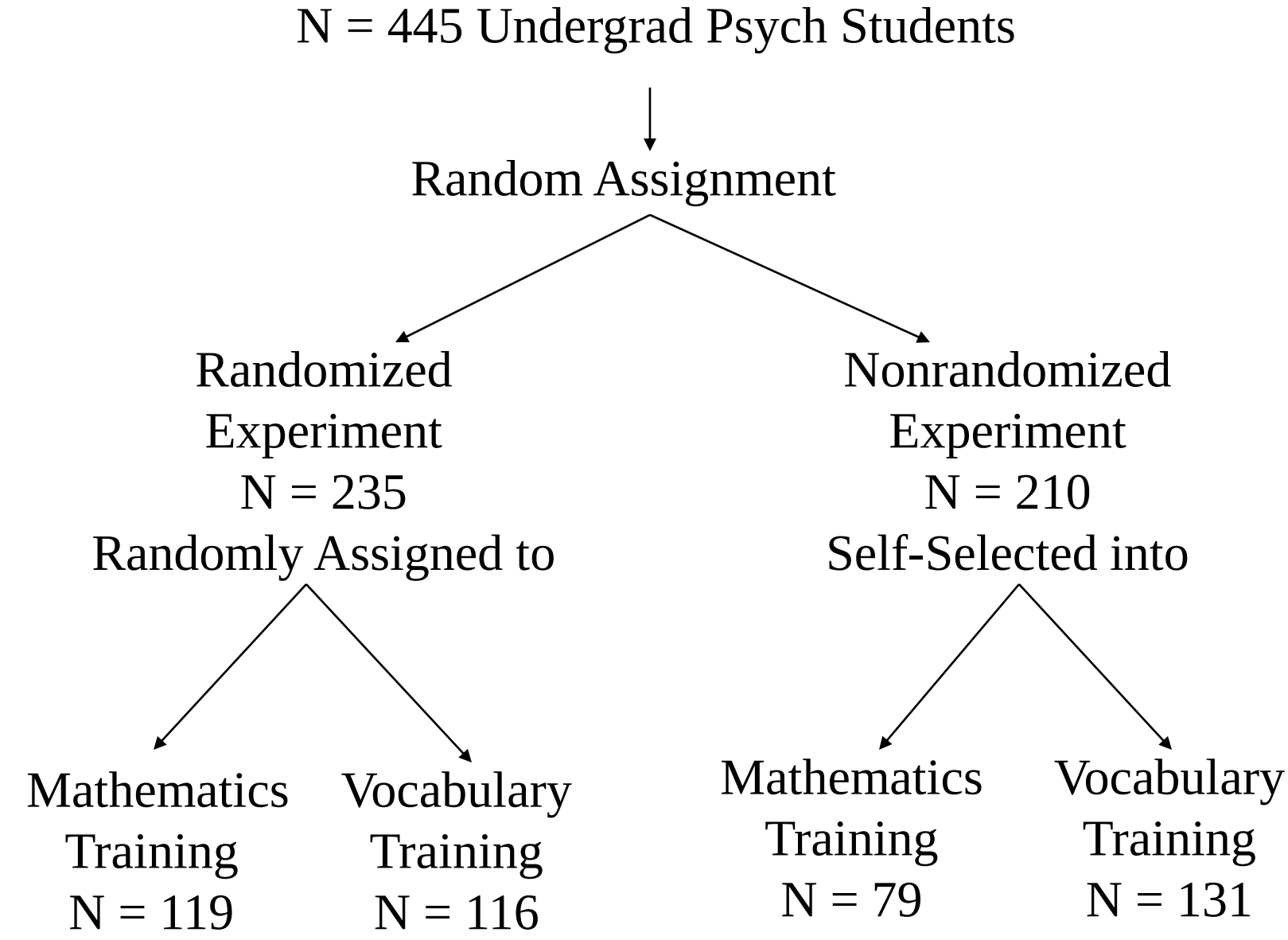
Propensity Score Analysis Example: The “Four Arm” Study

A Little Background

- Part of a program of research on the conditions under which nonrandomized experiments might approximate results from randomized experiments.
- But instead of just comparing RE and QE from past, started from scratch—randomly assign to RE and QE (a laboratory analogue)
- Was also the time PSA was just becoming popular, so I was learning about how to do it.
- Here is the design as we implemented it:



All Participants Post-tested on both Vocabulary and Mathematics Outcomes

More on the Design

- All participants pretested on a host of covariates
- All participants treated together without knowledge of the different conditions.
- All participants post-tested on both math and vocab outcomes.
- Complete treatment implementation and negligible attrition
- Used logistic regression to create propensity scores.
- Used several different analyses to see if they made a difference

More on the Covariates

- Carefully studying and measuring the selection process—why do people choose math or vocabulary?
  - We reviewed the literature—not much help.
  - We interviewed academic advising office
  - We ran a pilot study and then interviewed the students about why they chose
  - We used “common sense” theory
  - E.g., a lot of students will do anything to avoid math.
- Careful study of selection may be the most important thing
  - Remember the strong ignorability assumption (more in a moment)

More on Analysis

- We knew little about PSA
- We made mistakes
  - E.g., Sue Marcus/Paul Rosenbaum consult
    - no, the goal is not to maximize prediction of which conditions people chose.
    - It is to get balance on the pretest covariates
- We had to learn about everything from
  - Logistic regression to
  - Ensemble methods (bagging etc.)
- But in the end, it was an interesting study. Here are the results:

Basic Findings

- Analyst of QE masked to findings from RE
- Adjusted QE findings closely approximated RE findings
- The method of adjustment did not matter:
  - Propensity score stratification, weighting, covariance (but doubly robust matters)
  - Ordinary regression (no PSA)
  - Structural equation models[1]
- *Having good pretest measures did matter*

[1] Shadish (2013)

Results for Mathematics

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.
- Used Doubly Robust Adjustment (PSA and Covariates)
  - Adding covariates reduces standard error nontrivially.

Results for Vocabulary

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 also tested the effectiveness of propensity score adjustments based only on predictors of convenience (sex, age, ethnicity, marital status)
  - What happens if we ignore strong ignorability?

Bias Reduction Poor for Math Outcome

	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

And Poor for Vocab Outcome

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

- Do we need propensity score analysis at all?
- OLS ANCOVA results did as well as the more complicated propensity score methods.
- For example, look at the last row of the next two tables:

Results for ANCOVA adjustment for Mathematics

	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

Results for ANCOVA Adjustment for Vocabulary

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.
- None of our initially conceptualized models even remotely fit the data
- So we did atheoretical specification searches using CFI as the criterion for the best model (making sure that the model made conceptual sense, of course)
- Here is an example of a model:

Structural Equation Models Did Well in Subsequent Analyses

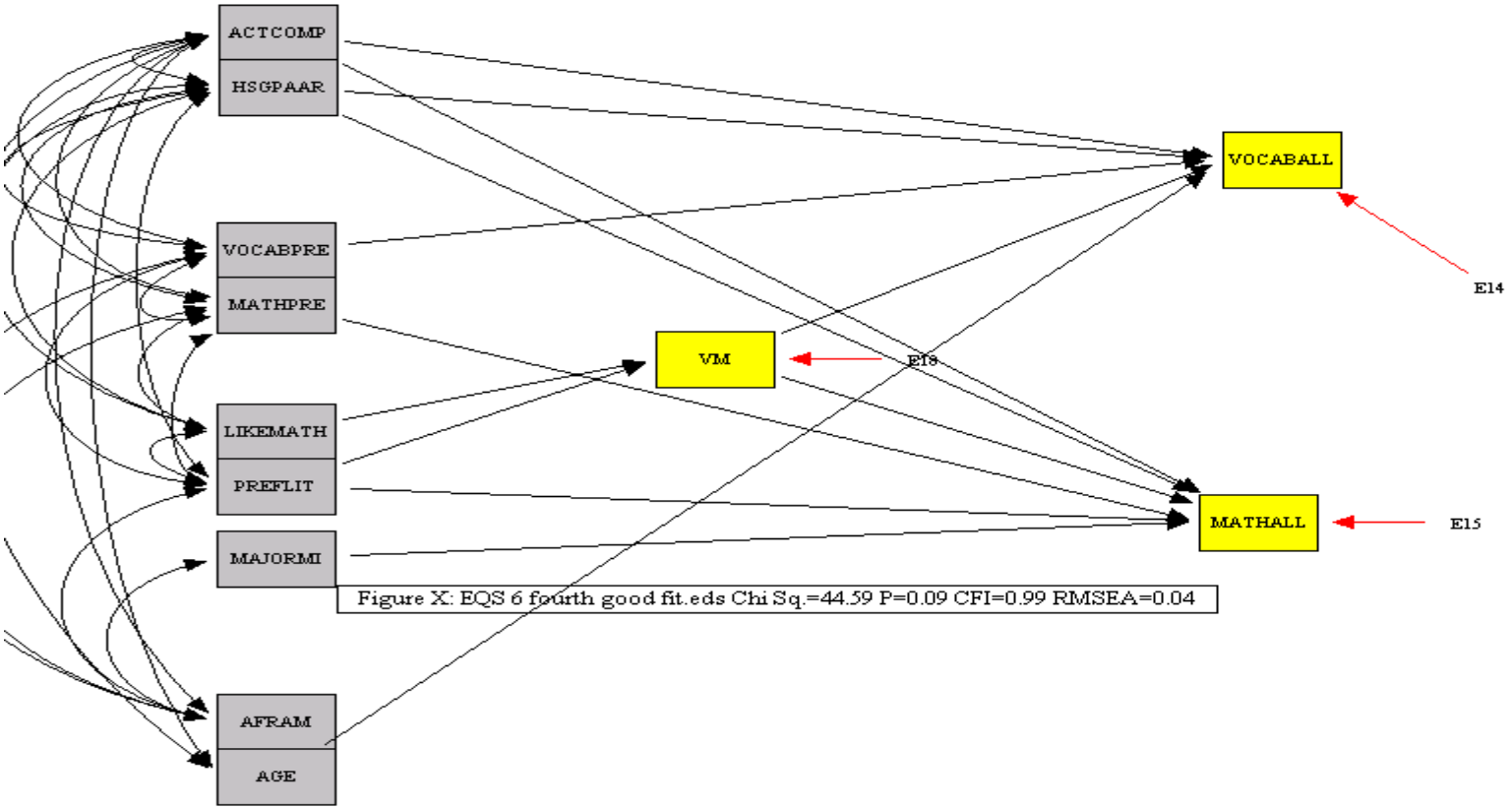


Figure X: EQS 6 fourth good fit.eds Chi Sq.=44.59 P=0.09 CFI=0.99 RMSEA=0.04

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

95% bias reduction

76% bias reduction

Observed Mediational Models

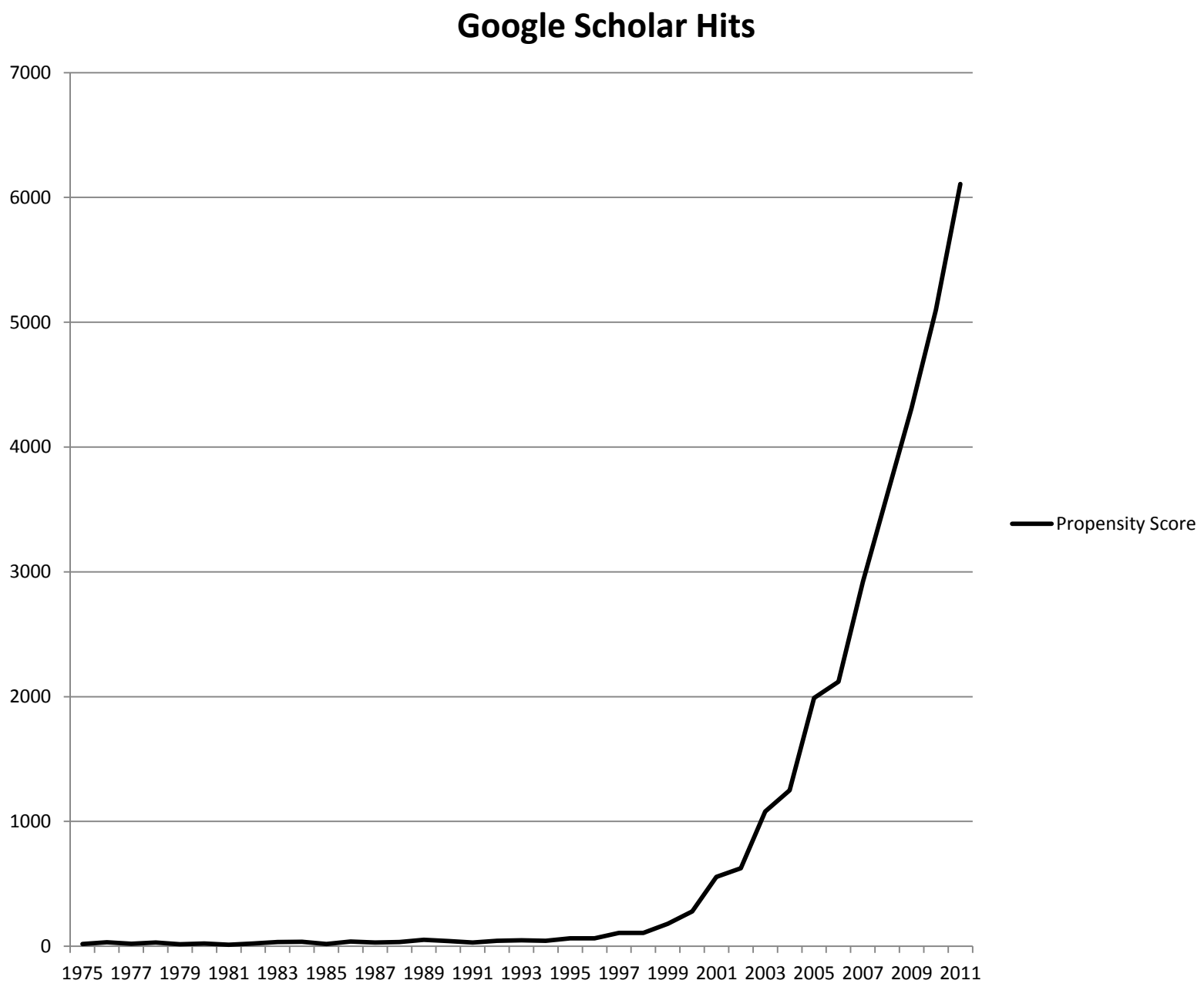
Fourth good fit	.983	3.96	8.43

Latent Variable Models

Model	CFI	Math Effect	Vocab Effect
Fourth good fit	.961	3.69	8.49

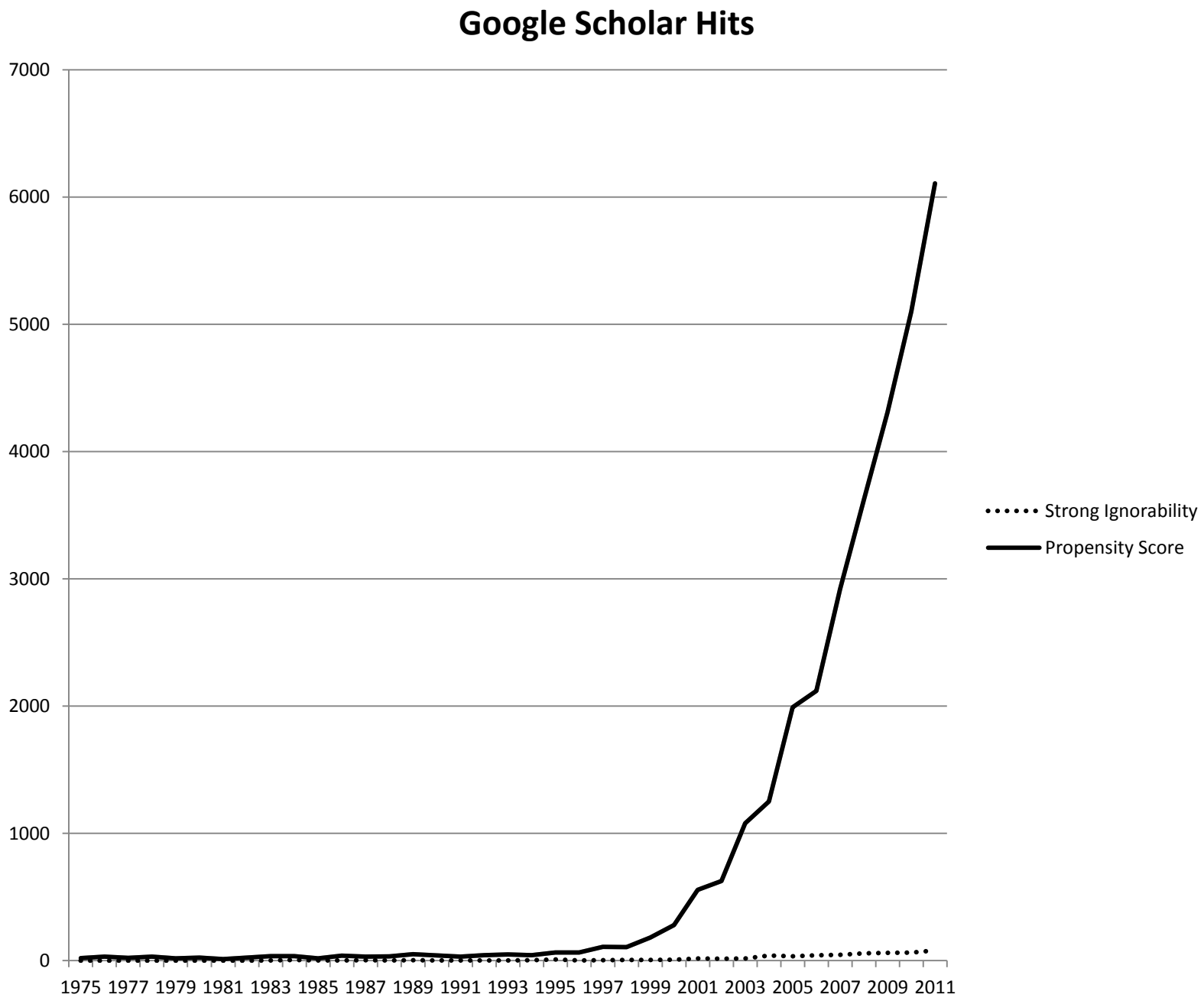
A digression into strong ignorability

- Good covariates are crucial to propensity score analysis (and I suspect for all other analyses as well) because of the strong ignorability assumption.
- We all know that propensity score analysis has become wildly popular in recent years:


Google Scholar Hits
7000
6000
5000
4000
3000
2000
1000
0
1975 1977 1979 1981 1983 1985 1987 1989 1991 1993 1995 1997 1999 2001 2003 2005 2007 2009 2011
Propensity Score

But what about strong ignorability

- Unbiased estimates from PSA depend on meeting the strong ignorability assumption.
  - Potential outcomes are unrelated to treatment assignment given the set of covariates X.
    - This assumption does not have a test
    - Balance ≠ strong ignorability
  - If not met, no guarantee you are reducing bias.
- Apparently, most PSA researchers strongly ignore strong ignorability:


Google Scholar Hits
7000
6000
5000
4000
3000
2000
1000
0
1975 1977 1979 1981 1983 1985 1987 1989 1991 1993 1995 1997 1999 2001 2003 2005 2007 2009 2011
Strong Ignorability
Propensity Score

Strong Ignorability and the "Right" Covariates

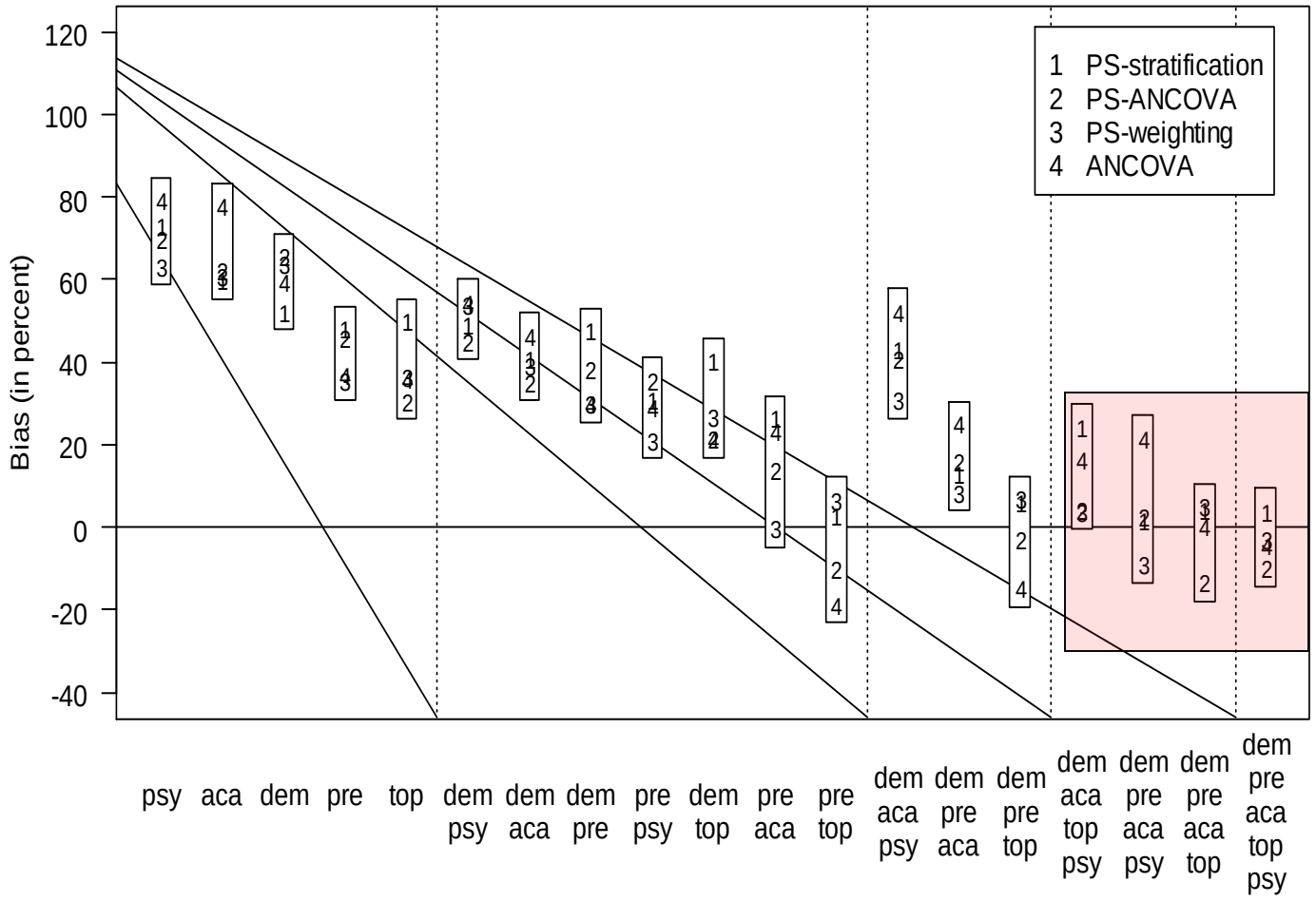
- Strong Ignorability implies you have covariates
  - Correlated with outcome and selection
  - That make assignment (condition) orthogonal to the potential outcomes at the start
- Covariate selection is the key
  - Either have the right variables that predict selection (hard to do—e.g., all boys schools), or
  - Have the right domains (latent constructs that the right variable is a part of).

Steiner, Cook, Shadish & Clark (2010), *Psychological Methods*.

A Reanalysis to See Which Covariates Make the Analysis Work

- Divide our 25 covariates into five classes
  - Demographics
  - Pretests
  - Psychological
  - Academic
  - Topical Preference
- Then see which make a difference
  - Singly
  - As constructs

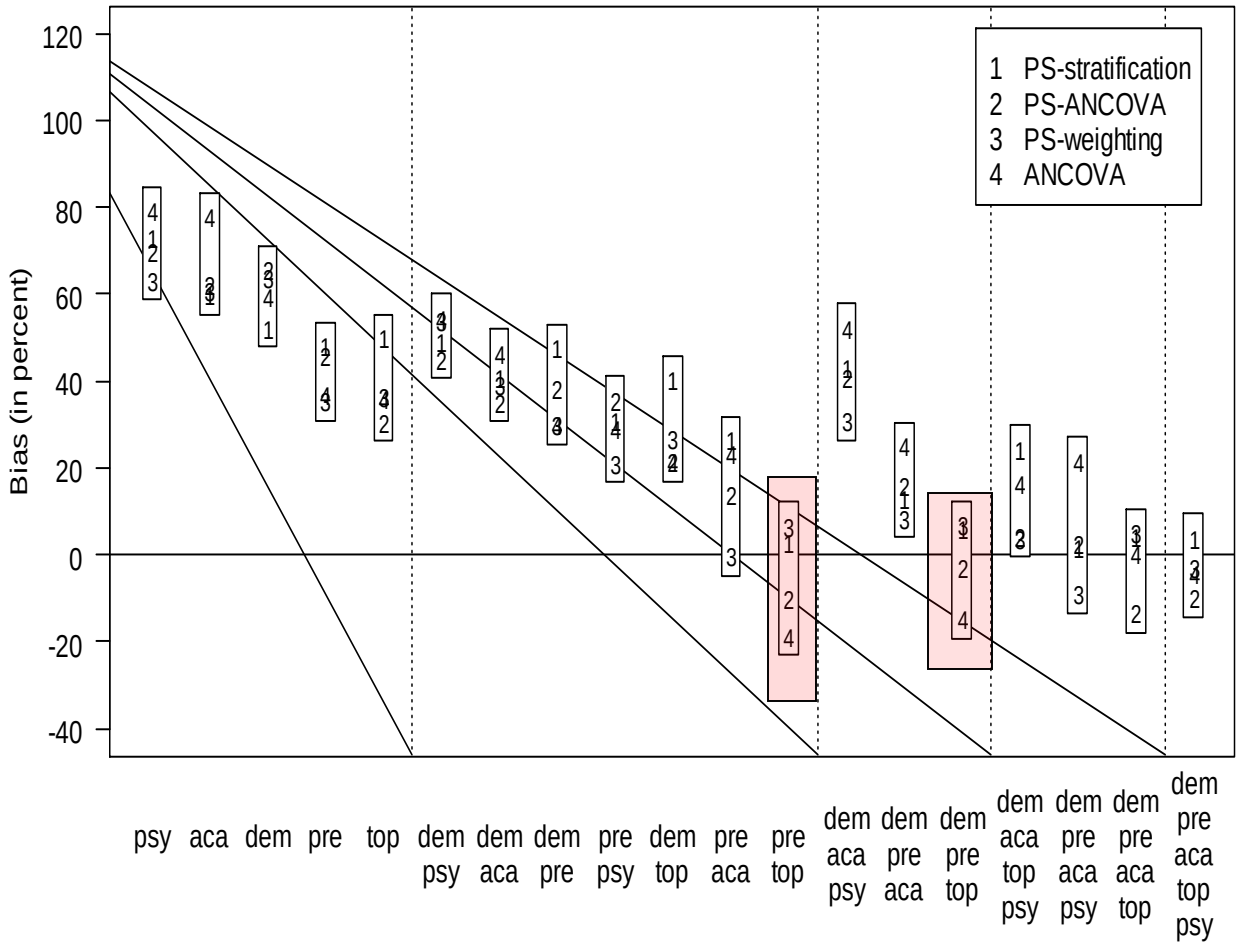
For the Effects of Vocabulary Training on Vocabulary Outcome:



Remember, we had done a good job of studying and measuring the selection process at the start

We get good bias reduction if we use most or all the covariates

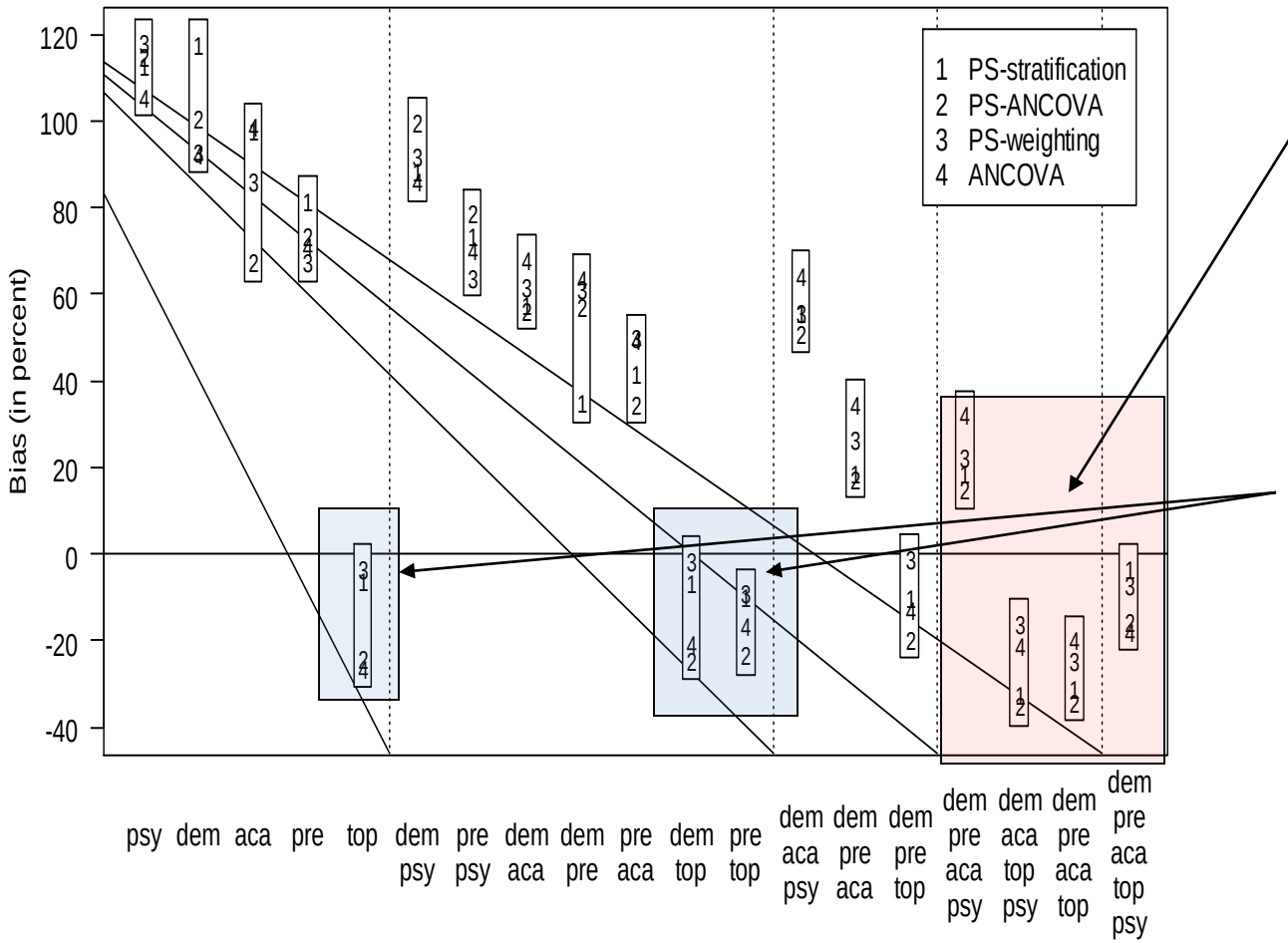
But



We also get good bias reduction using just a few of the "right" covariates.

- Topical Preference, and
- Proxy Pretest

For the Effects of Math Training on Math Outcome



Good bias reduction if you use all covariates.

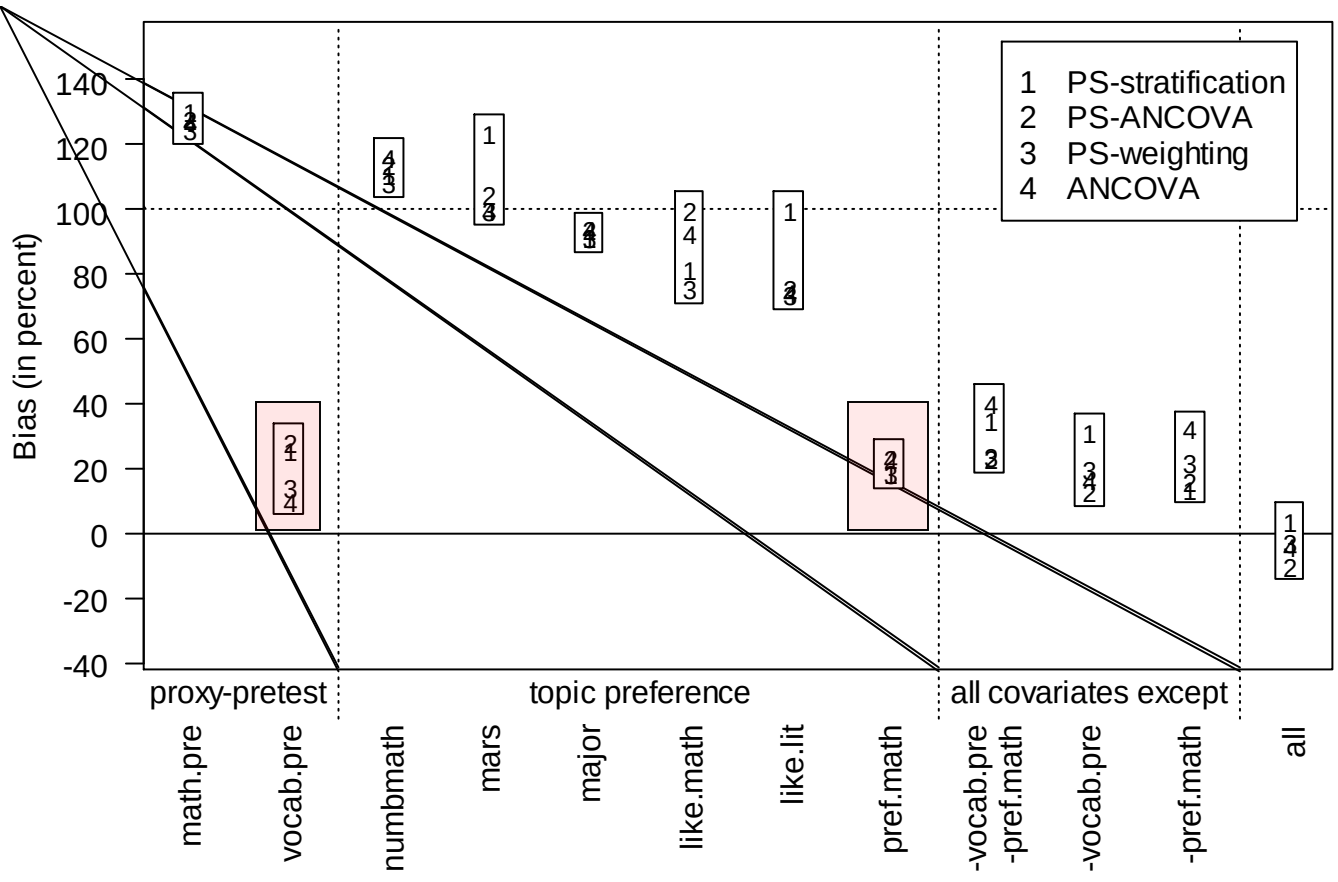
But also good bias reduction with the "right"ones.

In both math and vocab, topic preference and pretest were key.

How to Know the “Right” Covariate?

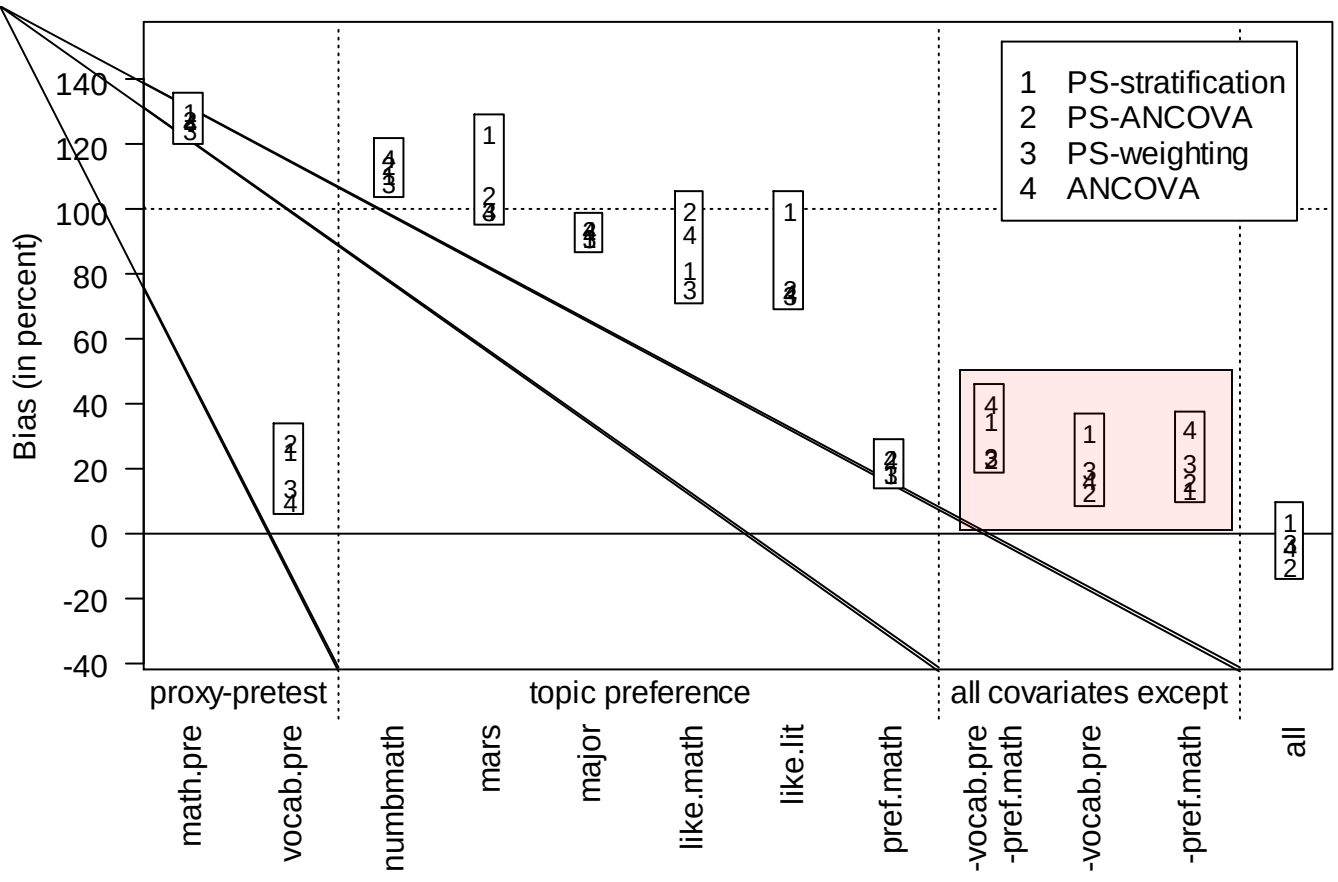
- Selection of pretest covariates should involve:
  - Past research on variables related to selection
  - Interviews with treatment providers and recipients (crucial!) about why they got the treatment they did
  - Common sense (the role of math anxiety)
- This process is hard and time consuming, but can generate a set of potential covariates.
- But what if that set does not include the “right” variables?

Vocabulary Bias Remaining using Individual Covariates



For bias reduction in vocabulary outcome, the key individual variables were preference for math and vocabulary pretest scores.

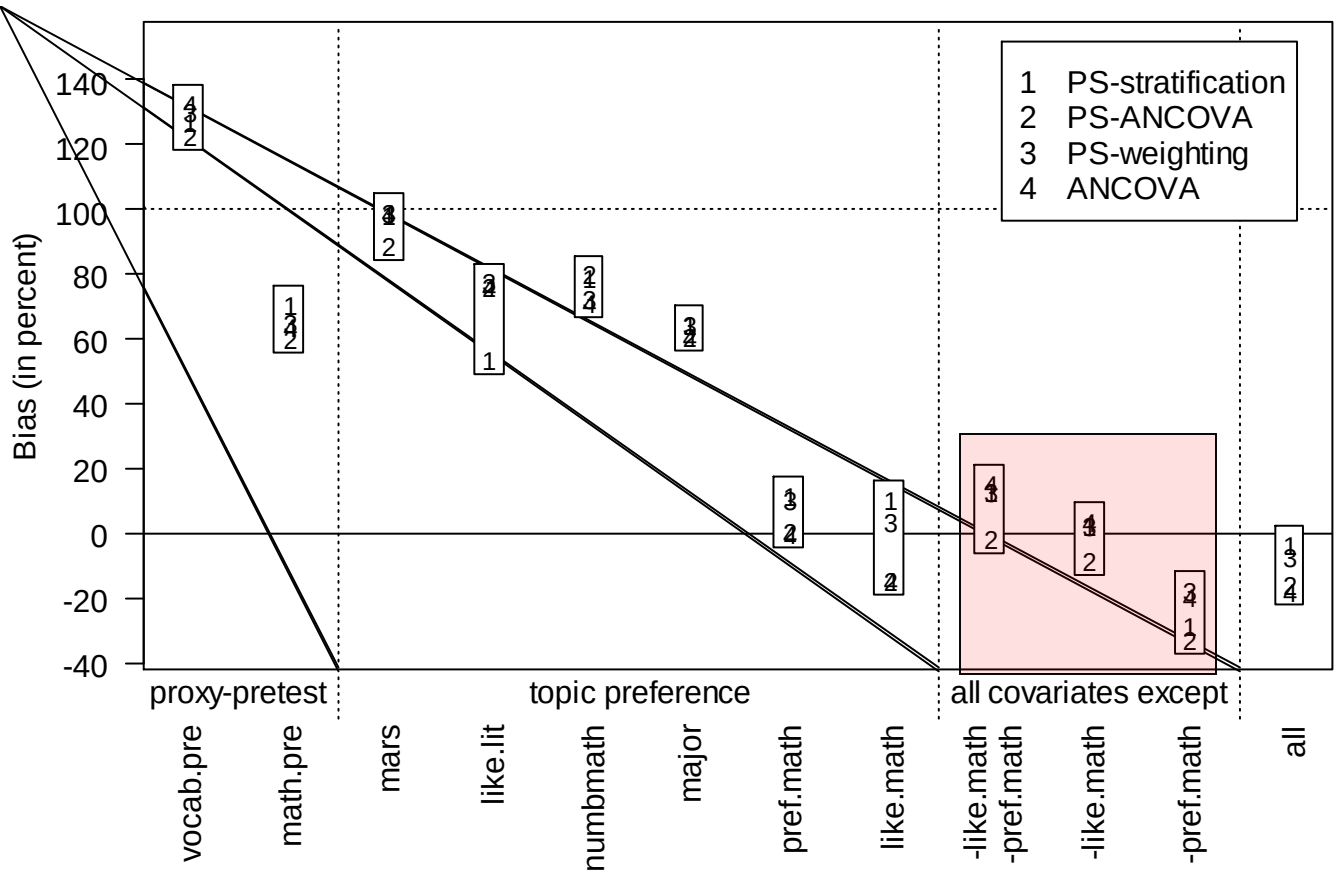
Vocabulary Bias Remaining using Individual Covariates



But bias reduction in vocabulary outcome was also good if a thoughtfully constructed kitchen sink didn't have the "right" variables.

Even if the kitchen sink contained only variables that individually did not reduce bias very well.

Math Bias Remaining using Individual Covariates



And the same was true for bias reduction in math outcome.

Lessons

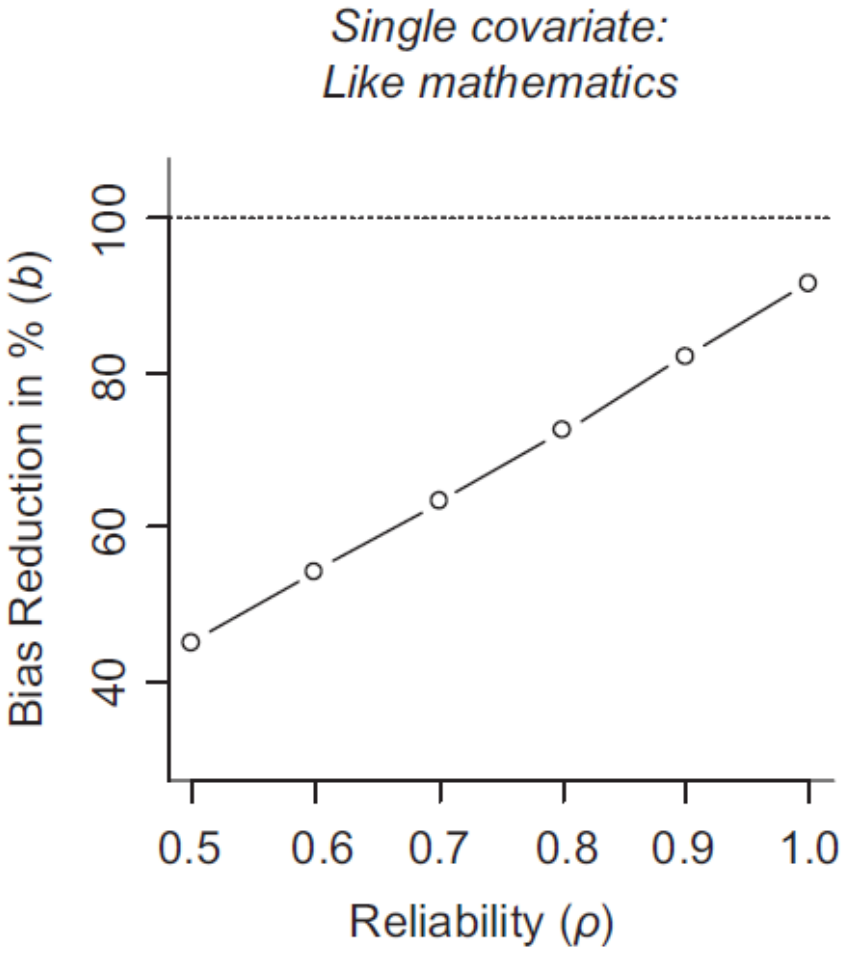
- Work hard to get the “right” variables
- But as important is getting *multiple measures* of the “right” constructs leading to selection, an easier task

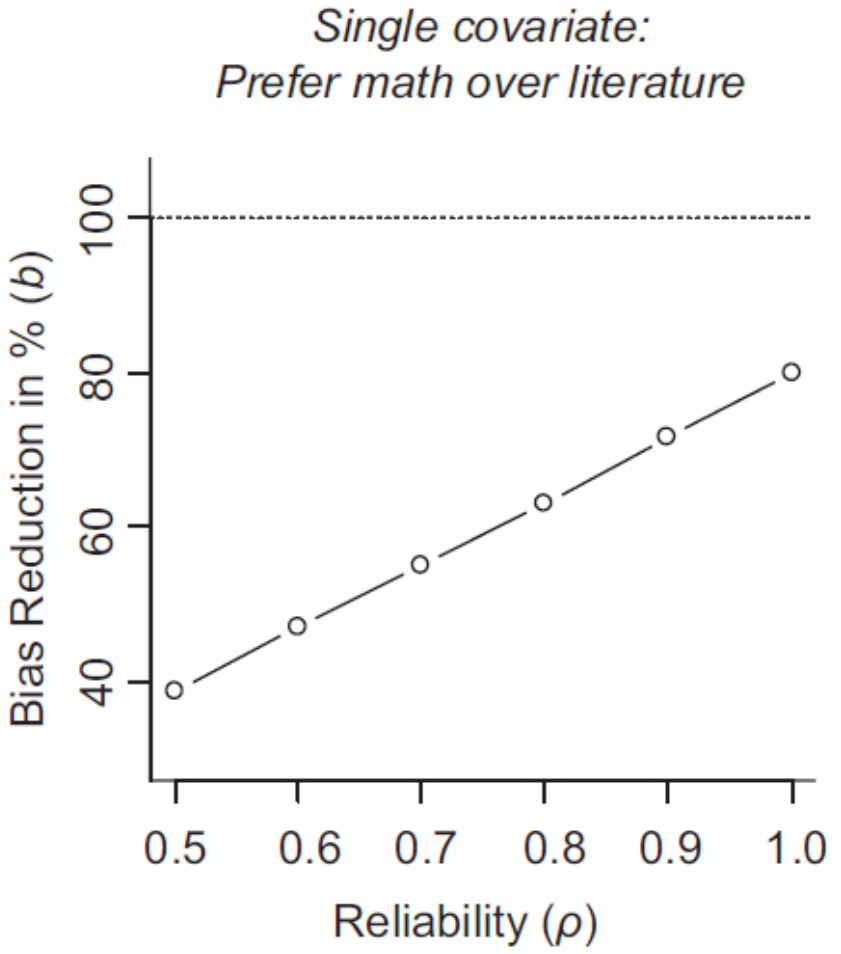
What About Measurement Error?[1]

- Some parts of the literature, especially propensity score analysis, have mostly ignored the role of measurement error.
- We used the data from the JASA study to simulate the effects of adding various amounts of measurement error to covariates.
  - 2000 replications
  - Re-estimating PS's each time
  - For three sets of
    - All covariates
    - Effective covariates
    - Ineffective covariates

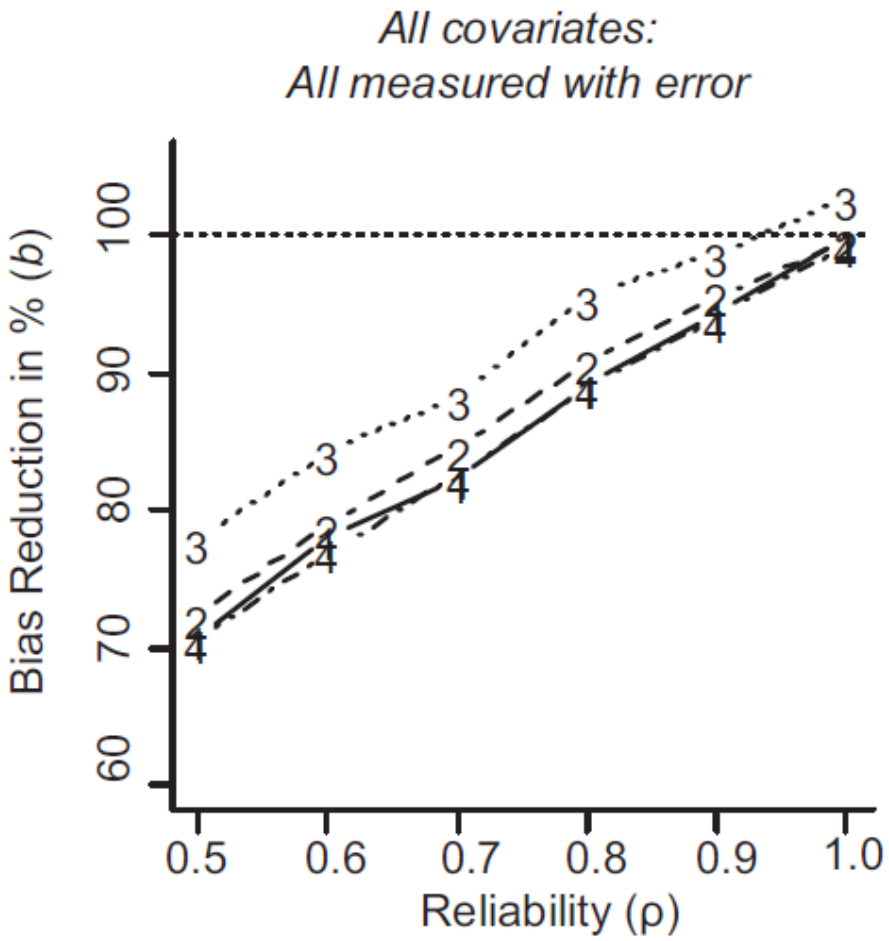
[1]This section based on Steiner, Cook & Shadish (2011).

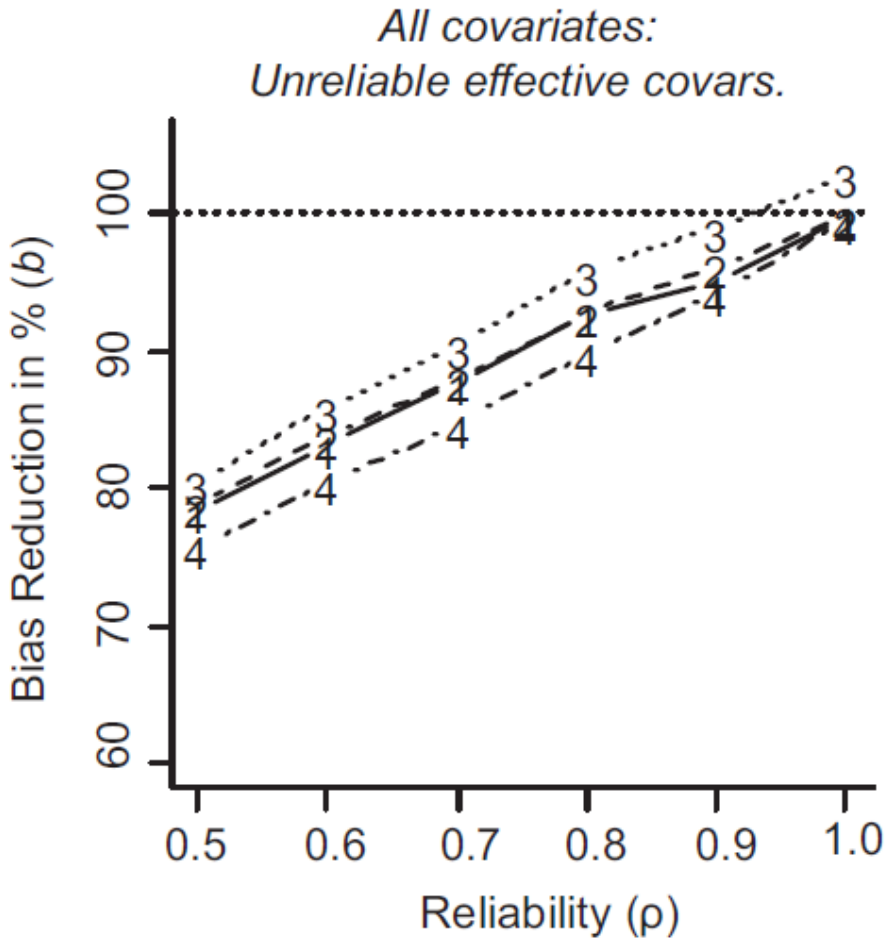
Effect of measurement error on bias reduction in the mathematics treatment effect (OLS regression)





Effect of measurement error on bias reduction in the mathematics treatment effect by analytic method





Comments for Reliability

- Results were the same for both vocab and math outcomes
- Results were the same for any subset of effective covariates
- However, for ineffective covariates, there is no bias reduction, so adding measurement error makes no difference
- Groups of covariates lose less bias reduction due unreliability than individual covariates
  - Aggregates are inherently more reliable than individual items.
  - PS's are an aggregate

Summary about Covariates

- Need excellent measurement of predictors of selection and outcome to reduce bias.
  - Cover *different construct domains* that are related to both treatment selection and outcome—administrative data or demographics alone will rarely suffice
  - Measure *several operationalizations of* each construct domain
  - *Reliably* measure variables
- No substitute for interviewing participants and others about selection.
  - Shadish et al had @25 covariates predicting treatment selection, gathered from systematic study of the selection process
  - How many QEs (and PSA studies) devote this much attention to careful measurement of selection?

Good Control Group

- The four arm study used a good control group:
  - Highly similar to treatment group in many respects: same university, same class, roughly same age, etc.
  - We sometimes call it a focal local control[1]
    - Local: from the same location
    - Focal: sharing the same substantive characteristics
- Uses good design to reduce the amount of bias that analysis has to remove.
- Cook Shadish Wong (2008) reviewed a number of other studies supporting this idea.

[1] Shadish and Cook, 2009.

Three Good Practices Converged in the Four Arm Study

- A good control group
- A rich set of covariates
- Including a pretest (proxy pretest but…)

Replication

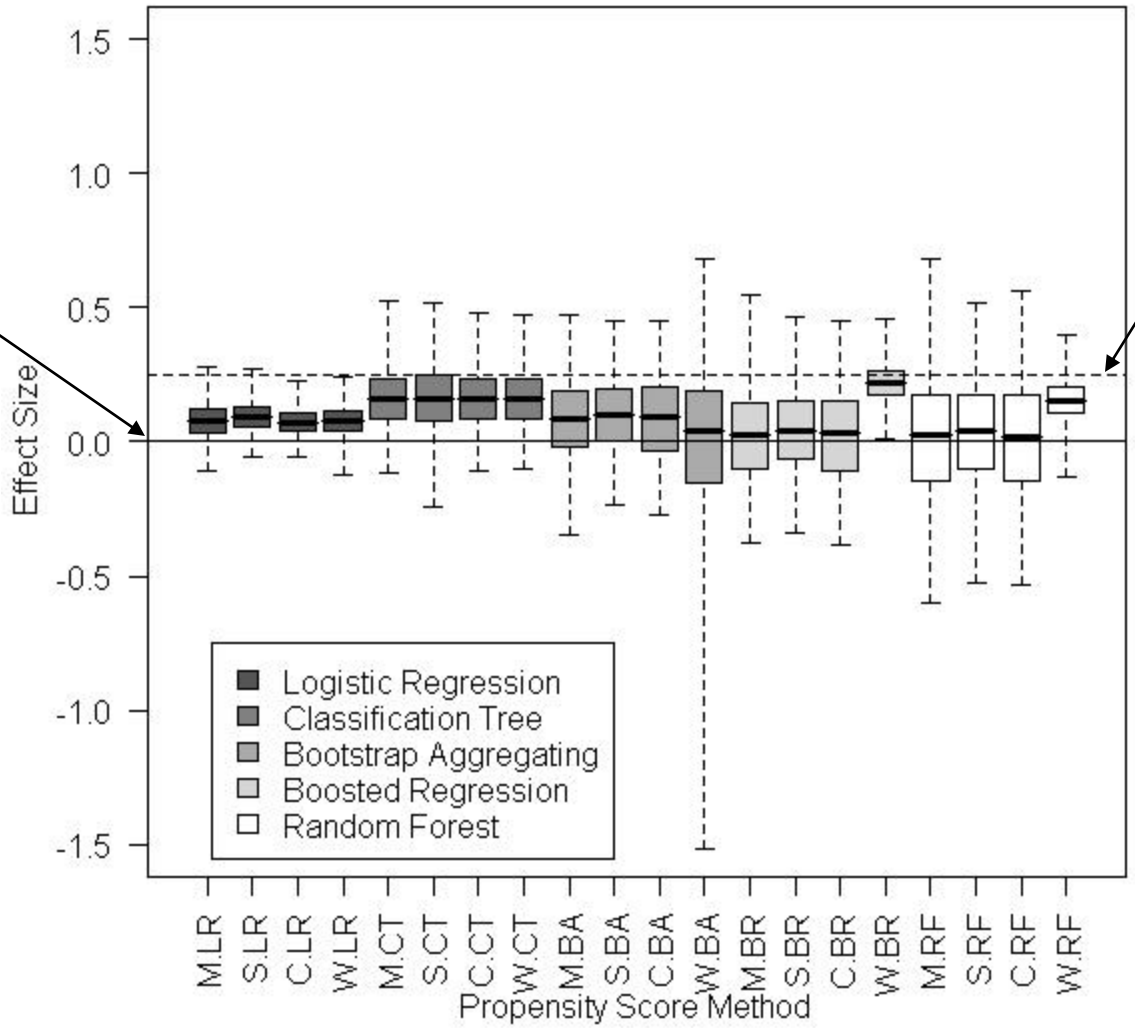
- All of this replicated by Steffi Pohl[1] in Germany
- Vocabulary condition revised
- Initial bias higher
- Results in general were the same, and the same for
  - No difference in analytic methods
  - Importance of covariate and reliability

[1] Cook Steiner Pohl (2009); Cook Pohl Steiner (2011); Pohl, Steiner, Eisermann, Soellner & Cook (2009)

PSA and Sample Size

- Luellen Dissertation (2006)
- Motivated by trying to understand which of the many options for PS analysis is most important to attend to in practice:
  - Different PS creation methods
  - Different analysis methods
  - What does it mean to say PS is large sample method?
- Computer simulation
  - Five sample sizes (200, 500, 1000, 1500, 2000)
  - Five PS construction methods (LR, Classification Trees, Bagging, Boosting, Random Forest).
  - Four analysis methods (matching, strat, cov, weighting)

N = 500: Typical of the Results

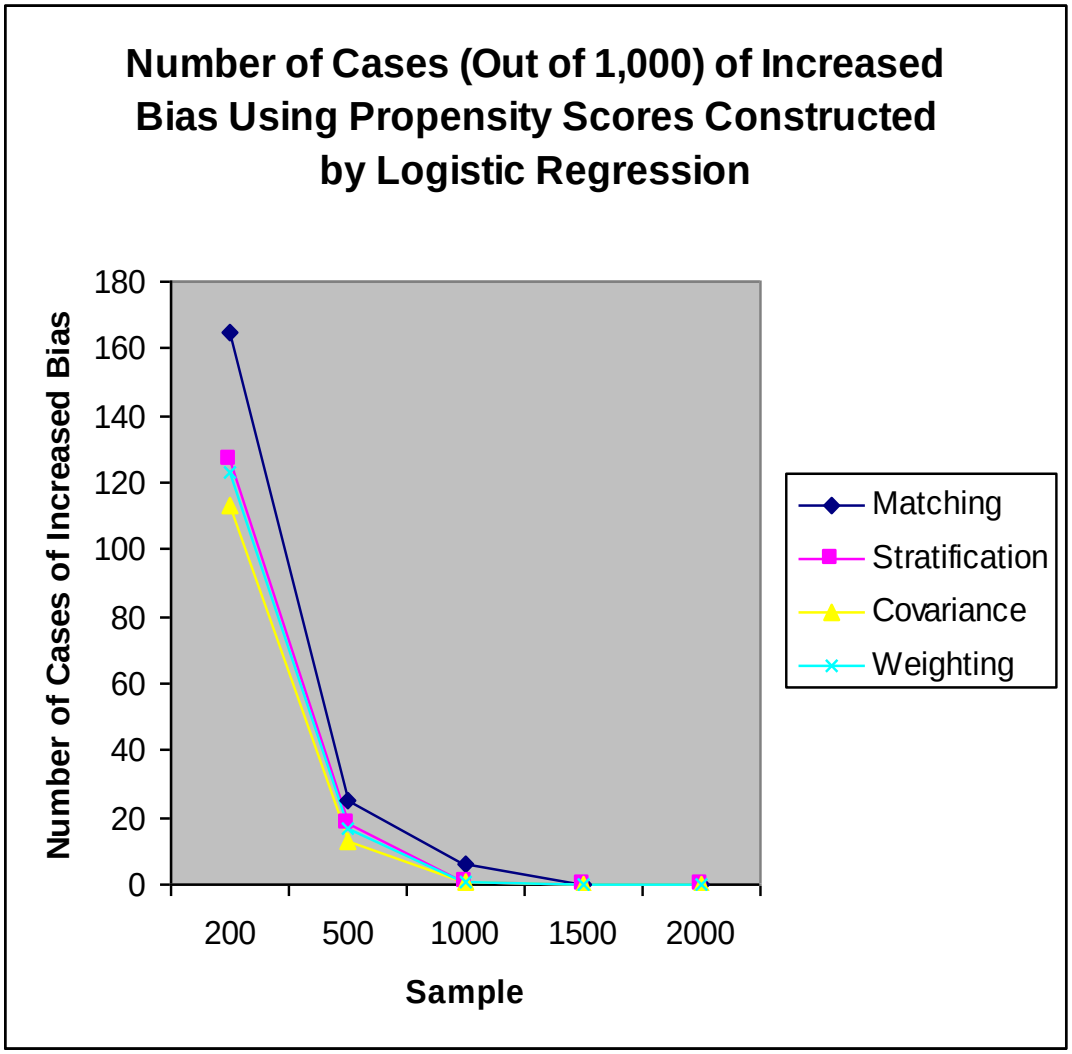


True Effect was zero

Bias of .25 programmed.

1. Logistic regression typically performed best
2. No analysis methods tended to perform best
3. You could get further away from the right answer. So:

Needed Sample Size to Ensure No Increase in Bias



1. N = 200 (100 per group) increased bias @15% of the time.
2. N = 500 much better, but still 2% increased bias
3. N ≥ 1000 best

Some Advantages of PSA

- We highlighted problems with PSA, but
- It has many advantages too:
  - Summarizes many covariates into a single score
    - Makes matching-stratifying easier
  - Separates “design” from analysis
    - Looking at balance and choosing final model before analyzing the outcome
  - Less reliant on functional form assumptions
  - Avoid relying on non-overlapping cases
  - Lends itself to doubly robust analysis--can be combined with additional covariate adjustments in the outcome model
- So we are fans of PSA despite is problems

Summary Lesson #1

- Empirically study the selection process into conditions to identify the key constructs in selection.

Summary Lesson #2

- Make multiple measures of each construct in designing prospective studies,
- or in the case of archival data, assess and report the extent to which the archive has measures of each construct.

Summary Lesson #3

- Use highly reliable measures, and report reliability levels.

Summary Lesson #4

- Use at least 500 people in the experiment, and preferably 1,000–1,500.

Summary Lesson #5

- Use a good focal local control group,
  - one that is from the same location
  - and shares as much in common with the treatment group as possible,
- in order to reduce the amount of bias that propensity scores have to address.

Question

- How many PSA studies do all five of these things?
- Unknown: Can doing well on one make up for the other?
  - Will rich covariates make up for small N (four arm study was small N)?
  - Will having more covariates compensate for unreliability in individual covariates?
- And, of course, these lessons probably apply just as well to any other approach, not just PSA

Bigger Question

- Can propensity score analysis substitute for randomized experiments?
  - In theory, and possibly in the best practice, the answer might be positive
  - in usual practice, the answer is probably not.
- The current popularity of propensity score analysis may reflect irrational exuberance.
- However, usual practice can change and become good or even best practice,
- and that exuberance can become more rational.