

Non-Equivalent Control
Group Designs
(Observational Studies)

The Plan

- Today
 - Overview of non-equivalent control group designs/observational studies (OS)
 - RE-OS WSCs
 - Purposes for RE-OS WSCs
 - Common Issues to Consider
 - Examples using the WSC of Shadish et al. (2008)
 - Discussion example: Peikes et al. (2008)
- Tomorrow:
 - WSCs in the context of multilevel data
 - Selected topics

RCM & treatment effects

Strong ignorability

Propensity score techniques

OVERVIEW OF NON-EQUIVALENT CONTROL GROUP DESIGNS (PS MATCHING)

Observational Study

- *Non-equivalent control group designs* (observational studies) are characterized by
 - deliberate *self-selection* and/or *third-person selection* of subjects into the treatment and control condition
 - *pretreatment group differences* in covariates (observed and unobserved)
 - *selection bias* in the unadjusted average treatment effect (due to differential selection)
- Independence (*strong ignorability*) can be established if the selection mechanism is completely known and observed

Strong Ignorability (SI)

- If all covariates $\mathbf{X} = (X_1, \dots, X_p)'$ related to *both* treatment assignment and potential outcomes are observed, and
- if the selection probabilities, given \mathbf{X} , are strictly between zero and one $0 < P(Z = 1 | \mathbf{X}) < 1$ holds
- then, potential outcomes are independent of treatment assignment given observed covariates \mathbf{X} :

$$(Y^0, Y^1) \perp Z | \mathbf{X}$$

and treatment assignment is said to be strongly ignorable (*strong ignorability*; Rosenbaum & Rubin 1983). → no hidden bias

Practice of Strong Ignorability

- Need to *measure all confounding covariates*! If *not* all covariates, that are simultaneously related to treatment selection and potential outcomes, are observed
 - the strong ignorability assumption is not met and
 - the average treatment effect will remain biased!
- Need to measure covariates *reliably* (with respect to the selection mechanism)
- Each subject must have a *positive probability (but less than one)* of being in the treatment group (overlap).

SUTVA

- Besides the strong ignorability assumption we need a second assumption (Rubin, 1990):
Stable-unit-treatment-value assumption (SUTVA)
- SUTVA implies
 - only *one treatment* (no variations of the treatment)
 - *no interference* among units (being aware of other subjects' assignment has no effect)
 - *type of assignment* does not matter

How to Estimate an Unbiased Treatment Effect?

- Assume that we observe all covariates \mathbf{X} such that SI holds, selection bias can be removed with different approaches
- With original covariates \mathbf{X}
 - Covariance adjustments (standard regression methods)
 - Case matching on observed covariates
 - Multivariate stratification
- With a composite of original covariates $b = f(\mathbf{X})$
 - *Propensity scores* (Rosenbaum & Rubin, 1983)
 - Other approaches we will not consider: first discriminant, prognosis scores

Propensity Scores (PS)

- The propensity score is the *conditional probability* that a subject belongs to the treatment group given the observed covariates \mathbf{X} :

$$e(\mathbf{X}) = P(Z = 1 | \mathbf{X})$$

- If treatment selection is strongly ignorable given an observed set of individual covariates \mathbf{X} , then it is also *strongly ignorable* when these individual covariates are combined into a propensity score $e(\mathbf{X})$, i.e.

$$(Y^0, Y^1) \perp Z | e(\mathbf{X}) \quad \text{with} \quad 0 < e(\mathbf{X}) < 1$$

(Rosenbaum & Rubin 1983)

Estimation of PS

- Propensity Score:
 - conditional probability of receiving treatment given all observed covariates \mathbf{X} : $e(\mathbf{X}) = P(Z = 1 | \mathbf{X})$
 - Weighted composite of all observed covariates
- Different methods for estimating PS can be used:
 - *Binomial models*: logistic regression, probit regression
 - Rely on functional form assumption
 - *Statistical learning algorithms* (data mining methods): classification trees, boosting, bagging, random forests
 - Do not rely on functional form assumptions
 - It is not yet clear, whether these methods are on average better than binomial models

Choice of PS technique

- Four groups of PS techniques (Schafer & Kang, 2008; Morgan & Winship, 2007; Steiner, 2012):
 - *PS-matching* (individual case matching)
 - *PS-stratification*
 - *Inverse-propensity weighting*
 - *Regression estimation using PS*
- Combination with an additional covariance adjustment (within the regression framework): *Mixed methods*
 - Makes estimates “*doubly robust*”, i.e., robust against the misspecification either of the PS model or outcome model
 - However, if both models are misspecified, doubly robust methods might perform worse (Kang & Schafer, 2007)

Purposes of WSCs

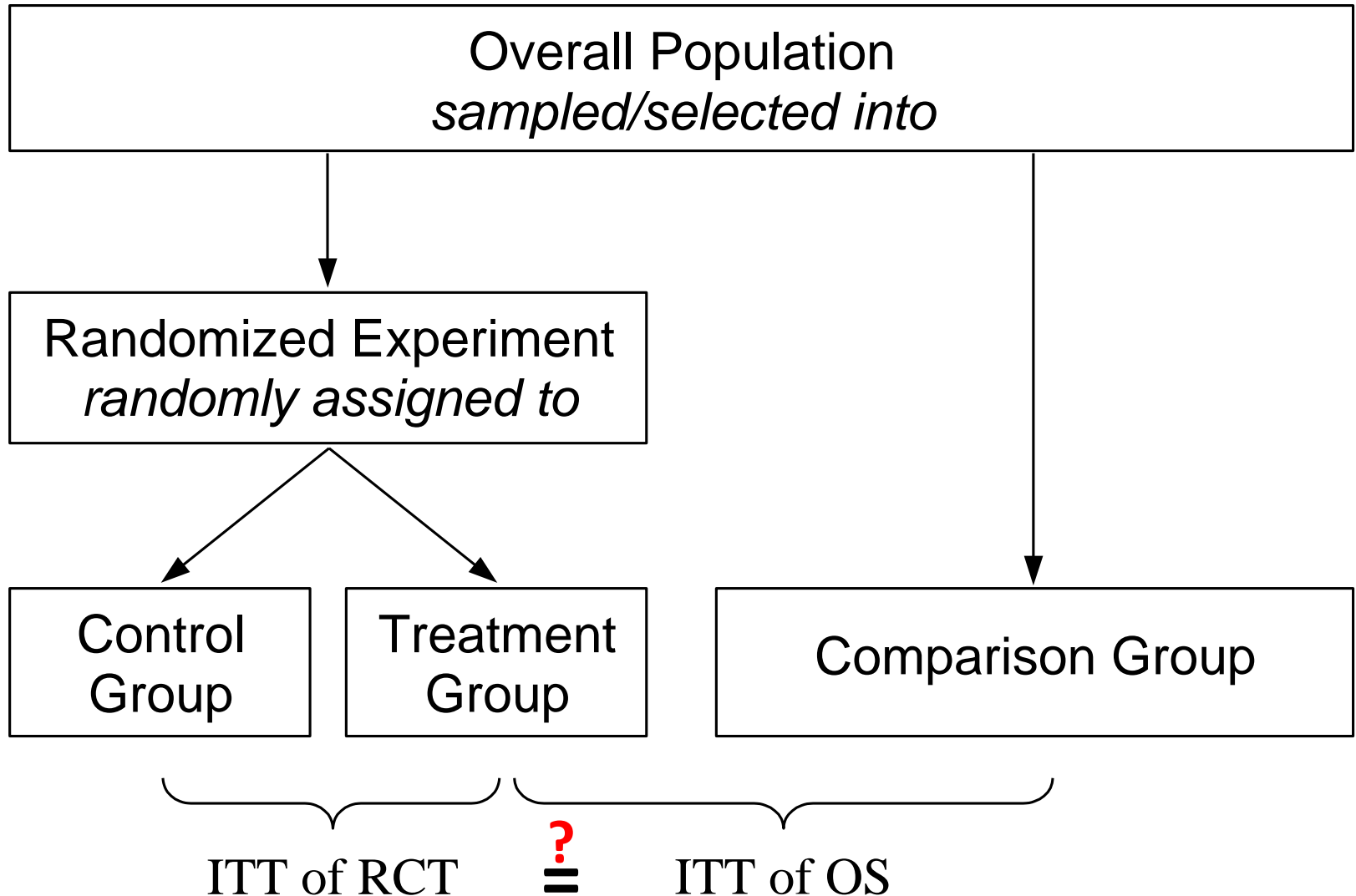
Common issues to consider

WITHIN-STUDY COMPARISONS WITH NON-EQUIVALENT CONTROL GROUP DESIGNS (OBSERVATIONAL STUDIES)

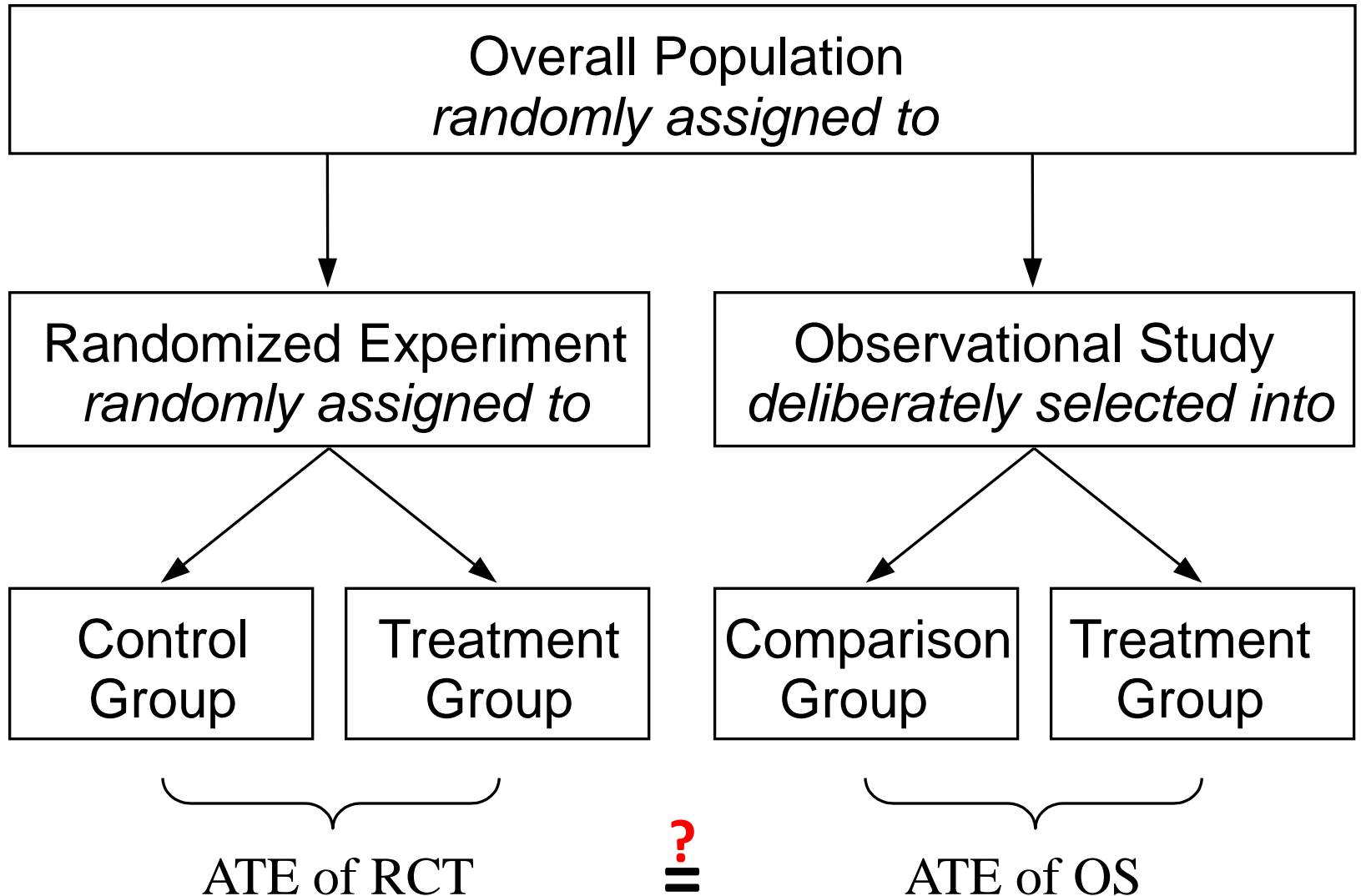
WSC in Matching Contexts

- The identification and estimation of causal treatment effects relies on strong *assumptions*: strong ignorability, SUTVA, correct specification of PS model
- WSCs can be used to probe these assumptions:
 - Are the assumptions actually met in practice?
 - Do our current methods (& design elements) for detecting violations of these (untestable) assumptions work?

Three-Arm WSC Design



Four-Arm WSC Design



Purposes of Matching WSCs

1. *Replication*: can matching methods produce unbiased results, as compared to an experimental benchmark?
2. *Analytic method*: which analytic methods work best?
3. *Covariate selection*:
 - which covariates need to be measured?
 - which covariate selection strategies work best?
4. *Matching strategies* for multilevel data: which matching strategies work best for multilevel data?
5. *Design elements*: Can design elements detect threats to validity?
6. *Other purposes*

Purpose 1: Can OSs Replicate RE Results?

- WSCs probe whether the additional design and analytic *assumptions are likely to be met* in practice
 - Do the observed covariates establish a *strongly ignorable selection* mechanism?
 - Is *SUTVA* met?
 - Are we able to *correctly specify the PS model* (i.e., achieve satisfactory balance on observed covariates)?
- Lalonde (1986), Dehejia & Wahba (1999), Michalopoulos et al. (2004), Peikes et al. (2008), Shadish et al. (2008), ...

Purpose 2: Which Analytic Methods Work Best?

- Independent of whether the design assumptions are met, we can test which *analytic methods* work best in practice
 - Which *PS technique* (matching, stratification, weighting, regression adjustment) reduces most effectively selection bias (Shadish et al., 2008)?
 - Do *doubly robust methods* (mixed methods) perform better than standard PS adjustments?
 - Are PS techniques better than *multivariate matching* methods (Mahalanobis distance matching) or *standard regression adjustments* (ANOVA, OLS) (e.g., Shadish et al., 2008; Pohl et al., 2009)?

Purpose 3: Covariate Selection

- The *strong ignorability* assumption is the most crucial assumption for identifying a causal treatment effect (i.e., for removing all the selection bias).

WSCs allow us to probe

- which *covariates* need to be measured in order to remove all the bias (Cook et al., 2010; Steiner et al., 2010)
- which *covariate selection strategies* work best (Steiner, Cook & Li, in prep)

Purpose 4: Matching Strategies for Multilevel Data

- With multilevel data, one can *match at different levels*—at the cluster level, individual level or both
 - For *cluster-level interventions* we can compare
 - matching of intact clusters with individual level matching and two-stage matching (first cluster, then individuals) (Hallberg et al., in prep)
 - For *individual-level interventions* we can compare
 - Individual level matching *within* clusters to matching *between* clusters
- (→ See more detailed discussion below)

Purpose 5: Probing Design Elements for Ruling Out Threats to Validity

- *Design elements* like non-equivalent outcomes or multiple comparison groups can be tested whether they actually work for ruling out threats to validity
(for a list of design elements see Shadish, Cook & Campbell, 2002, page 157)
- (A) OS replicates RE results:
 - Design elements should indicate no threat to validity
- (B) OS does not replicate RE results:
 - Design elements should indicate violations of assumptions

Purpose 5: Probing Design Elements for Ruling Out Threats to Validity

Example: *Non-equivalent outcome*

- Intervention: new math curriculum
- Outcome: math achievement score
- Non-equivalent outcome: reading score
 - Reading score is not affected by math intervention
 - Reading and math score are highly correlated
- Apply the same PS technique to the reading outcome (with the same PS as used for the math outcome)
 - (A) OS replicates RE math results: we would expect no significant effect of the math intervention on the reading score
 - (B) OS does not replicate RE math results: a significant reading effect might show up (most likely due to remaining selection bias)

Purpose 6: Other issues

- WSCs can be used to compare
 - different *PS estimation strategies*: (semi)-parametric logistic/probit models, statistical learning algorithms (e.g., random forests, boosted regression, support vector machines)
 - *PS model selection criteria*: Balancing criteria, model selection procedures
 - *matching designs* to *other quasi-experimental designs*, e.g., instrumental variable design or difference-in-differences design
- Preference effects & randomization effects

Purpose 6: Other issues

- WSCs can be used in simulation studies to investigate the effect of
 - covariate *measurement error* on OS estimates (Steiner, Cook & Shadish, 2011)
 - add measurement error to covariates
 - *small sample sizes* on OS estimates, particularly small samples of treated cases (Kolar & Vehovar, 2012)
 - draw small random samples of treatment and comparison cases from available WSCs
 - other

Common Issues with RE-OS WSCs

1. Which factors of a WSC should be controlled?
2. Causal estimands of interest
3. Comparability of target populations
4. Overlap issues
5. Doubly robust methods (for both RE and OS)
6. Standard errors
7. Metric of RE-OS comparison
8. Power of RE & OS

1. What Should be Held Constant?

- An RE and OS might deliberately differ or be controlled with regard to *third-variable confounds*, i.e., factors that are not directly related to treatment selection, for example,
 - underlying target population (location)
 - measurement particulars (time, instruments)
 - implementation of treatment (laboratory vs. in the field)
 - causal estimand & estimation method
 - missing data (avoidance, handling)
 - overlap in OS treatment & comparison group / RE

1. What Should be Held Constant?

Which third-variable confounds should be held constant depends on the research question

- From a *policy* point of view, we might not be interested in controlling for all possible third-variable confounds; We are more interested in comparing the best “real-world” OS to the best “real-world” RE

- target populations
- treatment implementation
- analytic methods
- handling of missing data
-

} very likely differ if two independent research teams would conduct an RE and OS

1. What Should be Held Constant?

- From a *methodological* point of view, we are interested in whether OSs can reproduce REs under ideal conditions, i.e., when all third-variable confounds are held constant
 - same target population
 - no difference in treatment implementation
 - no differential handling of missing data
 - comparable analytic method
 - ...only the removal of *bias due to selection* is studied
- Will discuss issues mostly from methodological point of view

2. Causal Estimands

- *Randomized experiment:*
 - full compliance → average treatment effect (ATE) (= ATT)
 - non-compliance → intent-to-treat effect (ITT)
→ complier (local) average treatment effect (CATE; LATE)
= ATT in case of one-sided non-compliance (only no-shows)
- *Observational study:*
 - Average treatment effect (ATE)
 - Average treatment effect for the treated (ATT)
 - (ITT in case of three-arm WSCs)

3. Target Population

- *Target populations* of RE and OS need to be comparable; They should refer to the
 - same *units*
 - same *treatment*
 - same *outcomes*
 - same *setting*
- e.g., same location and time; if participating in the RE was subject to some eligibility restrictions, the same restrictions should be applied to the OS)
- Attrition in the RE might change the originally intended target population (the RE might no longer represent a reliable benchmark)

4. Lack of Overlap

- If the treatment and comparison group of the OS *lack complete overlap* not all treated cases can be matched to a comparison case (or vice versa)
- *Non-overlapping cases* are typically dropped from the analysis which restricts the target population of the OS
- In order to maintain a comparable RE benchmark, corresponding cases need to be dropped from the RE

5. Doubly Robust Methods

- Doubly robust methods (mixed methods): in addition to the PS adjustment, a *regression adjustment* with observed covariates is made (e.g., after matching cases on the PS logit, a standard OLS regression is conducted on the matched dataset)
- *Protects against the misspecification* either of the PS model or the outcome model (but not both)
- Regression adjustments are also frequently made for *REs* in order to remove random imbalances and increase efficiency

5. Doubly Robust Methods

- Should doubly robust methods be used for WSCs?
 - Yes, but preferably compare pure design-based and regression-adjusted treatment effects separately
 1. Compare pure RE results to PS-adjusted OS results (without any additional regression adjustment)
 - tests the different *designs* (without further statistical adjustments)
 2. Compare regression-adjusted RE to PS & regression-adjusted OS
 - compares *designs* plus additional *statistical adjustments*

6. Standard errors

- Comparing RE and OS results requires the estimation of standard errors that appropriately reflect sampling uncertainty
 - *RE results*: regression-based standard errors can be used
 - *PS results*: one might need to take the uncertainty with respect to the estimated PS into account
 - Adjusted variance formulas (e.g., Schafer & Kang, 2008)
 - Bootstrapped standard errors (also avoids relying on distributional assumptions)

7. Metric of Comparison

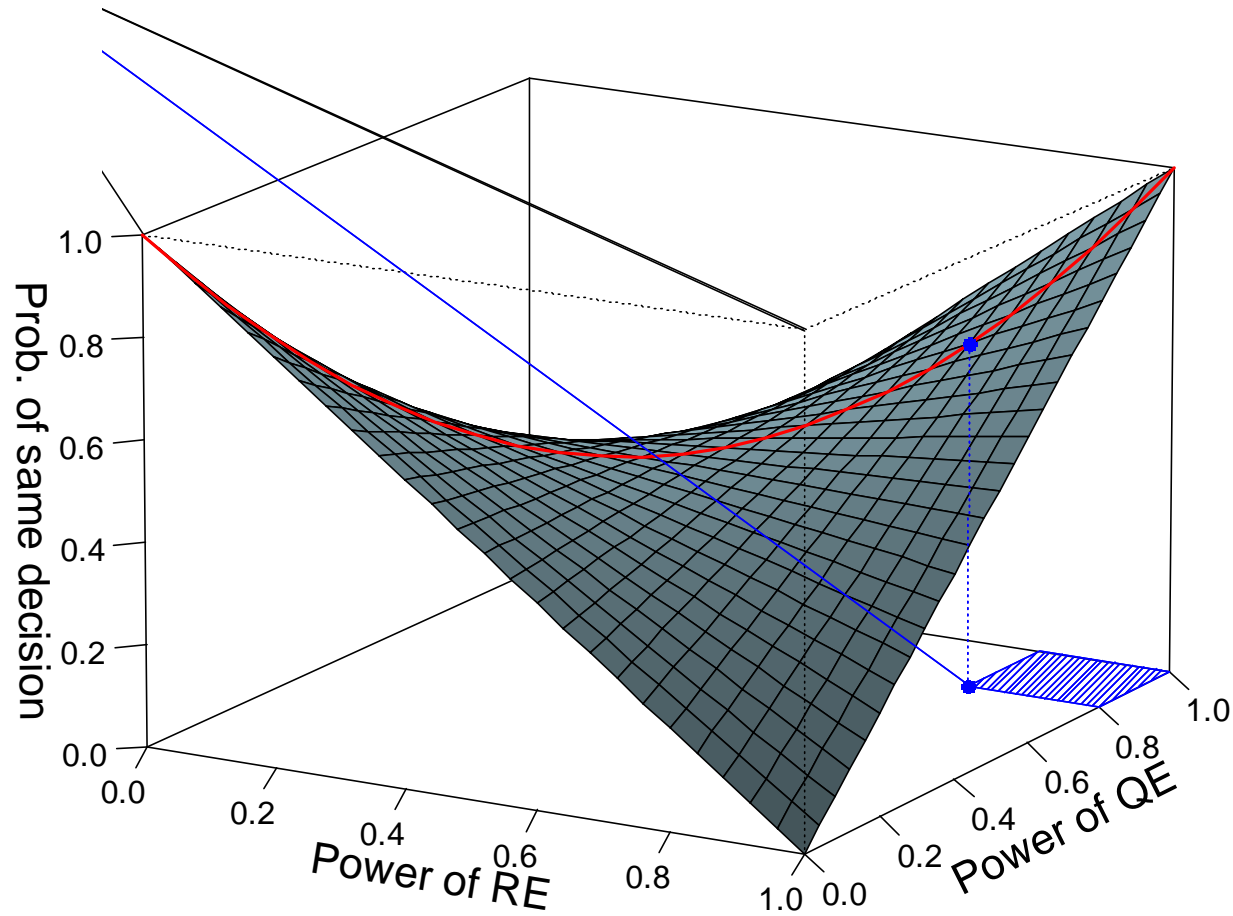
- Do OS and RE result in the *same conclusions*, i.e., the rejection or acceptance of a null hypothesis? (mostly of policy relevance)
- Do treatment effect estimates obtained from OS and RE *significantly differ*? (primarily of methodological relevance)
 - *Four-arm WSCs*: two sample *t*-test for treatment effects
 - *Three-arm WSCs*: two sample *t*-test of control group and matched comparison group means (since treatment groups are identical they cancel out)
 - Simultaneously *bootstrap* RE and OS data (estimate RE-OS difference for each bootstrap sample)

8. Power

- Do OS and RE result in the *same conclusions*, i.e., the rejection or acceptance of a null hypothesis?
 - Both, OS and RE need sufficient power
 - If power of OS and RE extremely high, the “same conclusions” criterion is not very discriminative
- Do treatment effect estimates obtained from OS and RE *significantly differ*?
 - Need sufficient power not only for showing a treatment effect but also for a minimum detectable *difference in OS and RE estimates*

8. Power

(Probability of Coinciding Decisions)



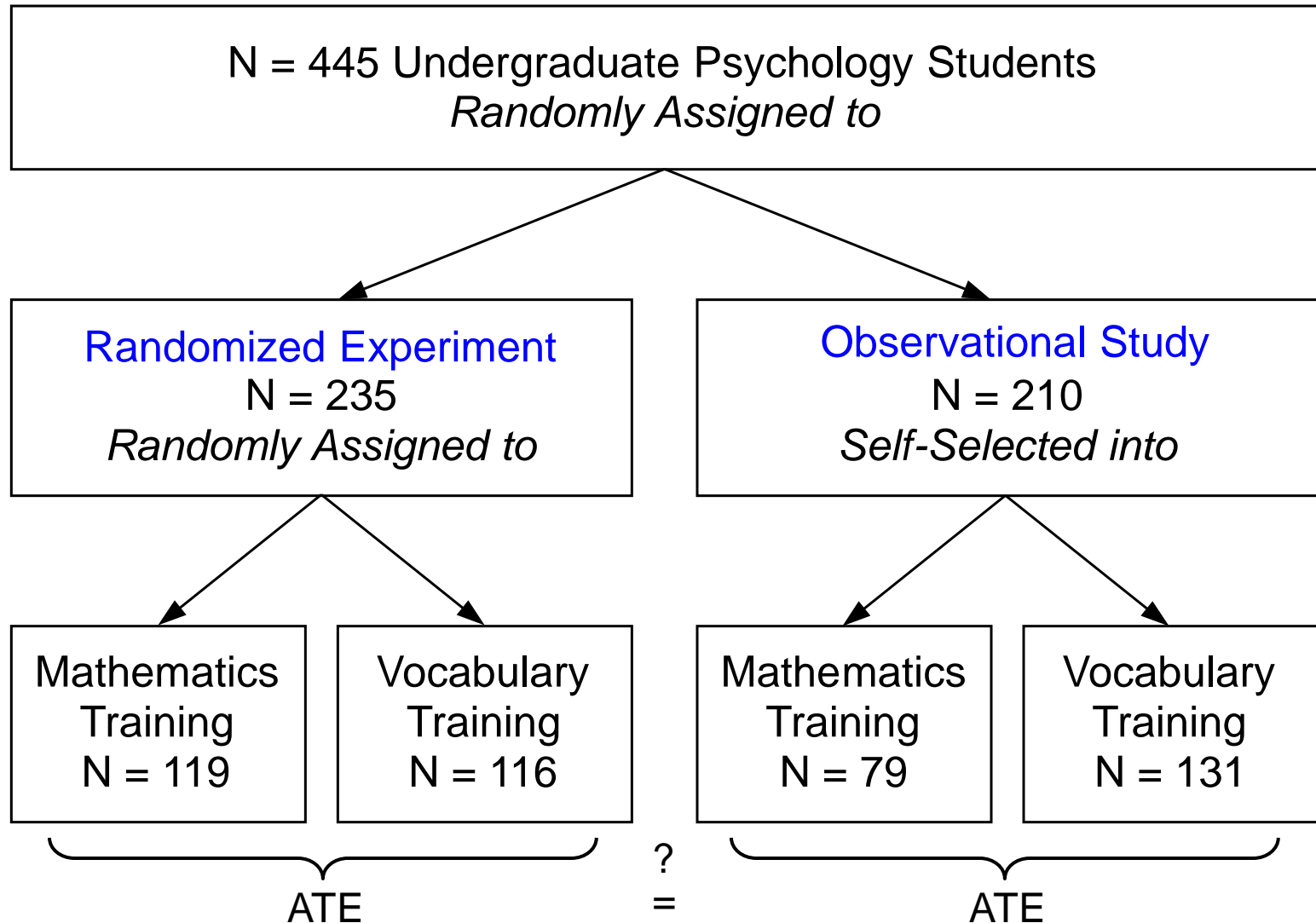
Can OSs replicate RE results?

Do PS techniques and regression adjustments differ?

EXAMPLE I: SHADISH, CLARK & STEINER (2008)

Four-Arm WSC Design

Shadish, Clark & Steiner (2008)



Treatments & Outcomes

- Two treatments and outcomes
 - *Two treatments*: short training either in *Vocabulary* (advanced vocabulary terms) or *Mathematics* (exponential equations)
 - *Two outcomes*: *Vocabulary* (30-item posttest) and *Mathematics* (20-item posttest)
- Treatment effect of interest:
 - ATE for both OS and RE

Covariates

- Extensive measurement of constructs (in the hope that they would establish strong ignorability):
 - 5 construct *domains* with
 - 23 *constructs* based on
 - 156 questionnaire *items*!
- Measured *before* students were randomly assigned to randomized experiment or observational study
 - Hence, measurements are not influenced by assignment or treatment

Construct Domains

23 constructs in 5 domains

- *Demographics* (5 single-item constructs):
Student's age, sex, race (Caucasian, Afro-American, Hispanic), marital status, credit hours
- *Proxy-pretests* (2 multi-item constructs):
36-item Vocabulary Test II, 15-item Arithmetic Aptitude Test
- *Prior academic achievement* (3 multi-item constructs):
High school GPA, current college GPA, ACT college admission score

Construct Domains

- *Topic preference* (6 multi-item constructs):
Liking literature, liking mathematics, preferring mathematics over literature, number of prior mathematics courses, major field of study (math-intensive or not), 25-item mathematics anxiety scale
- *Psychological predisposition* (6 multi-item constructs):
Big five personality factors (50 items on extroversion, emotional stability, agreeableness, openness to experience, conscientiousness), Short Beck Depression Inventory (13 items)

Analysis of RE

- The ATE of the *randomized experiment* was estimated in two different ways
 1. OLS regression (ANOVA) without any covariance adjustment

$$Y_i = \beta_0 + \tau Z_i + \varepsilon_i$$

2. OLS regression with additional covariance adjustment

$$Y_i = \beta_0 + \tau Z_i + \mathbf{X}'_i \boldsymbol{\beta} + \varepsilon_i$$

- Remove random imbalances
- Increase efficiency of treatment effect estimate
- used standard errors from regression analyses

Analysis of OS

1. Estimation of the *propensity score*
 - Logistic model
 - Model selection according to balancing criteria
2. Estimation of *treatment effect* (ATE)
 - PS stratification, inverse-propensity weighting & PS regression adjustment
 - with and without additional covariance adjustment in the outcome model
 - Bootstrapped standard errors

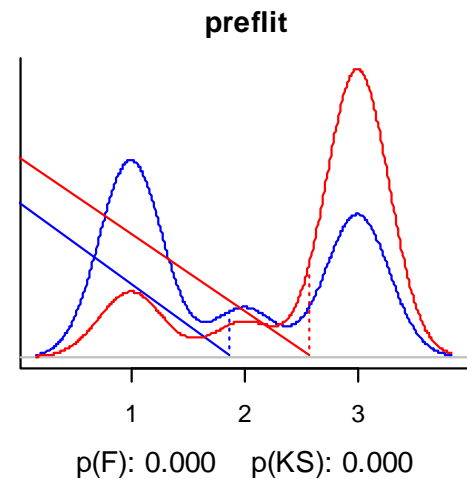
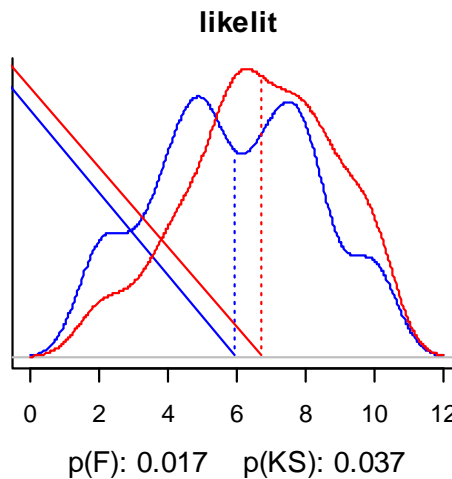
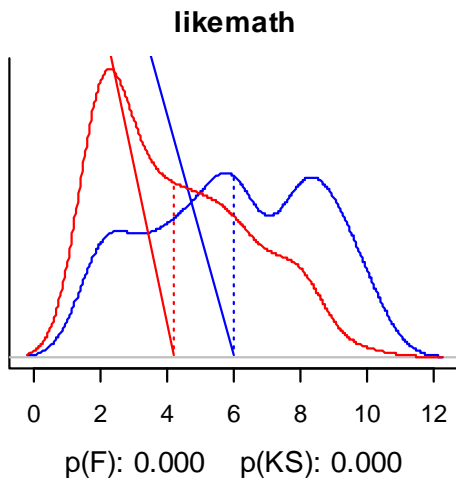
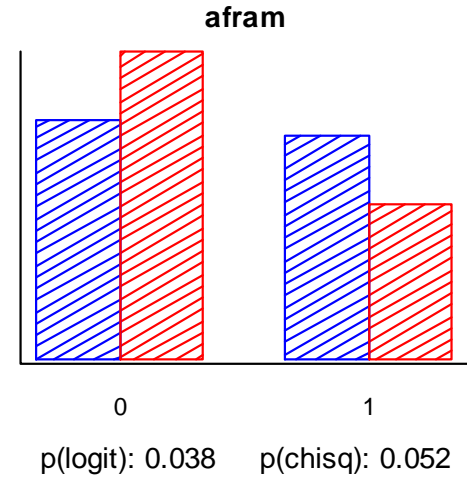
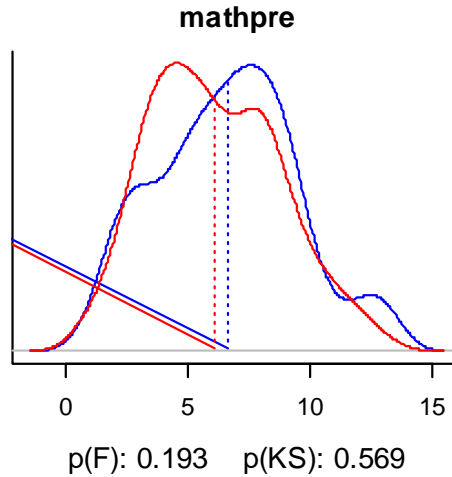
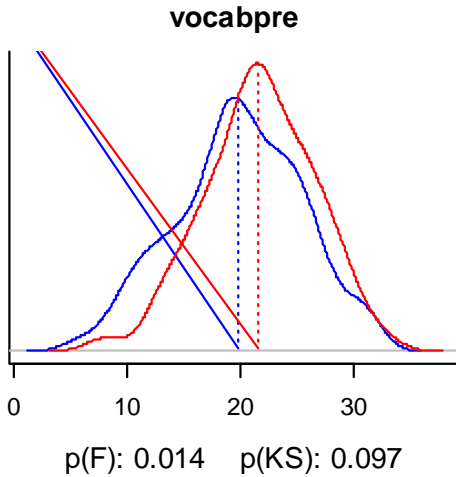
Estimation of PS

- *Logistic regression* of treatment Z on observed predictors \mathbf{W} (covariates \mathbf{X} or transformations thereof, e.g. polynomials, log, interactions)
 - Logistic model: $\text{logit}(Z) \sim \mathbf{W}\gamma$
 - Estimated *PS logit*: $\hat{l} = \mathbf{W}\hat{\gamma}$
 - Estimated *PS*: $\hat{e} = \frac{\exp(\hat{l})}{1 + \exp(\hat{l})}$
- PS-model was specified according to *balancing criteria* (with no outcome data in sight)

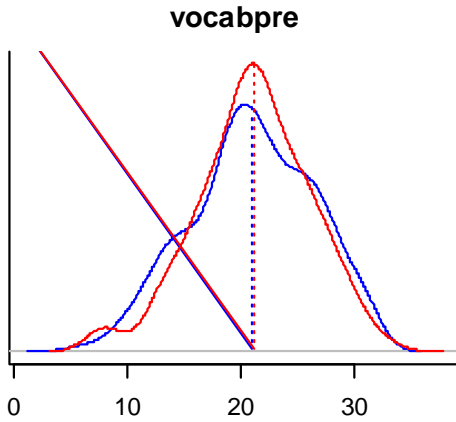
Balance

- *Balance* refers to the equivalence of the treatment and control groups' distribution on all observed covariates
- *Balancing criteria* compare the covariate distribution of the treatment and control group *before* and *after* PS-adjustment:
 - *Visual inspection*: Comparison kernel density estimates
 - *Descriptive statistics*: Standardized mean difference (Cohen's d), variance ratio
 - *Test statistics*: t -test, Kolmogorov-Smirnov test

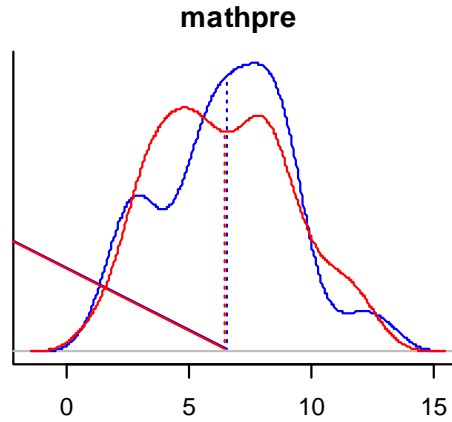
Initial Imbalance before PS adjust.



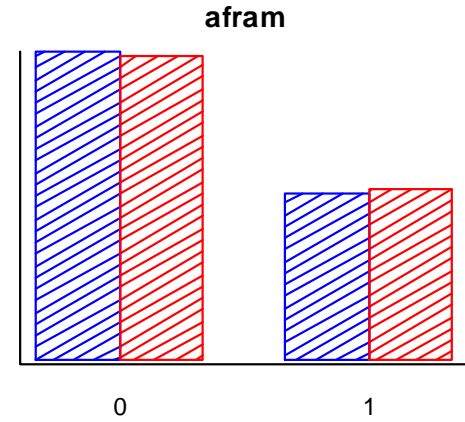
Balance after PS-Adjustment



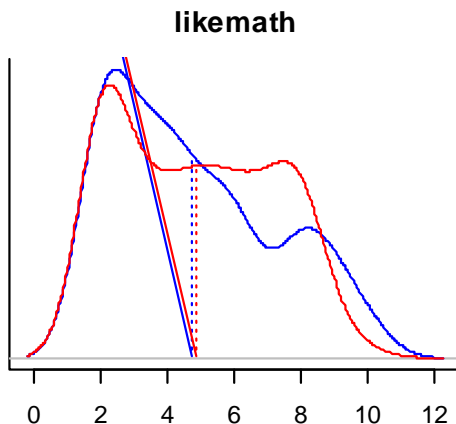
M/IA: 0.970 / 0.984 Mwt: 0.939



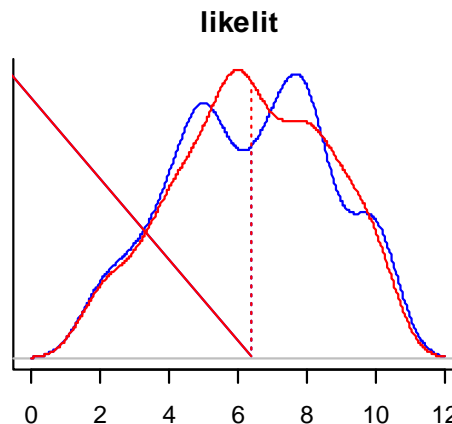
M/IA: 0.887 / 0.595 Mwt: 0.727



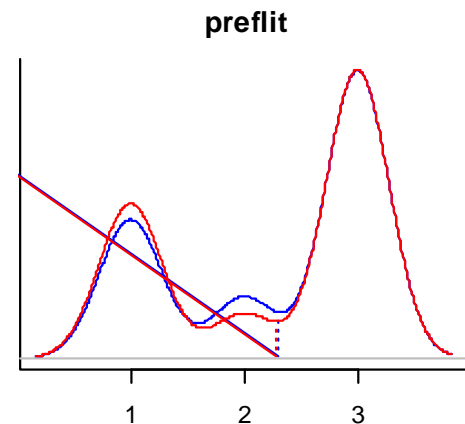
M/IA: 0.915 / 0.150 Mwt: 0.884



M/IA: 0.883 / 0.585 Mwt: 0.829



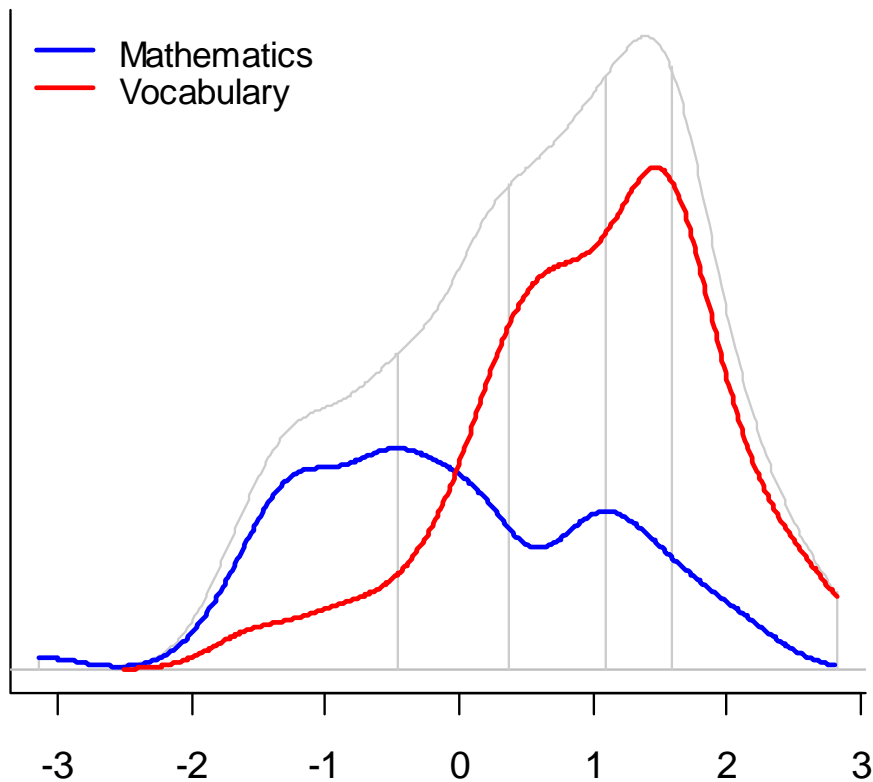
M/IA: 0.855 / 0.202 Mwt: 0.968



M/IA: 0.544 / 0.618 Mwt: 0.805

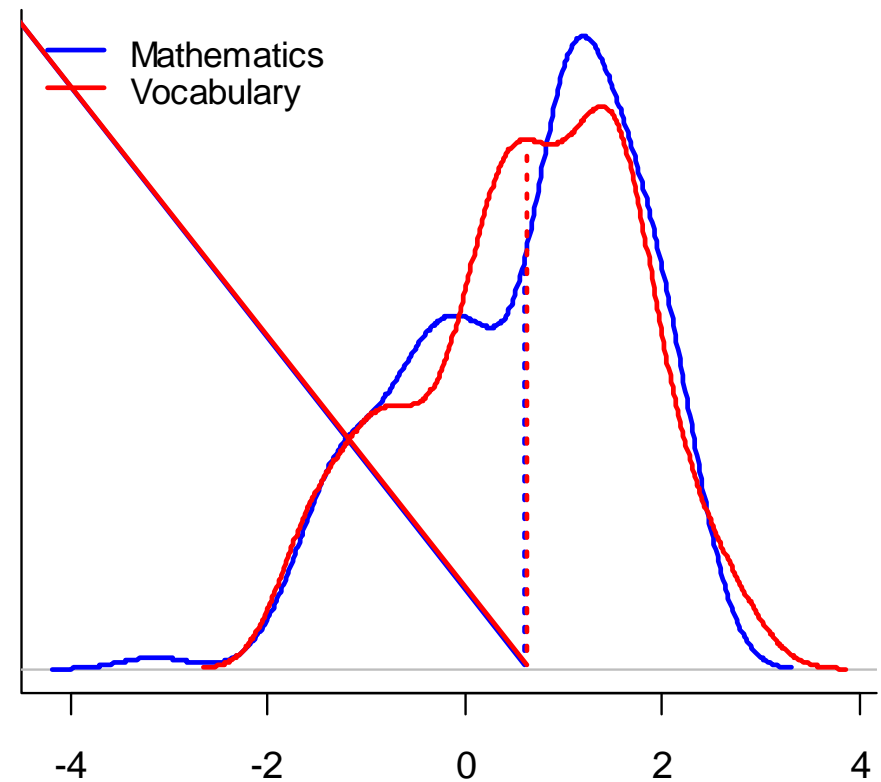
Overlap & Balance (PS-logit)

before PS-adjustment



Propensity Score Logits (by quintiles)

after PS-adjustment



Propensity Score Logits

PS Methods

- PS methods & effect estimators: without and with additional covariance adjustment
 - *PS stratification* on PS with $j = 1, \dots, 5$ quintiles

$$\hat{\tau}_j = \bar{Y}_1^j - \bar{Y}_0^j \qquad \hat{\tau} = \frac{1}{n} \sum_{j=1}^k \hat{\tau}_j n_j$$

or, alternatively, via *marginal mean weights* and *WLS* regression (Hong, 2011)

$$Y_i = \beta_0 + \tau Z_i + \varepsilon_i \qquad \text{without covar. adjustment}$$

$$Y_i = \beta_0 + \tau Z_i + \mathbf{X}'_i \boldsymbol{\beta} + \varepsilon_i \qquad \text{with covar. adjustment}$$

PS Methods

- PS methods & effect estimators (cont'd):
 - *Inverse-propensity weighting* in a WLS regression, with weights $w_i = z_i / \hat{e}_i + (1 - z_i) / (1 - \hat{e}_i)$ obtained from the estimated PS:

$$Y_i = \beta_0 + \tau Z_i + \varepsilon_i$$

$$Y_i = \beta_0 + \tau Z_i + \mathbf{X}'_i \boldsymbol{\beta} + \varepsilon_i$$

- *Covariance adjustment* using PS-logit, cubic polynomial (\mathbf{L})

$$Y_i = \beta_0 + \tau Z_i + \mathbf{L}'_i \boldsymbol{\lambda} + \varepsilon_i$$

$$Y_i = \beta_0 + \tau Z_i + \mathbf{X}'_i \boldsymbol{\beta} + \mathbf{L}'_i \boldsymbol{\lambda} + \varepsilon_i$$

Before We Look at the
Outcome/Results ...

.... review the design
features of the WSC

WSC Design Features

- Initial randomization ensured *comparable target populations*
- Students of both arms, RE and OS, where *treated in exactly the same way*:
 - Shared training sessions
 - Identical procedure of pre- and posttest measurements
- The few *missing data* on independent covariates were imputed for RE and OS simultaneously (single joint imputation model)

WSC Design Features

- *Well-executed RE:*
 - full compliance
 - almost no attrition (3 students)
 - minor degree of missing data
- *Well-implemented OS:*
 - heterogeneous set of covariates, including proxy-pretest measures of the outcome
 - tried to get reliable covariate measurements & all measured before assignment & treatment implementation
 - almost no attrition (3 students)
 - estimated PS balanced distribution of observed covariates
 - bootstrapped standard errors

WSC Design Features

- Identical *causal estimand*: ATE
- Comparable *analytic procedure*: RE and PS analyses with and without additional covariance adjustment (doubly robust)
- Partially *blinded analysis* of RE and OS:
 - PS was estimated without the outcome data in sight
 - PS analyst was not blinded to RE results
 - Overall, could have done a better job in blinding
- Overall, this WSC tried to *rule out most third-variable confounds* in order to assess whether the observed covariates establish a strongly ignorable selection mechanism (→ methodological point of view)

WSC Design Features

- Weaknesses:
 - *Moderate sample sizes* resulting in *moderate power* for comparing RE and OS (increased type II error)
 - *No perfect blinding* of analysts
 - *Laboratory-like setting* (though this might also be considered as a strength—due to better control over third-variable confounds)
 - *No true pretest* of the outcome (only proxy-pretests)
 - Others?

Results

| Mathematics | Mean Difference (standard error) | Absolute Bias (Δ) | Percent Bias Reduction |
|--|--|----------------------------------|------------------------------|
| Covariate-Adjusted Randomized Experiment | 4.01 (.35) | .00 | |
| Unadjusted Quasi-Experiment | 5.01 (.55) | 1.00 | |
| PS Stratification | 3.72 (.57) | .29 | 71% |
| Plus Covariates | 3.74 (.42) | .27 | 73% |
| PS Linear ANCOVA | 3.64 (.46) | .37 | 63% |
| Plus Covariates | 3.65 (.42) | .36 | 64% |
| PS Nonlinear ANCOVA | 3.60 (.44) | .41 | 59% |
| Plus Covariates | 3.67 (.42) | .34 | 66% |
| PS Weighting | 3.67 (.71) | .34 | 66% |
| Plus Covariates | 3.71 (.40) | .30 | 70% |
| ANCOVA Using Observed Covariates | 3.85 (.44) | .16 | 84% |

Results

| Vocabulary | Mean Difference (standard error) | Absolute Bias (Δ) | Percent Bias Reduction |
|--|--|----------------------------------|------------------------------|
| Covariate-Adjusted Randomized Experiment | 8.25 (.37) | | |
| Unadjusted Quasi-Experiment | 9.00 (.51) | .75 | |
| PS Stratification | 8.15 (.62) | .11 | 86% |
| Plus Covariates | 8.11 (.52) | .15 | 80% |
| PS Linear ANCOVA | 8.07 (.49) | .18 | 76% |
| Plus Covariates | 8.07 (.47) | .18 | 76% |
| PS Nonlinear ANCOVA | 8.03 (.50) | .21 | 72% |
| Plus Covariates | 8.03 (.48) | .22 | 70% |
| PS Weighting | 8.22 (.66) | .03 | 96% |
| Plus Covariates | 8.19 (.51) | .07 | 91% |
| ANCOVA Using Observed Covariates | 8.21 (.43) | .05 | 94% |

Comparison of OS and RE

- Hypothesis tests based on OS and RE result in *same conclusion*
 - in all analyses the null hypothesis (of a zero treatment effect) is rejected
- *No significant differences* between OS and RE but also between the different methods
 - but this is mostly due to the moderate initial bias and the insufficient power

Which covariates effectively remove selection bias?

Covariate selection strategy: heterogeneity of covariate measurements

Do PS techniques and regression adjustments differ?

**EXAMPLE II: STEINER ET AL. (2010)
& STEINER, COOK & LI (IN PREP)**

Importance of Covariate Selection

Steiner, Cook, Shadish & Clark (2010)

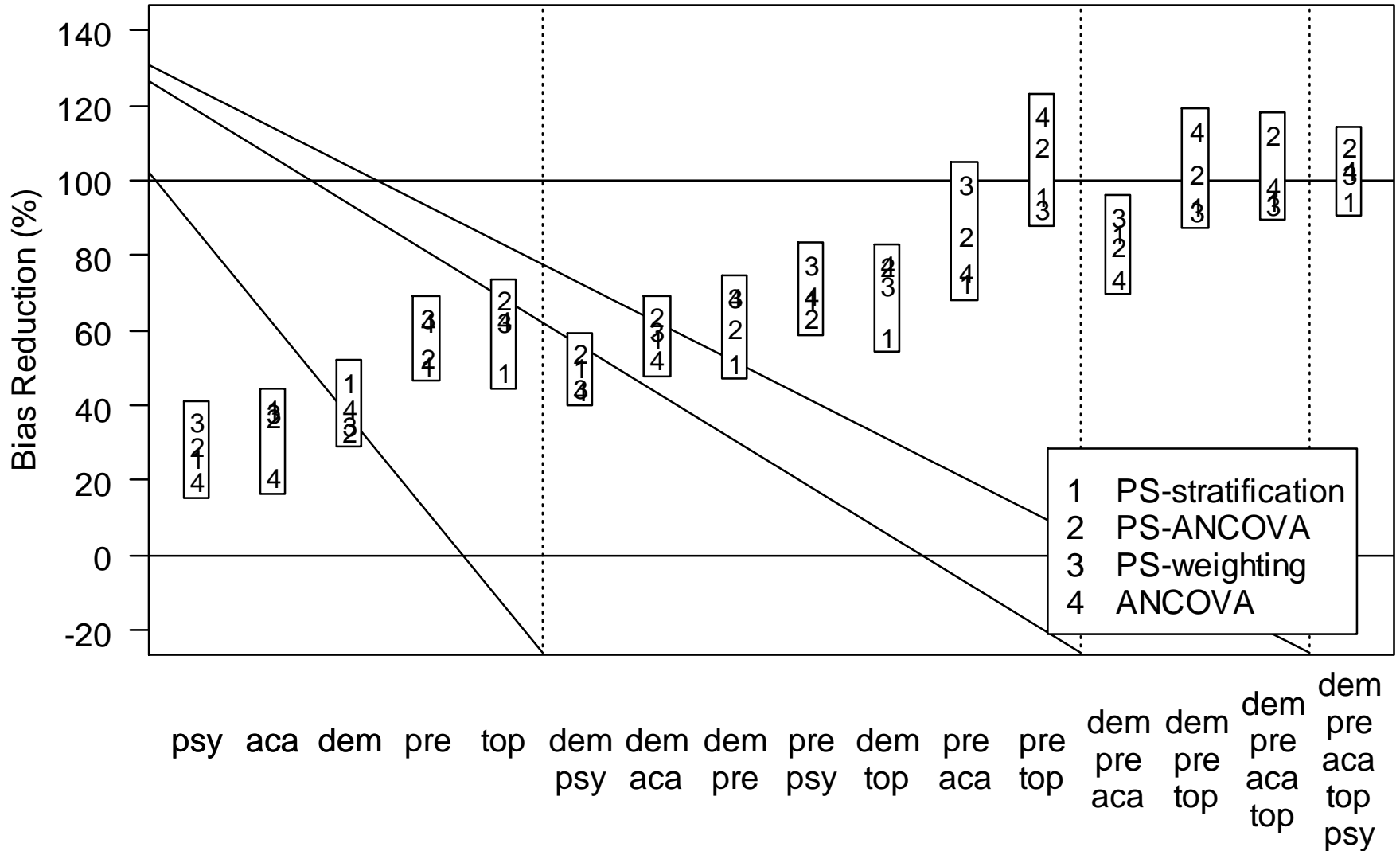
- Investigate the importance of
 - *Construct domains*
 - *Single constructs*for establishing strong ignorability
- Compare it to the importance of choosing a specific *analytic method*

Importance of Covariate Selection

Steiner, Cook, Shadish & Clark (2010)

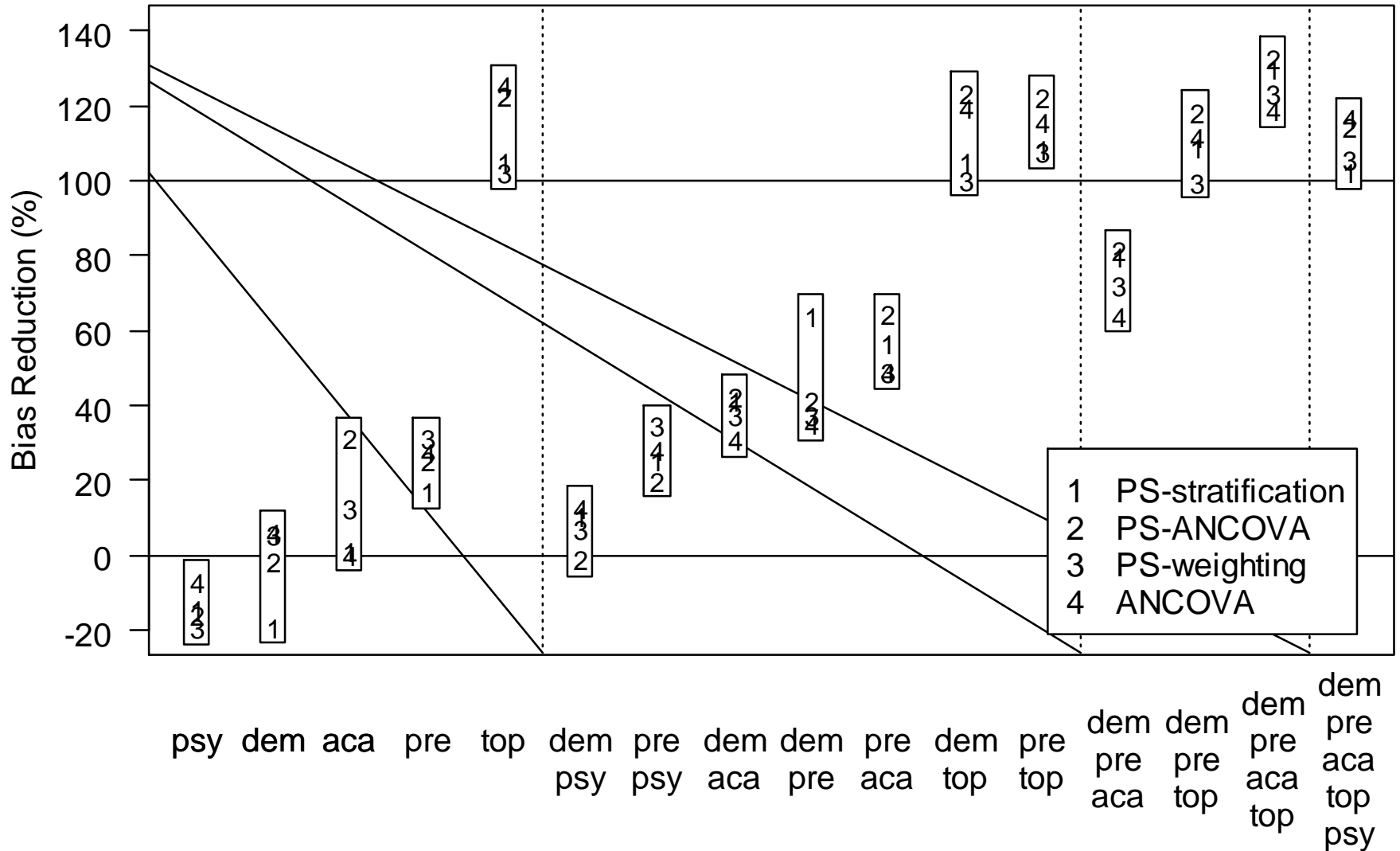
- Investigated *covariate domains* by selecting
 - Single domains
 - Combinations of 2, 3 and 4 domains
- Estimated *PS model* for each combination of domains separately
(used only selected domains for checking balance)
- Estimated treatment effect and remaining bias
- Did the same with single covariates or selected sets of covariates

Bias Reduction: Construct Domains Vocabulary

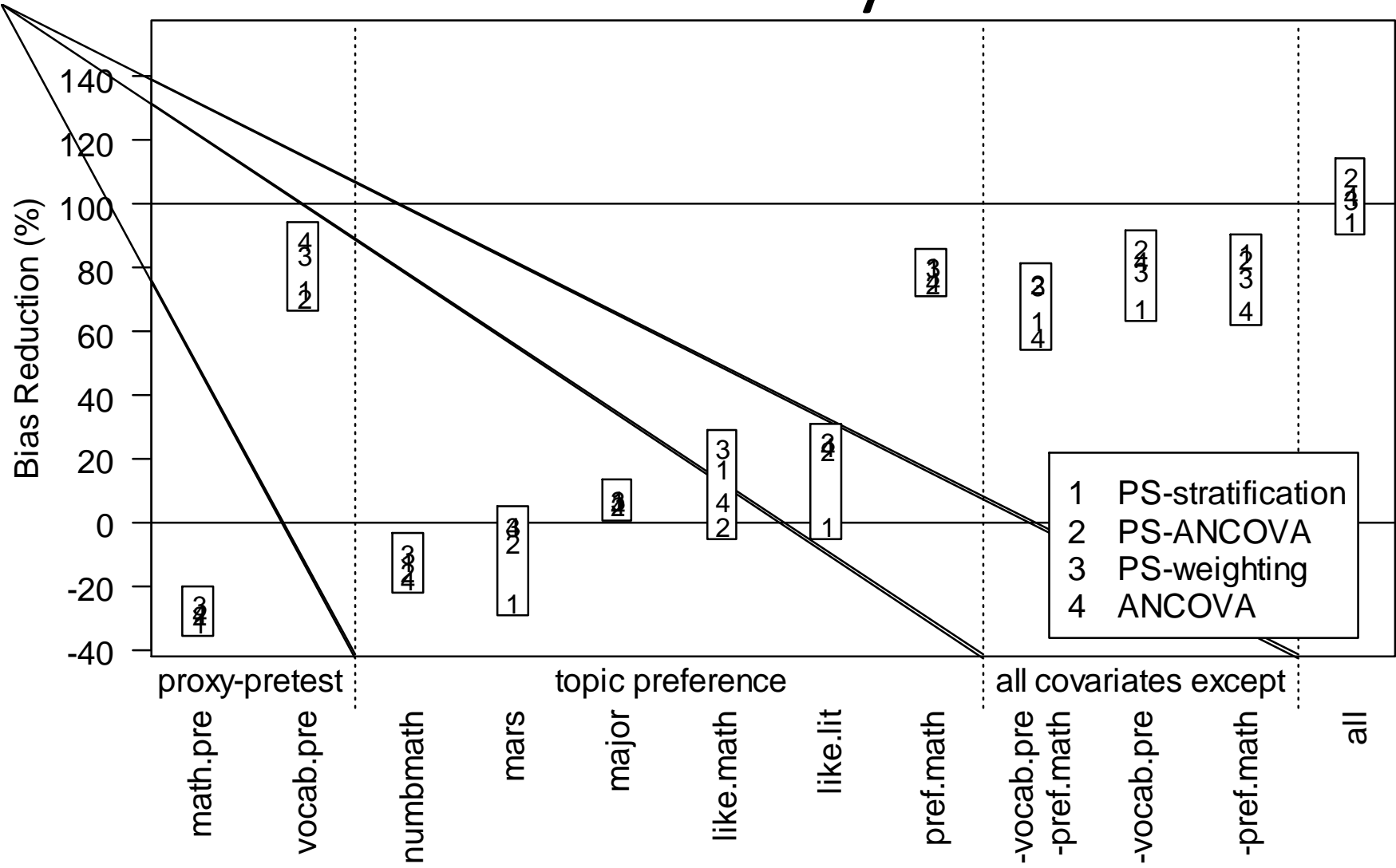


Bias Reduction: Construct Domains

Mathematics

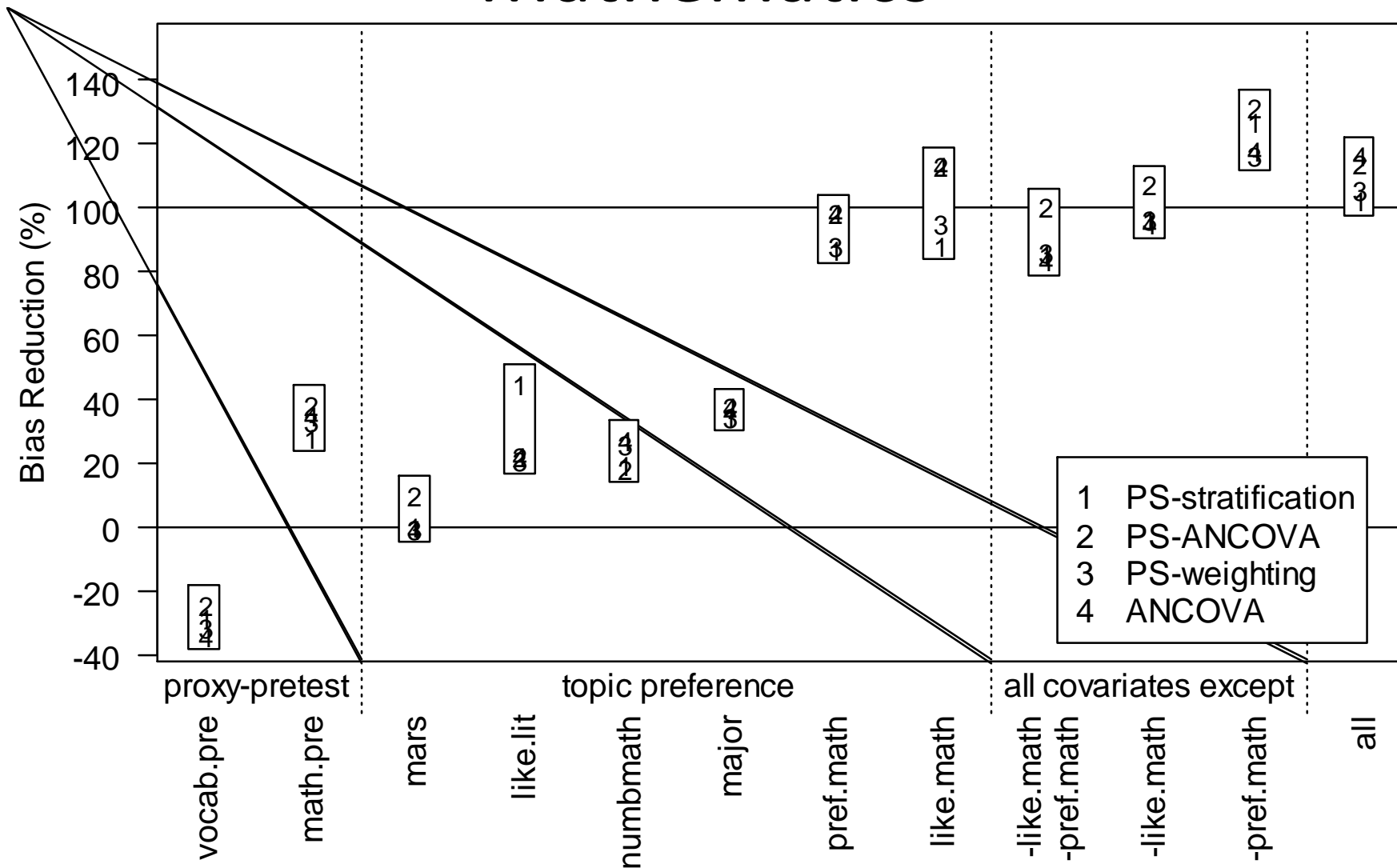


Bias Reduction: Single Constructs Vocabulary



Bias Reduction: Single Constructs

Mathematics



Constructs: Conclusion

- In establishing SI, selection of constructs matters
 - Need those *construct domains* that effectively reduce bias (those related to both treatment selection and outcome)
 - Need the right *single constructs* within domains because only a few covariates successfully reduce bias
- Choice of analytic method is less important (given its competent implementation)
 - No systematic difference between PS methods
 - ANCOVA did as well (at least in that case)

Number & Heterogeneity of Covariates

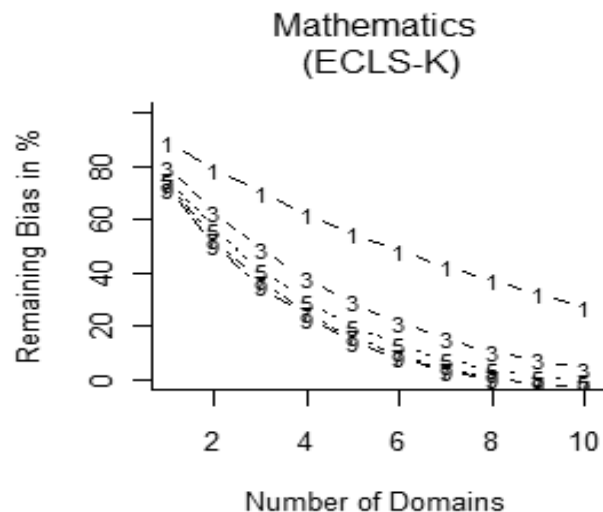
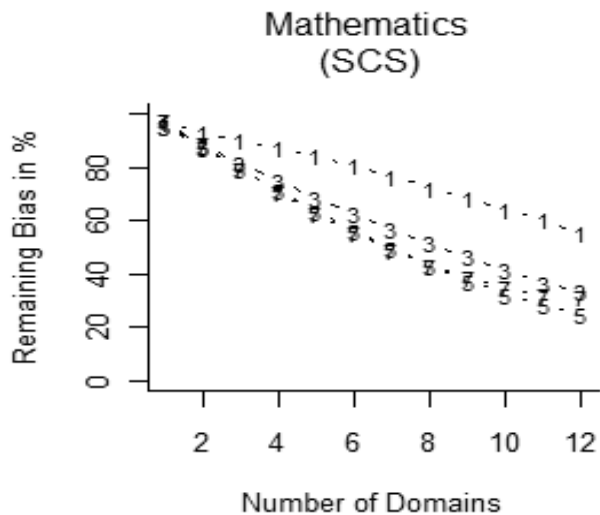
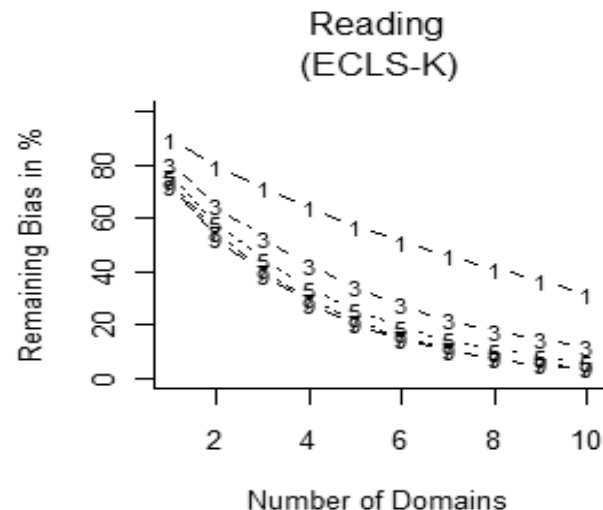
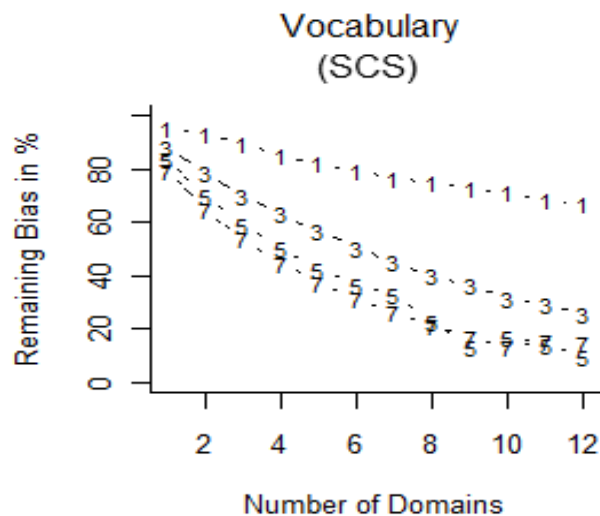
Steiner, Cook & Li (in prep)

- How important is the *number* and *heterogeneity* of covariates for removing bias?
 - Heterogeneity: measurements from different construct domains (e.g., pretests, demographics)
 - Number of covariates within construct domains
- Simulate the (worst case) scenario of having *no substantive theory* that could guide us in selecting covariates. Investigate two different scenarios:
 1. Randomly sample construct domains and covariates within domains
 2. Same as 1. but assume that the most effective covariates for reducing bias (pretests & motivational factors) were not measured

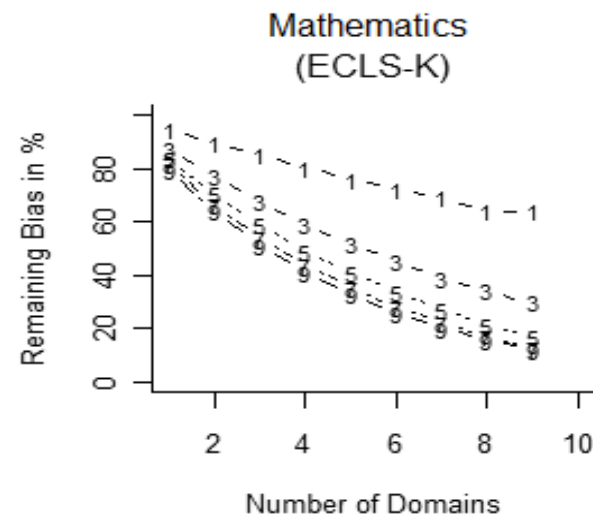
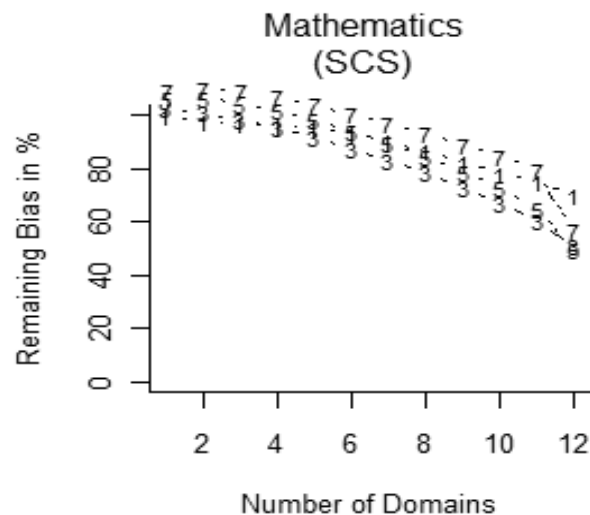
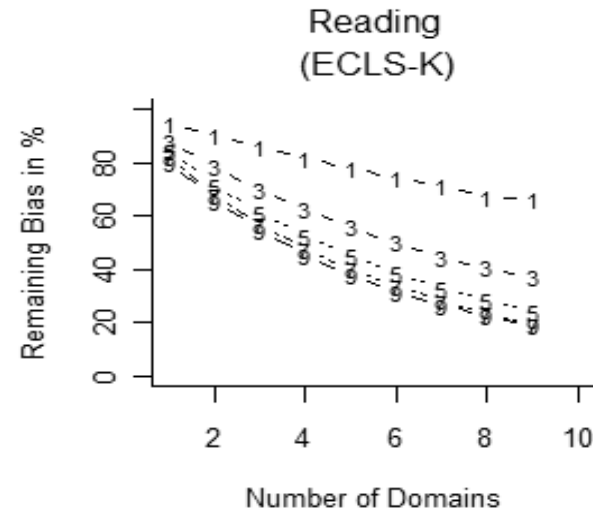
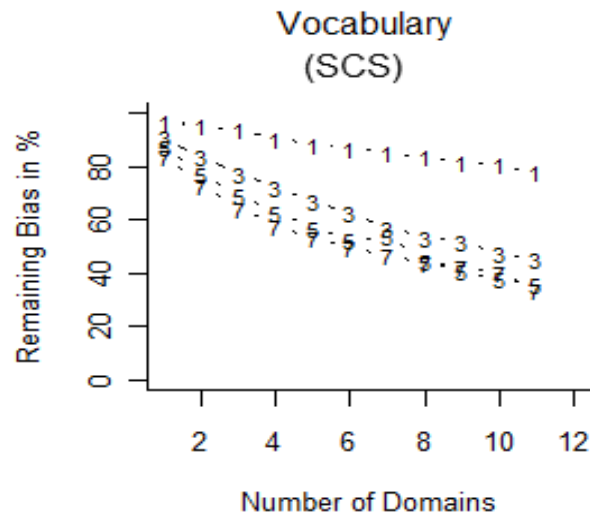
Simulation Design

- Use the data on the item level (156 items)
- Defined 12 construct domains:
 1. Demographics (5 items)
 2. Vocabulary proxy pretest (36 items)
 3. Math proxy pretest (15 items)
 4. Prior academic achievement (3 items)
 5. Mathematics anxiety test (25 items)
 6. Topic preference (10 items)
 7. Short Beck Depression Inventory (13 items)
 8. Big Five Personality Factors: extroversion (10 items)
 9. Big Five Personality Factors: emotional stability (10 items)
 10. Big Five Personality Factors: agreeability (10 items)
 11. Big Five Personality Factors: openness to experience (10 items)
 12. Big Five Personality Factors: conscientiousness (10 items)

Average remaining bias across random samples (all covariates)



Average remaining bias across random samples (without effective covariates)



Vocabulary

| | | All Samples | | | | Samples Without Critical Covariates | | | | |
|--|-----------------------|-------------------------------------|-----------------|-----------------|-----------------|-------------------------------------|-----------------|-----------------|-----------------|-----------------|
| | | Number of Covariates Per Domain (n) | | | | Number of Covariates Per Domain (n) | | | | |
| | | 1 | 3 | 5 | 7 | 1 | 3 | 5 | 7 | |
| Average results & spread (remaining bias in %): [10%, 90%] quantiles of 300 random samples of covariates | Number of Domains (m) | 1 | 96 [82, 109] | 88 [64, 113] | 84 [60, 113] | 80 [49, 111] | 97 [85, 110] | 91 [70, 114] | 87 [65, 115] | 84 [60, 114] |
| | | 2 | 93 [70, 112] | 79 [46, 110] | 70 [36, 105] | 65 [30, 102] | 95 [76, 113] | 84 [56, 113] | 77 [49, 110] | 73 [41, 108] |
| | | 3 | 90 [62, 113] | 70 [34, 106] | 59 [22, 96] | 54 [15, 92] | 94 [70, 115] | 77 [48, 110] | 69 [38, 102] | 64 [31, 99] |
| | | 4 | 85 [52, 113] | 63 [24, 100] | 50 [10, 88] | 45 [3, 85] | 91 [63, 115] | 72 [41, 105] | 63 [30, 96] | 58 [23, 94] |
| | | 5 | 82 [44, 113] | 57 [16, 94] | 42 [2, 82] | 37 [-6, 79] | 89 [60, 115] | 67 [36, 100] | 57 [24, 91] | 53 [17, 90] |
| | | 6 | 79 [39, 112] | 51 [9, 89] | 36 [-5, 77] | 31 [-13, 75] | 87 [57, 115] | 62 [31, 95] | 52 [19, 87] | 50 [13, 87] |
| | | 7 | 77 [36, 112] | 45 [3, 84] | 33 [-12, 73] | 26 [-20, 71] | 85 [54, 116] | 57 [27, 90] | 54 [15, 83] | 47 [10, 84] |
| | | 8 | 75 [33, 111] | 40 [-2, 80] | 23 [-17, 69] | 21 [-26, 67] | 84 [51, 115] | 54 [24, 86] | 45 [11, 80] | 44 [7, 81] |
| | | 9 | 73 [31, 110] | 36 [-6, 75] | 13 [-22, 66] | 17 [-31, 65] | 82 [49, 114] | 52 [22, 82] | 41 [8, 76] | 43 [6, 78] |
| | | 10 | 71 [29, 109] | 32 [-9, 71] | 17 [-25, 62] | 13 [-34, 61] | 81 [46, 114] | 48 [18, 77] | 38 [6, 71] | 40 [5, 74] |
| | | 11 | 68 [26, 106] | 29 [-13, 69] | 14 [-32, 59] | 15 [-36, 59] | 78 [42, 112] | 45 [16, 75] | 36 [4, 67] | 34 [-4, 71] |
| | | 12 | 67 [22, 104] | 26 [-18, 67] | 9 [-25, 54] | 15 [-37, 62] | | | | |

Covariate Selection: Conclusions

- A *heterogeneous* set of *numerous* covariates removes most of the selection bias—at least on average
- Heterogeneity is more important than having many covariates within construct domains
- Even if the covariates that most effectively remove selection bias are not measured, a large set of heterogeneous covariates may remove a major part selection bias
- Results only from one WSC → *need more WSCs*
- Investigate other *covariate selection strategies* (e.g., selection based on causal structural models)

WSC-based simulation study on measurement error

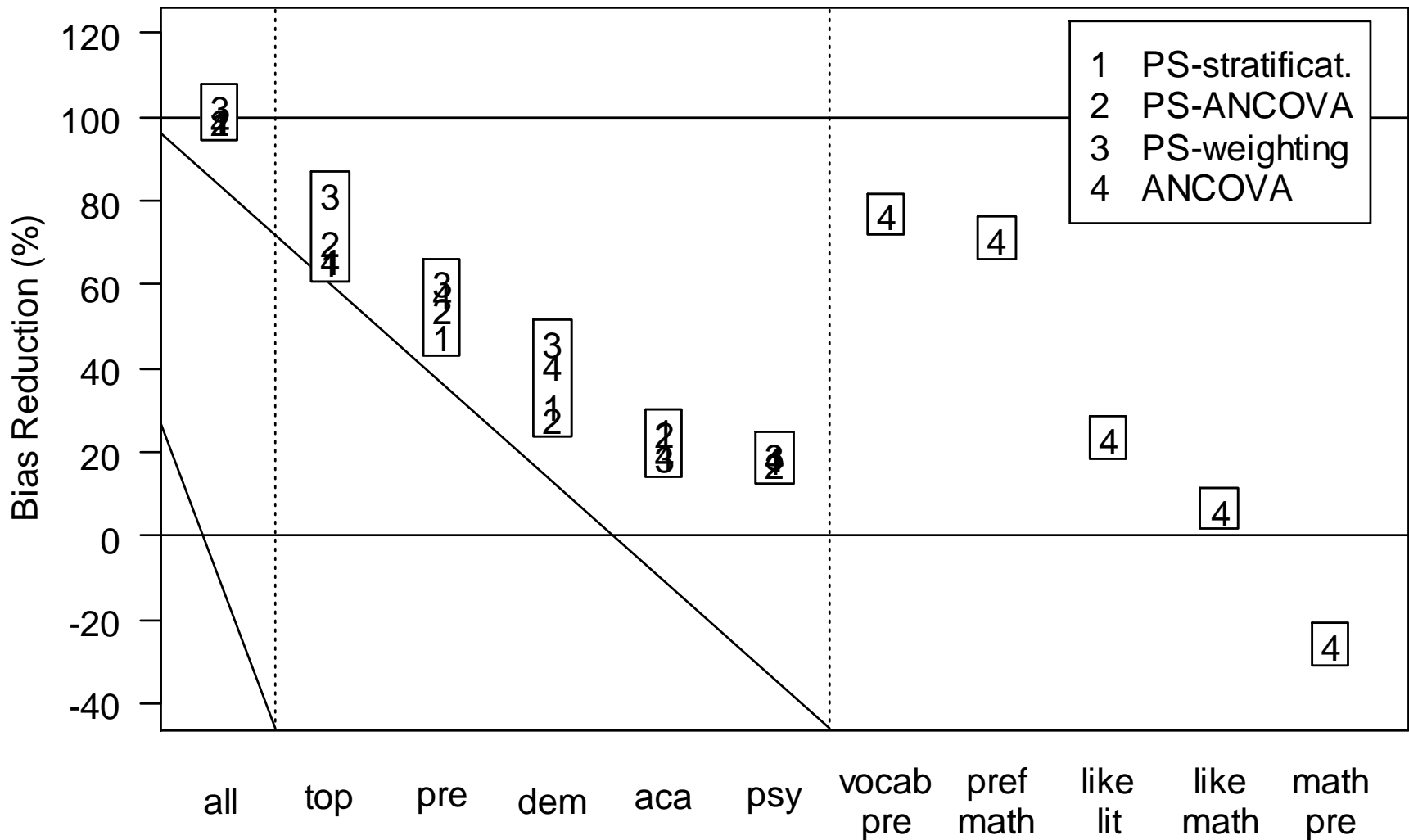
EXAMPLE III: STEINER, COOK & SHADISH (2011)

Reliability of Construct Measurement

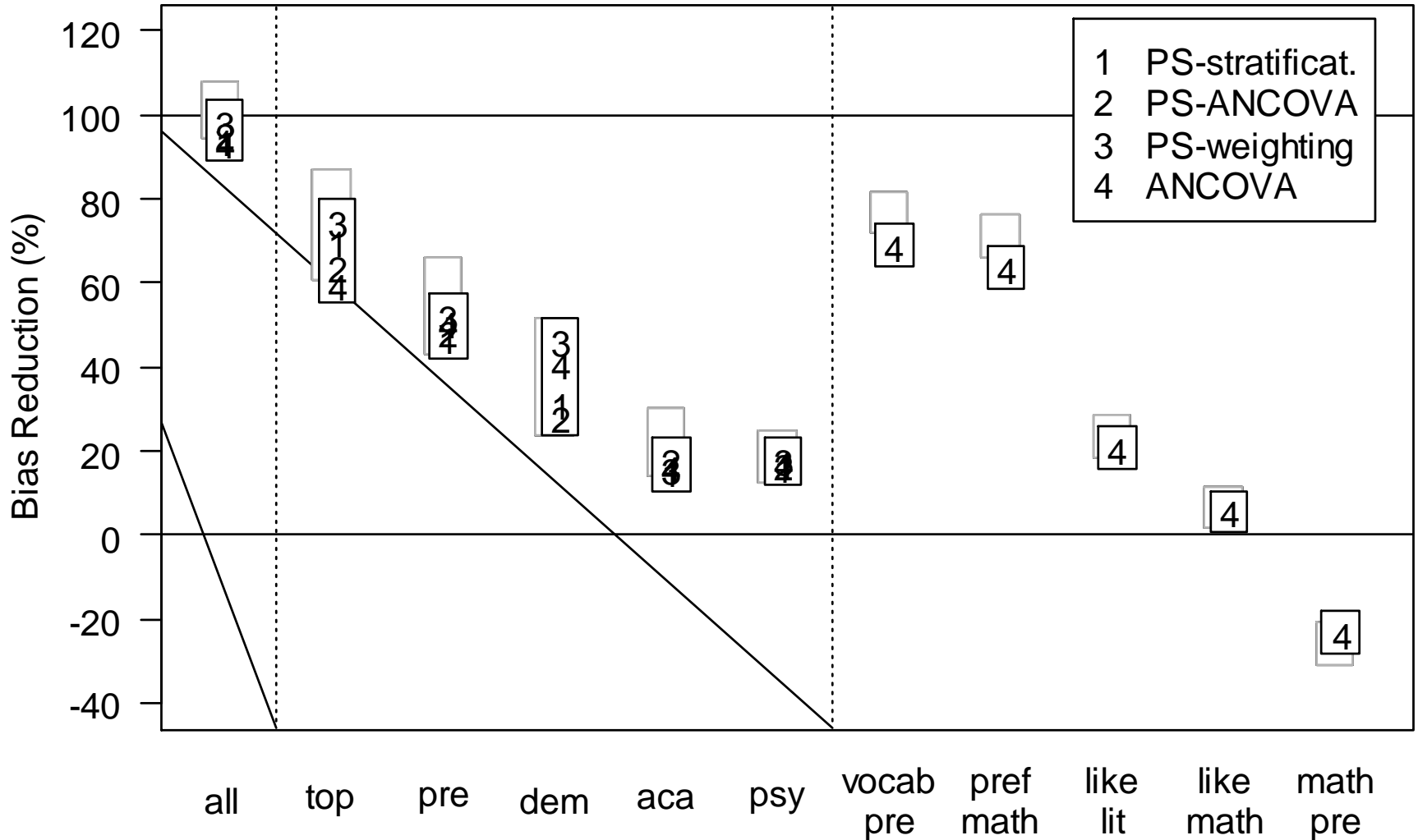
Steiner, Cook & Shadish (2011)

- How important is the *reliable measurement* of constructs (*given selection on latent constructs*)?
 - Does the inclusion of a large set of covariates in the PS model compensate for each covariate's unreliable measurement?
- Add measurement error to the observed covariates in a *simulation study*
 - Assume that original set of covariates is measured without error and removes 100% of selection bias
 - Systematically added measurement error such that the reliability of each covariate was $\rho = .5, .6, .7, .8, .9, 1.0$

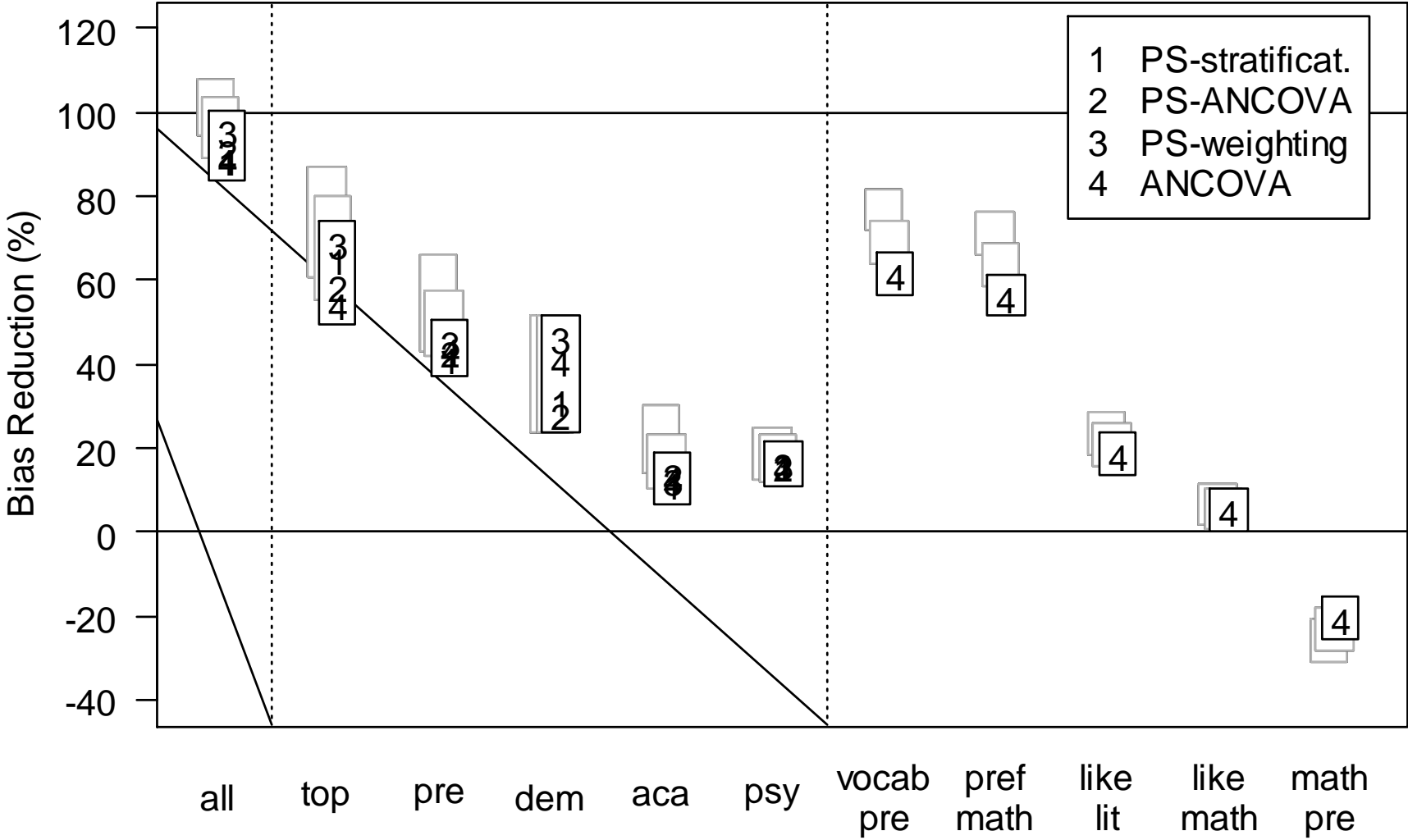
Vocabulary: Reliability 1.0



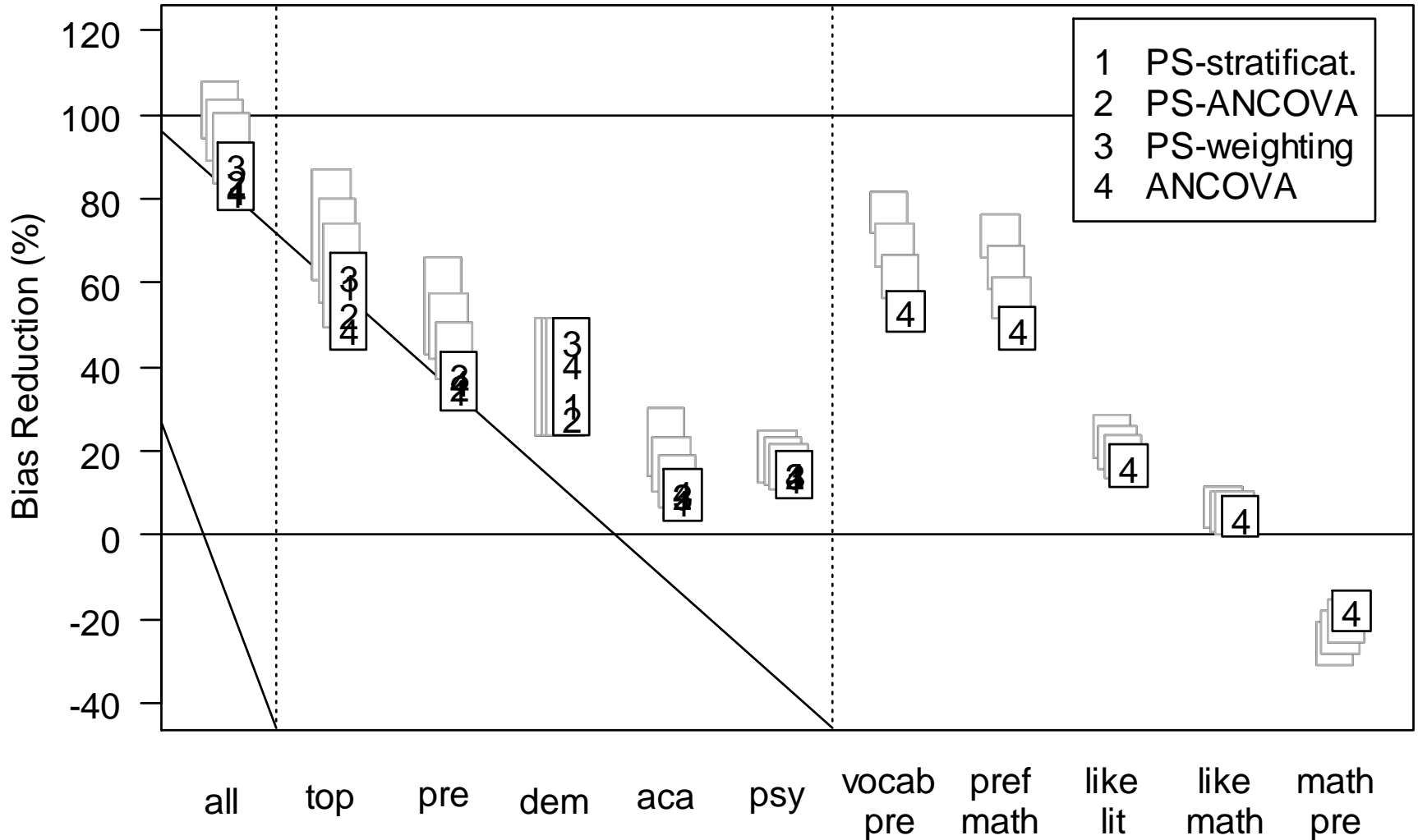
Vocabulary: Reliability .9



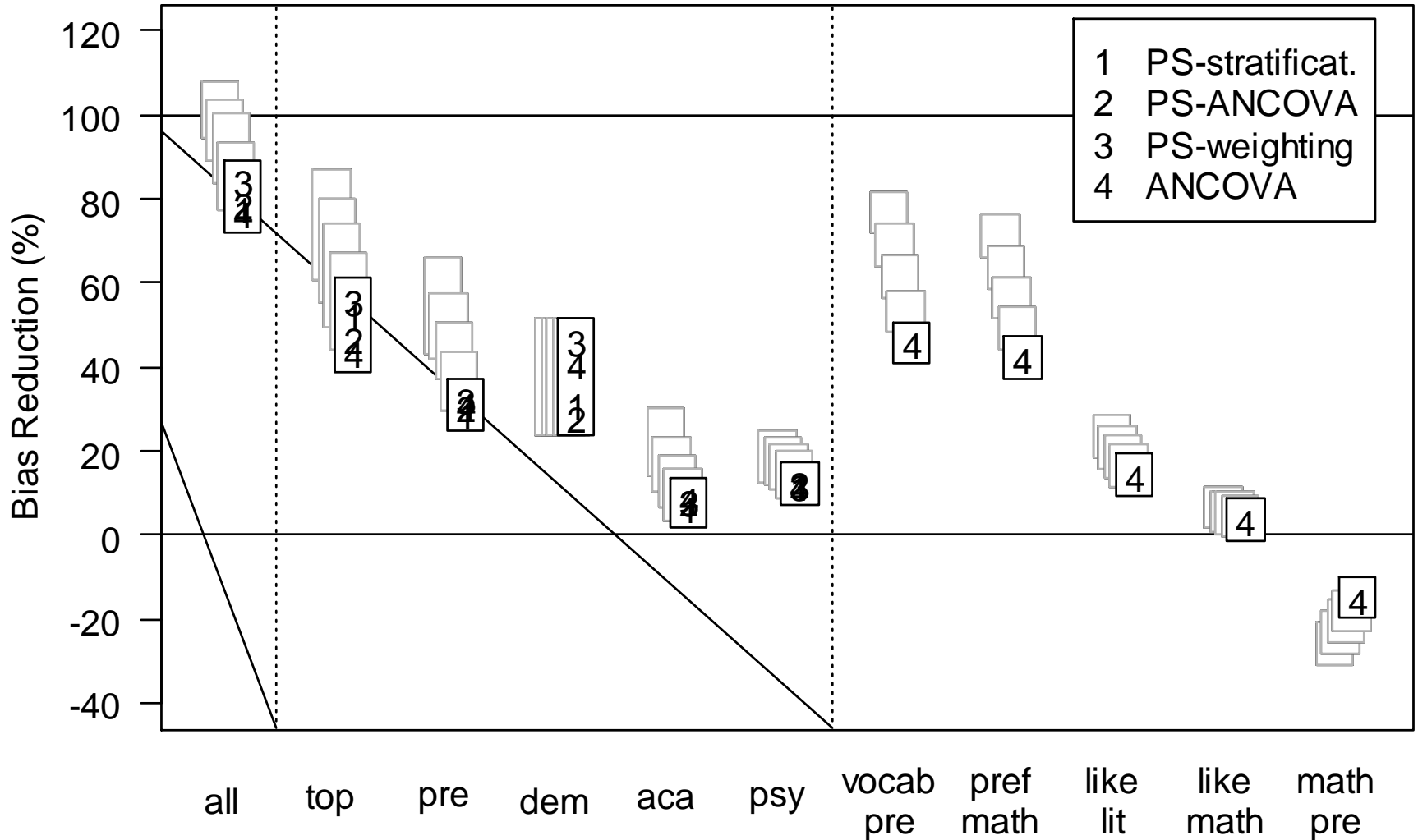
Vocabulary: Reliability .8



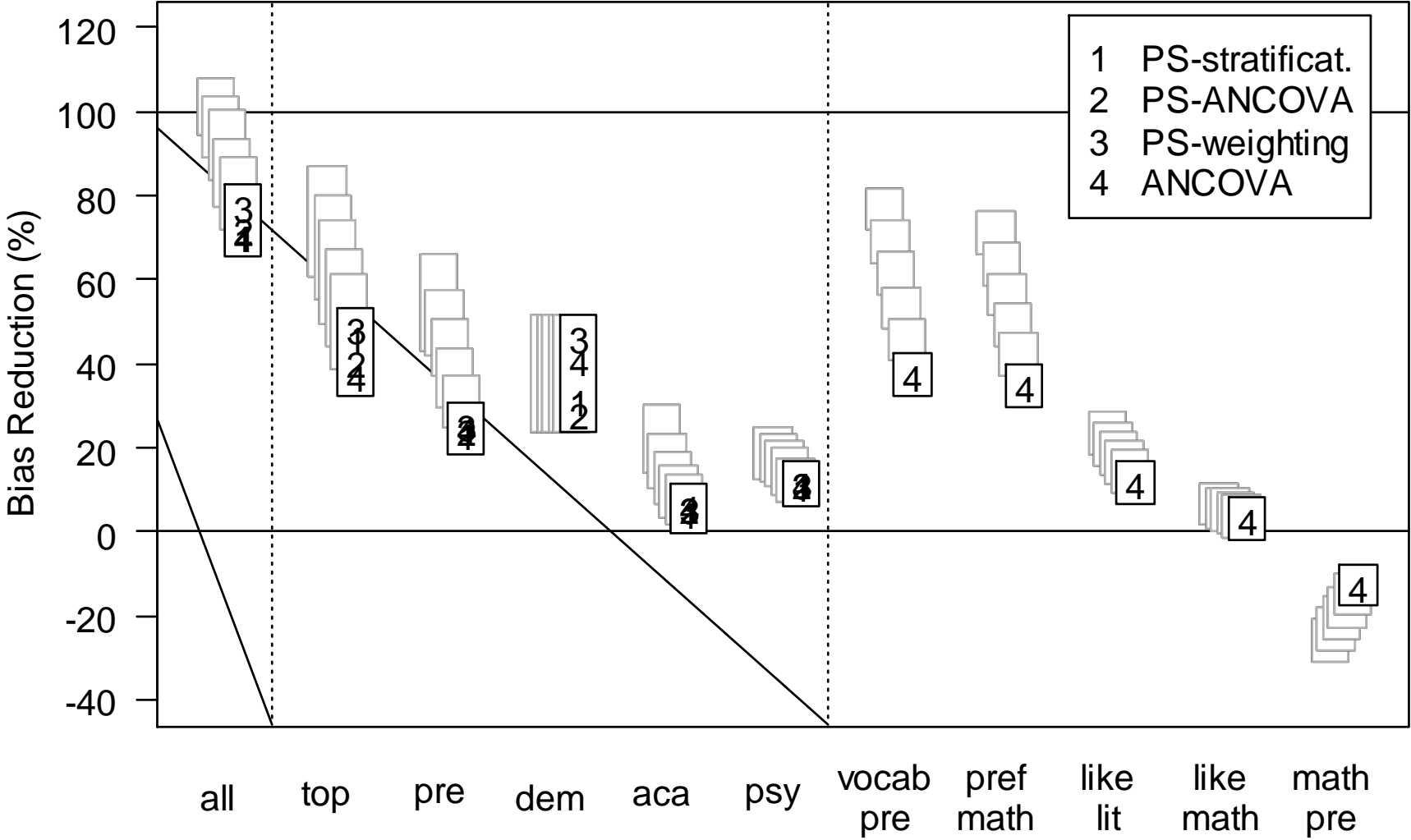
Vocabulary: Reliability .7



Vocabulary: Reliability .6



Vocabulary: Reliability .5



Reliability: Conclusions

- If selection is on latent constructs
 - Measurement error *attenuates a covariate's potential* for reducing selection bias
 - The measurement of a large set of *interrelated covariates* compensates for unreliability in each covariate—but does so only partially
 - The reliability of *effective covariates* matters. Measurement error in ineffective covariates has almost no influence on bias reduction.
 - Choice of *analytic method* is less important (no systematic difference between methods)

Discussion example

EXAMPLE IV: PEIKES, MORENO & ORZOL (2008)

Peikes, Moreno & Orzol (2008)

“Propensity Score Matching: A Note of Caution for Evaluators of Social Programs” (The American Statistician, 62(3))

- WSC design (three-arm design)
- Design and Analysis of RE
- Design and Analysis of OS
- Results & author’s conclusion
- Discussion of the WSC’s strengths
- Discussion of open questions

WSC Design (Three-Arm WSC)

- *Three-arm design* with *multiple sites*
(New York, New Hampshire, Oklahoma)
- Evaluation of State Partnership Initiative:
 - promote employment among beneficiaries with disabilities who receive Supplemental Security Income (SSI) or Social Security Disability Insurance (SSDI) benefits
 - Outcomes of interest: employment rates and earnings

WSC Design (Three-Arm WSC)

- Comparison of *PS matching* estimates of the three sites to corresponding RE estimates
- Evaluate whether PS and RE estimates
 - differ in magnitude and sign
 - result in different conclusions about the impact of the program

Randomized Experiments

| Sample sizes: | CTRL | TREAT | TREAT | Compliance |
|--|------|---------|----------|------------|
| | | assign. | received | . |
| • New York (two subgroups) | | | | |
| – A | 914 | 937 | 277 | 30% |
| – B | 914 | 932 | 301 | 32% |
| • New Hampshire (two subgroups) | | | | |
| – A | 27 | 22 | 22 | 100% |
| – B | 29 | 35 | 35 | 100% |
| • Oklahoma | 256 | 1440 | 314 | 22% |
| • One-sided <i>noncompliance</i> (no-shows)! | | | | |
| • Rather small sample sizes for New Hampshire! | | | | |

Randomized Experiments

- Causal estimand:
 - *Intent-to-treat* (ITT) divided by participation rate
= complier (local) average treatment effect (IV-CATE)
= *ATT* (due to one-sided compliance)
- Estimator:
 - *Difference-in-differences* estimator
 - Plus additional *regression adjustment*
- Outcomes:
 - change in the proportion employed
 - change in earnings (two versions)

Observational Studies

- Selection of eligible comparison group members:
 1. Identified pools of potential comparison group members in areas where the demonstration was not offered (*non-demonstration areas*)
 - Identification of comparison counties with similar economic and service environments as in demonstration counties
 2. Selected comparison group member that meet same *eligibility criteria* as for RE

Observational Studies

- PS was estimated for each site/subgroup using a *logit model* with binary outcome
 - 1 ... for participants (i.e., compliers with treatment assignment)
 - 0 ... for comparison group members
- More than *250 covariates* were considered (with a stepwise model selection procedure)
- *Balance checks* within strata (Dehejia & Wahba, 1999)
 - *t*-test of differences in treatment and comparison group means within each stratum (for all the 250 covariates)
 - 95% of the *t*-test indicated no covariance difference after matching

Observational Studies

- Causal estimand:
 - *ATT* (since participants that actually received treatment were used for estimating the PS model)
- Estimator:
 - *Matching* estimator:
One-to-one nearest neighbor matching (with replacement) on estimated PS
 - Combined with a *difference-in-differences* estimator
 - Plus additional *regression adjustment*
- Outcomes:
 - change in the proportion employed
 - change in earnings (two versions)

Observational Studies

| Sample sizes: | RE | | OS | |
|---------------------------------|----------------|------------------|-------------------|-------------------|
| | CTRL | TREAT assign. | TREAT Received | COMPAR. GROUP. |
| <hr/> | | | | |
| • New York (two subgroups) | | | | |
| – A | 914 | 937 | 277 | 253 |
| – B | 914 | 932 | 301 | 271 |
| • New Hampshire (two subgroups) | | | | |
| – A | 27 | 22 | 22 | 19 |
| – B | 29 | 35 | 35 | 34 |
| • Oklahoma | 256 | 1440 | 314 | 244 |
| Estimand | IV-CATE (=ATT) | | | ATT |

Selection of Results (I)

Change in proportion employed

| | PSM | | Random assignment | |
|---|---------|-----------------|-------------------|-----------------|
| | Impact | <i>p</i> -value | Impact | <i>p</i> -value |
| Change in proportion employed in the year after the randomization year relative to the year before (percentage points) | | | | |
| New York—SSI/concurrent | | | | |
| Benefits counseling and waivers | 14.3*** | 0.001 | 8.8 | 0.186 |
| Benefits counseling, waivers, and employment services | 25.9*** | <0.001 | 17.0*** | 0.008 |
| New Hampshire (<i>note small sample sizes</i>) | | | | |
| SSI/concurrent | −4.1 | 0.818 | −29.5* | 0.065 |
| SSDI only | 16.7 | 0.249 | −29.6** | 0.018 |
| Oklahoma | | | | |
| SSI/concurrent | 10.6*** | 0.005 | 17.0 | 0.152 |

Selection of Results (II)

Change in earnings

| | PSM | | Random assignment | |
|--|----------|-----------------|-------------------|-----------------|
| | Impact | <i>p</i> -value | Impact | <i>p</i> -value |
| Change in earnings in the year after the randomization year relative to the year before (dollars) | | | | |
| New York—SSI/concurrent | | | | |
| Benefits counseling and waivers | 1,214*** | <0.001 | −1,080* | 0.059 |
| Benefits counseling, waivers, and employment services | 1,209*** | 0.002 | −455 | 0.401 |
| New Hampshire | | | | |
| SSI/concurrent | 3,942*** | 0.009 | −709 | 0.511 |
| SSDI only | 339 | 0.694 | −1,633** | 0.045 |
| Oklahoma | | | | |
| SSI/concurrent | −75 | 0.715 | 451 | 0.448 |

Authors' Conclusion

- PS matching produced impact estimates that *differed* from the RE benchmark estimates—in magnitude and sign
- Compared to the RE, PS matching would have led to *different conclusions* about the program's impact
- Speculate that not all confounders were measured

Strengths of WSC

- *Comprehensive administrative data*: more than 250 variables
 - Key predictors of both program participation and outcomes (demographics, disability type, participation in SSI/SSDI/Medicaid, use of work incentives)
 - Multiple waves (5 years) of pretest measures on outcomes (employment and earnings) from Social Security and Internal Revenue Service (IRS) federal income tax records
- Two sites with at least moderately *large sample sizes*
- Large pool of potential comparison group members
 - Comparison/treatment group ratios from 8:1 to over 90:1
- Comparable *causal estimands*
- Combination with *difference-in-differences* approach
- *Replication* across sites & subgroups

Open Questions

(possible reasons for RE-OS difference)

- Standard errors?
 - IV-CATE requires a two-stage least squares estimation (2SLS) for getting appropriate standard errors
 - Peikes et al. used the ITT standard errors → underestimates variability of IV-CATE estimates
- Power?
 - For demonstrating an effect for RE and OS
 - For comparing RE and OS
 - Some small sample sizes
- Balance?
 - Within-stratum balance *t*-tests might lack power
 - Overall balance?
 - Sufficient overlap between treatment & comparison cases?

Balance (Selected Results)

| | New York SSI-concurrent (benefits counseling and waivers) | | | New York SSI-concurrent (benefits counseling, waivers and employment services) | | | Oklahoma SSI-concurrent | | |
|--|--|----------|-----------|--|----------|-----------|-------------------------|----------|-----------|
| | Comparison group | | | Comparison group | | | Comparison group | | |
| | Participants | Selected | Candidate | Participants | Selected | Candidate | Participants | Selected | Candidate |
| Earnings in the year before enrollment (\$) | \$1,665 | \$1,421 | \$1,210* | \$1,812 | \$1,600 | \$1,210** | \$1,030 | \$922 | \$645** |
| Employed in the year before enrollment (%) | 41.5 | 31.8** | 23.0*** | 38.2 | 36.2 | 23.0*** | 28.3 | 34.4 | 20.4*** |
| Age at enrollment | 41.0 | 40.7 | 41.6 | 41.1 | 41.1 | 41.6 | 41.1 | 41.6 | 41.5 |
| White (%) | 42.7 | 38.3 | 58.6*** | 36.9 | 37.5 | 58.6*** | 51.6 | 56.4 | 77.5*** |
| Disability: mental disorder (%) | 92.8 | 100*** | 100*** | 94.0 | 100*** | 100*** | 96.5 | 100*** | 100*** |
| Average SSI benefit per month in 2 years before enrollment | \$461 | \$469 | \$389*** | \$468 | \$452 | \$389*** | \$391 | \$389 | \$285*** |

Open Questions

(possible reasons for RE-OS difference)

- Covariates?
 - Did the 250+ covariates (available from administrative data) tap into the most important covariates domains (that are predictive of selection and the outcomes)?
Not measured (Agodini et al., 2002): e.g., household composition, occupation & industry, motivation to work
 - Waves of pretest measures: the repeated measures of employment rates and earnings might not be highly correlated
- Handling of missing data?

Open Questions

(possible reasons for RE-OS difference)

- Non-local comparison group?
 - Comparison group members come from a different geographical location (non-demonstration areas)
 - if the matching of comparison counties did not include all confounding covariates target populations would not be comparable
- ITT instead of ATE?
 - Could have estimated ITT for both RE and OS (use treatment assignment status instead of participation status for PS logit model)
 - Estimating ITT instead of IV-CATE for RE would rely on weaker assumptions (e.g., no finite sample bias)
- Blinding?

What is special with multilevel data

Issues to be considered with multilevel WSCs

WITHIN-STUDY COMPARISONS WITH MULTILEVEL DATA

What is Special with Multilevel Data?

- Data typically show a *nested data structure*
 - students : classrooms (teacher) : schools :
: school districts : states : countries
 - time (longitudinal data)
- *Interventions* can be implemented at different levels: student- or classroom-, or school-level
 - Long interventions / time-varying interventions
- *SUTVA* is more likely violated
 - within classrooms/schools students are not independent
 - also school might not be independent

What is Special with Multilevel Data?

- With observational multilevel data *selection processes* can be very complex because they might
 - involve many stakeholders (students, peers, parents, teachers, school management, parent teacher association)
 - take place at multiple levels and
 - differ from classroom to classroom, school to school, district to district, etc.
 - introduce bias in opposite directions at different levels

What is Special with Multilevel Data?

- Observational studies in educational research frequently have a *rich set of covariates* in order to control for selection bias
 - *Large number* of covariates measured at *multiple* points in time and different levels, particularly, direct or proxy *pretest measures* on the outcome of interest are frequently available (e.g. achievement scores)
 - Specifying the PS model (model selection) and achieving balance on all observed covariates is more challenging

Multilevel Structure & Treatment & Matching

- To simplify matters, focus only on *two-level structure*
 - Students are nested within school
- *Treatment* might be implemented either at the
 - *school level* (school level treatment) or
 - *student level* (student level treatment)
- Basic *matching strategy* for observational studies depends on the treatment level
 - if treatment is implemented at the school level match intact schools
 - if treatment is implemented at the student level match students

School Level Treatment

- Treatment at the *school level*:
 - Schools select themselves or are assigned to the treatment or control condition
 - all students within a school receive the same treatment
- RCT equivalent (school level):
 - cluster randomized controlled trial (*schools* are randomly assigned to the control or treatment condition)
- Matching strategies (school level):
 - *matching of intact schools*
 - standard PS model with school level covariates, incl. aggregated student level covariates

School Level Treatment

- Matching strategies (cont.):
 - *Local matching* using geographic information (Hallberg et al.):
 - match schools that are locally very close (in the same neighborhood or school district)—the hope is that local matching controls for unobserved neighborhood characteristics
 - PS model includes geographic information (e.g., distance measure based on the longitude and latitude) as covariates with/without caliper
 - Matching of schools followed by an *additional matching of students* within matched (pairs of) of schools—if matched schools show residual imbalance on student level covariates (Hallberg et al., in prep)

Student Level Treatment

- Treatment at the *student level*:
 - students select themselves or are assigned to the treatment or control condition within schools (student level selection)
 - moreover, schools might choose to participate or not (school level selection)
- RCT equivalent (student level):
 - randomized block design (randomization of *students* within schools)
 - overall sample of schools might be randomly or deliberately selected

Student Level Treatment

- Matching strategy:
 - *Match students within schools*
 - local matching: students in the treatment and control condition share the same learning, social, and geographic environment
 - Separate PS model for each school using student level covariates
 - School-specific estimates are then pooled/averaged across school

Student Level Treatment

- Matching strategy (cont'd):
 - *Match students within & between schools*
 - if sample sizes within schools are very small or lack of overlap within schools → need to “borrow” students from other schools
 - PS model is a two-level model with student and school level covariates
 - violates the idea of a randomized block design and of local matching
 - Does it work (i.e., result in unbiased estimates)?

Purposes of WSCs with Multilevel Data

Purposes of multilevel WSCs (in addition to those listed above)

- Different *matching strategies*
 - Intact cluster matching, local matching and two-stage matching for individual-level interventions
 - within vs. between cluster matching for individual-level interventions
- Do matching strategies work at *different levels*?
E.g., in education
 - Student-level (students nested within classrooms/schools)
 - School-level (schools nested within districts)
 - District-level (districts nested within states)
- Matching strategies for *three or more levels*

Purposes of WSCs with Multilevel Data

Purposes of multilevel WSCs (cont'd)

- *Random vs. fixed effects PS models* (for individual-level interventions)
- *Covariate selection*
 - Relative importance of individual- and cluster-level covariates
(particularly of pretest measures of the outcome)
 - Local/geographical covariates
- *Samples size requirements* at individual- and cluster-level

Requirements for WSCs with Multilevel Data

For three-arm WSCs

- RE:
 - cluster randomized trials, randomized block trials, multisite trials
 - Large enough sample sizes (*power*)
- OS:
 - RE data need to be *matchable/identifiable* in data base for OS (access to data, ID numbers at different levels)
 - At each level, RE and OS should have the *same outcome data and covariates* (measured with identical instruments at the same time)
 - Ideally, the unadjusted treatment effect of the OS shows some significant *initial bias*

Preference effects, randomization bias

Six-arm within-study comparisons

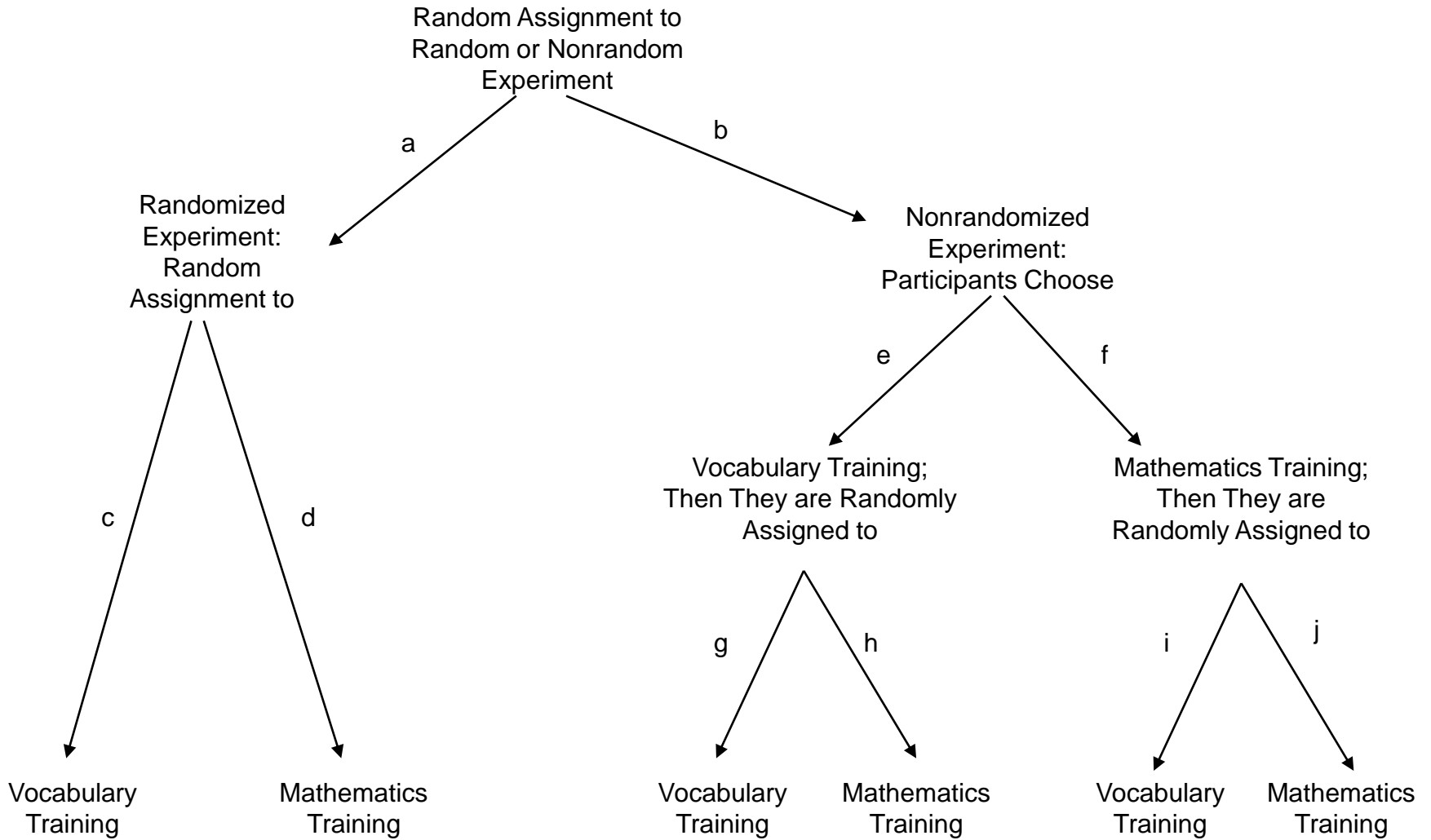
ALTERNATIVE USES OF WSC AND MORE COMPLEX DESIGNS

Alternative Uses of WSC Designs

Within-study comparisons (WSC) / Doubly randomized preference trial (DRPT) can be used for

- evaluating *non-experimental methods*
(Shadish et al., 2008; Pohl et al., 2009)
- estimating *preference effects*
(Long, Little & Lin, 2008a, 2008b; XXX)
- Estimating *randomization bias*
(Marcus, Stuart, Wang, Shadish & Steiner; 2011)

Six-Arm Design



Designs for Mediation Analysis

- Mediation analysis:
 - *Sequential ignorability* assumption (into treatment and into mediator)
 - SUTVA
- Compare experimental designs for studying causal mediation (e.g., sequential randomization) with mediation analyses from OS