboratory settings (Glazerman et al, AAAPSS, 2003) and (Cook, Shadish & Wong JPAM, 2008)

• **Shadish, Clark & Steiner (JASA, 2008)**

4-Arm within-study Comparison



4-Arm Comparison



More on the Design

- All participants pretested on a host of covariates - proxy pretest, past performance, topic preference, demographics, personality (big 5)
- Chose math and vocab training because
 - Good analogue to educational interventions
 - 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 to create a replication of the basic effect and chance to see if biased reduced.

Bias in the Unadjusted Q-E 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

Bias in the Unadjusted Q-E 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.

Adjusted Quasi-Experiments

- It is no surprise that randomized and nonrandomized experiments might yield different answers.
- The more important question is whether statistical adjustments can improve the quasi-experimental estimate
- Consider the use of propensity scores and of OLS (ANCOVA) to make those adjustments

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

Note: All estimates are based on regression analyses. For propensity score stratification studies.

Predictors of Convenience

- Bad practice: We also tested the effectiveness of propensity score adjustments based only on predictors of convenience (sex, age, ethnicity, marital status)
- Depending on how we did the analyses bias reduction ranged from 43% bias reduction to 5% bias increase.
- The importance of thoughtful selection of covariates in the design of the study.

Balance for Predictors of Convenience

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

Analysis	Propensity Score		Number of Covariates with Variance Ratio				
	B	R	≤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

Ordinary OLS Regression

- 84-94% bias reduction just by entering covariates as predictors in regression.
- What good are propensity scores, then?
 - When creating a control group by matching
 - To discover if there is enough balance to make adjustments valid.
 - When the assumptions of ANCOVA (e.g., linearity) are problematic.

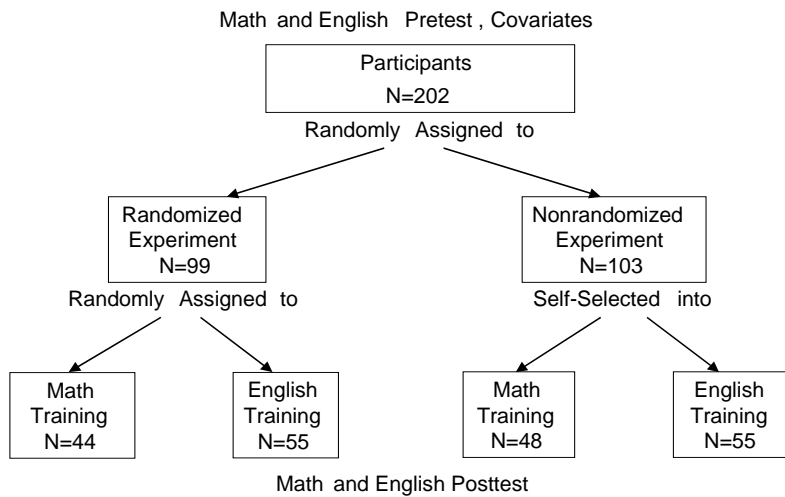
Summary of Shadish et al

- Elegant study to show bias and its reduction
- More like an analog experiment than an exp
- Less relevant populations and interventions
- Will it replicate even in similar situations
- Probably a simple (self-)selection process
- Initial bias quite small, .09 and .24 SDs

**Do these Findings Replicate
in a similar Study?**

**Pohl, Steiner, Eisermann,
Soellner & Cook**

Pohl, Steiner, Eisermann, Soellner & Cook (2008)



English (SD = .16)	Mean diff.	Standard error	Bias	Bias SD	in R squared
Random. experiment (covar.-adj.)	-0.016	0.025	0.000	0.00	0.63
Unadjusted quasi-experiment	-0.078	0.033	-0.061	-0.38	0.05
PS stratification	-0.005	0.037	0.012	0.07	0.00
plus covariate adjustment	-0.011	0.022	0.005	0.03	0.68
PS nonlinear ANCOVA	-0.005	0.033	0.012	0.07	0.24
plus covariate adjustment	-0.017	0.025	0.000	0.00	0.63
ANCOVA using observ. covariates	-0.013	0.024	0.004	0.02	0.64

Mathematics (SD = .22)	Mean diff.	Standard error	Bias	Bias SD	in R squared
Random. experiment (covar.-adj.)	0.126	0.047	0.00	0.00	0.47
Unadjusted quasi-experiment	0.139	0.040	0.01	0.06	0.11
PS stratification	0.140	0.042	0.01	0.06	0.10
plus covariate adjustment	0.150	0.035	0.02	0.11	0.47
PS nonlinear ANCOVA	0.126	0.045	0.00	0.00	0.10
plus covariate adjustment	0.137	0.040	0.01	0.05	0.40
ANCOVA using observ. covariates	0.142	0.038	0.02	0.07	0.38

Summary 1

In English, bias was greater than in Shadish et al--.39 SDs

- Bias was eliminated by the covariates
- Mode of data analysis made no difference
- Heterogeneity in nation, topic, subjects, size of bias, direction of bias.
- But not heterogeneous in basic procedure of university students and short intervention

Summary 2

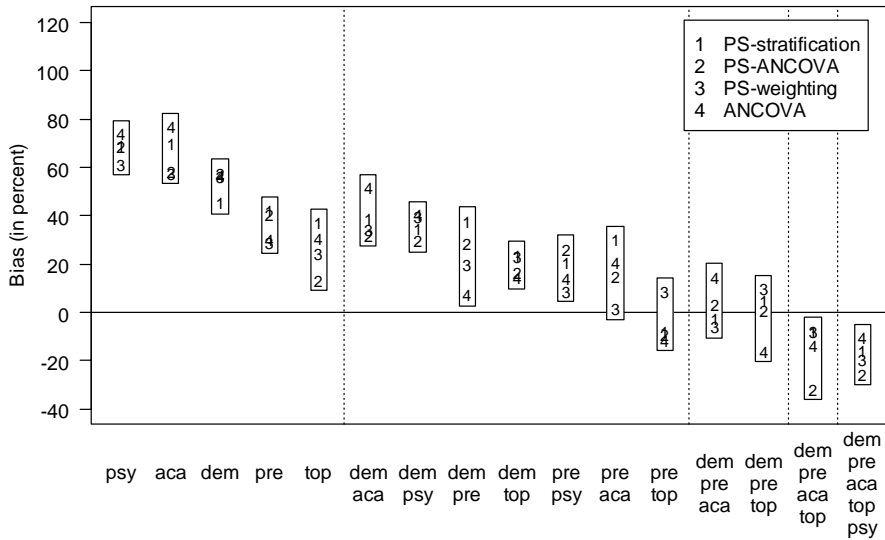
- In Math, there was no initial bias due to self-selection
- What happens if researchers deliberately select non-equivalent populations to maximally overlap with treatment, by being focal, local and intact?
- Three studies examined in Cook et al., and they produce same results in experiment and matched non-experiment
- Thus in these three cases no hidden bias from unobserved covariates

**Why was Bias
successfully reduced in
Shadish et al.?: Steiner,
Cook, Shadish & Clark**

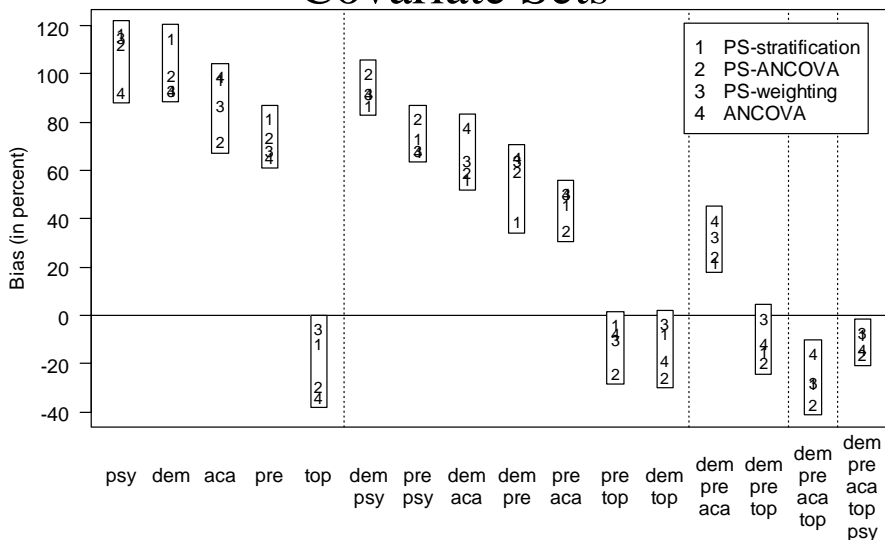
**Explanation 1: via Covariates
Used**

- Single Correct Theory Route in covariate selection
- Multiple plausible theories route in covariate selection
- Covariates in Shadish et al. can be assigned to different domains and sub-domains
- We break down covariates to ask which ones made most difference

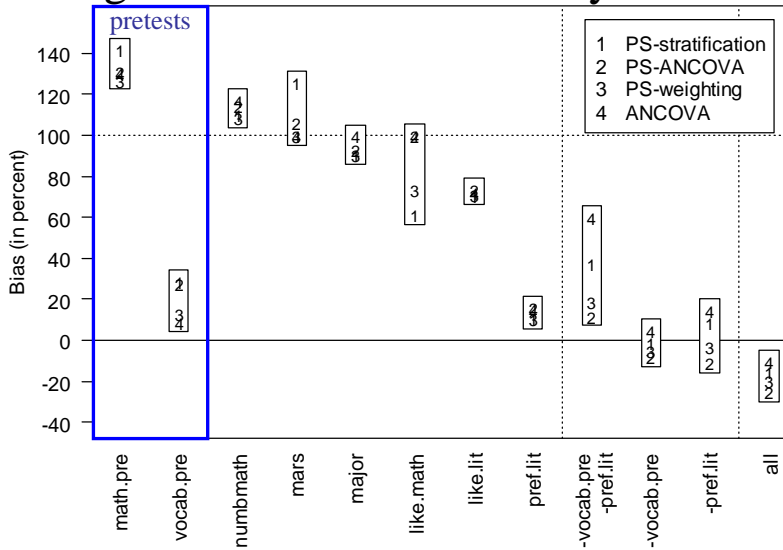
Remaining Bias: Vocabulary Covariate Sets



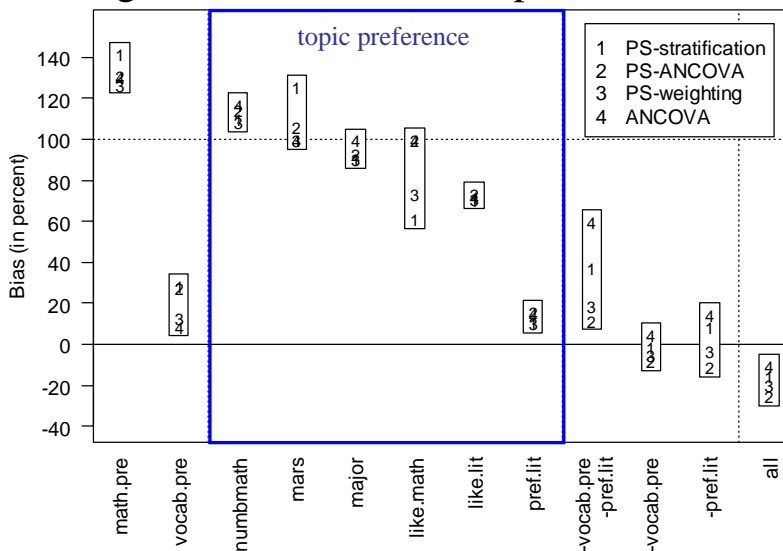
Remaining Bias: Mathematics Covariate Sets



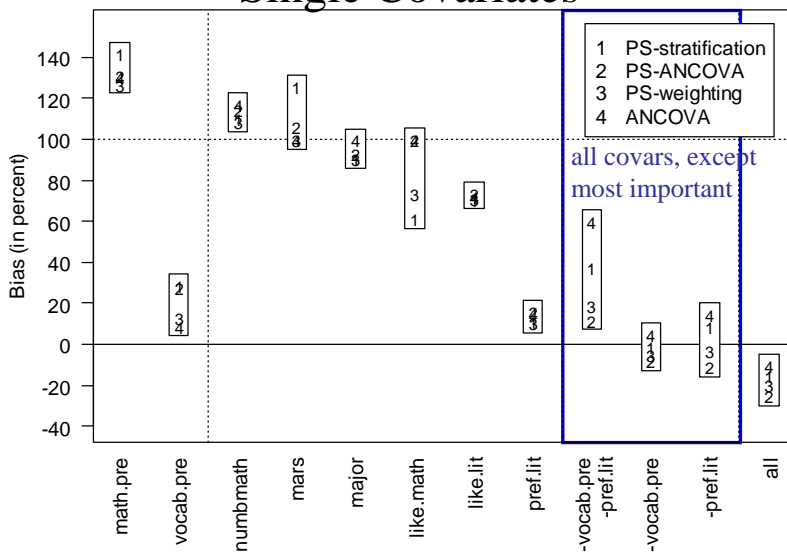
Remaining Bias: Vocabulary Single Covariates from Proxy-Pretests



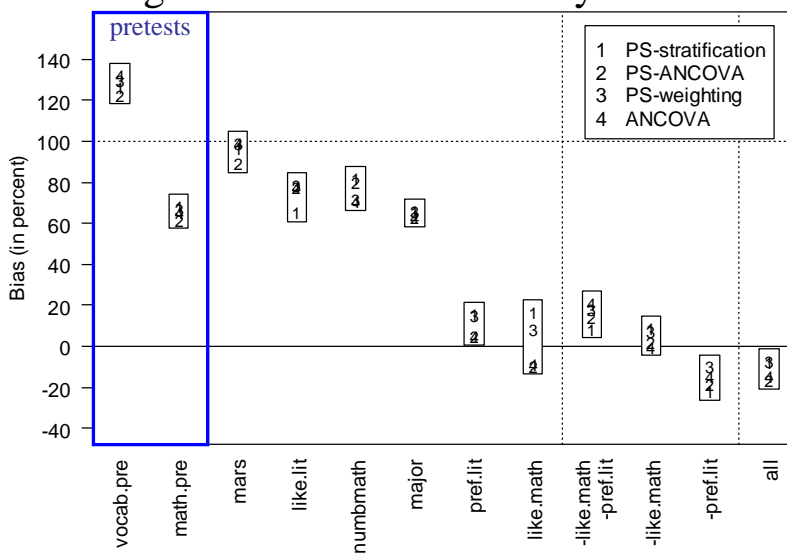
Remaining Bias: Vocabulary Single Covariates from Topic Preference



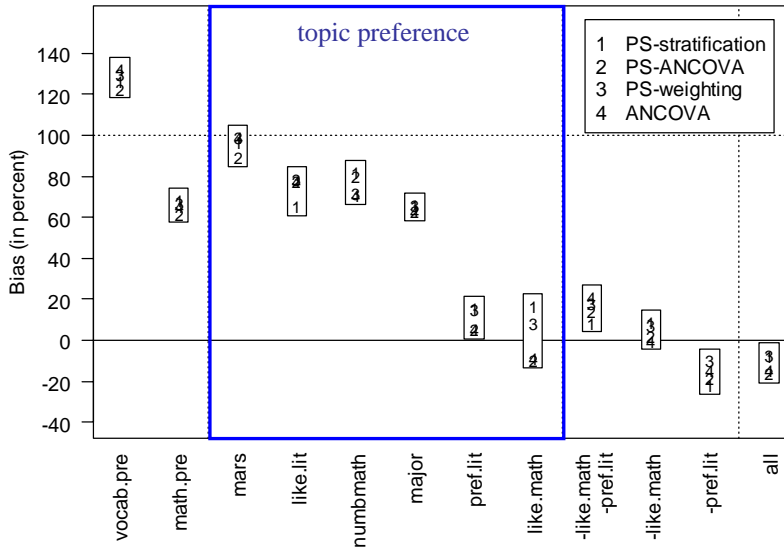
Remaining Bias: Vocabulary Single Covariates



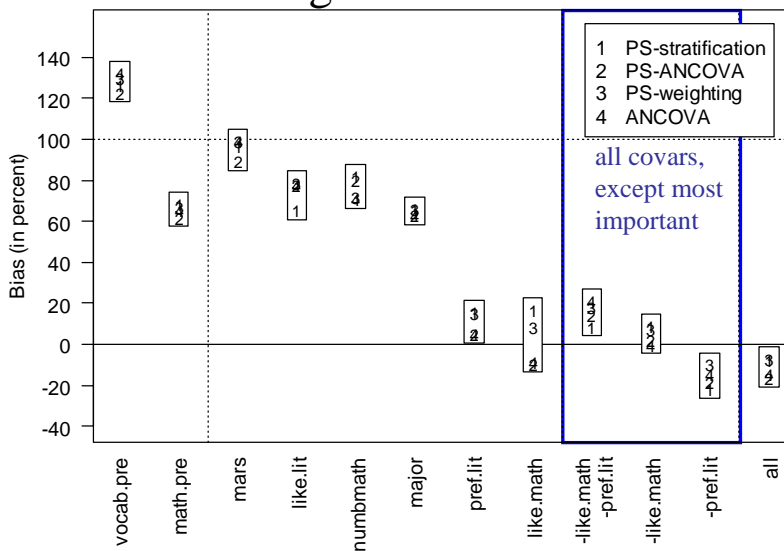
Remaining Bias: Mathematics Single Covariates from Proxy-Pretests



Remaining Bias: Mathematics Single Covariates from Topic Preference



Remaining Bias: Mathematics Single Covariates



Summary thus far:

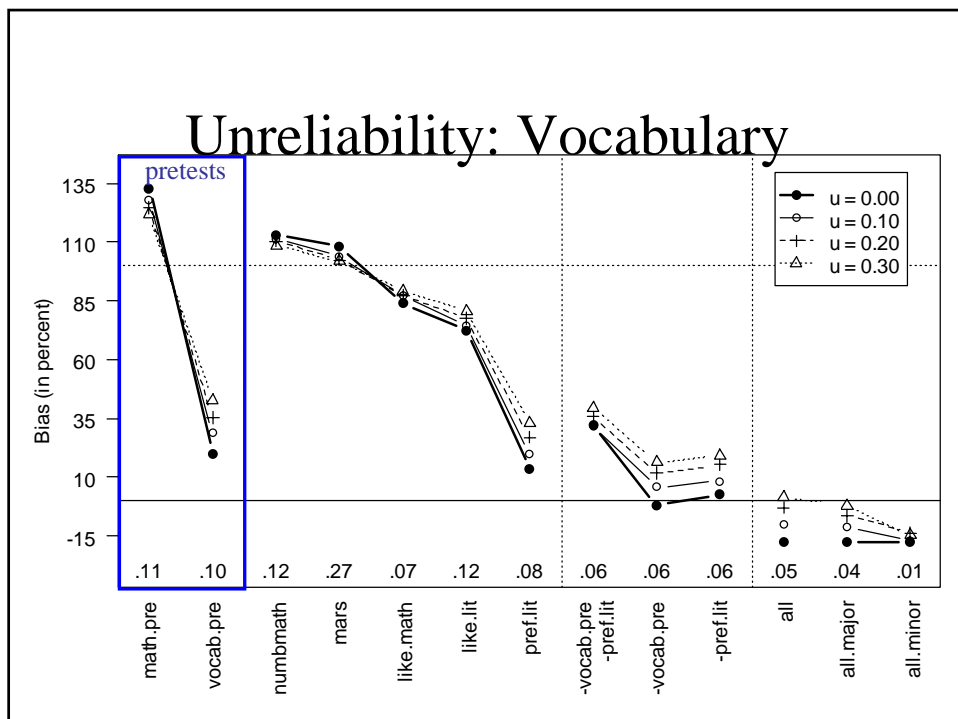
- Bias is reduced by the motivational variables and for vocabulary a little from proxy pretests
- More specifically, the question about preference among the two alternatives is the important one
- The other variables do not help much singly or as single domains
- But the less bias-reducing variables do fine as a group when used for the propensity score
- Implication is two ways to getting covariates that meet strong ignorability (hidden bias) assumption

Two Pathways to Collecting Good Covariates

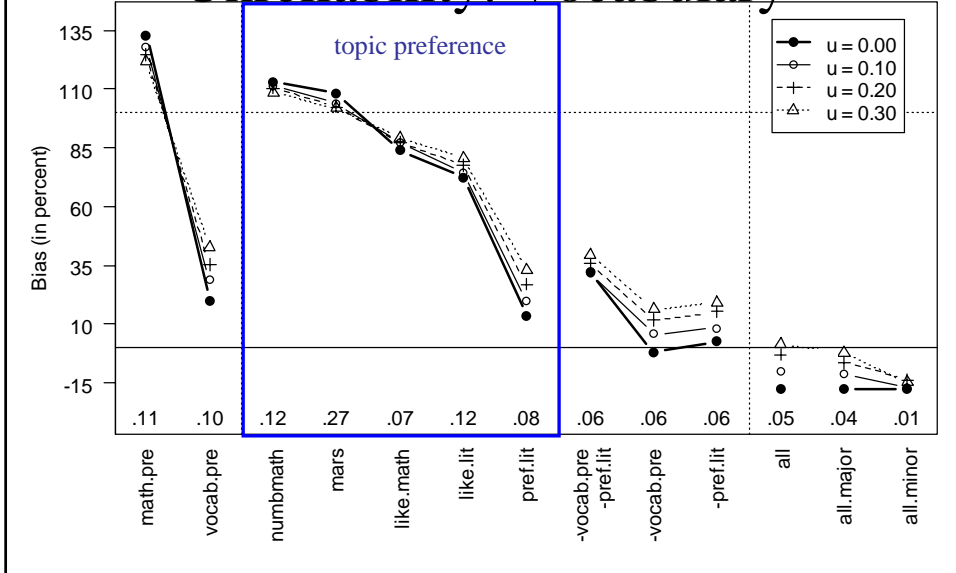
- Single true selection theory. But dilemma is one cannot be sure what this is.
- Multiple theories even if they exclude the true one
- The importance of combining multiple theories of selection and
- Measuring their components very well.

Explanation 2: Unreliability in the Covariates: Steiner & Cook

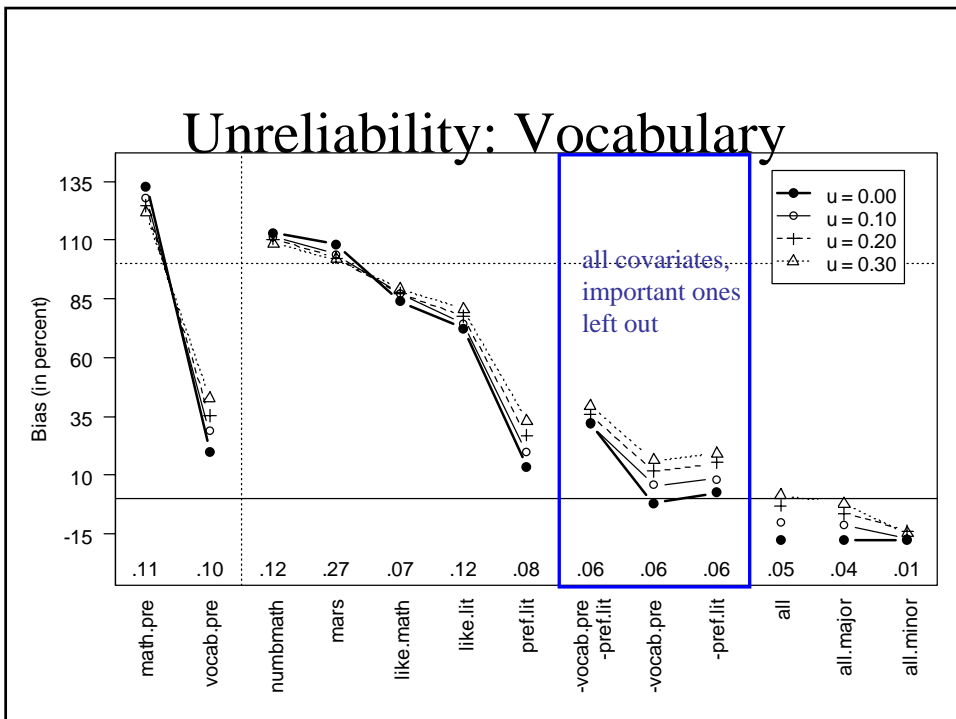
- Shadish et al measured particularly well, but of course not perfectly
- We add 10%, 20%, 30% unreliability to each of their individual items used as covariates to see if unreliability decreases the amount of achieved bias reduction
- One expectation is that it will
- Another is that the propensity score is a composite and so more reliable and so will have little effect



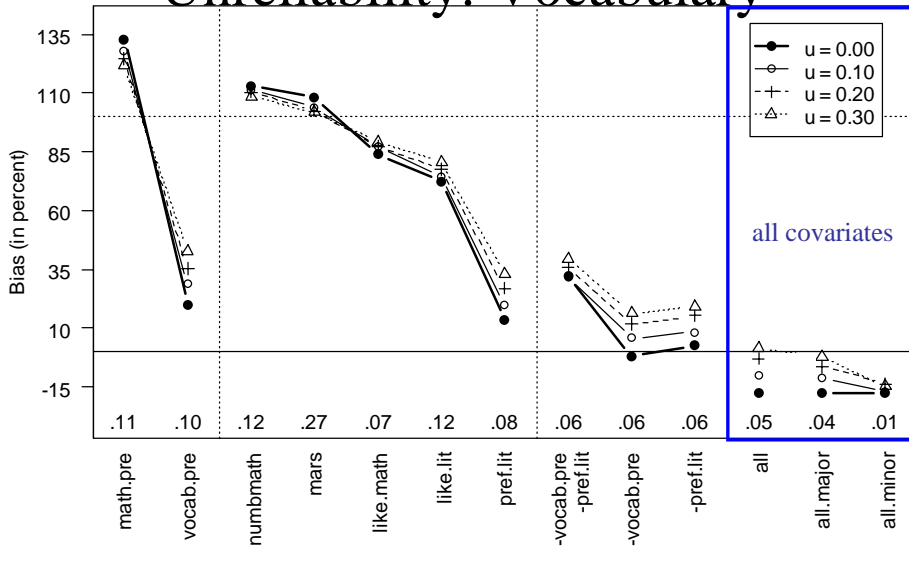
Unreliability: Vocabulary



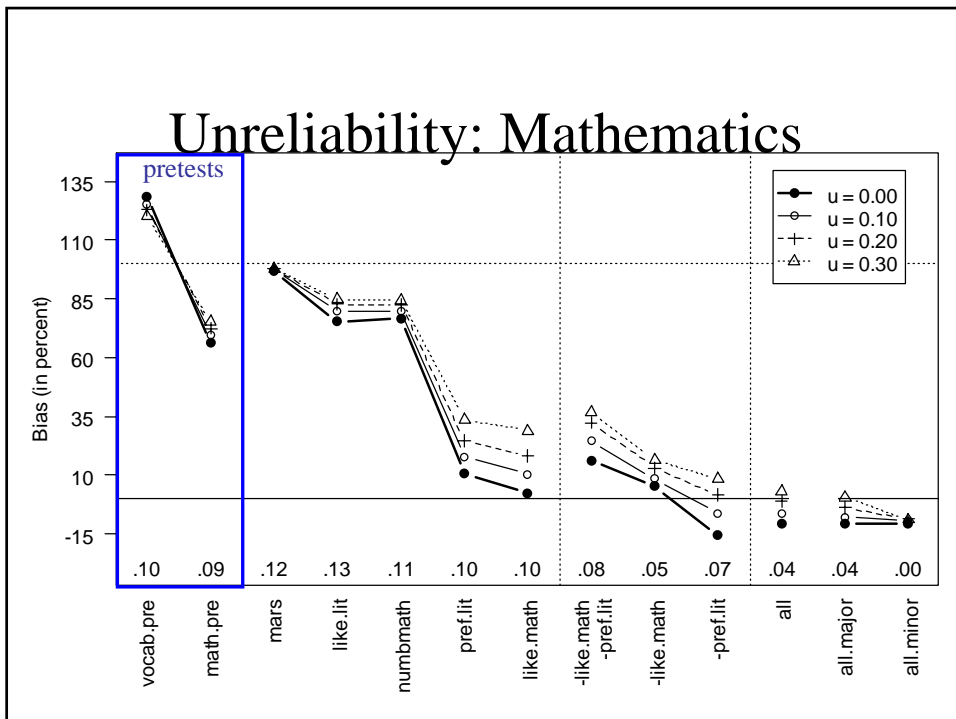
Unreliability: Vocabulary



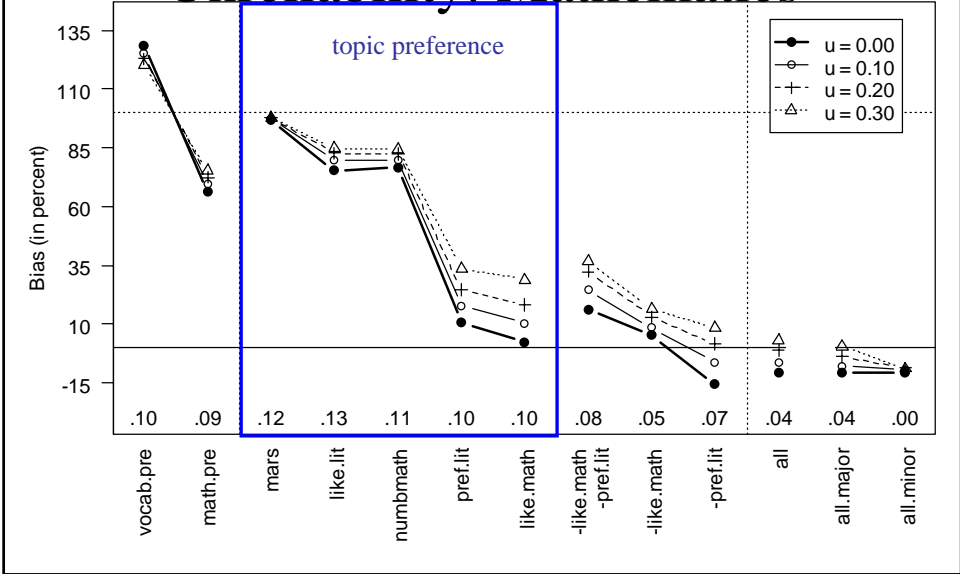
Unreliability: Vocabulary



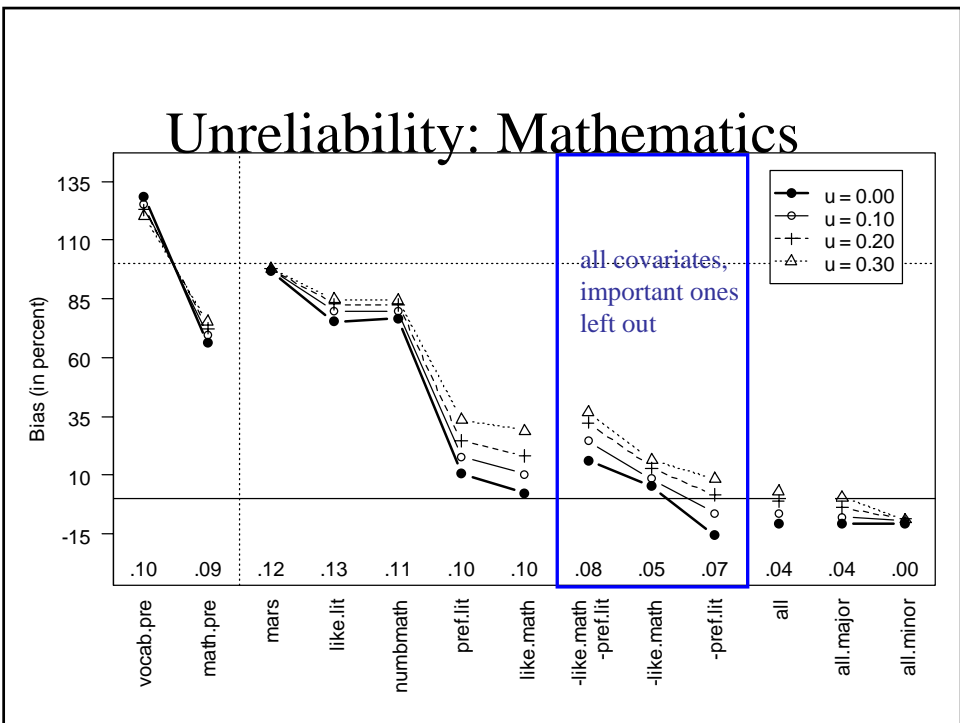
Unreliability: Mathematics



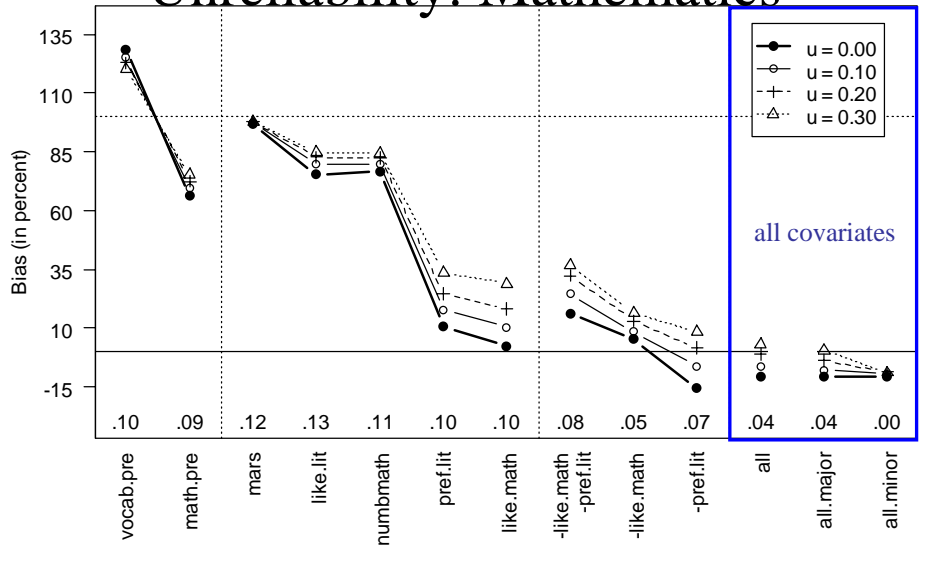
Unreliability: Mathematics



Unreliability: Mathematics



Unreliability: Mathematics



Summary re Unreliability

- Using a large number of covariates reduces the bias problem but does not eliminate it
- Bias is in the direction of the unadjusted difference in the observational study
- Unreliability might have played a minor role in Shadish et al. since the adjusted mean would have been closer to experiment
- But probably a minor role

Explanation 3: Role of the Student Sample

- Note how similar students were in the quasi-experiment and experiment
- Same city, same age, same university, same college class in same quarter
- Almost certainly increased overlap (area of common support) relative to other PS studies--e.g., national smokers and non-smokers

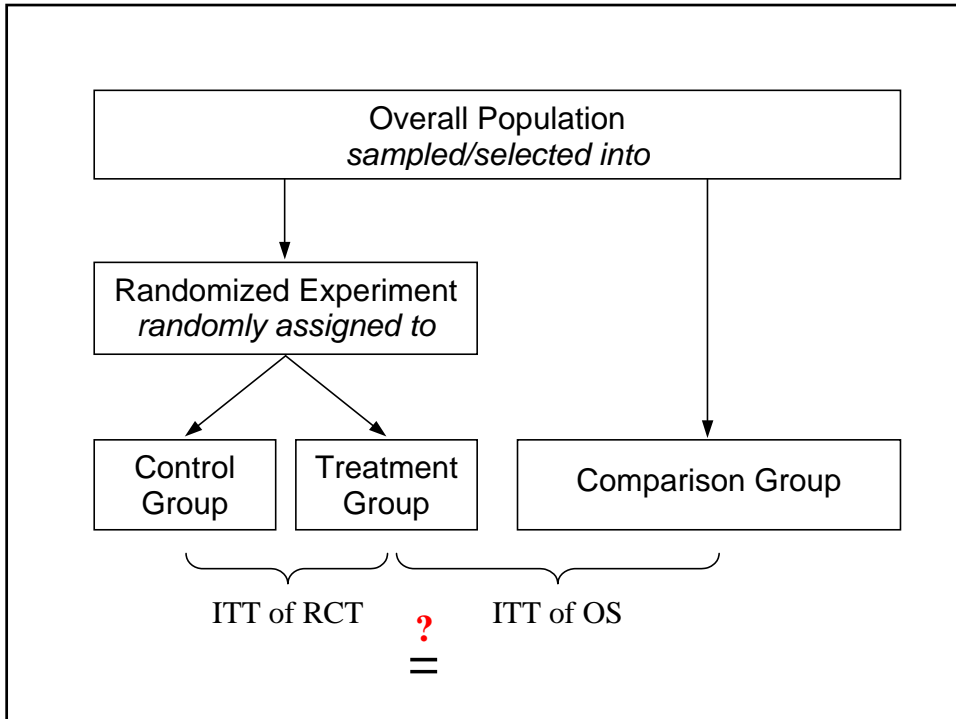
Section Summary

- Shadish et al got the results they did cos of:
- 1. Covariate selection--big one. They had the crucial motivational selection variable plus a second process: other less direct selection processes that cumulatively counted
- 2. High overlap between experimental and non-experimental populations on many background variables that might have caused hidden bias
- 3. Modest help from such reliable measurement
- 4. No help from the mode of data analysis

Which of these Findings Generalize to More Ecologically Relevant Settings?

Reviews of 3-Arm Within-Study Comparisons

- Glazerman et al (2003) in *Annals of the American Academy of Political and Social Science* limited to job training domain and 13 studies
- Cook, Shadish & Wong (2008) in *Journal of Public Policy and Management*. Multiple other domains and 10 studies but 12 contrasts of experiment and quasi-experiment



Glazerman et al Claims

- No Q-E reproduced the experimental estimates to a degree they believed to be adequate to conclude similarity
- Similarity was closer if a pretest, if local controls in Q-E (overlap), if a “rich” set of covariates - vague but in line with Steiner et al interpretation of Shadish et al.
- OLS and PS did not differ in estimates, but Heckman-type IV analyses did worse

Cook, Shadish & Wong Claims

- In 3 of 3 tests, RD and exps give same answer; in 3 of 3 tests selecting local and focal intact group matches same result, as also in 2 of 2 cases when selection process independently known
- In 4 other studies, get same results twice and different ones twice.

Also in Cook et al.

- OLS and PS did not differ
- Use of only demographic controls was only once effective out of many attempts

Conclusion 1: Revisiting why Shadish et al “worked”

- We argued it worked principally because the selection process into treatment was well conceptualized and measured. It was completely known.
- An even better example of this point is the three-arm within-study comparison of Diaz & Handa (2006)

Diaz & Handa (2006)

- Progresa where the assignment into treatment is completely known
- Non-equivalent richer villages with some eligible families within them
- These eligibles clearly different from eligibles in eligible villages on outcomes
- When adjustment for covariates known to control for selection all the bias disappears
- Knowing the selection process eliminates bias

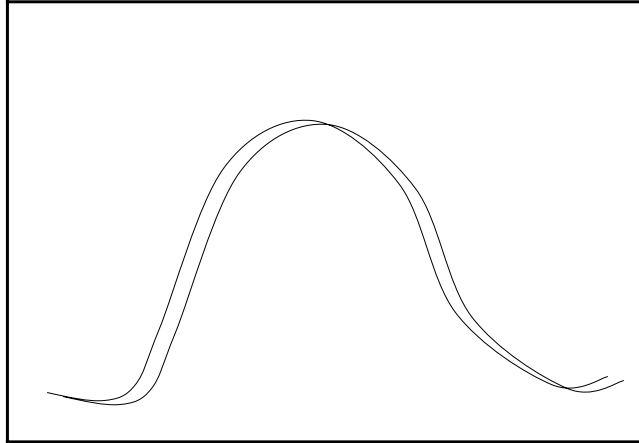
Conclusion 2: Knowing selection model not the only way to unbiased causal knowledge

- Diaz & Handa (2006) other sample of matched villages eligible but not getting Progresa. Here there is no initial bias on observables, nor on outcome. Hence, comparability through sampling designs that match on the pretest
- Also, RD vs experiments produces basically the same answer.
- Three routes are thus....

What you see

- Randomly and non-randomly formed counterfactuals hardly differ and high powered statistical tests confirm no differences in intercept or slope
- In these cases, equivalence of randomly and non-randomly formed comparison groups is achieved thru sampling design alone
- Thus, no need for statistical adjustments to render the non-equivalent counterfactual group “equivalent” to the randomly formed one

Matching for Population Comparability



T C

Intact Group Matching

The Value of the Sampling Design to reduce or eliminate bias on observables.

- Can it work to reduce most bias, including on unobservables?
- How would we know it works?
- If it recreates the results of experiments, of course.

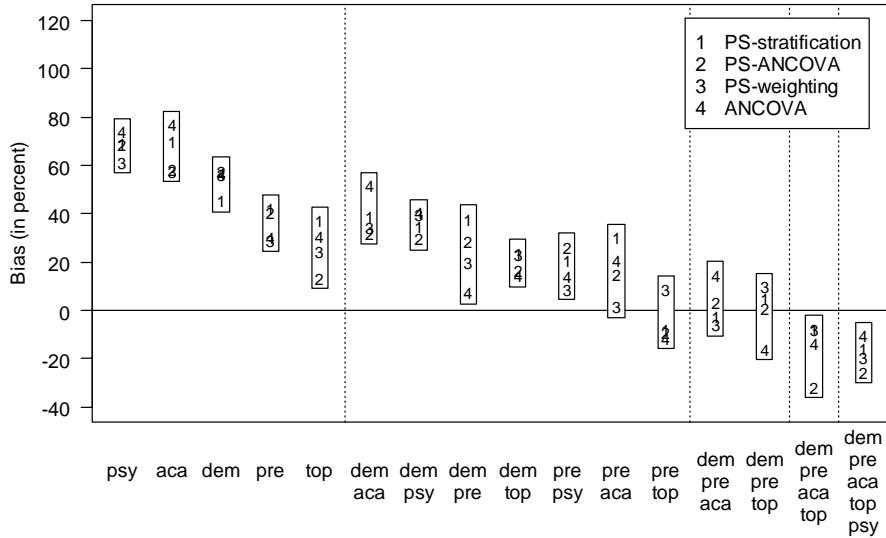
Criteria for Comparing Experiments and Q-Es

- Clear variation in mode of forming control group--random or not
- RCT merits being considered a “gold standard” because it demonstrably meets assumptions
- Experiment and non-experiment difference is not confounded with 3rd variables like measurement
- The quasi-experiment should be a good example of its type--otherwise one compares a good experiment to a poor quasi-experiment

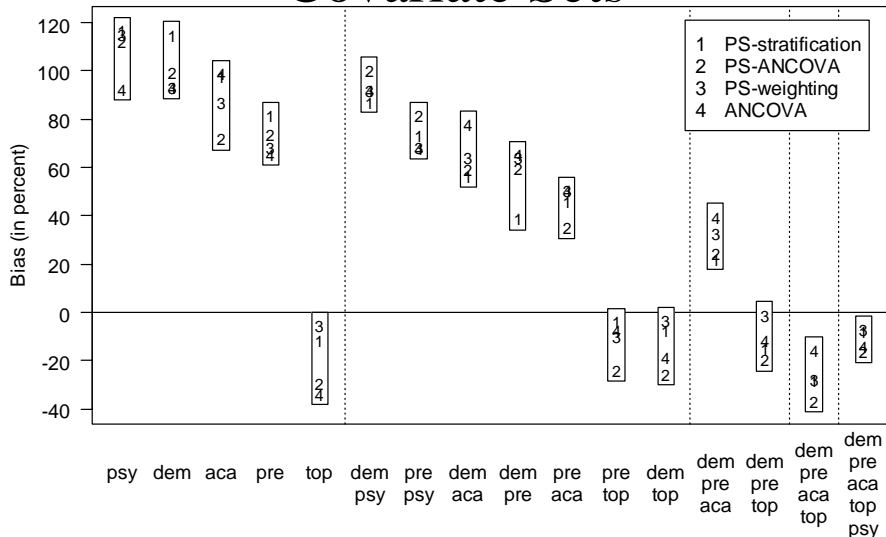
Criteria continued

- The experiment and quasi-experiment should estimate the same causal quantity--not LATE vs ATE or ITT vs TOT
- Criteria for inferring correspondence of results should be clear
- The non-experimental analyses should be done blind to the experimental results

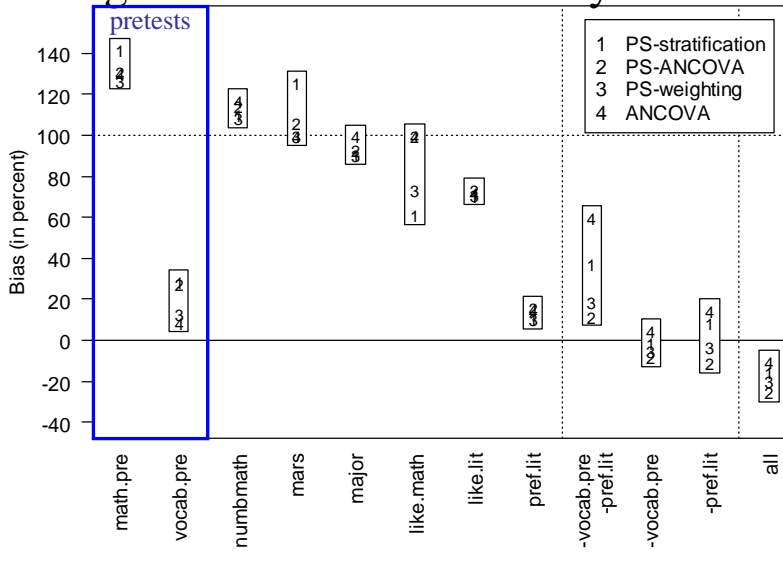
Remaining Bias: Vocabulary Covariate Sets



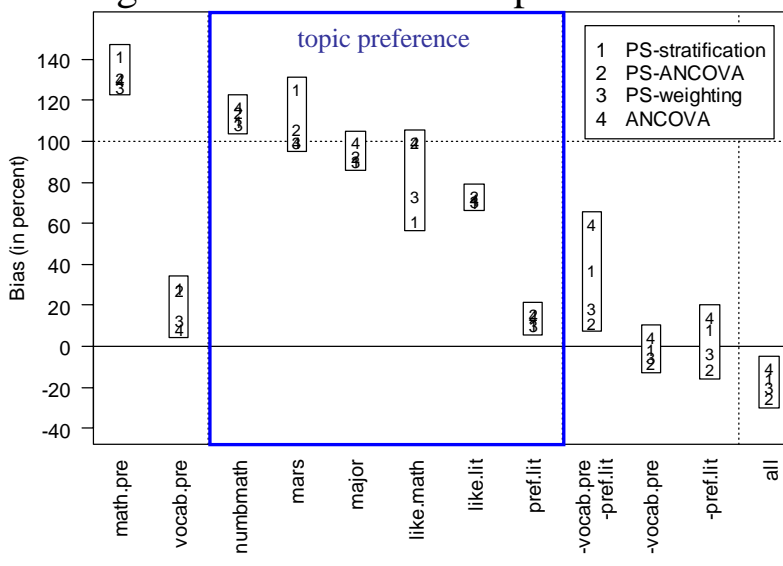
Remaining Bias: Mathematics Covariate Sets



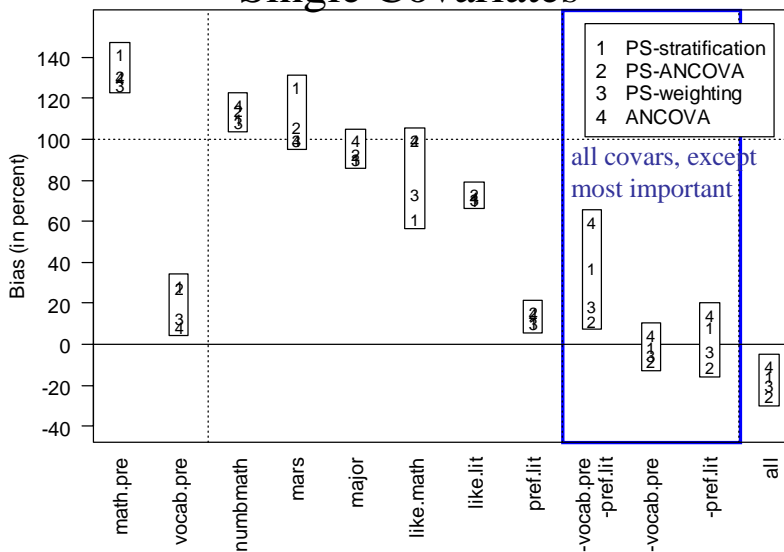
Remaining Bias: Vocabulary Single Covariates from Proxy-Pretests



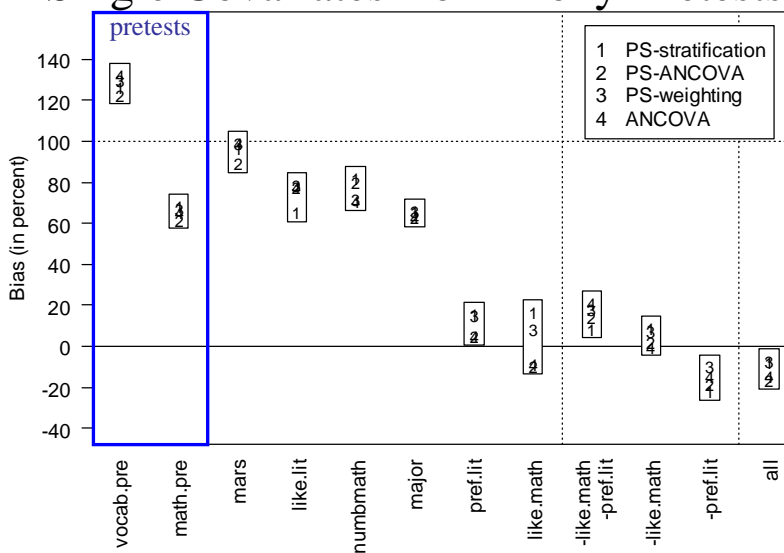
Remaining Bias: Vocabulary Single Covariates from Topic Preference



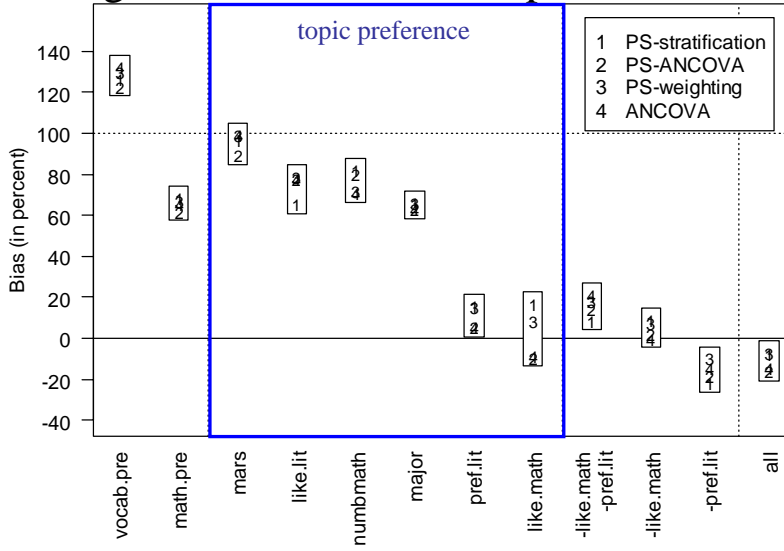
Remaining Bias: Vocabulary Single Covariates



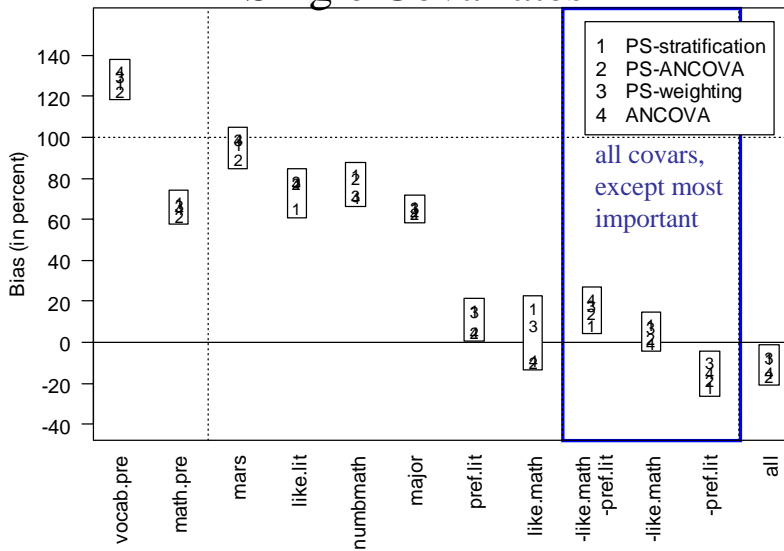
Remaining Bias: Mathematics Single Covariates from Proxy-Pretests



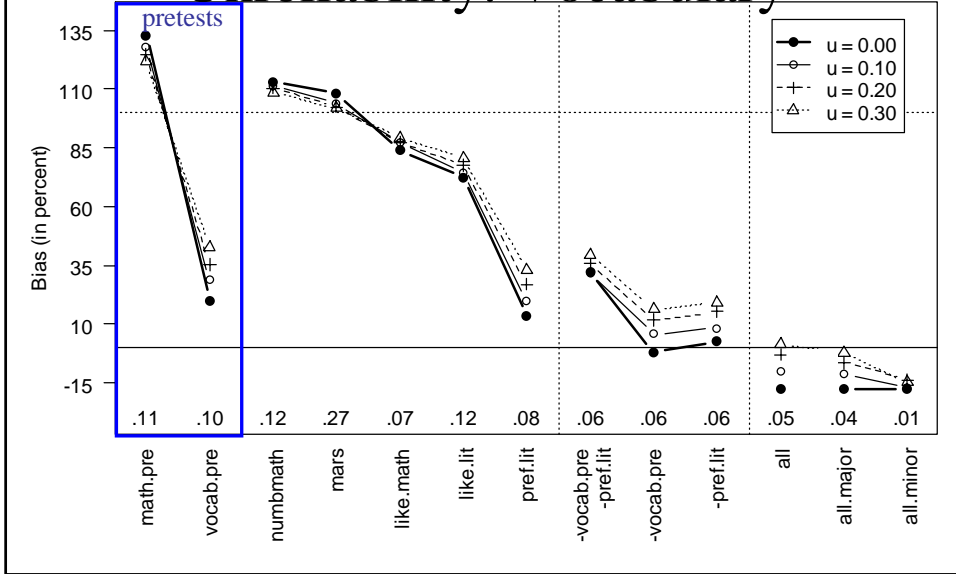
Remaining Bias: Mathematics Single Covariates from Topic Preference



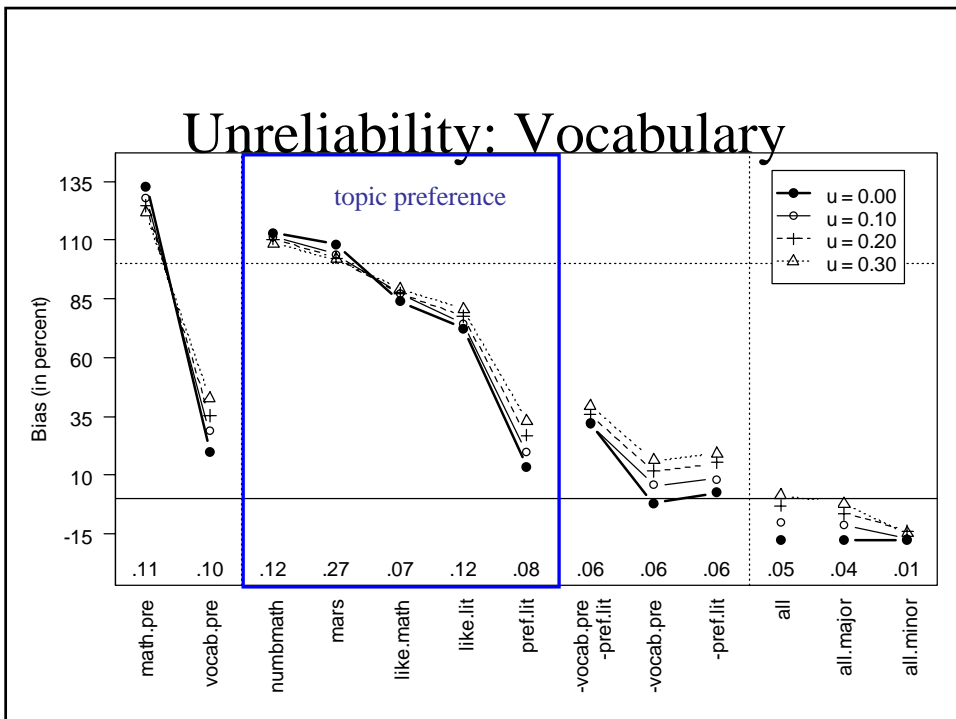
Remaining Bias: Mathematics Single Covariates



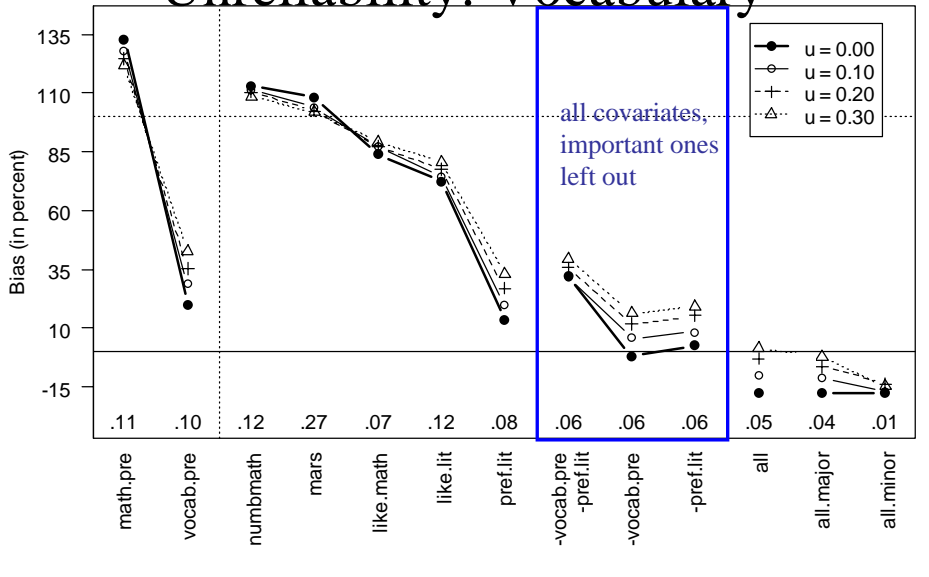
Unreliability: Vocabulary



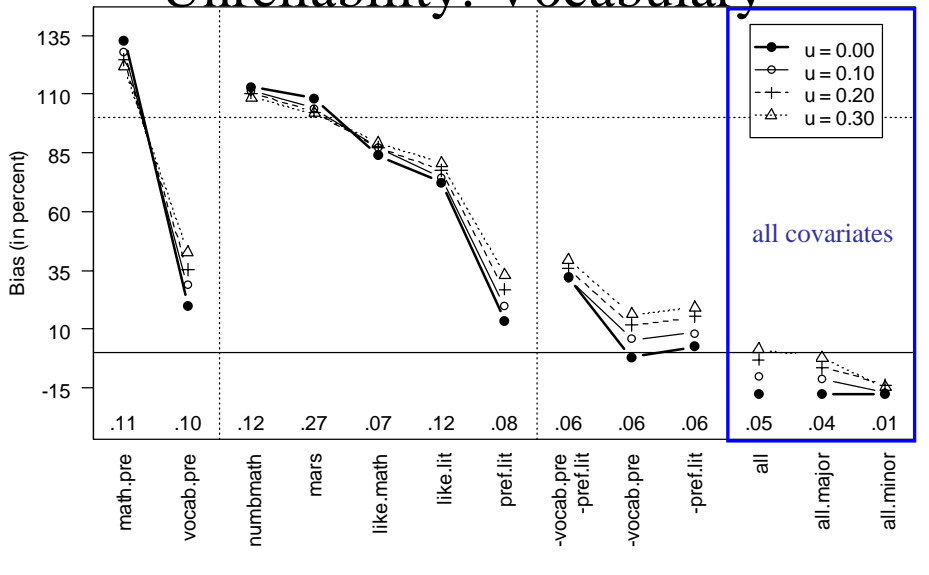
Unreliability: Vocabulary



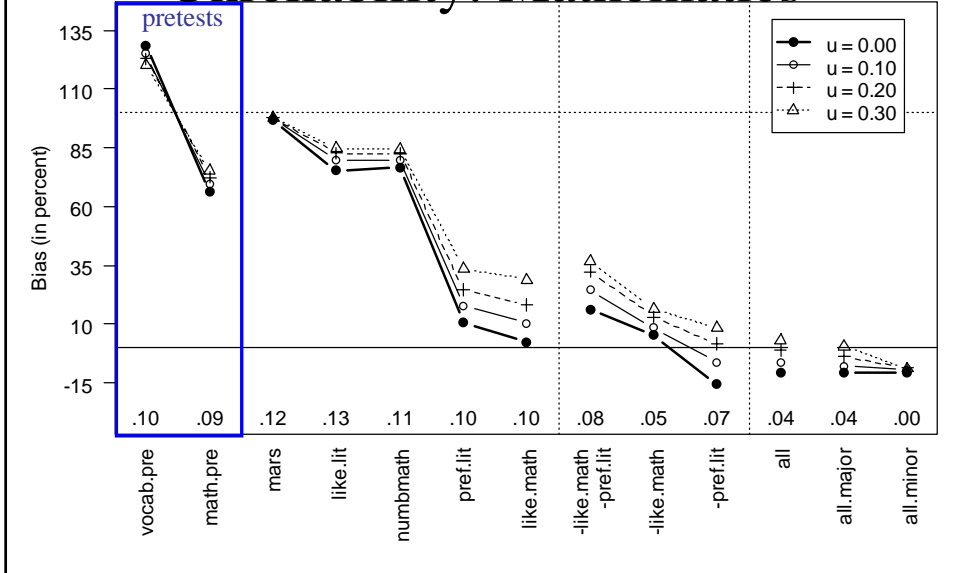
Unreliability: Vocabulary



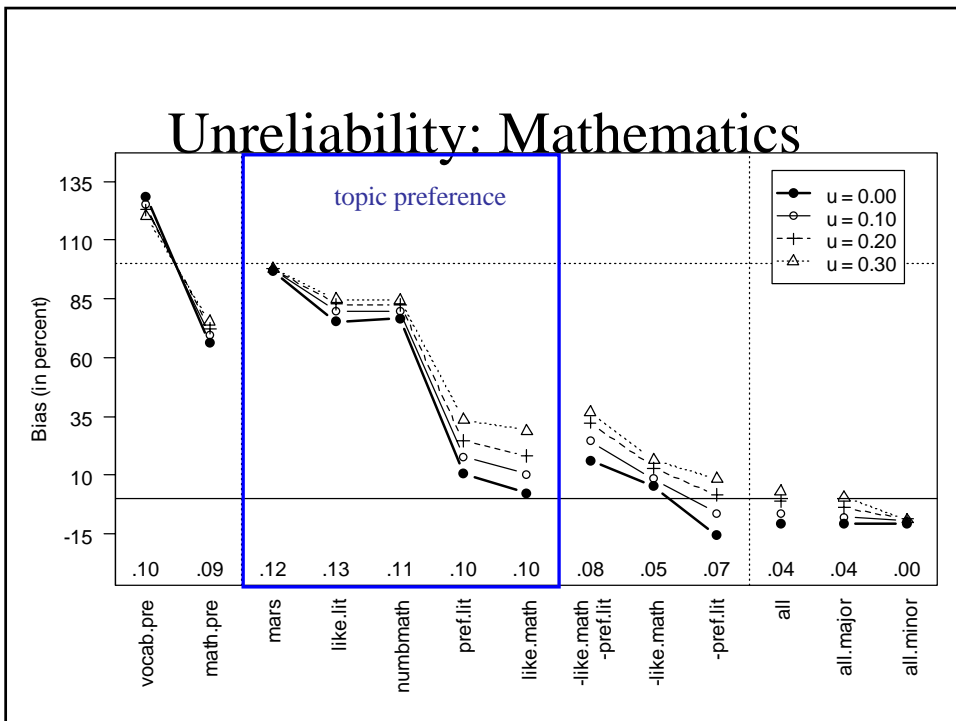
Unreliability: Vocabulary



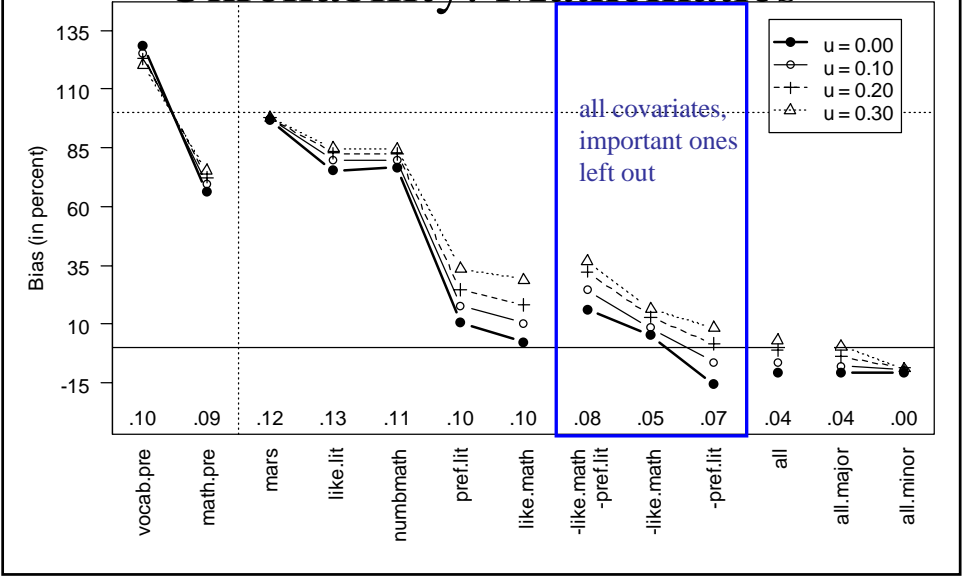
Unreliability: Mathematics



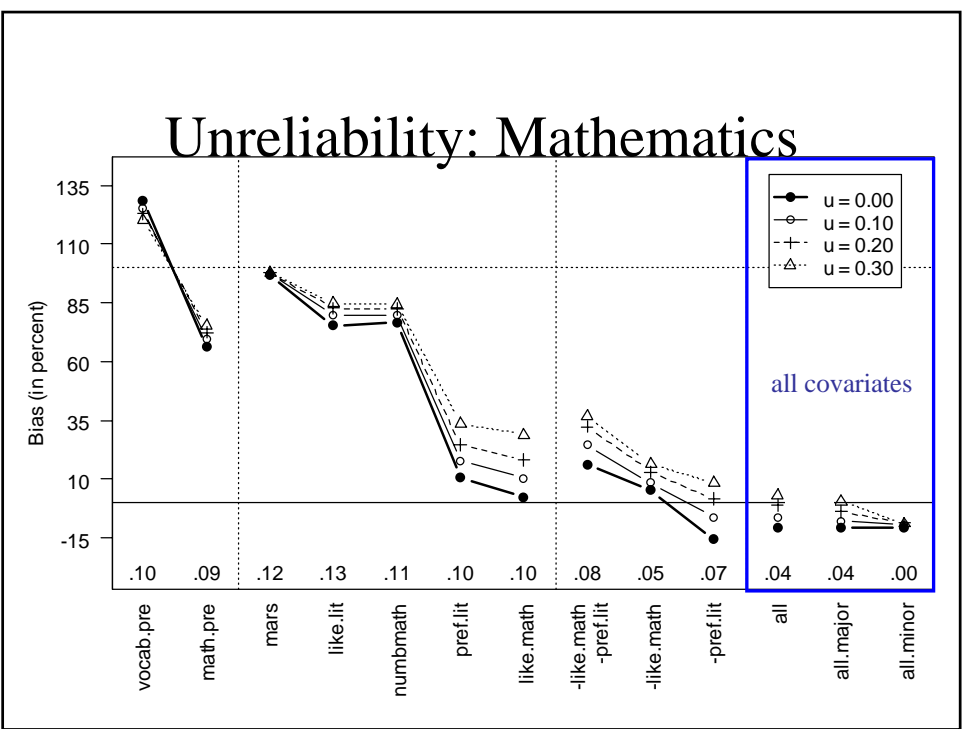
Unreliability: Mathematics



Unreliability: Mathematics



Unreliability: Mathematics



Consider 3 examples

- Aiken, West et al.
- Diaz & Handa
- Bloom, Michaelopoulos et al.

Selecting Intact Groups with Maximal Overlap: 1st Example

- Aiken et al. ASU--effects of remedial writing
- Sample selection in their Quasi-Experiment was from the same range of ACTs and SATs as in their experiment
- Differed by failure of researchers to contact them over summer and later registration
- What will the role of unobserved variables be that are correlated with these two features that differentiate randomly and non-randomly formed control units?
- Measurement framework the same in the experiment and quasi-experiment, as were the intervention and control group experiences

Results

On SAT/CAT, 2 pretest writing measures, the randomly and non-randomly formed comparison groups did not differ

- So close correspondence on observables w/o any need for statistical adjustment; and
- In Q-E, OLS test controls for pretest to add power and not to reduce bias
- Results for multiple choice writing test in SD units = .59 and .57--both sig.
- Results for essay = .06 and .16 - both non-sig

Diaz & Handa (2006)

- Progresá: Matched Villages with & w/o program
- One sample had to meet the village eligibility standards--in bottom quintile on test of material resources, but for several reasons not in RCT
- Eligible families in these matched villages were not different on pretest outcome measures but were on a few family characteristics
- But results for eligibles in the matched village analyses were similar whether covariates were added to control for family differences or not

Bloom, Michaelopoulos et al

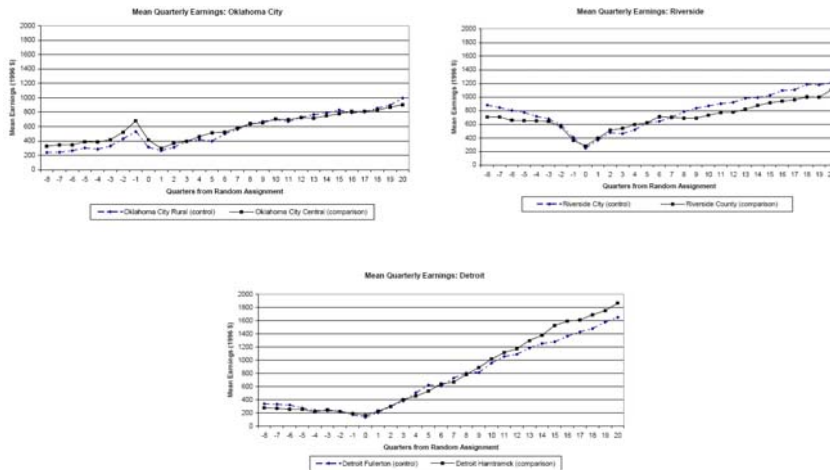
- RCT is on job training at 11 sites
- This analysis restricted to 4 within-city comparisons
- Non-random comparison cases chosen from job training centers in same city
- Measured in the same ways as treated at same times

Bloom, Michaelopoulos et al

Logic of the design is to compare ES from a randomly created control group with ES from a non-randomly formed comparison that shares the same treatment group

- Treatment group is a constant and can be ignored, comparing only the two types of control differently formed
- Issue is: Will the randomly and non-randomly formed control groups differ over 8 pretest observation points after standard statistical adjustments for any differences in mean or slope

Results: 3 within-city Samples



At non-Local Comparison Sites

- Randomly and non-randomly formed comparison groups were not equivalent at pretest
- OLS, propensity scores, Heckman selection models, random growth models--all failed to give the same results as the experiment under these conditions
- But the more the pretest time points, the less the bias

Selecting Intact Groups locally matched on pretest outcomes

- Bloom et al's within-city non-equivalent controls achieved comparability with the randomly formed experimental controls. That is, there was
- No bias across 3 of the 4 within-city samples; nor for the weighted average of all 4 sites
- So, overlap on observables was achieved through the sampling design alone, precluding need for statistical adjustments
- Moreover, there was bias in across-state comparisons, and it could not be adjusted away statistically with the data and models used

Example of Comparison Group Choice to limit Non-Equivalence

- Comer Detroit study as an example
- Sample schools in same district; match by multiple years of prior achievement and by race composition of school body--why?
- Choose multiple matches per intervention school, bracketing so that one close match above and the other below intervention schools

The Trade Offs here are...

- Identity vs. Comparability. We cannot assume that siblings are identical, for example. They have some elements of non-shared genes and environments.
- Comparability vs. Contamination. Closer they are in terms of space and presumed receptivity to the intervention, the greater the risk of contamination.
- To reduce an inferential threat is not to prevent it entirely.

What is a Local, Focal Non-Equivalent Intact Control Group

- Local is easy, as is intact; but focal?
- Highest correlates of outcome, especially pretest and especially over time, But also
- Identical Twins
- Fraternal Twins
- Siblings
- Successive Cohorts within same Organization

Conclusions of this Section

- Aiken et al, Bloom et al. and Diaz & Handa created non-equivalent control groups that were not different on pretest and other observables from the treatment group
- The populations so achieved did not affect study results, implying unobservables played no causal role. But we never know in other cases
- Sampling designs can often be created to limit population differences by being local, focal and also preferring to select intact groups over cases

Analysis of Workhorse Design when the Comparison Groups are clearly Different on Empirical Criteria

The Two Generic Strategies

- Modeling the outcome, like covariance analysis
- Modeling selection, like propensity scores

Modeling the Outcome

- One approach to the analysis of nonequivalent control group designs is to try to fully model the outcome, such as ANCOVA.
- This suffers from two problems
 - Specification error
 - Errors in Pretests

General Principle

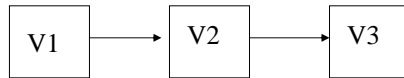
- It is generally known that you can obtain an unbiased estimate of a treatment effect if you can
 - Know the selection model or the outcome model completely
 - Measure it perfectly
- One of the reasons that regression discontinuity works is that it can meet these two criteria.
- Other designs cannot do so.

Specification Error

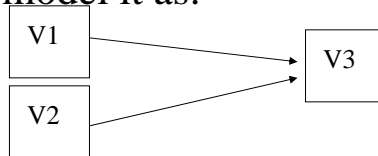
- Two forms
 - In any form of selection modeling (e.g., propensity score analysis), this is the omission of a variable correlated with outcome that is important to how people selected into conditions.
 - In any form of a complete model of the outcome (e.g., LISREL), this is the omission of a variable that is correlated with both treatment and outcome.

Example of Specification Error

- If the true relationship is:



- But you model it as:



- You will get a biased estimate of results.

Errors in the Pretest

- This refers to measurement error in the measurement of the pretest.
- Such measurement error is nearly always present, and it biases results:

All three figures show the relationship between two variables for each of two groups. The top figure shows two variables that have no measurement error, and so are perfectly correlated. Notice that the regression lines are parallel, that all dots are on the regression line, and that scores at pretest are the same as scores at posttest.

To generalize, the pretest can be a pretest on the posttest, but it can also be *any* covariate, and the logic in these three figures will apply.

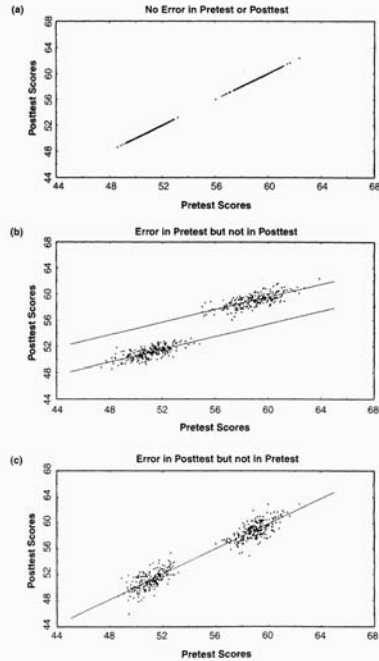


FIGURE 7.8 The effects of errors in pretests and posttests.

The middle figure shows what happens when we add random measurement error to the pretest but not to the posttest. Now, each dot is displaced horizontally to the right or the left by the error. As a result, the regression lines have different intercepts, when in fact the two groups are not different at all.

To generalize, efforts to use pretest covariates to reduce bias in quasi-experiments are problematic if the covariate has error.

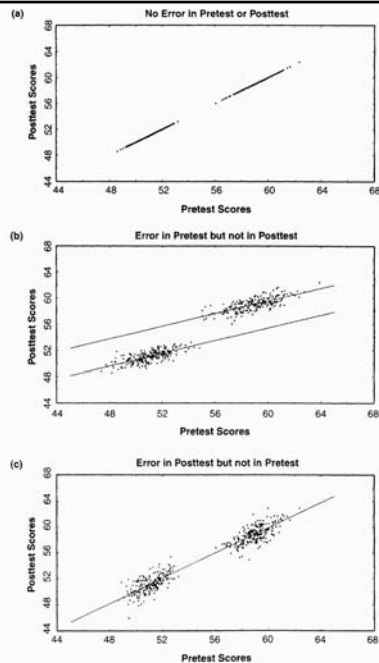


FIGURE 7.8 The effects of errors in pretests and posttests.

The bottom figure shows the same two groups, this time with errors in the posttest but not in the pretest. Now each dot is displaced vertically either up or down by error. But the two groups still have the same intercept (as they should since there is no effect).

The lesson: Errors in pretest covariates can bias results. So OLS is not good at correcting bias.

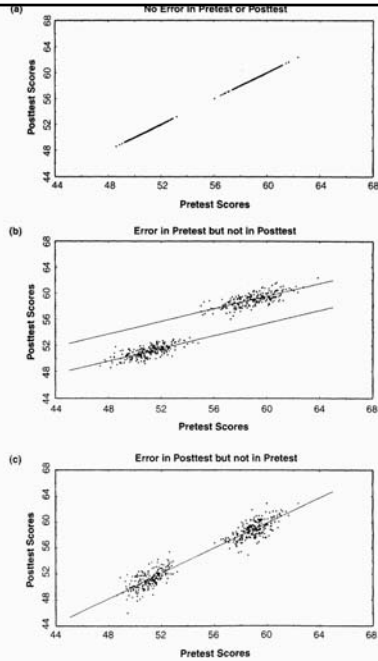
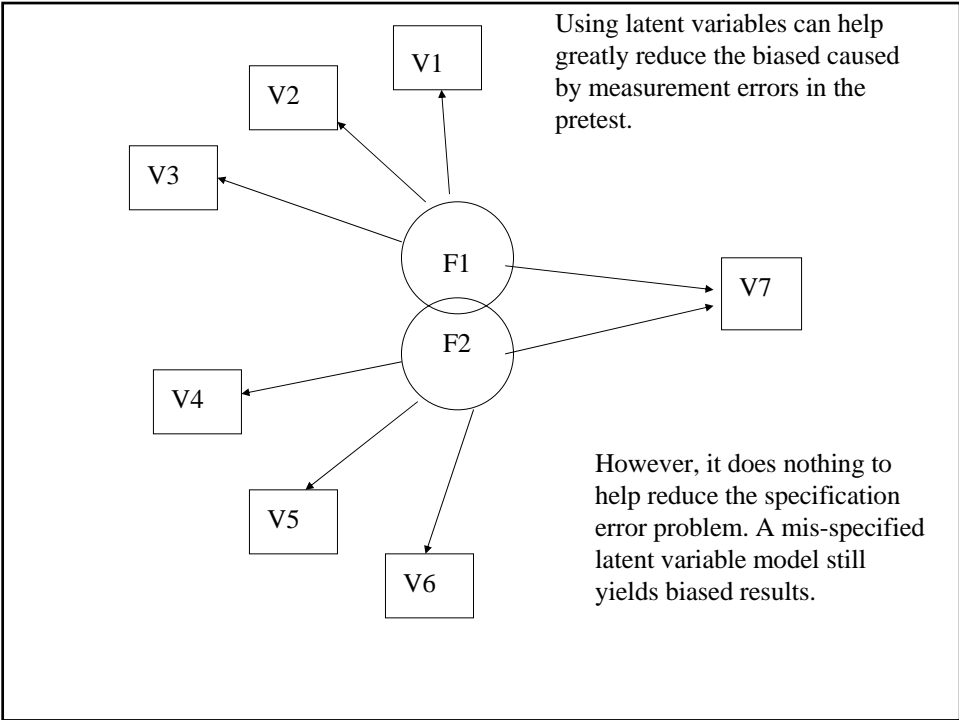


FIGURE 7.8 The effects of errors in pretests and posttests

226

Correcting for Measurement Error

- Using reliability coefficients to disattenuate correlations
 - But no reliability estimate is actually a perfect measure, so this will not work completely.
- Using Structural Equation Modeling to model relationships between latent variables rather than observed variables.
 - Requires multiple observed measures for each latent variable:



Propensity Scores and Quasi-Experiments:

What we will do

- 1. Describe Propensity scores
- 2. Design of Test of the validity of propensity scores using a 4-Arm Variant of the earlier Within-Study Comparison-- Shadish, Clark & Steiner JASA (in press)
- 3. Provide Test Results and link them to results from summarizing the 3-Arm tests

Propensity Scores: What are they?

- 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.

Propensity Scores

- In a randomized experiment, the true propensity score is .50 for each person.
 - In practice, it will vary from .50 due to sampling error.
- In a quasi-experiment, the true propensity score is unknown, but is presumed not to be .50.
 - If treatment = 1 and control = 0, then a propensity score closer to 1.00 (e.g., .83 is a prediction that the person is more likely to be in the treatment group, etc).

Estimating Propensity Scores

- Logistic Regression
 - Most widely used
 - Sensitive to nonlinearities in predictors
- Classification Tree
 - Not sensitive to nonlinearities in predictors
- Ensemble Methods
 - Bagging (Bootstrapped Aggregating)
 - Done on subset of people, classification on other subsets, repeatedly, assigned to branch by majority vote
 - Boosted Regression Trees
 - An iterating classification/regression strategy that iteratively improves estimates of the log odds of treatment assignment by adjusting weights of each case on each iteration
 - Random Forest
 - Classification tree approach that uses random subset of predictors, and iterates

Assessing Balance in Predictors

- The goal is NOT to get accurate prediction into groups.
- The goal is to create scores that, when used, create balance on predictors over groups within propensity score strata
- Crucial further assumption that the covariates are correlated with outcome

Old Approach (Rosenbaum & Rubin 1984)

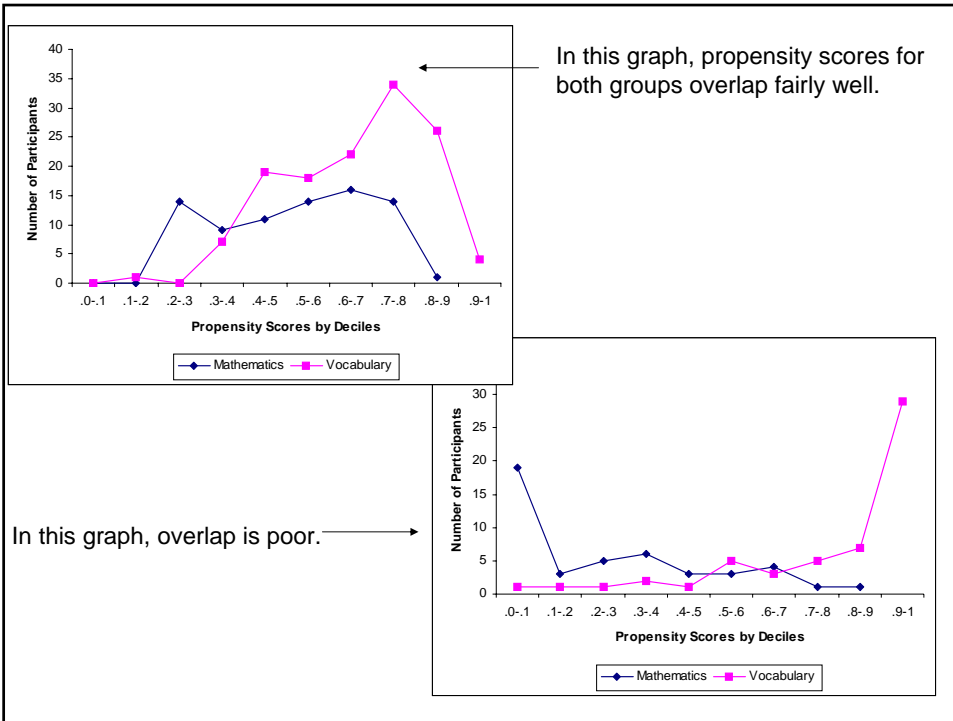
- A 2 x 5 ANOVA with
 - A treatment factor (treatment and control)
 - A propensity score strata factor (quintiles)
- Conduct the ANOVA for each pretest covariate and test interaction and main effect of treatment.
 - If more than 5% significant, then
 - Add covariates
 - Add nonlinear or interaction terms

New Approach (Rubin 2001)

- the standardized difference in the mean logit of propensity score in the two groups (B) should be near zero,
- the ratio of the variance of the propensity score in the two groups (R) should be near one, and
- after adjusting for the propensity score, the ratio of the variances of the covariates must be close to one, where ratios between 0.80 and 1.25 are desirable, and those smaller than 0.50 or greater than 2.0 are far too extreme.

Using Propensity Scores to Test Whether Nonequivalent Groups Should Be Compared

- If it is not possible to obtain balance in the covariates, then perhaps the groups are so nonequivalent that they should not be compared.
- One can graph the overlap in propensity scores to examine whether groups overlap enough to be worth comparing.



Estimation of Propensity Scores in Shadish et al.

- 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)

Balance: Old Criteria

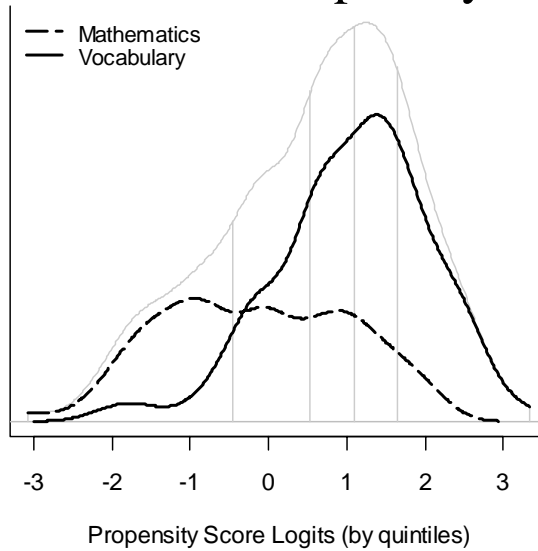
- Checked balance with 2 x 5 ANOVA, recomputed, rechecked.
 - Initially 10 of 30 covariates out of balance.
 - At the end, 0 of 30 interactions were significant and 0 of 30 main effects were significant.

Balance: New Criteria

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

Analysis	Propensity Score		Number of Covariates with Variance Ratio				
	B	R	≤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

Distribution of Propensity Scores



Reasons for Choosing Conditions

- N = 48 said they liked the condition (24 of whom chose vocabulary)
- N = 34 chose their condition to avoid the other condition (28 chose vocabulary, most were avoiding mathematics)
- N = 92 chose their condition for self-improvement (55 of whom chose vocabulary)
- N = 31 chose their condition because they had a high sense of self-efficacy that they could do the task (22 of whom chose vocabulary)
- The remaining N = 5 gave answers that could not be coded or were missing.
- Similarities to the achievement motivation literature?

Methods for Propensity Score Adjustments

- Matching
 - Selecting controls that match treatment subjects on propensity scores
 - Can have more than one match.
- Stratification on propensity score quintiles.
- ANCOVA
 - Sensitive to nonlinearities
- Weighting
 - Each observation is weighted by the inverse of its propensity score (tmt) or of $(1 - ps)$ for control, and then a standard weighted average is computed
- Here are results of the latter three adjustments

Results of the Test

Predictors of Convenience

- Bad practice: We also tested the effectiveness of propensity score adjustments based only on predictors of convenience (sex, age, ethnicity, marital status)
- Depending on how we did the analyses bias reduction ranged from 43% bias reduction to 5% bias increase.
- The importance of thoughtful selection of covariates in the design of the study.

Balance for Predictors of Convenience

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

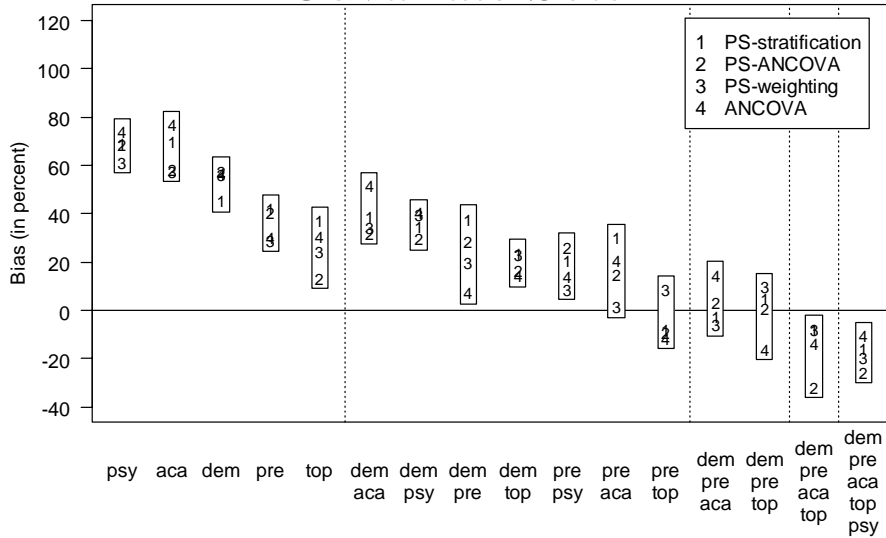
Analysis	Propensity Score		Number of Covariates with Variance Ratio				
	B	R	≤1/2	>1/2 and ≤4/5	>4/5 and ≤5/4	>5/4 and ≤2	>2
Before Any Adjustment	-1.13	1.51	0	2	17	6	0
After Stratification on Propensity Scores Constructed from All Covariates	-0.03	0.93	0	1	22	2	0
After Stratification on Propensity Scores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

Ordinary OLS Regression

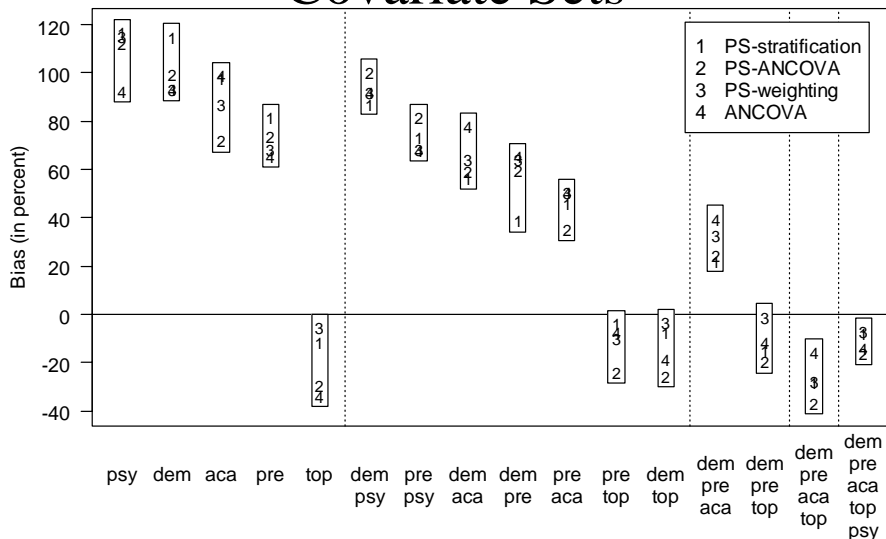
- 84-94% bias reduction just by entering covariates as predictors in regression.
- What good are propensity scores, then?
 - When creating a control group by matching.
 - To discover if there is enough balance to make adjustments valid.
 - When the assumptions of ANCOVA (e.g., linearity) are problematic.

Which Covariates Reduce the Bias? Two Pathways

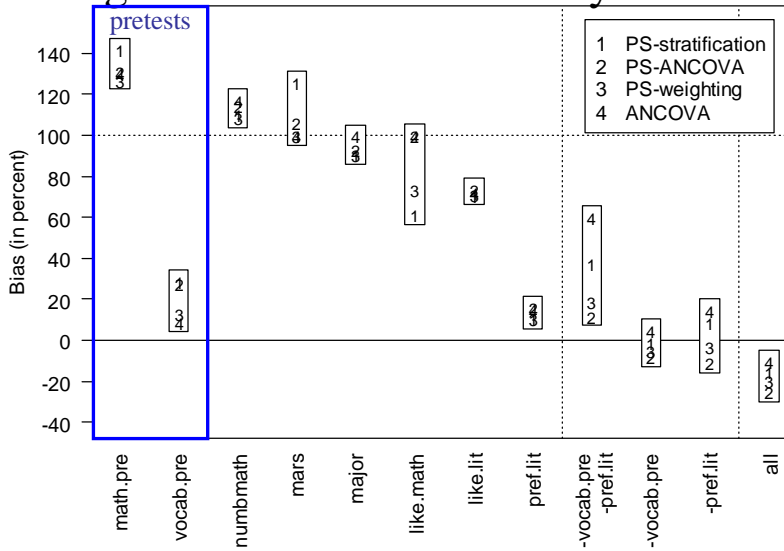
Remaining Bias: Vocabulary Covariate Sets



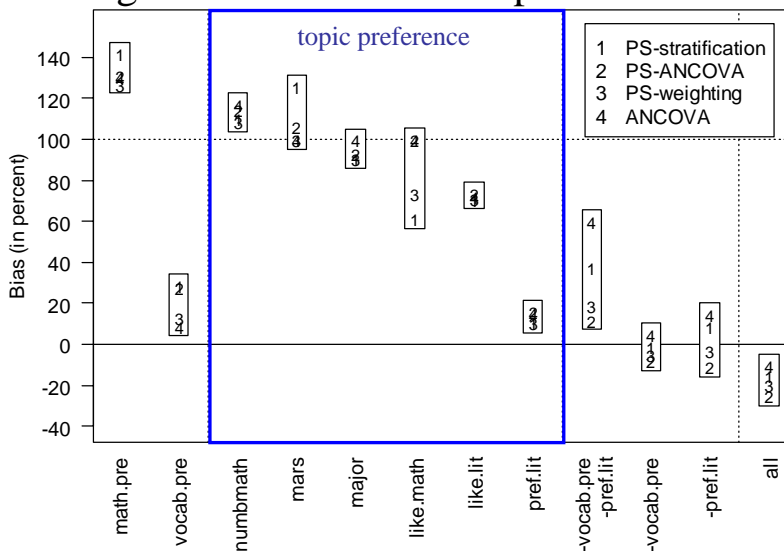
Remaining Bias: Mathematics Covariate Sets



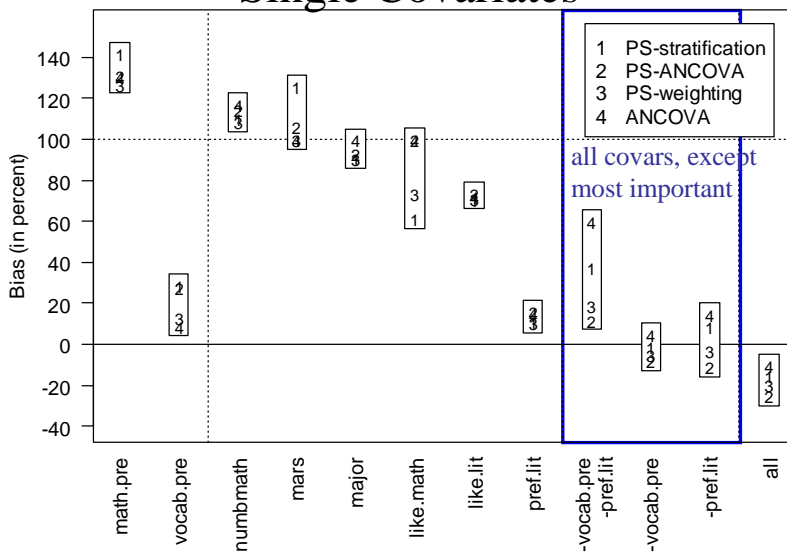
Remaining Bias: Vocabulary Single Covariates from Proxy-Pretests



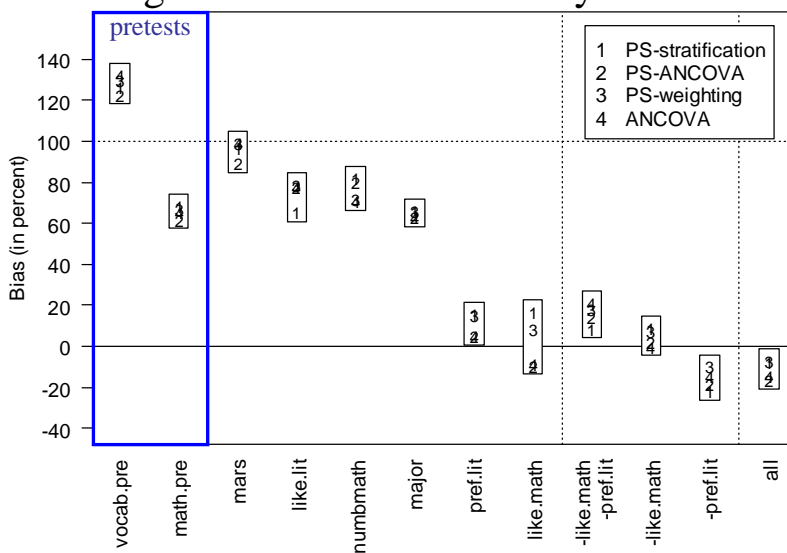
Remaining Bias: Vocabulary Single Covariates from Topic Preference



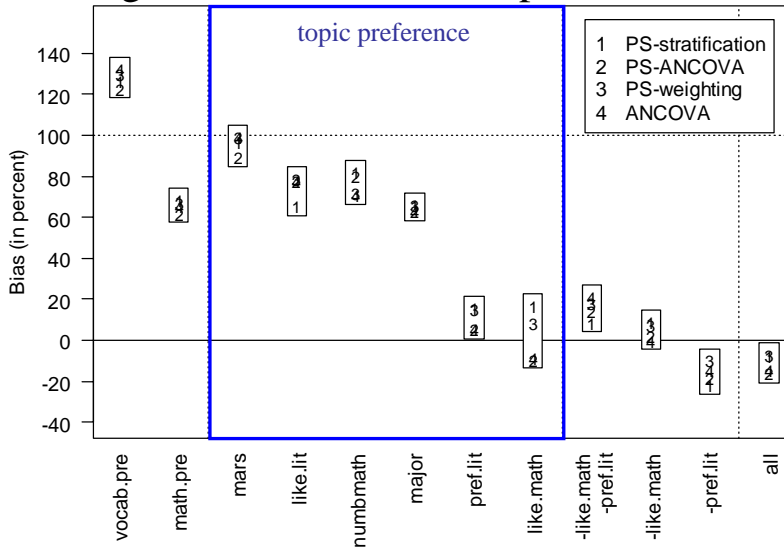
Remaining Bias: Vocabulary Single Covariates



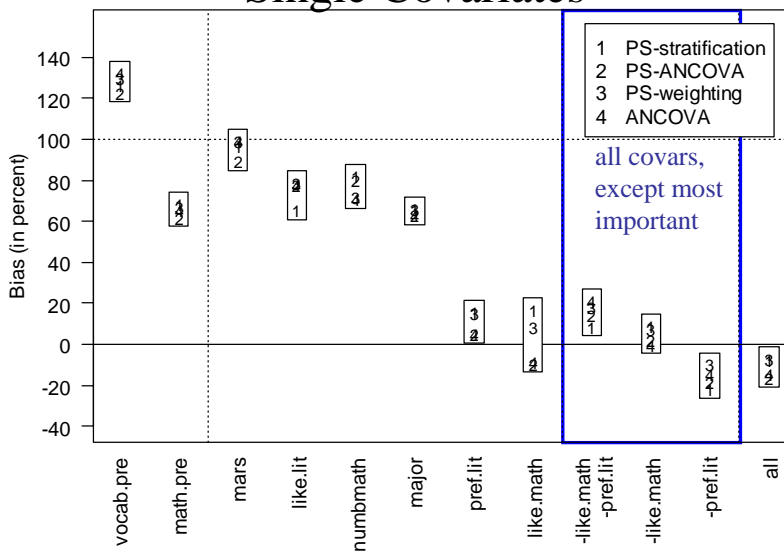
Remaining Bias: Mathematics Single Covariates from Proxy-Pretests



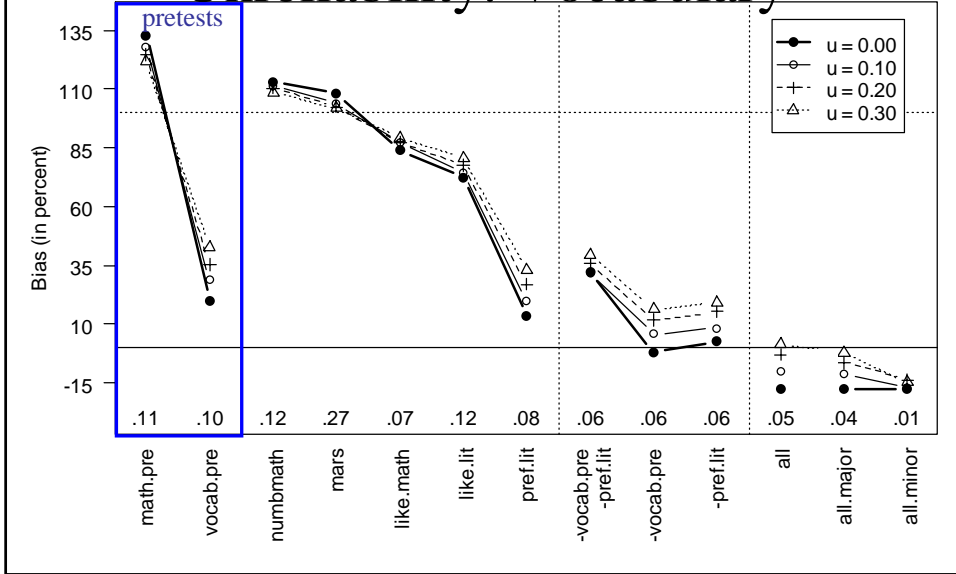
Remaining Bias: Mathematics Single Covariates from Topic Preference



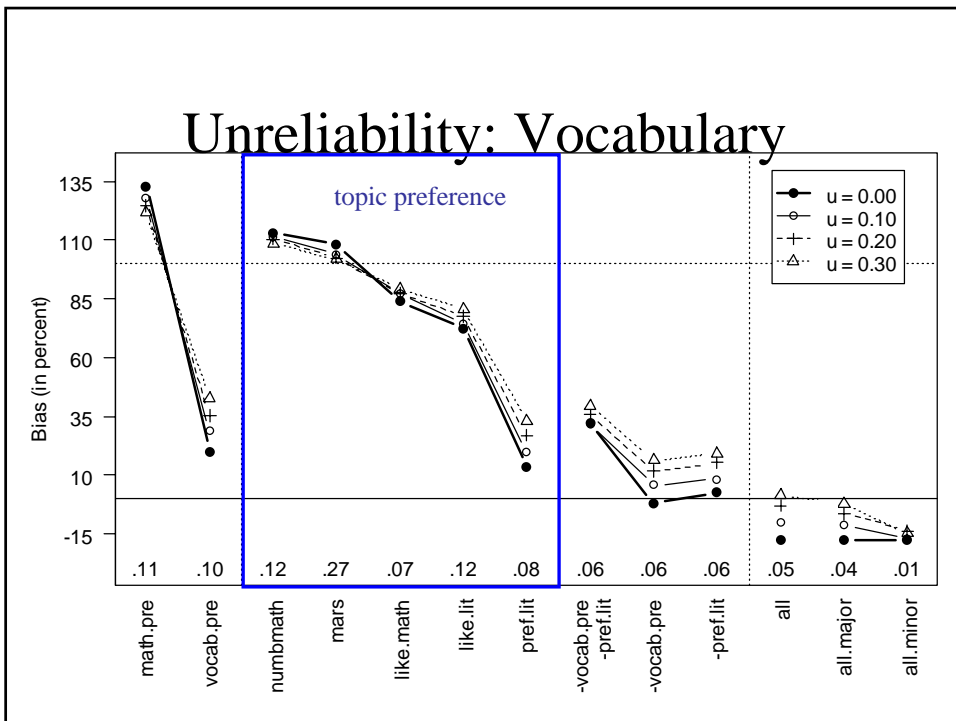
Remaining Bias: Mathematics Single Covariates



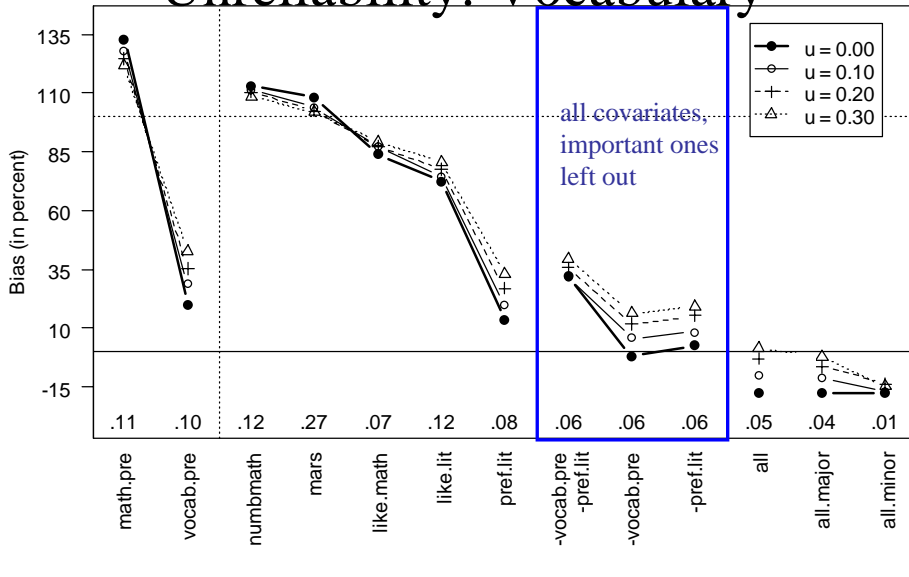
Unreliability: Vocabulary



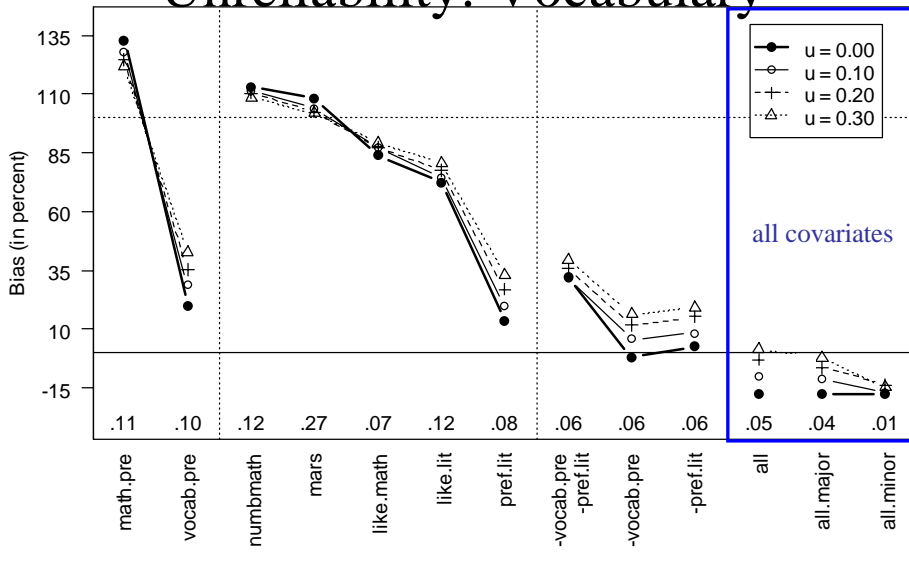
Unreliability: Vocabulary



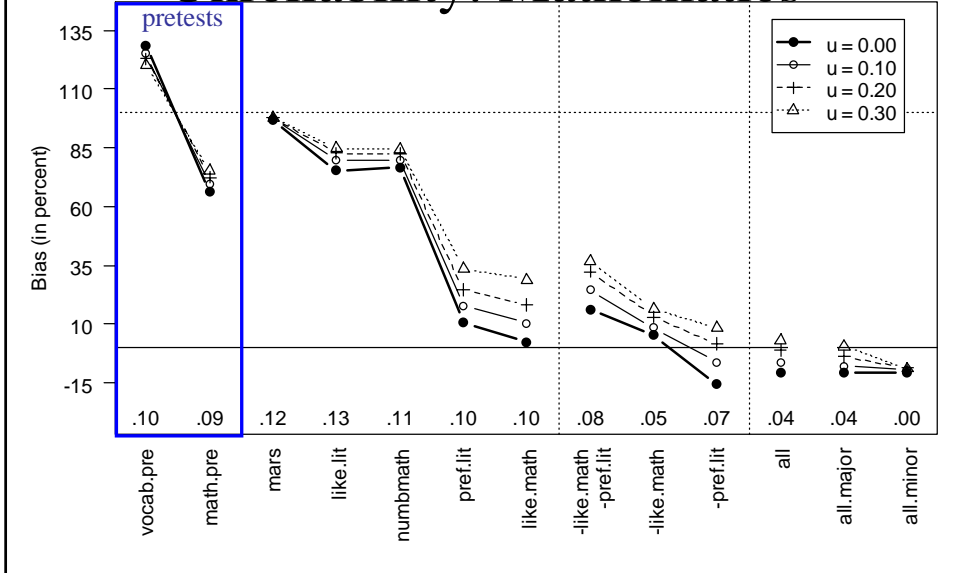
Unreliability: Vocabulary



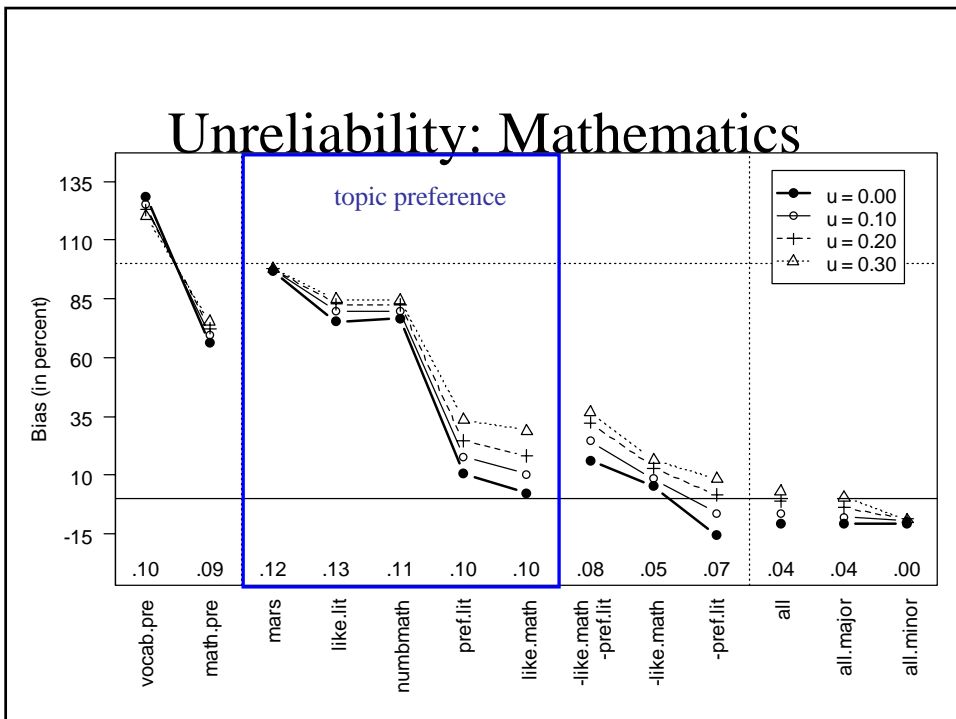
Unreliability: Vocabulary



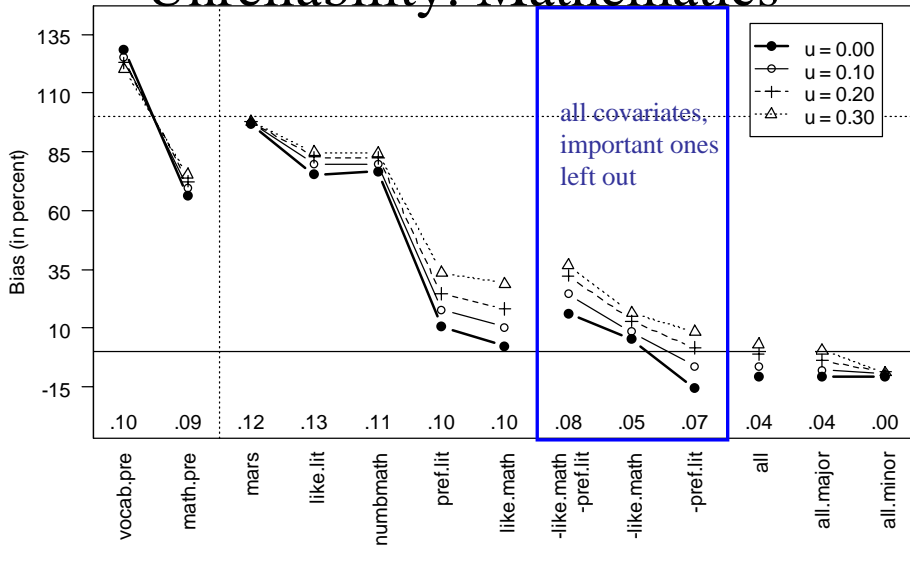
Unreliability: Mathematics



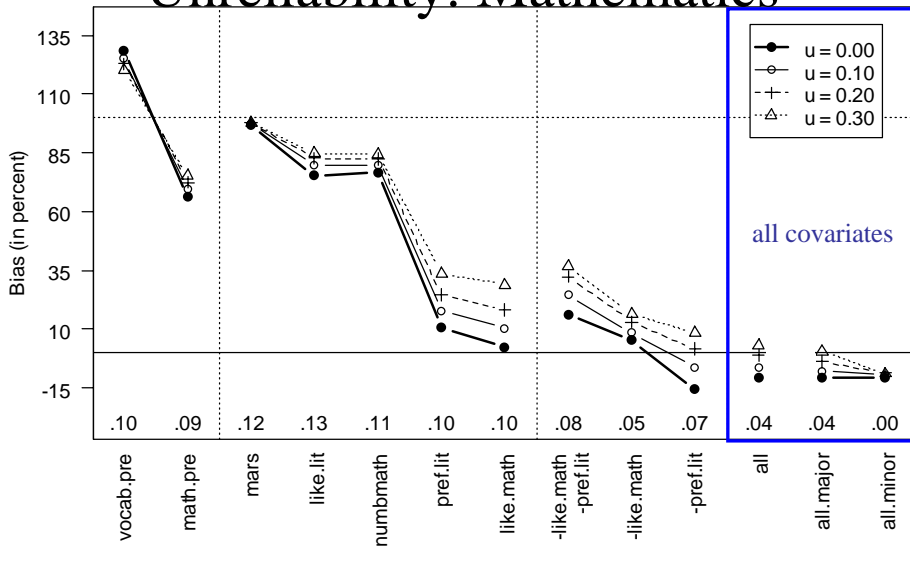
Unreliability: Mathematics



Unreliability: Mathematics



Unreliability: Mathematics



Discussion

- Results that propensity score adjustments to nonrandomized experiments might yield a reasonable estimate of what the effect would have been if these same participants had instead been randomly assigned to these same conditions using these same measures.
- However, OLS and LISREL did as well? Would yet other methods?

Limitations of PSM: Computational

- Is there a canonical methodology for constructing and analyzing propensity scores?
 - Rosenbaum and Rubin (1984), logistic regression, balancing, and stratification
 - Rapidly developing methodology
 - Construction by classification trees, boosted regression, bagging
 - Rubin (2001) tests for balance etc
 - Analysis by weighting, stratification plus weighting
 - How much balance is enough balance (5%)?
 - Important because experience is that the results are somewhat sensitive to these variations

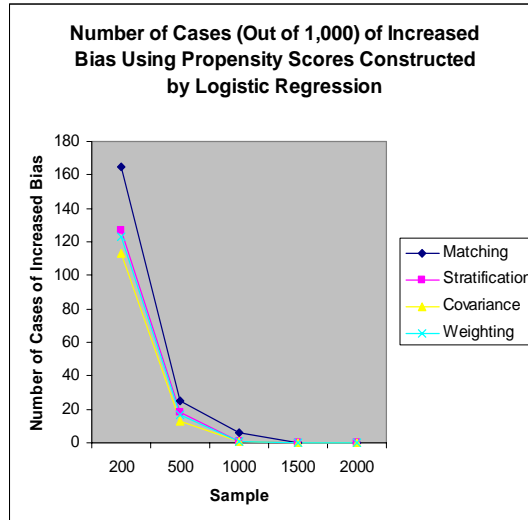
Limitations of PSM: Internal Validity

- Balance is necessary but not sufficient
- Strong ignorability is sufficient - hidden bias
- But not easily testable
- Depends on covariate choice
- Two related paths to covariate choice
- But never can be certain
- OLS seems to work as well in this and other applications

Limitations of PSM: Sample Size

- Propensity scores are said to work best with “large samples”, but there is little data about how large is large.
 - Rubin says his experience is $N > 300$
 - Luellen’s dissertation

Jason Luellen's Dissertation



Limitations of Shadish et al.

- Laboratory Setting
- Short-acting Treatment
- Perhaps simpler Selection Process than in many other situations
- Are Shadish et al results replicable in more representative studies?
- Examine two literatures summarizing within-study comparison literature with 3-Arm Designs

Glazerman et al (2003): Job Training

- Claim that quasi-experiment and RCT never give same result, but confounded by differences in location, time and manner of testing.
- Find that OLS and PMS not different in results
- Bias is reduced more when covariate set includes pretest and is richer in number
- Conclude that demographics is not enough

Cook, Shadish & Wong

- In 10 of 12 instances RCT and Q-E produce comparable results - summarized here - OLS and PMS comparable
- Demographics not work
- But effective are RD, focal, local matching and case matching when selection process is completely known

Beyond a Matching Perspective

- Towards one based on pattern matching, combining multiple design elements
- Begin with 3 examples
- Generalize from there

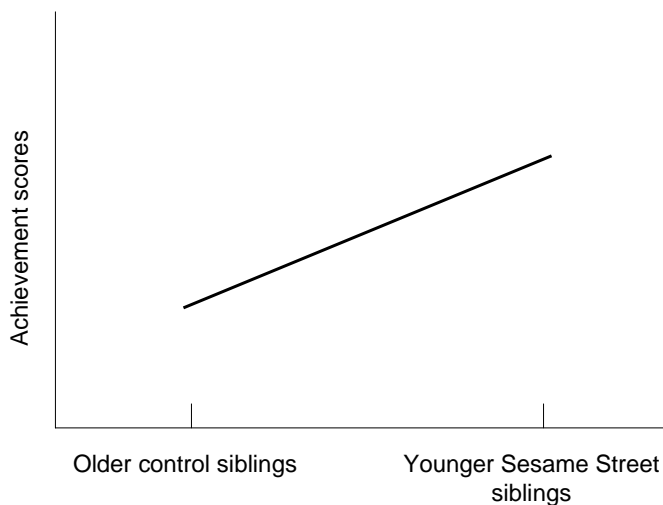
Minton's Dissertation

- Object: Evaluate Sesame St in 1st year
- Problem 1: Program already launched
- Problem 2: No pretest possible
- Problem 3: No money for original data coll.
- Setting: One kindergarten in NJ that built SS into its day regularly that has records on children and their families plus annual PPV assessment

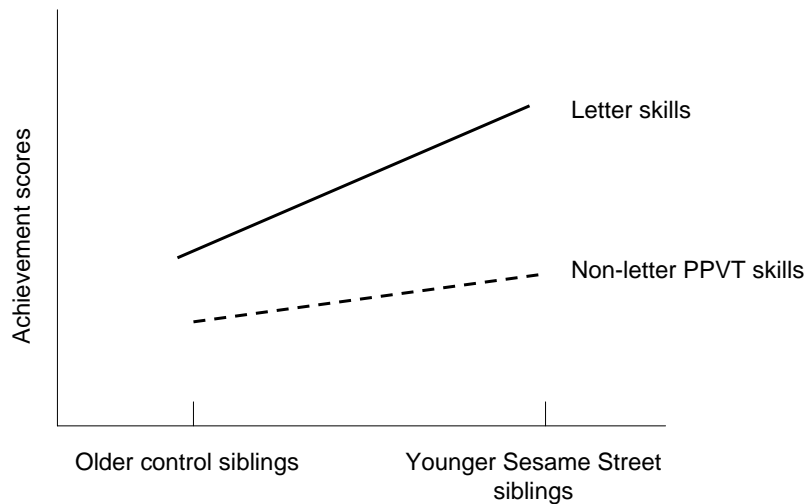
Question 1: What control group is possible?

- What control group to find, given program was very popular in its first year.
- Why is popularity a problem?
- Neighborhood kids who did not view
- Next-door kids of same age who not view?
- Older siblings in general
- Older sibs attending same kindergarten within last N years
- Older sibs attending same kindergarten last 2 years

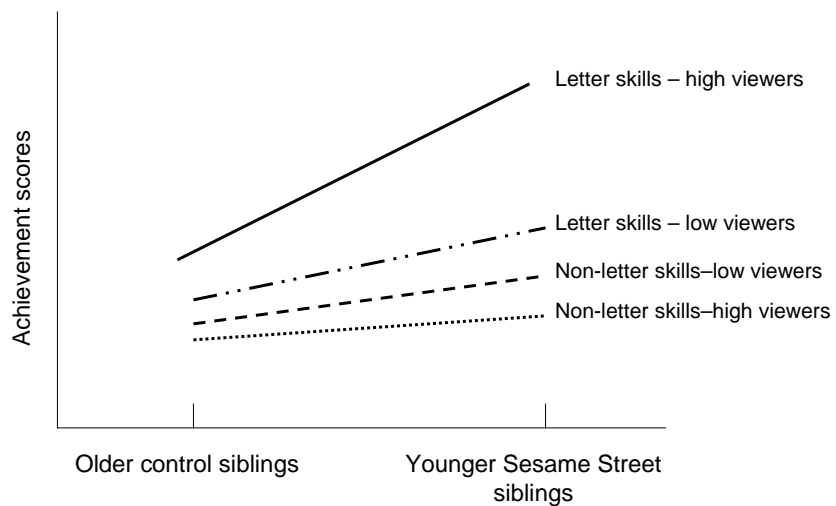
Minton (1975) Sesame Street Study - 1



Minton (1975) Sesame Street Study - 2



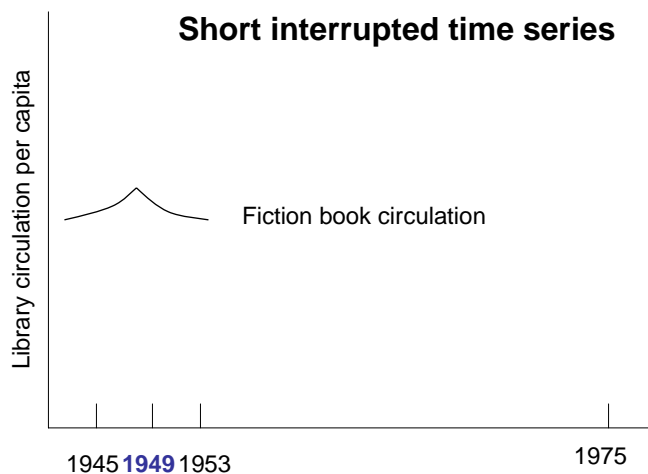
Minton (1975) Sesame Street Study - 3



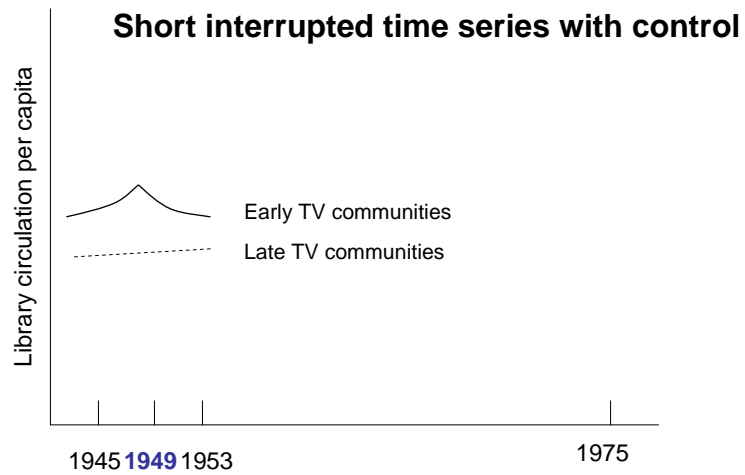
What has happened here?

- Single causal hypothesis of SS effective made to have multiple data implications
- These are meant to rule out alternative hypotheses and not to recreate same bias
- These implications here in the form of a difference in difference in differences
- Collect data and test hypothesis

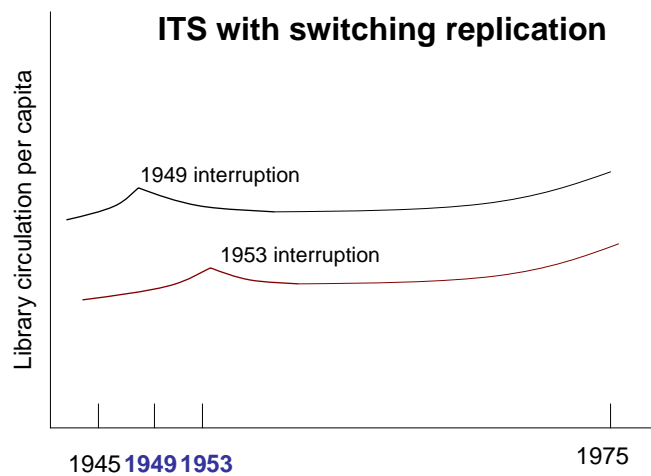
Parker et al. (1966) Effects of TV - 1



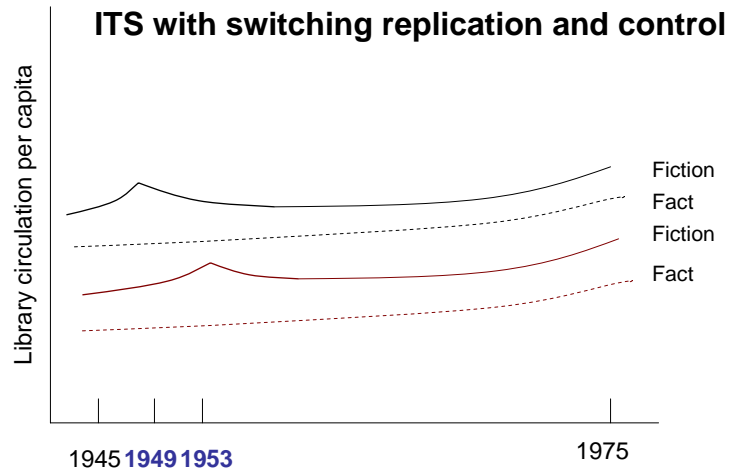
Parker et al. (1966) Effects of TV - 2



Parker et al. (1966) Effects of TV - 3



Parker et al. (1966) Effects of TV - 4



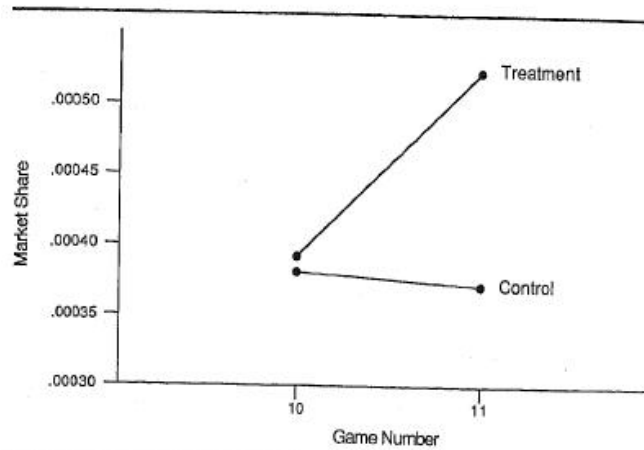
Reynolds and West's (1987) "Ask for the Sale" Experiment

From all stores selling lottery tickets, some stores volunteered (or not) to post a sign reading "Did we ask you if you want a Lottery ticket? If not, you get one free". So this is a basic nonequivalent control group design, with the control matched on zip, store chain, and pretest ticket sales.

<i>NR</i>	<i>O1</i>	<i>X</i>	<i>O2</i>

<i>NR</i>	<i>O1</i>		<i>O2</i>

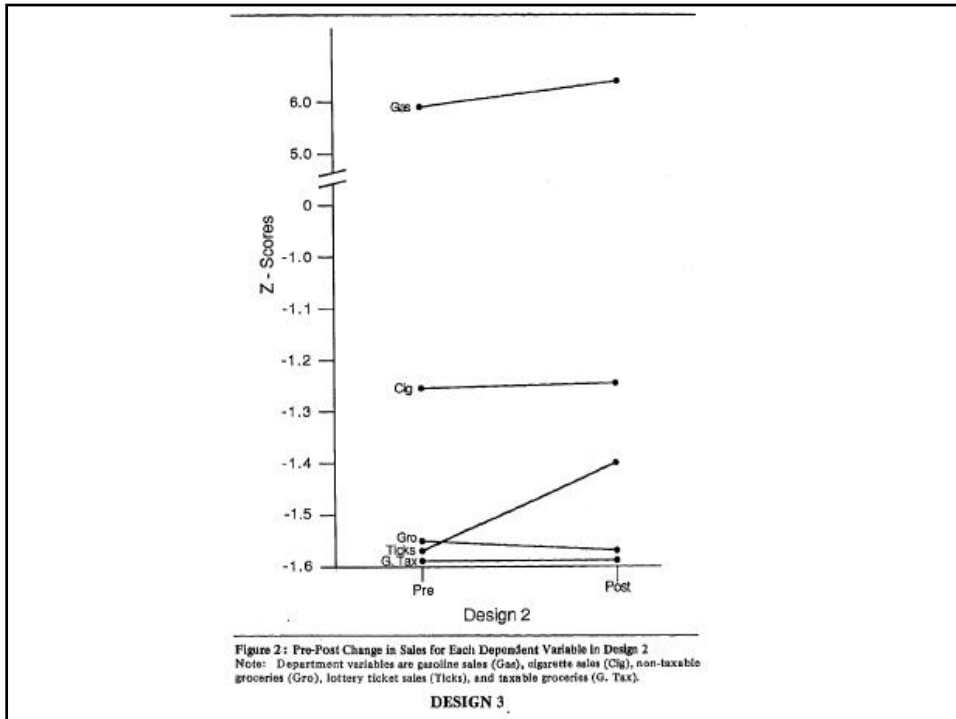
The Outcome of the Basic Design



But there might be many reasons besides treatment that caused treatment group sales to rise.

Adding a Nonequivalent DV

- They added three **nonequivalent dependent variables**, showing that the intervention increased ticket sales but not sales of gas, cigarettes, or grocery items.



Adding Multiple Baselines

- They located some stores in which the **treatment was initiated later than in other stores, or initiated and then removed**, and found that the outcome tracked the introduction of treatment over time while sales in the matched controls remained unchanged

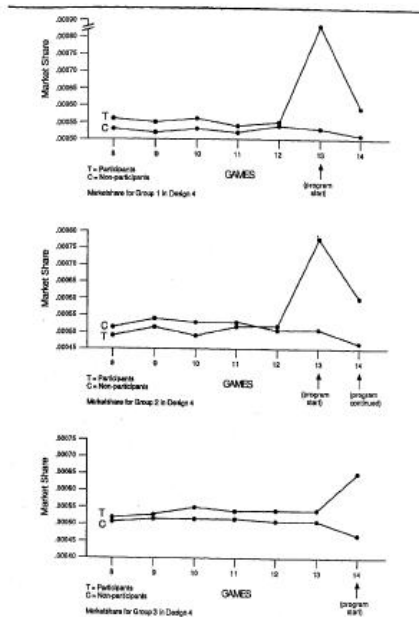


Figure 4: Mean Marketshare for Treatment and Corresponding Control Stores in Group 1, Group 2, and Group 3
 Note: Treatment stores in Group 1 began the program during game 13 and discontinued the program during game 14. Treatment stores in Group 2 began the program during game 13 and continued during game 14. Treatment stores in Group 3 began the program during game 14.

Adding Multiple Pretests and Posttests

- They added **multiple pretests and posttests** by examining mean weekly ticket sales for four weeks before and four weeks after the treatment started.

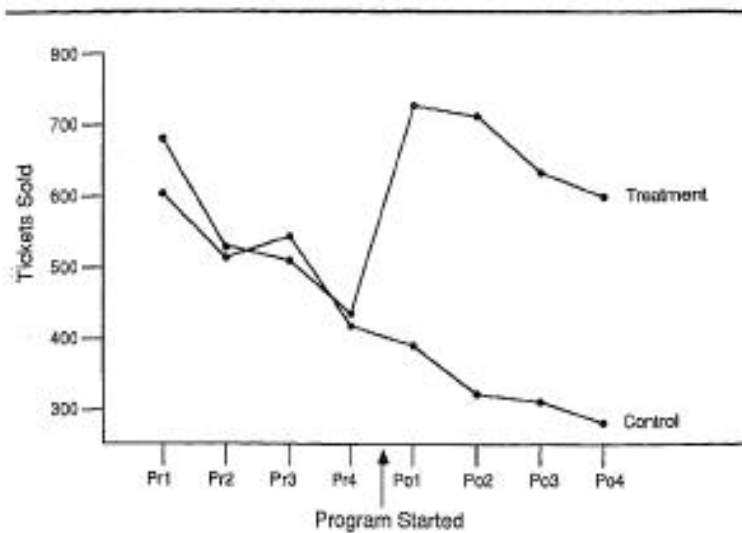


Figure 3: Mean Weekly Ticket Sales Pre- and Posttreatment
 Note: Raw mean weekly ticket sales for the 4 weeks prior to treatment (Pr1 – Pr4) and the 4 weeks after treatment (Po1 – Po4) are plotted for treatment (T) and control (C) stores.

Warning!

- This strategy requires:
- Sharp hypothesis - relevance to discontinuity
- Careful measurement - reliability and ceiling or floor effects
- Large samples (or large effects) because hypothesis is of a complex statistical interaction

Design Elements to be combined: Assignment

- Random Assignment
- Cutoff-Based Assignment
- Matching of many kinds

Design Elements to be combined:Treatments

- Switching Replications
- Reversed Treatments
- Removed Treatments
- Repeated Treatments

Design Elements to be combined: Measurement

- Single Pretest
- Pretest Time Series
- Proxy Pretests
- Retrospective Pretests
- Moderators with predicted Interactions
- Measuring Threats to Validity

Design Elements to be combined: Comparison Groups

- Single Non-Equivalent Groups
- Multiple Non-Equivalent Groups
- Twins/Siblings
- Cohorts
- Other Focal, Local Comparison Groups

Golden Rules

- You can't put right through statistics what you have done wrong by design
- Statistical adjustments work better the less non-equivalence there is to adjust away in the first place
- Since the work horse is so prevalent but so problematic, try to complexify the design through.....

So (1)

- Do an experiment; if not
- Do Regression-discontinuity study. If not,
- Do ITS with some sort of a comparison series. If not
- Do study combining multiple design element, including choice from focal local intact controls, case matching on pretest mean and slope, reintroduction of treatment at new time, non-equivalent DVs, etc.

So (2)

Don't be bamboozled by fancy models in Greek clothing.
Always translate them into structural design elements
before evaluating their utility for unbiased inference. That
will reveal what you have got