“Work Horse” Design: Description

- _O _X_O_
  O O
- Two components, pretest and control group not formed at random and hence non-equivalent on expectation
- These two components make up one slope that one wants to treat as though it were a perfect counterfactual
- But it isn’t known to be, and is not likely to be

Workhorse Design: Overview

- Chief Internal Validity Threats
- Bad Matching to Equate Groups
- Maximizing Overlap in Observables
- Case Matching-Propensity Scores and OLS
- Other Forms of Statistical Adjustment

Chief I. V. Threats with Design

With or without differential attrition, we struggle to rule out:

- Selection–Maturation
- Selection-History (Local History)
- Selection–Instrumentation
- Selection-Statistical Regression
- So why not match to eliminate all these different faces of selection? If groups can be made equivalent to start with, does not the problem go away, as it does with random assignment?

Bad Matching for Comparability

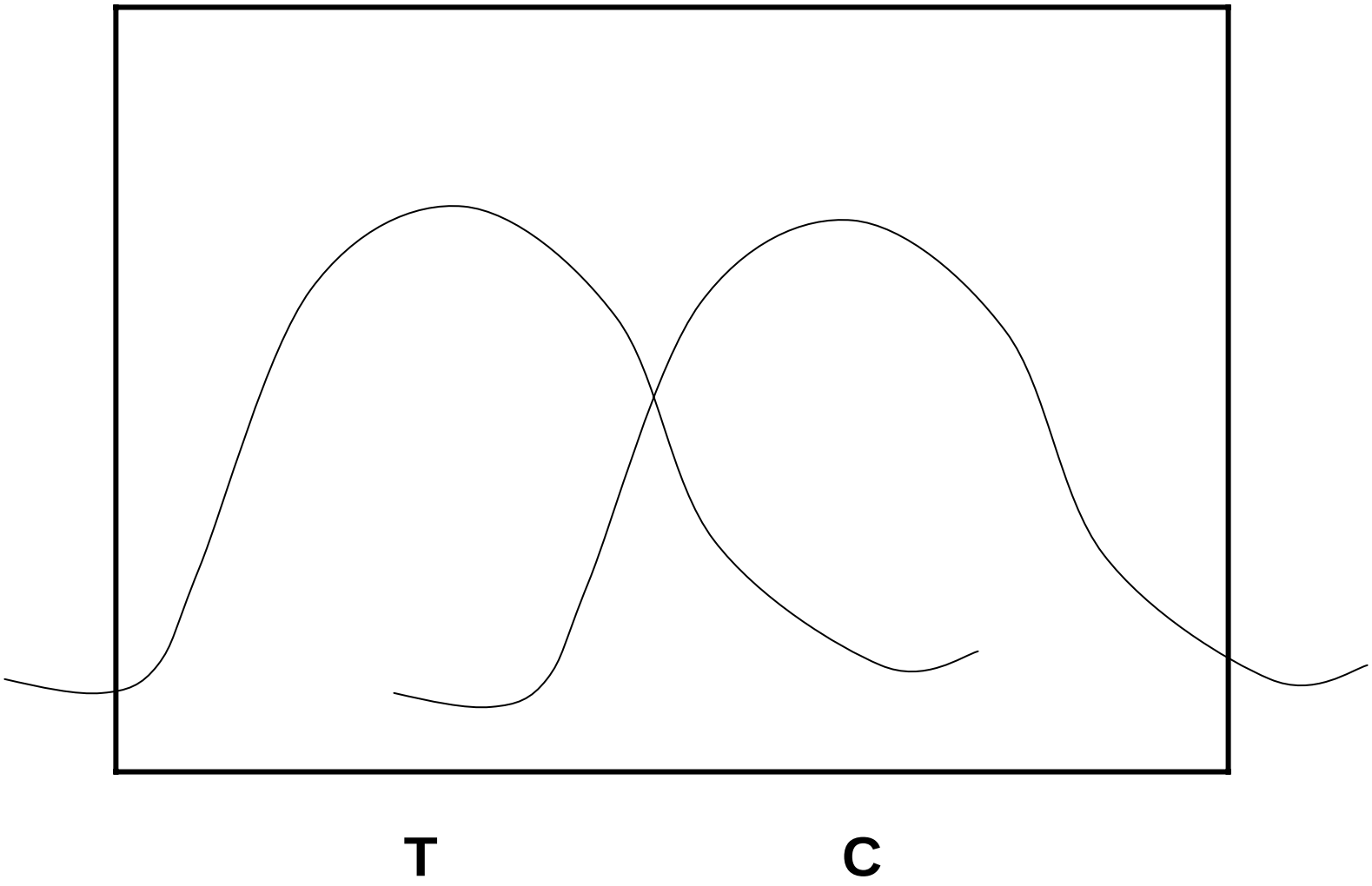
Simple Regression illustrated with one group

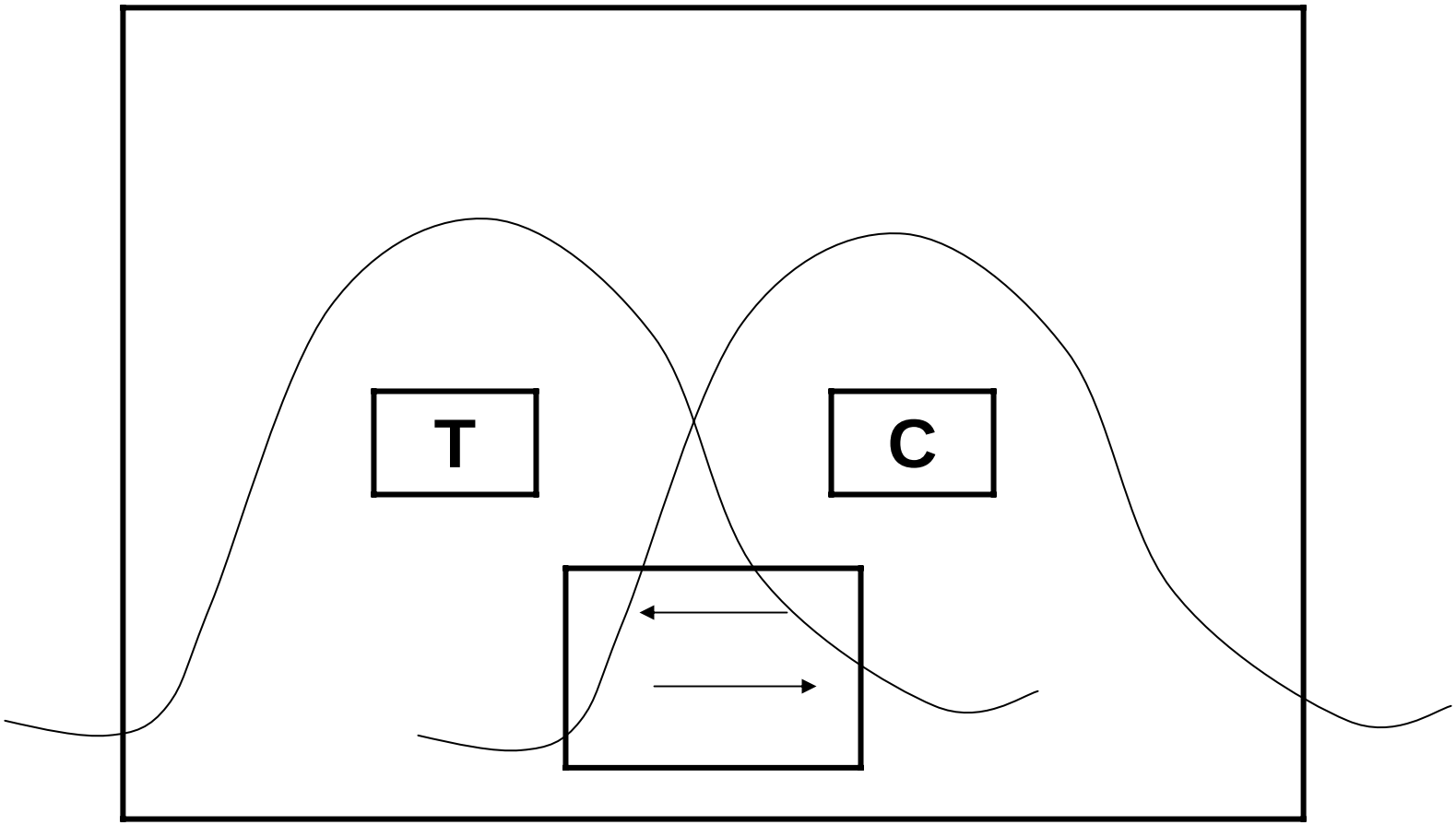
Frequency of such regression in our society

It is a function of all imperfect correlations

Size of Regression gross f(unreliability/population difference)

Simple one-on-one case matching from overlap visually described


T
C



T: Decreases C: Increases	Both forces makes T look ineffective

The Net Effect is…

- If either treatment decreases or controls increase due to regression, then bias results
- If both change in opposite directions, then the bias is exacerbated
- Matching individual units from extremes is not recommended

If you need to, then…

- The Cicirelli Head Start Evaluation had this problem, concluding Head Start was harmful
- LISREL reanalyses by Magidson using multiple measures at pretet led to different conclusion, likely by reducing unreliability.
- In theory, propensity scores might work if, as we see, a sufficiently rich covariate structure were available to that end
- Reliability is higher using aggregate scores like schools--but beware here as with effective schools literature.

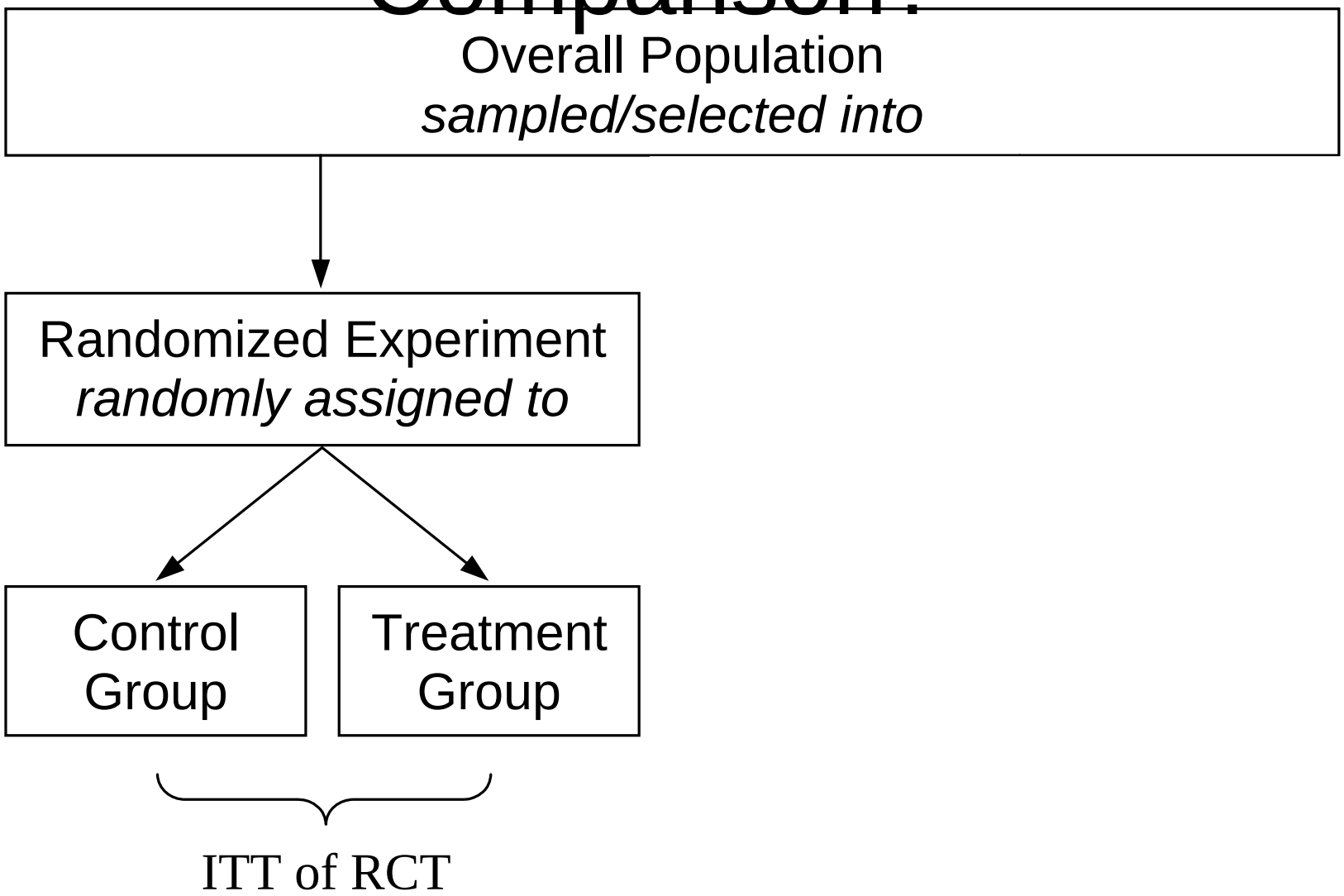
Better is to get out of the Pickle

- Don’t match from extremes! Use intact groups instead, selecting for comparability on pretest
- 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

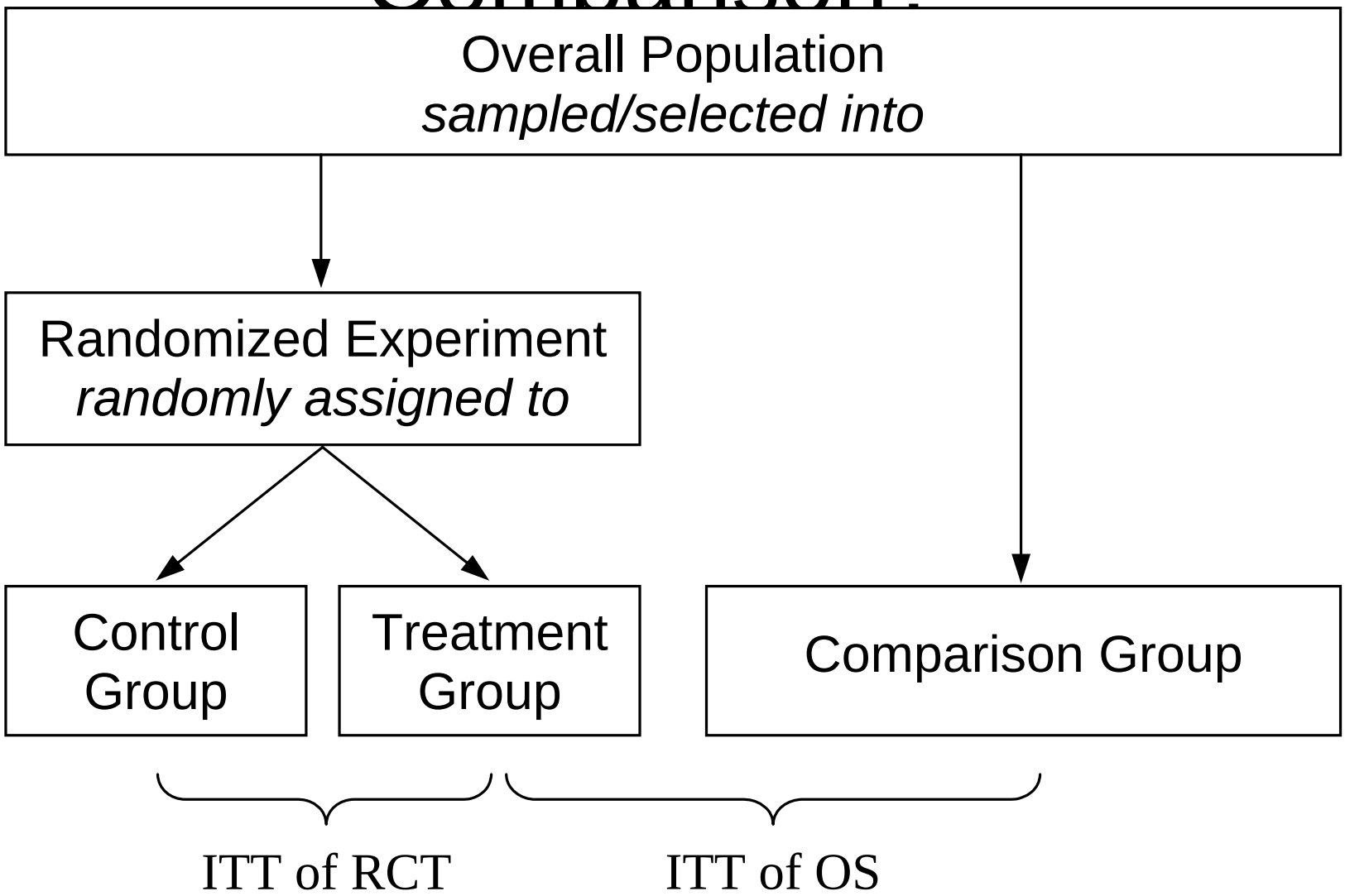
Intact Group Matching

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

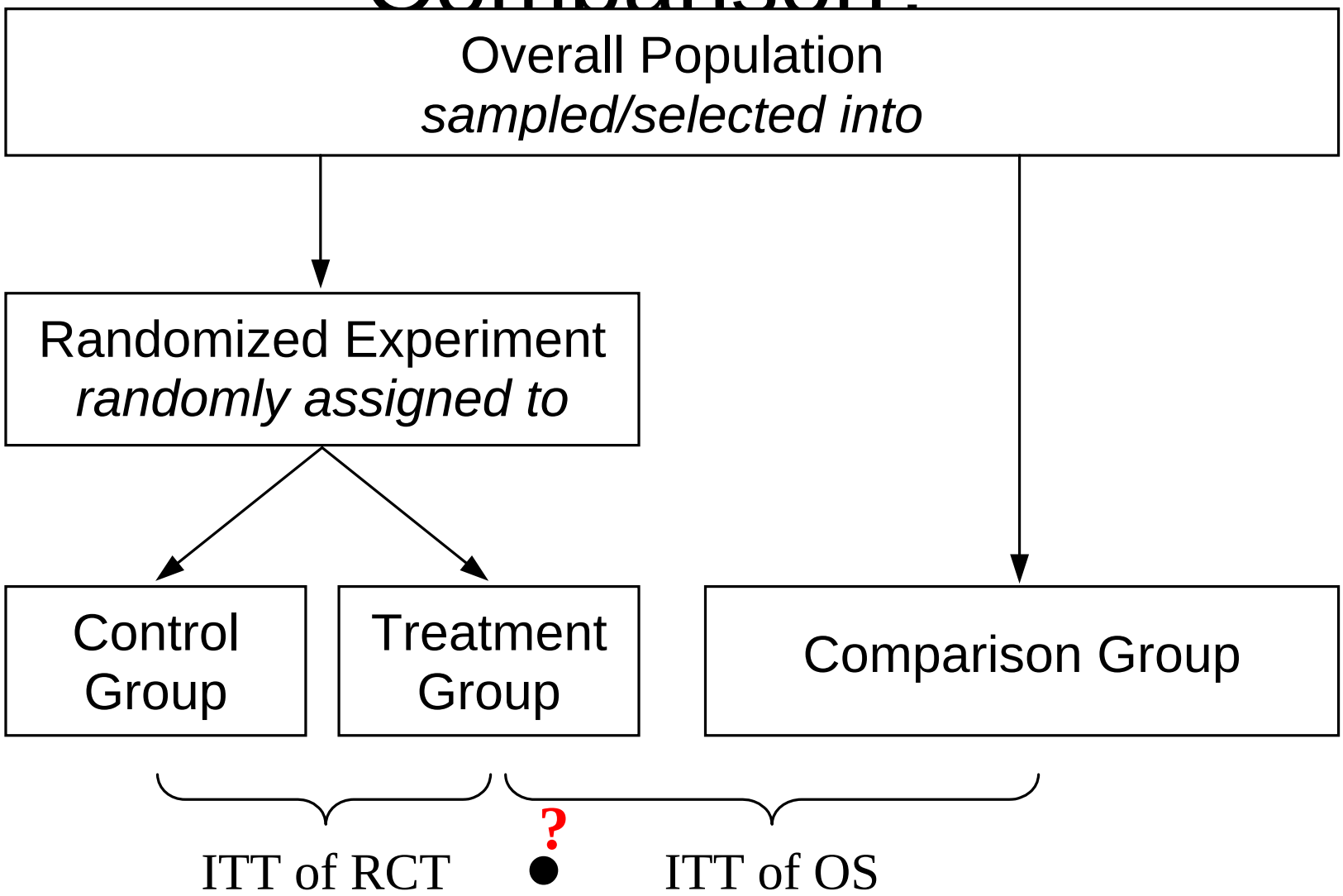
What is a Within-Study Comparison?



What is a Within-Study Comparison?



What is a Within-Study Comparison?



Consider 3 examples

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

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
- Historical change in meeting of criteria

Bloom, Michaelopoulos et al

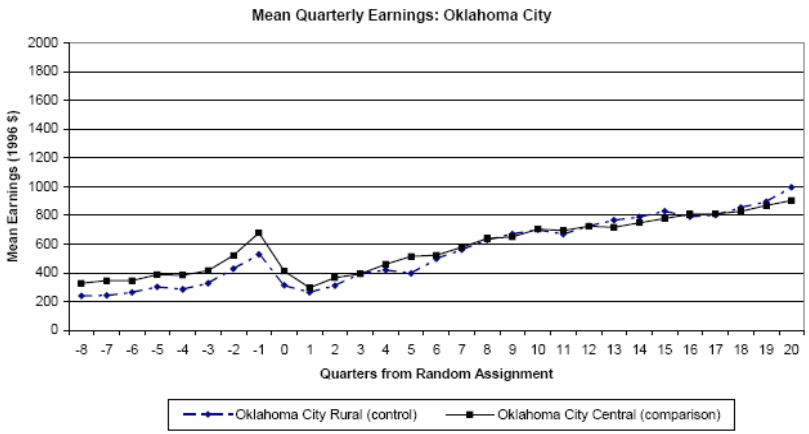
Logic of the design is to compare ES from a randomly created control group with ES from a non-randomly formed comparison that shares the same treatment group

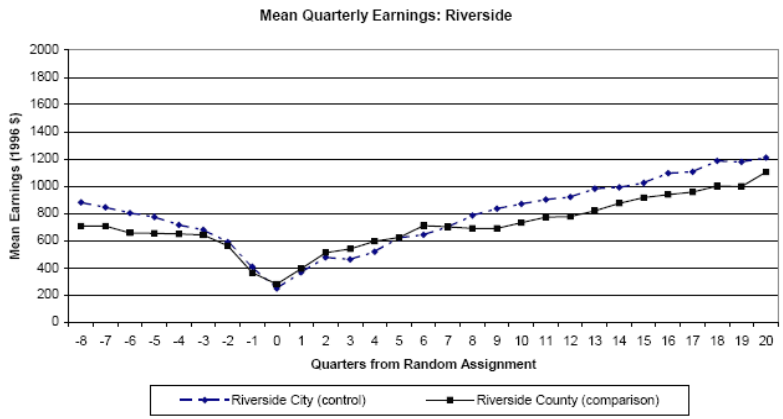
- Treatment group is a constant and can be ignored, comparing only the two types of control
- Issue is: Will the randomly and non-randomly formed control groups differ over 8 pretest observation points after standard statistical adjustments for any differences in mean or slope

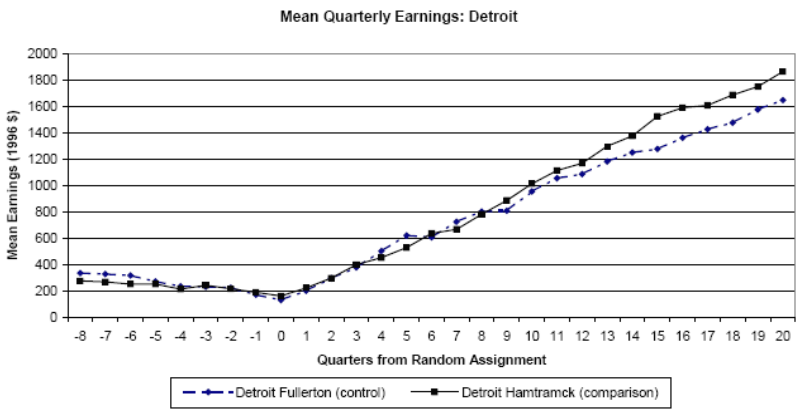
The Context

- RCT is on job training at 11 sites
- Bloom et al restrict the ITS to 5 within-state comparisons, 4 of them within-city
- Non-random comparison cases chosen from job training centers in same city
- Measured in the same ways as treated at same times

Results: 3 within-city Samples







What you see in Graphs

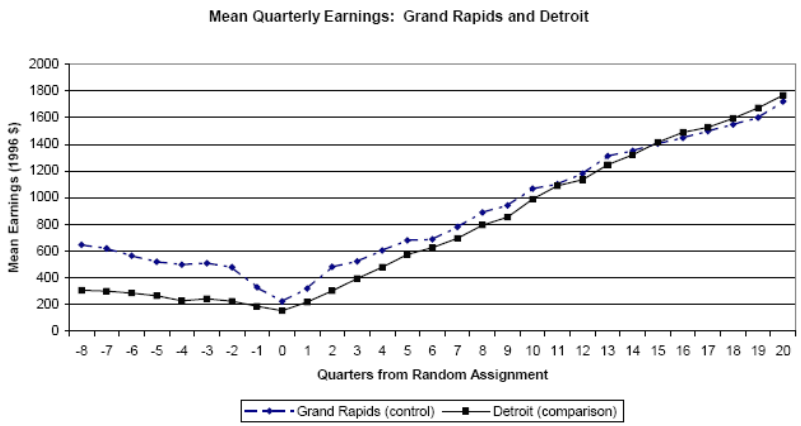
- Hardly differ at all --- one advantage of TS is that we can see group differences
- Statistical tests confirm no differences in intercept or slope
- In these cases, equivalence of randomly and non-randomly formed comparison groups is achieved thru sampling design alone
- Thus, no need for statistical tests to render them “equivalent”

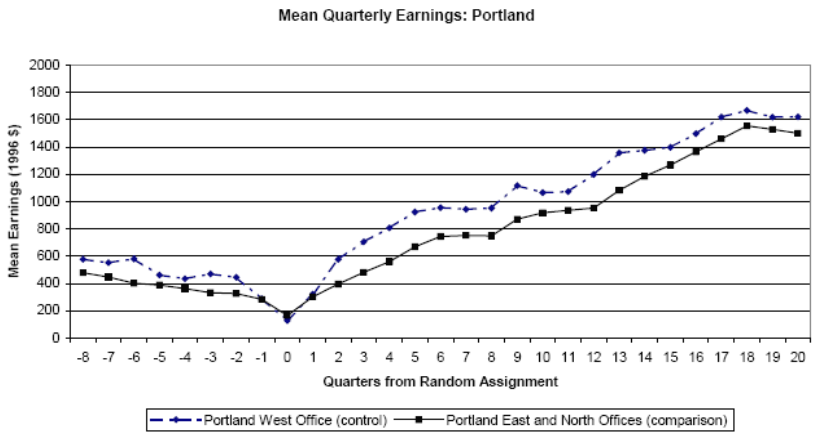
Two other Sites

Portland--sample size smallest and least stable

Detroit vs Grand Rapids--a within-state but not within-city comparison. Hence, this is not a very local comparison

Bloom et al. Results (2)





Here you see

- TS are not equivalent overall
- TS especially not equivalent around the crucial intervention point
- Thus use of a random or non-random control group would produce different results
- 10 types of statistical analyses were used to make the series equivalent:

Results of these Analyses

- 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
- Only the random growth model took advantage of the TS nature of the data
- Why did it fail too?

Selecting Intact Groups locally matched on pretest outcomes

- Without intending it, Bloom et al's choice of 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
- Remember: There was bias in across-state comparisons, and it could not be adjusted away statistically with the data and models used

Selecting Intact Groups with Maximal Overlap: 2nd 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

3rd Example: Diaz & Handa (2006)

- Progresa: Matched Villages with and without the program
- One sample of villages had to meet the village eligibility standards--in bottom quintile on test of material resources, but for a variety of reasons not in experiment
- The eligible no-treatment comparison families in these villages were not different on outcomes from the randomly created comparison group

But there were different on a few family characteristics

- Nonetheless, the results of the matched village analyses were similar whether covariates were added to control for these differences or not

Implications of all Three Studies

- Aiken et al and Bloom et al. created non-equivalent control groups that were not different on observables from the treatment group.
- These observables included a pretest on the same scale as the outcome
- Diaz and Handa created much overlap but some differences that did not affect outcome
- But what about unobservables? We never know. But if there are real differences, or real unadjusted differences, then we know to worry

What is a Local, Focal, Non-Equivalent Intact Control Group 1

- Local because…
- Focal because…
- Non-Equivalent because…
- Intact because…

What is a Local, Focal, Non-Equivalent Intact Control Group 2

- Identical Twins
- Fraternal Twins
- Siblings
- Successive Grade Cohorts within the same School
- Same Cohort within different Schools in same District
- Same Cohort within different Schools in different districts in same state
- Same Cohort within different schools in different states, etc.

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.

Analysis of Workhorse Design Data when group Differences

- Modeling the outcome, like covariance analysis
- Modeling selection, like Propensity Scores
- Empirical Validation literature

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

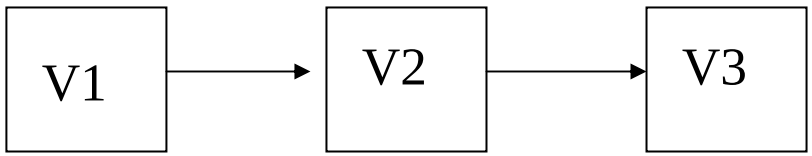
- It is generally known that you can obtain an unbiased estimate of a treatment effect if you can
  - Know the selection model or the outcome model completely
  - Measure it perfectly
- One of the reasons that regression discontinuity works is that it can meet these two criteria.
- Other designs cannot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

Specification Error

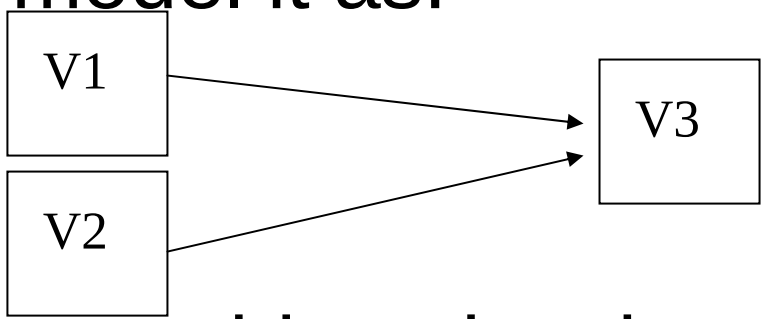
- Two forms
  - In any form of selection modeling (e.g., propensity score analysis), this is the omission of a variable correlated with outcome that is important to how people selected into conditions.
  - In any form of a complete model of the outcome (e.g., LISREL), this is the omission of a variable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

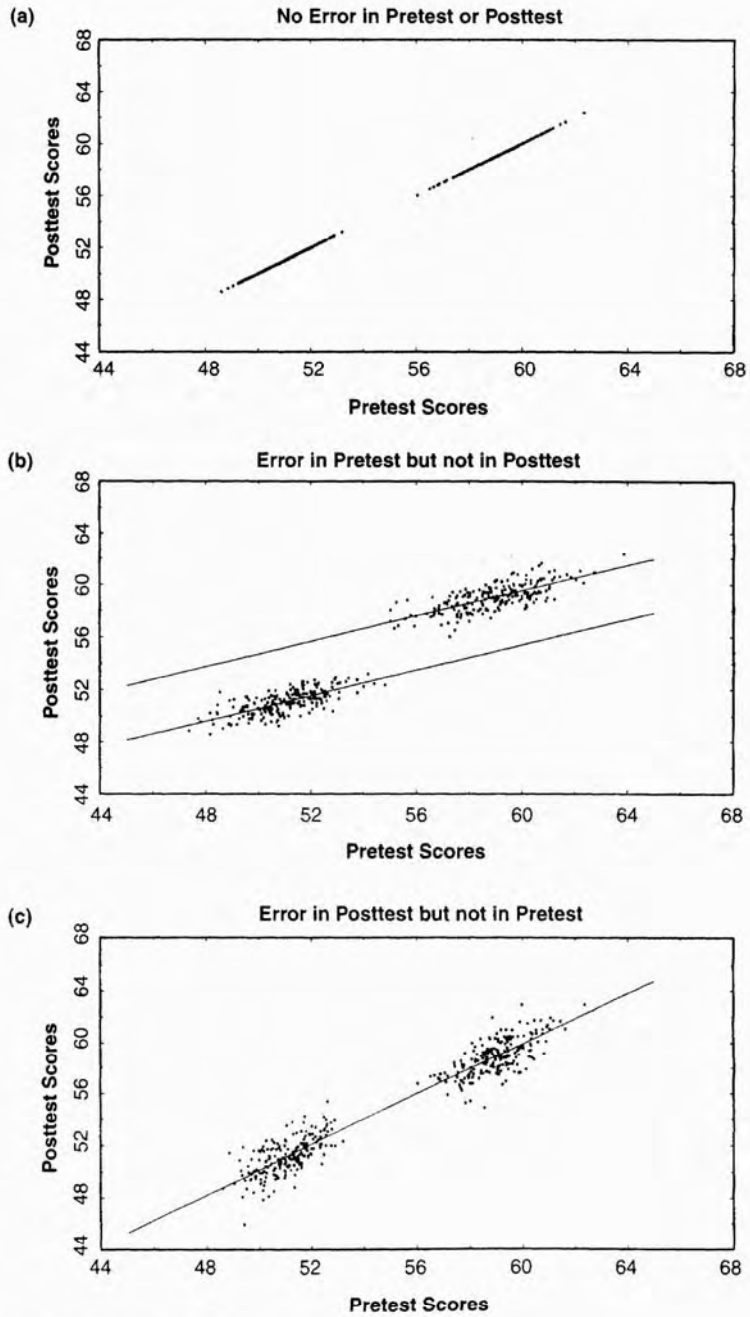


FIGURE 7.8 The effects of errors in pretests and posttests.

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.

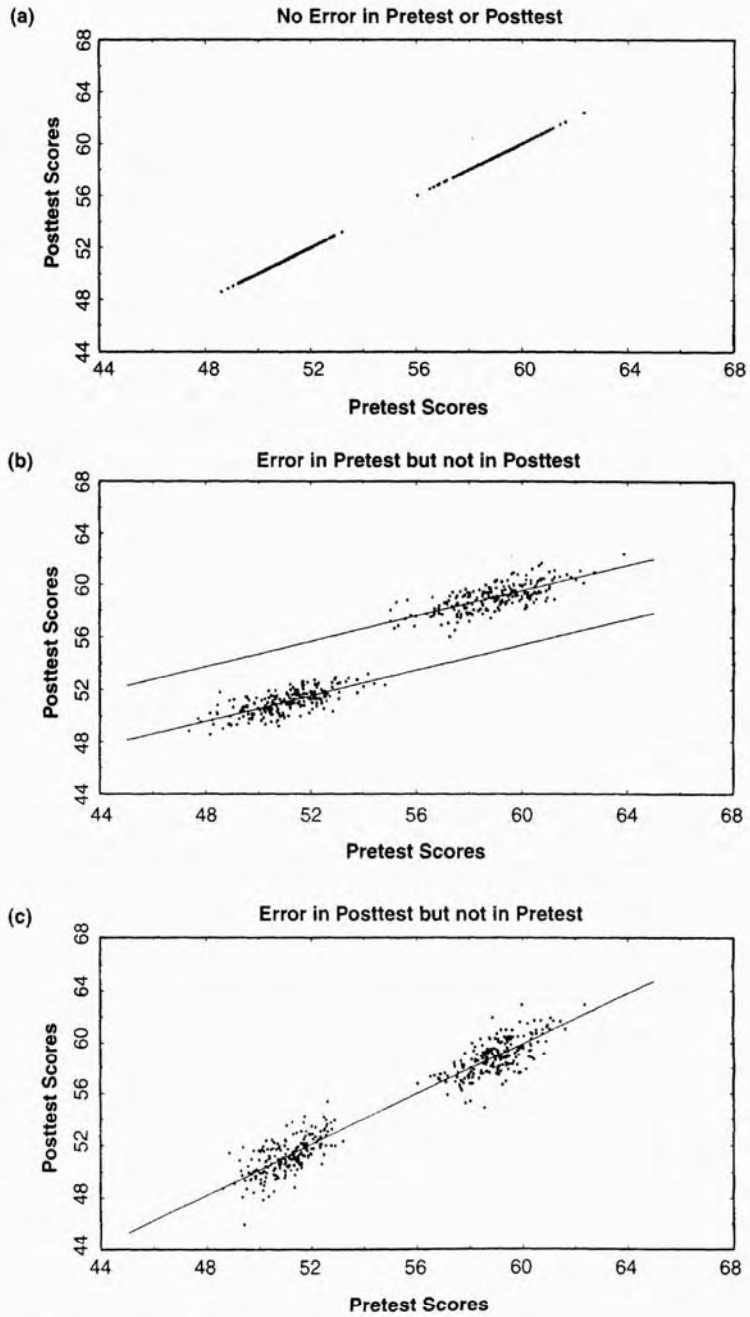


FIGURE 7.8 The effects of errors in pretests and posttests.

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.

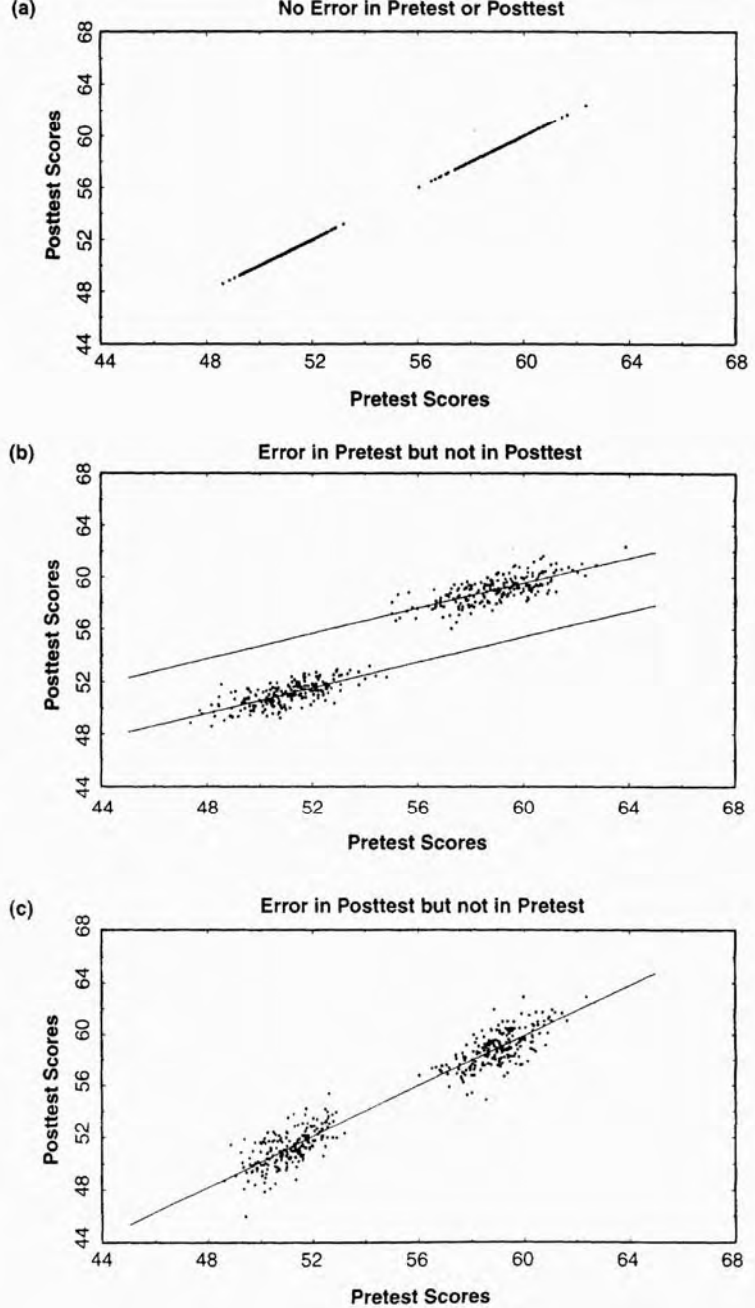
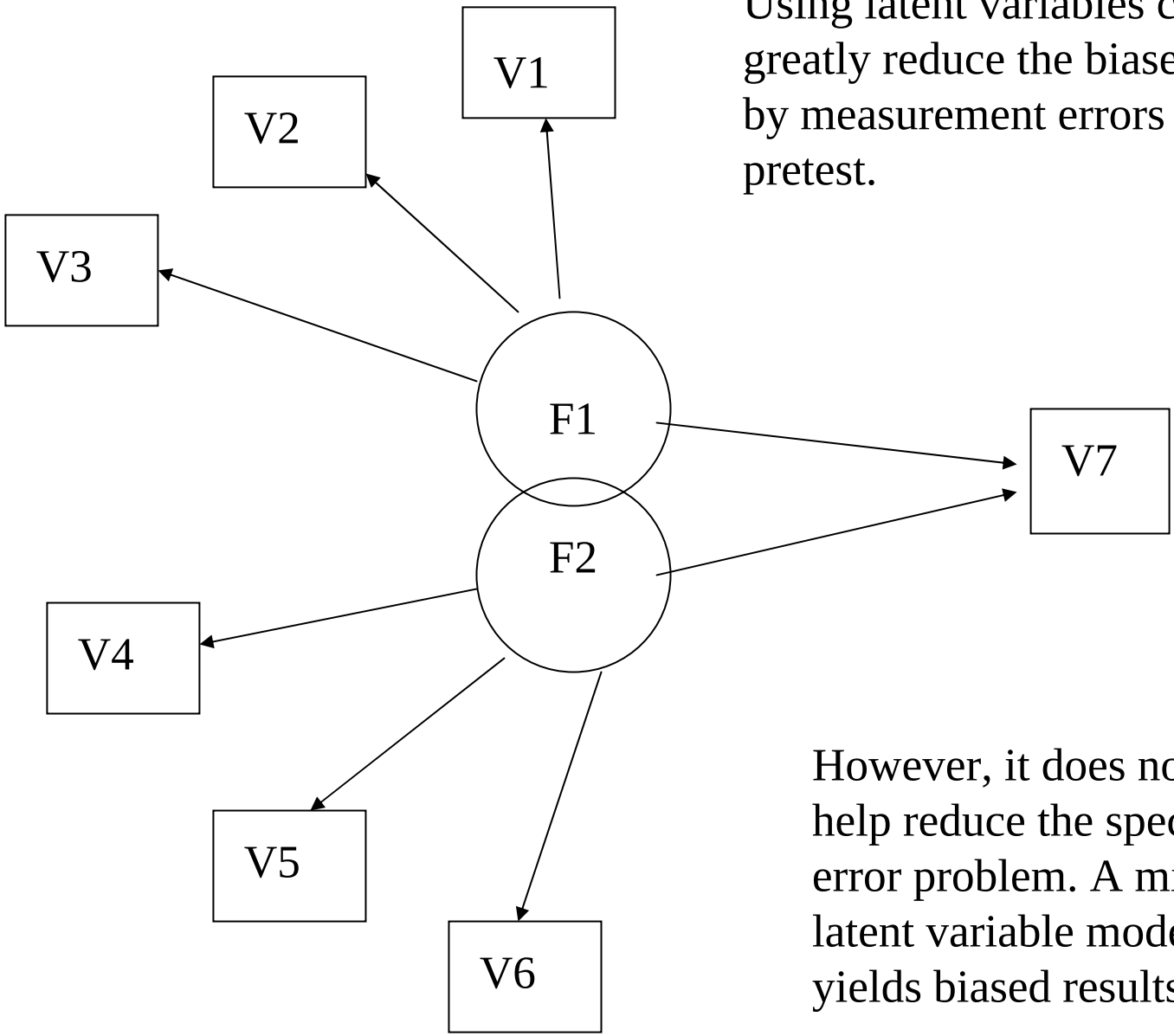


FIGURE 7.8 The effects of errors in pretests and posttests.

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:



Using latent variables can help greatly reduce the biased caused by measurement errors in the pretest.

However, it does nothing to help reduce the specification error problem. A mis-specified latent variable model still yields biased results.

Propensity Scores and Quasi-Experiments:

What we will do

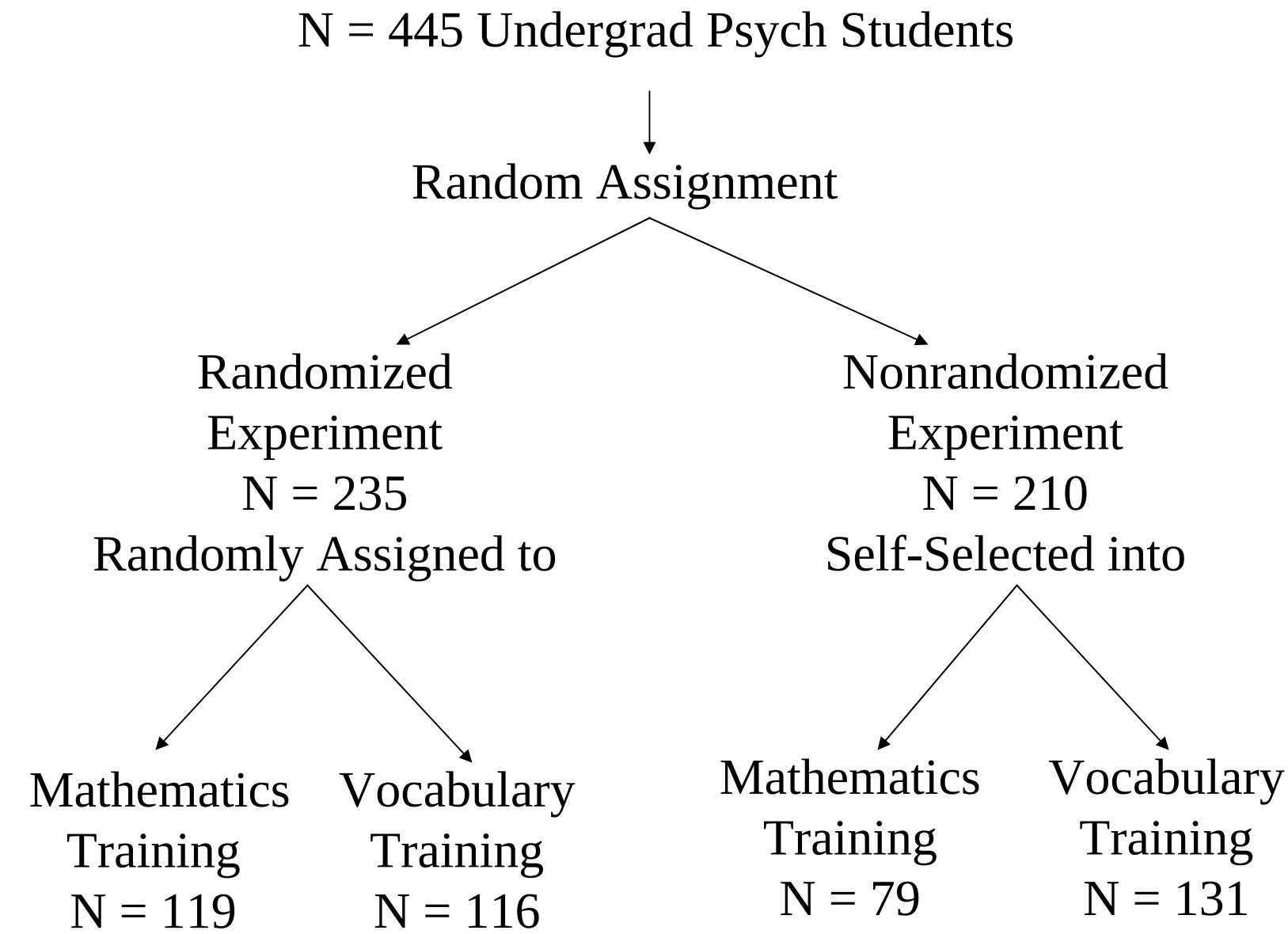
- 1. Begin with Design of Test of propensity score
- 2. Describe Propensity scores
- 3. Results of Test

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?

Randomly Assign People to Random or Nonrandom Assigment

- One way to test this is to randomly assign participants to being in a randomized or nonrandomized experiment in which they are otherwise treated identically.
- Then one can adjust the quasi-experimental results to see how well they approximate the randomized results.
- 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
- 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.

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 to make those adjustments

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 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
- The ratio of the variances of the covariates after adjusting for the propensity score 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.

Computer Programs: STATA (www.stata.com)

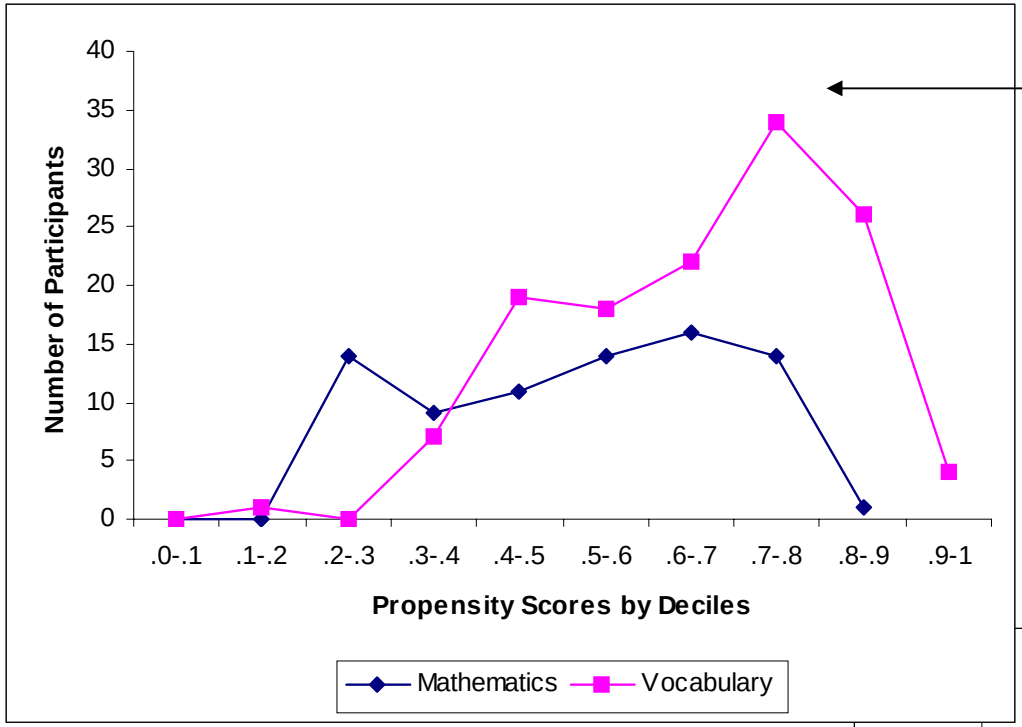
- pscore
- psmatch2 (http://www.stata.com/meeting/7uk/sianesi.pdf)
- match (http://emlab.berkeley.edu/users/imbens/estimators.shtml)
- sensatt (http://ideas.repec.org/c/boc/bocode/s456747.html)
- attnd: nearest neighbor with randomization to resolve ties
- attnw: nearest neighbor with equal weighting of ties
- attr: radius matching
- atts: stratification matching (or interval matching, using the blocks chosen by the balancing test)

Computer Programs: R

- R (http://cran.r-project.org/)
  - Estimation
    - Classification trees (rpart)
    - Bagging (ipred)
    - Boosted regression (gbm; AdaBoost)
    - Random forest (randomForest)
  - Analysis
    - Matching (MatchIt, Matching)
    - Stratification (twang)
- Some of these programs do both estimation and analysis, and some assess balance

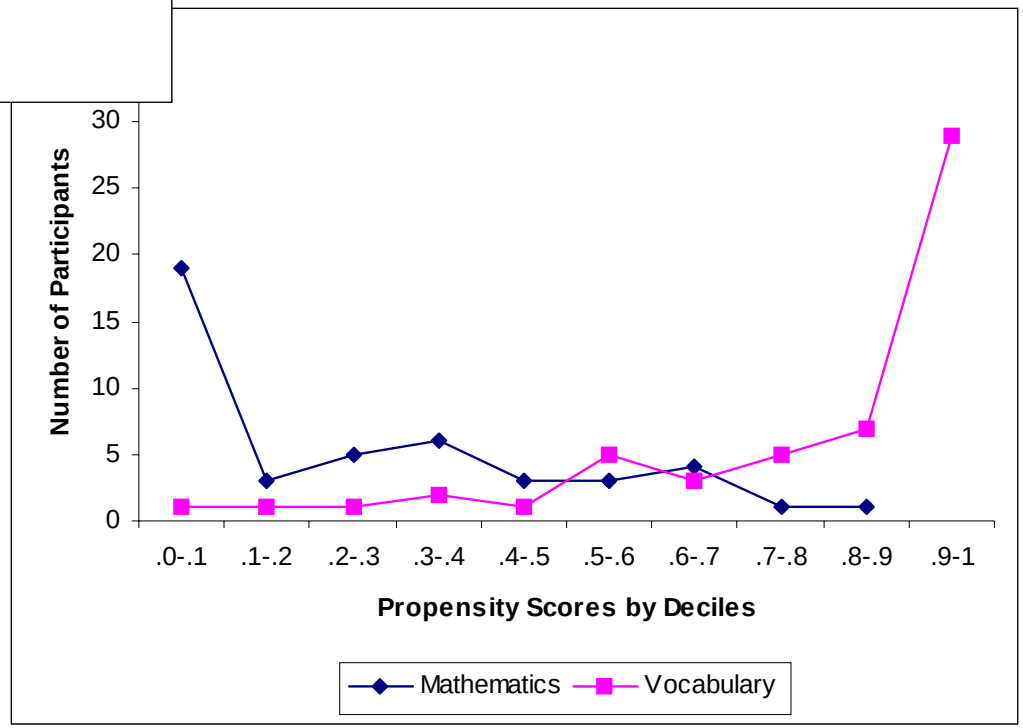
Using Propensity Scores to Test Whether Nonequivalent Groups Should Be Compared

- 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.



In this graph, propensity scores for both groups overlap fairly well.

In this graph, overlap is poor.



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)

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

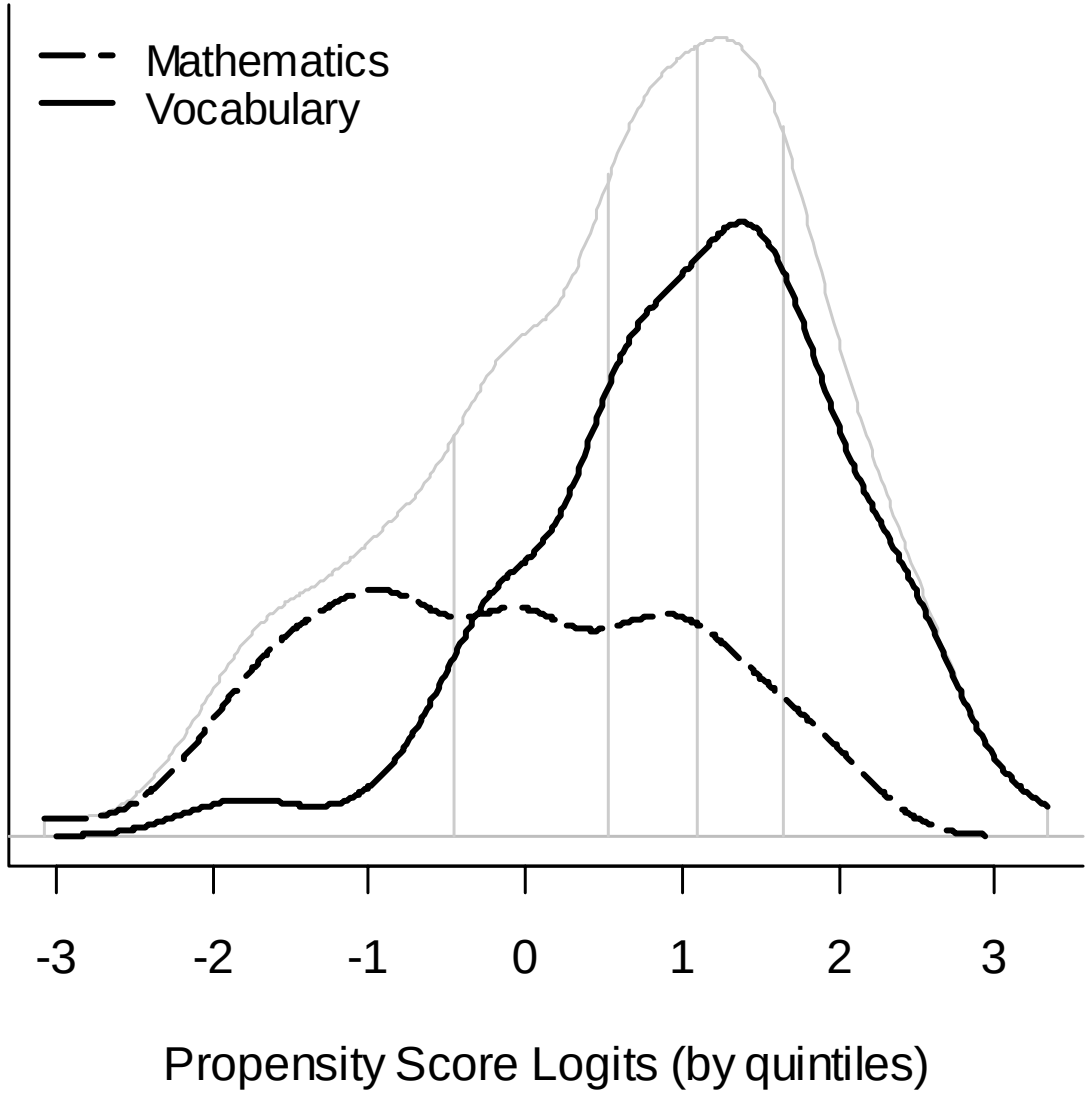
Table 3. Rubin's (2001) Balance Criteria Before and After Propensity Score Stratification

	Propensity Score		Number of Covariates with Variance Ratio				
Analysis	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

Sample SPSS Syntax

```
LOGISTIC REGRESSION VAR=vm
  /METHOD=ENTER vocabpre numbma_r
   likema_r likeli_r prefli_r pextra pconsc
  beck cauc afram age momdeg_r daddeg_r
   majormi liked avoided selfimpr
  /SAVE PRED (ps).
```

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

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 strata

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 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

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:

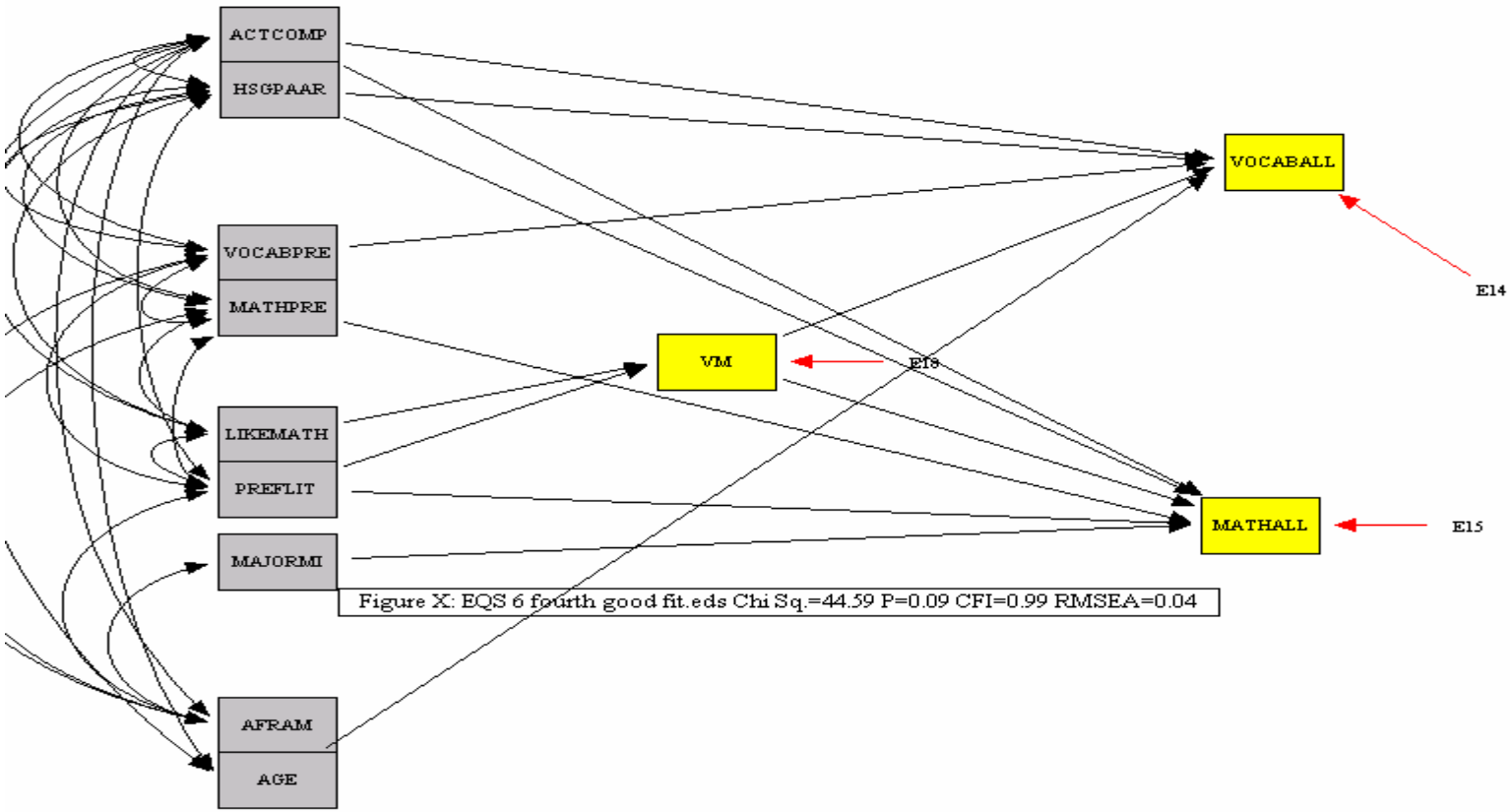


Figure X: EQS 6 fourth good fit.eds Chi Sq.=44.59 P=0.09 CFI=0.99 RMSEA=0.04

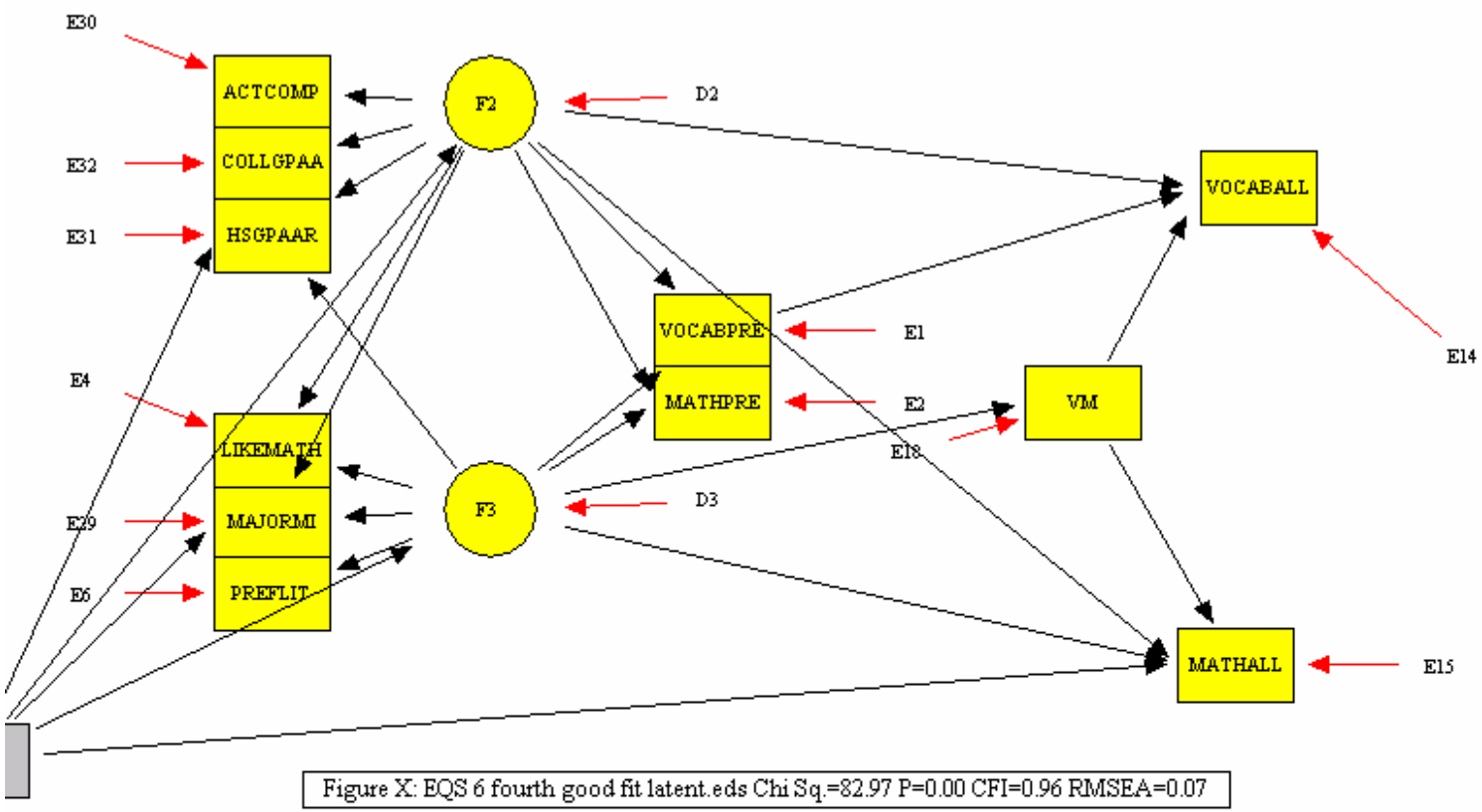


Figure X: EQS 6 fourth good fit latent.eds Chi Sq.=82.97 P=0.00 CFI=0.96 RMSEA=0.07

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.

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.

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, we did not test other methods (selection bias modeling, LISREL). Perhaps they would have worked as well?

Limitations: Replication

- Will it replicate?
  - The method is easy enough to implement at any place with access to a subject pool
  - Several researchers are currently gathering data with this method.
  - Particularly interested in
    - how sample size affects this method.
    - Can we create selection biases that make a bigger difference to the treatment effect?

Limitations of Propensity Score Methodology: Hidden Bias

- In theory, they work if you measure all the predictors of group membership.
  - In practice
    - We can never be sure we have all predictors
    - Missing data
  - Hidden Bias Analysis

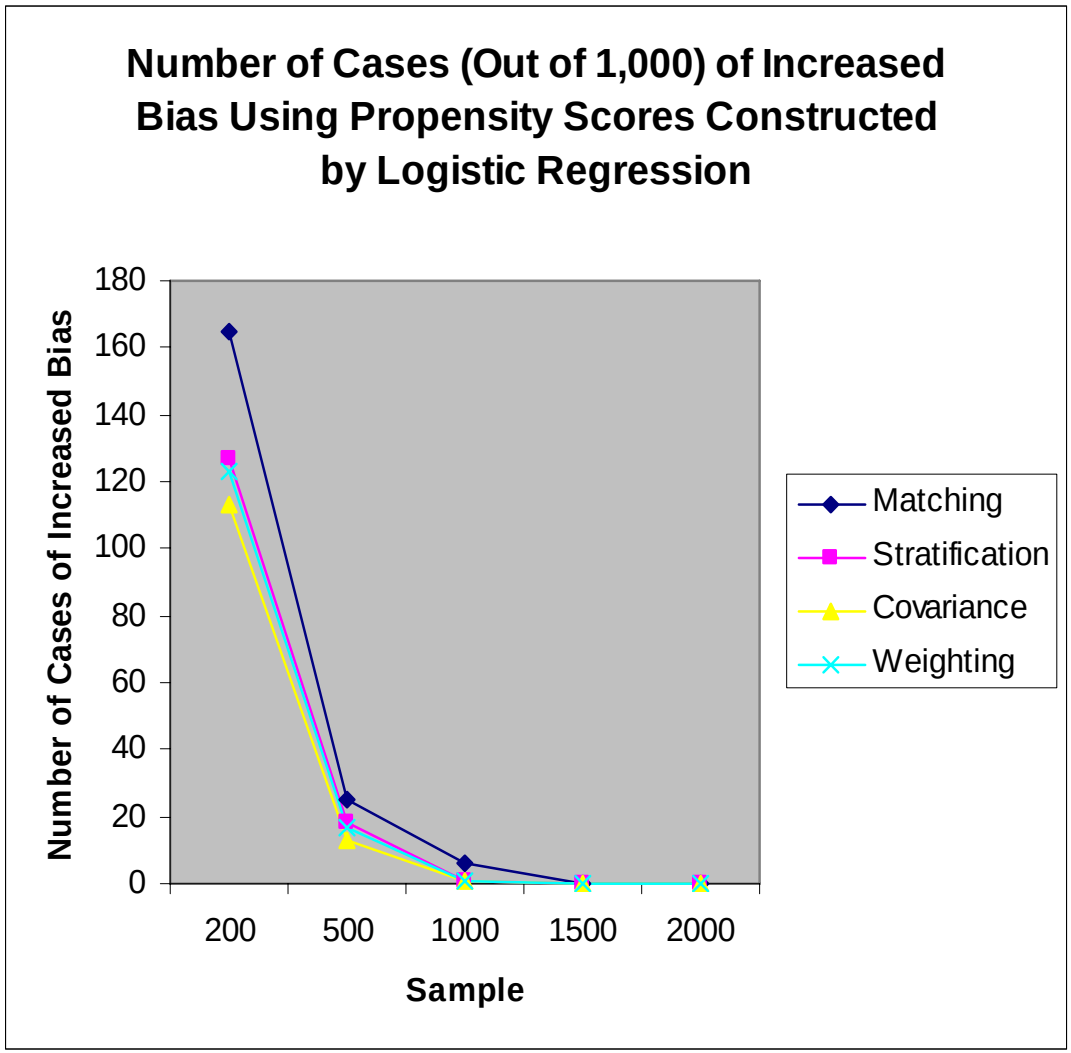
Limitations of Propensity Score Methodology: How to Construct Propensity Scores?

- 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 my experience is that the results are somewhat sensitive to these variations (though we always got closer to the randomized experiment)
  - E.g., accurate prediction vs balancing

Limitations of Propensity Score Methodology: 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
  - My experience is it may not work well with sample size of 150 (our pilot study)
  - Perhaps consider stable matching on outcome measure at pretest when N is low (but also consider using balance tests)
  - Luellen’s dissertation

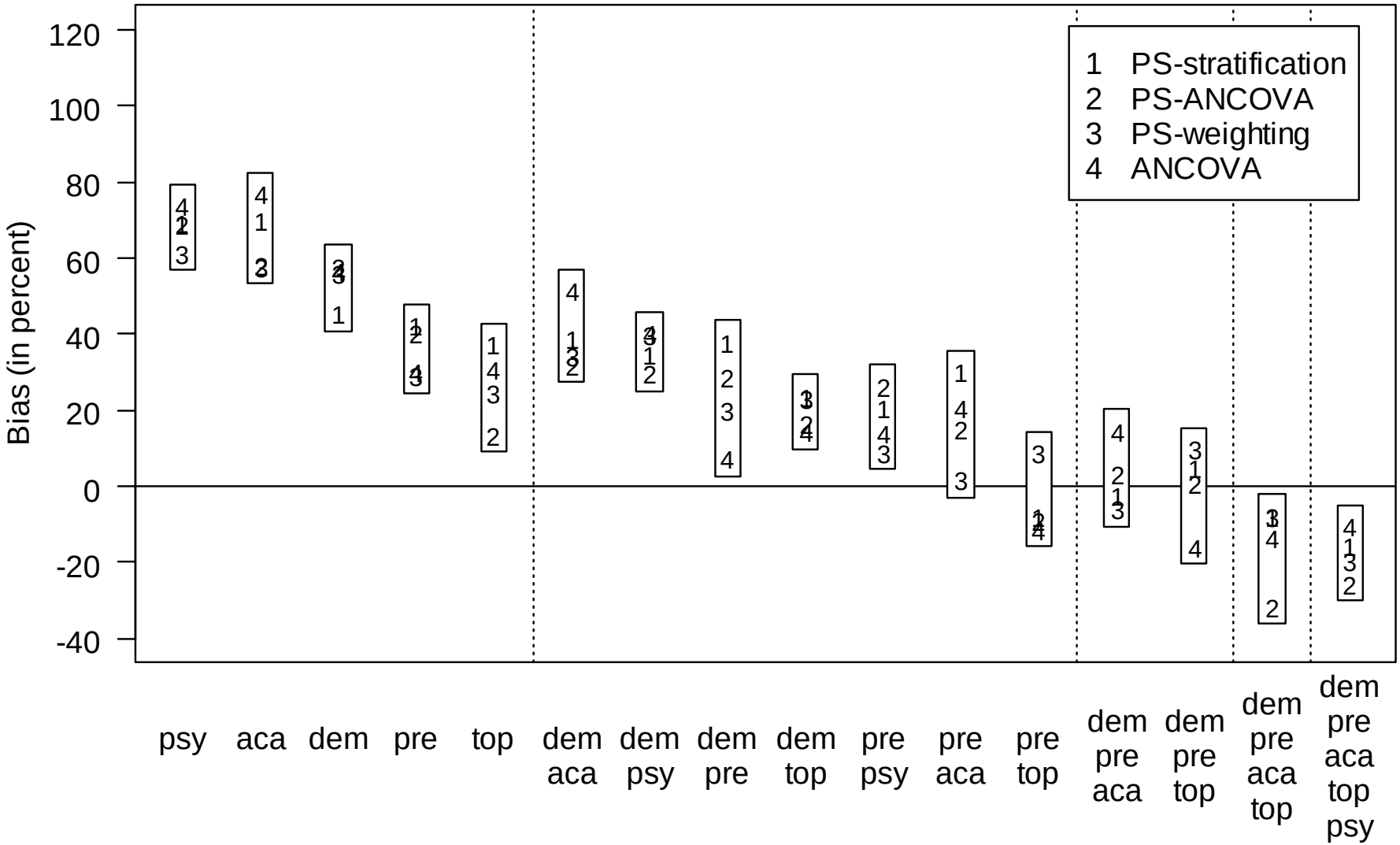
Jason Luellen's Dissertation



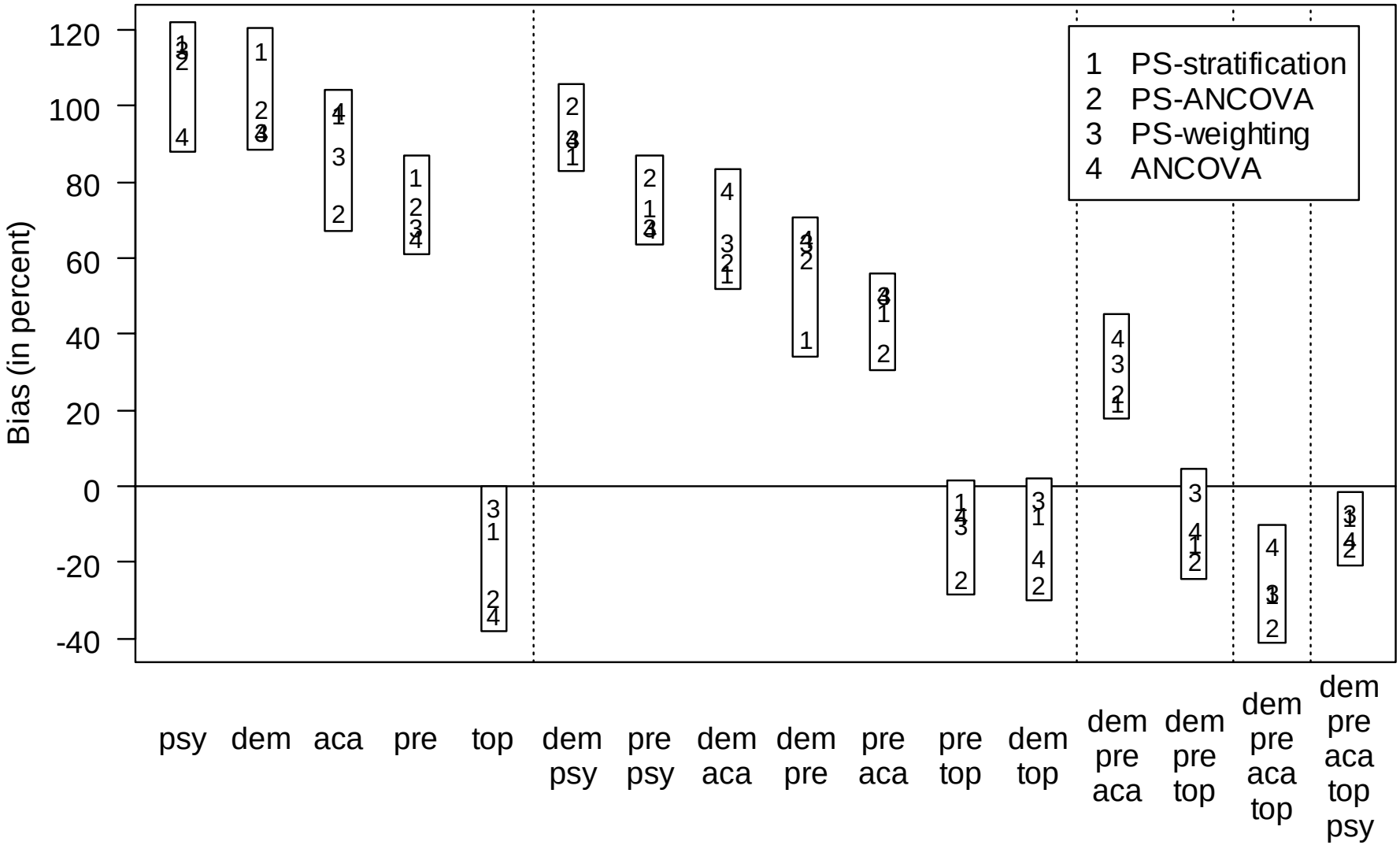
Further Analyses of the Shadish et al Data by

Steiner, Cook & Shadish

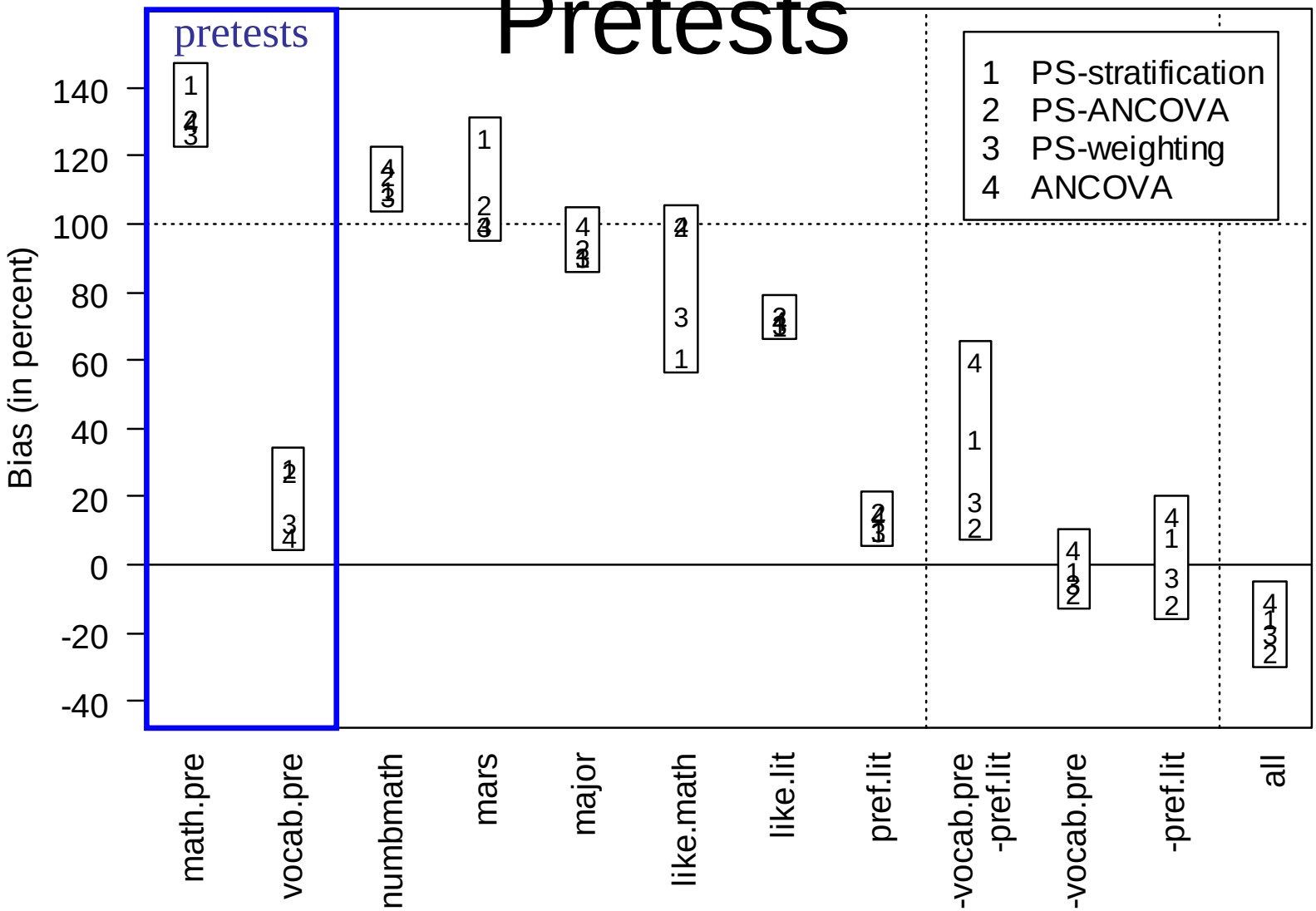
Remaining Bias: Vocabulary Covariate Sets



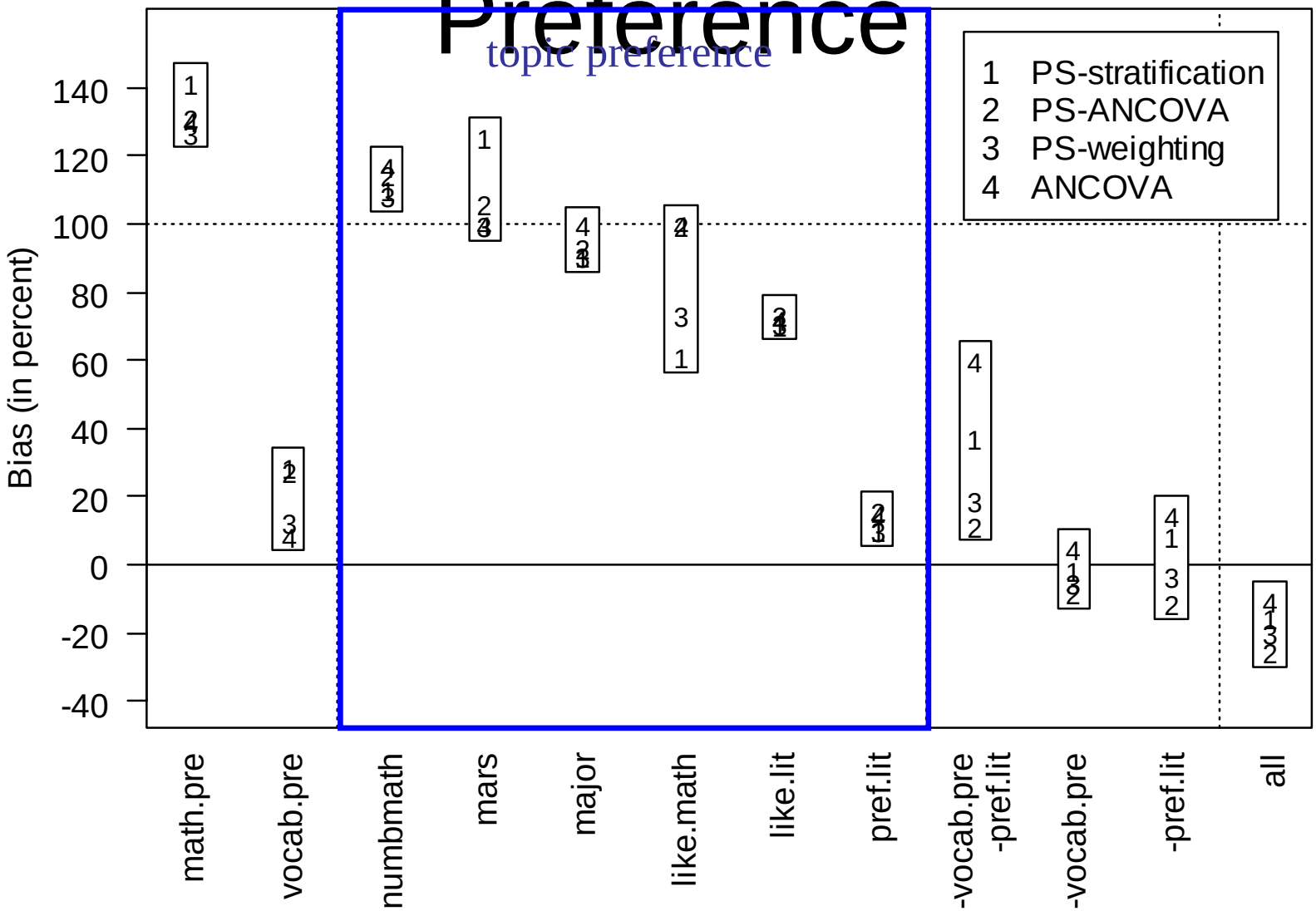
Remaining Bias: Mathematics Covariate Sets



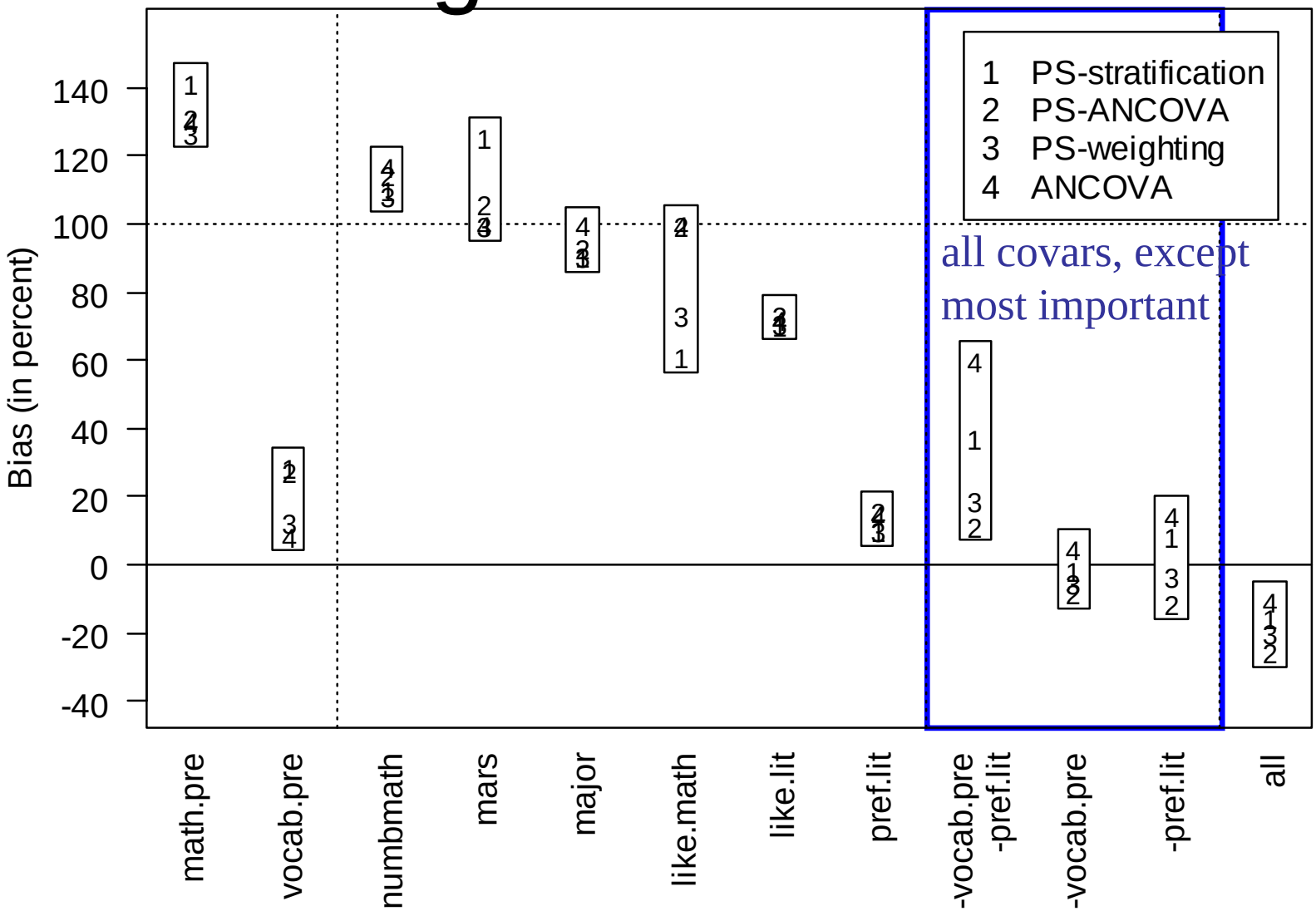
Remaining Bias: Vocabulary Single Covariates from Proxy-Pretests



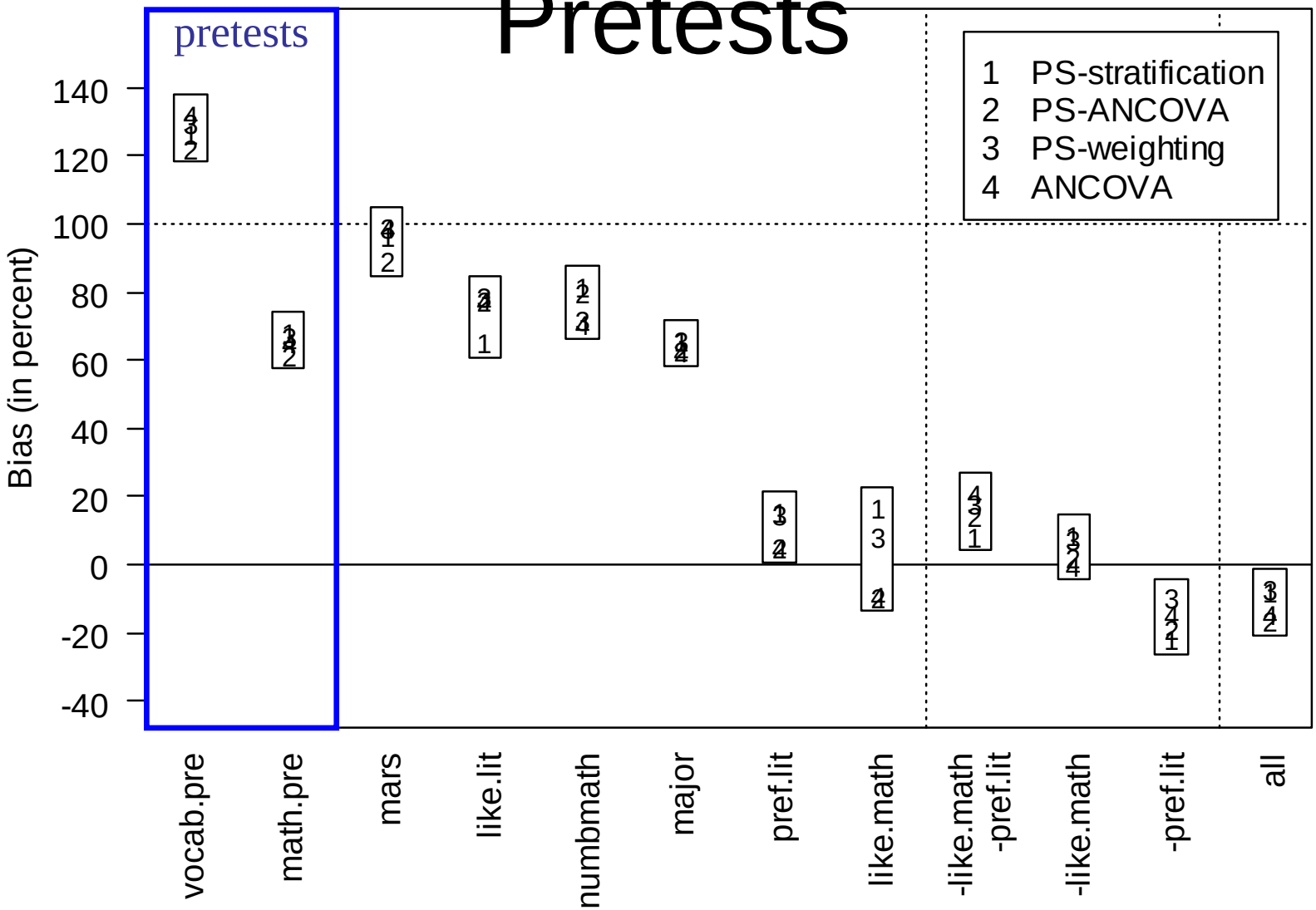
Remaining Bias: Vocabulary Single Covariates from Topic Preference



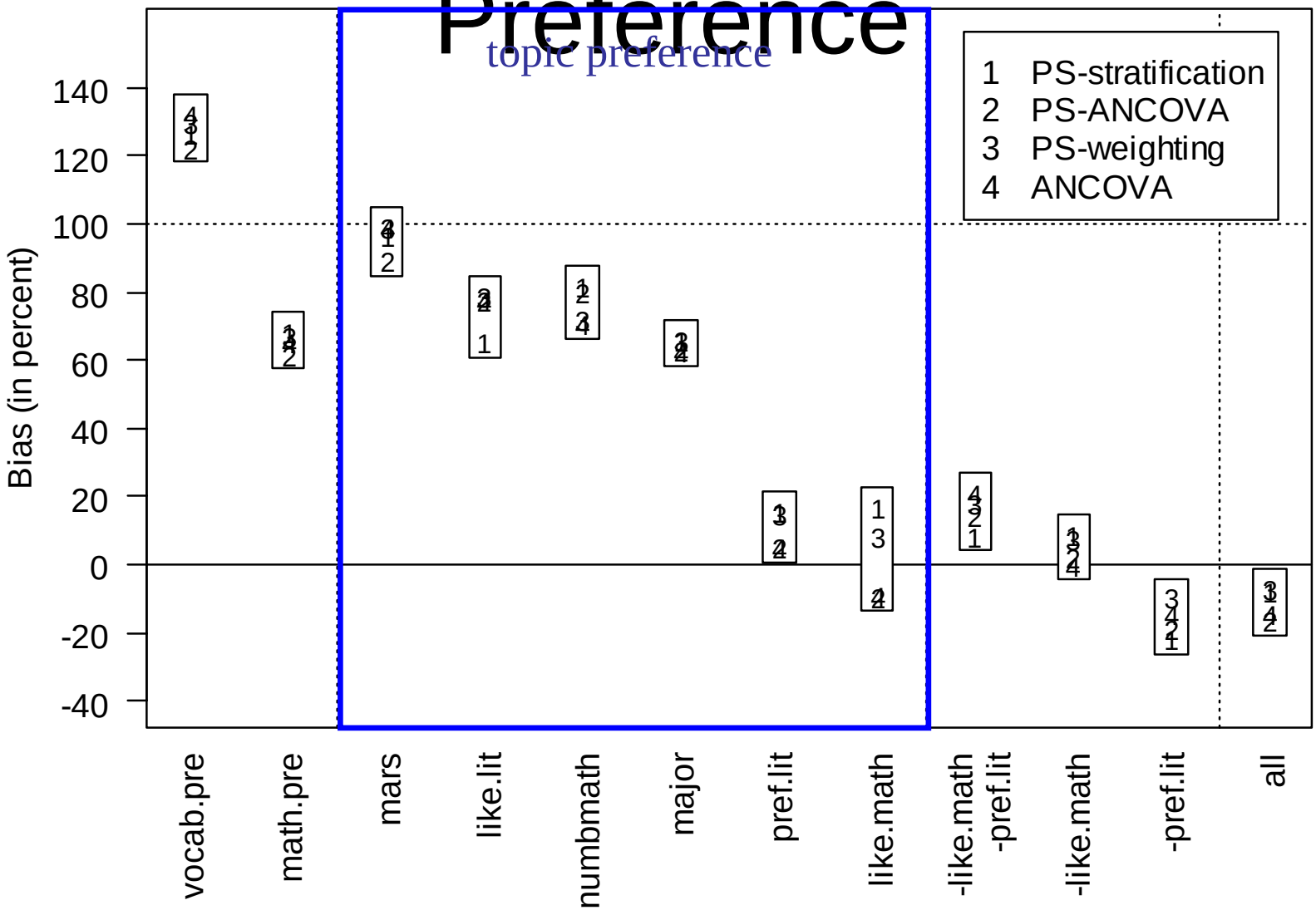
Remaining Bias: Vocabulary Single Covariates



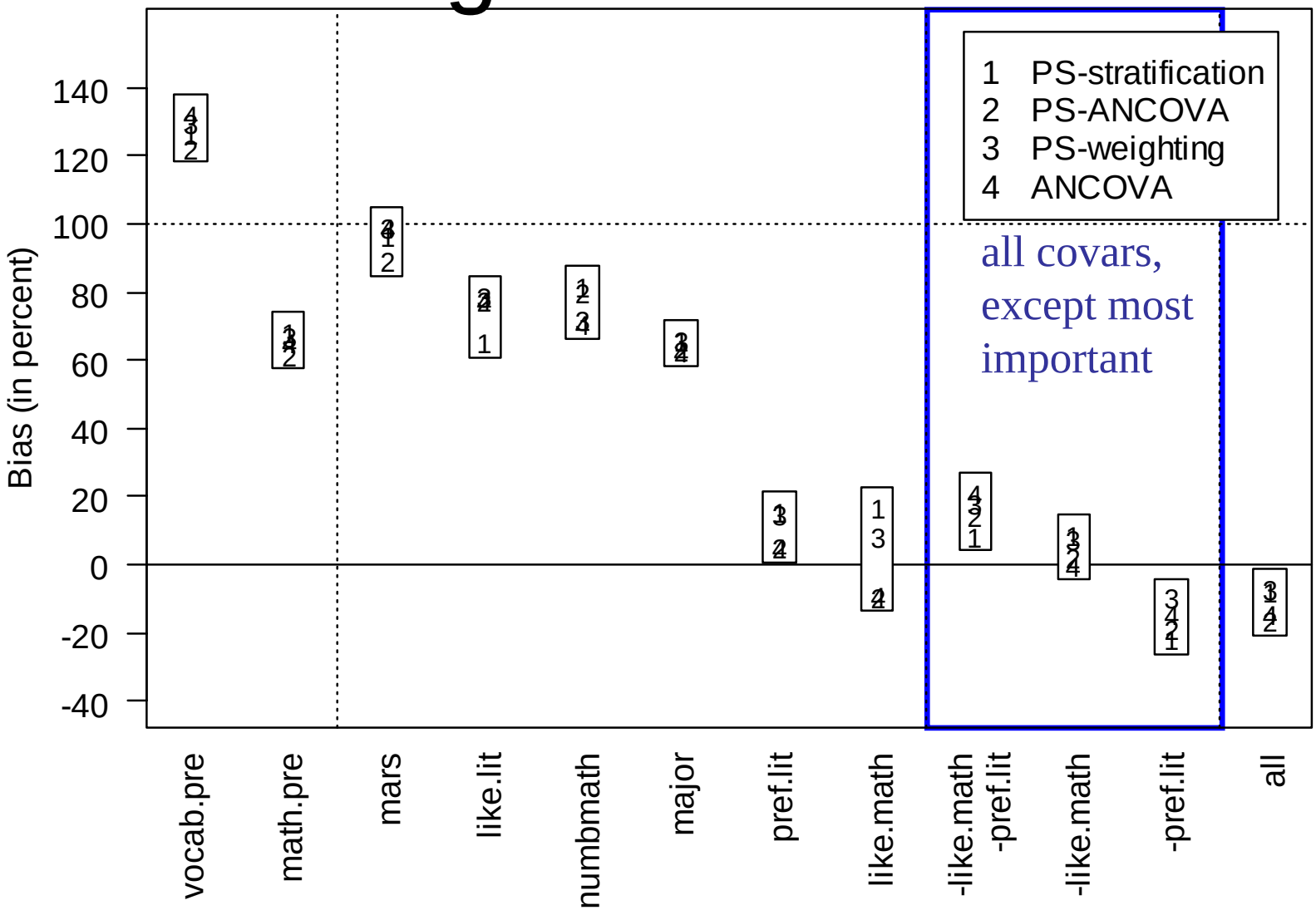
Remaining Bias: Mathematics Single Covariates from Proxy-Pretests



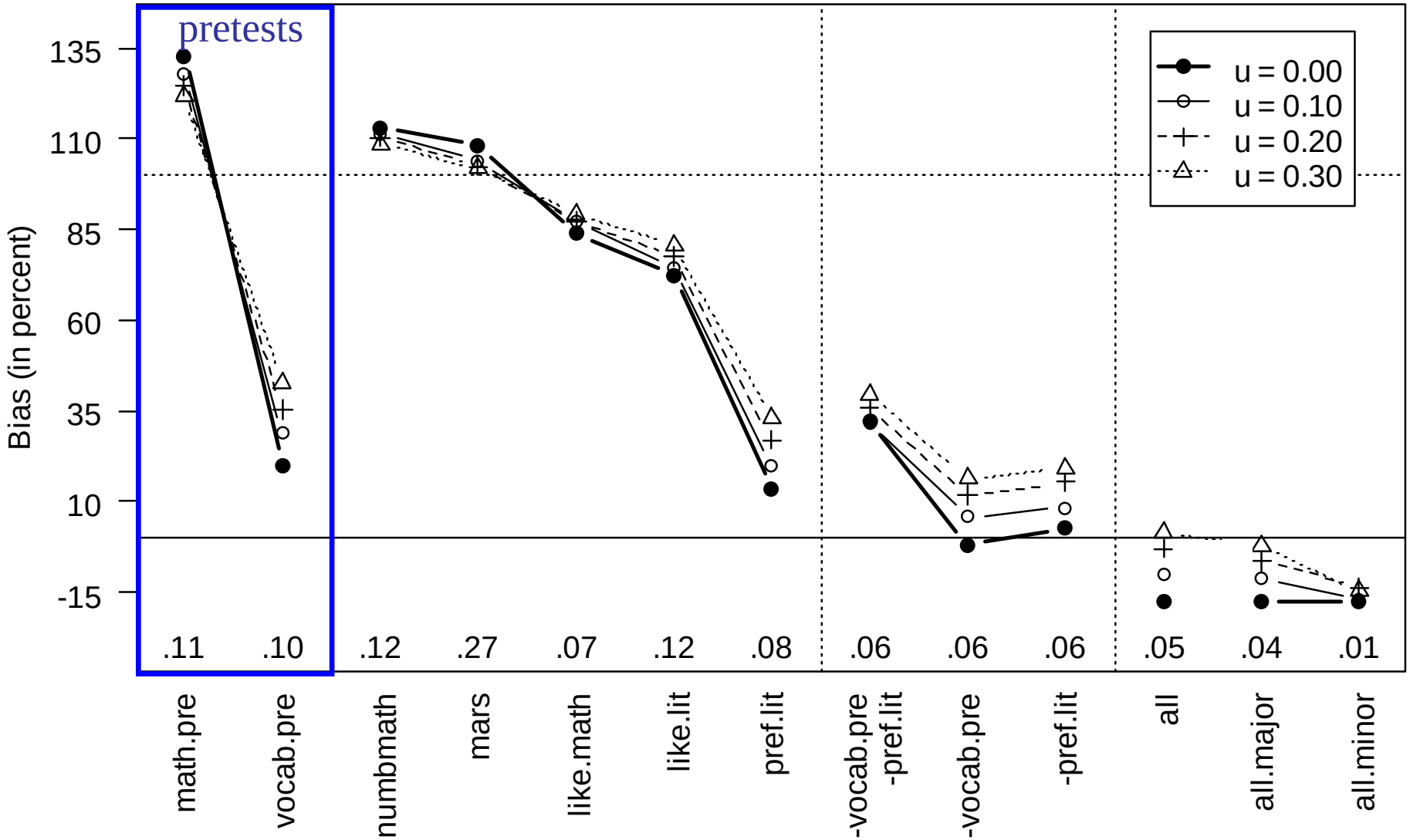
Remaining Bias: Mathematics Single Covariates from Topic Preference



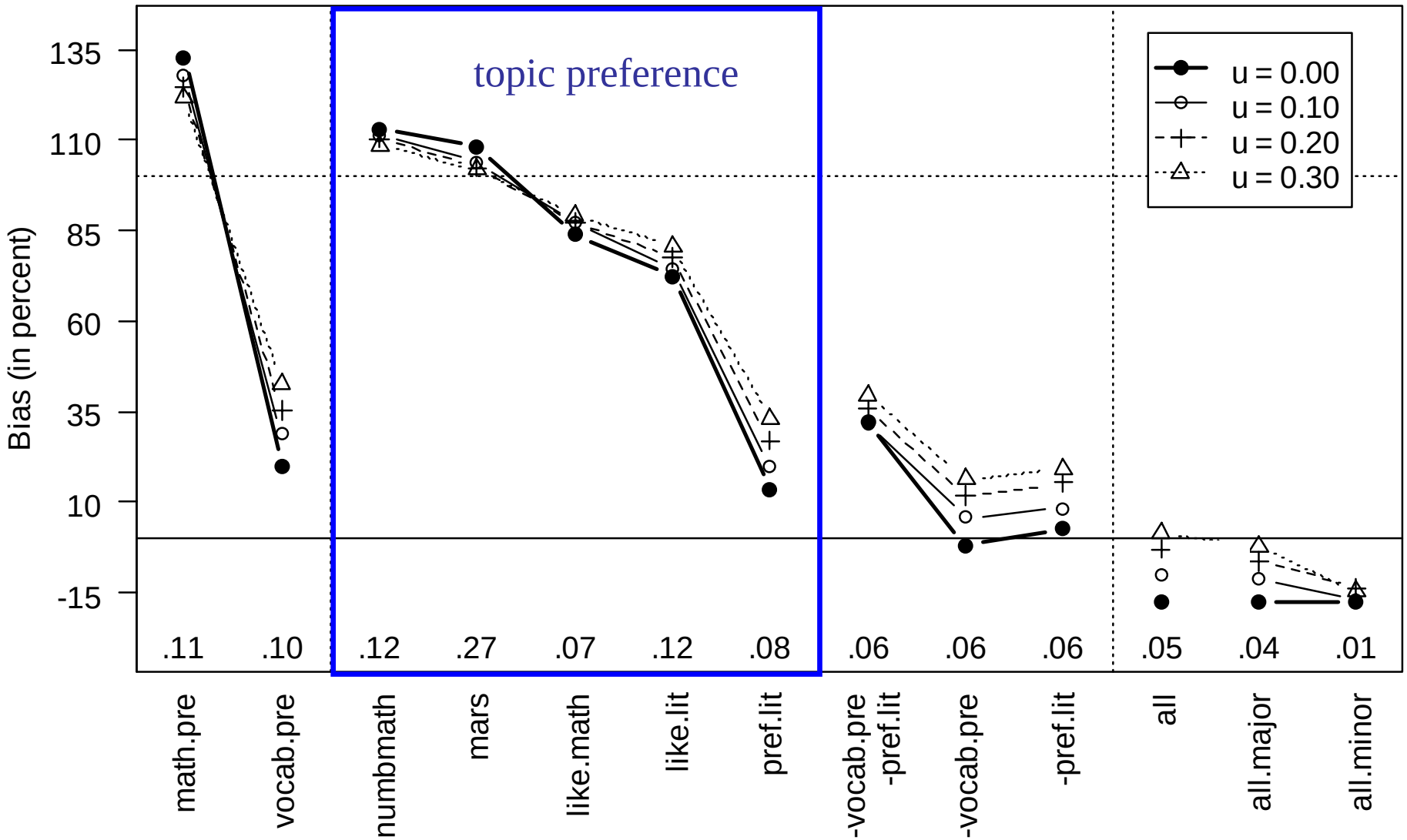
Remaining Bias: Mathematics Single Covariates



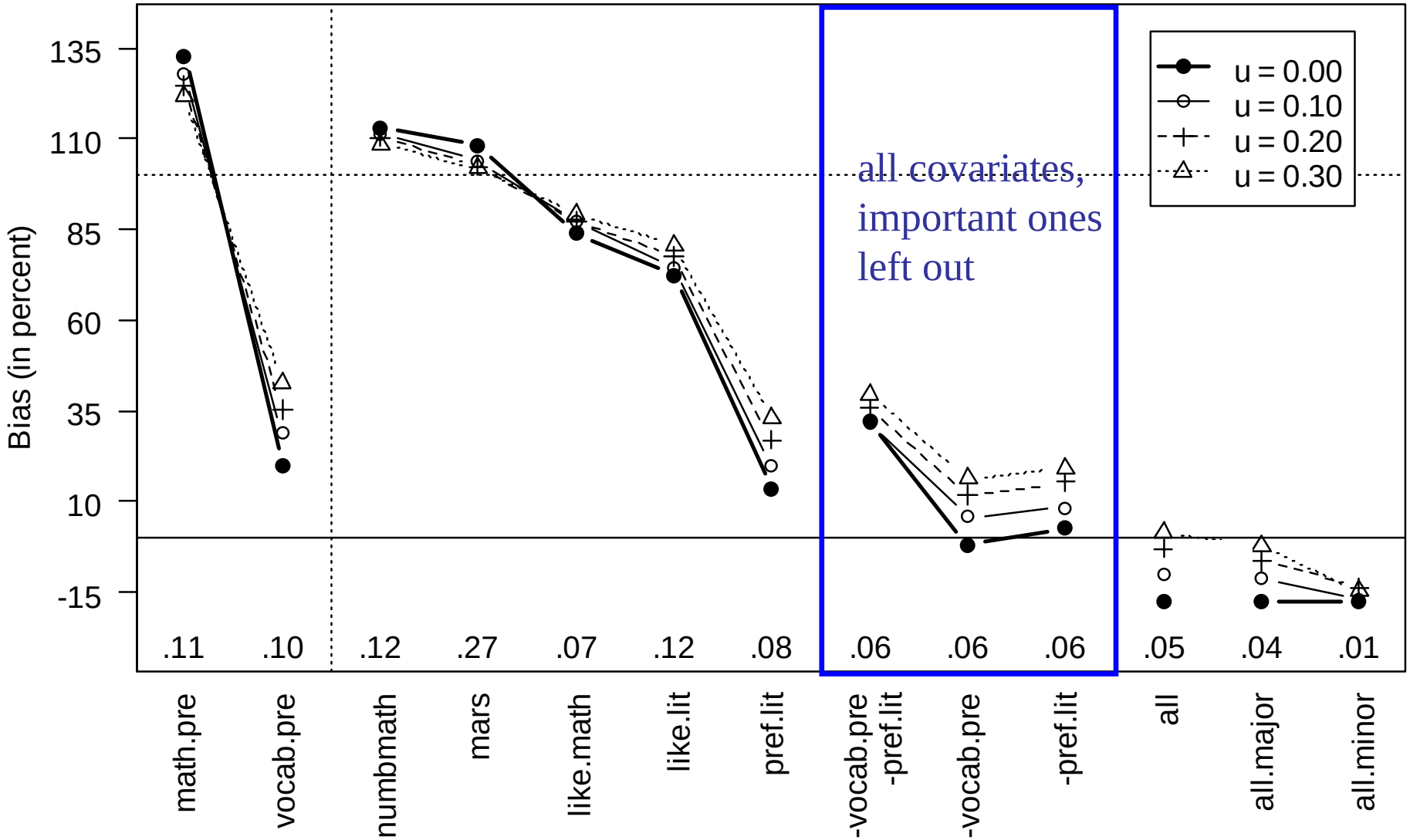
Unreliability: Vocabulary



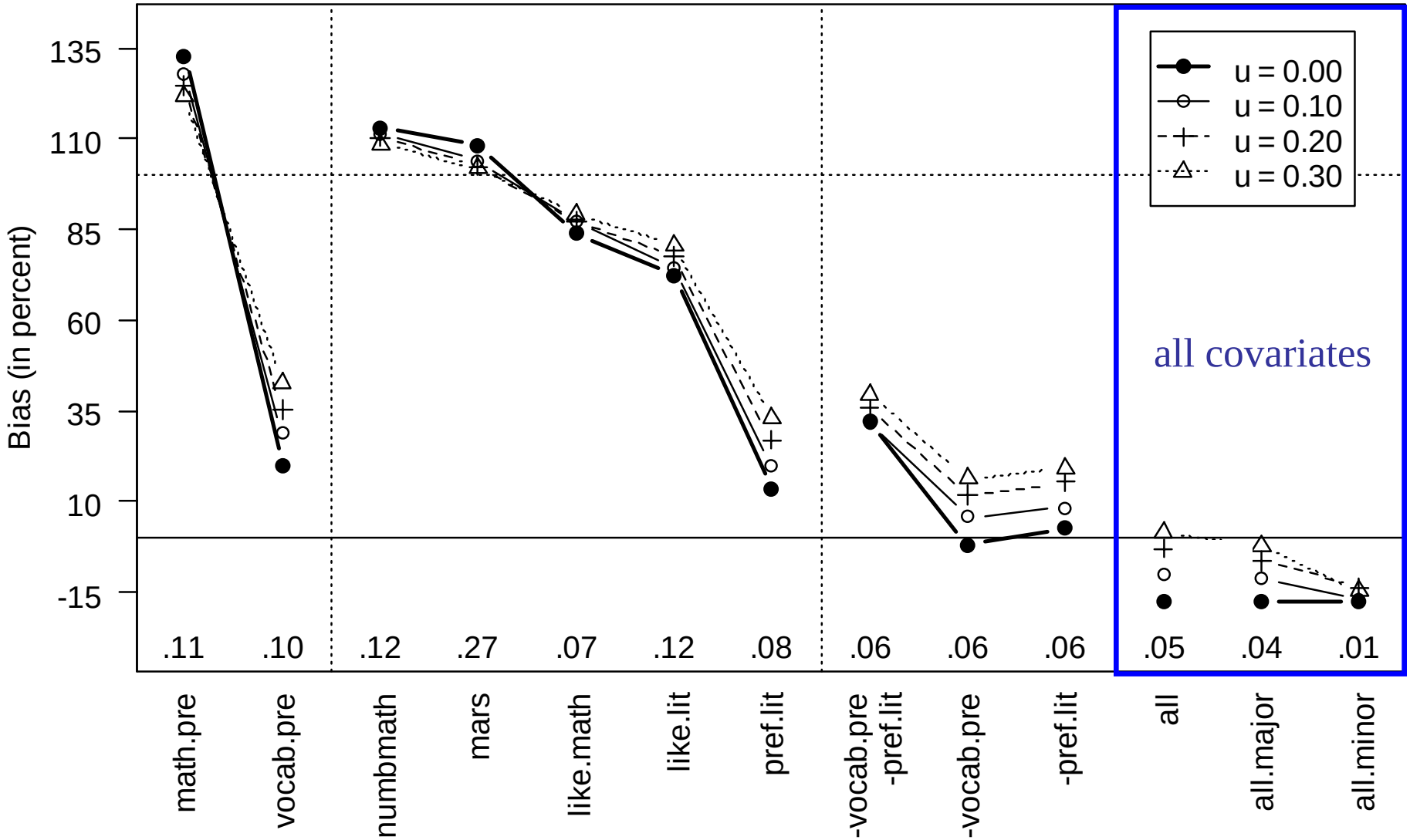
Unreliability: Vocabulary



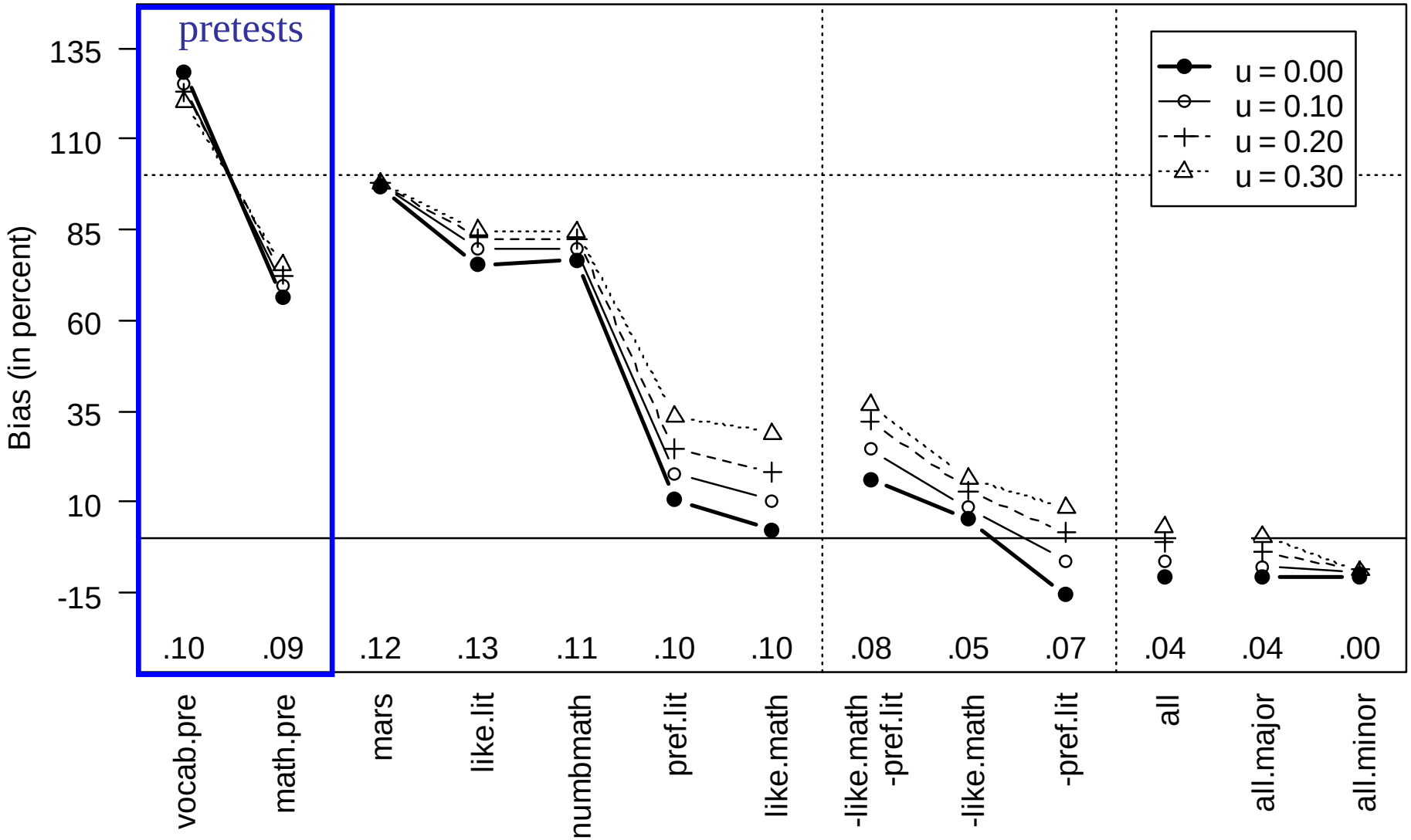
Unreliability: Vocabulary



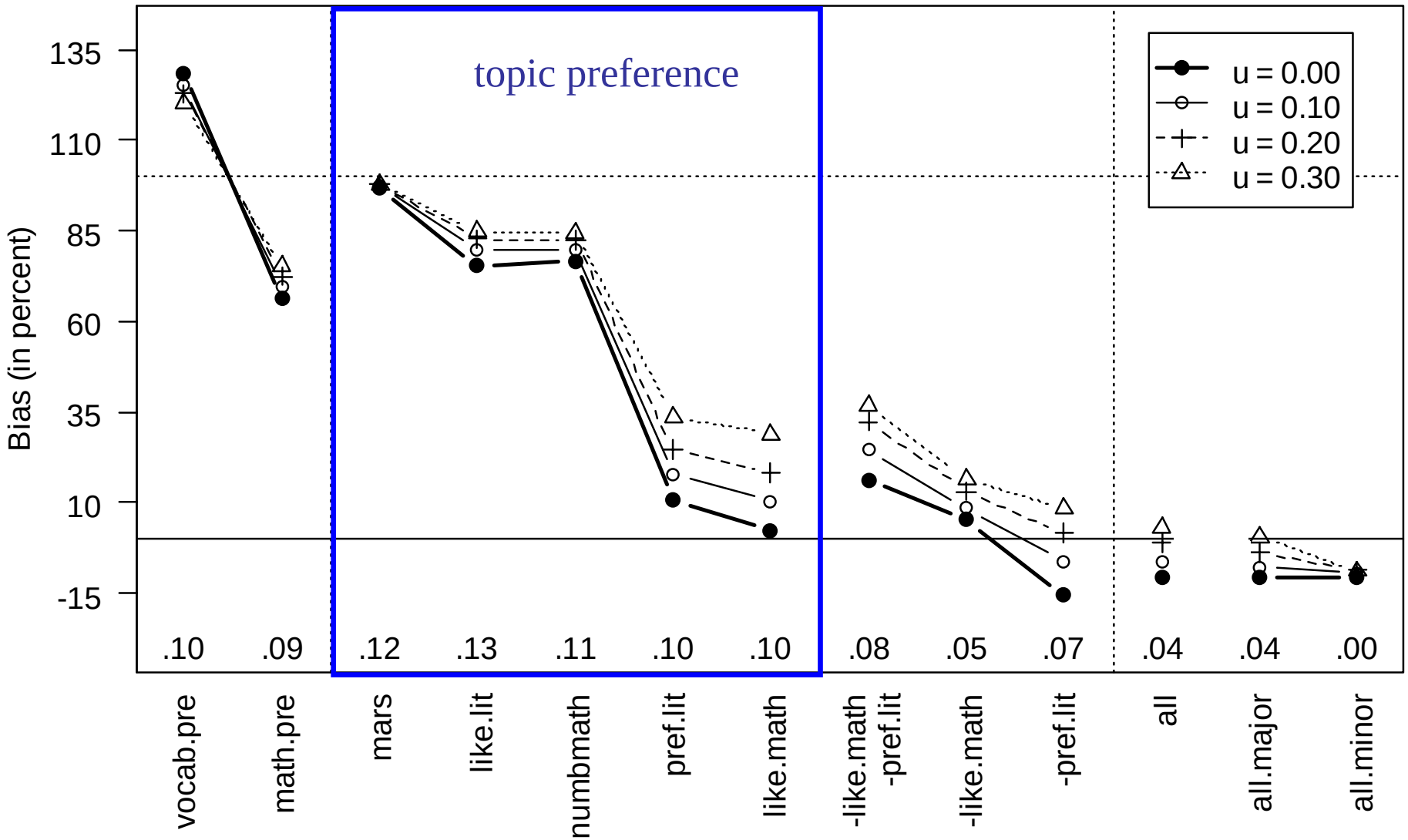
Unreliability: Vocabulary



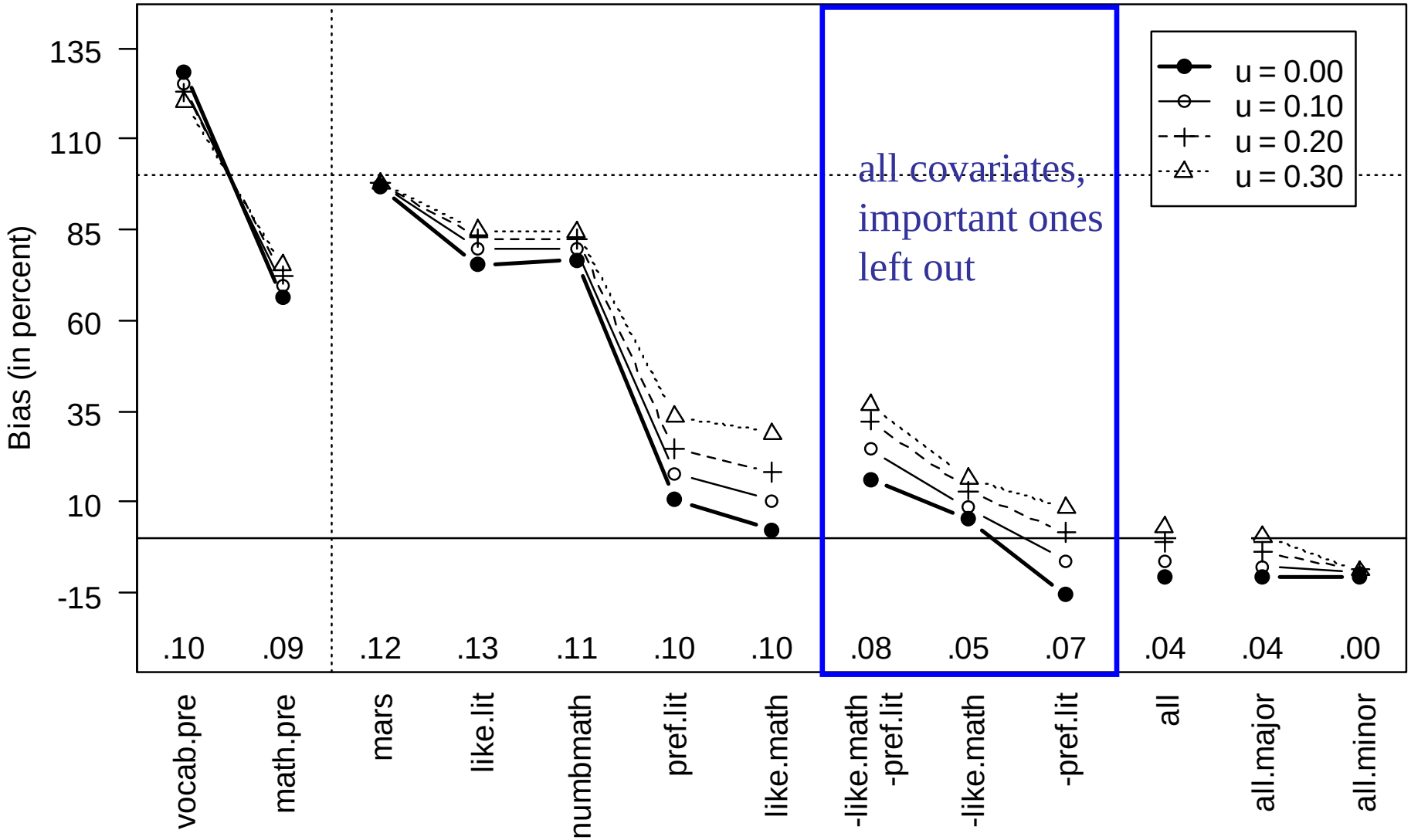
Unreliability: Mathematics



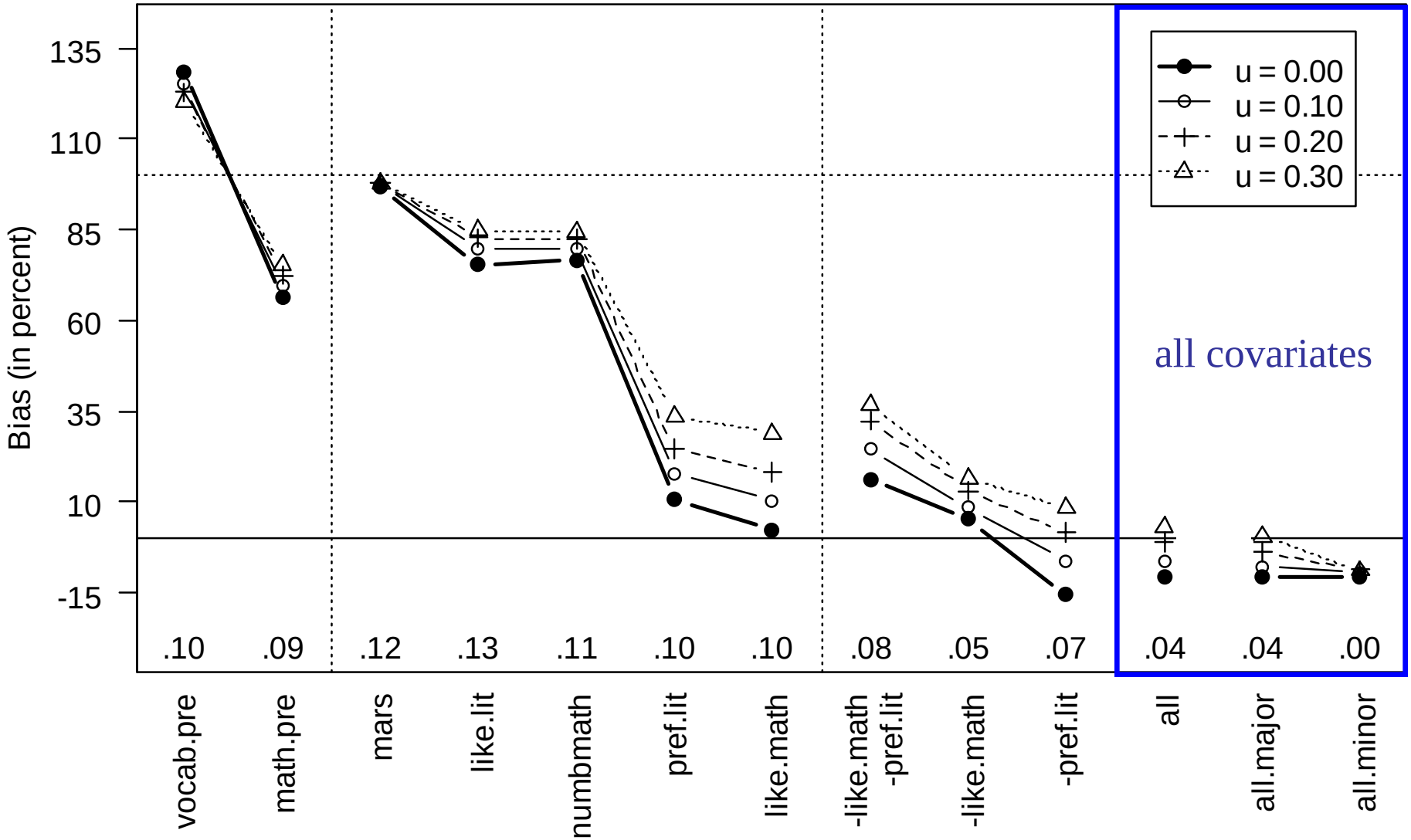
Unreliability: Mathematics



Unreliability: Mathematics



Unreliability: Mathematics



Diaz & Handa: Other Non-Equivalent Sample

- Sample of non-equivalent and clearly richer villages
- Big differences between randomly formed eligible families and non-eligible ones, but some overlap
- Use of same means of material welfare as used for treatment assignment
- When used as covariates, all bias is non-experiment explained away

Diaz & Handa and Shadish et al

- Complete knowledge of selection process in Diaz & Handa
- Highly plausible knowledge in Shadish et al
- Shadish et al second route to bias reduction

Within Study Comparison in Job Training

- 12 studies
- No close correspondence of results
- Identification of some facilitating conditions
- Deep pessimism

Within Study Comparison in Cook, Shadish & Wong

- Multiple domains
- 12 comparisons in 10 studies
- RD
- Intact Group Matching
- Selection process known
- 4 others
- More optimistic implications

Education: Wilde & Hollister

- The Experiment is Project Star in 11 sites
- The non-equivalent comparison group formed from other Tenn. sites via propensity scores
- No pretest, but proxy background variables and some school data
- Analysis of 11 exp vs non-exp comparisons
- Conclusions are: (1) no equivalence in individual site comparisons of exp and non-exp results
- (2) pooled across sites, each significant but they differ in magnitude (.69 vs. 1.07)

What's debatable here?

- Design first: Who would design a quasi-experiment on this topic w/o pretest? with non-local and non-intact matches?
- Analysis: How good is a propensity score analysis with mostly demographic data?
- How valid is it to examine separate sites
- Does this study compare a good experiment with a bad quasi-experiment?

Agodini & M.Dinarski

- The experiment is on dropout at individual level at middle and high school
- The workhorse design uses propensity scores
- They are constructed separately from two sources--one another study at the middle school level and the other national data
- Findings: Few balanced matches are possible (29 of 128), given covariate data available and overlap achieved;
- Where balanced experiment and non-experiment do not produce same results

Commentary

How good was the experiment? 2 of 5 sites

- Control cases not from high school
- Testing at different times
- Pretest measures mostly not available
- How rich was the covariate structure? No theory of dropping out used, merely what was available in archive
- Modest exp contrasted with poor non-experiment

Summary on Fixed Effects

- Much discussion that workhorse design is empirically not validated
- True for low quality quasi-experiments that no-one trained in Campbell tradition would ever do
- Not universally true--e.g., local focal matching can reproduce the results of experiments, as with Bloom et al and Aiken et al.
- Not true if rich covariates are available that assess assignment process well--e.g. Shadish
- The mantra: We have a problem of unknown selection; not of selection per se