Die Kausalität in der Evaluation: Normative und empirische Betrachtungen

Thomas D. Cook

Luxemburg

16. September, 2010

Normative Betrachtungen betreffend die diversen Evaluationsfragen und ihre Beziehungen zur Methodenwahl

1. Aufgabe der Evaluation

- Das Evaluationsobjekt feststellen
- Die Theorien dessen Wirksamkeit klarlegen und dann überprufen, ob sie in Einzelheiten schon falsch bewiesen sind
- Dazu brauchen wir theoretische Analysen und Kenntnisse der relevanten Literatur
- Die Bedeutung dieser Aufgabe ist hoch und unumstritten; wir haben die Mittel dazu; und also für heute ist diese erste Aufgabe uninteressant

2. Aufgabe der Evaluation

Die Implementation beschreiben, um zweierlei festzustellen: 1.Welche Aspekte des Programms befriedigend und unbefriedigend implementiert sind? und 2. Was wir machen können, um die Implementation zu verbessern?

- Dazu sind sowohl qualitative wie auch quantitative Methoden geeignet
- In den USA ist diese Aufgabe kaum umstritten und macht eine “mixed-method” Evaluation unentbehrlich

3. Aufgabe der Evaluation

Kausalzusammenhänge zwischen dem Programm und dessen möglichen Konsequenzen festlegen

Politisch ist jedes Programm nur durch seine Fähigkeit gerechtfertigt, das Leben von Individuen besser zu machen

Besser machen ist kausal - d.h.,wenn das Programm da ist, so ist das Leben besser, mindestens im Durchschnitt wenn nicht für einen jeden Einzelnen

4. Aufgabe der Evaluation

- Feststellen, warum oder mittels welcher Prozesse die Ursache zur Wirkung führt
- Dazu sind qualitative und quantitative Methoden der “Modelling” angebracht
- Obwohl Prozesskenntnisse von grösster wissenschaftlicher Bedeutung sind, sind sie für die Social Policy nur dann relevant, wenn sie wesentlich zur Programmbesserung beitragen. Aber
- Wozu das Programm bessern, wenn es grundsaetzlich unfaehig ist, wirksam zu sein? “Mediation supposes efficacy”.

5. Aufgabe der Evaluation

- Summativ entscheiden, ob das Evaluationsobjekt gut oder nicht gut ist
- Dazu stehen mehrere Methoden zur Verfügung: Kosten-Nutzen Analyse; empirische Analysen der langfristigen Konsequenzen und derer Nebenkosten; Meta-Analyse; oder sogar Diskussionen mit Stakeholders, die unterschiedliche Interessen vertreten
- Egal wie die 5. Aufgabe ausgeführt ist, setzt sie die *bewiesene* Kausalität voraus

Na, und?

- Es wäre schön, wenn wir all diese Fragen antworten könnten - aber in einer Studie haben wir weder die Mittel noch die Zeit dazu. Wir müssen Prioritäten vorschlagen und rechtfertigen. Das mache ich gleich.
- Die Aufgaben waren frueher in logischer Reihe aufgestellt. Warum sollen wir sie also nicht reihenweise beantworten? Erfahrung in den USA zeigt aber, dass wir meistens so viel Zeit mit den ersten Aufgaben verbringen, dass wir die letzteren selten angreifen

Welche Aufgabe verdient die höchste Priorität?

- Aufgabe 1: Nein, weil…
- Aufgabe 2: Nein, weil…
- Aufgabe 3: Ja, vor allem weil Evaluation und kausale Behauptungen von der Wenn/Dann Art logisch zusammenhängen
- Weil Aufgabe 4 mehr mit der “Science” als mit “Social Policy” zu tun hat
- Weil Aufgaben 4 und 5 akzeptable kausale Kenntnisse voraussetzen
- Augabe 3 ist also der Schluessel.

Normative Betrachtungen: Zusammenfassung

- Die Evaluation muss sich mit diversen Fragen befassen
- Also muss sie sich mit diversen Methoden auch befassen
- Aber nicht jede Art von Evaluationsfrage ist gleich wichtig
- Deswegen muss man Prioritäten in der Evaluationsfragestellung klar machen und rechtfertigen, die dann notwendigerweise zu Prioritäten in der Methodenwahl führen.

Normative Betrachtungen: Zusammenfassung 2

- Begrifflich: Ohne Kausalität gäbe es keine Evaluation, nur Management Consulting in dem öffentlichen Sektor
- Praktisch: Das öffentliche Gut und die politische Realität verlangen, dass wir die Wirksamkeit von Programmen feststellen, die von Steürgeldern subventioniert sind
- Das führt also zur wichtigen Frage: Da kausale Behauptungen so wichtig sind, wie begründet man sie am besten?

Empirische Betrachtungen

2 Rahmenbedingende Bemerkungen

- Es geht um einen bestimmten Begriff von Kausalität - wenn/dann und nicht prozesserklärend wie z.B. bei Modeling
- Epistemologisch grundlegend ist das Ziel, alle zur Zeit denkbare alternative Erklärungen als ungültig zu beweisen, damit ein einziger Schlusss möglich ist: Dass der Zusammenhang zwischen A und B hoechstwahrscheinlich kausal ist

Dementsprechend

- Kann jede Methode prinzipiel kausal sein, wenn sie es fertigbringt, alle Alternative als ungueltig zu beweisen.
- Kann jede Methode dazu beitragen, einige (wenn nicht alle) der alternativen Erklärungen als ungueltig zu beweisen, Jede Methode ist also mindestens zum Teil kausal relevant und keine Methode ist also völlig irrelevant
- Die praktische Kernfrage ist: Welche Methoden sind besser, da sie mehr Alternative ausschliessen als andere und zwar häufiger?

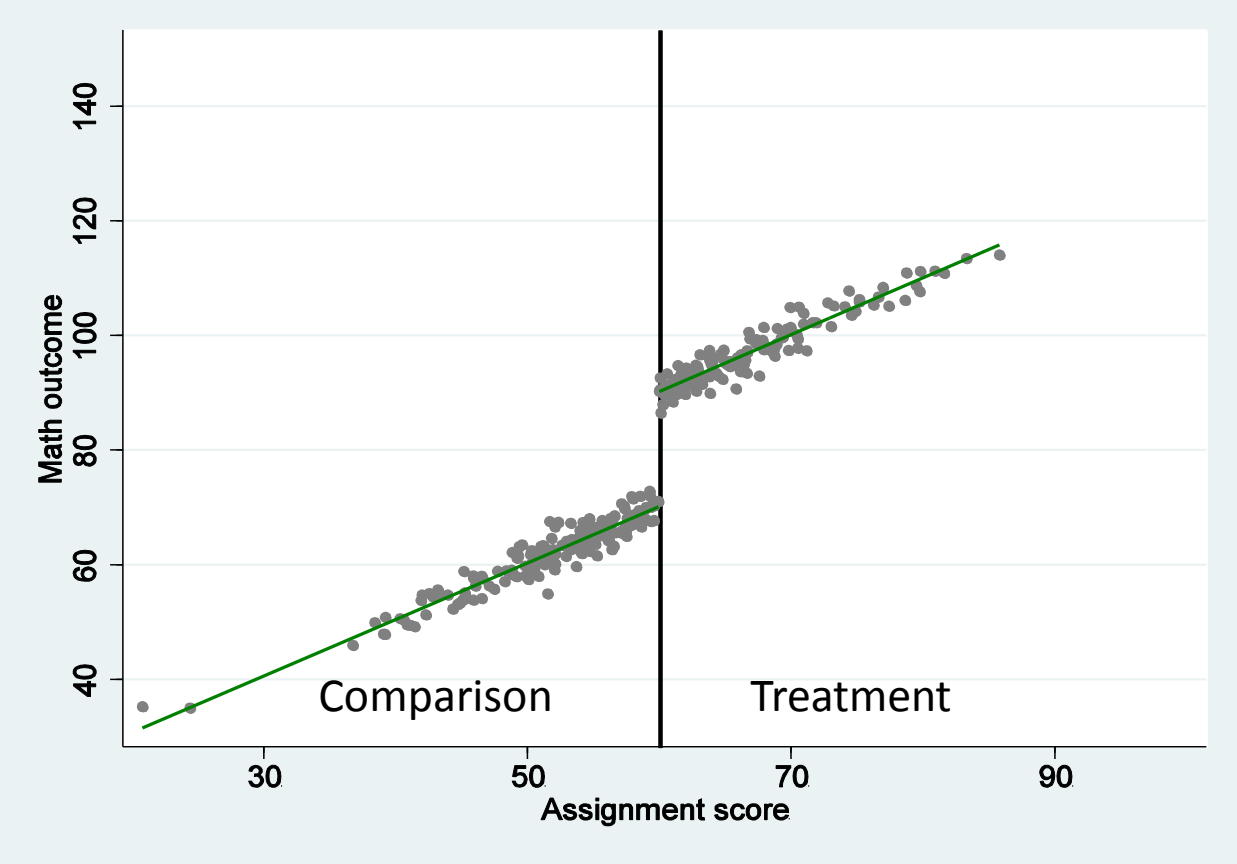
Experimente

- In der Wissenschaft gilt das Experiment als das beste Mittel, kausale Hypothesen zu testen
- Voraussetzungen des Experiments sind klar und empirisch einfach zu überprüfen
- Kann viel häufiger verwendent werden, als Kritiker behaupten, aber immerhin nicht immer
- Nachteil: auf Freiwillige beschränkt
- In den USA zunehmend vom Kongress verlangt, wenn ein Sozialprogramm wichtig ist

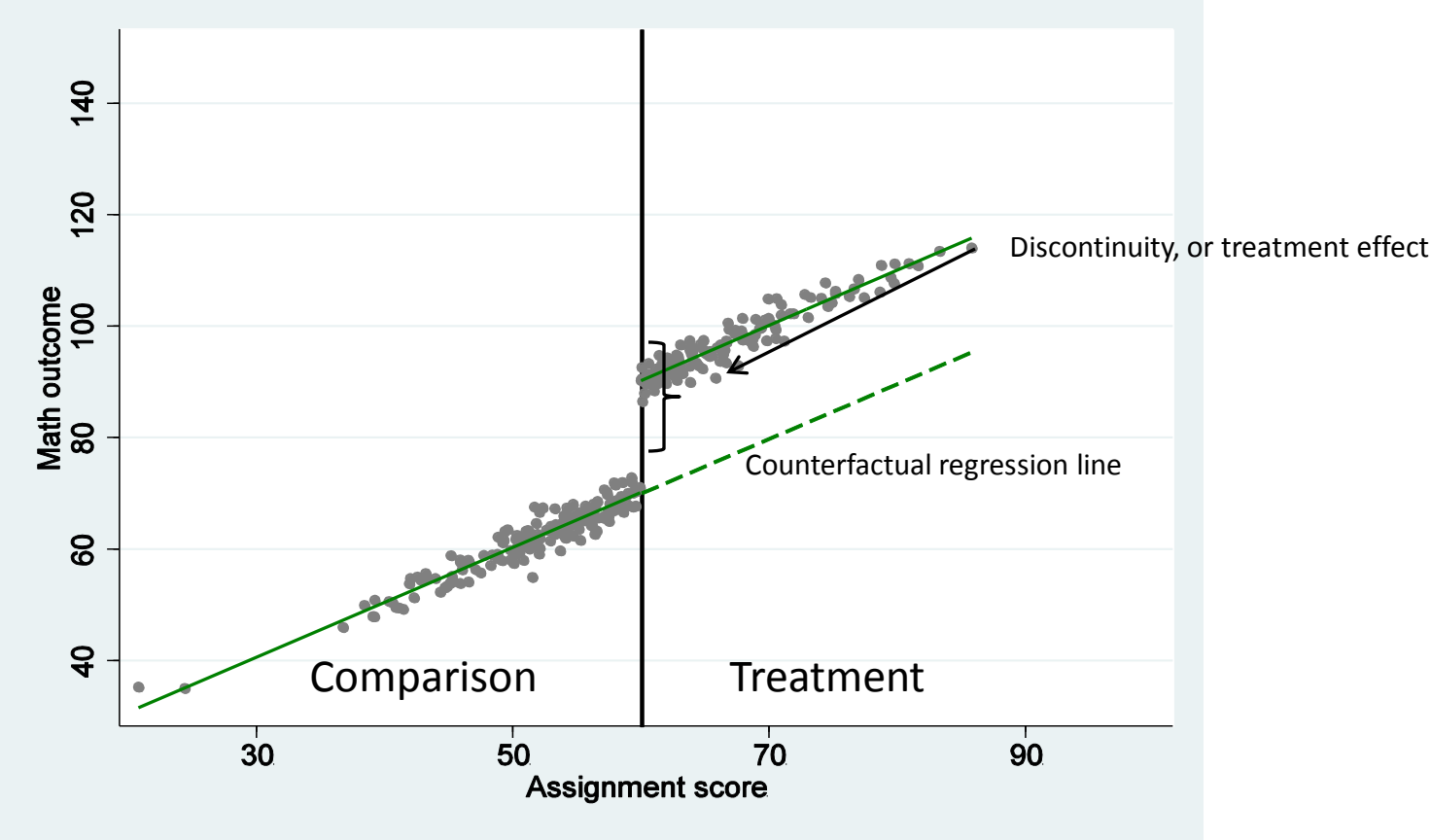
Quasi-Experimente: Regression-Discontinuity

- Viele Programme sind ausschiesslich für diejenigen geeignet, die besonders verdienstwürdig oder nötig sind
- Verdienstwuerdig und noetig sind von einem bestimmten Wert an einer Verteilung abhaengig.
- Dann kann man Regression-Discontinuity verwenden

RDD Visual Depiction



RDD Visual Depiction



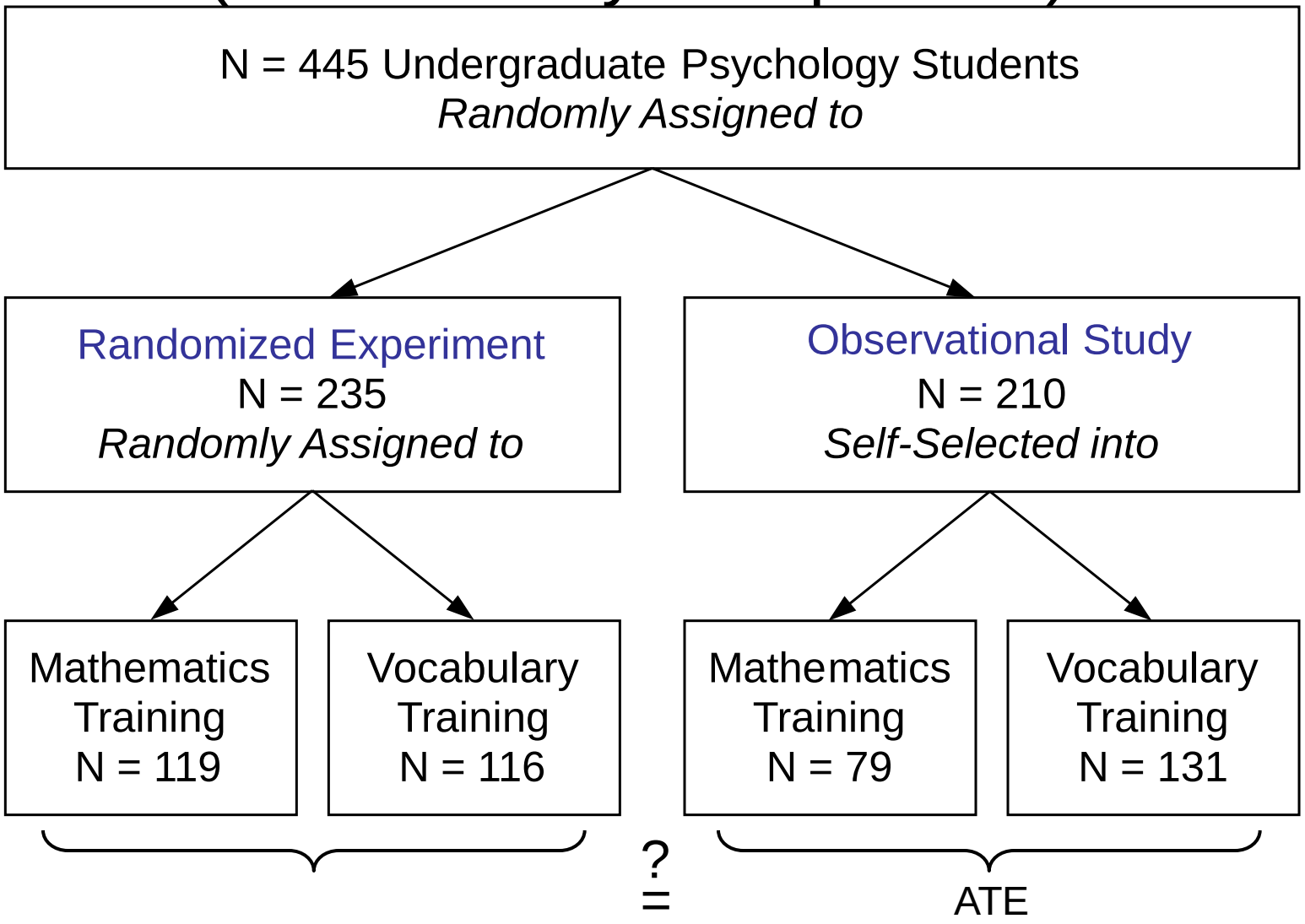
Regression-Discontinuity (RD)

- RD hat klare Voraussetzungen, die empirisch prüfbar sind
- RD führt zu biasfreien kausalen Antworten, obwohl die standard errors grösser als bei Experimenten sind
- Das ist in 5 Vergleichen von experimentellen und RD Ergebeniseen empirisch bewiesen worden
- RD soll häufiger gebraucht werden

Wie stellt man fest, dass eine nicht experimentelle Methode kausal gültig ist?

Shadish, Clark & Steiner (2008)

(Within-Study Comparison)



Shadish et al.: Treatments & Outcomes

- Two treatments and outcomes
  - *Two treatments*: short training either in *Vocabulary* (advanced vocabulary terms) or *Mathematics* (exponential equations)
    $\rightarrow$ All participants were treated together without knowledge of the different conditions
  - *Two outcomes*: *Vocabulary* (30-item posttest) and *Mathematics* (20-item posttest)
- Treatment effect:
  - ATE: *average treatment effect* for the overall population in the observational study

Shadish et al.: 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 or quasi-experiment
  - Hence, measurements are not influenced by assignment or treatment

Shadish et al.: 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

Shadish et al.: 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)

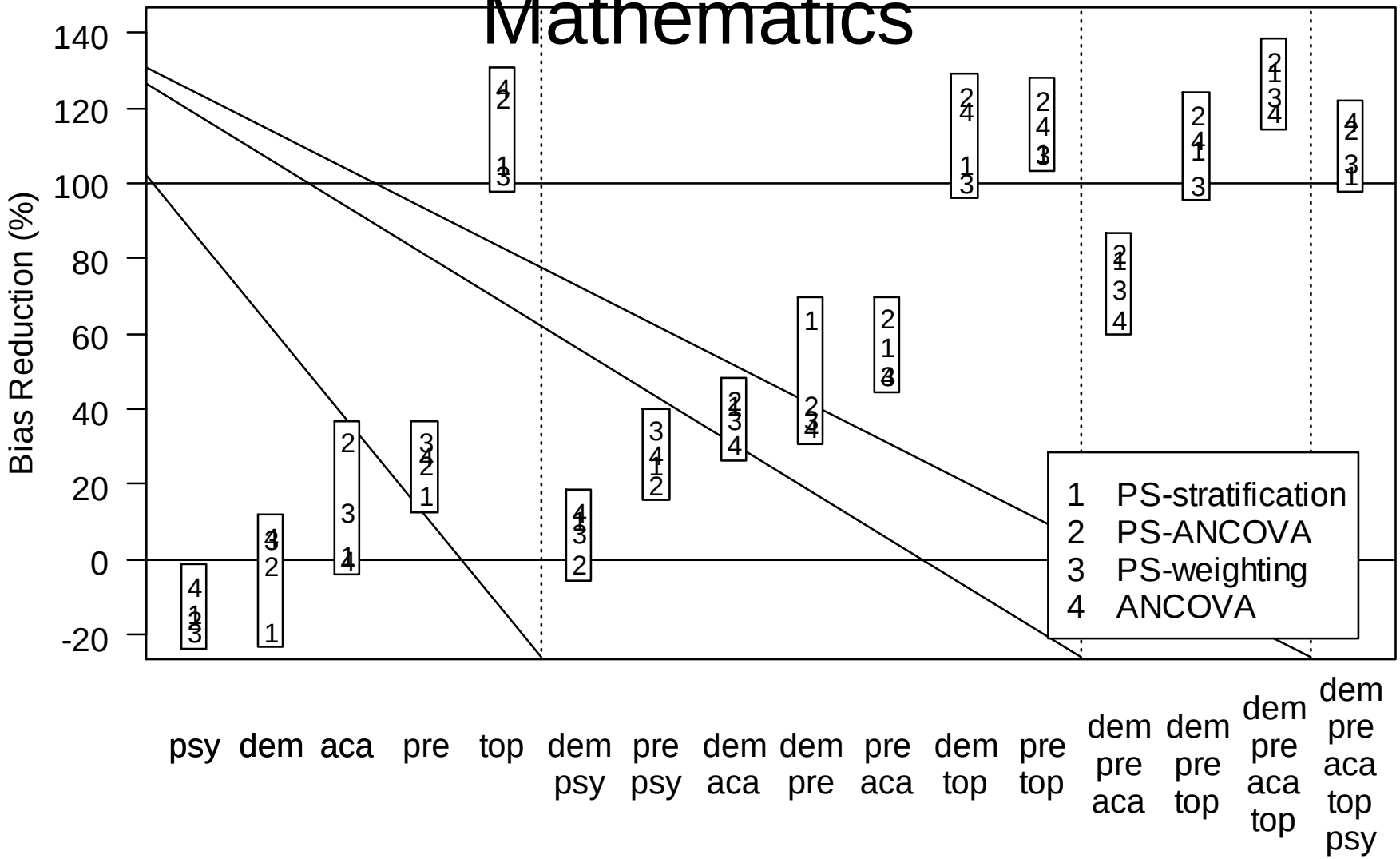
Shadish et al.: Unadjusted Results
Effects of Math Training on Math Outcome

	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

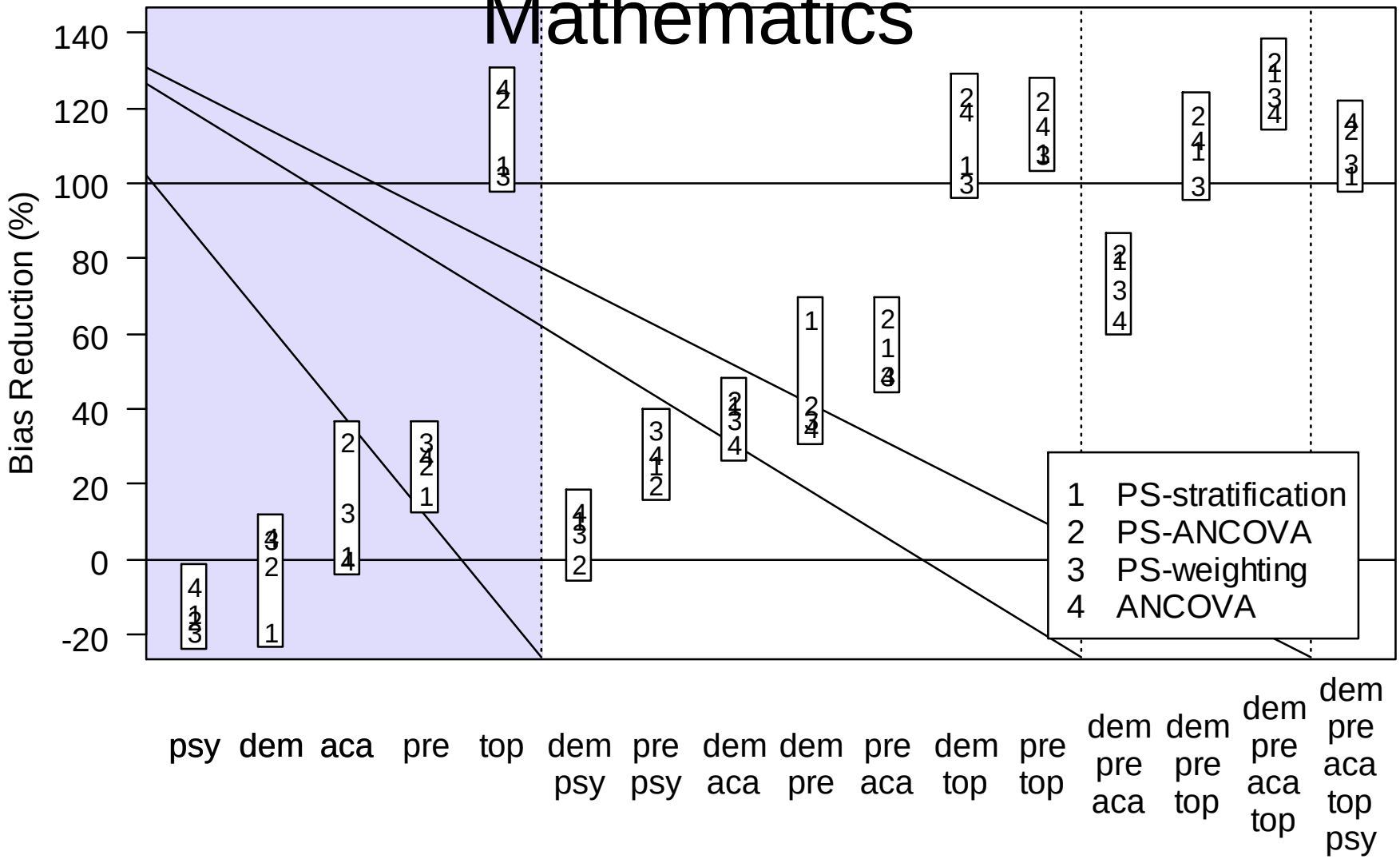
Conclusions:

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

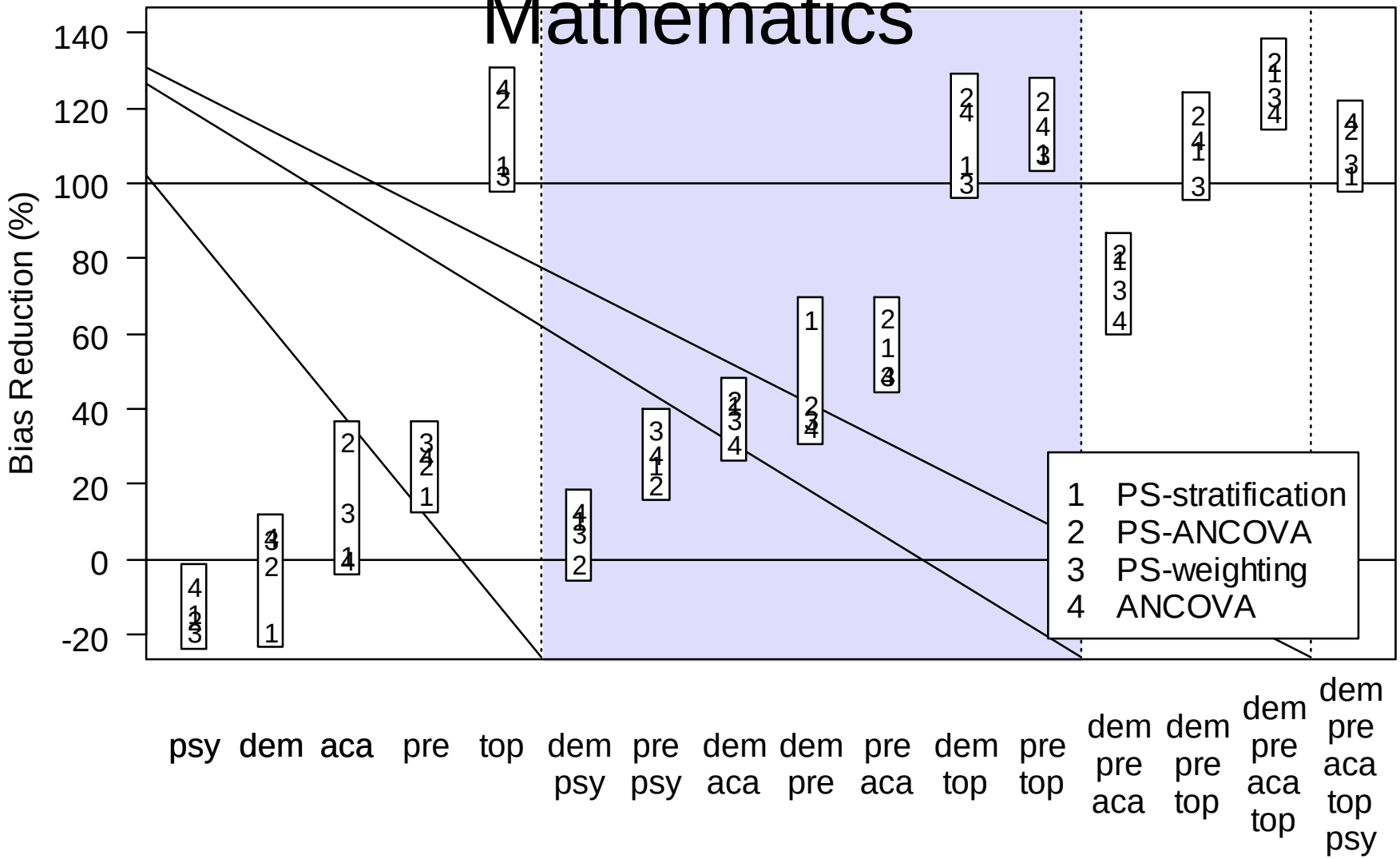
Bias Reduction: Construct Domains Mathematics



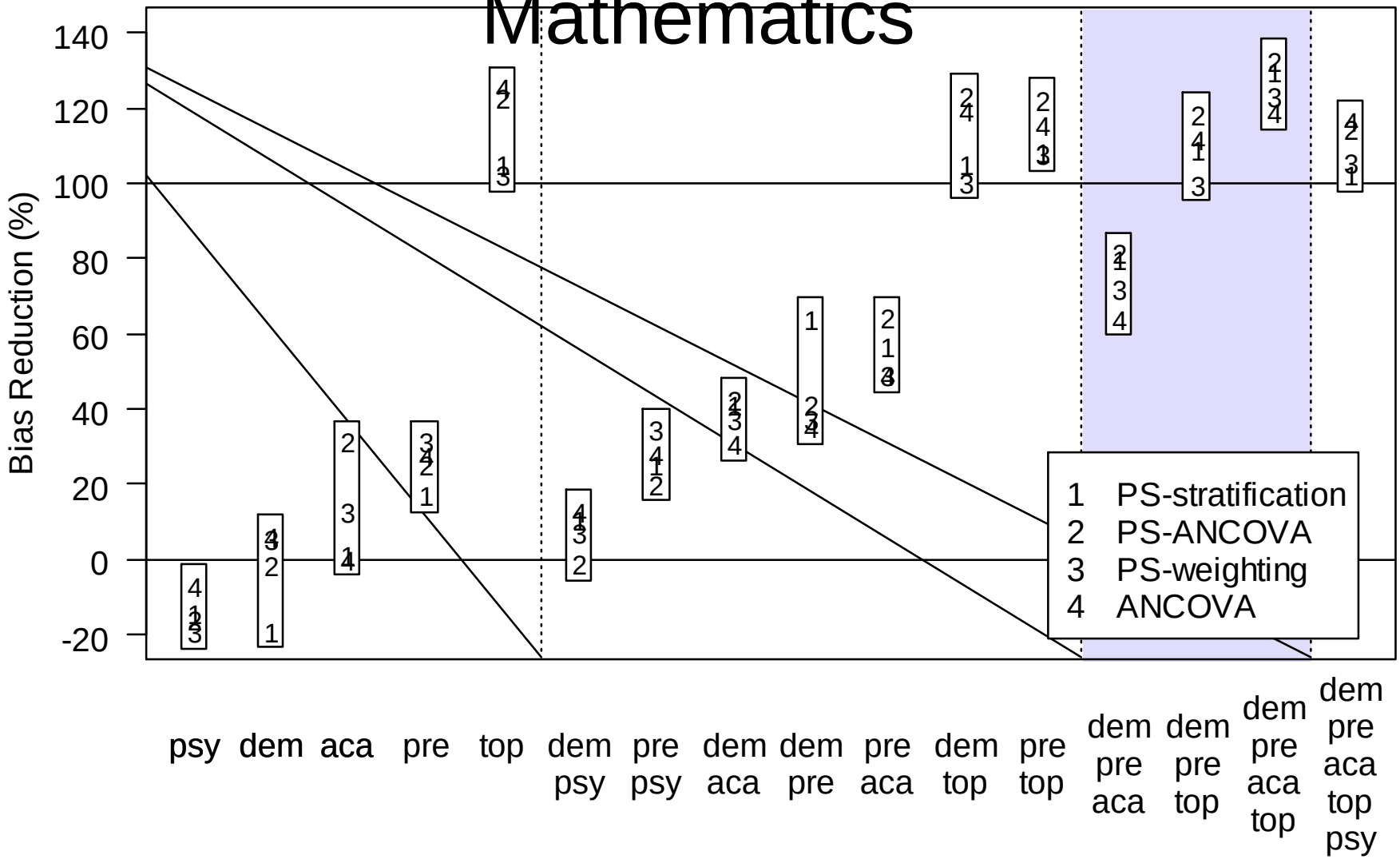
Bias Reduction: Construct Domains Mathematics



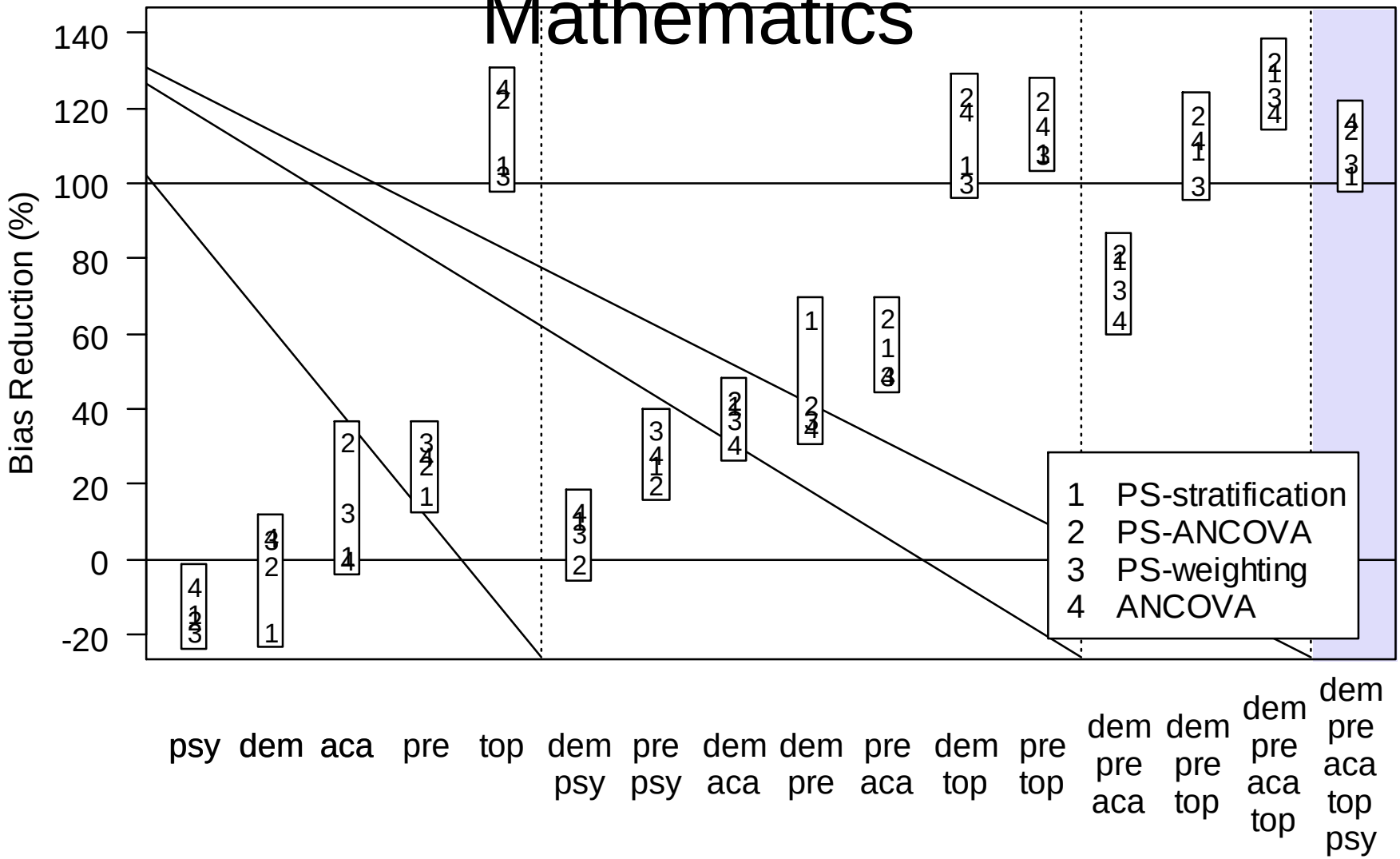
Bias Reduction: Construct Domains Mathematics



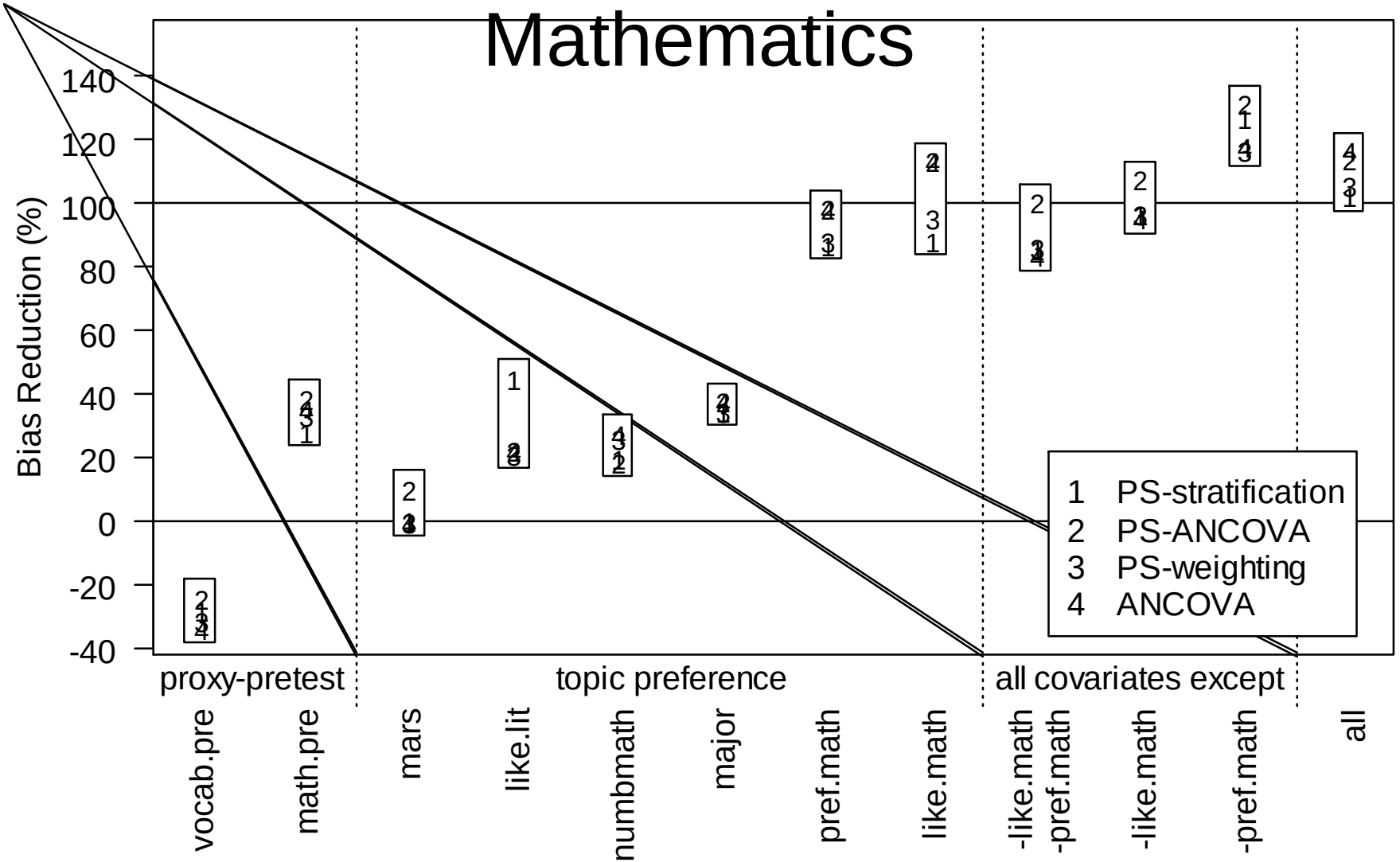
Bias Reduction: Construct Domains Mathematics



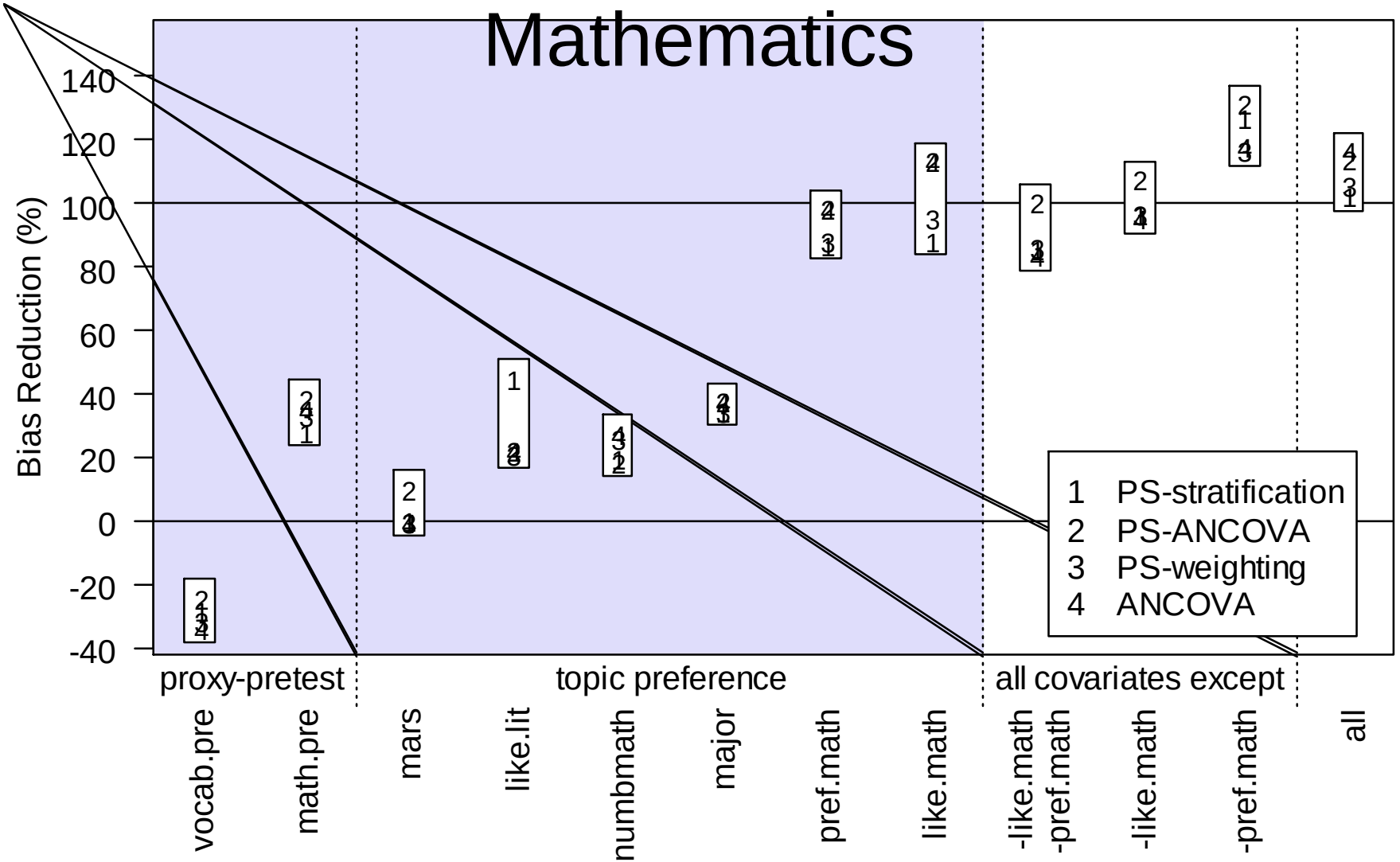
Bias Reduction: Construct Domains Mathematics



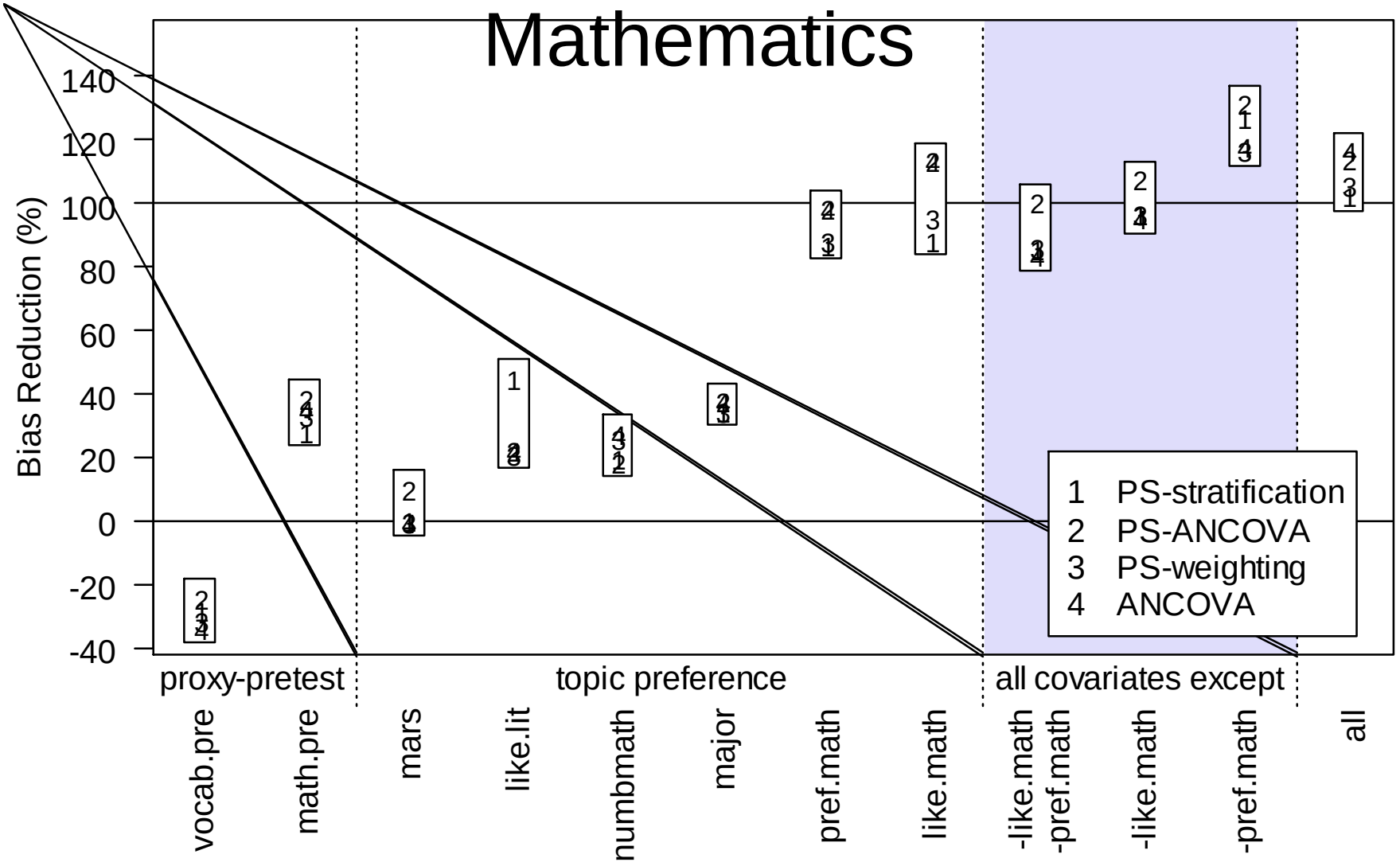
Bias Reduction: Single Constructs Mathematics



Bias Reduction: Single Constructs Mathematics



Bias Reduction: Single Constructs Mathematics



Ergebnisse

- 1. Wünschenswert waere es, die Selektionprozesse genau zu wissen - das kommt selten vor, aber mit Hilfe von Theorie und qualitative Methoden kann man das manchmal annähren
- 2. Möglicher ist, dass man auf sehr vielen Bereichen gleichzeitig matchen kann
- Beides führt oft zu demselben kausalen Ergebnis wie in einem Experiment

Ein anderes Ergebnis: Multi-level Matching

- 1. Schritt: auf der Ebene der Schule oder Nachbarschaft matchen - mittels Variablen, die mit dem Endpunkt am höchsten korrelieren
- 2. Schritt: auf der Ebene der Individuen matchen, je nach dem obigen Prinzip oder mit Variablen von vielen, unterschiedlichen Bereichen
- Das kann oft zum selben kausalen Ergebnis wie beim Experiment führen

Was nicht zum selben Ergebnis führt

- Alle anderen quantitativen Alternative, die bei weitem die Mehrzahl heutiger Praxis/Anwendungen bezeichnen
- Es steht uns also bloss eine geringe Wahl von nichtexperimentellen Methoden zur Verfügung, die zu demselben Ergebnis wie bei einem Experiment führen
- Man soll diese nichtexperimentelle Alternative also bevorzugen

Zusammenfassung 1

- Die Evaluation muss mixed method sein, weil sie Fragen verschiedener Art stellt, zu denen unterschiedliche Methoden relevant sind
- Aber die Kernfragen der Evaluation sind kausal oder setzen etablierte Kausalzusammenhänge voraus
- Deswegen sind kausale Methoden von besonderer Bedeutung für die Evaluation

Zusammenfassung 2

- Kausale Methoden koennen ihrem Wert nach eingeordnet werden:
- 1. das Experiment; 2. Regression Discontinuity; 3. Matching, wo der Selektionsprozess unabhängig bekannt ist, wo Kovarianzvariablen aus vielen unterschiedlichen Bereichen vorhanden sind, oder wo Matching auf mehreren Ebenen zugleich stattfindet
- Die Evaluationspraxis soll diese Methoden bevorzugen