LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Research Designs

Wintersemester 2025/26

Dienstag, 16:15-17:45 Uhr, Raum 308

Dr. Fabian Kalleitner

Allgemeiner Hinweis zum Skript: Ein Teil der Folien wurde von Katrin Auspurg und Daniel Krähmer entworfen!

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Research Designs: Eine Einführung

Vorstellungsrunde

Dozierendenteam

- Interessen sowie Erfahrung in Forschung und Lehre
- Motivation für das Seminar

Und nun zu Ihnen…

- Belegung über welches Modul?
- Hauptfach im BA?
- Thema der Bachelorarbeit?
- Vorkenntnisse?

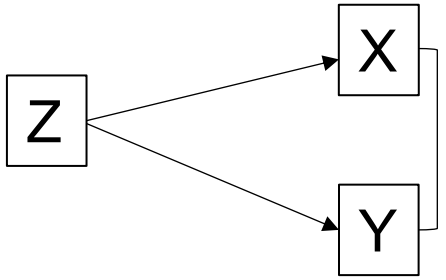
Inhaltlicher Einstieg: Forschungsziele und Forschungsdesigns

Forschungsziele in der Sozialforschung

1. Deskriptionen: Was ist der Fall?
   (z.B. Sozialberichterstattungen)
2. Erklärungen: Warum ist etwas der Fall?
   - Entwicklung möglichst guter Theorien
   - Evaluations- und Wirkungsstudien

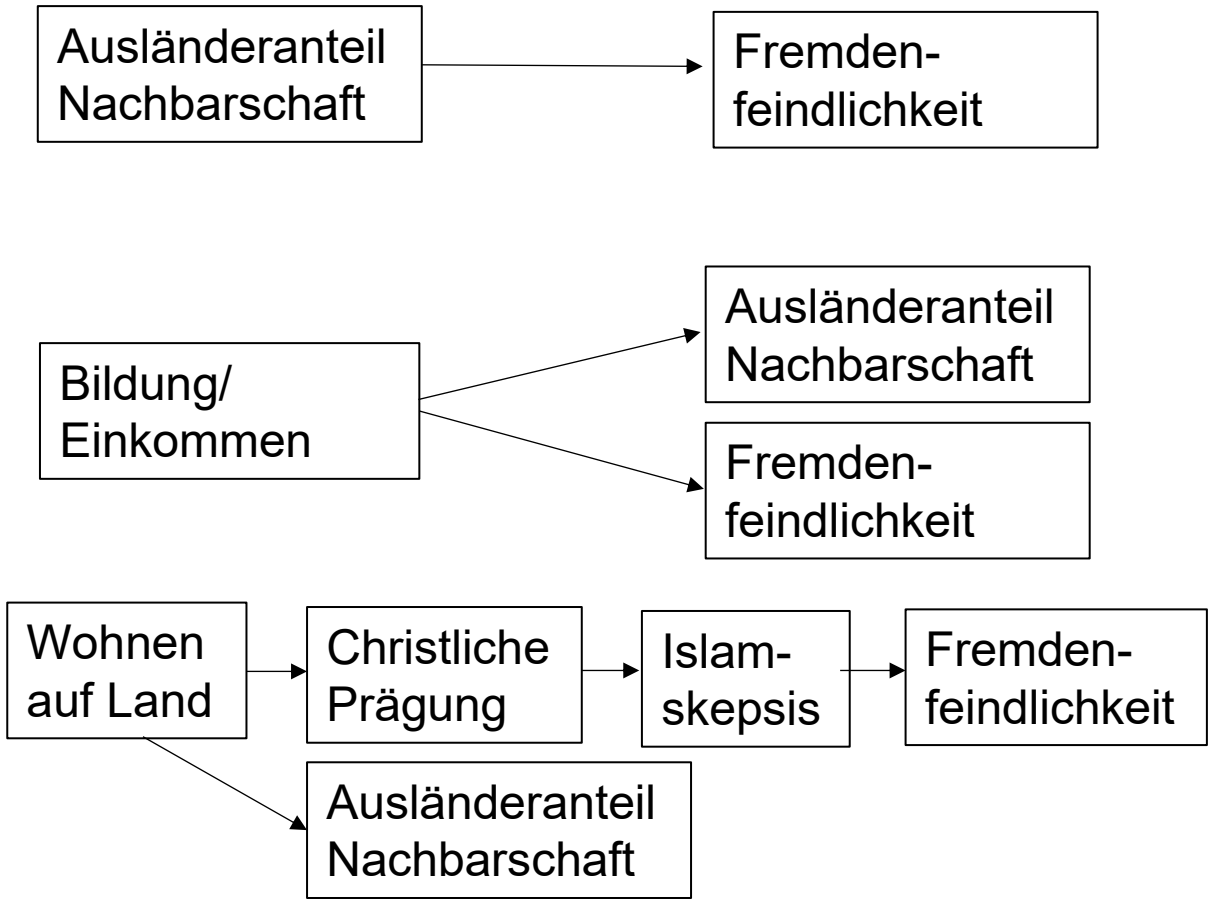
- Deskriptionen sollten Erklärungen immer vorangehen
  („One should not theorize in advance of the facts“)

- Erklärungen sind besonders anspruchsvoll
  - Es gibt in der Regel viele mögliche Ursachen
  - Korrelation ≠ Kausalität
    Beispiel Confounder/Scheinkorrelation



Beispiel Kontaktthese

- Je mehr Kontakt zu Ausländern, desto weniger fremdenfeindliche Einstellungen
- Alternative Erklärungen für eine beobachtete Korrelation?
  - Selektion
  - Umgekehrte Kausalität?



Kausaler Effekt

- Mindestbedingungen für Kausalität
  - Es gibt einen Zusammenhang zw. X und Y
  - Ursache (Treatment) X geht Outcome Y voraus
  - Es ist wirklich X, das Y bewirkt
    - Y ist anders, wenn X (nicht) vorliegt
    - X (und nicht andere Variablen) ist ursächlich für den Unterschied in Y
- Ideal der kontrafaktischen Kausalität
  - Vergleich derselben Einheiten zur gleichen Zeit mit/ohne Treatment (Parallelwelt)
  - Das ist *per se* nicht zu beobachten
- Forschungsdesigns versuchen kontrafaktische Kausalität möglichst gut anzunähern
  - Isolation von X
  - Anders gesagt: „Gleiches mit gleichem“ vergleichen: nur X sollte den Unterschied machen können
  - Ausschluss von möglichst viel anderer Heterogenität

Forschungsdesign

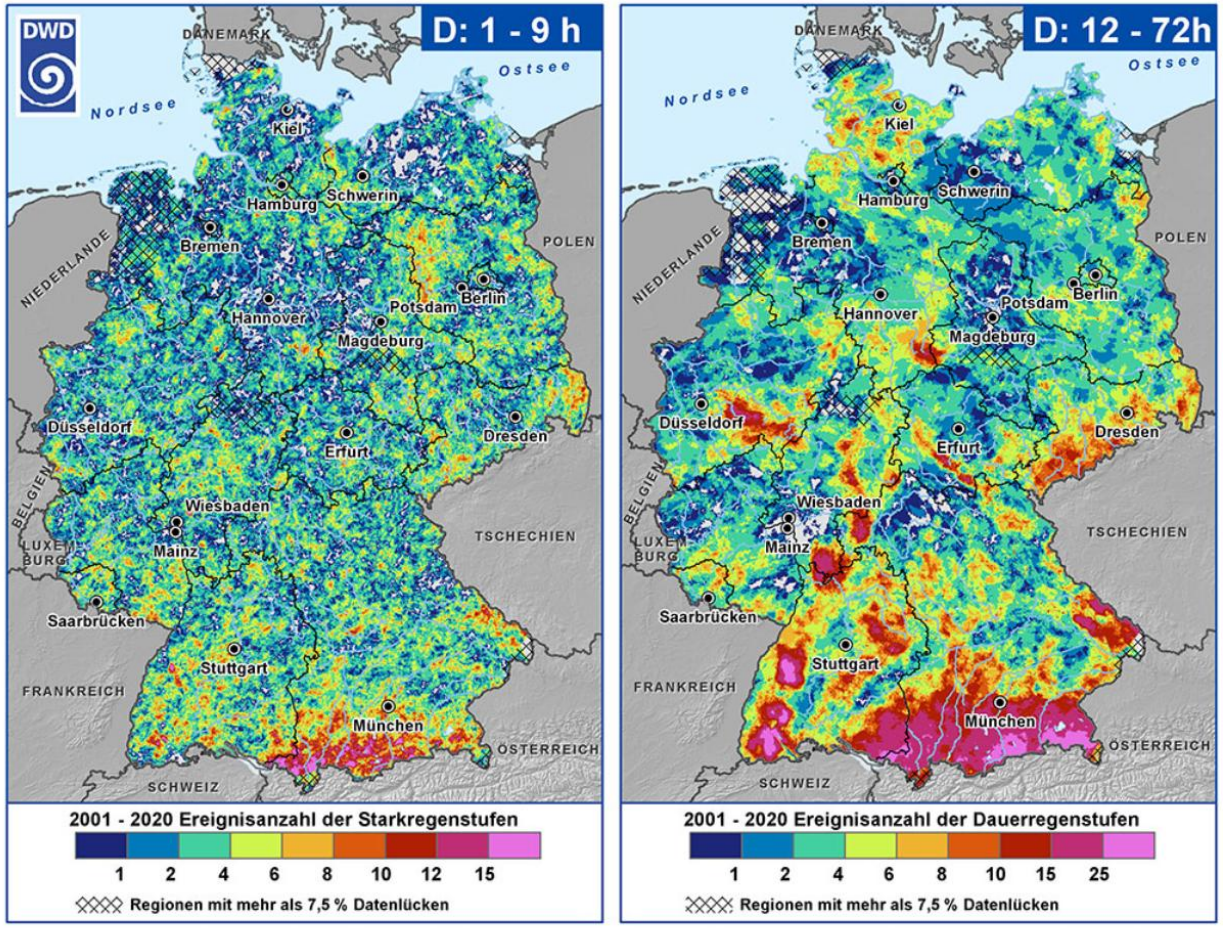
- Ziel: Möglichst überzeugende Antwort auf Forschungsfrage: Was/warum ist etwas der Fall?
- Bei Kausalfragen: Qualität kausaler Inferenz verbessern

➢Welche empirische Information ist optimal, um eine möglichst zuverlässige (unzweifelhafte) Antwort zu finden?

Beispiele – und nun zum Wetter…

- Einfluss von Extremwetter auf Einstellungen zum Klimaschutz
- Ideal: Vergleich von Regionen, die sich nur im Extremwetter unterscheiden
  - Zufällig getroffen werden oder nicht

- Das wird nicht ganz möglich sein – aber mehr oder weniger gut anzunähern
- Was Sie lernen sollen: Wann ist z.B. das eher der Fall und sind damit (Kausal-) Effekte verlässlicher identifizierbar?



Klimadaten und Darstellung: © DWD 2021 (CatRaRE Daten: 10.5676/DWD/CatRaRE_W3_Eta_v2021.01); Geodaten: © GeoBasis-DE/BKG 2020 (Stand: 01.01.2020).

https://www.ardalpha.de/wissen/umwelt/klima/starkregen-dauerregen-dwd-104.html

Forschungsdesigns lassen sich unterscheiden nach

- Anzahl und Art der Vergleichsgruppen: „between“ vs. „within“
- Falls between: Verwendung von Messung vor Treatment ja/nein
- Art der Zuteilung zu Gruppen
- Art des Treatments: durch Forschende gesetzt oder durch die „Natur“ gegeben
- Anzahl der Treatments/Interventionen: Werden auch kumulative Effekte gemessen? Interaktionen?

- Hinzu kommen unterschiedliche Messungen und Datenerhebungen
  - Surveys, Beobachtung
  - Qualitativ, quantitativ

- Zunächst sollte aber die grundsätzliche Logik (Design) überlegt werden!

Zwei Strategien Ausschluss alternativer Erklärungen

(1) „ex ante“

- Durch das Design der erhobenen Daten
  - z.B. Experimente, die durch Randomisierung Störfaktoren eliminieren

Seminar Research Designs, Forschungsmethoden, Forschungspraktikum

(2) „ex post“

- Durch Verfahren der Datenanalyse
  - z.B. Kontrolle von Störfaktoren durch multivariable Regressionen, Matching,…

VL und Übungen im kleinen SP „Quantitative Methoden“, Forschungspraktikum

Im dritten Semester Zusammenführung und weitere Details in der „Kausalanalyse“

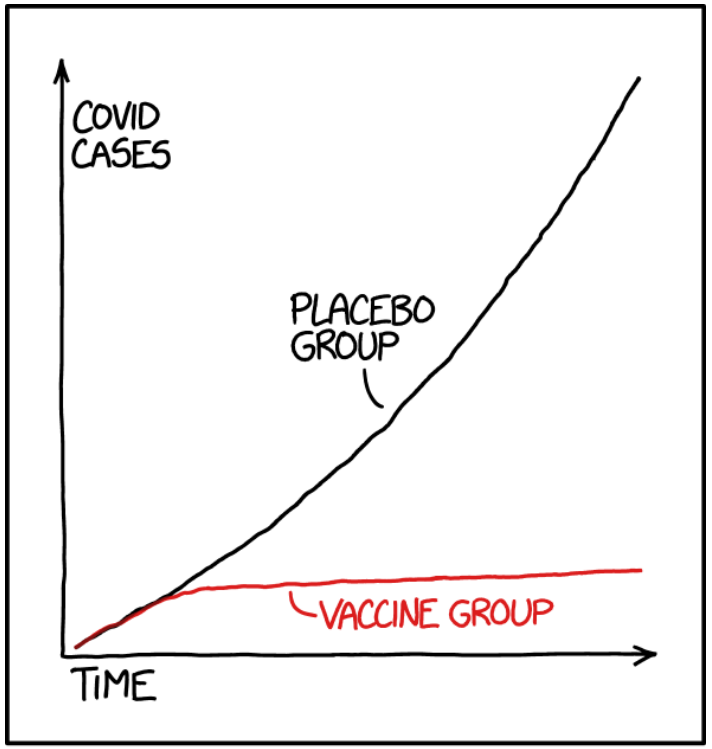
(1) ist vorzuziehen, aber nicht immer möglich und ausreichend

Seminarorganisation: Wer, was, wann, und wie?

Bildquelle: Dall-E

Was lernen Sie hier?

- Grundlagen für „gute“ Forschungsfragen und Research Designs
- Was sind Kriterien für die Einschätzung der Güte (Validität)?
  - Somit auch: Anhand welcher Aspekte kann man vorliegende Studien kritisieren?
  - Was gilt es zu beachten, wenn man selbst Studien plant?
- Wissen zu speziellen Designs: Experimente, Surveys, Big Data, …
- Fokus liegt dabei auf Datenerhebungen!



STATISTICS TIP: ALWAYS TRY TO GET DATA THAT'S GOOD ENOUGH THAT YOU DON'T NEED TO DO STATISTICS ON IT

XKCD

Kursorganisation: Moodle

Falls noch nicht geschehen: **Schreiben Sie sich in den Moodle-Kurs ein!**

Titel: [WiSe 2025/26] 15206 Research Designs

Passwort: Des!gns25

→ Über Moodle haben Sie Zugriff auf sämtliche Materialien (Folien, Literatur, etc.)

→ Die Abgabe sämtlicher Prüfungsleistungen erfolgt ebenfalls über Moodle!

Kursorganisation: Leistungsanforderungen

Aufschlüsselung der ECTS-Punkte

- Regelmäßige und aktive Teilnahme:
  ca. 30 Stunden (1 ECTS)
- Regelmäßige Vorbereitung (Lektüre Basistext):
  ca. 2 h/Woche (1 ECTS)
- Vorbereitung eines Referats:
  ca. 30 Stunden (1 ECTS)
- Abgabe von 2 Übungsaufgaben während des Semesters: ca. 30 h (1 ECTS)
- Verfassen eines Forschungsproposals:
  ca. 60 Stunden (2 ECTS)

Notenbildung

- 50% Forschungsproposal
- 25% Übungsaufgaben
- 25% Referat/mündliche Beiträge

Kursorganisation: Prüfungsleistungen

Übungsmappe, 3 Teile

2

während des Semesters

- ÜA A5
- Freie Wahl aus ÜA B
- Je ÜA + 1-2 Folien
- s. Aufgabenblatt und
- Details im Syllabus

+

1

zum Semesterabschluss:

- Forschungsproposal
- Skizze bis 31.01. + 1-3 Folien
- ca. 20.000 Zeichen (Abgabefrist: 15.03)

+

Referat

1

Wahltermin (im B-Block)

- Gruppen à max. 2 Pers.
- 15 min. + Diskussion
- Verpflichtende Vorbesprechung (Übung)

Kursorganisation: Veranstaltungsprogramm

	Datum	Thema
Grundlagen (A)	14.10.	Einführung
	21.10.	Grundlagen: u.a. Gütekriterien
	28.10.	Kausalität und Experimente
	04.11.	Bedrohungen der Validität
	11.11.	Stichproben, Generalisierbark.
	25.11.	Anwendung auf Studien

	Datum	Thema
Vertiefung (B)	02.12.	Feldexperimente
	09.12.	Natürliche Experimente
	16.12.	Meta-Analysen, Replikationen
	13.01.	Digitale Beobachtungsdaten
	20.01.	Digitale Experimente
	27.01.	Räumliche/Geo-Daten
	03.02.	Wrap Up, Elevator Pitch

Kursorganisation: Referatsvergabe

- Die Themenwahl für die Referate erfolgt über Moodle (Kapitel „Organisatorisches“)
- Windhundverfahren, d.h. „First come, first served“
- Beginn der Vergabe: **Morgen, Mittwoch (15.10.), 18 Uhr**

Hinweise:

- Wer bis Ende der Frist (Montag, 27.10., 23:59 Uhr) kein Thema wählt, wird zugelost
- Innerhalb der Frist können Sie beliebig häufig zwischen freien Terminen wechseln

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Research Designs: Grundlagen

Agenda

1. Das Ausgangsproblem: Beobachtung von Kausaleffekten
2. Was sind „gute“ Fragestellungen?
3. Ausschluss alternativer Erklärungen, Falsifikation statt Konfirmation
4. Gütekriterien für empirische Forschung und Designs

Kausaleffekte sind nicht beobachtbar – Warum?

- Interesse an Kausaleffekt einer "Treatment Variable" D auf "Outcome Variable" Y
  - Was wäre, wenn Treatment (Ereignis) nicht passiert wäre?

Notationen:

- D: Treatment Variable

$$\mathrm{D} = \begin{cases} 1 \text{ treated (Treatmentgruppe T)} \\ 0 \text{ nicht getreated (Kontrollgruppe C)} \end{cases}$$

- Potenzielle Ereignisse (Outcomes):
  - Y^1: potentielles Outcome mit Treatment
  - Y^0: potentielles Outcome ohne Treatment
- Realisierungen für Individuum i: d_i, y_i^1, y_i^0

Individueller Kausaleffekt:

- Treatmenteffekt für Individuum i
- Definiert als:

$$\delta_i = y_i^1 - y_i^0$$

- Dieser ist per se nicht beobachtbar! (Problem kontrafaktischer Kausalität – man kann Individuum zum selben Zeitpunkt nur in einem Zustand beobachten)

Beispiel

- Was ist der Effekt von Aspirin auf die Wahrscheinlichkeit Kopfschmerzen zu entwickeln?
- Beispiel für eine Person: Tom

Was sind die “Potential Outcomes” (potentielle Ergebnisse)?

- Y_{1Tom} : Das potential outcome realisiert wenn Tom das Treatment (Aspirin) genommen hat.
- Y_{0Tom} : Das potential outcome realisiert wenn Tom das Treatment (Aspirin) nicht genommen hat.

Was ist der kausale Effekt von Aspirin auf Toms Kopfschmerzen?

- $\tau = Y_{1,Tom} - Y_{0,Tom}$

Von einer Person zu vielen

- Kausaler Effekt von Person i: $\tau_i = Y_{1i} - Y_{0i}$ → Problem?
- Individuen tragen nur für die Gruppe, in der sie beobachtet werden, Informationen über die Ergebnisse bei. Realisierte Ergebnisse enthalten nur einen Teil der Informationen, die wir benötigen, um kausale Effekte für alle Einheiten direkt zu berechnen.

	Y_{1i}	Y_{0i}
Treatment Group $(\mathrm{D} = 1)$	Observable as Y	Counterfactual
Control Group $(\mathrm{D} = 0)$	Counterfactual	Observable as Y

- Aus diesem Grund kann das fundamentale Problem von kausaler Inferenz (Holland) auch als **Problem von unvollständigen Daten** verstanden werden. Wir können nicht beide potentielle Ergebnisse beobachten und dadurch ist $\tau_i = Y_{1i} - Y_{0i}$ nicht berechenbar

Kurzer Reminder: mathematisch / statistische Notation

- Informell E kann interpretiert werden als “beste Schätzung”
  - Zumeist ist damit der Bevölkerungsmittelwert gemeint: die Erwartung (Expectation) von Y

Definitionen von Erwartungen

- Lasse Y_i eine zufällige Variable sein die über die Beobachtungen $i = 1, \dots, N$ hinweg Werte annimmt. Dann ist $E[Y_i]$ der Durchschnitt in der Bevölkerung für diese Variable. Zwei Beispiele:
  - Wenn Y_i durch einen Zufallsprozess erzeugt wird (z. B. durch Würfeln), ist $E[Y_i]$ der Durchschnitt in unendlich vielen Wiederholungen dieses Prozesses
  - Wenn Y_i aus einer Erhebung stammt, ist $E[Y_i]$ der Durchschnitt, den man erhält, wenn alle Personen der Grundgesamtheit, aus der die Stichprobe gezogen zusammengezählt werden
- Der erwartete Wert von Y gegeben X wird mit einem “|” Symbol angezeigt
- Alles zusammen kann demnach die Notation $E(Y|X = x)$: als. Konditionale Erwartung von Y gegeben der Wert von X ist x gelesen werden

ATE und Schätzer

- Statt den Effekt einer Person verwendet man den durchschnittlichen Treatmenteffekt
- Average Treatment Effect (ATE):

$$\mathrm{ATE} = \mathrm{E}[\delta] = \mathrm{E}[\mathrm{Y}^1] - [\mathrm{Y}^0]$$

- Aber auch das ist nicht beobachtbar!
- Fehlende Information muss daher mit Annahmen ersetzt werden
- Geschätzt wird der NATE
  Naive Average Treatment Effect (NATE)

$$\mathrm{NATE} = \mathrm{E}[\mathrm{Y}^1 | \mathrm{D} = 1] - [\mathrm{Y}^0 | \mathrm{D} = 0]$$

- Dabei zwei Möglichkeiten
  - Vergleich verschiedener Individuen in T und C zu demselben Zeitpunkt
    - $\widehat{\mathrm{ATE}}$ = Differenz der Gruppenmittelwerte
    - Zentrale Annahme: Homogenität der Individuen (compare like with like)
  - Vergleich derselben Individuen zu unterschiedlichen Zeitpunkten in T und C
    - $\widehat{\mathrm{ATE}}$ = Mittelwert der Vorher/Nachher Differenzen
    - Zentrale Annahme: Stabilität Y^0 über Zeit
    - Schätzung nur mit treated Individuen möglich!

Kausaleffekte auf einen Blick

- Average Treatment Effect (ATE)

$$\mathrm{E}[\delta] = \mathrm{E}[\mathrm{Y}^1] - [\mathrm{Y}^0]$$

- Average Treatment Effect on the Treated (ATT)

$$\mathrm{E}[\delta \mid \mathrm{D} = 1] = \mathrm{E}[\mathrm{Y}^1 \mid \mathrm{D} = 1] - [\mathrm{Y}^0 \mid \mathrm{D} = 1]$$

- Average Treatment Effect on the Untreated (Control) (ATC)

$$\mathrm{E}[\delta \mid \mathrm{D} = 0] = \mathrm{E}[\mathrm{Y}^1 \mid \mathrm{D} = 0] - [\mathrm{Y}^0 \mid \mathrm{D} = 0]$$

- Naive Average Treatment Effect (NATE)

$$\mathrm{E}[\mathrm{Y}^1 \mid \mathrm{D} = 1] - [\mathrm{Y}^0 \mid \mathrm{D} = 0]$$

rot: ein kontrafaktisches (unbeobachtbares) Ergebnis

Interpretation of Treatment Effects

Think of the following example: The effect of Medicaid on Americans' health status.

- ATE
  - The effect of Medicaid on the health status of the average American.
- ATT
  - The effect of the Medicaid on the health status for those who are actually enrolled. Would those who subscribed to Medicaid have a better or a worse health condition if they had not subscribed?
- ATC
  - The effect of the Medicaid on the health status for those who are not enrolled. Would those who have not subscribed to Medicaid have a better or a worse health condition if they had subscribed
- Example: What would be the ATE, ATT & ATC of **Azzollini (2023):**
  - Imagine this universal result: Individuals with unemployment scars are less likely to vote than individuals without those scars.

https://onlinelibrary.wiley.com/doi/full/10.1111/1468-4446.13039

Research Designs

- Die Beobachtung von Kausaleffekten ist also immer annahmebasiert!
- Ziel von Research Designs: Erforderliche Annahmen möglichst plausibel machen
- Damit möglichst zuverlässige Identifikation von (Kausal-)Effekten
- Zudem: Auch zuverlässige Deskriptionen erfordern gute Designs

- Wichtig für kausale Fragestellungen
  - Ausschluss alternativer Erklärungen
    - Unterschiede zw. C, T nur durch Treatment
    - Somit keine Konfundierung des Treatments
  - Keine Messfehler
- Wichtig für deskriptive Fragestellungen
  - Stichprobe die auf Population schließen lässt
  - Keine Messfehler

In den Sozialwissenschaften sind Kausalanalysen besonders komplex

- Vorausschauendes Handeln
- Strategisches Antwortverhalten
- ….

nature human behaviour

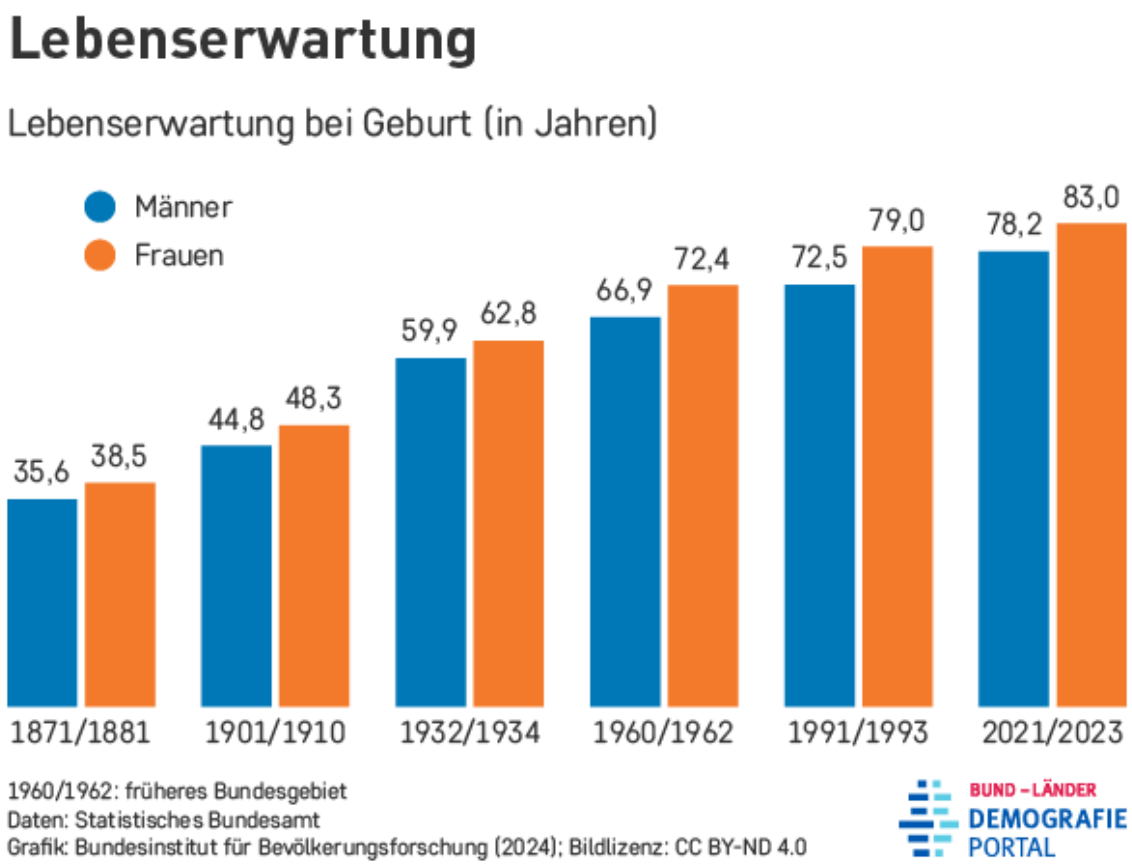
Review article

https://doi.org/10.1038/s41562-024-01939-z

Causal inference on human behaviour

Beispiel: Bewirkt wirklich das biologische Geschlecht unterschiedliche Lebenserwartungen?

- Wie könnte man das herausfinden?
- Was sind hier mögliche alternative Erklärungen?
- Welches Design könnte diese bestmöglich ausschließen?



https://www.demografie-portal.de/DE/Fakten/lebenserwartung.html

Eine Möglichkeit: „Klosterstudien“

- Vergleich von Frauen und Männern mit möglichst ähnlichen Lebensbedingungen
  - Tagesablauf
  - Ernährung
  - Tätigkeiten
  - …
- Schlussfolgerung Luy (2011)
  - Größter Teil männlicher Übersterblichkeit nicht auf biologische Unterschiede rückführbar
  - Verbleibende Confounder
    - Männern (aber nicht Frauen) ist in Klostern gestattet zu rauchen;
    - etwas andere Tätigkeiten (mit untersch. Ausmaß Mobilität per Auto)
    - ??

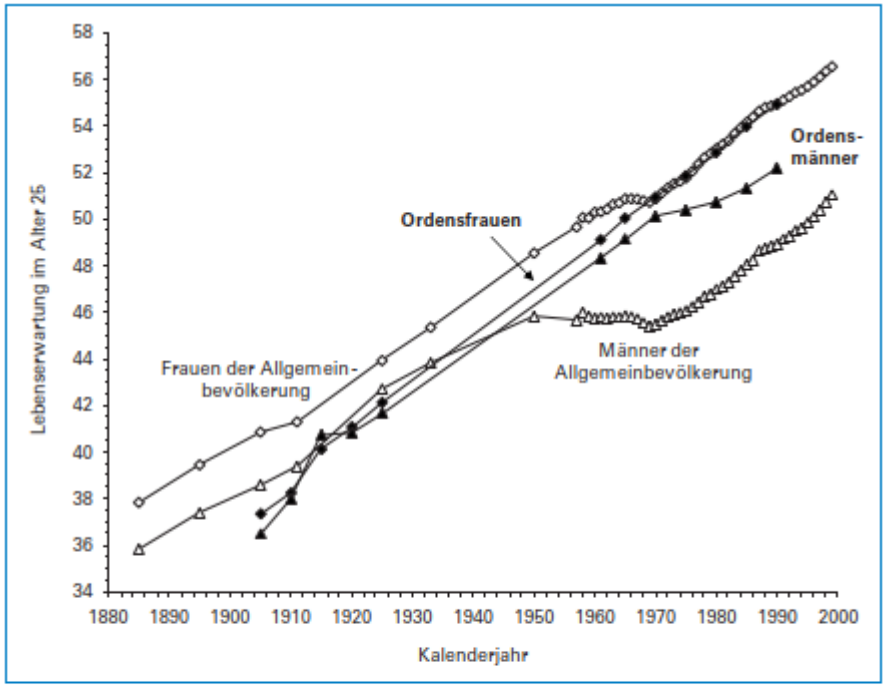


Abbildung 1
Fernere Lebenserwartung im Alter 25 für bayrische Kloster- und westdeutsche Allgemeinbevölkerung, 1880–2000. Die Beobachtungszeiträume für die Perioden-Sterbetafeln umfassen bei den westdeutschen Frauen und Männern drei (vor 1910 zehn) und bei den Ordensmitgliedern jeweils 30 Kalenderjahre; die markierten Punkte repräsentieren die Mitte der Beobachtungszeiträume (eigene Berechnungen mit Daten der Klosterstudie; Daten der Allgemeinbevölkerung: Statistisches Bundesamt Wiesbaden).

Marc Luy (2011): S. 581

Was will man eigentlich wissen?

Gute Forschungsfragen

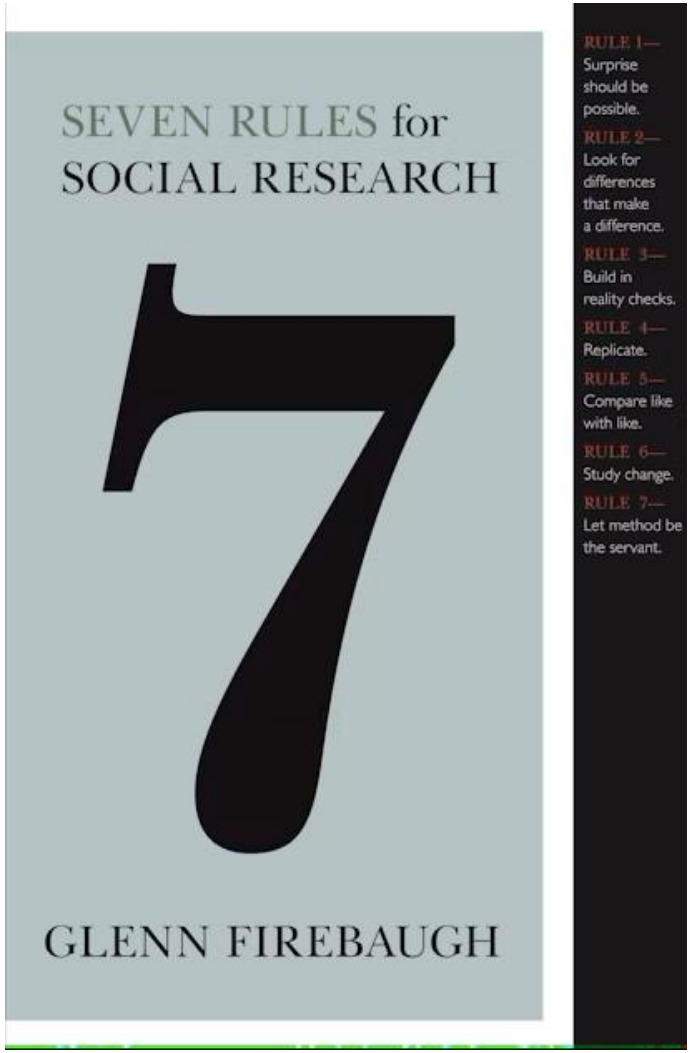
“We're great at giving answers without knowing the question”.
(Felix Elwert).

Was sagen Sie zu den folgenden Forschungsfragen?

1. Die Studie zielt darauf ab, die Einstellungen linkshändiger Personen zum Gender Wage Gap zu erforschen. Das ist relevant, weil das bislang noch gar nicht erforscht wurde, und es eine größere Gruppe an linkshändigen Personen gibt.
2. Die Studie will wissen, welche Aspekte Radikalisierung fördern. Das ist ein wichtiges Thema für Gesellschaften, weil Extremismus zunimmt.
3. Wir wollen fremdenfeindliche Einstellungen erforschen.

Auch Sie brauchen gute Forschungsfragen!

- Etwa für die Masterarbeit
- Literaturempfehlungen dazu:
  - Firebaugh, Kap. 1
  - Van Tubergen, Kap. 1
  - de Vaus, Kap. 2



„Gute“ Fragestellungen für Forschungsvorhaben

Interessant und relevant

- Beantwortung der Frage bringt Erkenntnisgewinn
  - für die Wissenschaft (und/oder
  - evidenzbasierte Politik)
- Somit: von allgemeinem Interesse
  - Damit: i.d.R. gerade nicht ein völlig neues Forschungsgebiet
  - Sondern Erweiterung bestehender Forschung
    - Ist diese intern valide/robust?
    - Vertiefung durch Prüfung von Mediatoren
    - Erweiterung / externe Validität: andere Subgruppen, Moderatorvariablen

Präzise und praktikabel

- Es muss klar sein
  - Was soll herausgefunden werden?
  - Auf welche Population will man verallgemeinern?
- Ohne präzise Fragen ist die Relevanz, der Fokus, die relevante Literatur unklar
  - Alles könnte relevant sein oder nicht
  - Es lässt sich kein Forschungsdesign entwickeln und beurteilen
  - Letztlich gewinnt man dann Antworten ohne Frage

Präzise

- Setzt immer klare Definitionen und Operationalisierungen von Konzepten voraus!
- Zudem:

Deskriptionen

- Interessierende Population: Raum- & Zeitbegrenzung
- Analyseeinheit
- Interessieren Parameter insgesamt und/oder für Subgruppen

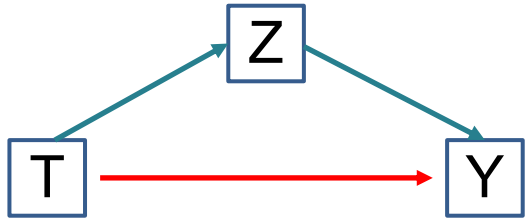
Kausale Fragen

- Ursachen – alle oder eine? Folgen?
- Was sind das Treatment und Outcome? Interessieren direkte, indirekte, oder totale Effekte? (S. später) Eine oder rivalisierende Erklärung(en)?
- Idealerweise Präzisierung in Diagramm!
- Zudem angestrebte Verallgemeinerung auf welche Population?

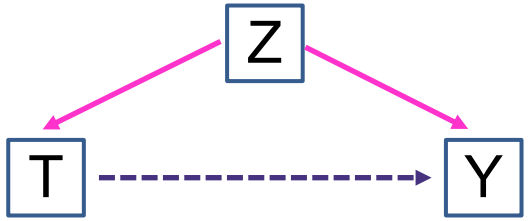
Diagramme: Unterschiedliche (Kausal-)Effekte

- Totaler Effekt

- Ergibt sich als Summe aus:
  - indirekter Effekt
  - direkter Effekt



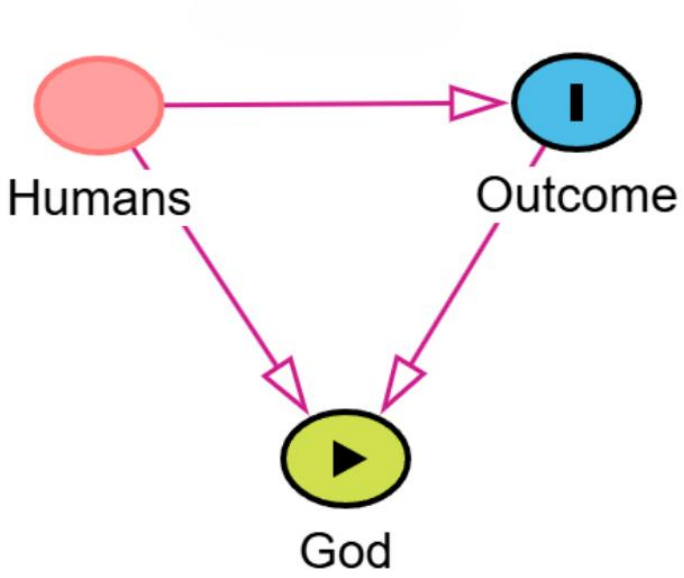
- Confounder



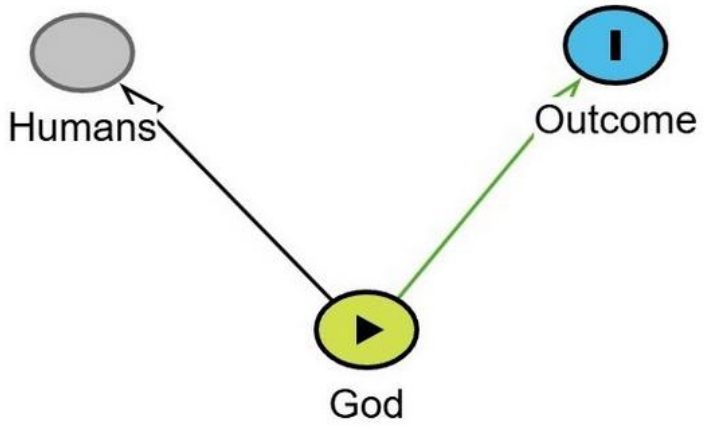
- Z.B Effekt von Gender auf Lohn: Unbereinigte Lücke
- Effekt von
  - Bildung, Berufswahl, Arbeitszeiten und anderen gewählten Arbeitsmarktfaktoren
  - Gender (bereinigte Lücke, oft als Messung von „Diskriminierung“ gedeutet)
- Für Gender schwierig Beispiele zu finden! Bei anderen Fragen aber i.d.R. sehr plausibel

Does God Matter?

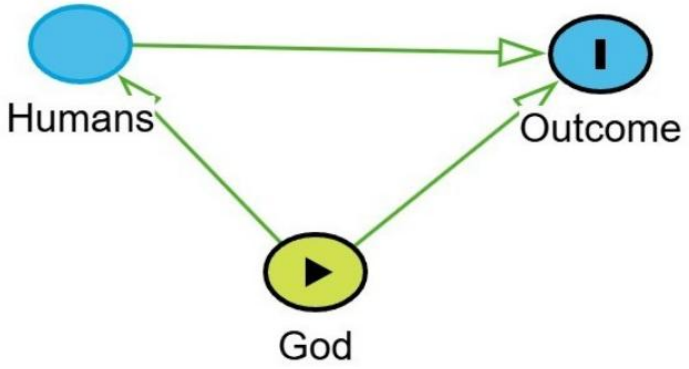
The Atheist



Laplace‘s Demon



The Calvinist



https://www.linkedin.com/pulse/does-god-matter-depicting-belief-systems-through-matt-422zf/

Was untersuchen diese Forschungsfragen?*

- Sind hier die X- und Y-Variablen klar? Direkte oder totale Effekte? Ist es eine kausale Fragestellung? Was wäre ggf. noch zu konkretisieren?

1. ‘Why are some people happier than others?’
2. ‘Does happiness predict a person’s later health status?’
3. ‘What are the economic returns to completing an additional year of schooling?

*Quelle: Bailey et al. 2024. Causal inference on human behaviour. Nature human behaviour. https://doi.org/10.1038/s41562-024-01939-z

Dennoch falsch?

Alternative Erklärungen, Falsifikation
statt Konfirmation

Gute Designs antizipieren alternative Erklärungen (um sie dann möglichst gut zu eliminieren)

Theoretisch/substanziell

- Mögliche Confounder
- Ideen dafür: Literatur, Theorien, Nachdenken
- Zudem sollte der angenommene Effekt selbst plausibel sein (plausibler Mechanismus) – s. auch nächste Folie
  - Besser: theoriegestützte Forschung! (Theorie: Antwort auf Warum-Frage)
  - Und nicht einfach Testung aller möglichen Interventionen

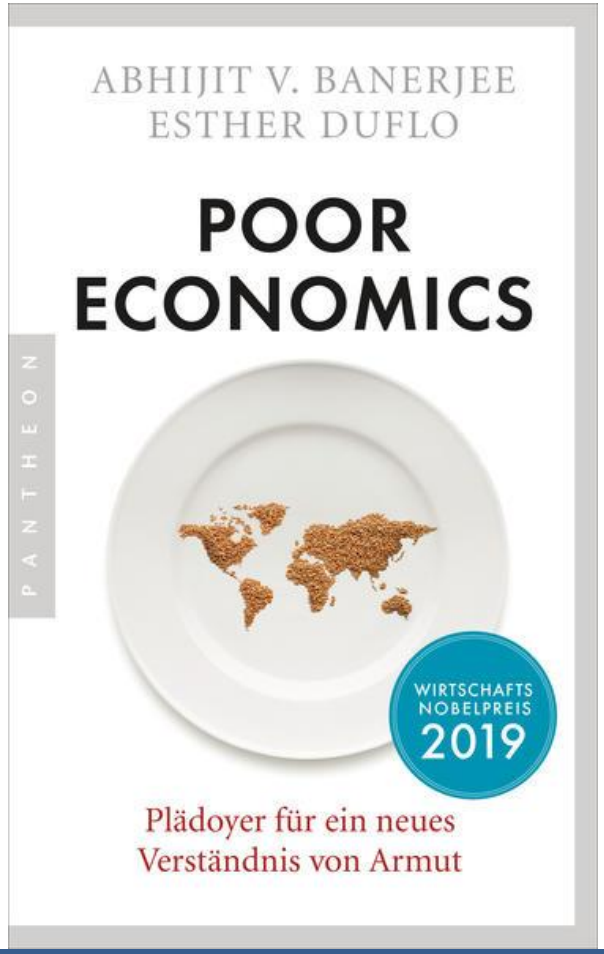
Messfehler/technisch

- Viele Fehlerquellen möglich
  - Verzerrte Samples
  - Falsch verstandene Fragen
  - Unaufmerksame Befragte
  - Falsch kodierte Antworten
  - Analysefehler, Reporting Errors
  - ….
- Zufällige und systematische Fehler
- Vermeidung vorab, aber auch Testung für vorhandene Studien ([Re-]Analysen, Replikationen)

Forschung basierend auf möglichst präzisen Theorien bringt mehr Erkenntnisgewinn

- Viele Studien sind „Effekt-Experimente“
  - Es wird ein (kurzfristiger) Effekt (Intervention) getestet
  - Z.B. rote vs. grüne Tafeln in Klassenzimmern
  - Vergünstigte Anti-Mücken Netze*
- Auch wenn ein Effekt valide untersucht wurde, kann der Erkenntnisgewinn gering sein
  - Das ist z.B. die Kritik an Interventions-studien durch die „Randomistas“
  - Was könnte hier das Problem sein?

- S. dazu auch:



Probleme reiner Effekt-Experimente

Verallgemeinerbarkeit?

"The problem of generalizing from one concrete instance to another depends on a theory: it requires definition of what constitutes an instance of the phenomenon and of its scope — that is, the **conditions under which a result is or is not applicable**. Effect experiments and programs, because they do not define the instantiation and scope conditions of their effects, leave the question of their generalizability unanswered."
(Zelditch 2014: 191)

Policy Implikationen?

Hoher Grad an Abstraktion
(oft „fat-handed interventions" – Maßnahmebündel)

+

Zweifelhafte Verallgemeinerbarkeit

=

Wenig Wissensgewinn für Policy Interventionen

(unklare Effektivität, d.h. hohes Risiko)

Parachute Use to Prevent Death and Major Trauma When Jumping from Aircraft:
Randomized Controlled Trial

"**Conclusions:** Parachute use did not reduce death or major traumatic injury when jumping from aircraft in the first randomized evaluation of this intervention. However, the trial was only able to enroll participants on small stationary aircraft on the ground, suggesting cautious extrapolation to high altitude jumps."

Yeh, Robert W., et al. 2018. 'Parachute Use to Prevent Death and Major Trauma When Jumping from Aircraft: Randomized Controlled Trial'. *BMJ* 363:k5094. doi: 10.1136/bmj.k5094.

Besser testet man also „Theorien“

- Menge miteinander verknüpfter Hypothesen zu einem Phänomen, das mit bestimmter Regelmäßigkeit auftritt (Prozess)
  - ***Wann*** tritt ***warum*** ein Effekt auf

Scope-Conditions
Externe Validität

Mechanismus/kausaler Effekt,
Interne Validität

- Erklärungen, statt reine Variablenassoziationen
- Und: Mehr als Definitionen!

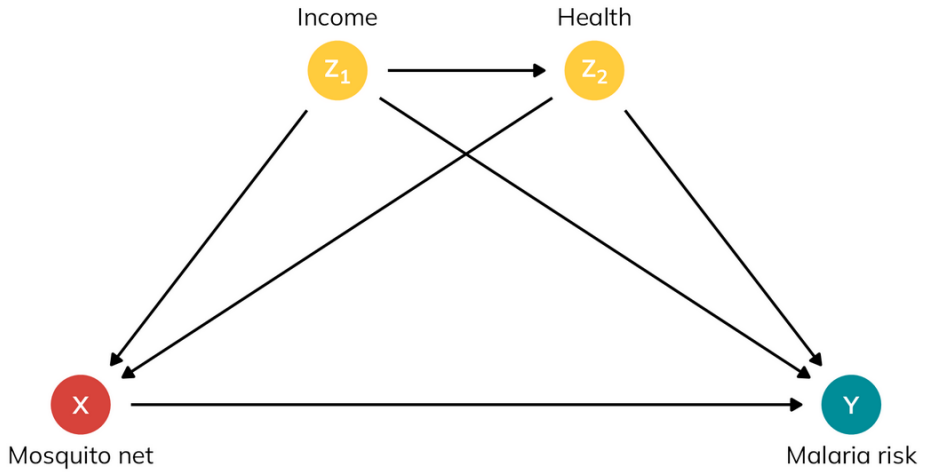
- Keine Theorien: Kategorienschemata, Definitionen, reine Beschreibungen etc.

Warnung: Karikaturen können Spuren von Satire enthalten und sind für Andersdenkende nicht geeignet.

Nelcartoons

Nur präzise Theorien sind testbar (falsifizierbar)

- Widerspruchsfrei / nicht tautologisch
- Klare Definitionen
- Idealerweise Formalisierung
  - Graphische Darstellung (z.B. DAG)
  - Logik
  - Mathematischer Zusammenhang
- Bei zugleich Abstraktion von unwesentlichen Details



https://www.andrewheiss.com/blog/2024/03/21/demystifying-ate-att-atu/

Gute Forschungsdesigns = Falsifikationsversuche

- Es ist relativ einfach, positive Evidenz für Theorien zu finden
  - Beispiel: Die Modernisierungstheorie besagt, dass primär finanziell deprivierte Menschen AfD wählen.
  - Beobachtung: Viele Geringverdiener wählen die AfD.
  - Ist das ein überzeugender Theorietest?
- Theorien lassen sich nie endgültig beweisen
- Besonders glaubwürdig sind Theorien, die möglichst viele Anstrengungen überstanden haben, sie zu widerlegen
  - Bei widersprüchlicher Evidenz:
    - Prüfung auf Messfehler und Weiter-/Neuentwicklung von Theorie
    - Dann erneute Testung

Konfirmation vs. Falsifikation

Konfirmation

- Annahme: Falls A, dann B
  (z.B.: Private Schulen bringen Schülern mehr bei, daher bessere Noten)
- Man beobachtet für A auch B
  (Bessere Noten in privaten statt öffentlichen Schulen)
- Somit: Die Annahme stimmt?
  Private Schulen sind lehrreicher?
- Problem: Beobachtetet wird nur:
  Falls A [oder konfundierte Aspekte]

Falsifikation

- Falls A, dann sollte B folgen
- B folgt *nicht*
- Somit: A ist nicht korrekt.

Zusammenfassende Evaluierung - Mehr oder weniger gute Designs:

Gütekritierien

Gütekriterien: Was ist „gute“ Empirie?

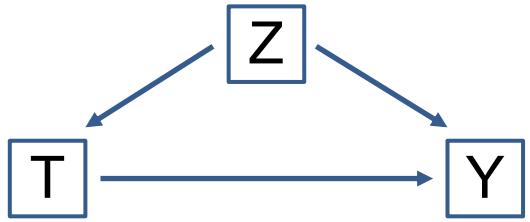
- Und damit auch ein gutes Forschungsdesign
- Welche Kriterien werden hierfür üblicherweise genutzt? Ideen?
- Gut: = „valide“
  - Interne Validität: Wird der Treatmenteffekt/ die Zielgröße richtig identifiziert?
  - Externe Validität: Ist der Effekt/der Zielwert verallgemeinerbar?
    - Andere Populationen/Zeitpunkte/Umgebungen, Operationalisierungen
  - Vermeidung von Messfehlern
    Bilden die Messungen die Zielparameter valide ab?

Beispiele für Bedrohungen?

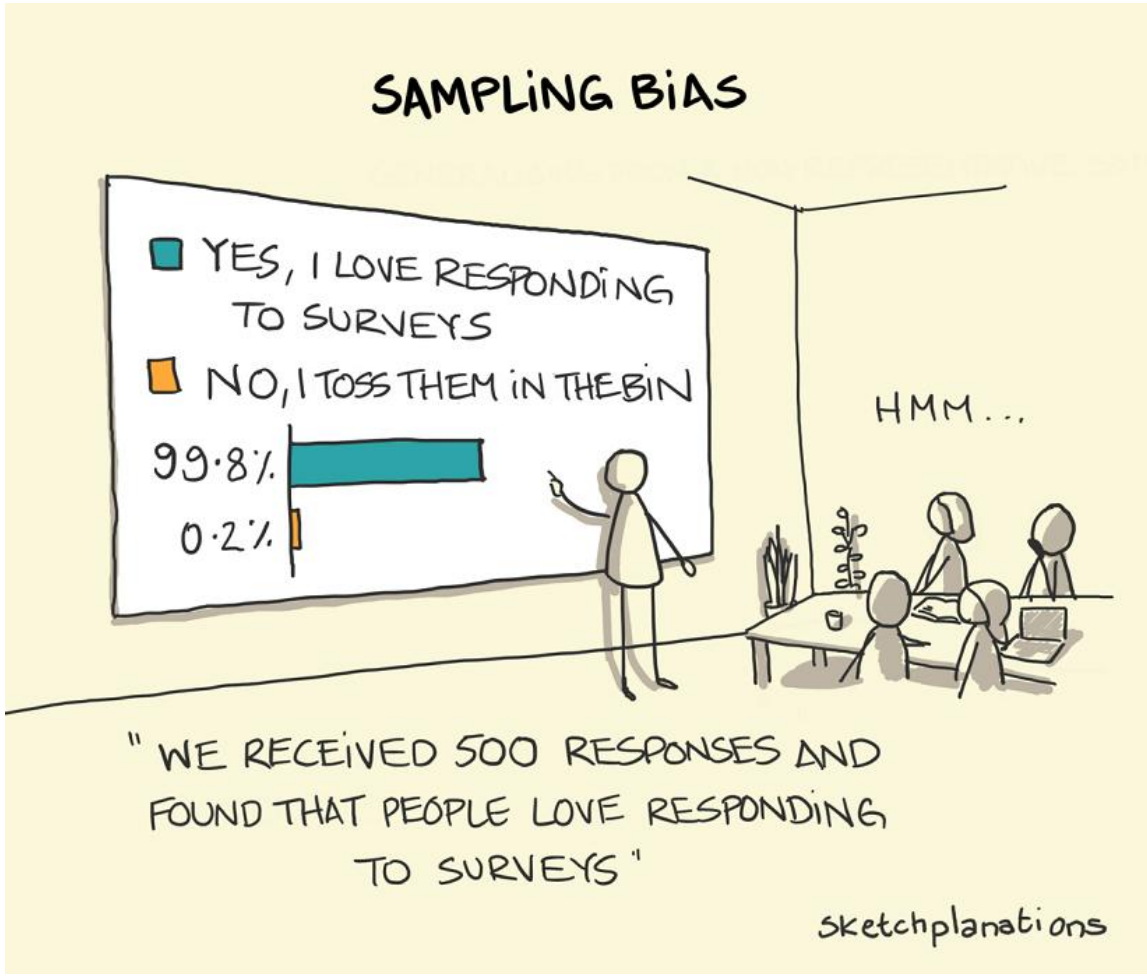
- (Gütekriterien umfassen oft auch gute Theorien/Argumentationen – bei Interesse: Otte et al. 2023: "Gütekriterien in der Soziologie: Eine analytisch-empirische Perspektive“. Zeitschrift für Soziologie 52 (1): 26-49.)

Bedrohungen Interner Validität

- Confounder / Scheinkorrelationen

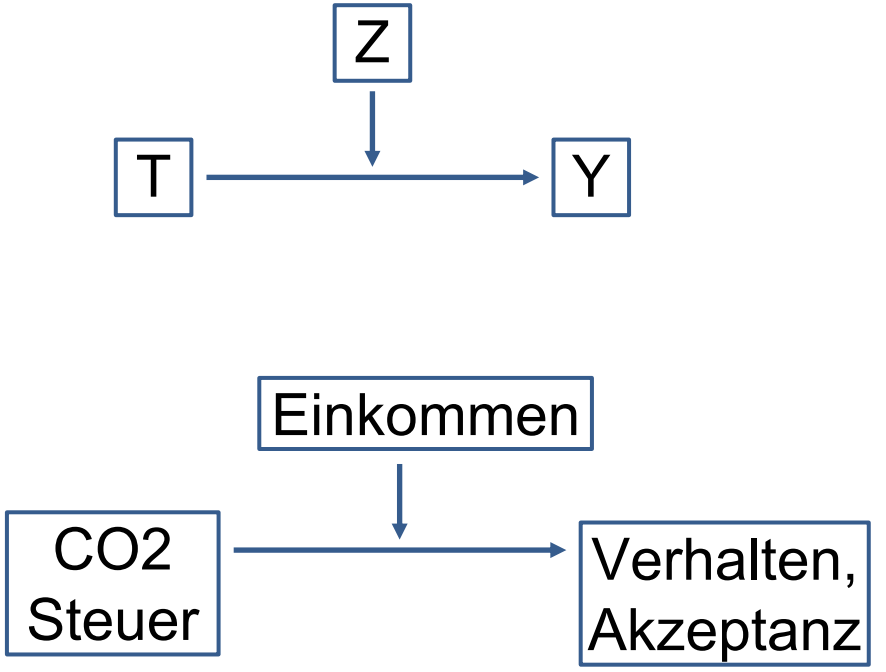


- Möglicher Grund: selektive Teilnahmen / Abbrüche bei Studien – diese unterminieren auch valide Deskriptionen
- Treatmenteffekt misst anderes Konstrukt/Mechanismus
  - Z.B. Soziale Erwünschtheit anstatt Diskriminierung
  - Handlungsabsichten ≠ Handeln
  - …



Externer Validität

- Sample/Setting unterscheidet sich von Zielpopulation in Eigenschaften, die den Treatmenteffekt beeinflussen
  - Systematische Moderatoreffekte und Zufallsschwankungen
  - Beispiel: Es nehmen primär nur Personen aus der Mittelschicht an Umfragen teil
  - Effekte von Treatments wie CO2 Steuern werden damit falsch geschätzt (z.B. Unterschätzung der Auswirkung auf Niedrigeinkommensgruppen)

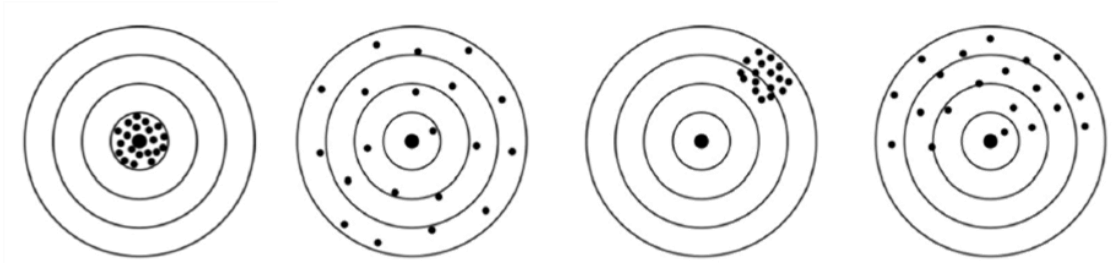


Confounder / Scheinkorrelationen

Messfehler

- Sind spezielle Bedrohung der internen Validität (es wird etwas anderes gemessen als angestrebt)
- Beeinträchtigen reliable und/oder valide Messungen. Was war das nochmal?

➢Was symbolisiert Reliabilität, was Validität?



Sehr reliabel und sehr valide (ideale Messung)

Valide, aber nicht sehr reliabel

Reliabel, aber nicht sehr valide

Nicht sehr reliabel und nicht sehr valide

Quelle: Eigene Darstellung in Anlehnung an Babbie 1986 und De Suouza et al. 2017

Zusammenfassend Forschungsdesigns

- Ziel: Möglichst überzeugende Antwort auf Forschungsfrage: Was/warum ist etwas der Fall?
- Bei Kausalfragen: Qualität kausaler Inferenz verbessern

- Forschungsdesigns sind umso besser,
  - Je mehr alternative Erklärungen (Theorien) sie ausschließen
    - Ausschluss von Korrelaten bzw. Confoundern (hohe *interne Validität*)
    - Ausschluss von Messfehlern
  - Je allgemeingültiger die mit ihnen gewonnenen Antworten sind (hohe *externe Validität*)

Anhang Forschungsfragen: Checks, Empfehlungen

Interessant und relevant

- Denkt man sich: „So what“? Who cares?
- „Novel“ – ja, aber nicht im Sinne, dass man der Erste ist der/die hier forscht
  - Das wären Sie z.B. auch, wenn Sie die Einstellungen Ihrer Tante untersuchen. But: who cares?
  - Alles, was noch NIE erforscht wurde, ist vermutlich nicht allgemein relevant.

➢Wie finde ich relevante Fragen?

- Literatursichtung, insb. auch Fazit von Artikeln mit Limitationen / Ausblick
- Kritischer Blick beim Lesen: Ist das alles unzweifelhaft robust? Valide?

Präzise und praktikabel

- Ist klar, was herausgefunden werden soll?
- Welche Literatur passend ist?
- Ist das (im Umfang) machbar?
- Können Sie Fragestellung und Relevanz in 1-2 Sätzen Anderen klar erläutern?

➢I.d.R. sind Fragestellungen zu groß und vage

- Bereits eine kleine überlegte Neuerung im Design ist i.d.R. sehr aufschlussreich
- Das kann in Form von Replikationen, Reanalysen, oder ggf. auch eigenen kleinen Datenerhebungen erfolgen

Experimente als Goldstandard, Validität

https://commons.wikimedia.org/wiki/File:CDU_Wahlkampfplakat_-_kaspl019.JPG

Agenda

1. Experimente als Königsweg für die Identifikation von Kausaleffekten
2. Arten von Experimenten

Einschub: Warum deskriptive Graphen auch bei Kausalen Forschungszielen wichtig sind

The solution

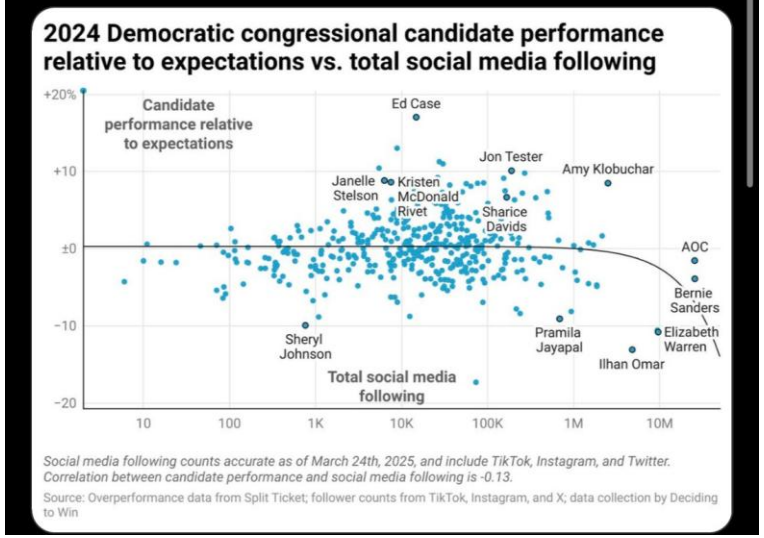
Lakshya Jain reposted

Simon Bazelon ✓ Follow
@simon_bazelon

Attention is not the problem.

The Democrats with the biggest social media followings did WORSE in 2024.

Our party's problems are substantive, not cosmetic.

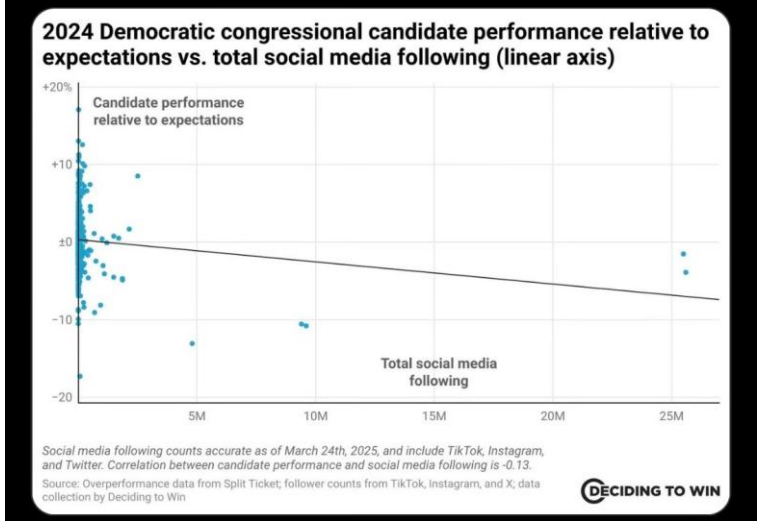


The problem

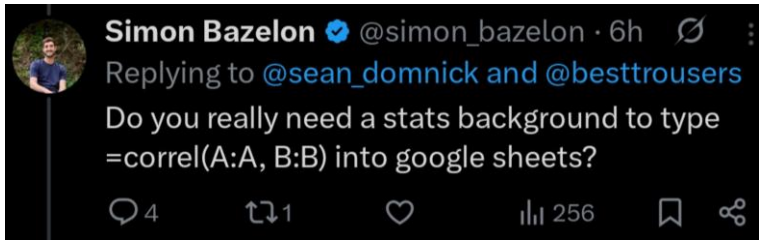
Simon Bazelon ✓ Follow
@simon_bazelon

Some people are mad about the x-xis in the graph below being logarithmic. I did that because otherwise it would be impossible to see most of the candidate names, not for any much deeper reason.

Here's the linear scale. I don't think this changes the general point at all!



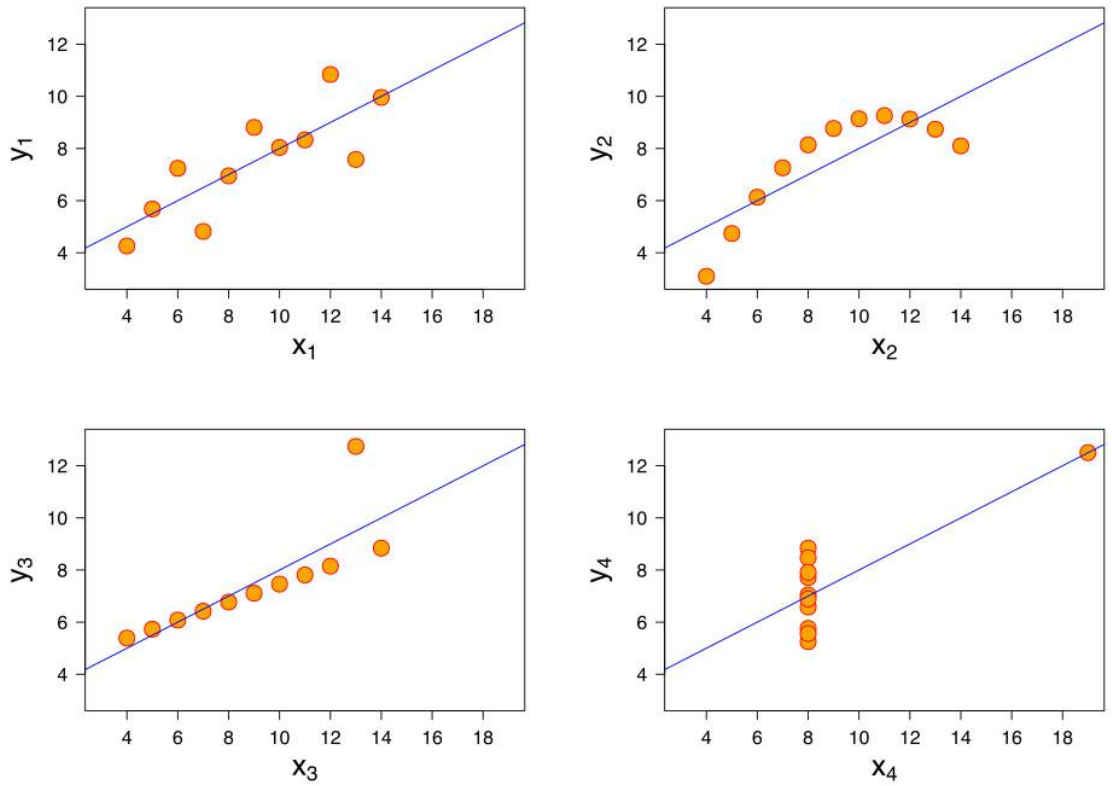
The misunderstanding

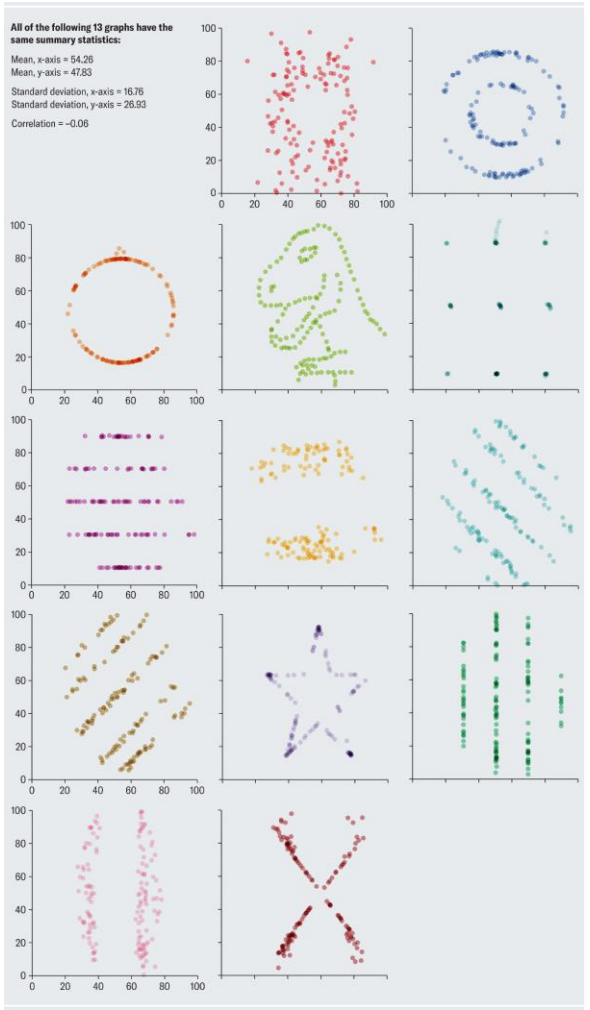


https://bsky.app/profile/sallylhudson.bsky.social/post/3m47h4ajsvk2u

Unterschiedliche Formen gleiche Korrelation

Anscombe's quartet

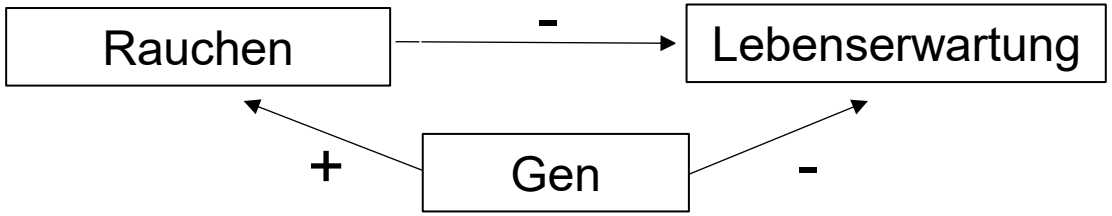




(Anscombe, 1973)
https://bsky.app/profile/sallylhudson.bsky.social/post/3m47h4ajsvk2u

Kausaler Effekt

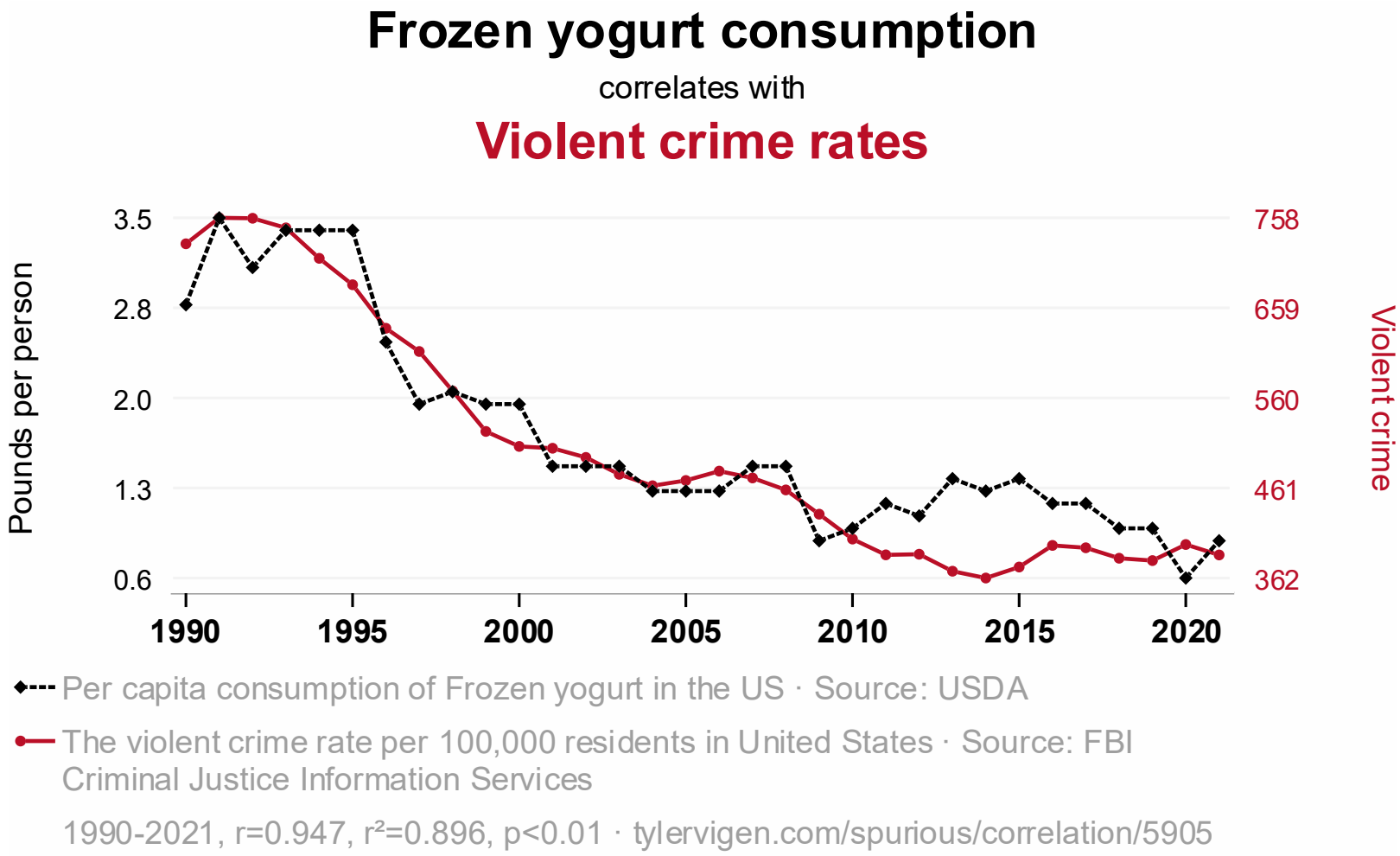
- Forschungsfrage: Verkürzt Rauchen die Lebenserwartung bei durchschnittlichen Personen der erwachsenen Allgemeinbevölkerung?
- Wie würden Sie versuchen, das herauszufinden?
- Könnte es nicht auch sein, dass es ein Gen gibt, dass gleichermaßen die Neigung zum Rauchen positiv beeinflusst und die Lebenserwartung negativ?
  Somit: Rauchen gar keinen kausalen Effekt hat? (Weil Personen mit dem Raucherneigungs-Gen ohnehin kürzer leben würden, egal ob sie rauchen oder nicht?



- Wie könnte man das ausschließen?

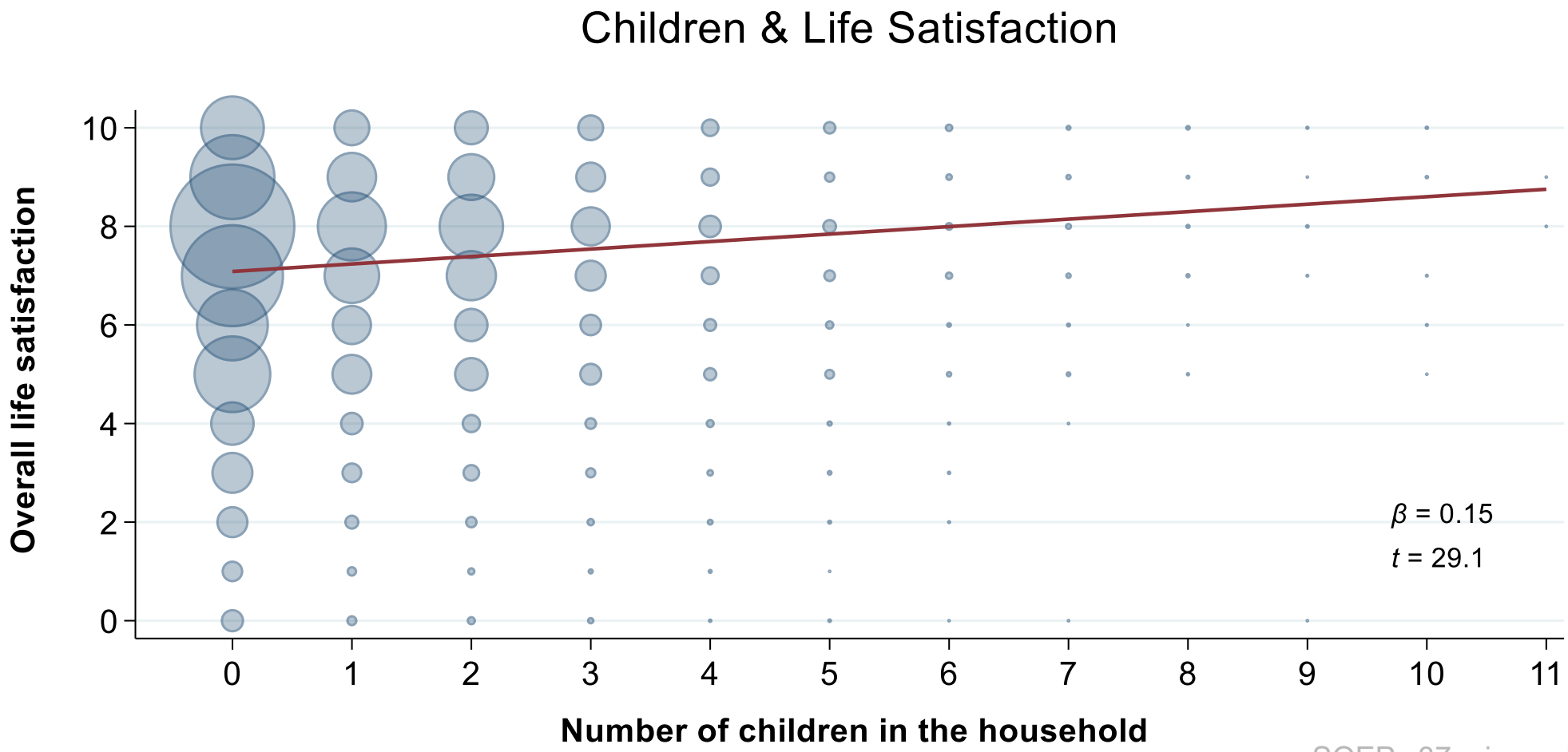

BREAKING
NEWS

Vorsicht vor „FroYo“!



Tyler Vigen

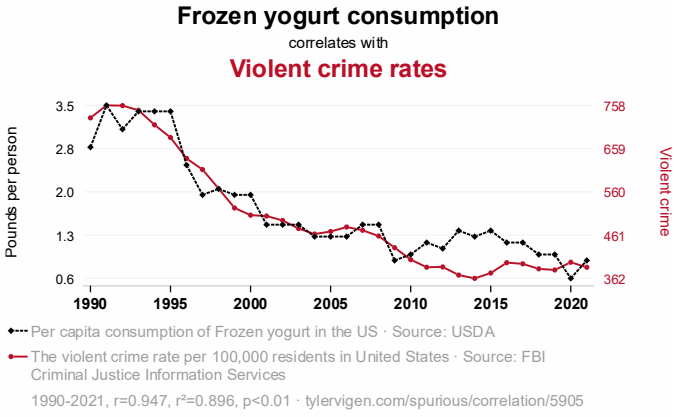
Glücklichsein leicht gemacht: Bekommen Sie Kinder!



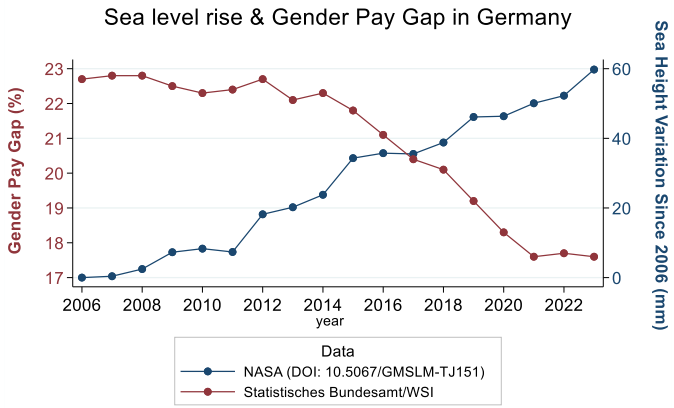
SOEP v37, eigene Darstellung

Korrelation ≠ Kausalität!

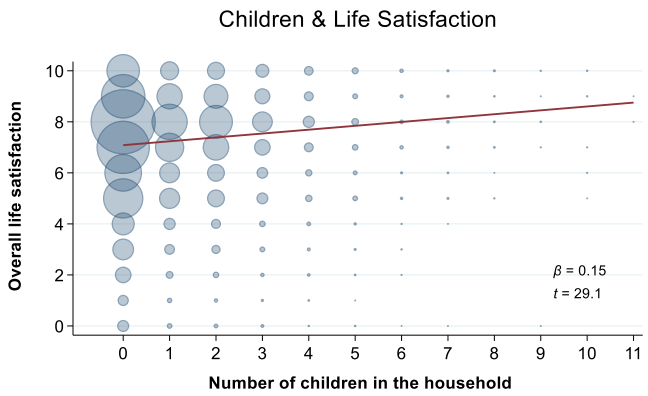
Zufall



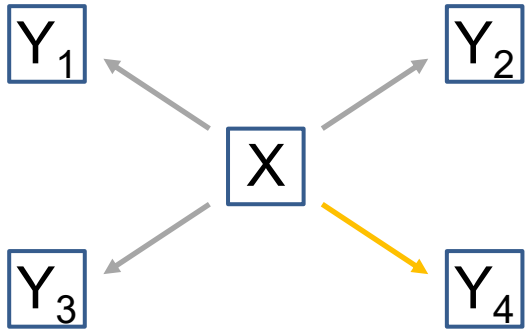
Scheinkorrelation



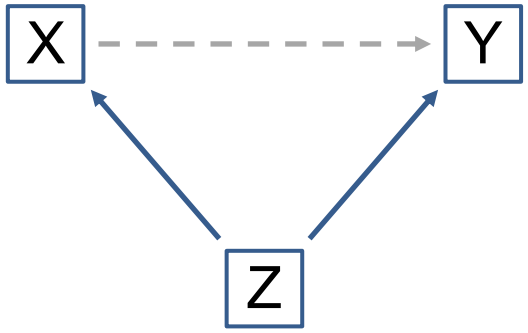
Umgekehrte Kausalität



Data Dredging



Confounding



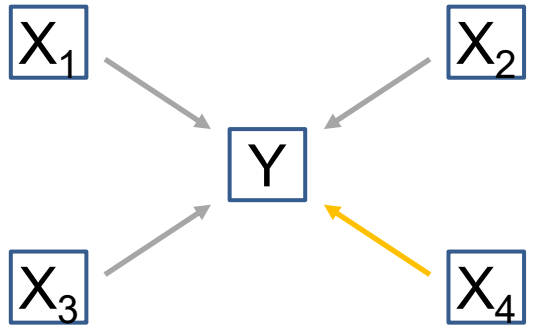
Reziprozität

Unvollständige Lösungen

Zufall

Testen von theoriegeleiteten Hypothesen

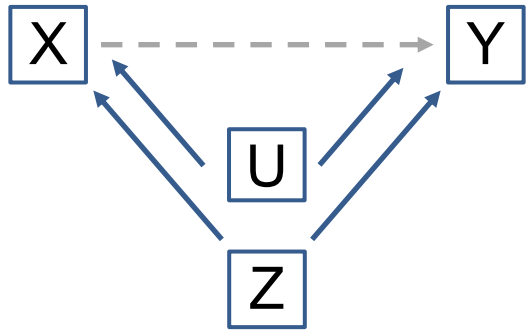
Problem: Ex-post Rationalisierung von Ergebnissen (p-hacking)



Scheinkorrelation

Kontrolle von potentiellen Confoundern

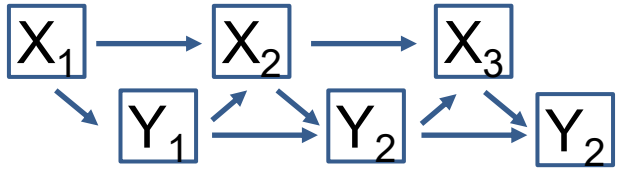
Problem: Viele Confounder werden nicht bedacht oder wurden nicht gemessen & Problem von Überkontrolle



Umgekehrte Kausalität

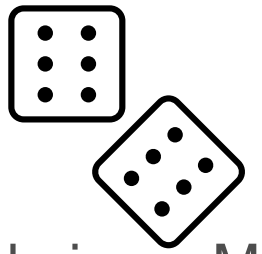
Messung vor und nach einem Ereignis

Problem: Teile der Effekte von X sind Folgeeffekte der Veränderung von Y (keine strikte Exogenität)



Experimente

1. Aufteilung der Beobachtungseinheiten in mindestens 2 Gruppen: Treatmentgruppe (T) vs. Kontrollgruppe (C)
2. „Verabreichung“ des Treatments (*Stimulus*) an Gruppe T durch Forschende
3. Zuteilung in T bzw. C erfolgt **zufällig** (*randomisiert*)



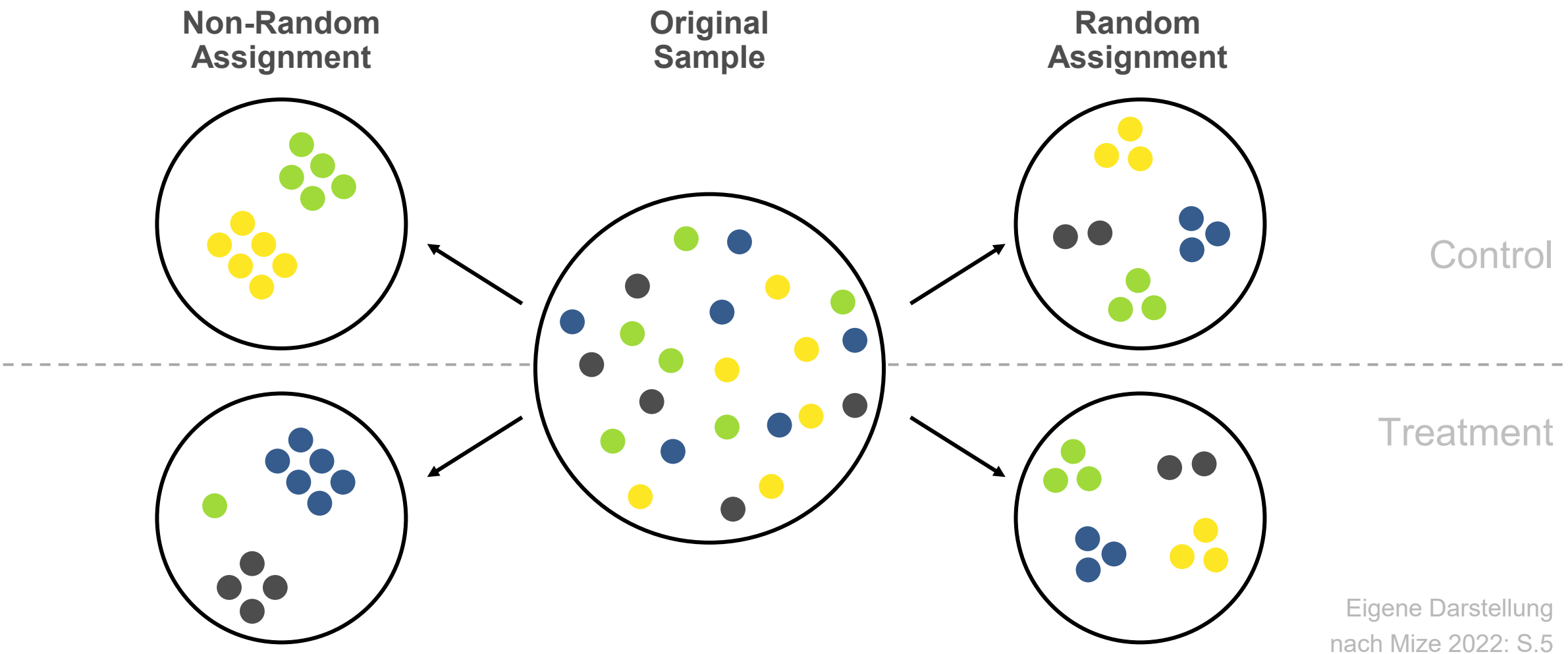
→ Durch Randomisierung unterscheiden sich T und C in keinem Merkmal systematisch

Potential outcomes sind unabhängig von D: $Y_{1i}, Y_{0i} \perp D_i$

→ Unterschiede im Outcome lassen sich daher eindeutig auf das Treatment zurückführen

→ Allerdings: **In der Praxis viele Bedrohungen dieser Annahmen!**

Randomisierung: Das Ideal



Eigene Darstellung
nach Mize 2022: S.5

Was bedeutet das für die “potential outcomes”

- Bei vollständig zufälliger Zuteilung des Treatments können 2 Stichproben (Beispiel Kontrollgruppe C und Treatmentgruppe T) als asymptotisch ident behandelt werden
- Das bedeutet auch die potentiel outcomes sind ident:

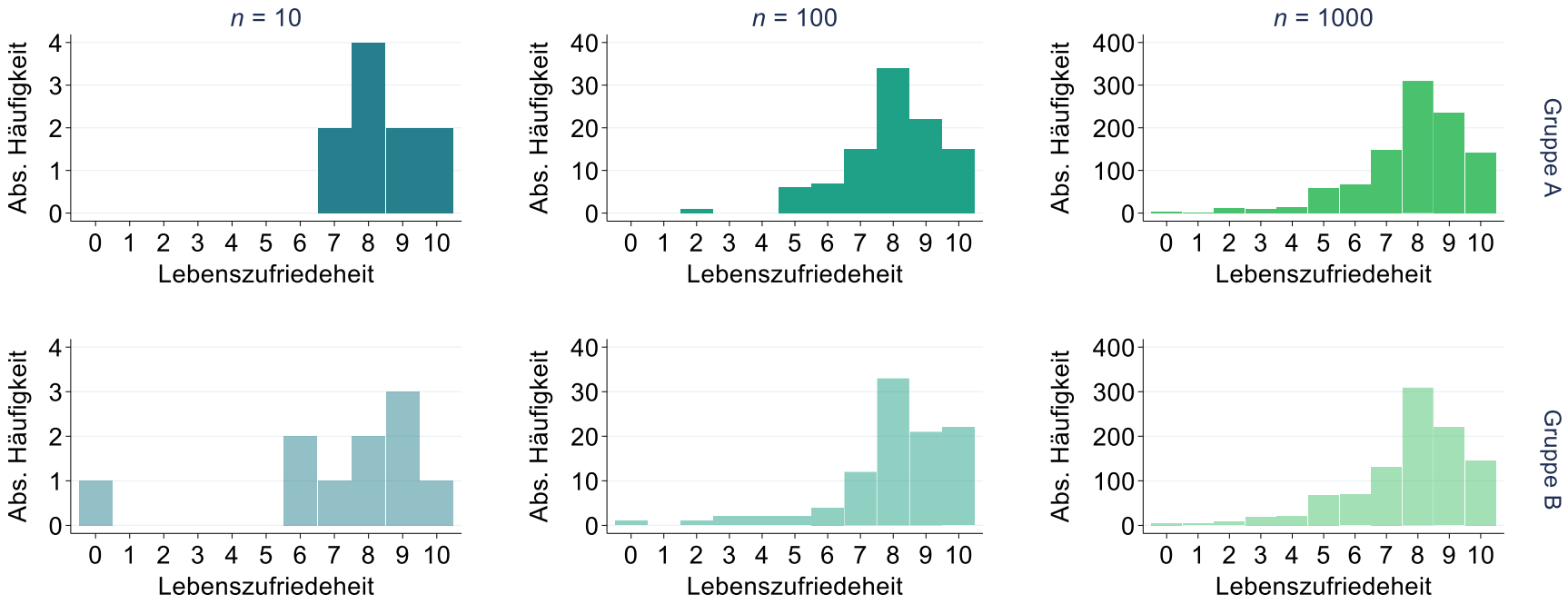
$$E[Y_{0,i}|i \in C] = E[Y_{0,i}|i \in T] = E[Y_{0,i}]$$
$$E[Y_{1,i}|i \in C] = E[Y_{1,i}|i \in T] = E[Y_{1,i}]$$

	Y_{1i}	Y_{0i}
Treatment Group $(\mathrm{D} = 1)$	Observable as Y	Counterfactual equivalent to Y for $D = 0$
Control Group $(\mathrm{D} = 0)$	Counterfactual equivalent to Y for $D = 1$	Observable as Y

- Aus diesem Grund können wir das fundamentale Problem der kausalen Inferenz (fehlende Daten) umgehen, da die potentiellen outcomes unabhängig davon sind ob Personen getreated wurden oder nicht.

Und die Praxis

Vollständige Kontrolle auf Heterogenität gelingt nur asymptotisch



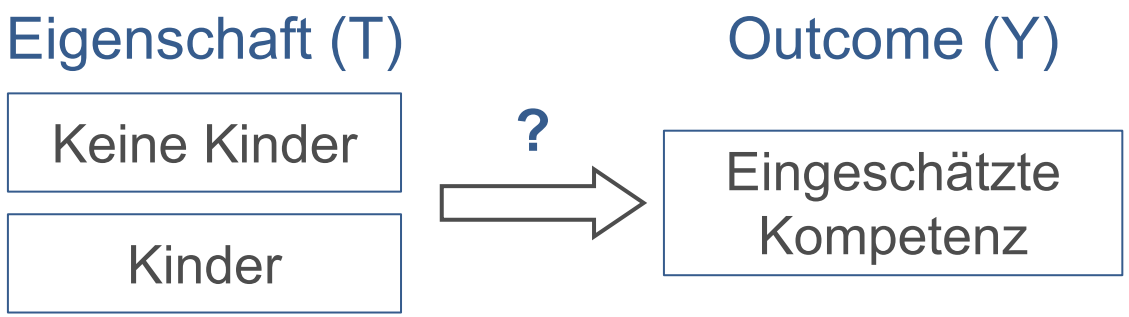
Quelle: ALLBUS 2018 (eigene Berechnungen)

Perfekte kontrafaktische Kausalität kann es (rein logisch) nie geben!

Terminologie

Wir verwenden lieber Y statt O
(Um zu verhindern: 0 vs. O, siehe Reisepässe)

- Beispiel: Werden Frauen mit Kindern weniger kompetent wahrgenommen als andere Bewerberinnen?*
  - Mutterschaft = Unabhängige Variable (X) bzw. *Treatment T*
  - Wahrgenommene Kompetenz = abhängige Variable (Y) bzw. *Outcome O*
- Test des *Treatmenteffekts* von T auf O: Erstellen von Lebensläufen mit bzw. ohne Kinder (= *Experimentalbedingungen*) und Bewertung durch Versuchspersonen



- Experimentelle Bedingungen sollten abgesehen von T möglichst ähnlich sein
- Oft: eine Kontrollgruppe C als „baseline“ (Abwesenheit von Eigenschaft)
- Aber: Auch Vergleich unterschiedlicher Treatments möglich!
- Wie könnte ein Experiment aussehen?

Was ist das Treatment?

Vorgehen

Paid undergraduate volunteers rated a pair of equally qualified, same-gender, same-race fictitious job applicants, presented as real, who differed on parental status.

Materialien

Participants inspected an applicant file for each of the two applicants.

The files contained: a short memo, a “fact sheet,” and a résumé.

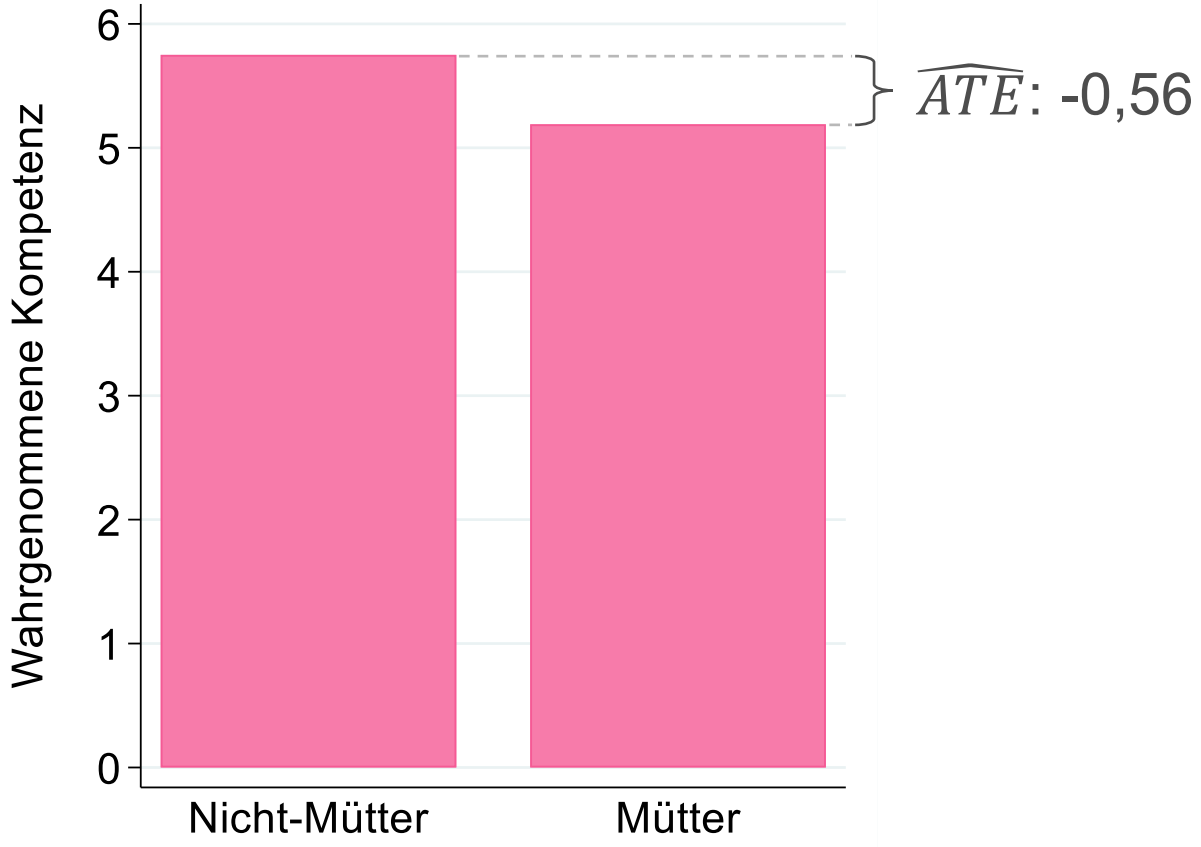
Treatment & Control

- The résumé for the parent member listed “Parent-Teacher Association coordinator”
- The memo for the parent member included the phrase: “Mother/father to Tom and Emily.”
- The control listed other unrelated activities and memberships

Bildbeispiel: Foster und Neugebauer 2024;
Text: Correll et al. 2007

Average Treatment Effect (ATE)

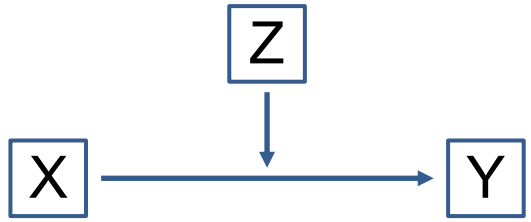
- Durchschnittlich erwarteter Effekt
- Ideal, das man oft identifizieren will
- Üblicherweise einfach als Differenz der Mittelwerte des Outcomes mit und ohne Treatment geschätzt (NATE), hier:
  $\widehat{ATE}$: = $\bar{Y}_{Kinder} - \bar{Y}_{Keine\ Kinder}$

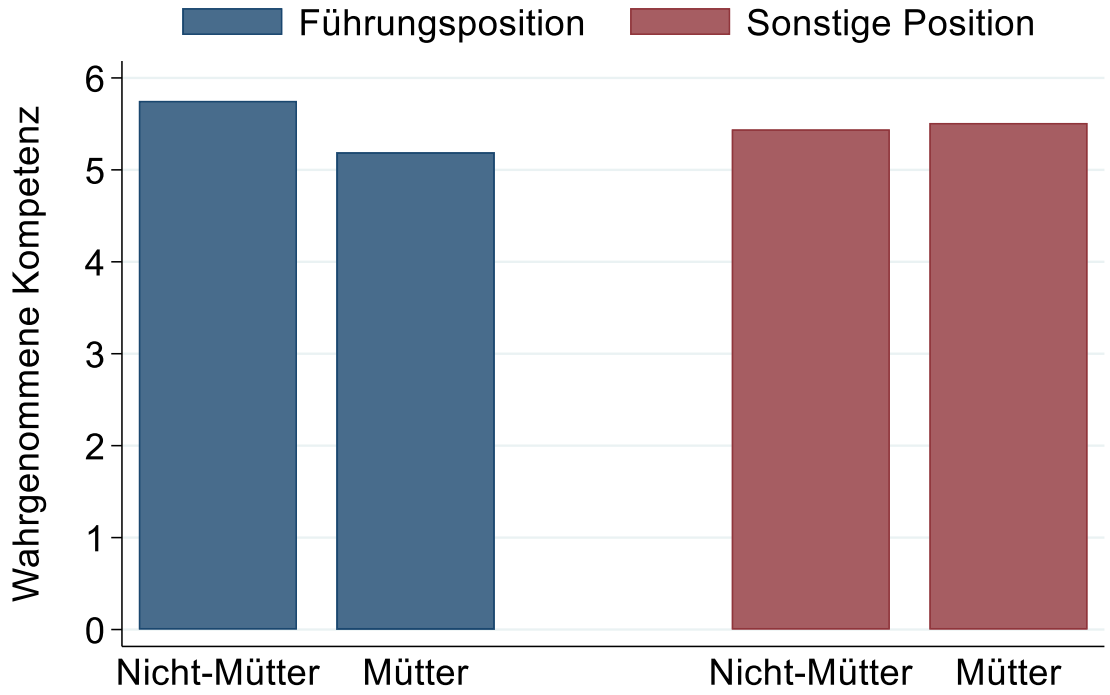


Correll et al. 2007, eigene Darstellung

Treatmenteffekt (TE): Heterogenität

- Oft variiert der TE nach Gruppen – was interessant sein kann
- Beispiele?
  - Führungspositionen vs. reguläre Positionen
  - Bewertung durch Frauen versus Männer
- Technisch handelt es sich dann um eine Interaktion bzw. Moderation durch die Drittvariable Z (z.B. Geschlecht VP)





Fiktive Daten

Arten von Experimenten
Vor- und Nachteile

Mit/ohne Vorhermessung

- Sinn von Vorhermessungen?
  - Kontrolle der Randomisierung
  - Ausgangsniveau – damit ggf. auch Hinweise auf Generalisierbarkeit
- Mögliche Nachteile?
  - Lerneffekte, mehr Künstlichkeit, Demand-Effekte
  - Kann das den Treatmenteffekt verzerren?
    - Ja, sofern Lerneffekte mit T interagieren
    - Beispiele?
- Solomon‘s 4 Gruppendesign
  - Treatment und Kontrollgruppe mit und ohne Vorhermessung (2x2 Designs)
  - Bei unterschiedlichen Kausaleffekten: Entweder hat die Randomisierung nicht geklappt, und/oder es gibt Lerneffekte

Arten von Experimenten: Between vs. Within

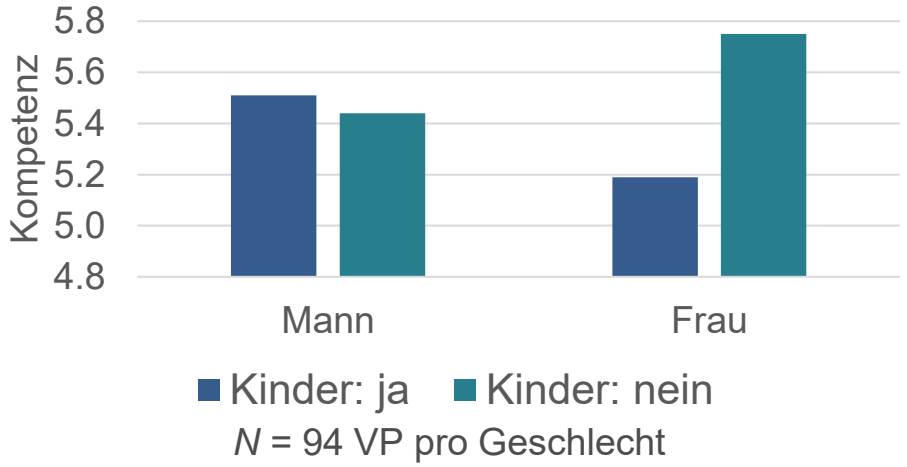
	Between	Within
Experimentalbedingungen pro Versuchsperson	Eine	Mehrere
Vergleich…	…zwischen („between“) VP	…innerhalb („within“) VP
Confounder	Potenziell problematisch, falls Randomisierung missglückt	Eher unproblematisch, noch mehr Konstanthaltung
Stärken/Schwächen	+ Subtilere Manipulation + Ggf. ethisch adäquater	– Lerneffekte, Ermüdung – Manipulation offensichtlicher

Arten von Experimenten: „mehrfaktoriell“/„factorial“

- Verwendung mehrerer experimenteller Stimuli
- Beispiel:
  - Elternschaft (ja/nein)
  - Geschlecht (Frau/Mann)

	♂	♀
👶	1	2
🧍	3	4

- 2 Faktoren à 2 Levels: „2x2 Design“
- Anzahl unterschiedlicher Treatments (Konditionen) als Produkt aller Levels, hier: 2 x 2 = 4 (entspricht der Zellenanzahl der Tabelle)
- Auch mehr Faktoren und Levels möglich, z.B. „3x4x2 Design“



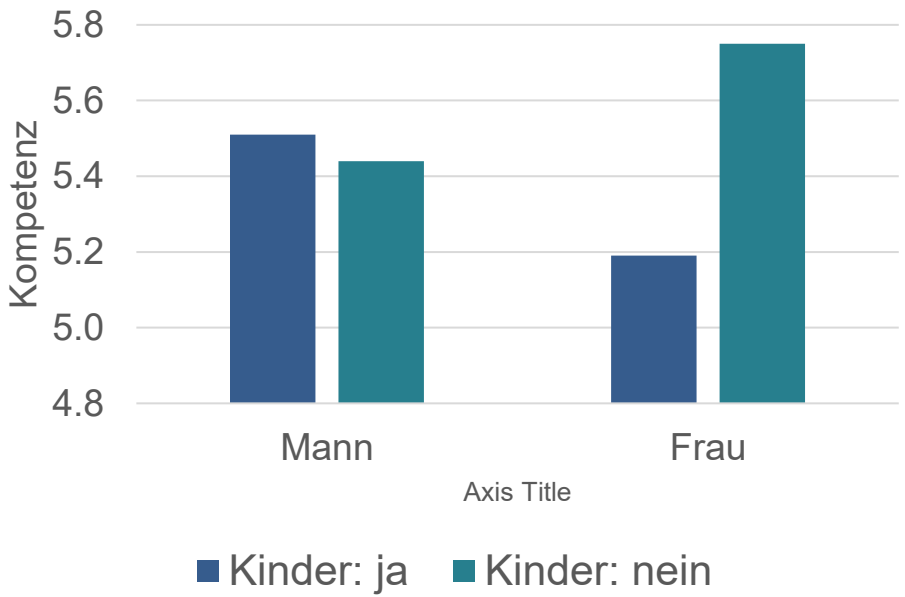
- Geschlecht „between“, Elternschaft „within“ → Gründe für diesen Mix?

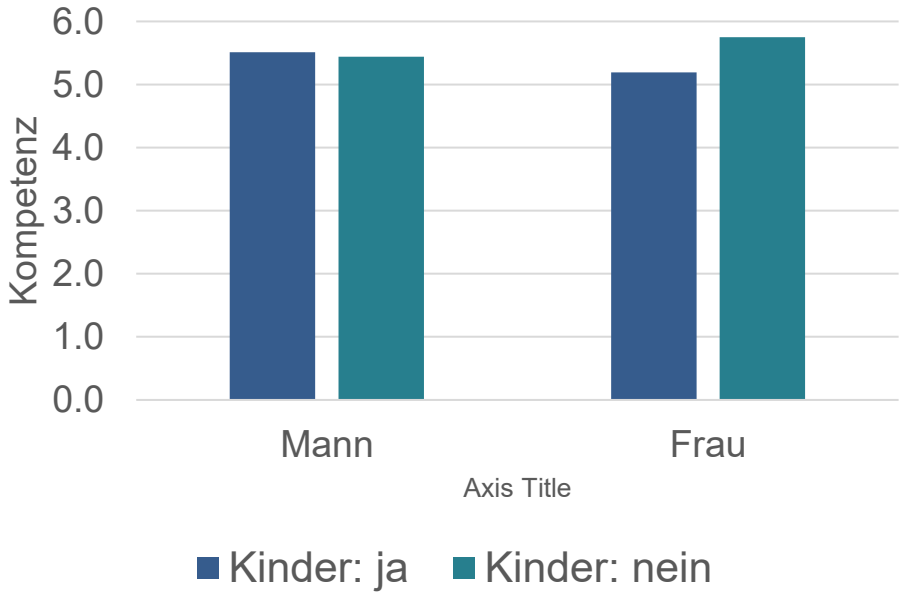
Correll et al. 2007, eigene Darstellung

Exkurs: Von Mücken und Elefanten

- Man sollte immer auch auf Effektstärken (d.h. Größe des Treatmenteffekts) achten:
  - Ist der Effekt substanziell? → muss mit Argumenten belegt werden
  - Auch sehr kleine Unterschiede können „statistisch signifikant“ sein
  - Vor allem ein Problem bei größeren Stichproben da der Standardfehler mit der Stichproebengröße abnimmt

Statistische Signifikanz ≠ Relevanz



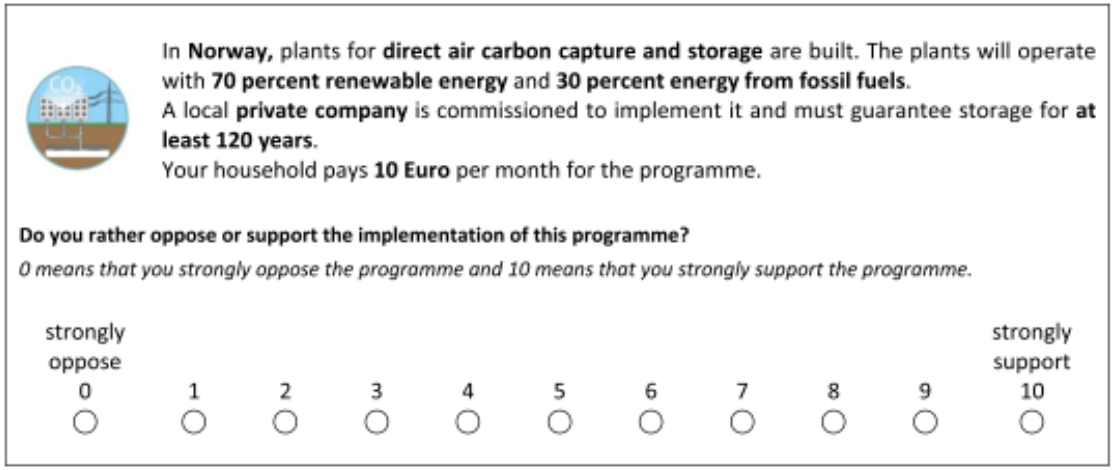


Correll et al. 2007, eigene Darstellung

Ausweitung: Insb. in Survey-Experimenten

- Dort Experimente mit vielen Faktoren und Levels einfacher implementierbar
- Grund: hohe Teilnehmerzahl, damit viele Kombinationen testbar
- Vorteil:
  - Interaktionen und Trade-offs schätzbar
  - Stärkere Standardisierung
  - …
- Vielfältiger Einsatz! (Conjointanalysen, Factorial Surveys, Choice-Experimente)

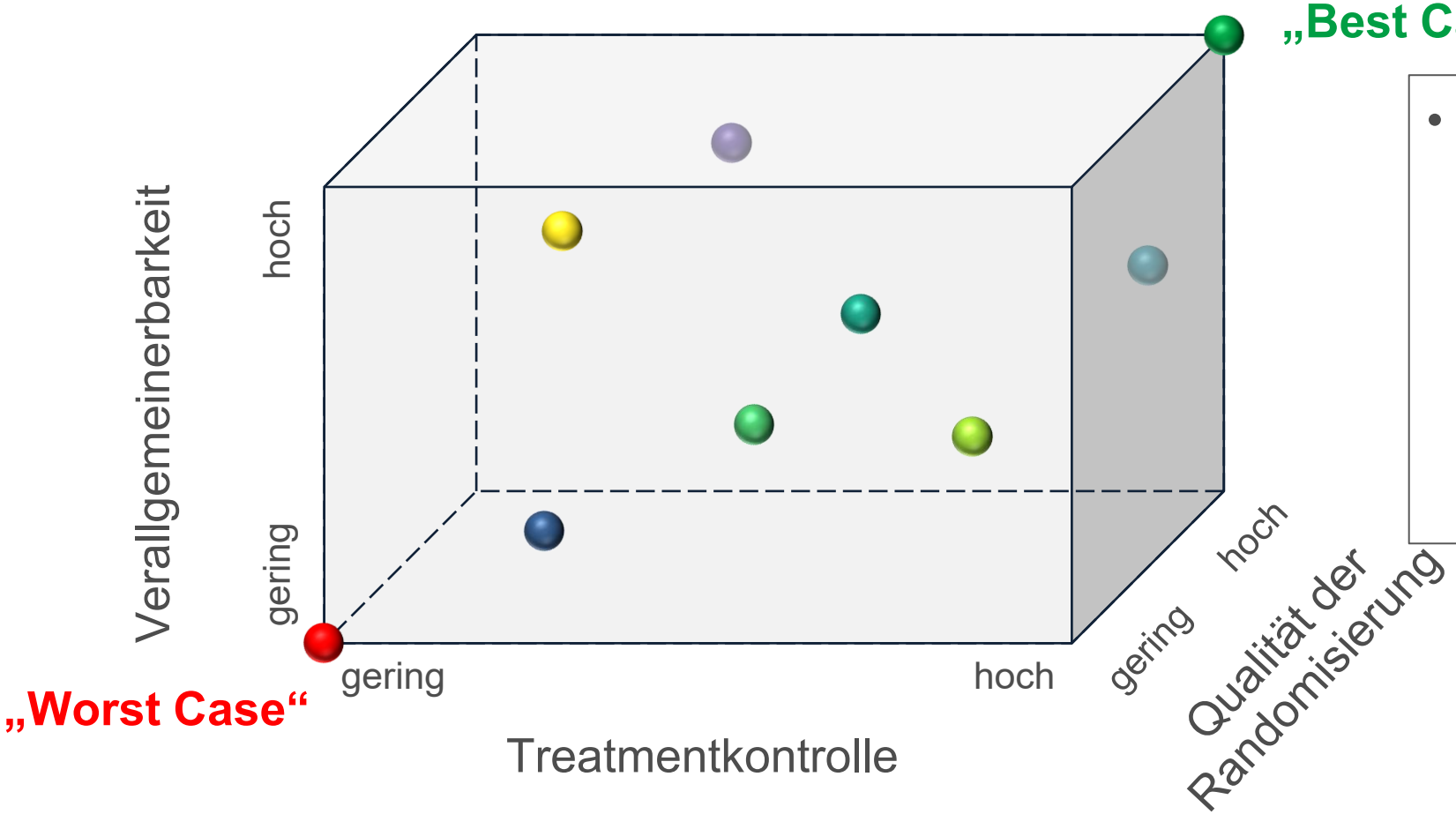
- Beispiel aus Akzeptanzforschung



- Potentielle Nachteile?

- Spezielles Seminar dazu im SoSe!

Varianten von Experimenten



- Unterschiedliche Optimierung in
  - Laborexperimenten
  - Feldexperimenten/Randomized Controlled Trials
  - Natürlichen Experimenten (s. Übersicht, spätere Kapitel)

Angelehnt an Dunning (2008: 31)

Wiederholung: Labor, Feld, natürliche Experimente

Labor-Experiment

- Maximale Kontrolle Treatment und Setting: Kontrolle d. Forschende
- Aber auch
  - Künstlich
  - Nur kurzfristige Effekte
  - Selektive Samples
- Damit hohe interne, geringe externe Validität

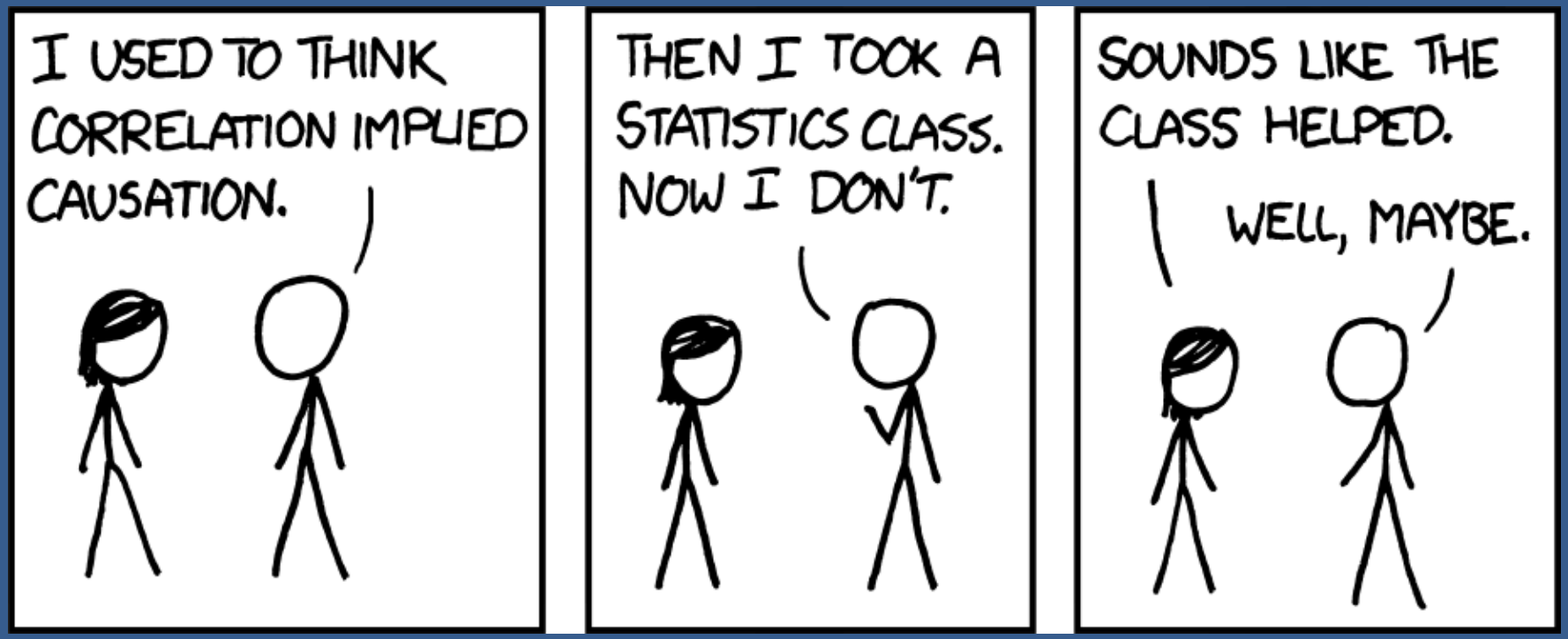
Feld-Experiment

- Treatment durch Forschende, aber in natürlicher Umgebung
- Damit etwas weniger interne, aber mehr externe Validität
- Einschränkung: Ethisch oft fragwürdig
  - Kein informed consent

Nat. Experiment

- Treatment durch andere (Natur, Institutionen) in natürlicher Umgebung
- Nur quasi-Randomisierung, damit noch weniger interne Validität
- Aber: Ethische/praktische Probleme ausgeräumt
  - Experiment findet ohnehin durch Andere statt

Kurze Pause


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Interne und externe Validität

Interne

- Misst man den TE richtig?
  - Hier: Effekt von Mutterschaft?
  - Erfordert kausalen Zusammenhang (keine Confounder!)
  - Erfordert korrekte Operationalisierung (keine Messfehler)

Externe

- Ist der TE generalisierbar?
  - Andere Settings/Situationen/Kontexte
  - Andere Untersuchungseinheiten (VP)
  - Andere Operationalisierungen T und Y
  - Falls „Scope-Bedingungen“ gegeben sind: TE ist theoretisch erwartbar

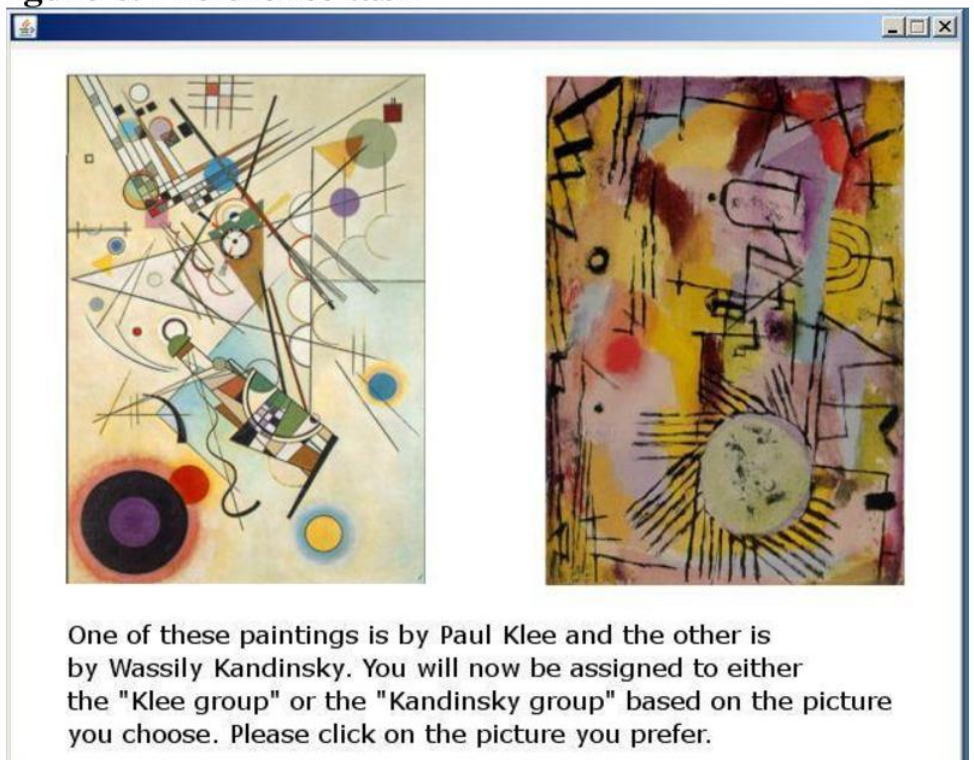
- Beispiele für Bedrohungen? Mögliche empirische Prüfung?
- Geht interne Validität zu Lasten der externen? Oder lässt sich beides optimieren?

→ Für beides sind Replikationen wichtig! (siehe späteres Kapitel)

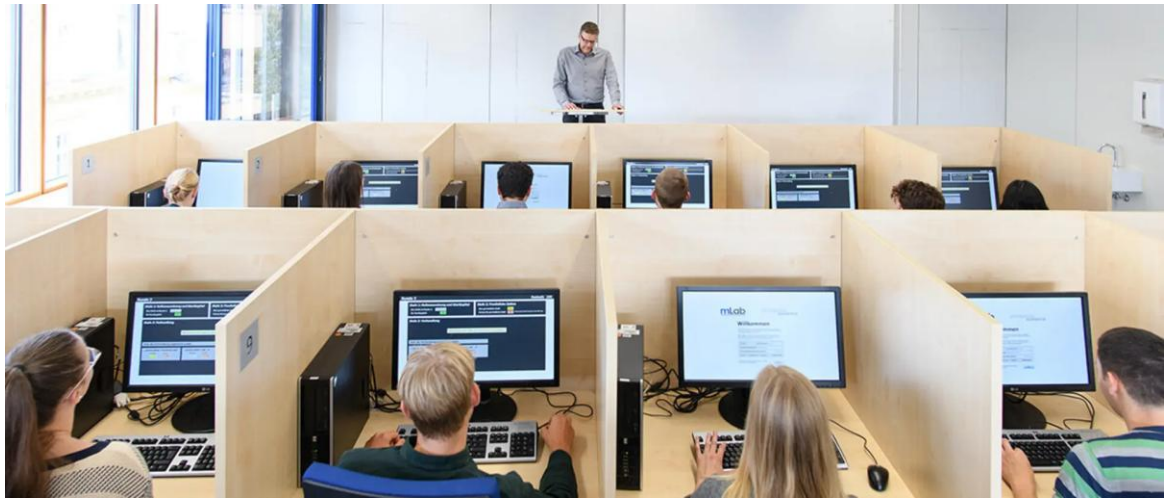
Beispiel: Maximierung von interner Validität

Treatment

Figure 6. Preference task



Umgebung



https://www.vwl.uni-mannheim.de/en/mlab/
Pillay, L. (2014). Investigating ingroup bias in an interactive minimal group environment.

Wurde das Treatment überhaupt empfangen?

- Annahme: Sie wollen den Effekt von Wahlerinnerungen auf Wahlteilnahme testen.
- Und werfen dazu Flyer in zufällig ausgewählte Briefkästen.
- Was könnte hier passieren? Erhalten alle das Treatment?
- ITE: Intention to Treat Effect δ^*

$$\mathrm{ITE} = \mathrm{E}[\delta^*] = \mathrm{E}[\mathrm{Y}^{1*}] - [\mathrm{Y}^0]$$

➢Mögliche Unterschiede? Testung?

➢Weitere Beispiele, warum ggf. nur ITE geschätzt werden kann?

- Möglichkeit ist z.B. Vergleich von Personen mit/ohne erfolgreichem Manipulationscheck
- Fraglich ist, welcher Effekt sinnvoll ist: ITE oder ATE konditioniert auf erfolgr. Treatment (= LATE: Local average treatment effect)

Diskussionsfragen/Verständnisfragen

Angenommen, wir wollen den kausalen Effekt von finanziellen Ressourcen auf Zufriedenheit untersuchen.

1. Wir versuchen das zunächst mit allg. Bevölkerungsdaten (z.B. ALLBUS). Was treten hier für Probleme auf?
2. Was wäre, wenn wir stattdessen Lotto-Gewinner mit der üblichen Bevölkerung vergleichen? Gibt es dann auch noch Confounder-Probleme?
3. Angenommen, wir bekommen alle Confounder-Probleme in Griff. Kann man aus den Ergebnissen zu Lottogewinnern dann folgern, dass finanzielle Ressourcen generell die Zufriedenheit beeinflussen?

Ausblick

- Nächste Woche: Warum Experimente (möglicherweise) ein Flop sind
- Aufbau der Sitzung ähnlich zu heute
- Unbedingt teilnehmen – Sie erhalten wichtige Hinweise für die Übungsaufgaben und Referate, u.A.:
  - An welchen Stellen kann man Experimente und andere Studien kritisieren?
  - Warum?
- Fragen/Anregungen?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Weiterführende Folien

Nochmals: Kausaleffekte

- Eigentlich will man oft den Effekt in einer Population schätzen:
  PATE: Population Average Treatment Effect
- Aber beobachtet wird nur Effekt in Subpopulation, oft CATE bezeichnet:
  CATE: Conditional Average Treatment Effect (Conditional: nur für Subgruppe, S = 1)

$$\mathrm{CATE} = \mathrm{E}[\delta \mid \mathrm{S} = 1] = \mathrm{E}[\mathrm{Y}^1 \mid \mathrm{S} = 1] - [\mathrm{Y}^0 \mid \mathrm{S} = 1]$$

- Alternative Bezeichnung: LATE: Local Average Treatment Effect
  (Local: nur für Subgruppe)

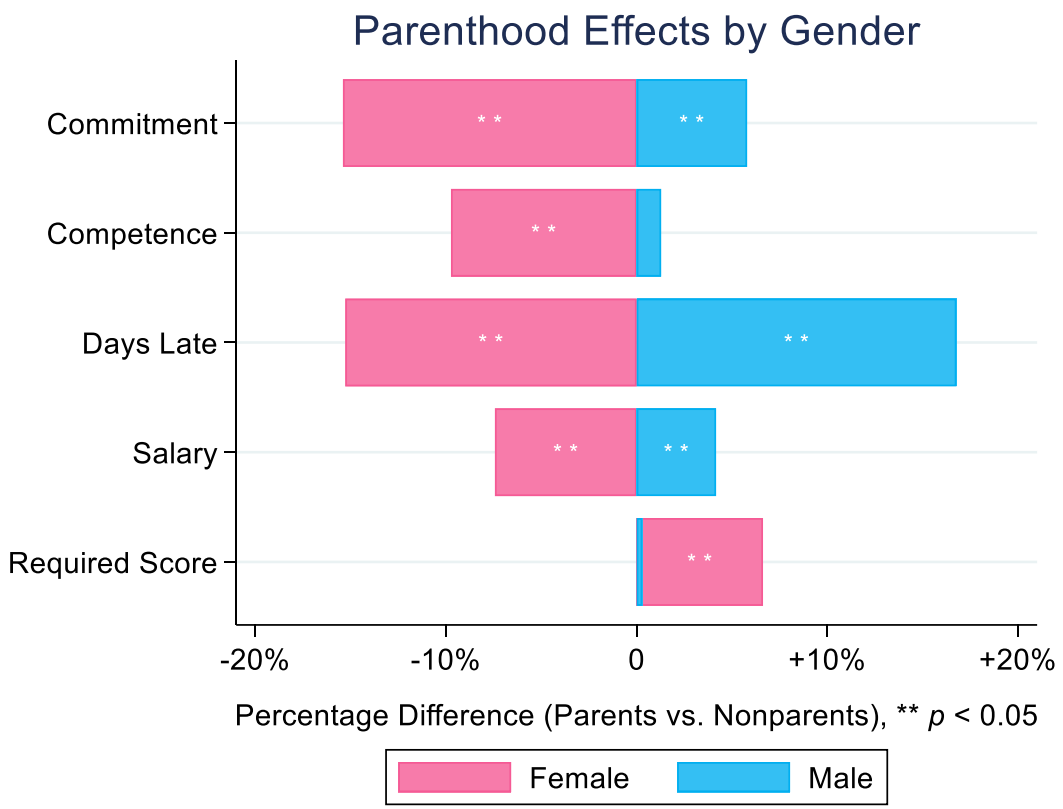
➢Was könnten Gründe für Unterschiede sein?

➢Wie könnte man diese testen?

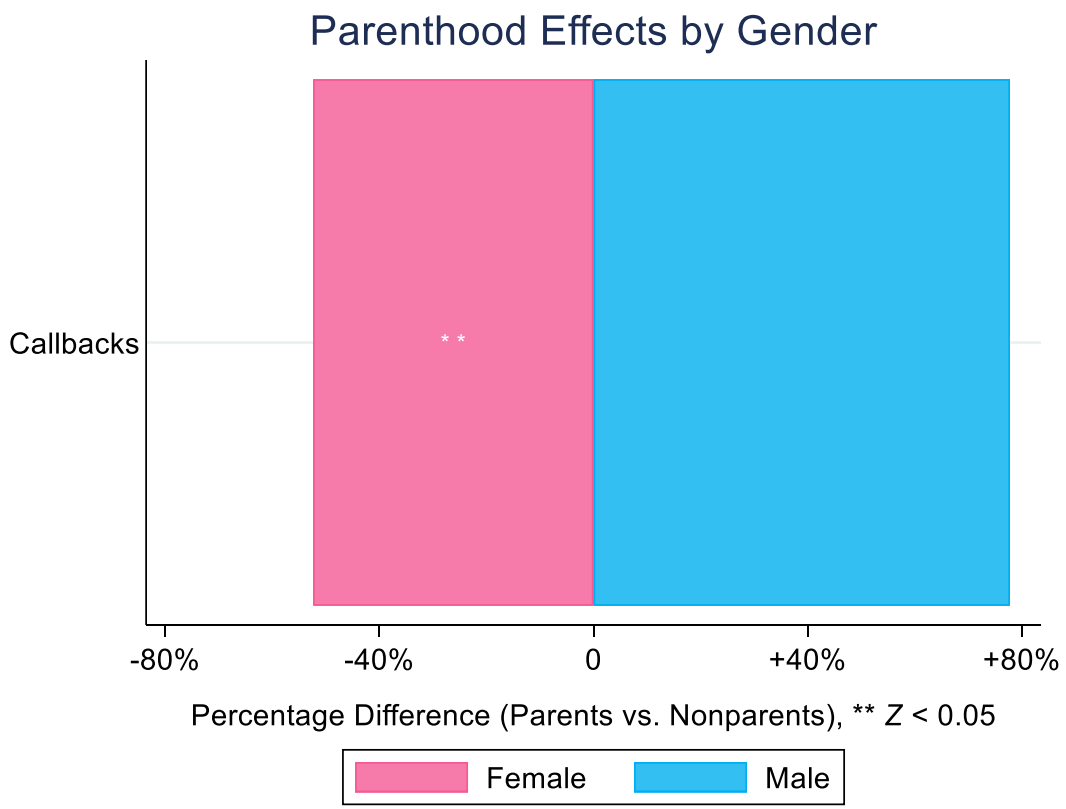
- Gründe sind Effektheterogenität bzw. Moderationseffekte!

Empirie: Motherhood Penalty (Correll et al. 2007)

Laborexperiment



Audit Studie



Eigene Darstellung

Nicht alles, was wie Zufall aussieht, ist (reiner) Zufall…

„Zunächst führten fünf Bachelor-Studenten jeweils mindestens 15 000 Münzwürfe durch und notierten das Ergebnis. Anschließend beteiligten sich 35 Freiwillige an zwölfstündigen ‚coin flipping marathons'. Und schließlich folgten weitere Teilnehmer einem Aufruf zum Münzenschnipsen, der über soziale Netzwerke geteilt worden war. Insgesamt wurde 350 757 Mal eine Münze geworfen. […]

Bei der Analyse zeigte sich zunächst, dass die geworfenen Münzen in etwa gleich häufig auf die eine oder die andere Seite fielen […] - es gab also keinen Hinweis auf einen ‚heads-tails bias' […] Einen Einfluss auf das Ergebnis hatte jedoch, welche Seite zu Beginn des Münzwurfs oben lag: Diese Seite liegt mit höherer Wahrscheinlichkeit am Ende des Wurfes wieder oben. Fachleute sprechen vom ‚same-side bias'."

Statistik

"Kopf oder Zahl" ist nicht fair

26. Oktober 2023, 17:43 Uhr | Lesezeit: 3 min

Im Fußball gehören Münzwürfe seit Jahrzehnten dazu - hier bei einem Spiel zwischen Borussia Dortmund und dem FC Schalke 04 im Februar 1971. (Foto: imago sportfotodienst)

Von wegen purer Zufall: Exakt 350 757 Mal haben Mathematiker Münzen geworfen - und Erstaunliches festgestellt.

(mehr zu „echten" Problemen bei der Randomisierung in kommenden Sitzungen)

Süddeutsche Zeitung

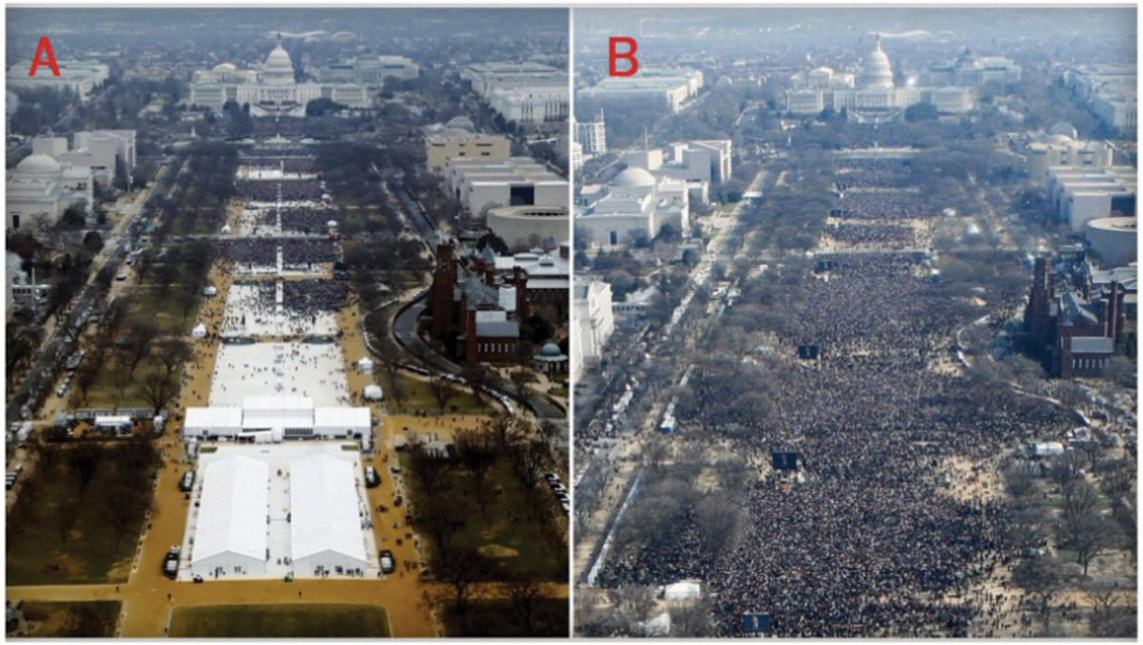
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Probleme von Experimenten: Bedrohungen der Validität

A picture says more than a thousand words

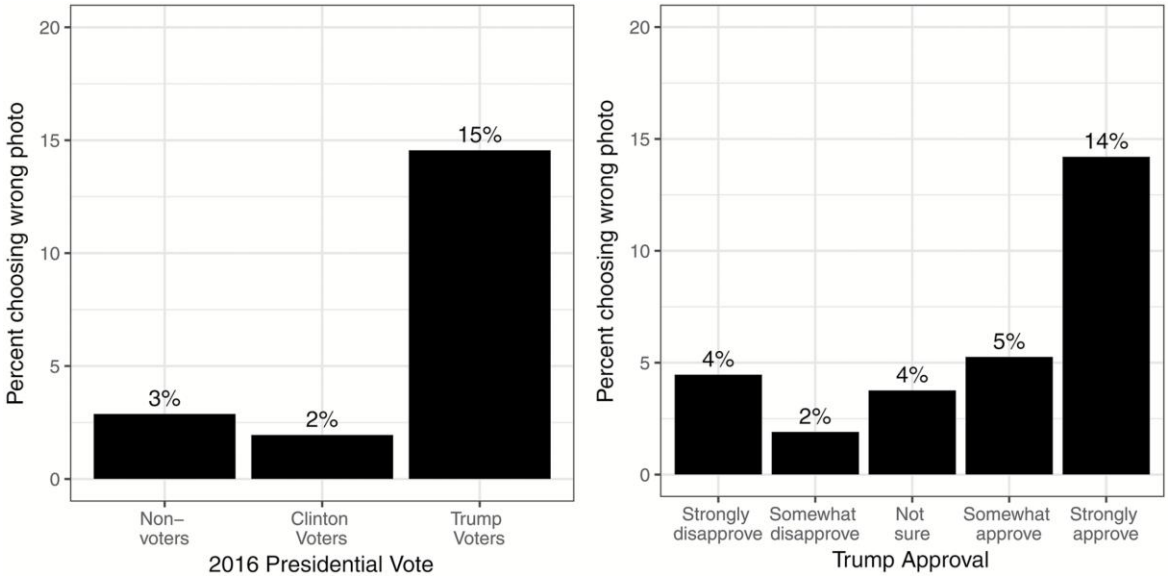
Misinformation or Expressive Responding?

Please look at the following two photos: Photo A and Photo B.



Which photo has more people?

Photo A has more people | Photo B has more people



https://academic.oup.com/poq/article/82/1/135/4868126

Grundsätzlich: Was ist der Mechanismus?

- *Z.B. Warum* kommt es zu „Diskriminierung“?

- Mediatoren nicht direkt manipulierbar
  - → Grund: simultane Zuweisung Treatments
  - → Experimente eignen sich nur bedingt für die Untersuchung von Wirkungsketten
- Lösung I: Komplexere Designs
  - → Mehr/weniger Information zu Mediator, z.B. Info zu Beschäftigungsstatus

Lösung II: Kombination mit Surveys

→ Messung von Vorurteilen, Präferenzen,…

Lösung III: Kontextinfos

→ Z.B. Marktsituation, Frauenanteil, etc.

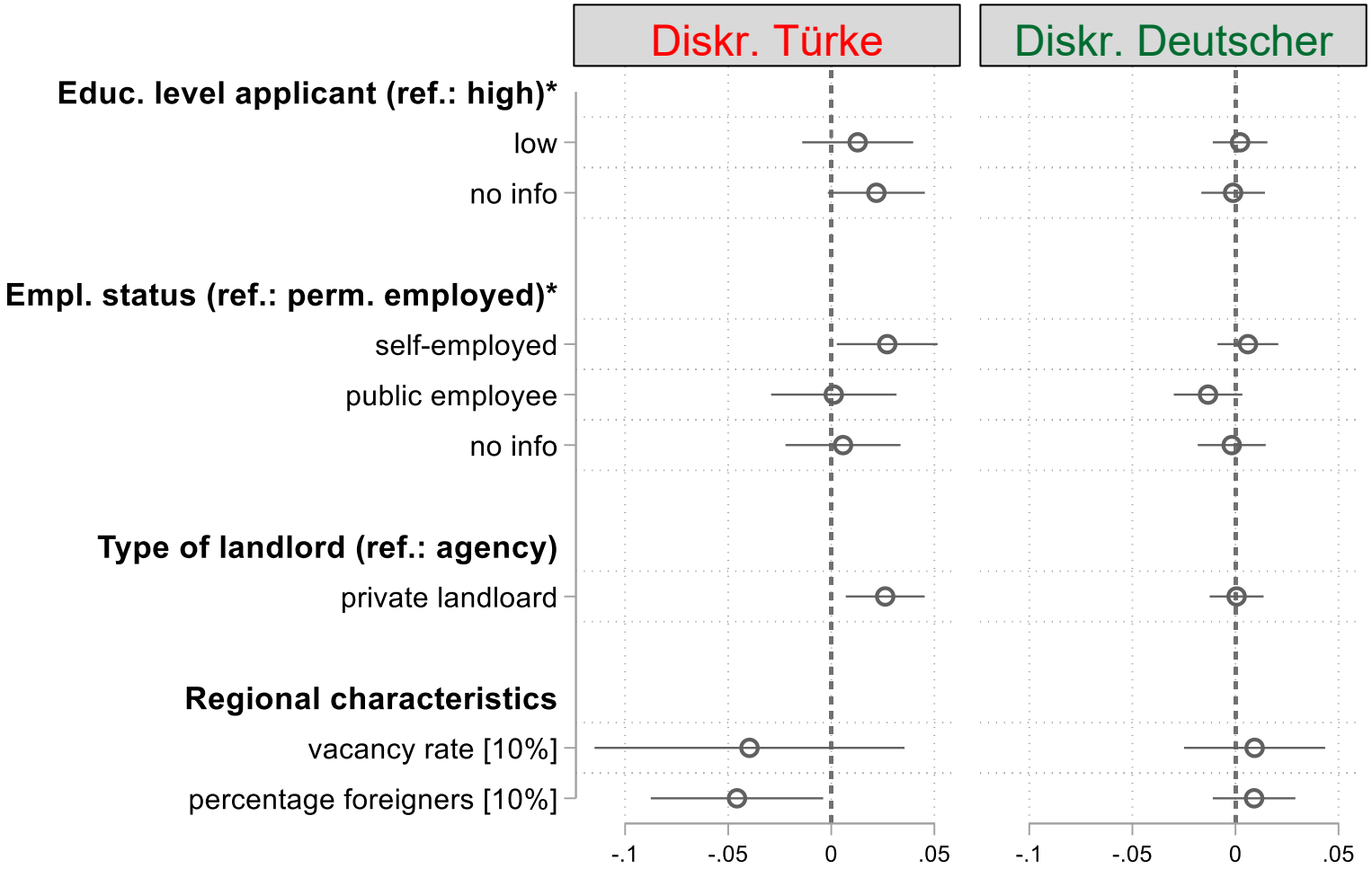
→ Geokodierungen und “Big Data“ bieten Optionen für prozessproduzierte Daten!

Mehrfaktorielles Design (Auspurg et al. 2017):

- Treatment 1: Nationalität
- Treatment 2: Beschäftigung (Info Ja/Nein + Art)

Hypothese: Info → weniger Diskriminierung

Diskriminierung auf dem Wohnungsmarkt



Note: *For outcome 'Discr. T' ('Discr. G'), Characteristics of T (G) are shown
Further controls: monthly rent, percent unemployed, east Germany, city; estimation with cluster robust standard errors

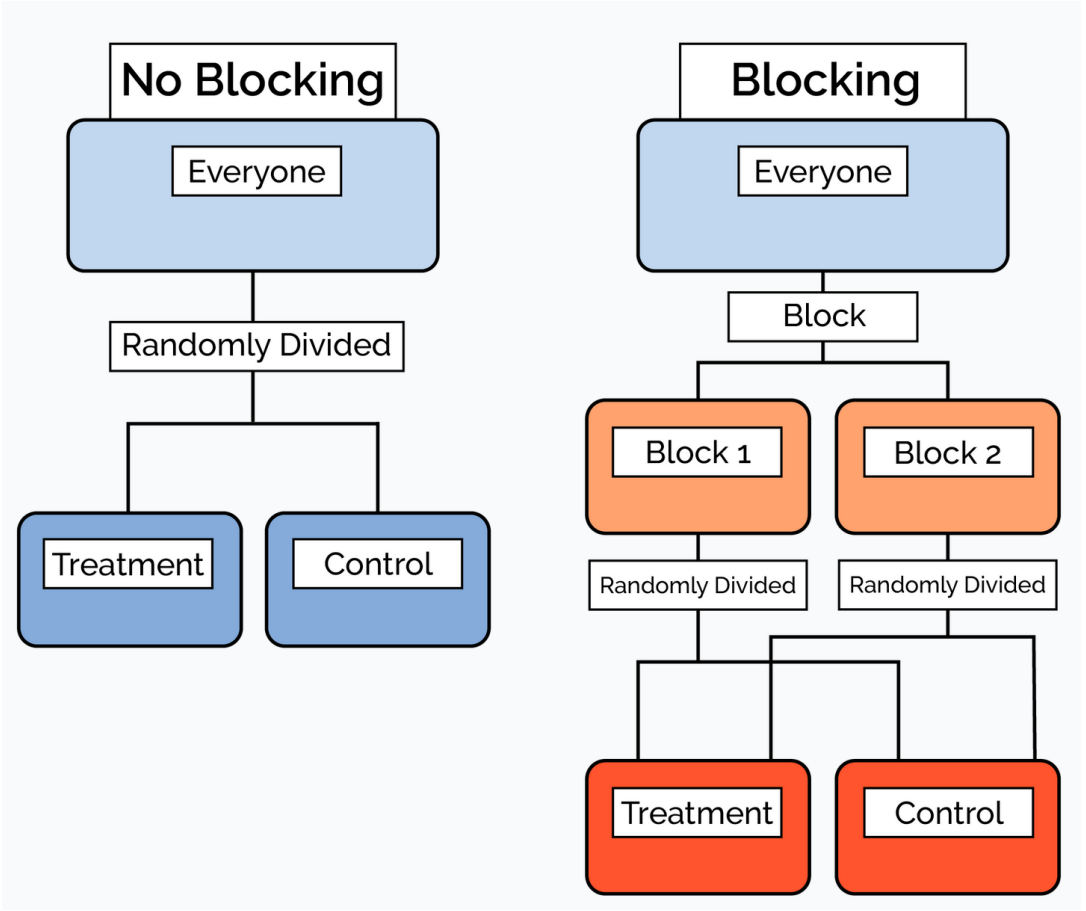
Ähnliche Kritik

- „Explanatory Narrowness“
  - Lediglich Fokus auf wenige Einflussfaktoren
  - Oftmals aber vermutlich Ursachenbündel
- Experimente messen i.d.R. nur sehr spezielle direkte Effekte („all else equal“)
  - Also z.B. direkten Gender-Effekt (Diskriminierung)

- Zu hinterfragen ist der Erkenntnisgewinn für die Forschungsfrage
  (s. dazu auch Kapitel zu natürlichen Experimenten)
  - Feststellung des kausalen Zusammenhangs (max. interne Validität)
  - Feststellung der Relevanz des Zusammenhangs zur Erklärung aktueller Phänomene (max. externe Validität)
- Designs mit ggf. mehr Erkenntnisgewinn
  - Theoriegeleitete Spezifikation (s. Kap. 2)
  - Mehrfaktorielle Experimente
  - Triangulation mit anderen Methoden

Bedrohungen Validität: Keine richtige Randomisierung …insbesondere bei kleinen Fallzahlen!

- Sowie Experimenten mit weniger Kontrolle (Feldexperimente, natürliche Experimente)
- Folge: Verlust interner Validität
- Abhilfen
  - Randomisierung prüfen, Robustheitsanalysen mit Kontrollvariablen (möglichen Confoundern)
  - Neutralisierung von erwarteten Confoundern durch balancierte Aufteilung sicherstellen („randomized block Design“)
    - Z.B. gleiche Geschlechterverteilung in Experimental- und Kontrollgruppe



Illinois Data Science Discovery

Zeitliche Veränderungen parallel zur Intervention

- Beispiele
  - Historische Ereignisse
  - Reifung („Mit Medikament dauert Grippe eine Woche, ohne 7 Tage“)
  - Antizipation
  - Lerneffekte/Reaktivität auf Vorhermessung
  - Regression zur Mitte (Extremgruppen können nicht noch extremer werden)

- Ist das (gegebenenfalls) eine Bedrohung der internen Validität? Der externen?
- Interne: Falls
  - C und T diesen Veränderungen unterschiedlich ausgesetzt sind
    - Ist – sofern Randomisierung geglückt ist – kaum plausibel?
  - T mit der Veränderung interagiert
    - Z.B.: Historisches Ereignis verstärkt T
    - Fallen Ihnen Beispiele ein?
- Andernfalls:
  Einschränkung externer Validität

Reaktivität / Experimenter Demand-Effekte

- **Definition**
  - Anderes Verhalten Versuchspersonen (VP) sobald sie wissen, dass sie Studienteilnehmer sind
- **Ursachen**
  - Streben nach angemessenem bzw. sozial erwünschten Verhalten
  - Streben nach plausiblen Ergebnissen im Einklang mit Hypothesen
- **Demand-Effekte**
  - (Unbewusste) Beeinflussung / Verhalten zugunsten der Hypothesen
- **Folgen**
  - Bedrohungen interner und externer Validität: künstliche Ergebnisse, Konfundierung von Treatmenteffekten
- **Abhilfen?**
  - Doppel-Blind Versuche
  - Experimente ohne…

    …Wissen der Beteiligten

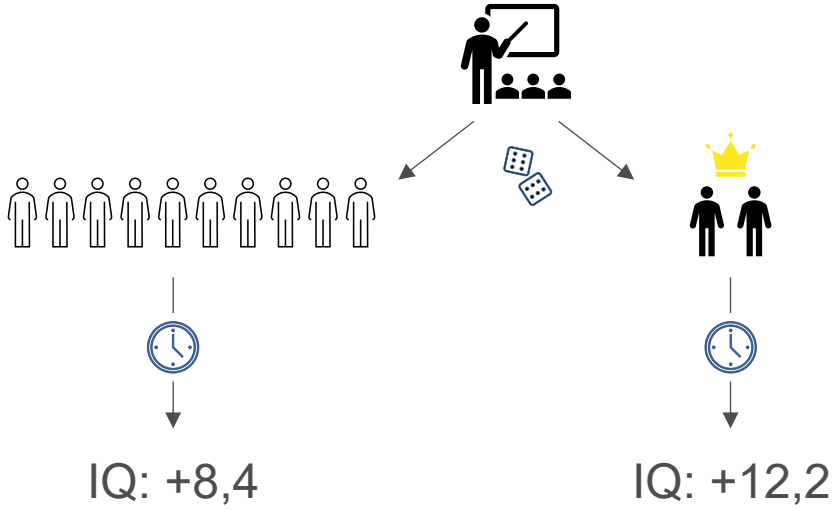
    → Praktische Umsetzung?

    …Offenlegung der Forschungsziele

    → Forschungsethik? (s. spätere Folien)

Beispiel I: Pygmalion-Effekt

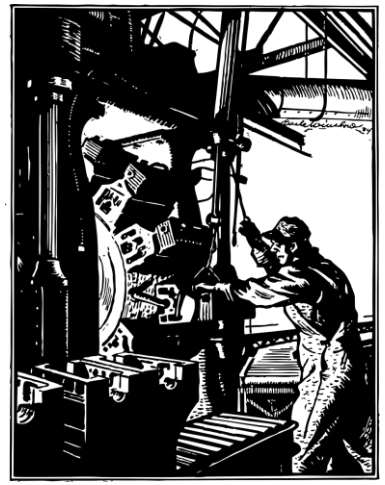
- Ursprung: Ovids *Metamorphosen*
  - Erweckung einer leblosen Marmorstatue durch Pygmalions Traum „der Künstler von der Beseelung ihrer Schöpfung“
- Empirie: Rosenthal & Jacobson (1966)



→ „Self-fulfilling prophecy“

Legenden - Pygmalion

Beispiel II: Hawthorne Effekt



- Oftmals plausibel und gut belegt
- Die Daten der Originalstudie sind aber weniger belastbar, als oft angenommen
  - Es werden etliche Confounder diskutiert!
  - Auch „klassische“ Experimente sollten gut geprüft werden (s. Kapitel „Total Error“)

Discussion paper

Scand J Work Environ Health 2006;32(5):402-412 pdf
https://doi.org/10.5271/sjweh.1036 | Issue date: 31 Oct 2006

The “Hawthorne effect” is a myth, but what keeps the story going?

by Kompier MAJ

Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments†

By Steven D. Levitt and John A. List*

Was There a Hawthorne Effect?[1]

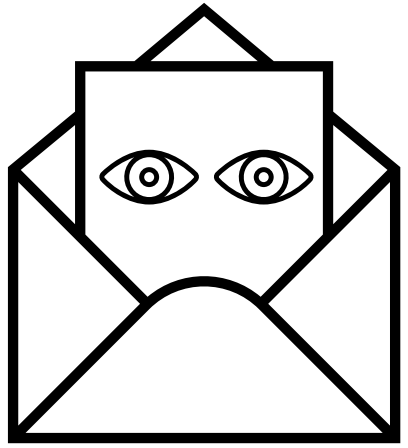
Stephen R. G. Jones
McMaster University

Beispiel III: Reaktivität (Schwarz et al. 2013)

- Postkarte mit Information, dass Energiekonsum für experimentelle Studie beobachtet wird
- Aber: *kein* Treatment in Form von Appellen zu mehr Energiesparen o.ä.
- Ergebnis: Absenken Stromverbrauch um im Mittel 2,7 % gegenüber der Kontrollgruppe ohne Postkarte

➢ Within-Experimente zur Wirkung von z.B. Informationen brauchen Kontrollgruppe mit neutraler Intervention, um wirklich Treatmenteffekt identifzieren zu können!

- Sonst keine Trennung von Reaktivität möglich



Weitere Form von self-fulfilling prophecy

- Confirmation /Selection bias durch Forschende
  - Kann zu Messfehlern führen
  - Selektiven Interpretationen
  - Selektivem Publizieren

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains

Stephanie Mertens, Mario Herberz, Ulf J. J. Hahnel, and Tobias Brosch Authors Info & Affiliations

Edited by Susan Fiske, Psychology Department, Princeton University, Princeton, NJ; received April 27, 2021; accepted November 24, 2021

December 30, 2021 | 119 (1) e2107346118 | https://doi.org/10.1073/pnas.2107346118

LETTER | PSYCHOLOGICAL AND COGNITIVE SCIENCES

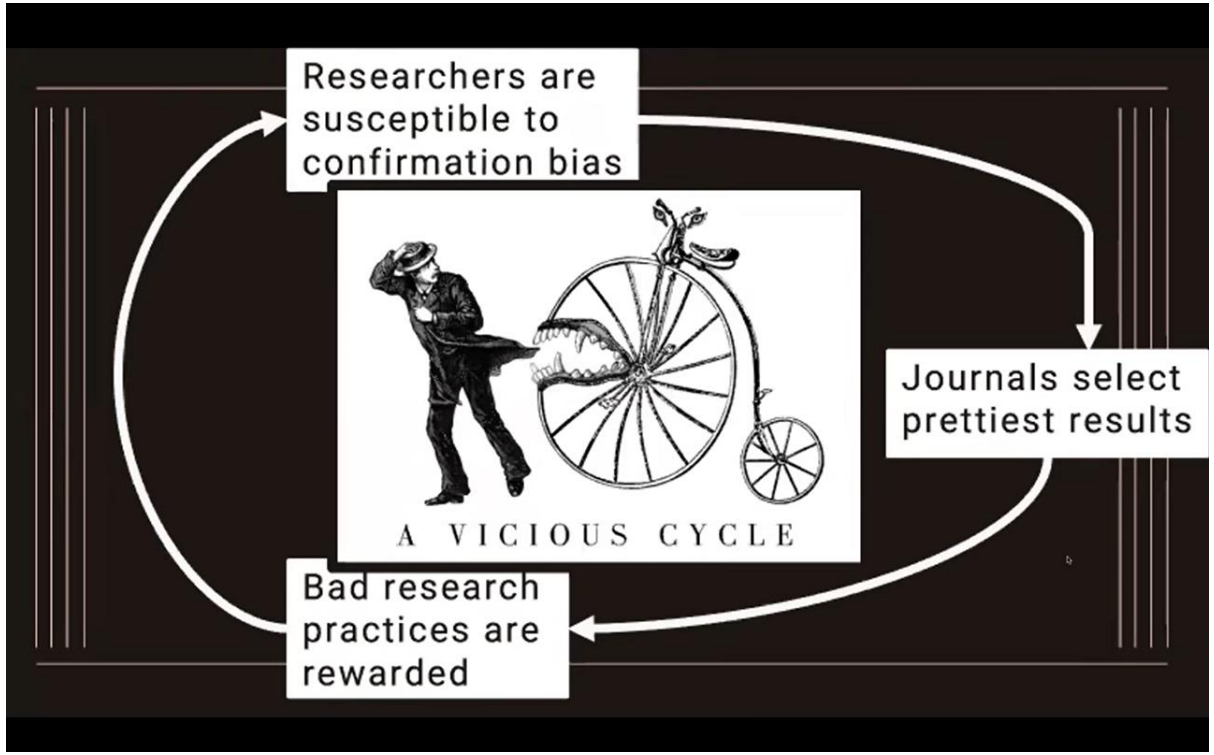
No evidence for nudging after adjusting for publication bias

Maximilian Maier, František Bartoš, T. D. Stanley, +2, and Eric-Jan Wagenmakers Authors Info & Affiliations

July 19, 2022 | 119 (31) e2200300119 | https://doi.org/10.1073/pnas.2200300119

Choice architecture research is a relatively young field that comes with some methodological “growing pains.” (Mertens et al. 2022)

- s. spätere Kapitel zu „Total Error“ und „publication bias“



Simine Vazire, Presentation at VSF Workshop

Wurde das Treatment überhaupt empfangen?

- Annahme: Sie wollen den Effekt von Wahlerinnerungen auf Wahlteilnahme testen.
  - Treatment: Einwerfen von Flyern in zufällig ausgewählte Briefkästen.
- Was könnte hier passieren? Erhalten alle das Treatment?
- ITE: Intention to Treat Effect δ^*

$$\mathrm{ITE} = \mathrm{E}[\delta^*] = \mathrm{E}[\mathrm{Y}^{1*}] - [\mathrm{Y}^0]$$

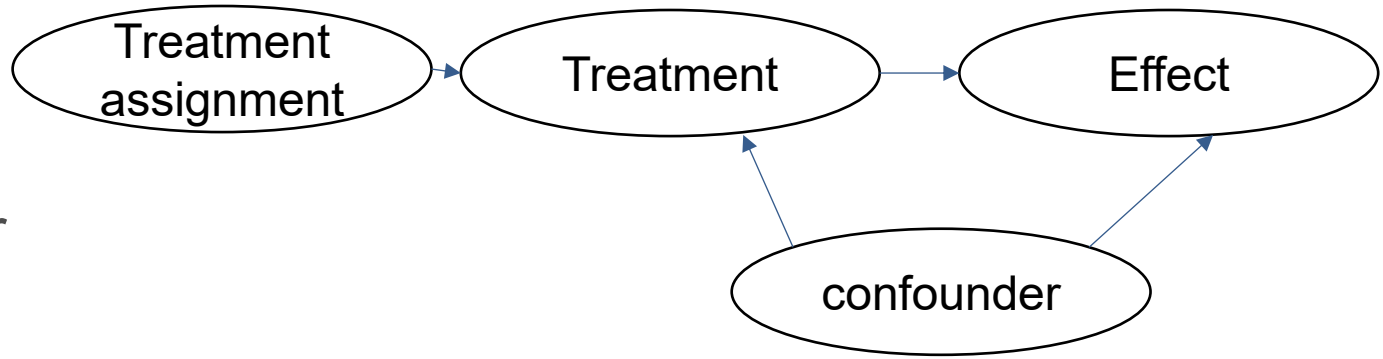
➢Mögliche Unterschiede? Testung?

➢Weitere Beispiele, warum ggf. nur ITE geschätzt werden kann?

- Fraglich ist, welcher Effekt sinnvoll ist: ITE oder ATE konditioniert auf erfolgr. Treatment (= LATE: Local average treatment effect)

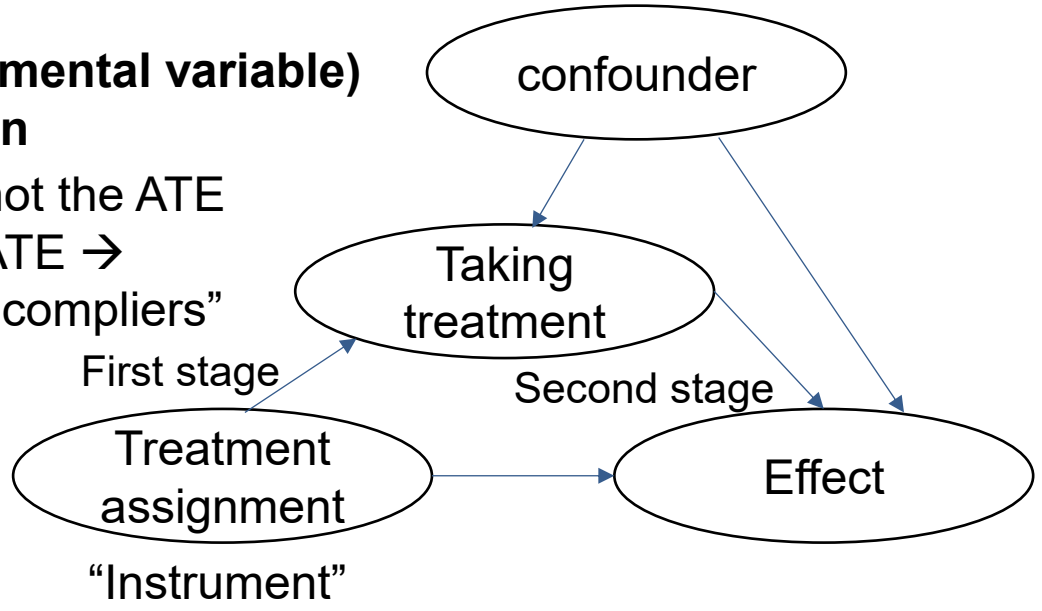
Beispiel: Informationsexperiment bei Arbeitslosen

- Idee: Arbeitslose würden rascher einen Job finden wenn sie mehr Information haben, wie sie ihre Jobsuche effizienter gestalten (unemployment friction).
- Experiment: Info-Mail mit Video an ein zufällig ausgewähltes Sample an Arbeitslose + Kontrollgruppe ohne Mail
- Analyse: Anteil Arbeitslose in Kontroll- und Treatmentgruppe nach 6 Monaten (ATE)
- Was könnten die Probleme sein?



IV (Instrumental variable) estimation

Get's us not the ATE but the LATE → effect on "compliers"



Mühlböck et al. 2021. Information, reflection, and successful job search: A labor market policy experiment

Weitere Beispiele für ggf. erfolgloses Treatment

1. Personen lesen zu unaufmerksam, verstehen Treatment falsch

   → Manipulationschecks

   Aber: Wie umgehen mit dem Ergebnis?

   - Wenn Check entlarvt, dass Treatment *per se* nicht funktioniert hat? (Beispiel?)
   - Wenn Check entlarvt, dass Treatment *für einzelne VP* nicht funktioniert hat?

2. VP brechen (selektiv) Teilnahme ab, sog. a*ttrition*

   → Messung/Dokumentation von Merkmalen

   Aber: Folgen für Interpretation/Analysen?

3. VP in Kontrollgruppe suchen sich selbst Treatment (Beispiel?)

   → Umgang damit? Folgen?

Kontrollgruppe erhält/holt sich auch Treatment?

- Beispiele?
  - Impfungen
  - Nachhilfe durch VP in Treatmentgruppe
  - Sensibilisierung durch VP in Treatmentgruppe
  - …

Herdenimmunität: Ausweg aus der Corona-Pandemie

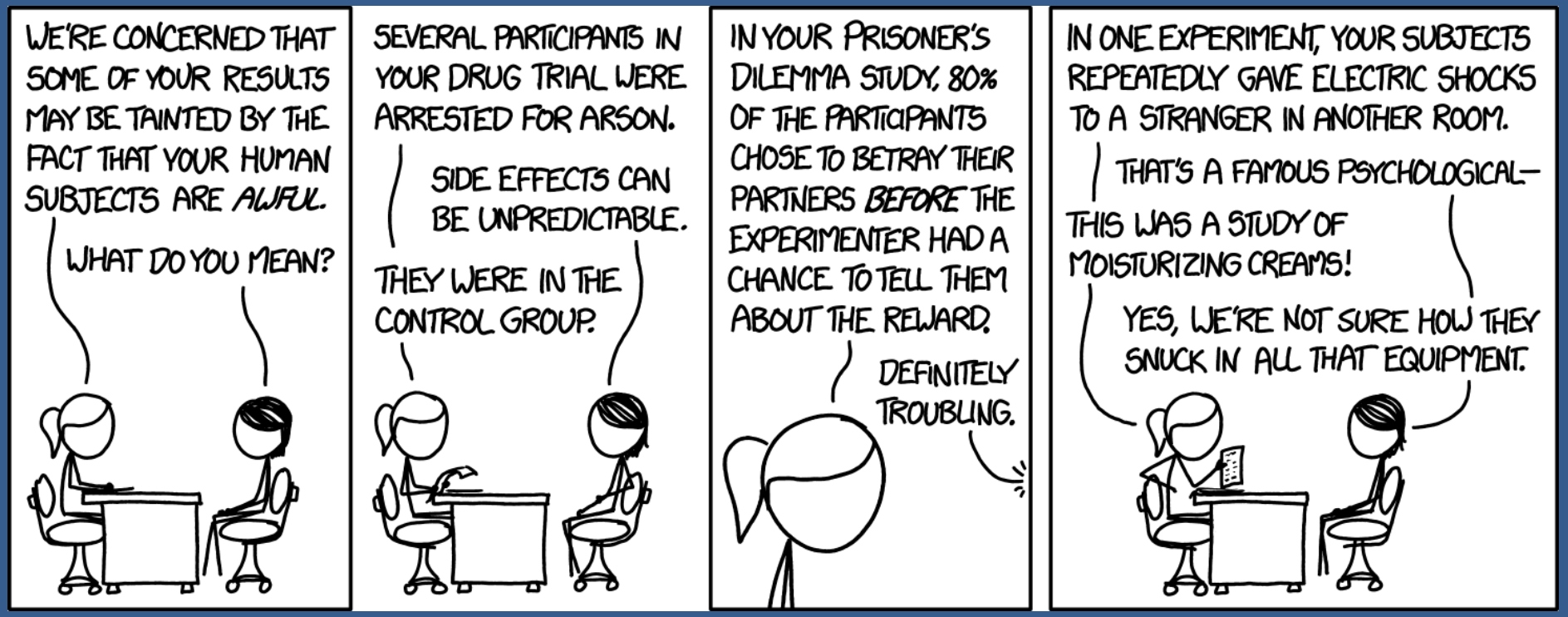
Stand: 09.06.2021 14:58 Uhr

Expertinnen und Experten gehen davon aus, dass in Deutschland 80 bis 85 Prozent der Menschen geimpft sein müssten, um eine sogenannte Herdenimmunität gegen das Coronavirus zu erreichen.

- Folgen?
  - Ggf. Bedrohung interner Validität: keine saubere Trennung von C und T
  - Teilweise aber auch Teil des interessierenden Treatmenteffets
    - „Spillover-Effekte“, die im Rahmen der SUTVA-Annahme zu diskutieren sind
    - Diese betrifft primär die externe Validität (spätere Folien)

NDR

Kurze Pause



XKCD

SUTVA

- **S**table **U**nit **T**reatment **V**alue **A**ssumption
- Treatmenteffekt für A ist unabhängig vom Treatmentstatus von B (und C…):

  "potential outcomes for a given observation respond only to its own treatment status; potential outcomes are invariant to random assignment of others"
  (Gerber and Green 2010)

- Gründe für die Verletzung (Beispiele folgen)
  - Spillover-Effekte
  - Exklusivität des Treatments beeinflusst
    - Wettbewerbsstruktur
    - Wahrnehmung / Reaktivität auf Treatment

- Folgen: Beeinträchtigung der Validität
  - Interne: Falls keine saubere Trennung Treatment- und Kontrollgruppe mehr
  - Externe: Falls Effekte von Anteil/Anzahl Personen mit Treatment abhängen

SUTVA: Beispiele Spillover-Effekte

Mechanismen

- Ansteckungseffekte
- Verdrängungseffekte
- Kommunikation / Interaktion
- Soziale Vergleichsprozesse
- Signalwirkung

Beispiele für Treatments

- Impfung
- Maßnahmen gegen Kriminalität
- Informationen (Produkte, Weiterbildung)
- Verbesserung Wohnbedingungen
- Policy-Interventionen wie Subventionen

Nach Gerber/Green 2010 bzw. Mize 2023; Mechanismen teils nicht trennscharf

SUTVA

- Warum könnten Wettbewerbsstrukturen relevant sein?
  - Denken Sie etwa an Experimente auf dem Arbeitsmarkt
- Wahrnehmung/Exklusivität des Treatments?
  - Beispiel: Nudging Experimente
  - Forschungsprojektseminar 2011 von Pelle G. Hansen in Roskilde
  - 46% weniger „Littering“

Beispiel

- Angenommen, Sie wollen in einer Pilotstudie mit interessierten Unternehmen den Effekt einer Arbeitszeitverkürzung schätzen auf
  - Die Zufriedenheit des Personals
  - Die Chancen, besonders qualifizierte und motivierte Fachkräfte gewinnen zu können
- Von den Unternehmen wird die Hälfte zufällig für die Implementation einer 4-Tages-Woche (Mo-Do) ausgewählt (Treatmentgruppe), die anderen Unternehmen bilden die Kontrollgruppe (weiterhin 5-Tages-Woche)

➢ Was wäre hier (bei der Implementation oder Schätzung der Effekte) ggf. zu beachten?

➢ Welcher Effekt wird geschätzt?
Ist die SUTVA hier womöglich verletzt?

Zusammenfassung: Merkposten interne Validität

Bedrohungen: Korrelation Y und T

- Missglückte Randomisierung
- Soziale Erwünschtheit / Reaktivität
  - Experimenter Demand Effekte, Hawthorne- / Pygmalion Effekte
- Treatment zu subtil/missverständlich; weitere Messfehler
- Personen aus T oder C springen selektiv ab / vermeiden Treatment
- Kontrollgruppe sucht sich selbst eine Art „Treatment“; keine saubere Trennung von C und T

Abhilfen

- Hohes N_{VP}, Prüfung
- Doppel-Blind-Versuche, anonyme Versuchsbedingungen
  - Feldexperimente, natürliche Experimente
- Manipulation Checks; ITE vs. ATE/ATT; Validität des Treatments diskutieren
- Durch Incentives vermeiden; dokumentieren; ITE vs. ATE/ATT
- Ggf. „nur“ SUTVA-Annahme verletzt; dokumentieren und reflektieren

Zusammenfassung: Merkposten externe Validität

Bedrohungen: Effektheterogenität

- Problematisch? Je nach Forschungsziel
  - Schätzung eines generellen Mechanismus (Ziel: Theorieprüfung)
    → Spezielle Samples sind unproblemati
  - Schätzung der Effektstärke (Ziel: Effektivität einer I
    → Selektive Sa
    Verzerrunge
- Ähnliches für ande kte externer Validität (Operationalisierung, SUTVA)

Abhilfen

- Generalisierb t beachten/diskutieren
  - Id Effekt vs. Effektstärke
    ulation und zu
    ktheterogenität
    . Verallgemeinerbarkeit für verschiedene N_{treated}
- Replikationen mit anderen (gezielt ausgewählten!) Samples vorteilhaft
  - Stärkt Vertrauen in interne Validität
  - Gibt Aufschluss über Effektheterogenität

Diskussion in Kapitel zu Samples

Merkposten insgesamt

- Sie sollten immer kritisch hinterfragen, ob es sich um eine glaubwürdige Schätzung handelt; somit:
  - Ob es alternative Erklärungen gibt
  - U.a. in Form von Messfehlern
  - Der Effekt wie gewünscht generalisierbar ist
- Und hier nicht „blind“ den Argumenten der Autoren vertrauen!
  - Häufig werden dort „Misserfolg“ und Einwände unterschlagen (s. späteres Kapitel zu „Total Error“)

Sieht nicht gut aus für unser Antidepressivum.

Nicht zuletzt: Forschungsethik!

- Wann sind Experimente ethisch problematisch?
- Wie sollte man damit umgehen?

Täuschung

Informed Consent

Wohlbefinden

Debriefing

Transparenz

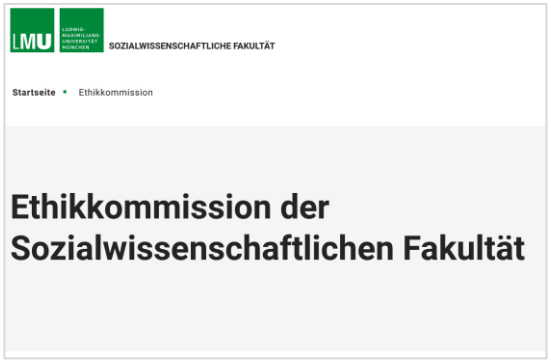
Anonymität

- Nähere Hinweise: spätere Sitzungen


akademie für soziologie

Ethik-Kodex der Akademie für Soziologie (AS)[1,2]

31. Januar 2022


SOZIALWISSENSCHAFTLICHE FAKULTÄT

Startseite • Ethikkommission

Ethikkommission der Sozialwissenschaftlichen Fakultät

Diskussionsfragen/Verständnisfragen

Angenommen, wir wollen den kausalen Effekt von finanziellen Ressourcen auf Zufriedenheit untersuchen.

1. Wir versuchen das zunächst mit allg. Bevölkerungsdaten (z.B. ALLBUS). Was treten hier für Probleme auf?
2. Was wäre, wenn wir stattdessen Lotto-Gewinner mit der üblichen Bevölkerung vergleichen? Gibt es dann auch noch Confounder-Probleme?
3. Angenommen, wir bekommen alle Confounder-Probleme in Griff. Kann man aus den Ergebnissen zu Lottogewinnern dann folgern, dass finanzielle Ressourcen generell die Zufriedenheit beeinflussen?

Diskussionsfragen/Verständnisfragen

Angenommen wir schenken 5 zufällig ausgewählten Studierenden im Seminar 100 Euro, um den Effekt von finanziellen Ressourcen auf Zufriedenheit zu messen.

1. Wäre hier die interne Validität hoch? Was könnte man dagegen einwenden?
2. Was wäre, wenn wir Sie bitten sich vorzustellen, sie bekommen ein Hiwi-Monatsgehalt geschenkt; und sie sollen davor und danach jeweils angeben, wie zufrieden Sie sind?
3. Was ist, wenn wir das Experiment mit den 100 Euro in einem Seminar zur Weiterbildung von BMW-Managern wiederholen und dann andere Effekte finden. Ist dann das erste Experiment nicht valide gewesen? Welche Validität ist ggf. bedroht, warum?

Weitere Ergebnisse aus unserer Studie

- Mehr Migration, mehr Diskriminierung?
  - Mehr Konkurrenz
  - Mehr Fremdenfeindlichkeit
  - Aber: Diskriminierung „kostet“
- Vergleich vor/nach 2015 („natürlicher“ Cut-Off)

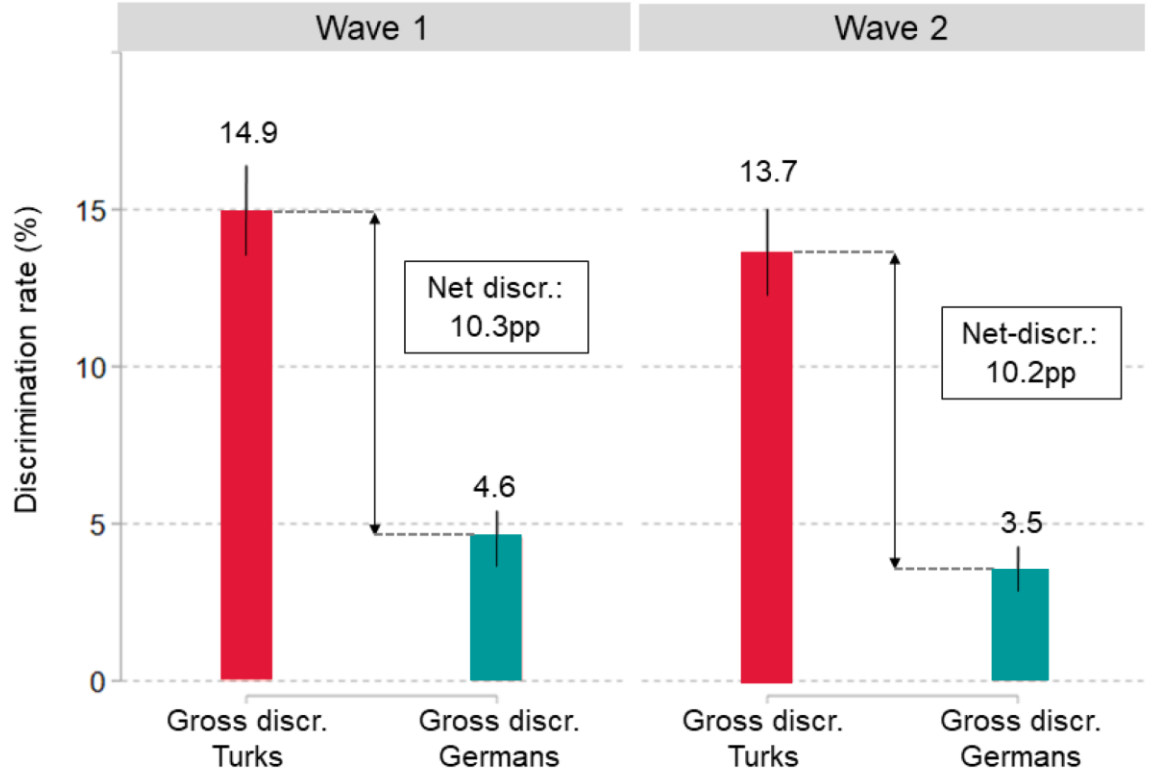


Figure 4: Gross and Net Discrimination Rates by Wave.
Note: The bars show the gross discrimination rates in percent. The net discrimination, which is the difference between the gross discrimination rate of Turks and Germans, is indicated in percentage points (pp). The sample comprises 2,389 tested housing units in the 1^{st} wave and 2,410 housing units in the 2^{nd} wave.

ITE oder ATE/ATT für Personen mit sicher wahrgenommenen Treatment – knifflige Frage!

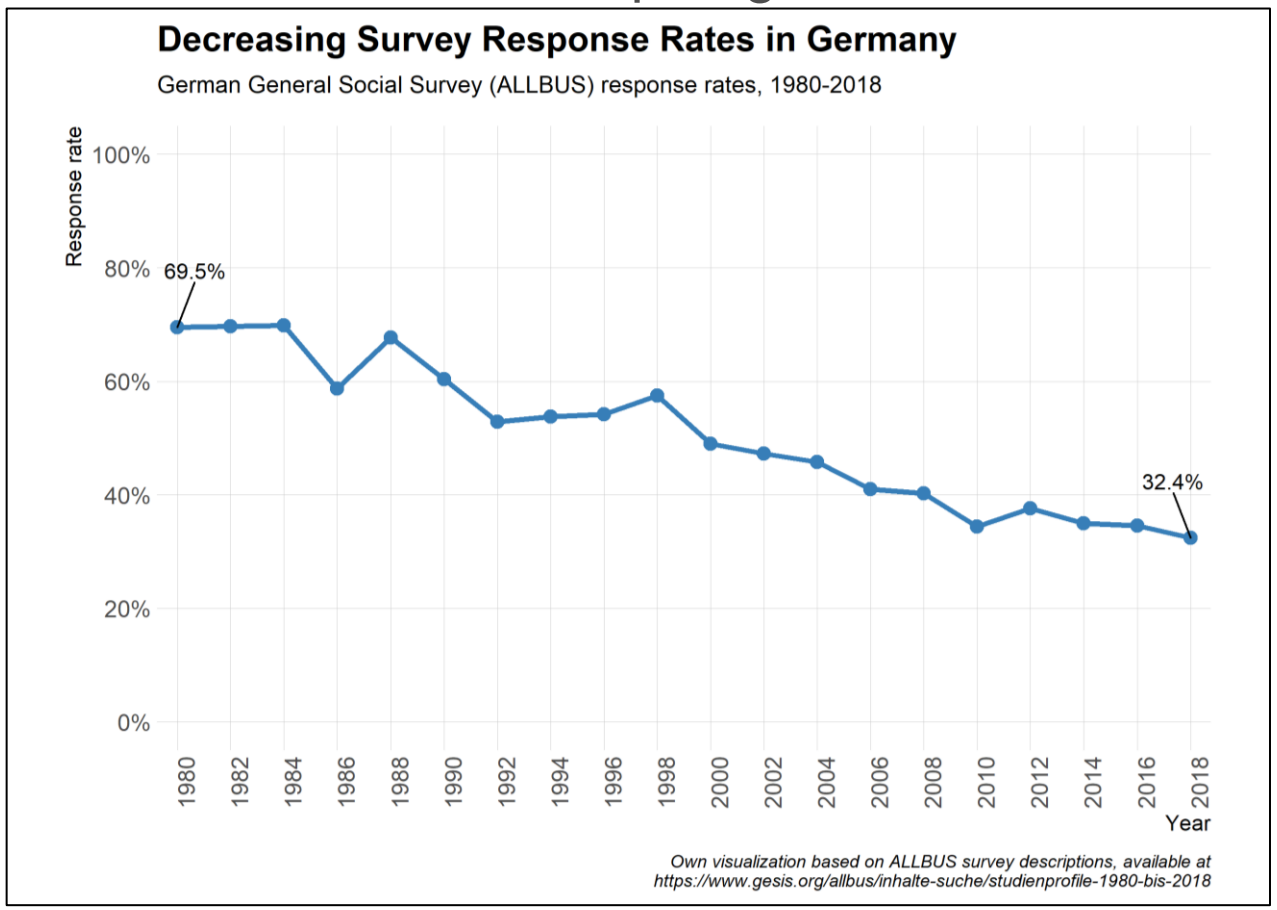
- ITE kann Treatmenteffekt gegen Null verzerren
  - Extremfall: Kein Effekt, bloß weil Probanden unaufmerksam waren
- Bei LATE oder ATT allerdings externe Validität eingeschränkt – es handelt sich um eine spezifische Gruppe (z.B. Personen, die einen Manipulationscheck bestanden haben)
- Bei Interpretationen ist jedenfalls zu beachten, dass man ggf. nur den ITE beobachtet; der ATE ist oft ein unerreichbares Ideal!
- Man sollte ggf. beide Effekte berichten und mögliche Unterschiede transparent diskutieren
  - Sowohl im Hinblick auf die interne und externe Validität
- Insbesondere selektive Attrition ist transparent zu machen und diskutieren
  - Medizinische Studien tun dies zum Teil nicht; selektive Abbrüche bedeuten dann oft Überschätzung der Wirksamkeit und Unterschätzung der Nebenwirkungen!
  - Abhilfen: Strikte Protokolle/Präregistrierungen

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Stichproben, Non-Response und Compliance

Zum Einstieg

- Sinkende Ausschöpfungsraten ALLBUS

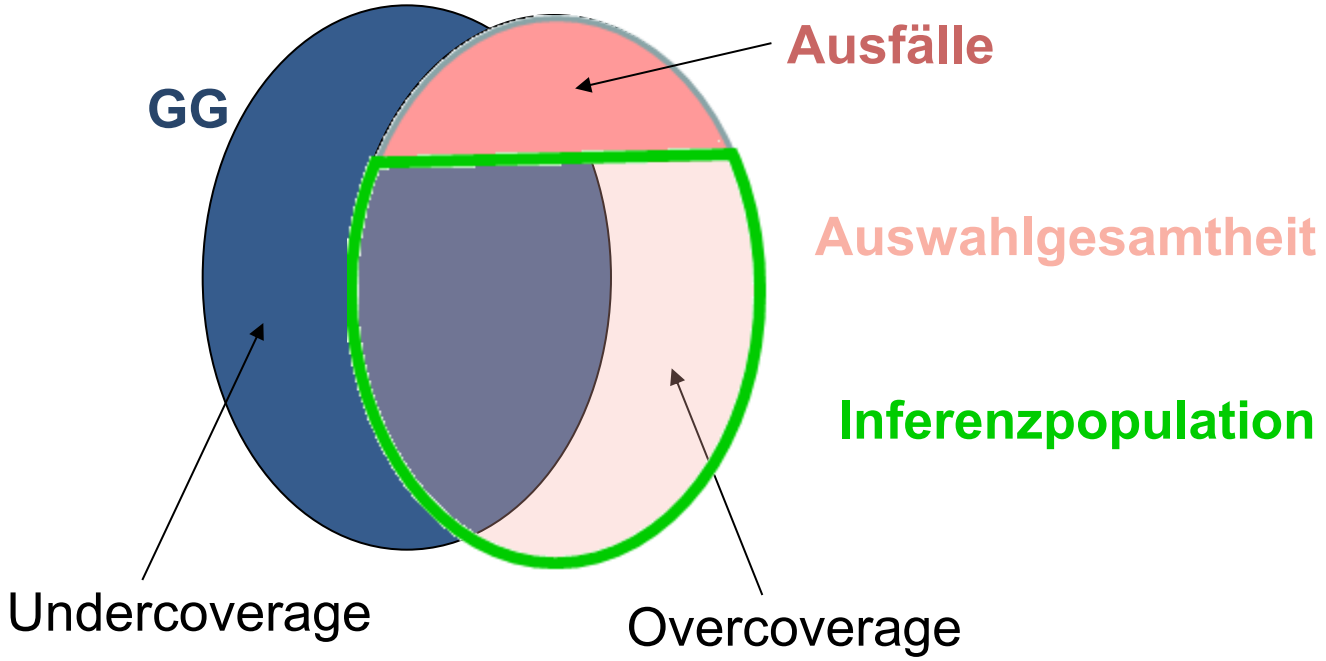


Schwemmer 2022: https://github.com/cschwem2er/allbus_responserates/

- „Ausschöpfungsrate“ ?
- Ist das ein Problem für valide Schätzungen? Wenn ja/nein, warum?
- Alternativ besser nichtzufällige Stichproben verwenden?
  - z.B. Online Access Panel?
  - Da etwa deutlich günstiger und schneller?

Abweichung Grundgesamtheit (GG) und Stichprobe

- **Fehlerhafte** Abbildungen der Grundgesamtheit durch
  - Coverage Errors (insb. Undercoverage problematisch)
  - Nonresponse Error (Ausfälle)
  - Sampling Error (Zufallsschwankungen über Stichproben)
  - (Fehlerhafte Korrekturen / Adjustment Errors)
- Ausschöpfung: Responserate in der bereinigten Bruttostichprobe (ohne Overcoverage)



Erfolgskriterien / Mögliche Abweichungen

- Correctness: $\hat{\theta}_r = \theta$
- Uncertainty: $\sqrt{Var(\hat{\theta})} = SE(\hat{\theta})$
- Unbiasedness: $\mathrm{E}(\hat{\theta}) = \theta \rightarrow Bias(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta$
- Mean squared error (MSE): $Bias(\hat{\theta})^2 + Var(\hat{\theta})$

➢Was meint uncertainty, was bias?

➢Was ist ggf. problematischer?

➢Wozu braucht es zusätzlich den MSE?

Mit r = replication;

θ = paramter of interest [theta]

$\hat{\theta}$ = estimate of parameter of i.

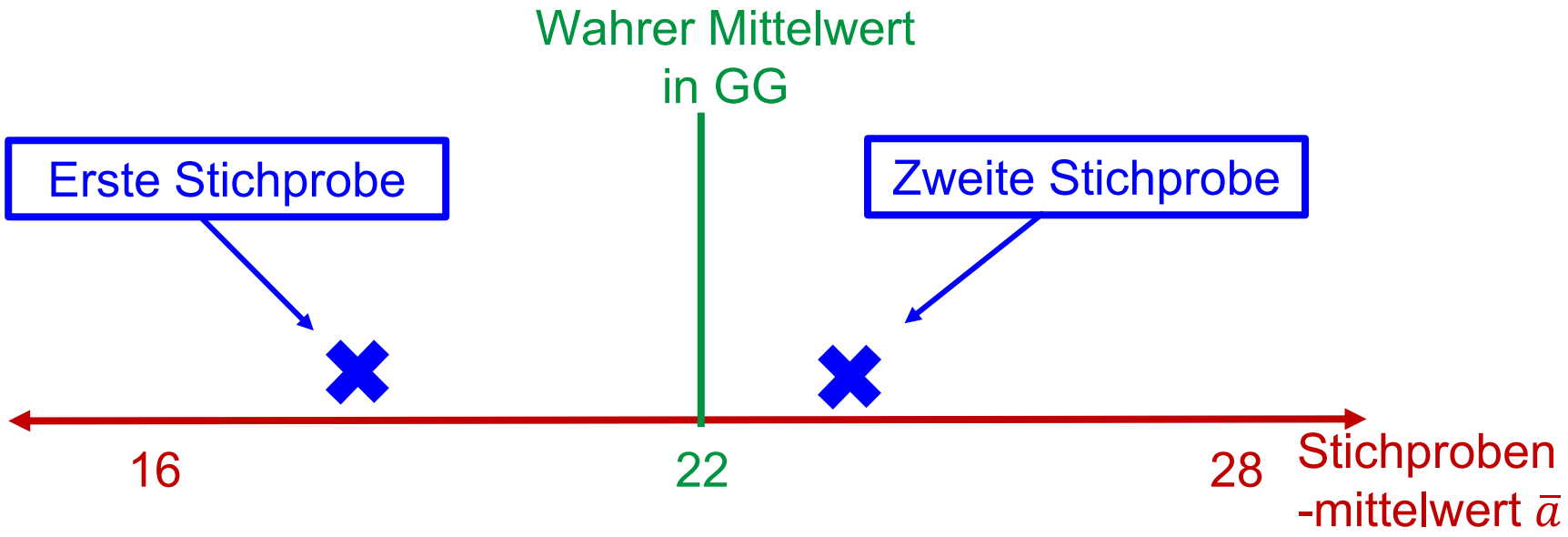
SE = Standard error

E = Erwartungswert

Uncertainty: Die Stichprobenkennwerte-Verteilung (Wiederholung aus Grundstudium)

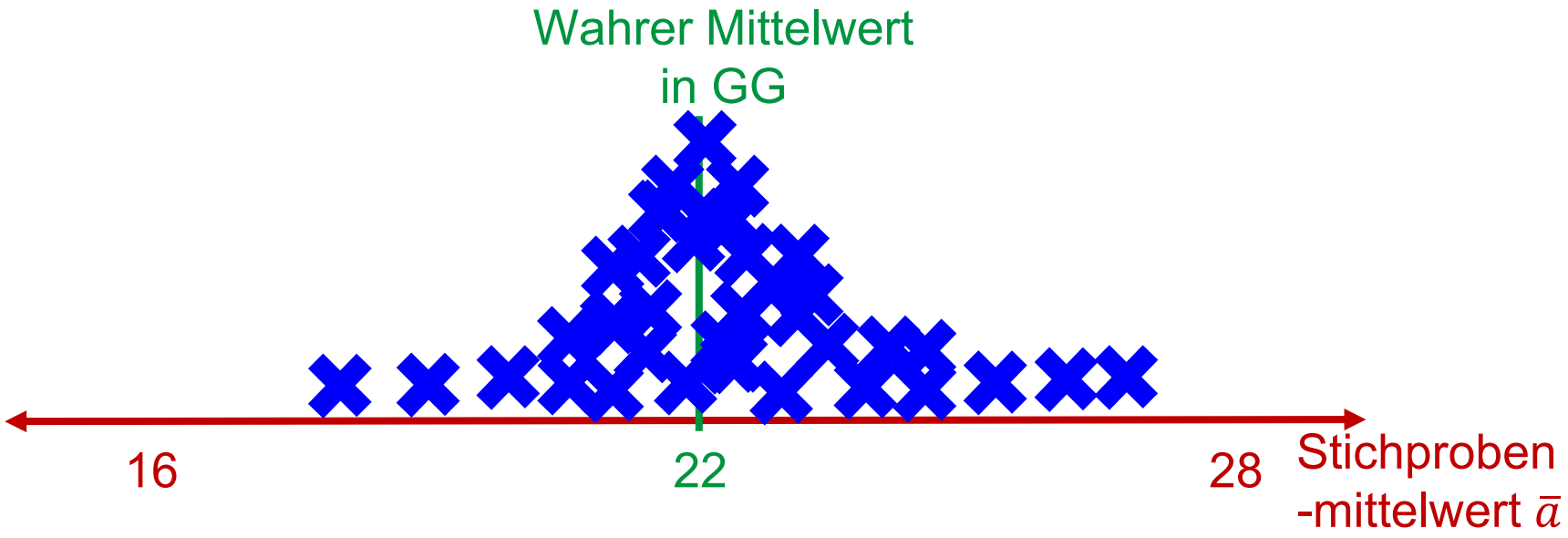
- Angenommen, wir möchten das Durchschnittsalter aller 500 Teilnehmer einer Vorlesung (=GG) berechnen
  - Und wir wissen den wahren Mittelwert (= 22 Jahre)
- Jetzt ziehen wir aus der GG (N = 500) mehrere Zufallsstichproben im Umfang von n = 40 („mit Zurücklegen“)
- Für jede Stichprobe berechnen wir das Durchschnittsalter

Uncertainty: Die Stichprobenkennwerte-Verteilung



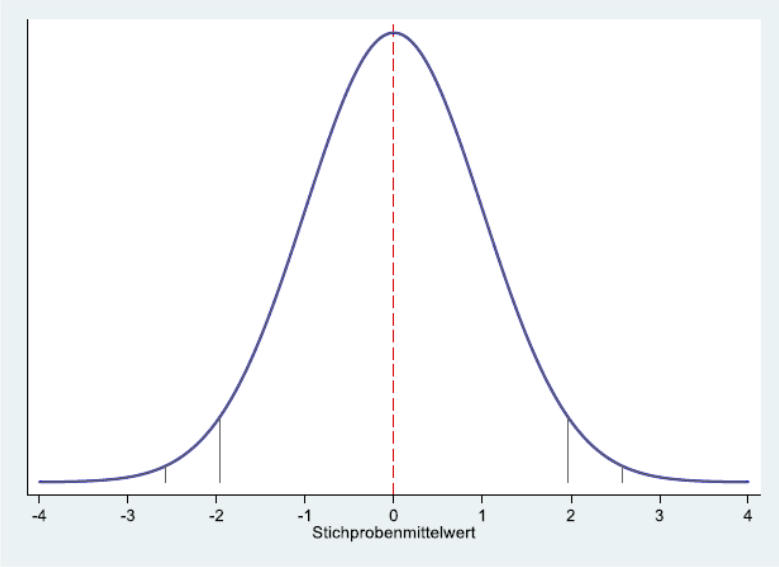
Jetzt ziehen wir unendlich viele Stichproben. Was meinen Sie: Wie verteilen sich die Mittelwerte in den gezogenen Stichproben um den wahren Wert herum?

Uncertainty: Die Stichprobenkennwerte-Verteilung



Die Stichprobenkennwerte-Verteilung

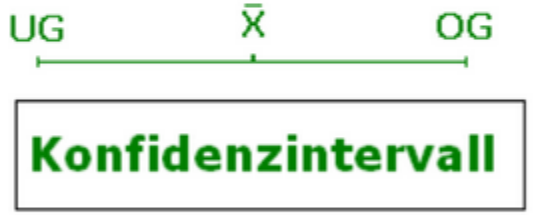
- Bei N Zufallsstichproben verteilen sich die Kennwerte aus der Stichprobe $\hat{\theta}_r$ in Form einer Normalverteilung (NV, „Glockenkurve“) um den wahren Mittelwert
  - Der Mittelwert der Verteilung aller $\hat{\theta}_r$ ist der wahre Kennwert θ
  - Die Streuung ist der Standardfehler $SE(\hat{\theta})$ äquivalent zur Standardabweichung der Verteilung aller $\hat{\theta}_r$!
  - Das kann genutzt werden, um Konfidenzintervalle (KI) zu berechnen



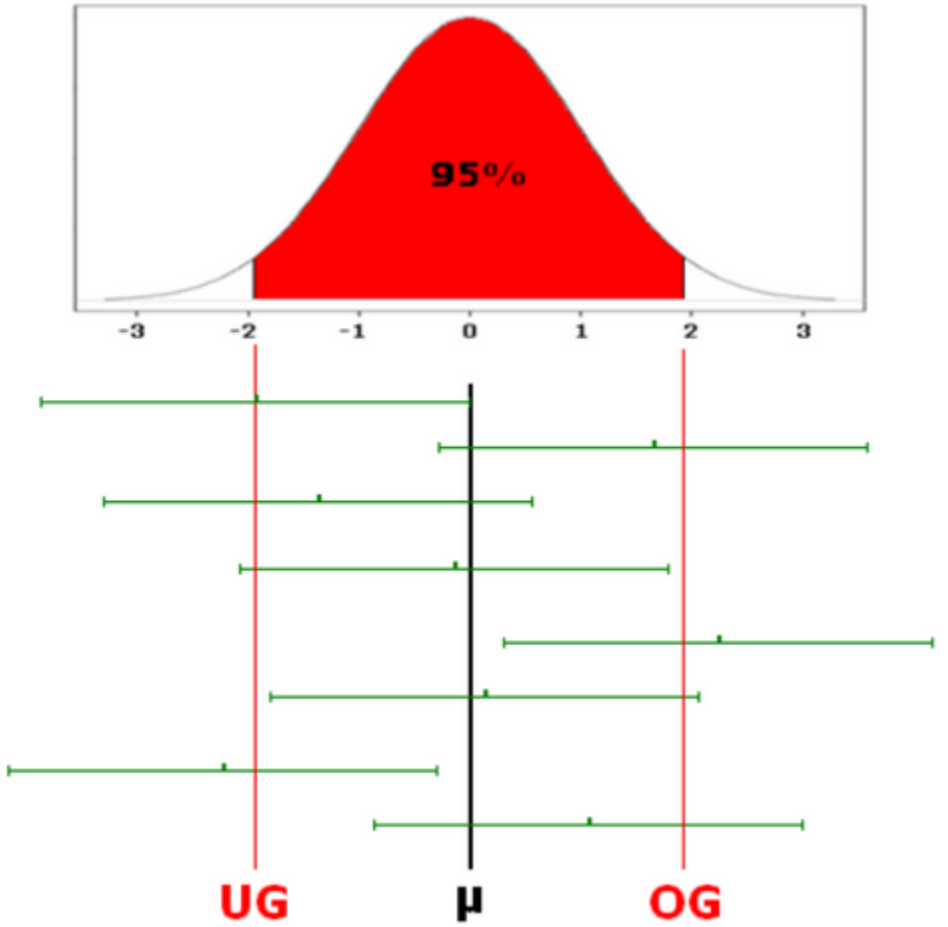
- Es gilt:
  - 68% der Werte liegen im Intervall ± 1 Standardfehler $SE(\hat{\theta})$
  - 95% der Werte im Intervall ± 2 $SE(\hat{\theta})$
  - 99% der Werte im Intervall ± 3 $SE(\hat{\theta})$
- Der Standardfehler $SE(\hat{\theta})$ ist bei großen Stichproben kleiner → genauere Schätzung!

Stichprobenkennwerte-Verteilung und KI

Interpretation Konfidenzintervall: das Intervall beinhaltet den wahren Wert mit 95 % Wahrscheinlichkeit.



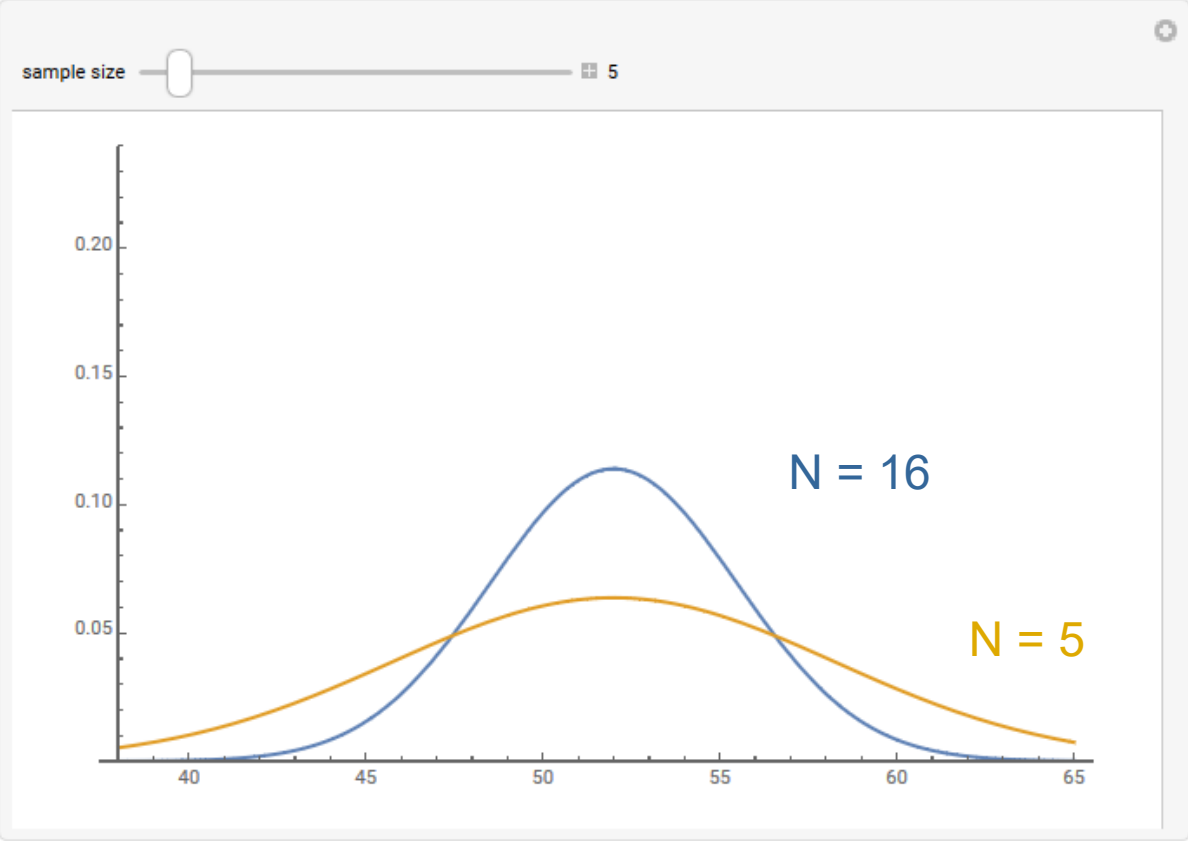
Interpretation Uncertainty: Wie stark schwanken die Schätzwerte $\hat{\theta}_r$ um den wahren Wert θ (wie klein ist der Konfidenzintervall der Verteilung aller $\hat{\theta}_r$)



Quelle: https://www.repetico.de/card-64948882

Stichprobenkennwerte-Verteilung

- Probieren Sie selbst unterschiedlich große Stichproben aus:
  https://demonstrations.wolfram.com/DistributionOfNormalMeansWithDifferentSampleSizes/



Stichprobenkennwerte-Verteilung bei unterschiedlich großen Stichproben
(hier n = 16 bzw. n = 5;
wahrer Mittelwert = 52,
Standardabweichung =14)

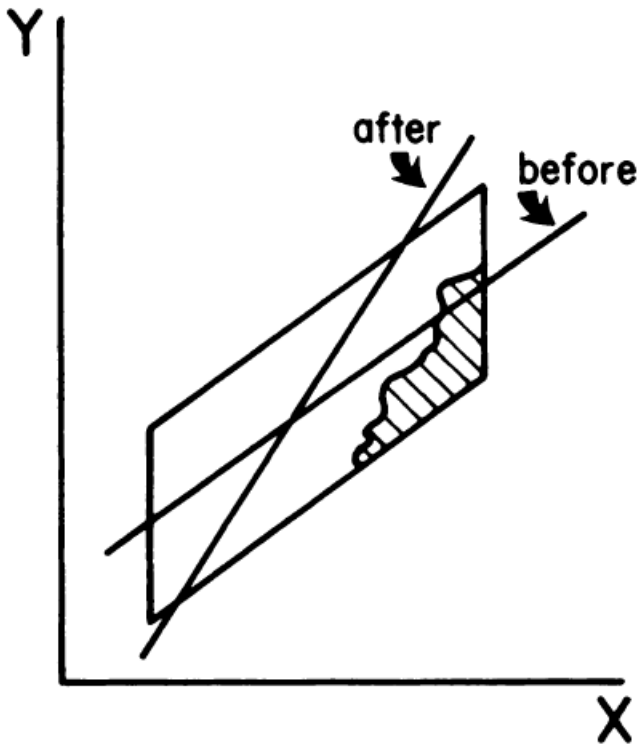
WOLFRAM Demonstrations Project

Bias

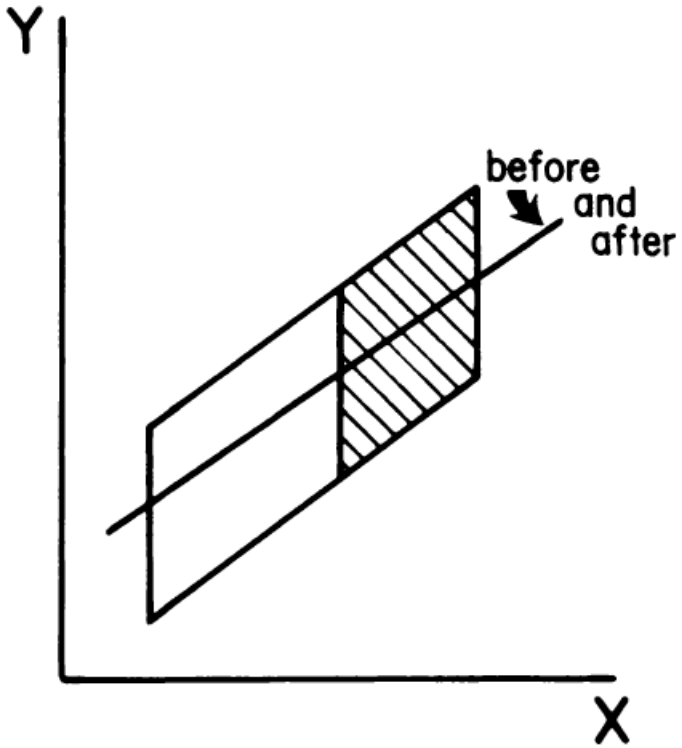
- Entsteht, wenn die Auswahl- oder Teilnahmewahrscheinlichkeit systematisch vom Outcome (der abhängigen Variable) abhängt!
  - Und man das nicht durch Kontrollvariablen/Gewichtung korrigieren kann (die Kausalität der Abweichung bekannt ist)
  - Dann ist die „Conditional Independence Assumption“ verletzt
    - Deskriptive Kennwerte wie Mittelwerte sind verzerrt
    - Treatmenteffekte sind verzerrt, wenn Selektion in die Treatment- bzw. Kontrollgruppe vom Outcome abhängt (bsp.: Personen, die gesünder sind selektieren sich in T mit medizinischem Treatment)
  - Auch große Stichproben können stark verzerrt sein!
    (Trotz hoher certainty; Achtung: KI messen nur certainty, nicht Abwesenheit von Bias!)
- Selektion nach der unabhängigen Variable X führt dagegen zu keiner Verzerrung
- Aber reduziert die Generalisierbarkeit

Bias: Selektion nach X versus Y

Selektion nach Y



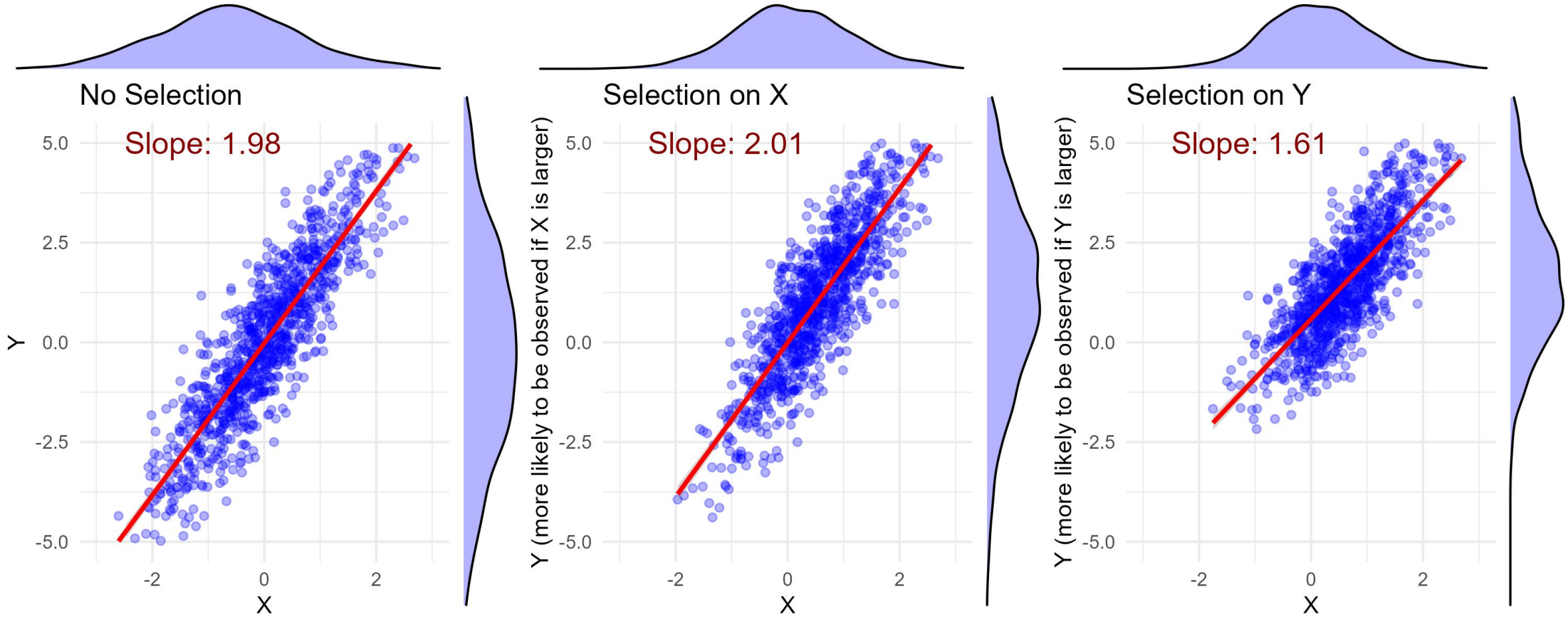
Selektion nach X



Berk 1983

- Bias, sofern die Selektion nicht post-hoc korrigiert werden kann (↓ int. Validität)
- Kein bias, aber fragliche Verallgemeinerbarkeit für unbeachteten Range von X (↓ ext. Validität)

Selektion nach X versus Y



Problem: Bei Selektion nach Y ist die konditionale Unabhängigkeit nicht mehr gegeben da die Korrelation zwischen Y und X jetzt auch durch die Selektion beeinflusst wird

Übungsaufgabe: Movie Stars

Eine häufig gemachte Vorstellung in den Medien ist die Vorstellung dass stärker attraktive Personen weniger talentiert sind in dem, was sie machen.

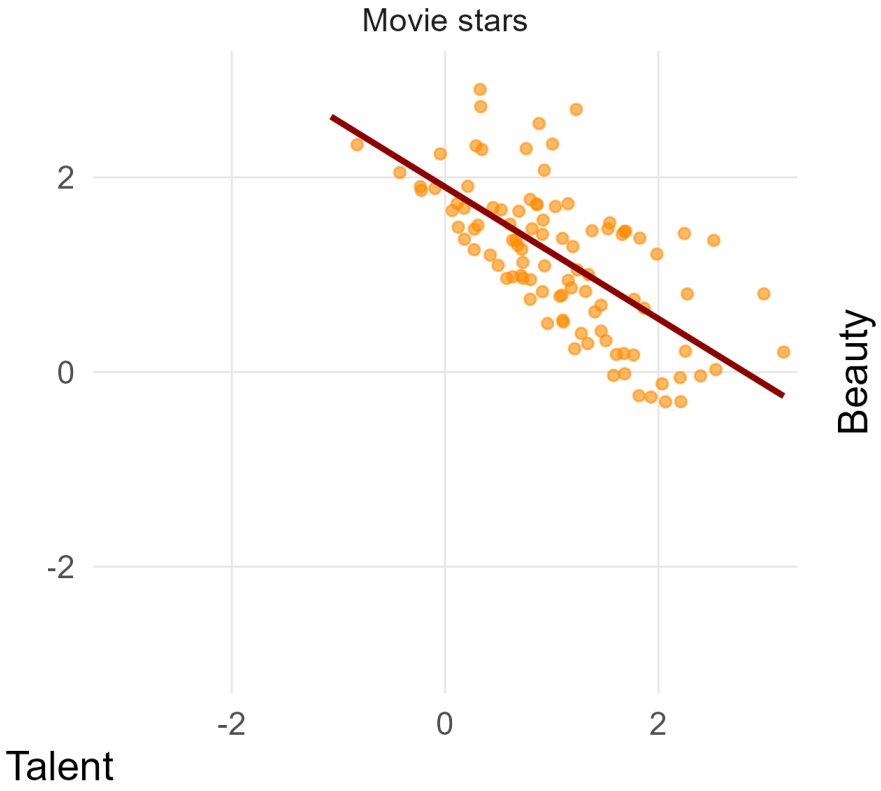
Als Beispiel werden hierfür oft Filmstars genannt (siehe CNN-Blogpost, 2009).

Wie könnte man die Frage beantworten, ob attraktivere Personen tatsächlich weniger talentiert sind?

- Könnte zur Beantwortung dieser Frage die Beziehung zwischen Talent und Attraktivität bei Filmstars hilfreich sein?
- Welche anderen Gründe könnten erklären, warum eine negative Beziehung zwischen Talent und Attraktivität entstehen könnte?

Siehe auch Cunningham Causal Inference (2021: 110 ff)

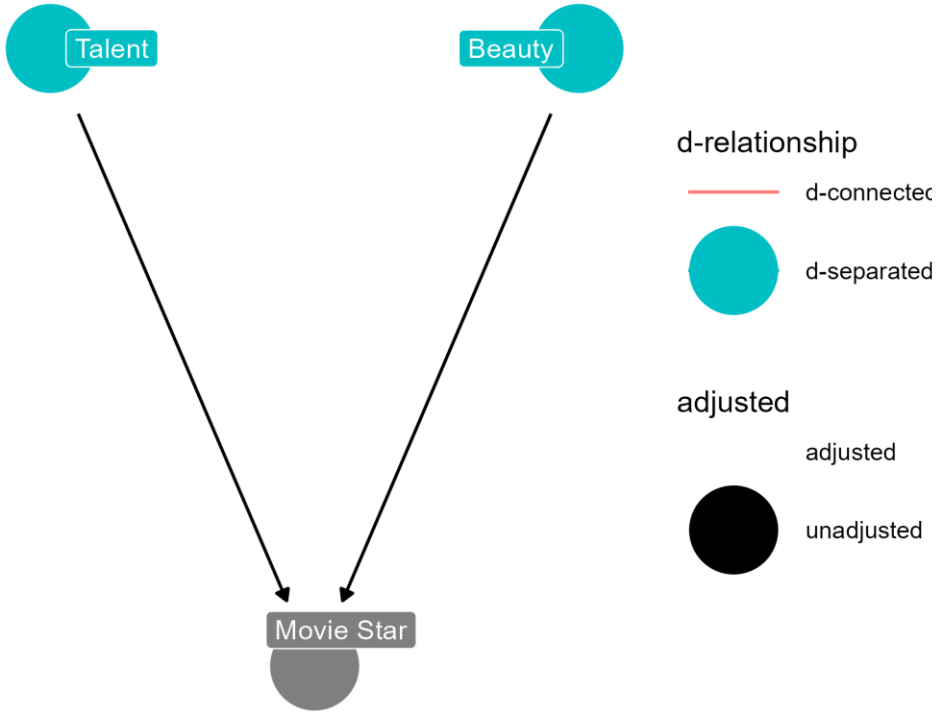
Collider bias und nonresponse bias



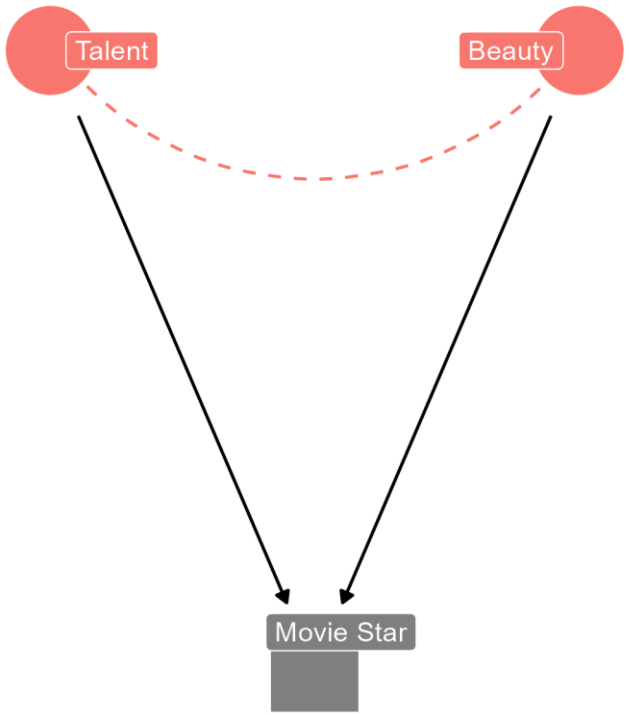
https://colliderbias.herokuapp.com/

Collider bias und nonresponse bias als DAG

Relation ohne Kontrolle auf Movie Star



Relation mit Kontrolle auf Movie Star



	(1)	(2)
(Intercept)	-0.019	-0.273***
	(0.020)	(0.019)
beauty	-0.006	-0.285***
	(0.020)	(0.020)
groupMovie stars		1.673***
		(0.055)
Num.Obs.	2500	2500
R2	0.000	0.272

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Übungsaufgabe: Facebook-Delphi-Survey

Während der Corona-Pandemie wurde versucht, Impfraten oder auch das Verhalten der Bevölkerung mit Surveys zu schätzen. Dazu kamen Probability Samples zum Einsatz, vor allem aber auch Nonprobability Samples wie der „Delphi–Facebook Survey“ (siehe z.B. https://madoc.bib.uni-mannheim.de/58686/1/7761-Article%20Text-25251-3-10-20200603.pdf).

Hier wurden z.B. weltweit Befragte zufällig unter Facebook-Nutzern ausgewählt und über die Plattform zu einem Survey dazu eingeladen, ob sie bereits geimpft sind. Geschätzt werden sollte u.a. die Impfquote. Die Stichproben waren größer als in anderen Studien.

- Was bedeutet das Vorgehen für einen möglichen „Bias“ and Uncertainty“?
- Was gibt es für mögliche weitere Gründe für Abweichungen von der tatsächlichen Impfquote?
- Was könnten trotz möglicher Abweichungen vom wahren Wert Gründe für das Design gewesen sein?
- (Könnte man Abweichungen ggf. (nachträglich) korrigieren?)

Facebook-Delphi-Survey: Ausgewählte Ergebnisse

- Methodik: Vergleich mit anderen Surveys und Benchmark
  - (1) Korrekter Wert: Amtliche Daten (Center for Disease Control & Prevention)

  Wird verglichen mit Schätzungen aus verschiedenen Surveys:
  - (2) Delphi–Facebook Survey (N ~ 250,000 pro Woche)
  - (3) Census Household Pulse (N ~ 75,000 alle 2 Wochen)
  - (4) Axios–Ipsos Online panel (N ~ 1,000 pro Woche)
- Dabei folgt nur (4) weitgehend “Best Survey Praxis” (etwa in Form möglichst zufälliger Stichproben/Offline-Rekrutierungen)
- Hauptergebnis:

*

Nature. 2021 Dec 8;1-6. doi: 10.1038/s41586-021-04198-4. Online ahead of print.

Unrepresentative big surveys significantly overestimated US vaccine uptake

Valerie C Bradley [# 1], Shiro Kuriwaki [# 2], Michael Isakov [3], Dino Sejdinovic [1], Xiao-Li Meng [4], Seth Flaxman [5]

Facebook-Delphi-Survey: Ausgewählte Ergebnisse

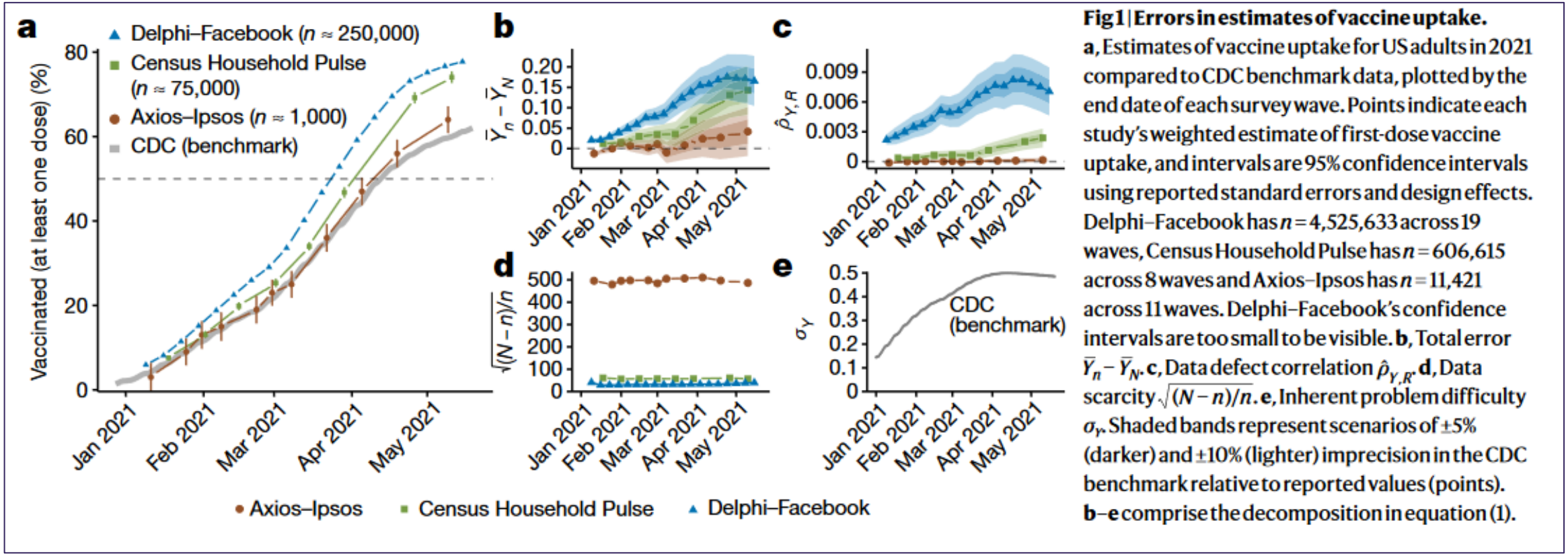


Fig 1 | Errors in estimates of vaccine uptake. **a**, Estimates of vaccine uptake for US adults in 2021 compared to CDC benchmark data, plotted by the end date of each survey wave. Points indicate each study's weighted estimate of first-dose vaccine uptake, and intervals are 95% confidence intervals using reported standard errors and design effects. Delphi–Facebook has $n = 4{,}525{,}633$ across 19 waves, Census Household Pulse has $n = 606{,}615$ across 8 waves and Axios–Ipsos has $n = 11{,}421$ across 11 waves. Delphi–Facebook's confidence intervals are too small to be visible. **b**, Total error $\bar{Y}_n - \bar{Y}_N$. **c**, Data defect correlation $\hat{\rho}_{Y,R}$. **d**, Data scarcity $\sqrt{(N-n)/n}$. **e**, Inherent problem difficulty σ_Y. Shaded bands represent scenarios of ±5% (darker) and ±10% (lighter) imprecision in the CDC benchmark relative to reported values (points). **b**–**e** comprise the decomposition in equation (1).

- Zentrale Take-Home-Message: Daten-Qualität wichtiger als -Quantität!

Bias bei Deskriptionen – Zufallssample (PS)

- Nonresponse-Bias

$$Bias(\bar{y})_{PS} \approx \frac{Corr(Y,\rho) \cdot SD(Y) \cdot SD(\rho)}{\bar{\rho}}$$

Wobei $\bar{\rho} \approx \hat{\bar{\rho}} = \frac{n_{Resp}}{n_{Resp}+n_{Nonresp}}$

➢ Somit umso höher, je …?

- Höher die Korrelation Antwortwahrscheinlichkeit mit Outcome Y
- Höher die Varianz in Y und Antwortwahrscheinlichkeit
- Geringer die Antwortwahrscheinlichkeit (Ausschöpfungsquote) bzw. höher die Nonresponse

Mit

$\bar{y}$	= interessierender Mittelwert
ρ	= Antwortwahrscheinlichkeit
Corr	= Correlation
SD	= Standard deviation
n_{Resp}	= Anzahl Respondents
$n_{\text{Non}-Resp}$	= Anzahl Nonrespondents

Bias bei Deskriptionen – Nicht-Zufallssample (NPS)

- Self-selection-Bias

$$Bias(\bar{y})_{NPS} \approx \frac{Corr(Y,\pi) \cdot SD(Y) \cdot SD(\pi)}{\bar{\pi}}$$

Wobei $\bar{\pi} \approx \hat{\bar{\pi}} = \frac{n_{Resp}}{N}$

➢ Somit umso höher, je…?

- Höher die Korrelation Auswahlwahrscheinlichkeit mit Outcome Y
- Höher die Varianz in Y und Auswahlwahrscheinlichkeit
- Geringer der Anteil der Befragten im Vergleich zur Größe der Population

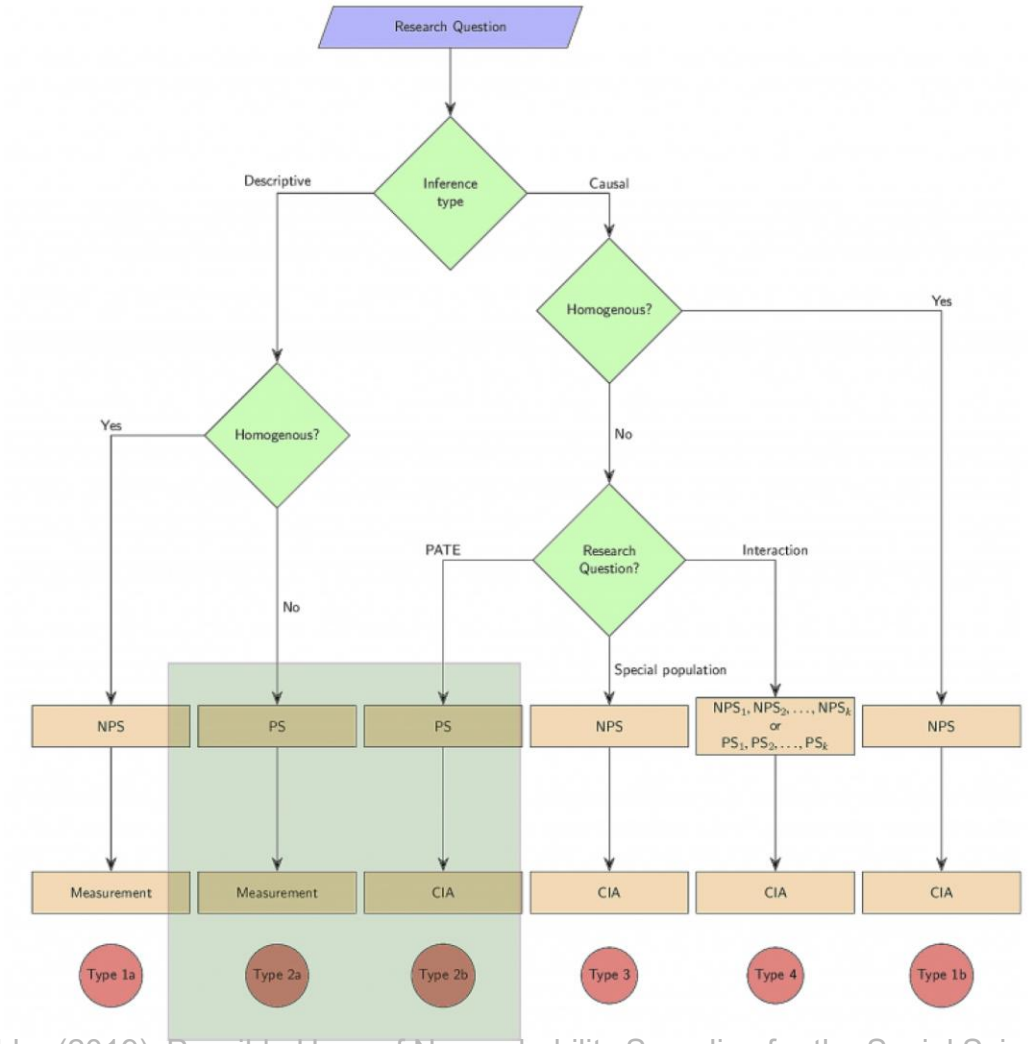
Insbesondere das vergrößert oft den Bias gegenüber PS!

Vergleichbare Schätzgüte nur bei kleinen Spezialpopulationen, die man gut abdecken kann

Mit

$\bar{y}$	= interessierender Mittelwert
π	= Auswahlwahrscheinlichkeit
Corr	= Correlation
SD	= Standard deviation
n_{Resp}	= Anzahl Respondents
N	= Größe der Population

Welches Sample?

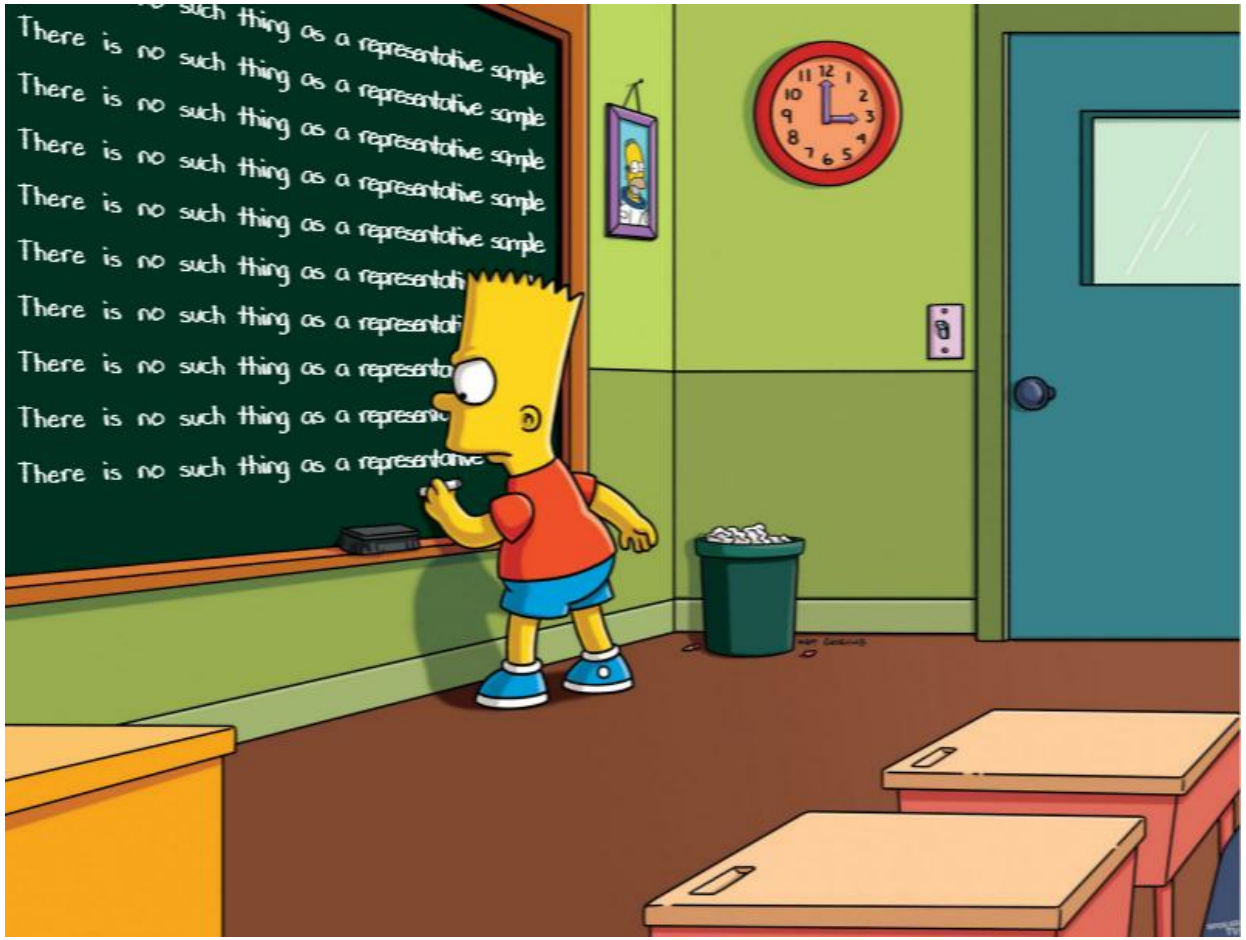


- Relevant sind die Forschungsfrage und Annahmen über die Homogenität von θ
- Bei deskriptiven Fragestellungen sind Zufallsstichproben (PS) nichtzufälligen Stichproben (NPS) vorzuziehen
- Bei kausalen Fragestellungen kommt es drauf an:
  - Bei Homogenitätsannahme oder wenn lediglich interessiert, ob Effekt in bestimmter Population auftritt: NPS
  - Bei Interesse an Effektheterogenität: Samples mit Variation in Moderatoren
  - Bei Interesse an PATE: PS

Kohler (2019); Possible Uses of Nonprobability Sampling for the Social Sciences. Survey Methods: Insights fromthe Field, 1-12.

Nicht zuletzt…

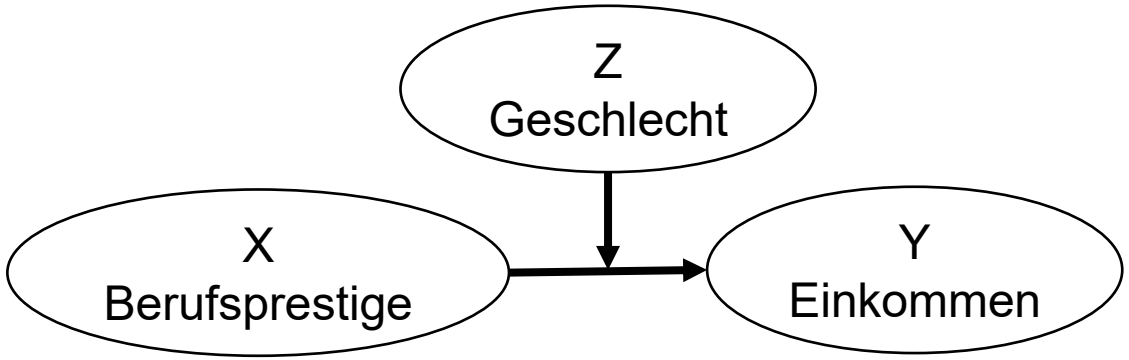
- Es gibt keine „repräsentative“ Stichprobe!
- Das ist lediglich eine „Nebelkerze“
  - Ebenso: „bildet die Bevölkerung gut ab“
- Man sollte transparent beschreiben,
  - Wie die Stichprobe gebildet wurde (Grundgesamtheit, Auswahlverfahren)
  - Teilnahmerate (Ausschöpfung, N)
  - Anzeichen für mögliche Verzerrungen (deskriptive Übersicht, ggf. nach C und T)
- (Siehe auch die Checkliste der AS: https://osf.io/mw59u)



The Simpsons /Picture published by Raphael Nishimura on Twitter

Stichproben: Take-Home Messages Kausalanalysen

- **Zentral ist die Randomisierung des Treatments, nicht die zufällige Stichprobe**
- Stichprobe muss aus Population sein, für die Treatmenteffekt theoretisch zu erwarten ist
- (Wann) ist eine zufällige Stichprobe dennoch sinnvoll?
  - Falls man an der Stärke des ATE für eine Population interessiert ist
  - Dann hat man ähnliche Verteilung von möglichen Moderatoren Z in der Stichprobe
- Was ist, wenn man Effektheterogenität nach Z möglichst genau schätzen will?
  - Kann man besser identifizieren, wenn man viel Varianz in Z hat
    → „purposive Sampling“
    (also z.B. Quoten für Ethnien, Geschlecht, falls man danach variierende Treatmenteffekte testen will)



Stichproben: Take-Home Messages

- **Vorsicht mit Generalisierungen über die getestete Population hinaus!**
  - Diese sind meist spekulativ
  - Etwa Aussagen über ethnische Diskriminierung allgemein, wenn nur ein (Arbeits-)Markt in München zu einem speziellen Zeitpunkt untersucht wurde
- Hier versprechen Studien oft zu viel!

- Allerdings sollte man Kritik auch theoretisch fundiert vorbringen – gibt es plausible Gründe, anderorts andere Effekte zu erwarten?
  - Etwa nicht: Linkshändige Jobanbieter hätten womöglich anders entschieden
  - Aber: In Arbeitsmärkten mit weniger Nachfrage sind Effekte womöglich anders, legen gängige Diskriminierungs-theorien doch eine Moderation mit der Nachfrage nahe

Zusammenfassung: Merkposten externe Validität

Bedrohungen: Effektheterogenität

- Problematisch? Je nach Forschungsziel
  - Schätzung eines generellen Mechanismus (Ziel: Theorieprüfung)
    → Spezielle Samples sind unproblematisch
  - Schätzung der Effektstärke (Ziel: Effektivität einer Intervention)
    → Selektive Samples können zu Verzerrungen führen
- Ähnliches für andere Aspekte externer Validität (Operationalisierung, SUTVA)

Abhilfen

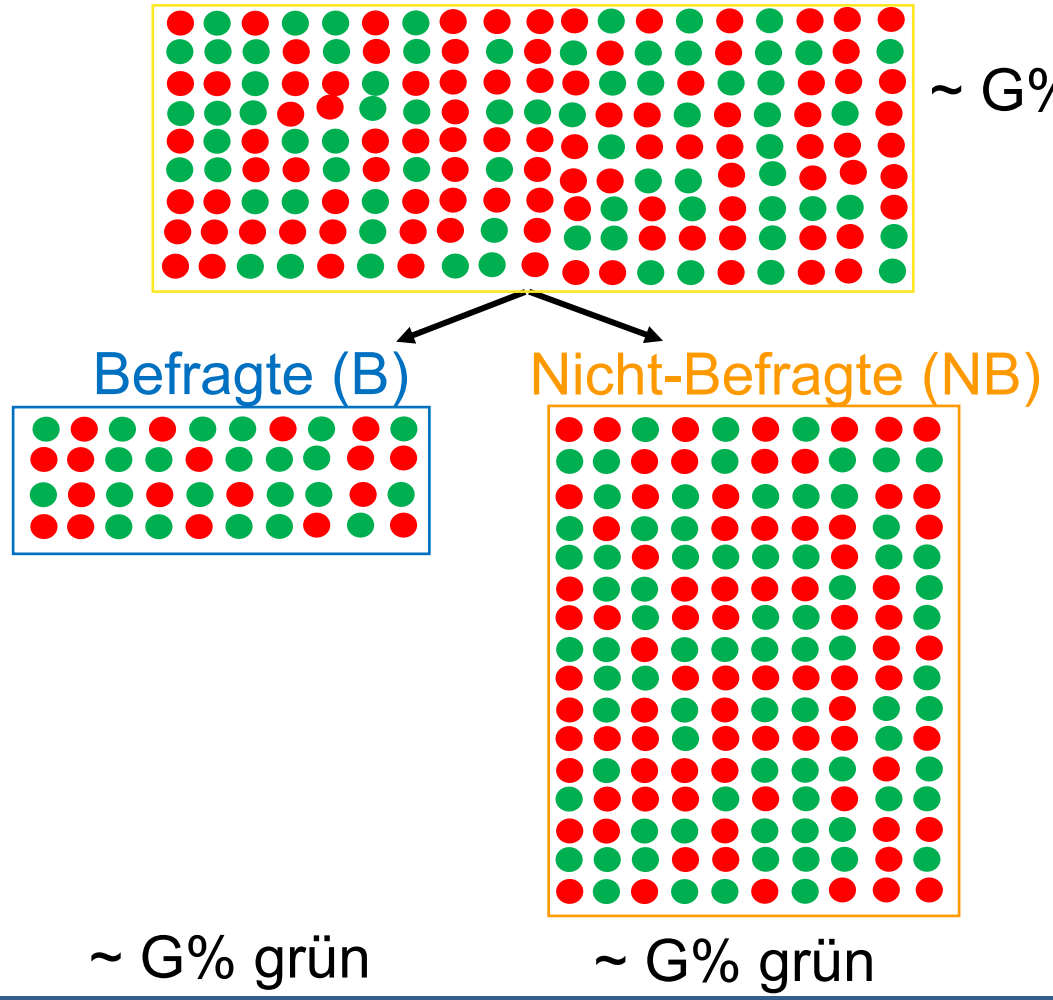
- Generalisierbarkeit beachten/diskutieren
  - Identifikationsziel: Effekt vs. Effektstärke
  - Interessierende Population und zu erwartende Effektheterogenität
  - SUTVA: Verallgemeinerbarkeit für verschiedene N_{treated}
- Replikationen mit anderen (gezielt ausgewählten!) Samples vorteilhaft
  - Stärkt Vertrauen in interne Validität
  - Gibt Aufschluss über Effektheterogenität

Ausblick

- Kommende Woche keine Seminarsitzung
- Übernächste Woche: Diskussion der ersten Übungsaufgabe
  - Abgabe bis 24.11., 11:00 Uhr über Moodle
  - Für alle verpflichtend!
  - Aufgabe: Diskussion von **einer** von zwei Studien, die auf Moodle bereitgestellt sind
  - Details: Siehe Aufgabenblatt oder Beschreibung auf Moodle!
- Zudem: Erste Referatsvorbesprechung (Lea Kreppold, Natalia Velic)

Unbiasedness: Response unabhängig von Y

- Response unabhängig von Merkmal grün/rot

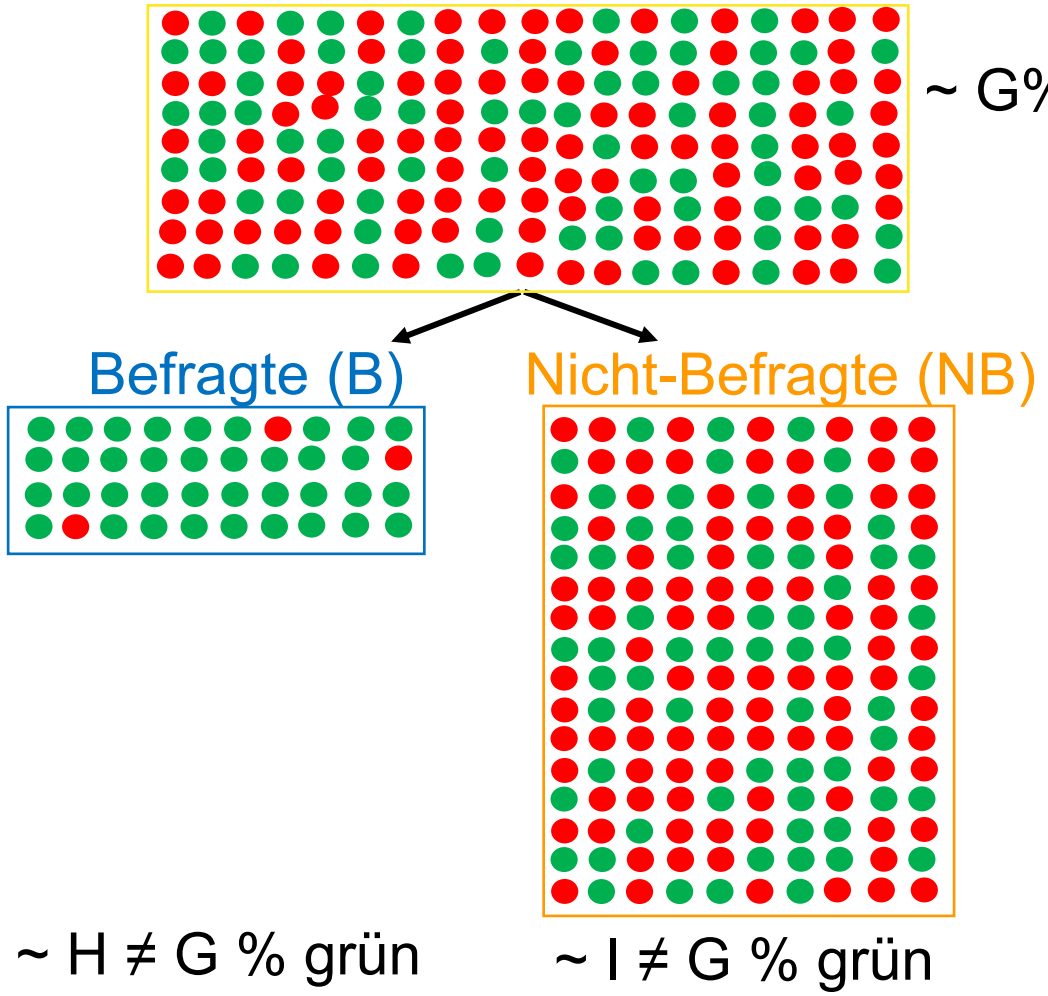


→ Schätzung basierend auf B leidet nur unter zufälligem „Sampling Error“ (Ø = 0)

→ Stichprobe gutes Abbild der Grundgesamtheit

Bias: Response abhängig von Y (Selelektion nach Y)

- Response abhängig von Merkmal grün/rot



→ Schätzung basierend auf B bedeutet *systematische* Über-/ oder Unterschätzung!

→ Stichprobe kein gutes Abbild der Grundgesamtheit

„Labor-Ratten“ und andere spezielle Samples

- Zumindest die meisten psychologischen Experimente wurden mit den „WEIRDest“ People durchgeführt (Henrichs et al.)
- Speziell bei Laborexperimenten sehr selektive Teilnehmende
  - Verpflichtung durch Studium
  - Spaß an Experimenten, Interesse an Incentives, Wissenschaftsaffine, …
  - Erfahrene VP
  - …

→ Ist das ein Problem?

→ Wenn ja, warum?

- Generalisierbarkeit ist jedenfalls zu diskutieren
  - Effekte in Gesamtpopulation oft anders, falls selektives Sample getestet wurde
  - Oft Überschätzung der Effekte!
  - Interventionen dann nicht skalierbar

Beispiel: Diktatorspiel nach Land/Geschlecht

- Rechts: Anteile, die in Diktatorspielen geteilt wurden
  („Violinplots“ zur Dichteverteilung;
  Linie = Median, Punkt = Arithm. Mittel)
- Varianz ist in der Regel Treatmenteffektheterogenität, die inhaltich interessant ist!
  (Nicht unbedingt Zeichen mangelnder interner Validität)

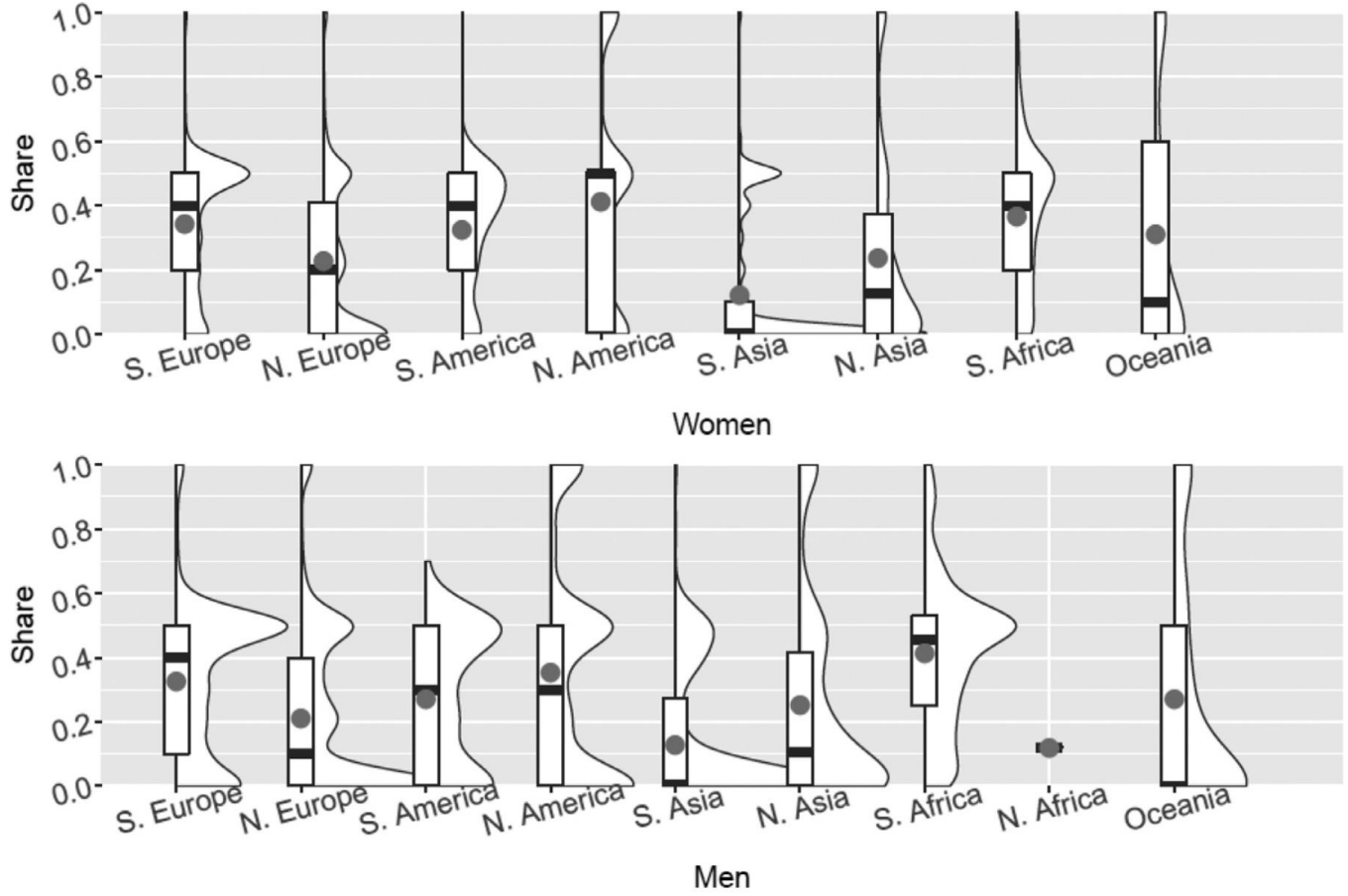


Fig. 6. Violin plot showing the share by location within each continent and gender.

Meta-Analyse von N = 136 Studien; Doñate-Buendía et al. 2022

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Anwendung auf zwei exemplarische Studien

Galos 2024

Whilson/Whitt 2006

Organisatorisches

Zeitraum für die Prüfungsanmeldung: **25.11. - 11.12.2024**

Abgabe Übungsaufgaben

- Titelfolie mit Namen
- Datei mit Namen
- .ppt statt .pdf

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Leitfragen

1. **Fragestellung**: Was ist die Fragestellung und welcher Effekt soll für welche Grundgesamtheit identifiziert werden? (DAG)
2. **Design**: Welches Forschungsdesign wurde gewählt und wie soll der welche Art von Treatmenteffekt (ATE, ITE, …?) identifiziert werden? Wie wird eine Verallgemeinerung über die untersuchte Stichprobe hinaus sichergestellt?
3. **Innovationsgehalt**: Warum ist die Fragestellung in Kombination mit dem Forschungsdesign interessant und relevant?
4. **Bedrohungen der Validität**: Gibt es Einschränkungen der internen bzw. der externen Validität? (z.B. Randomisierung, Messfehler, etc.)

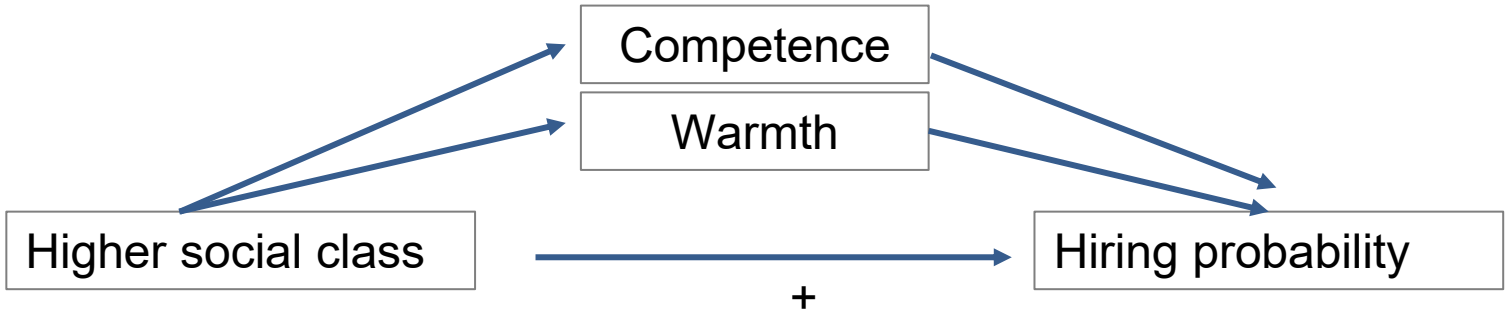
Galos, Diana Roxana. 2024. Social media and hiring: a survey experiment on discrimination based on online social class cues.

1. Theoretical background / research question (RQ)

- **Information asymmetry** accompanies the hiring process and employers seek to acquire optimal information about job seekers
- The information available from social networking platforms may plausibly generate social **class cues**, through individuals' **cultural consumption in terms of their taste and interests**
- RQ: "I test whether social networking platforms contribute to discrimination based on social class"
- Social mechanism?
  - For what should class information be used for?
  - To infer performance (statistical discrimination) or cultural match (taste based discrimination)? (proxied using perceived competence and warmth)

1. Research question more precise, visualization of assumed treatment effects (DAG)

- Research question / hypotheses to test



- ATE or ITT?
  - „the primary treatment is perceived social class in terms of cultural consumption on the social networking platform ‘Twitter’”

2. Design: The Treatment

- Sample: Participants in Online Access Panel, Quota-Sample

2. Design: Sample, outcome measure, identification of treatment effect

- Focus on: Twitter, IT Sector, Men
- Online experiment administrated in the US in May 2020 using ‘Lucid’: 995 respondents
- Respondents are asked to choose between the **job application of two fictitious candidates**
  - Candidate A was one of the two treatment conditions (upper-class-signalling or lower-class-signalling Twitter profile; validated in separate survey)
  - Candidate B was the control condition (random Twitter profiles with no manipulation of social class)
- The effect expresses the difference between the upper-class profiles and the lower-class profiles (subtracting the control profile in both, thus yielding the difference-in-differences estimate).

- 2nd experiment experimentally testing the effect of Candidate A vs. B on perceived warmth and competence

2. ATE or ITT?

- „the primary treatment is perceived social class in terms of cultural consumption on the social networking platform ‘Twitter’”
- Measurement of ITT, because validation survey was with different participants; some participants might have been inattentive etc.
- One more optimal solution would be to calculate the LATE (Local average treatment effect) using a 2SLS (two stage leased squares regression)

- The 2SLS calculates the effect of people’s beliefs about the social class of the candidate that changed solely because of treatment assignment on hiring probability

3. Innovationsgehalt? Gain of Knowledge?

- Experiment allows separating treatment from confounders that likely exist in observational data
  - Education
  - Unobserved confounders
- Additional experiment probes on possible mechanisms
- However, there are some limitations…

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Results

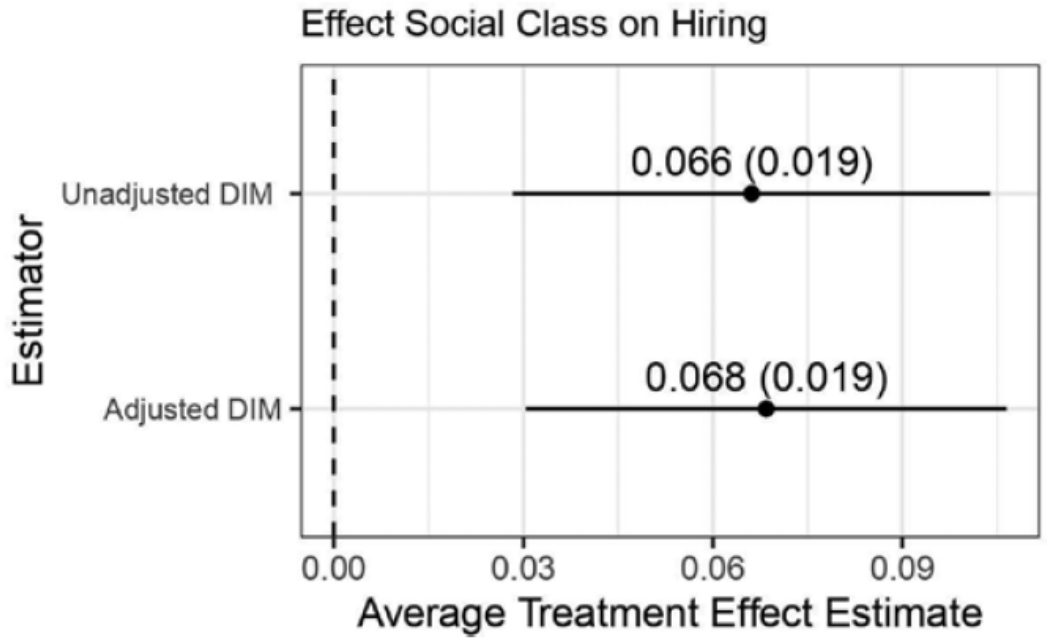
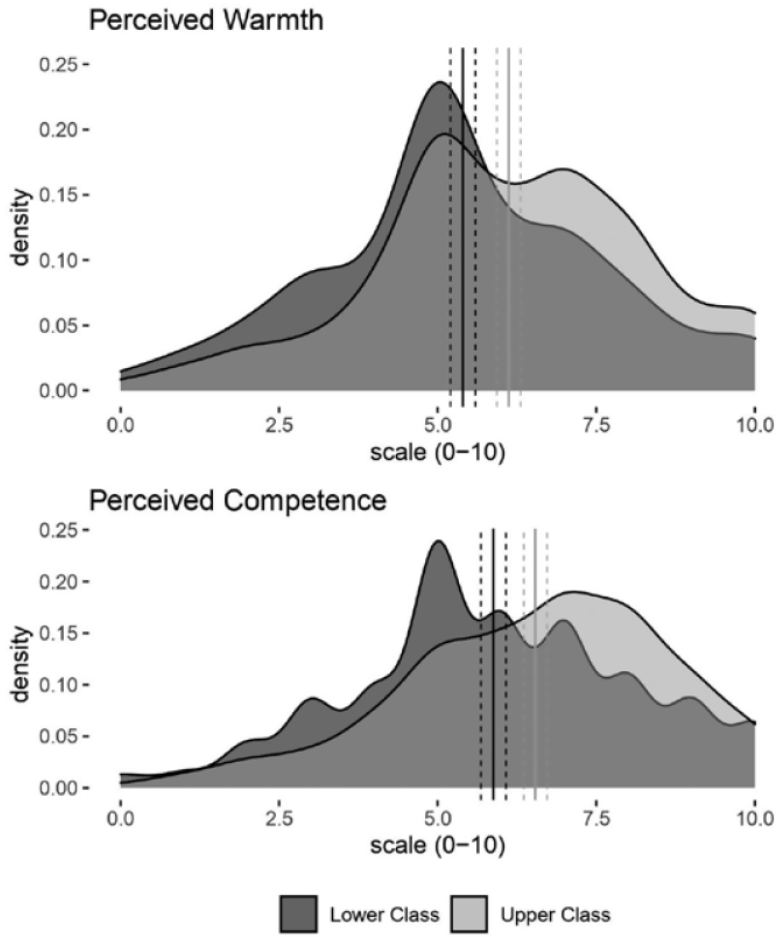


Figure 3 Estimated average treatment effect.



4. Validity? (Further) issues?

„exclusion restriction“

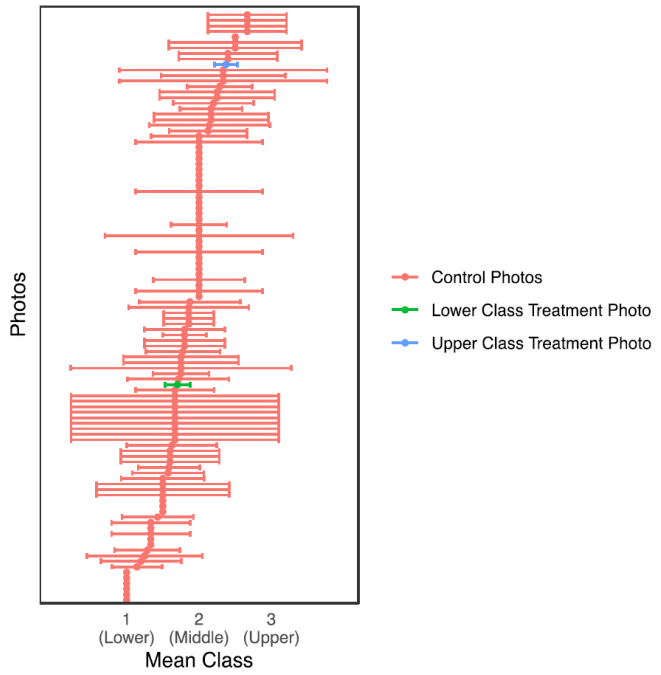


Figure AIII. Perceived Social Class of Names
Note: Mean photos and standard error.

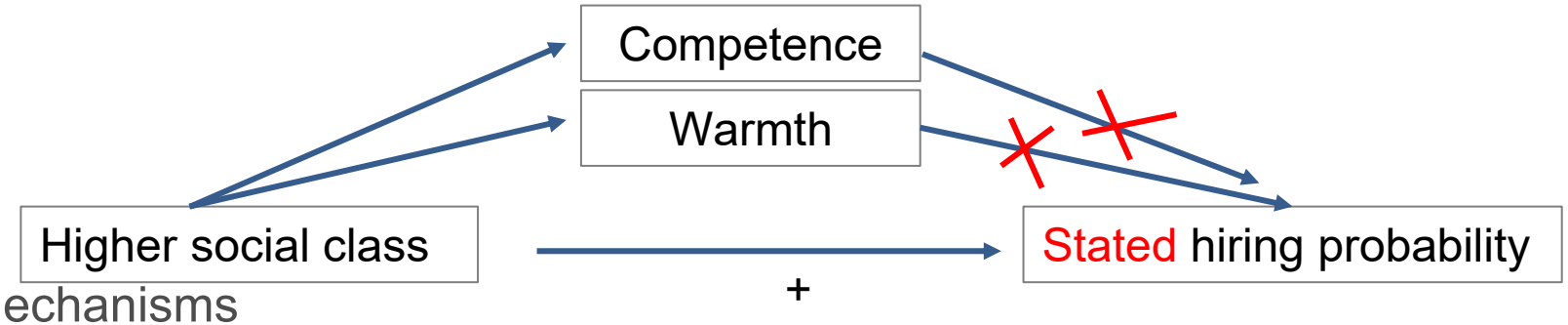
Does the treatment effect the outcome only through the variation of class?

Internal and external validity

- “Strikingly, this is ostensibly the only piece of information respondents use in their evaluation. As Supplementary Appendix, Table AII shows, neither the universities nor high schools of the fictitious candidates nor their characteristics (state of residence, gender, age, and income) seem to influence their chances of being hired.”

How is this result compatible with employers’ struggle to gather information?

4. Validity? (Further) issues?



- Internal validity
  - Social desirability bias?
  - Authors can only test part of mechanisms
  - Is warmth a good proxy for trustworthiness? Other measurement errors?
- External validity
  - Does this generalize to real employers and their behavior?
    - Can one for example assume that employers will check social media plattforms? Be able to identify the right candidates there?
    - Real discrimination comes with costs, survey responses not
    - Different experience with hiring
  - Other segments of the labor market? Can one assume lower effects for women?
  - The authors do not specify the target population very clearly; and only vaguely discuss generalizability, e.g. by assuming an upper bound effect (lower effects for women who would not be discriminated against in the IT sector. Does this make sense here?)

(Further) issues?

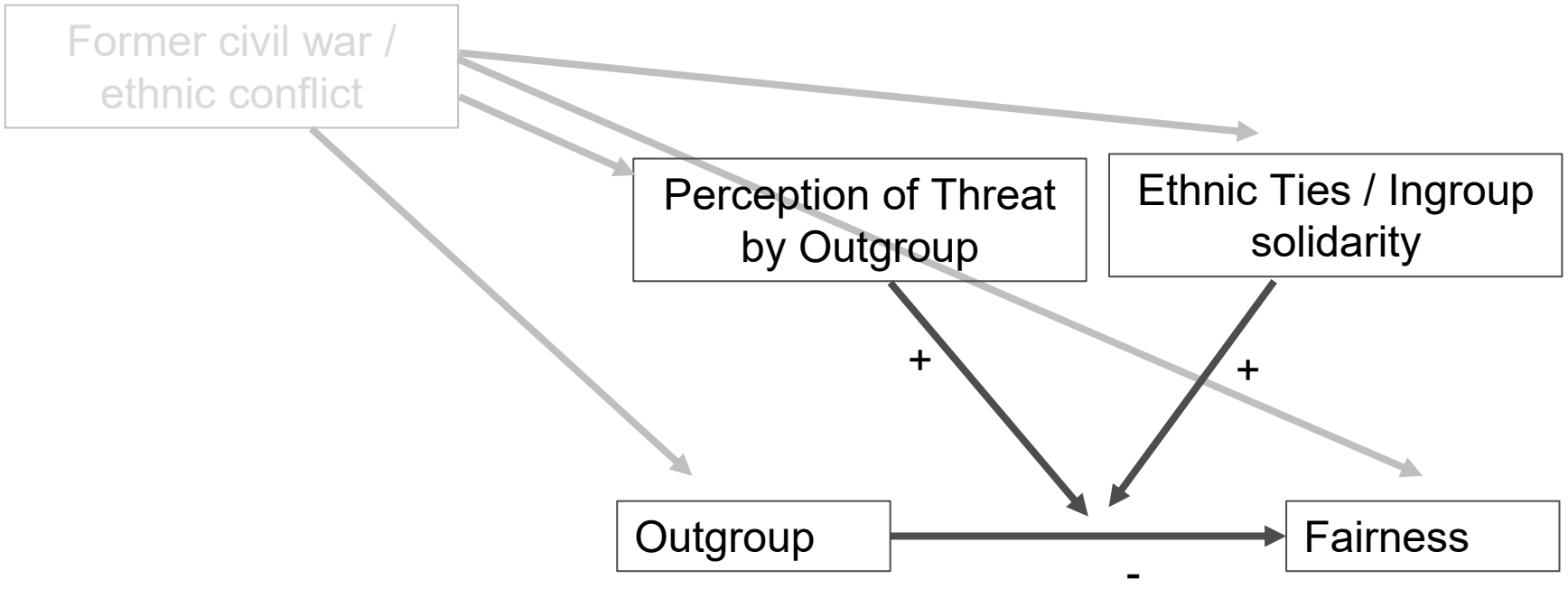
- Violation of research ethics?
  - Design in line with ethic guidelines: survey with consent to participate, no strong deception
  - But usage of other Twitter profiles?
- Other issues? SUTVA?
  - Probably violated, class signals work only when they are exceptional

Whitt, Sam, und Rick Wilson. 2007. The Dictator Game, Fairness and Ethnicity in Postwar Bosnia.

1. Theoretical background / Research question

- Violence might affect the basis for cooperative action – fairness norms
  - Social norms can prohibit or enable individuals to differentiate between ingroup and outgroup
- Research question?
  - "This study considers the effects of ethnic violence on norms of fairness"
  - "Do ethnic groups from Bosnia exhibit norms of fairness that are different for their own ingroup than for an outgroup?"
- Social mechanism?
  - Perceived threat of outgroup – might accelerate ingroup favouritism / outgroup discrimination
  - (inner) Ethnic "ties" – might accelerate ingroup favouritism

1. Causal Diagram



2. Design: The Treatment

In the **first dictator game**, D-1, Player A (i.e., the "Allocator") and Player B (the "Recipient") are of the **same ethnicity and reside in the same federal entity of Bosnia**.

The **second dictator game**, D-2, is similar to D-1, except that in this game, the anonymous **recipient is ethnically different from the subject**.

- The subject is given 10 Bosnian Convertible Marks (KM) and 10 blank slips of paper that were the same size as the bank notes.
- Subjects decided how to allocate the money and the blank slips between themselves and an anonymous recipient.
- Subjects are instructed to place 10 items in both the KEEP and SEND envelopes.
- The **SEND envelopes are marked as going to an individual of a specific ethnicity and place of residence**.
- Subjects are told the recipient will participate in a future experiment.

2. Design: Sample, setting, further details

- Focus on: Post-war Bosnia, Ethnic groups, Fairness as altruistic behaviour, dictator game, ethnic discrimination as ingroup favouritism
- Five stage stratified random sampling with private firm
- Experiments took place in hotel conference rooms, local cultural centres, or schools

- High earnings (~ a daily wage for the experiments)
- Questionnaire before the experiment to measure
  - Threat perceptions, attachment of ethnic ingroup, sociodemographics

3. Relevance

- Lack of research in this area
- Survey data might be „cheap talk“; here: measurement of behavior, experiment with high monetary incentives that undermine social desirability bias
- Experimental data allow eliminating confounders
- Additional survey data allow studying moderators
  (but also trigger social desirability bias?)

Results

FIGURE 2 Overall Distribution of What Was Sent to Ingroup versus Outgroup Counterparts in Decision 1 and Decision 2

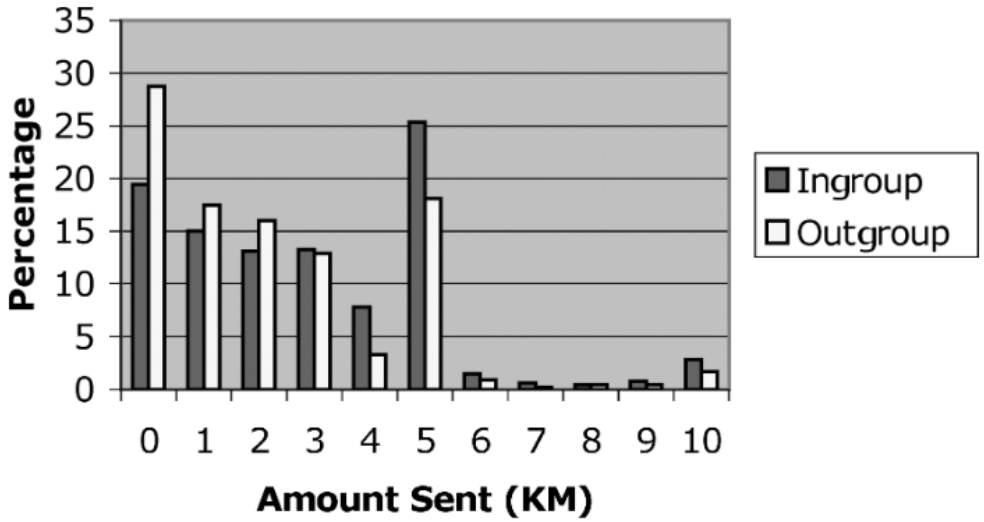


TABLE 3 Tobit Estimates of What Is Sent to a Non-co-ethnic (D-2) Broken Down by Ethnicity of the Allocator

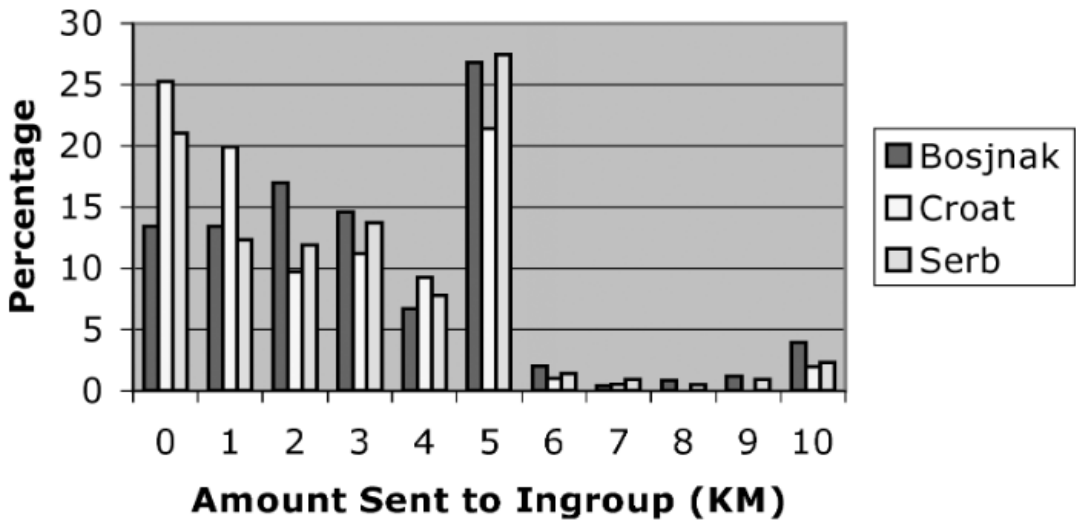
	Bosnjaks		
	Model 1	Model 2	Model 3
Decision 1	.807***	.807***	.799***
	(.054)	(.054)	(.054)
Age	−.003	−.007	−.007
	(.010)	(.010)	(.010)
Female	−.095	−.045	−.057
	(.249)	(.247)	(.247)
Threat	−.262*		−.183
	(.147)		(.149)
Ethnic Ties		−.475***	−.421**
		(.182)	(.187)
Serb Other	.078	.048	.058
	(.248)	(.247)	(.246)
Constant	.148	.726	1.045
	(.578)	(.646)	(.693)
LL	−442.216	−440.444	−439.693
r^2	.17	.17	.17
N	252	252	252

Notes: These models are estimated for each ethnic group. Standard errors are in parentheses.
***Significant at $p \leq 0.01$, **Significant at $p \leq 0.05$, *Significant at $p \leq 0.10$.

4. Validity: (Further) issues?

Minorities give less?

FIGURE 1 Distribution of Amounts Sent in Decision 1 to Ingroup Recipients



What do the offers in dictator game 1 (within ethnic group) measure?

Internal and external validity

- Internal
  - Place of experiments & drop-outs / learning effects
  - „55% of subjects do exactly the same thing in both decisions“: how could this bias results?
  - Evidence of ingroup solidarity or outgroup opposition? Moderators are observational
  - Trust in anonymity?
- External / are conclusions and generalizations based on data?
  - Selective sample and setting
  - „Findings indicate that a norm of reciprocity can emerge“; „We conclude that a norm of fairness is stronger than expected“

➢ Many conclusions exaggerated! There was, for example, no observation of social change!

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Feldexperimente

Galos, Diana Roxana. 2024. Social media and hiring: a survey experiment on discrimination based on online social class cues.

Table AII. Linear Probability Models predicting Candidate Choice

	Unadjusted DIM			Adjusted DIM		
	Coef.	S.E.	p-value	Coef.	S.E.	p-value
Media Profile (Candidate A)						
Lower-class (ref)						
Upper-class	0.066	0.019	0.000	0.068	0.019	0.000
Media Profile (Candidate B)						
Control profile 1						
Control profile 2				0.123	0.034	0.000
Control profile 3				0.111	0.035	0.001
Control profile 4				0.102	0.034	0.003
Control profile 5				0.064	0.036	0.072
University (Candidate A)						
Durham University (ref)						
New York University				-0.027	0.030	0.351
University of Chicago				0.022	0.030	0.454
University of North Carolina at Chapel Hill				-0.005	0.030	0.866
University (Candidate B)						
Indiana University (ref)						
North Carolina State University				-0.002	0.030	0.944
University of Iowa				0.034	0.029	0.231
University of Rochester				-0.006	0.030	0.835
High School (Candidate A)						
Asheville High School (ref)						
Ravenscroft High School				-0.056	0.030	0.067
Roycemore High School				-0.051	0.029	0.083
Trinity High School				-0.084	0.029	0.004
High School (Candidate B)						
Bloomington Hills High School (ref)						
Gilbert High School				-0.014	0.029	0.629
Marvin Ridge High School				-0.006	0.030	0.848
Stuyvesant High School				0.016	0.029	0.590
Gender (Respondents)						
Women (ref)						
Men				-0.039	0.023	0.090
Other				0.092	0.155	0.612
Age (Respondents)				0.001	0.000	0.011
Education (Respondents)						
BA degree (or more) (ref.)						
Some college (no degree)				-0.009	0.030	0.751
High school degree (or less)				-0.036	0.030	0.226
Constant	0.646	0.015	0.000	0.699	0.145	0.000
R^2	0.004			0.031		
N (individuals/observations)	995/1965			995/1965		

Note: Respondents' income and state of residence are also included in the analysis.

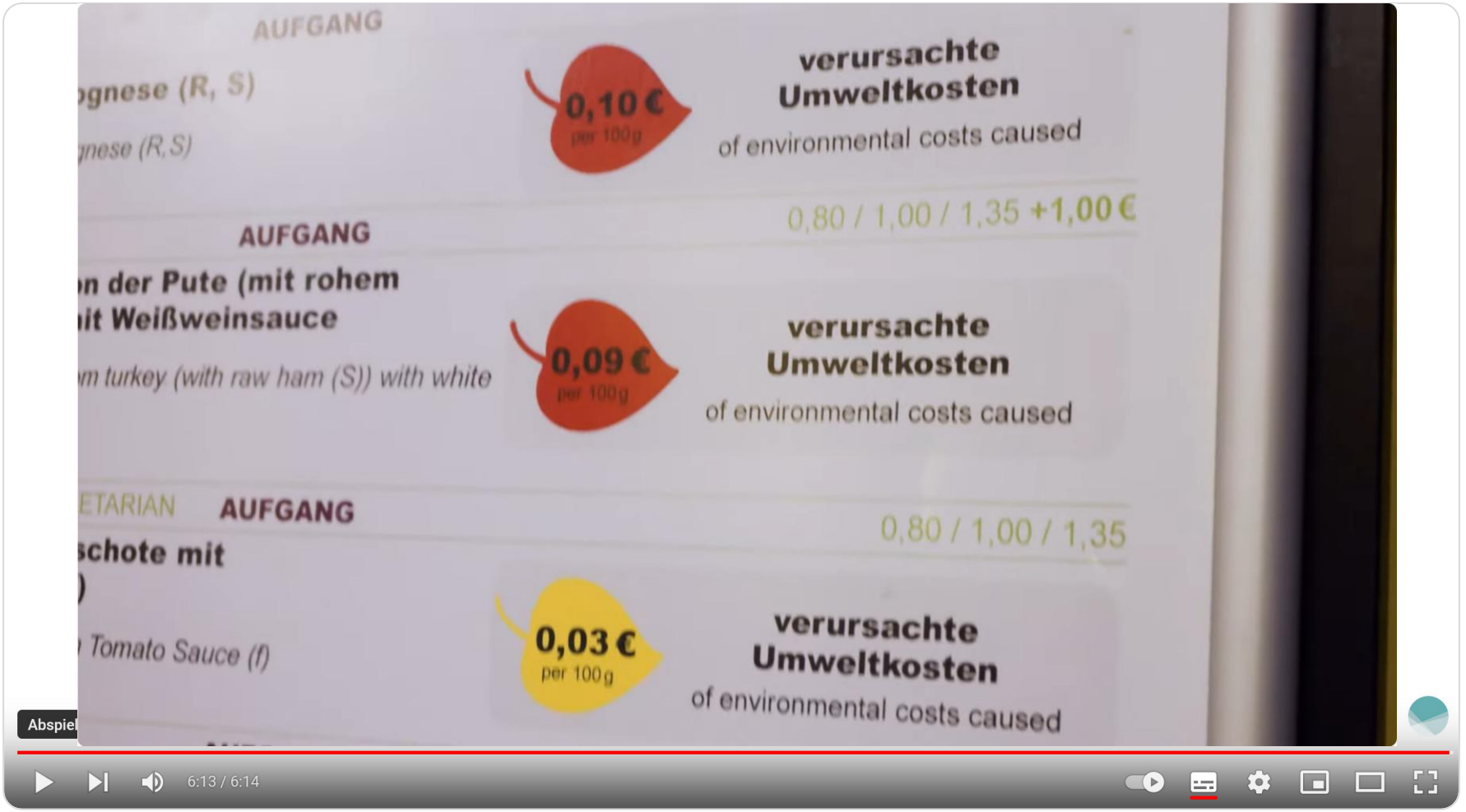
Mini-Umfrage

Frage 1: Wer hat schon einmal an einem Feldexperiment teilgenommen?

Frage 2: Wer war im November 2022 in der LMU Mensa?

Dall-E

Nachhaltiger essen dank CO_2-Label?



Feldexperimente: Was und wozu?

- Definition: Experimente in „natürlichen“ Kontexten
  → Kontexte von theoretischem Forschungsinteresse
  → Oftmals ohne Wissen der Teilnehmenden
- Versuch, zwei Gütekriterien zu optimieren:
  → Interne Validität (experimenteller Stimulus)
  → Externe Validität („real world behavior“)
- Unterschiedliche „Natürlichkeitsdimensionen“

U nit
T reatment
O utcome
S etting

Maike Krauss

Beispiele

- Hilfeexperimente
- „Lost Letter“ Experimente
- Vertrauensexperimente
- Effekte von „Broken Windows“
- Status-Experimente („Hup“-Experimente etc.)
- Effekte sozialer Normen / „Nudging“
- Profiling an Grenzübergängen
- ….

Verhalten auf der Rolltreppe

Aus dem Weg!

12. Juli 2013, 15:12 Uhr | Lesezeit: 2 min

Stau auf der Rolltreppe: Frauen werden schneller und unhöflicher gebeten, Platz zu machen. (Foto: Claus Schunk)

(Wolbring et al. 2013)

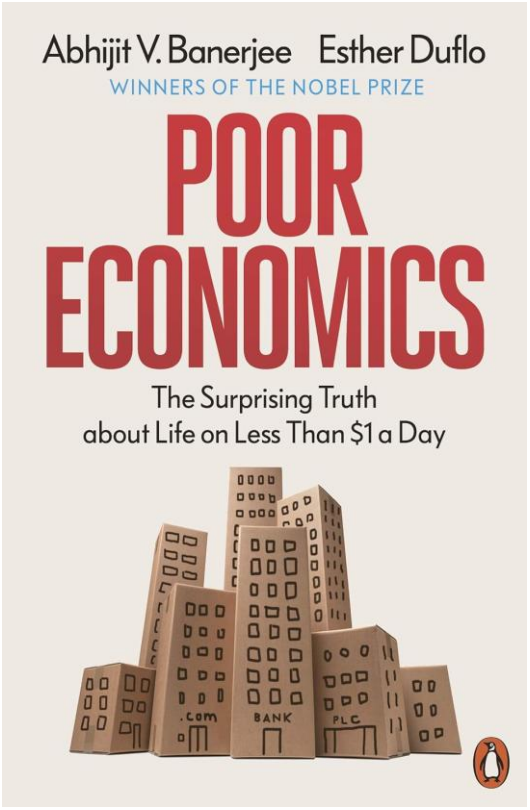
Was ist die Forschungsfrage?

Fig. 1: An example of the experimental setting, with Batman and a woman simulating pregnancy stand in a crowded metro.

- "study tested whether an **unexpected event**, such as the presence of a person dressed as Batman, could increase **prosocial behavior** by disrupting routine and **enhancing attention** to the present moment"
- "In the experimental condition, an additional experimenter dressed as Batman entered from another door. Passengers were significantly more likely to offer their seat when Batman was present"

https://www.nature.com/articles/s44184-025-00171-5

Spezialfall: Randomized Controlled Trials (RCTs)



- Ziel: Evaluation von Interventionen
  - "What works?" → Kosteneffizienz
- „*Whether*" statt „*why*" (d.h. weniger theoriegeleitet als andere Experimente)
- Beispiele
  - Wirkung von Grundeinkommen
  - Maßnahmen zur Bekämpfung von Armut
  - Mentoring
  - Aufklärungskampagnen
  - …

Feldexperimente: Vor- & Nachteile

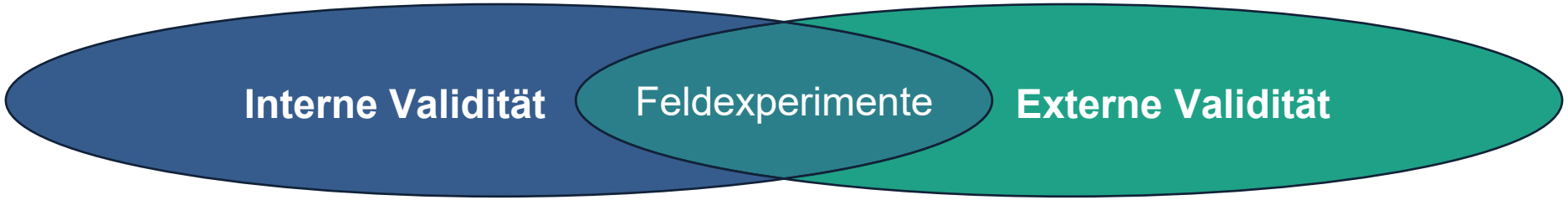
Vorteile

- Weniger/keine soziale Erwünschtheit
- Berücksichtigung von…
  - … Treatment-Effektheterogenität (durch Varianz in Gruppen, Kontexten)
  - … Längerfristigen Effekten (durch wiederholte Messung)
  - … Spillover-Effekten (z.B. Effekte auf Haushaltsebene)

Nachteile

- Geringere Kontrolle über Treatment und Randomisierung, insb.:
  - Treatment (un)auffällig genug
  - Selektion (bei „vorteilhaftem“ Treatment)
  - Abbrüche (bei „unvorteilhaftem“ Treatment)
- Externe Validität & Skalierbarkeit
  - Lokale Märkte → nationaler Kontext?
  - Generalisierbarkeit theorieloser RCTs?
- Forschungsethik

Auf einen Punkt…



Merkmale:

→ Randomisierung

→ Experimenteller Stimulus

→ Natürlichkeit (UTOS)

→ „Real world behavior“

Korrespondenz- & Audit-Studien

oft schwer beobachtbar
(Survey/Prozess-Daten?)

Eigenschaft von A
(z.B. Geschlecht)

→

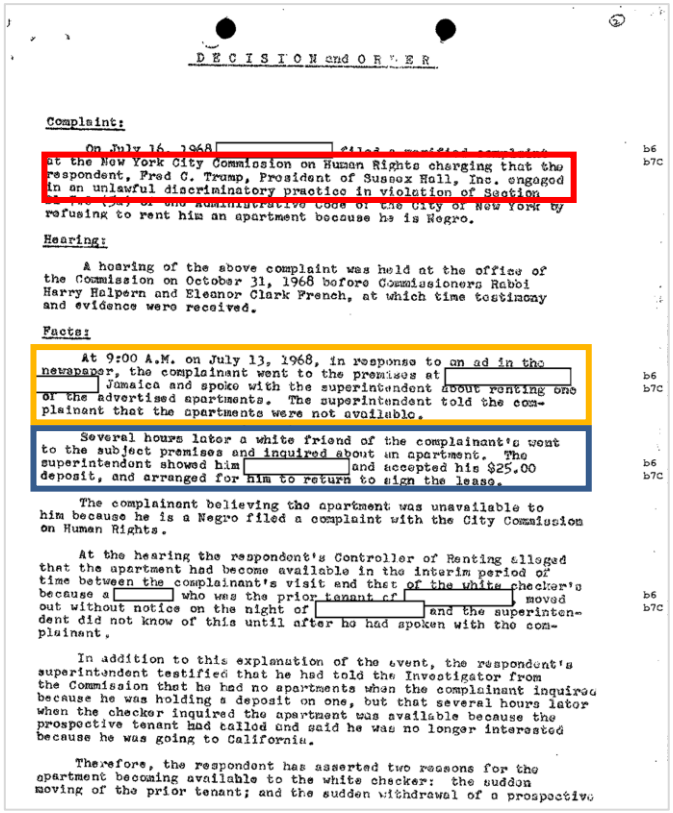
Verhalten von B
(z.B. Diskriminierung)

Korrespondenz-Studien

- Fiktive Personen
- E-Mails oder schriftliche Bewerbungen
- Automatisierung durch Web-Scraping u.Ä.

Audit-Studien

- Reale Personen
- Persönliche Interaktion (F2F, Telefon)
- Erfordert geschultes Personal!



DECISION and ORDER

Complaint:

On July 16, 1968 [redacted] filed a verified complaint at the New York City Commission on Human Rights charging that the respondent, Fred C. Trump, President of Sussex Hall, Inc. engaged in an unlawful discriminatory practice in violation of Section [redacted] of the Administrative Code of the City of New York by refusing to rent him an apartment because he is Negro.

Hearing:

A hearing of the above complaint was held at the office of the Commission on October 31, 1968 before Commissioners Rabbi Harry Halpern and Eleanor Clark French, at which time testimony and evidence were received.

Facts:

At 9:00 A.M. on July 13, 1968, in response to an ad in the newspaper, the complainant went to the premises at [redacted] Jamaica and spoke with the superintendent about renting one of the advertised apartments. The superintendent told the complainant that the apartments were not available.

Several hours later a white friend of the complainant's went to the subject premises and inquired about an apartment. The superintendent showed him [redacted] and accepted his $25.00 deposit, and arranged for him to return to sign the lease.

The complainant believing the apartment was unavailable to him because he is a Negro filed a complaint with the City Commission on Human Rights.

At the hearing the respondent's Controller of Renting alleged that the apartment had become available in the interim period of time between the complainant's visit and that of the white checker's because a [redacted] who was the prior tenant of [redacted] moved out without notice on the night of [redacted] and the superintendent did not know of this until after he had spoken with the complainant.

In addition to this explanation of the event, the respondent's superintendent testified that he had told the Investigator from the Commission that he had no apartments when the complainant inquired because he was holding a deposit on one, but that several hours later when the checker inquired the apartment was available because the prospective tenant had called and said he was no longer interested because he was going to California.

Therefore, the respondent has asserted two reasons for the apartment becoming available to the white checker: the sudden moving of the prior tenant; and the sudden withdrawal of a prospective

b6 b7C

Korrespondenz- & Audit-Studien

oft schwer beobachtbar
(Survey/Prozess-Daten?)

Eigenschaft von A
(z.B. Geschlecht)

→

Verhalten von B
(z.B. Diskriminierung)

> The Commission finds that the respondents have engaged in an unlawful discriminatory practice [...]
>
> The testimony and evidence submitted show that the respondent's treatment of the white checker was different from the treatment afforded the Negro complainant. The apartment in question was unavailable to the complainant, but it was available to the white checker who was allowed to leave a deposit with the superintendent.

- „aktivistischer" Einsatz von Audit-Studien
- Aber: Auch „wertneutrale" Korrespondenzstudien (mit Systematik, Methode)

Aufgabenstellung

Um ethnische Diskriminierung auf dem Wohnungsmarkt zu messen, werden oftmals sog. Korrespondenztests durchgeführt: Wohnungsanbieter erhalten E-Mail-Bewerbungen, in denen der Name der Bewerber/innen experimentell variiert wird (etwa Verwendung eines typisch deutschen vs. türkischen Namen). Anschließend werden Unterschiede in den Rücklaufquoten verwendet, um das Ausmaß ethnischer Diskriminierung zu messen. (Kürzlich wurde eine solche von Datenjournalisten durchgeführt: https://www.hanna-und-ismail.de/.)

1. Vorteile/Nachteile, diese Experimente per E-Mail und nicht mit realen Testpersonen (sog. Audit-Studien) durchzuführen?
2. Mögliche Bedrohungen der internen oder externen Validität solcher Studien?
   Diskussion mit idealerweise Fachbegriff und Beispiel.
3. Mögliche Vorteile/Nachteile, jeweils mehrere Bewerbungen (mindestens eine mit türkischem, eine mit deutschem Namen) pro Wohnungsanbieter zu versenden?
4. Sinnvoll, für Messung der Diskriminierungsquoten Fälle, bei denen keiner der Bewerber eine Antwort erhalten hat, als stichprobenneutrale Ausfälle aus den Analysen zu nehmen?

Lösungsansätze

1. Vor-/Nachteile von Experimenten per E-Mail?
   - Vorteile:
     - Audit-Studien sind keine echten Experimente (keine Randomisierung!); schriftliche Tests (Korrespondenztests) bieten mehr Kontrolle über das Treatment (s. vertiefend: Literatur von J. Heckman)
     - E-Mails sind zudem günstiger, ethisch vertretbarer, höhere Fallzahlen, somit Power (für mehrfaktorielle Experimente)
   - Nachteile: Nur erste Stufe des Bewerbungsverfahrens!
2. Mögliche Bedrohungen der internen/externen Validität?
   - Interne: Namen werden nicht also Proxy für Ethnizität gesehen (sondern z.B. für was?); E-Mails landen in Spamfilter; missglückte Randomisierung
   - Externe: Nur bestimmte Anbieter im Sample; bestimmter Wohnungsmarkt, Zeitraum, nur spezielle Nationalitäten etc.

Weitere Beispiele Namen – Treatment Geschlecht?

Gedankenexperiment

1) Denken Sie an Personen mit dem Namen **Alexander**.

2) Denken Sie an Personen mit dem Namen **Edeltraud**.

3) Vergleichen Sie beide Gruppen vor Ihrem inneren Auge.

Worin unterscheiden sich die beiden Gruppen?

Nur in ihrem Geschlecht?

Data Colada (Post Nr. 36)

Namen vielleicht eher nicht…(?)

“Everything that matters (income, age, location, religion) correlates with people’s names, hence comparing people with different names involves comparing people with potentially different everything that matters.”

“Because the applicant was female (Jennifer instead of John), she got a lower offer”

Jennifer

John

→ “Johns vary more in age, appearance, affluence, and presidential ambitions.”

Data Colada (Post Nr. 36)

Namen: Was sind wirksame (und verlässliche) Signale?

- Gefahr der Konfundierung
  - Signal für soziale Herkunft oder Migrationshintergrund?
- Validierungsstudien sinnvoll
- Eindeutige Namen ggf. nicht repräsentativ für soz. Gruppe

  → Forschungsinteresse: Diskriminierung von eindeutig zuordbaren Personen vs. „typischen“ Personen?
- Trennung unterschiedlicher Stimuli durch mehrfaktorielle Experimente

Namensforschung

"Ich kann nur allen Kevins Mut machen"

16. September 2015, 13:43 Uhr

Die Namen einer ersten Klasse in Niedersachsen. (Foto: Peter Steffen/dpa)

SZ

Lösungsansätze

3. Vor-/Nachteile mehrere Bewerbungen pro Wohnungsanbieter?
   - Vorteil: Within-Vergleich → Elimination Störfaktoren
   - Nachteil: Risiko der Entdeckung; Gefahr Konfundierung m. Reihenfolge; stärkerer Eingriff in den Markt (kann inhaltlich/ethisch problematisch sein)

Lösungsansätze

4. Fälle, bei denen keiner der Bewerber eine Antwort erhalten hat, als stichprobenneutrale Ausfälle werten?
   - Kann stichpobenneutraler Ausfall sein (Wohnung nicht mehr zu haben), muss aber nicht (dann Fall von Gleichbehandlung)
   - Löschen der Fälle ist Selektion nach Y: Fälle mit Gleichbehandlung werden aus dem Sample genommen, damit Gefahr der Überschätzung von Diskriminierung
     (f. ein Beispiel: Studie der Datenjournalisten)
   - Gängige Empfehlung und Praxis: man sollte die Fälle in Analysen behalten
     - Damit konservative Schätzung
     - Bessere Vergleichbarkeit über Studien
   - Ausnahmen sind inhaltlich zu rechtfertigen
     - Etwa Vermeidung einer Konfundierung mit dem Ausmaß von Konkurrenz auf Märkten

Ausblick

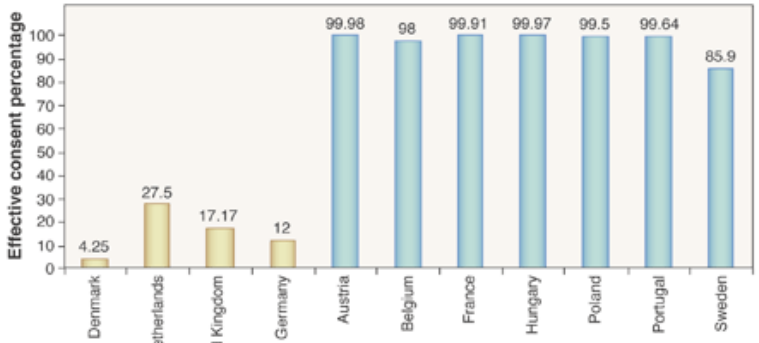
- Nächste Woche: Natürliche Experimente
- Grundlagentext: Bauer 2015

- Anregungen? Fragen?

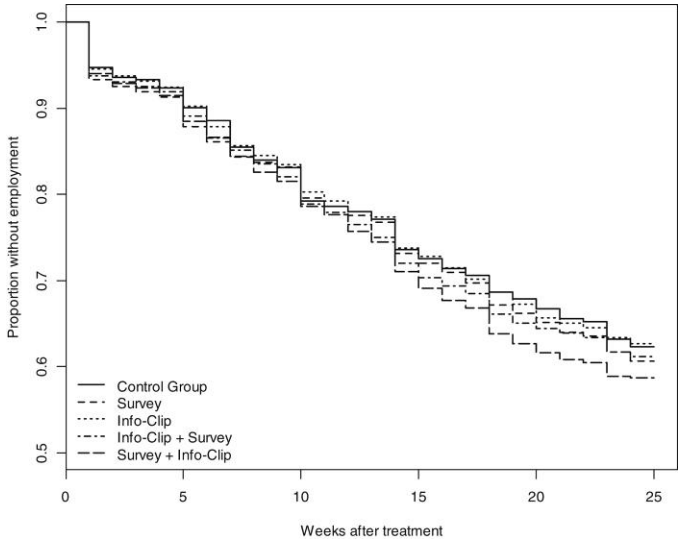
Beispiel RCT Feldexperiment:
Nudging von Arbeitslosen

Nudging-Idee: Verhalten von Menschen beeinflussen, ohne Handlungsoptionen zu verbieten oder die wirtschaftlichen Anreize erheblich zu verändern (Thaler und Sunstein 2008)

- Beispiel: Feldexperiment zur Wirkung von Information und Reflexion auf das Verhalten von Arbeitslosen (Erfolg bei der Jobsuche) (Mühlböck et al. 2021)
  - Idee: Idee teurer Bewerbungskurse umlegen auf kosteneffiziente und zeitsparende Nudges (Informationsvideo + Reflexion durch Fragebogen)
  - 35.333 Arbeitslose in Österreich (registergezogen) 4 Treatmentgruppen + Kontrollgruppe
  - Fokus auf Gruppe mit niedrigem Bildungsgrad
  - 12-18% „response“ rate (treated): ITE



Organ donations → opt in vs. opt out as a nudge

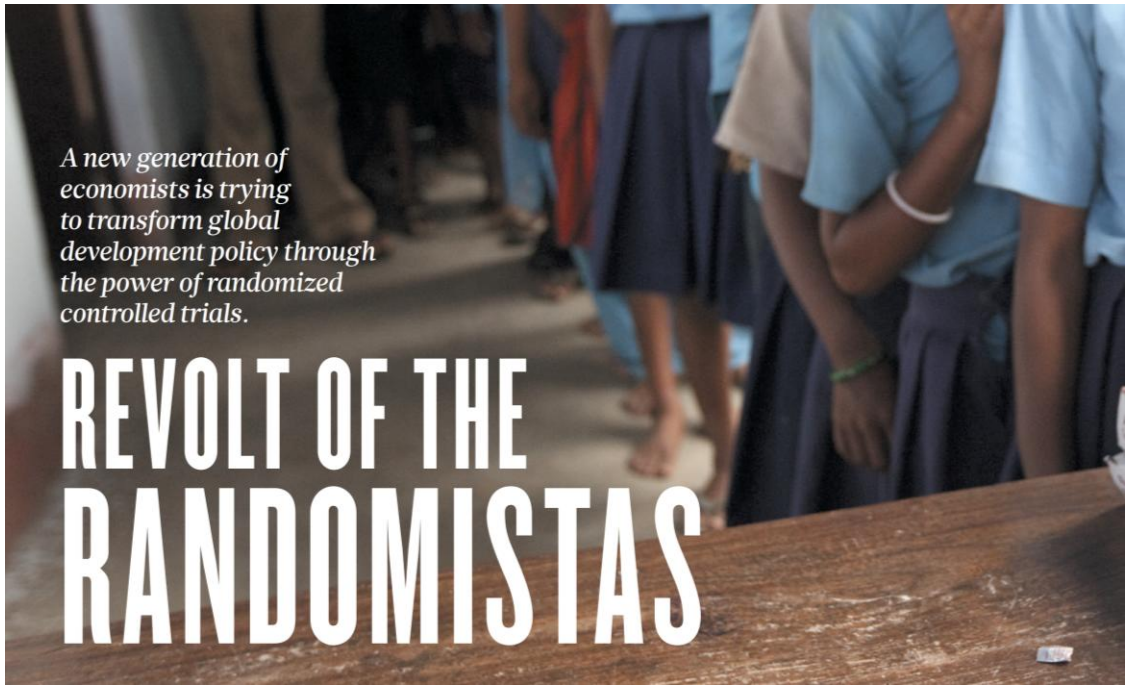


Dauer der Arbeitslosigkeit nach Treatmentgruppe

https://onlinelibrary.wiley.com/doi/10.1111/spol.12754

Weitere Kritik an Feldexperimenten / RCTs

- Überschätzung des Nutzens für Innovation/Wissensfortschritt durch die „Randomistas“
- Fokussierung auf Mikro-Mechanismen
  - → Effizienzverlust durch Ausblendung von Makro-Effekten? (Policy Empfehlungen)
  - → Entwicklungsökonomie/Politik: Aspekte wie Korruption sind Leerstellen



News Feature | Published: 12 August 2015

Can randomized trials eliminate global poverty?

Jeff Tollefson

(Tollefson 2015)

(Wann) Sind Experimente ethisch zulässig?

Ethische Standards

- „Informed Consent“
- Keine Täuschung („no deception)
- Kein Schaden durch…
  - … Verletzung der Privatsphäre
  - … Kosten (zeitlich/monetär)
  - … Reputationsverluste
  - … Illegale Handlungen
- Minimierung der Risiken für Versuchs- und Experimentalpersonen (HiWis)

Warum Experimente (& wie)

- Gesellschaftlicher Wert von Forschung
  → z.B. für Interventionen!
- Linderung durch nachtr. „Debriefing“
- Schadensbegrenzung, d.h.:
  - Alltagsähnliche Situationen
  - Kurze/minimalinvasive Interventionen
  - Profis statt Privatpersonen
  - Aggregierte Auswertung, Datenschutz
  - etc.

Forschungsethik: Experimente ohne Einwilligung?

Ethische Standards

Warum Experimente (& wie)

Abwägung durch Ethikkommission

- Antrag ist vor der Durchführung der Studie zu stellen!
- Gilt ggf. auch für Forschungsarbeiten im Studium (BA-Arbeit, Seminararbeit)
- Richtlinien der Akademie für Soziologie
- Falls unethisch: Kommen andere Experimente in Frage? (Survey-, Lab, natürliche Exp.)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Natürliche Experimente

Merkmale von Experimenten

	2+ Gruppen (T vs. C)	Randomisierung	Stimulus durch Forschende
„Echte“ Exp.	✓	✓	✓
„Natürliche“ Exp.	✓	[✓]	✗

Natürliche Experimente

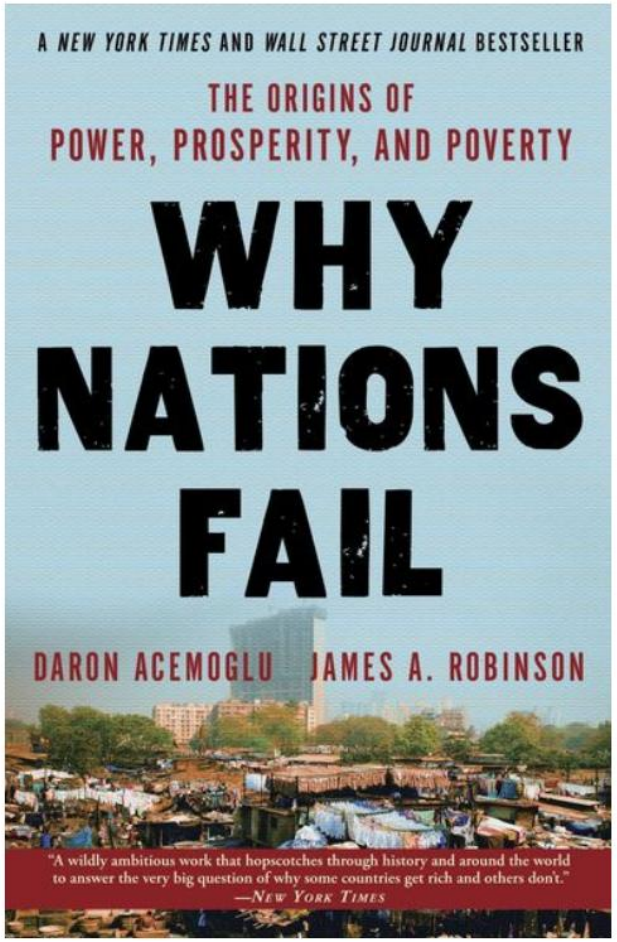
- Treatment und Zuteilung in T/C nicht durch Forschende, sondern durch die „Natur“
  - Menschliche/nicht-menschliche Umwelt
  - Natur, Gesetzgeber, Organisationen, …
- Kritische Voraussetzung für Kausalschlüsse:
  („as-if“) Randomisierung
  - z.B. Lotterien, Willkür/Quasi-Zufall
    → Plausibilität unbedingt prüfen (Fallkenntnis!)
- Können nicht designed, sondern nur „entdeckt“ werden
  → weniger Kontrolle über das Design

Bildquelle: Wikipedia

Beispiele für Stimuli in natürlichen Experimenten?

- Wettereinflüsse
- Schwellenwerte
- Regionale Gesetzesänderungen
- Siehe auch spätere Folien

„Why Nations Fail“: Regionen als nat. Experimente



- Fragestellung: Welche Rolle spielen politische Institutionen? (z.B. für Entwicklung, Wohlstand, etc.)
- Problem I: Institutionen sind nicht zufällig!
  - Potenzielle Confounder: Geschichte, Lage, Klima, usw.
- Problem II: Umgekehrte Kausalität?
  - z.B. Korruption ↔ Autokratien

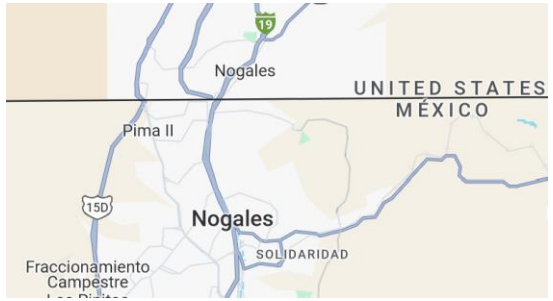
„Why Nations Fail“: Regionen als nat. Experimente

Idee:

- Vergleich möglichst ähnlicher Regionen
  (z.B. hinsichtlich Klima, Häfen/Handelswege, etc.)
  → Kovariatenbalance
- Aber: mit unterschiedlichen Institutionen

Beispiele:

- Arizona und Mexiko
- Botswana und benachbarte afrikanische Staaten
- Nord- und Südkorea



Bildquelle: NASA, Google Maps

Soap Operas und Fertilität

- Beeinflusst Fernsehen die Fertilität in Brasilien
- Idee: Telenovelas stellen „kleine Familien“ dar
- Natürliches Experiment: Unterschiede im Zeitpunkt des Markteintritts von „Globo“, dem wichtigsten Novela-Produzenten
- Resultat: Frauen, die in von Globo abgedeckten Gebieten leben, haben eine signifikant niedrigere Fertilität
- Evidenz dafür, dass Novelas und nicht nur Fernsehen allgemein individuelle Entscheidungen beeinflussten, gestützt auf Namensgebungsmuster von Kindern und den Inhalt der Novelas

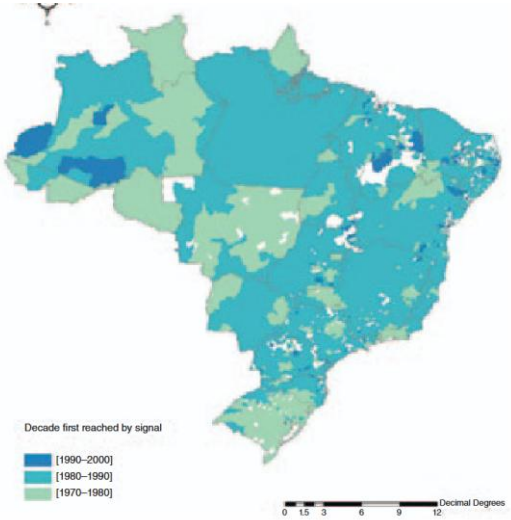


FIGURE 2. REDE GLOBO EXPANSION ACROSS SPACE

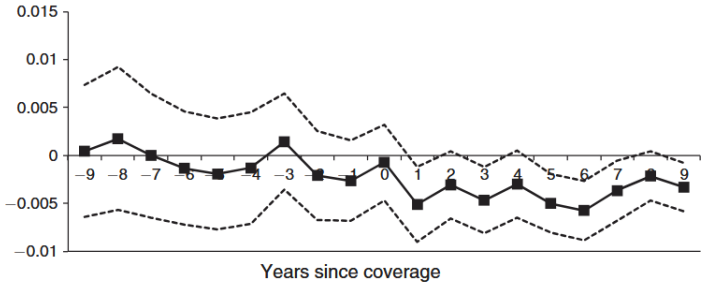


FIGURE 4. TIMING OF FERTILITY DECLINE AROUND YEAR OF GLOBO ENTRY

Note: Estimated coefficients and 95 percent confidence interval from a regression of the probability of giving birth on a set of dummies from $t-9$ to $t+9$, where $t=0$ is the year of Globo entry.

La Ferrara, Eliana, Alberto Chong, and Suzanne Duryea. 2012. "Soap Operas and Fertility: Evidence from Brazil." American Economic Journal: Applied Economics 4 (4): 1–31.

Vor- und Nachteile

Vorteile

- Natürliche Situationen
- Keine Kosten für Durchführung
- Ethische Probleme sind gegenüber anderen Experimenten geringer (Experiment findet ohnehin statt)
- Damit etliche Treatments und Outcomes untersuchbar, hohe externe Validität

Nachteile

- Randomisierung nicht unter Kontrolle, oft nur quasi-zufällig
  - Damit Minderung interner Validität, Gefahr von Konfundierungen
  - Abhilfen: Balance-Tests, Placebo-Tests & ex-post-facto-Kontrollen
- Zum Teil lediglich Spezialpopulationen
- Nicht gezielt planbar (damit selten replizierbar; Relevanz?)

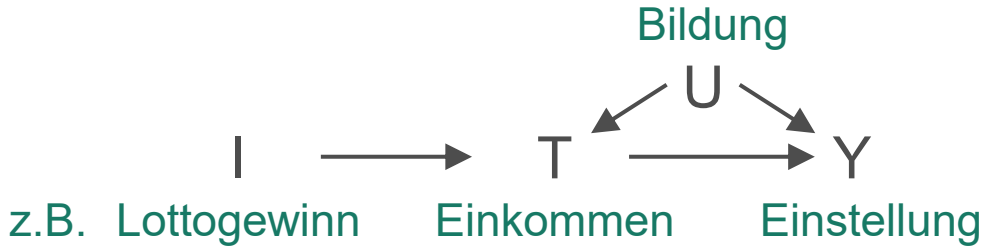
Varianten natürlicher Experimente

Art der Randomisierung

- Zufällig: Zufallsexperiment mit bekannter Wahrscheinlichkeitsverteilung
  - z.B. Lotterien
- Quasi-zufällig: Wahrscheinlichkeits-verteilung unbekannt (nur schätzbar)
  - z.B. institutionelle Regelungen, Wetter, Wasserversorger

Ziel der Randomisierung

- Treatment: Natürliches Standardexperiment
- Position um Schwellenwert: RD-Design („Regression Discontinuity“)
  - Idee: Quasi-zufällige Aufteilung rund um Schwellenwert (z.B. Note, Grenze, etc.)
- Vorgelagerte Variable: IV-Design („Instrumentalvariablendesign“)



Fehlerteufel!

Abbildung 1: Grundidee eines natürlichen Experiments mit Instrumentalvariable (Z)

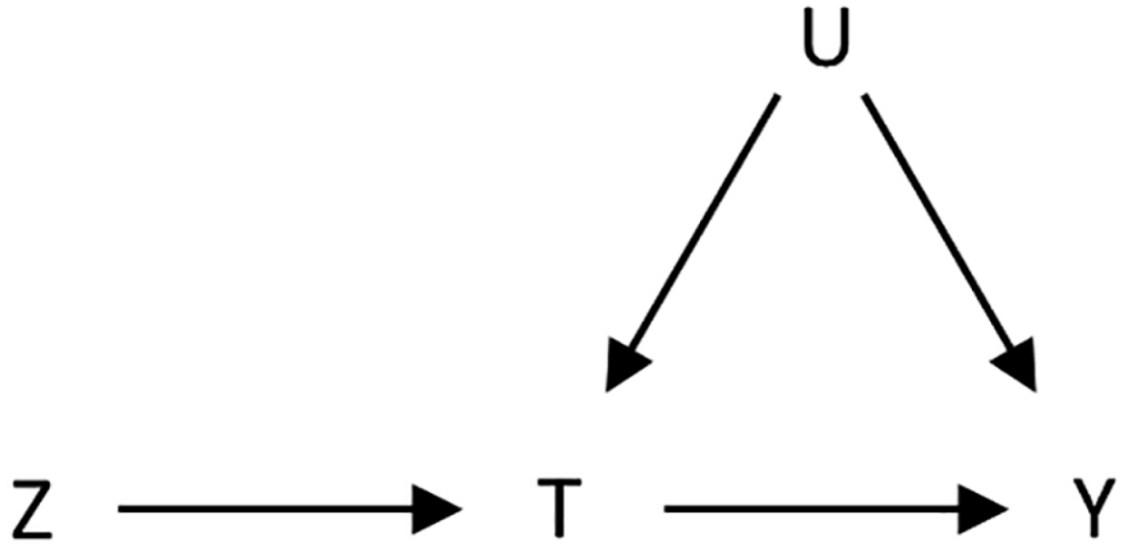
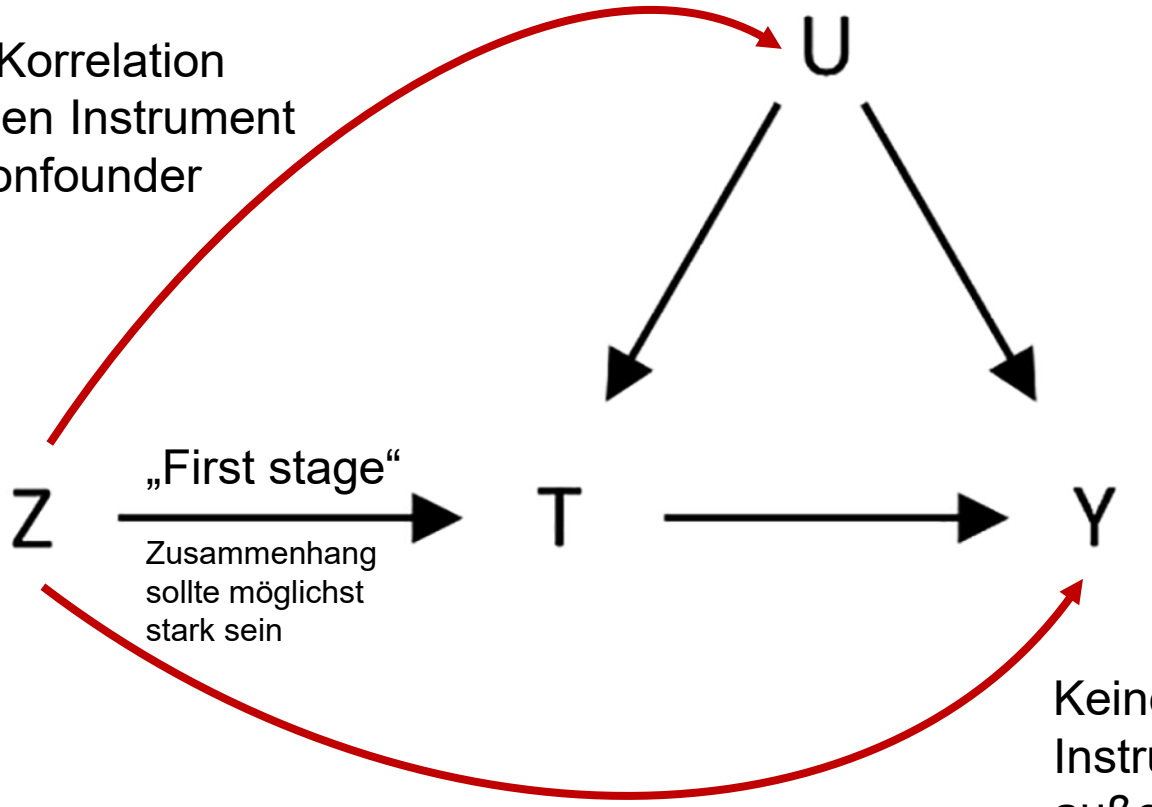


Abbildung 1 verdeutlicht das Instrumentalvariablen-Design. In diesem Beispiel ist der kausale Effekt des Treatments (T) auf eine Zielgröße (Y) von Interesse. Eine empirisch beobachtete Korrelation zwischen T und Y ist aber nur dann kausal, wenn unbeobachtete Drittvariablen (U) die Korrelation hervorrufen. Für beobachtete Drittvariablen lässt sich hierbei

Voraussetzung Instrumentalvariable

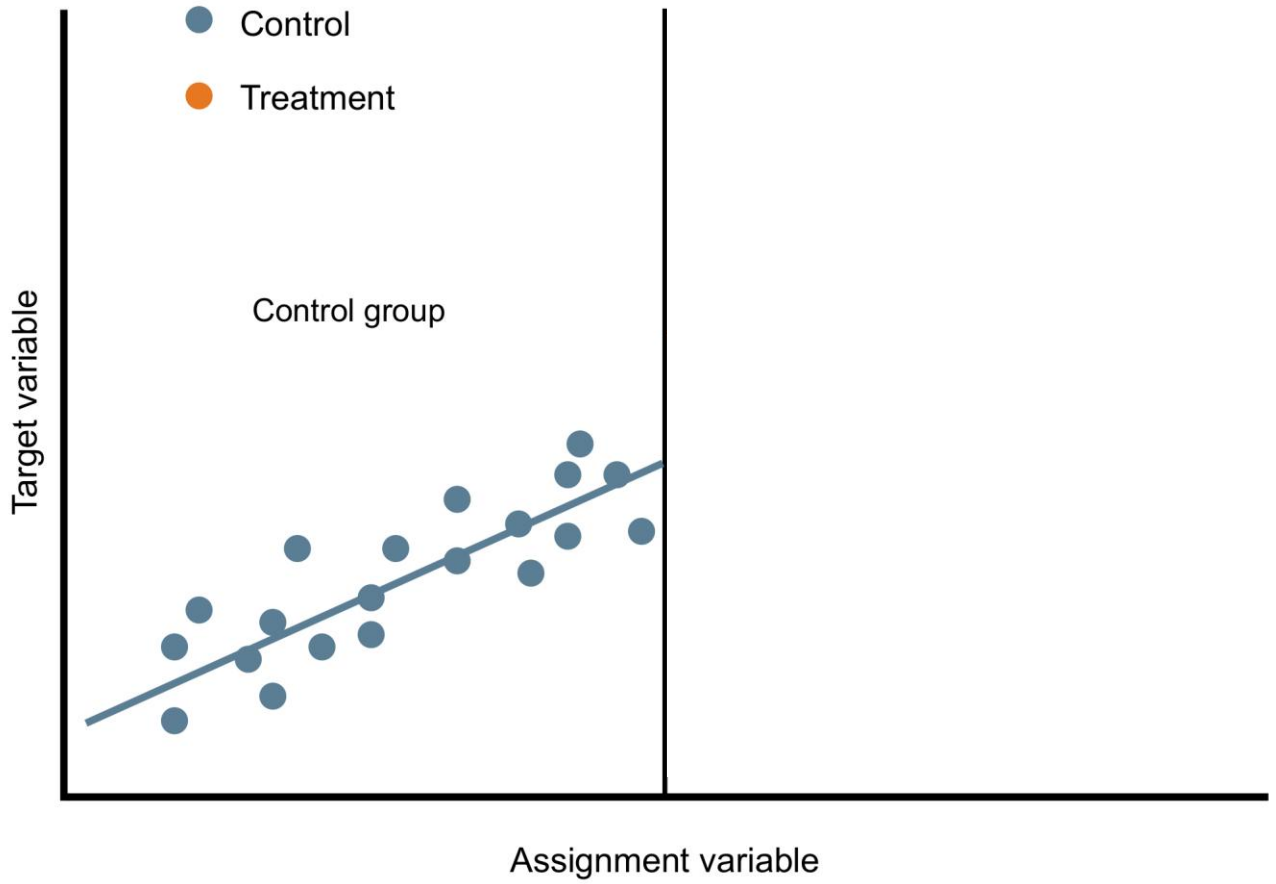
Abbildung 1: Grundidee eines natürlichen Experiments mit Instrumentalvariable (Z)



- Instrumente ermöglichen es einen „Local Average Treatment Effect“ zu berechnen
- Dieser bezieht sich nur auf Personen deren Treatment status durch das Instrument beeinflusst werden kann (keine always takers, never takers und defiers)

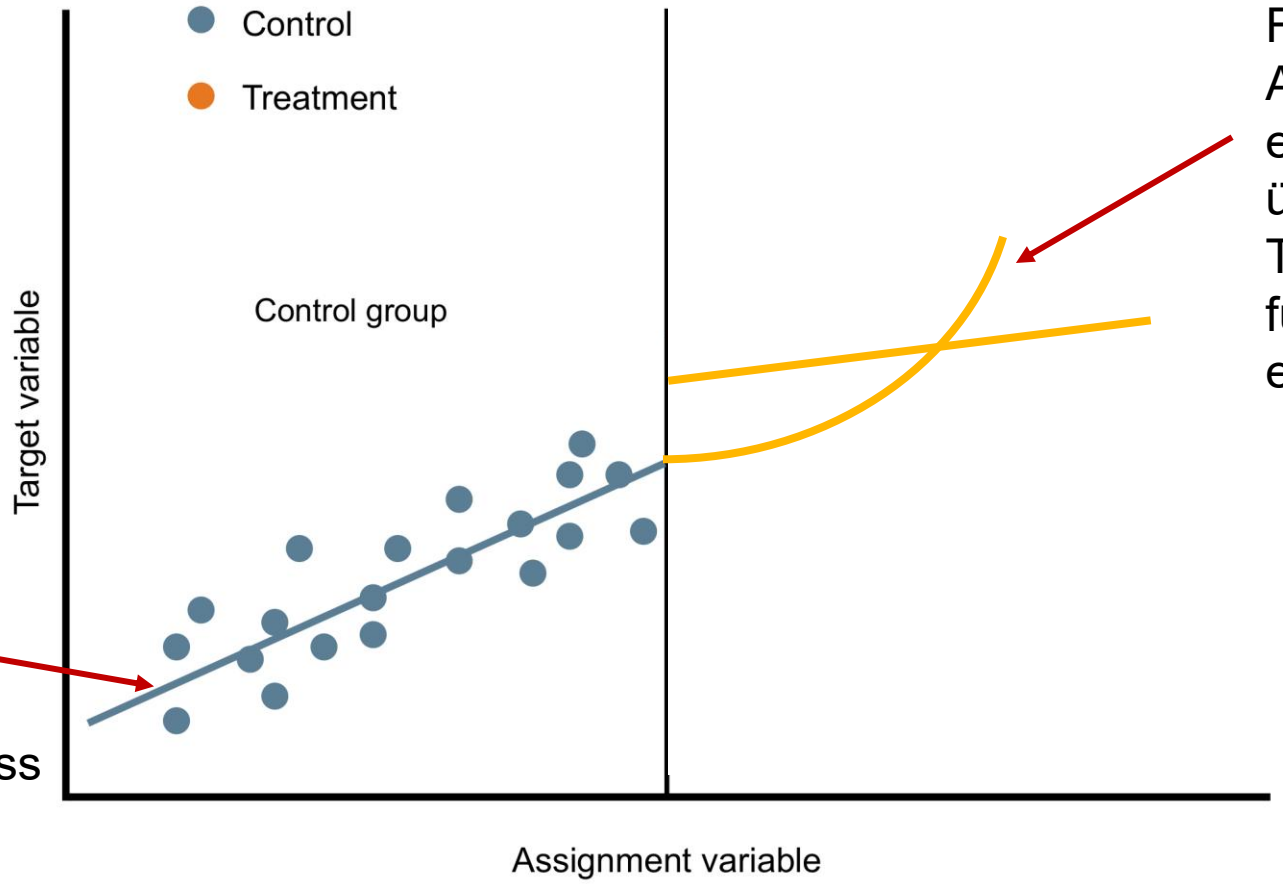
Regression Discontinuity Design (RDD)

Wie sähe hier ein möglicher Kausaleffekt aus?



Liu et al. 2022: 14

Voraussetzung: Regression Discontinuity Design (RDD)



Gute Kenntnis über die funktionale Form des Zusammenhangs unabhängig vom Treatment (+ Ausschluss von Antizipation)

Fundierte theoretische Annahmen und empirische Kenntnisse über die zu erwartende Treatmentfunktion (step-function, kumulierend, exponentiell, etc.)

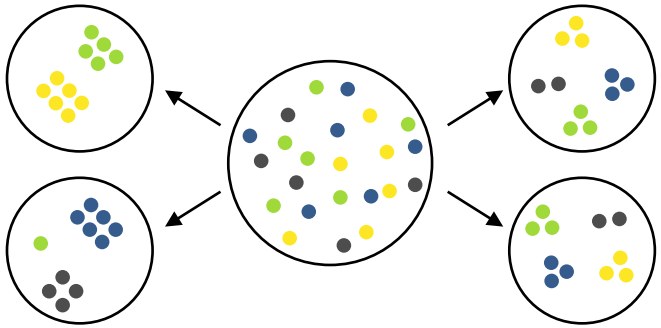
Liu et al. 2022: 14

Beispiele

Design/Typ	Randomisierung	
	Zufällig	Quasi-zufällig
Natürliches Standardexperiment	Lotteriegewinne Lotterie Wehrpflicht Vietnamkrieg	Wasserversorgung (Snow) Verlegung von Wahllokalen
Instrumentalvariablendesign	Vietnamkriegslotterie → Kriegseinsatz Strafmaß & Rückfallquote	3. Kind Westfernsehen
Regression Discontinuity Design	?	Wahlbeteiligung/-verhalten (Schwellenwerte: Alter, Ort, Regel-Kategorie)

Beurteilungskriterien natürlicher Experimente

- („as if“) Randomisierung
  → Selektion in T bzw. C muss unabhängig vom Outcome sein
  - Information/Anreize/Fähigkeit für Selektion?
  - „Balance-Tests“, „Placebo-Tests“ (RDD)



→ Wissen über datengenerierenden Prozess!
Qualitativ, anekdotisch, historisch…

- Relevanz des Treatments
  → Erkenntnisgewinn, d.h. Nähe zum „idealen“ Experiment?
  - Interne Validität (Randomisierung?)
    → „compound treatment problem“
  - Externe Validität (Generalisierbarkeit?)

- Glaubwürdigkeit des stat. Modells
  → Einfluss von Kontrollvariablen?

Übungsaufgabe: Präsentation/Diskussion Beispiel

1. Beschreiben Sie kurz die Forschungsfrage der Studie. Warum es für ihre Beantwortung vorteilhaft, mit einem quasi-experimentellen Design zu arbeiten? Nutzen Sie dazu auch ein DAG.
2. Worin besteht das Design des natürlichen Experiments?
3. Diskutieren Sie, inwieweit die Randomisierung geglückt ist und ggf. auch, ob es weitere Bedrohungen der internen Validität gibt.
4. Wie ist die externe Validität einzuschätzen?

Verfügbare Studien:

- Dinas et al. 2018: Exposure to the Refugee Crisis and Support for Extreme-right Parties
- Frey et al. 2024: Bridging the Digital Divide Narrows the Participation Gap: Evidence from a Quasi-Natural Experiment
- Hainmueller et al. 2017: Does Naturalization Promote the Social Integration of Immigrants?
- Islam et al. 2020: How do social distancing interventions work on covid-19 incidences?
- Legewie 2013: Terrorist Event and Attitudes towards Immigrants

Lösungsansätze: Design, interne Validität

- Forschungsfrage: Treatments in den verschiedenen Studien: Exposure zu Anzahl Flüchtlingen, Terroristischen Anschlägen etc.; Wirkung auf Outcomes wie Xenophobie;
  oftmals Ziel Testung von Theorien
- Design: Annahme einer Randomisierung (z.B. weitgehend zufällige Verteilung von Flüchtlingen auf Orte) oder Regression Discontinuity Design (etwa um Personen mit/ohne Einbürgerung zu vergleichen)
- Zentral für kausale Schlüsse/interne Validität: Ist die Randomisierung gelungen?
  - Dann keine Korrelation irgendeiner Drittvariable mit dem Stimulus
  - Keine Unterschiedliche Verteilung von Drittvariablen bei Fällen unter/oberhalb des Schwellenwertes bei RD Design
- Diskussion: Ist das bei den Studien realistisch?

Dinas et al. 2018

- Does exposure to the refugee crisis fuel support for extreme-right parties?
- Natürliches Experiment in Griechenland:
  - Lediglich einige Inseln haben viele syrischen Flüchtlinge erfahren
  - Abgesehen davon Annahme ähnlicher Bedingungen auf den Inseln

The geography and timing of the refugee crisis in the Aegean Sea

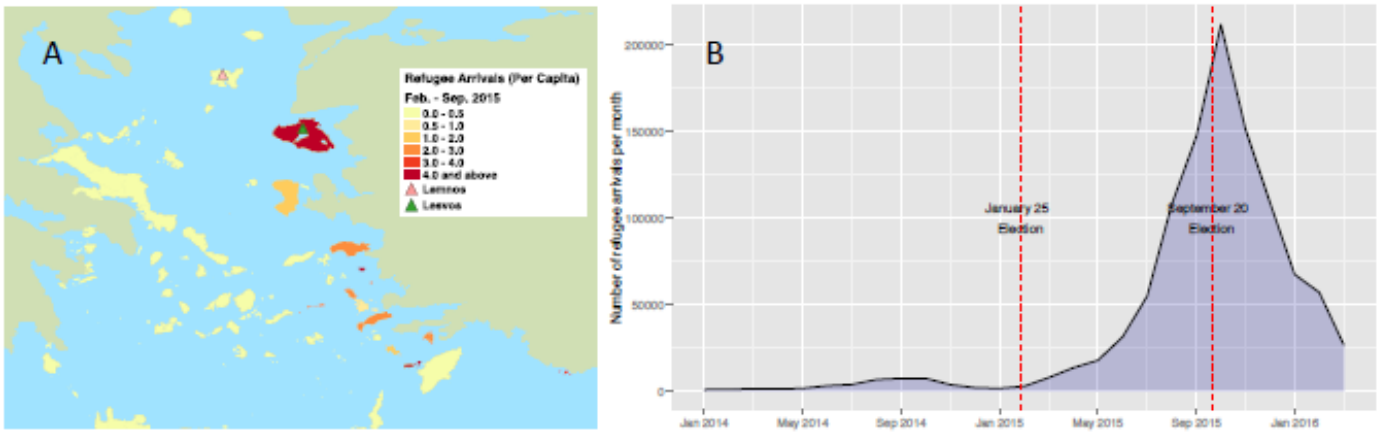
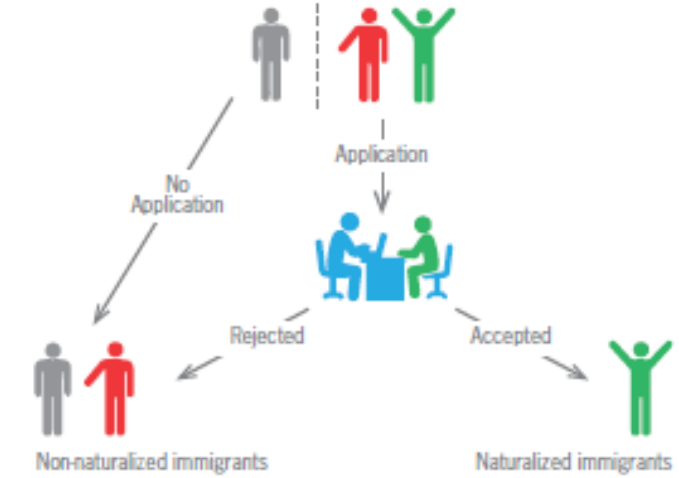


Figure 1: Panel A shows that islands close to the Turkish border received the most refugee arrivals per capita. Panel B shows the monthly number of asylum-seekers arriving at Greek islands over the period from January 2014 to March 2016. During the study period, the first election took place just before the onset of the refugee crisis on January 25, 2015. A second election took place at the height of the refugee crisis on September 20, 2015.

Hainmüller et al. 2017

- Einfluss der Einbürgerung auf die langfristige Integration in das Aufnahmeland
- Schweiz: Einige Kantone stimmen über Einbürgerung ab
- Regression Discontinuity Desgin
  - Gemeinden in CH die per Referendum Einbürgerung entscheiden
  - Vergleich Einwanderer, die knapp akzeptiert/abgelehnt wurden

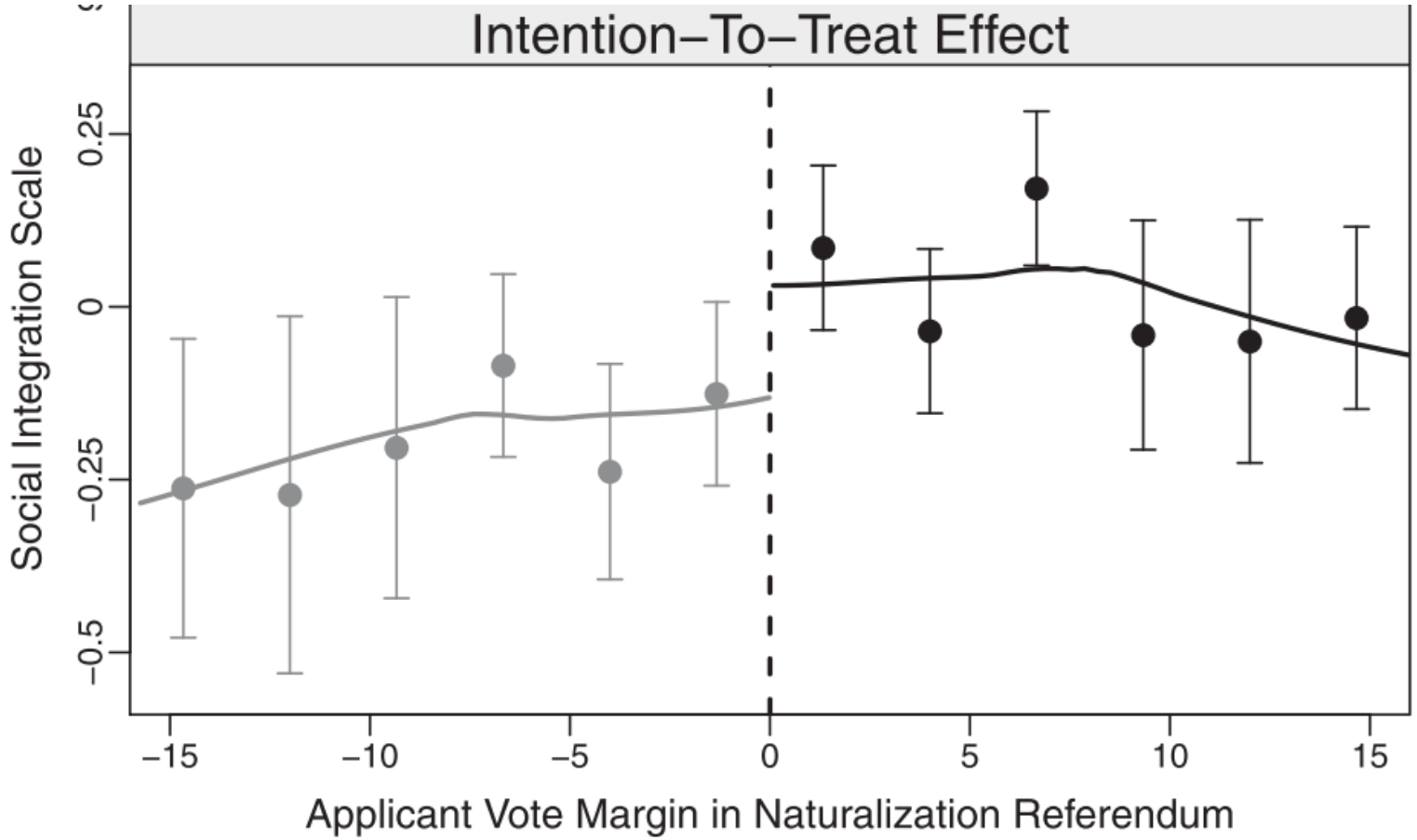
FIGURE 1. Double Selection Bias



Note: Illustration of the double selection bias that confounds the comparison of naturalized and non-naturalized immigrants.

Hainmüller et al. 2017

Knapp **nicht** eingebürgert Knapp eingebürgert



Frey et al. 2024

- Hat Internetzugang Einfluss auf Bürgerbeteiligung?
- Treatment: Ausstattung zufällig ausgewählter Personen ohne Internetzugang mit Zugang (durch Survey-Company)
- Outcome: Wahlbeteiligung / politische Beteiligung laut Survey

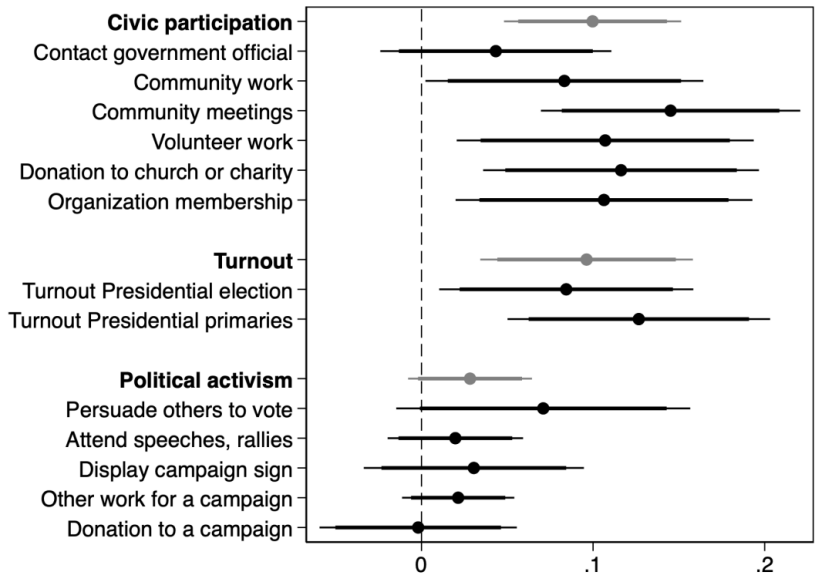
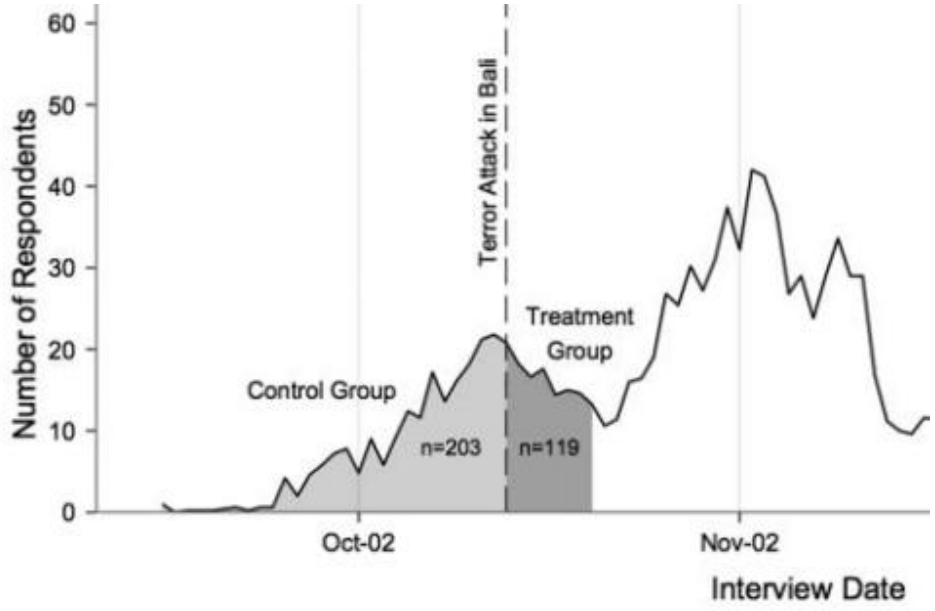
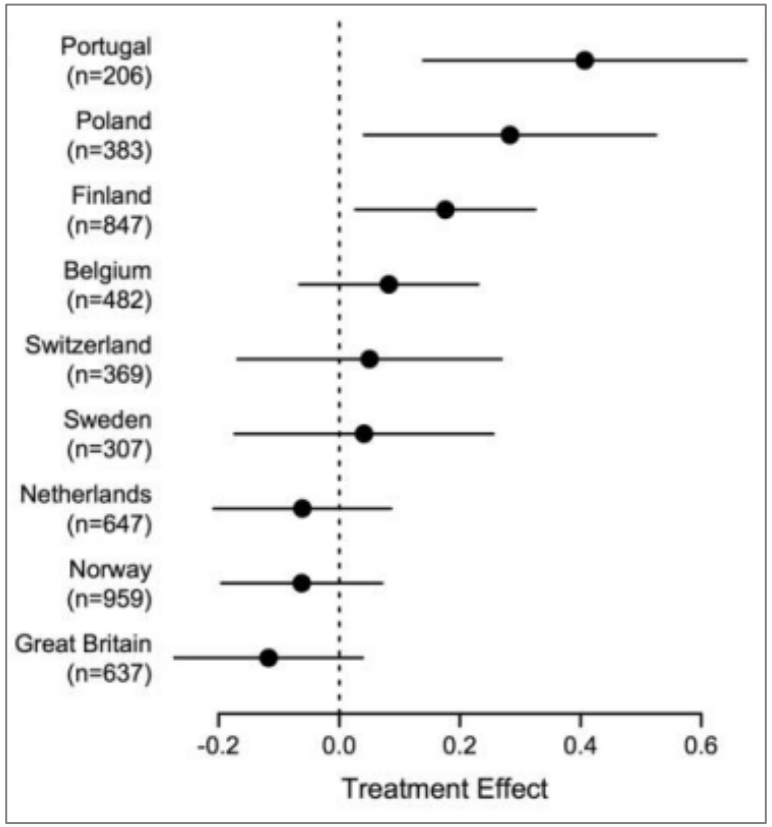


Figure 1: Estimates of the effect of gaining access to the internet on the likelihood that potential late adopters engage in thirteen participation behaviors (black markers). Gray markers show estimates of effects on indices constructed as mean scales of the binary items under the respective heading. Point estimates and confidence intervals (90 and 95%) from linear models. Online supplement A contains the associated regression tables.

Legewie 2013

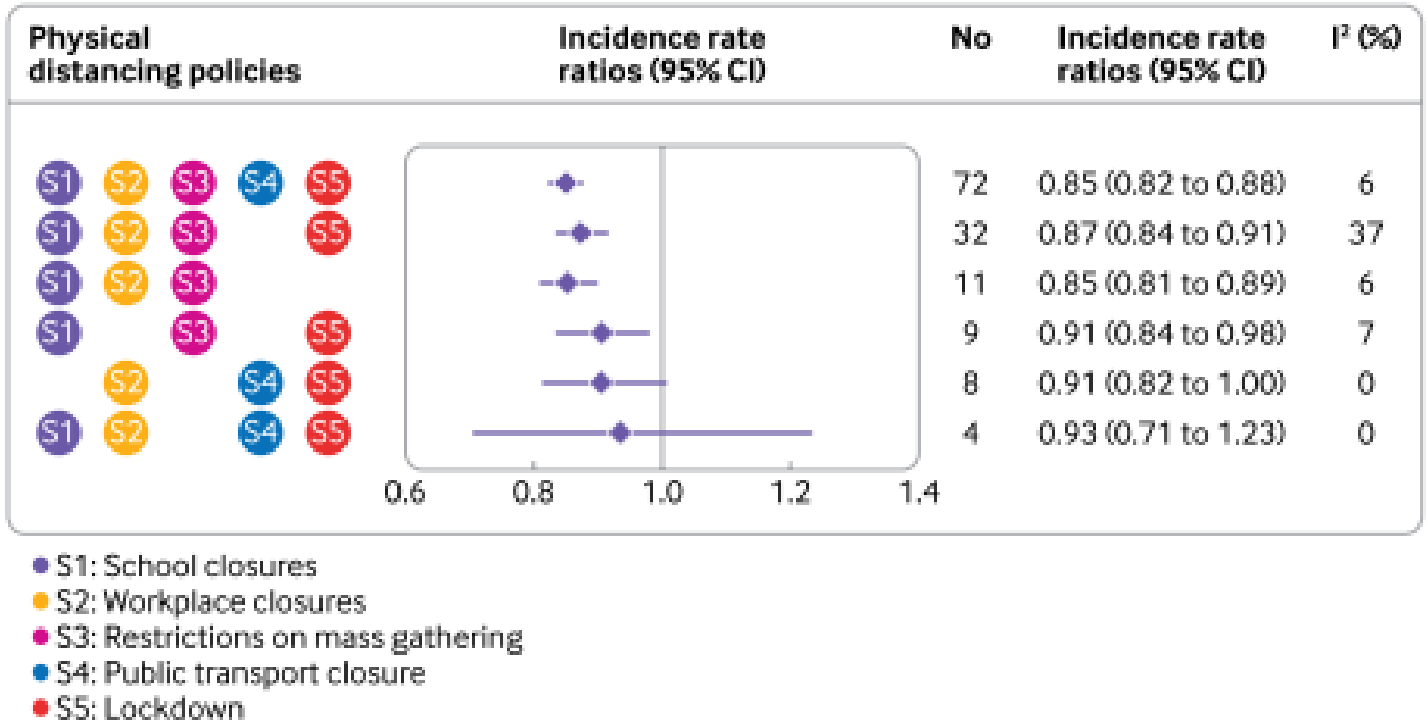
- Effekt von Terroranschlägen auf Einstellung zu Immigration
- Natürliches Treatment: Terroranschlag 2002 in Bali
- Messung der Einstellung im European Social Survey vor und nach dem Anschlag (Feldphase) 2002



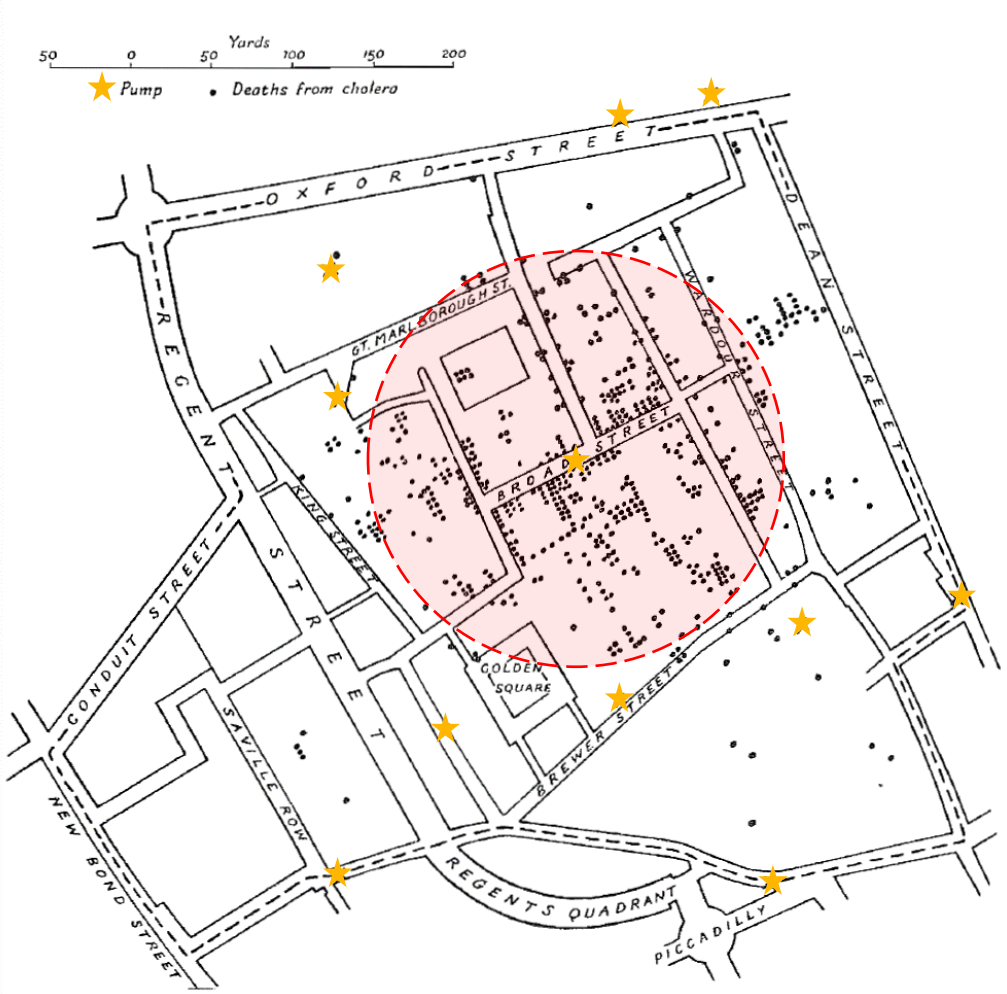


Islam et al. 2020

- Effekt versch. Interventionen des social-distancing auf COVID-19 Fallzahlen; Vergleich 149 Länder/Regionen weltweit
- Treatment:
  Einführung der Intervention



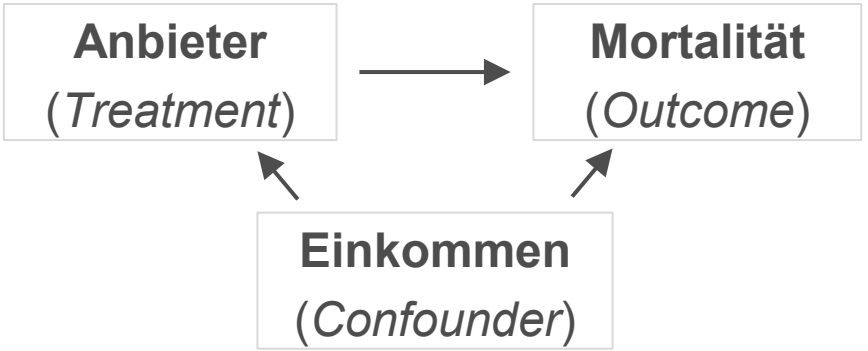
Ein Klassiker: John Snow‘s „Ghost Map“ (1855)



Zwei Trinkwasseranbieter in London:

- S & V: **37** Todesfälle pro 10.000
- Lambeth: **315** Todesfälle pro 10.000

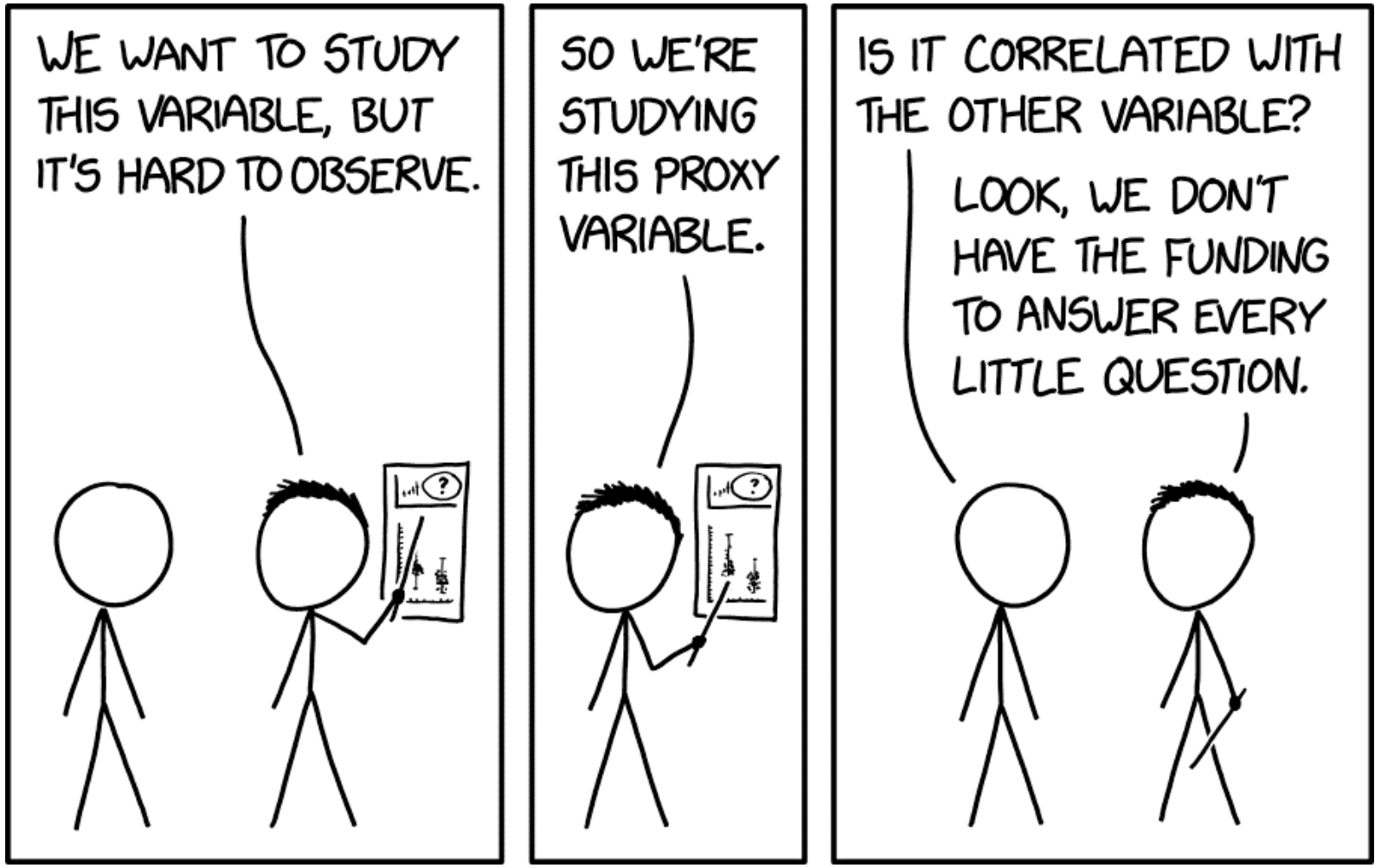
Kausalität?



Aber: Anbieter „quasi-zufällig“ zugeteilt (Balanciert bzgl. Einkommen, Berufe, …)

!

Relevanz: Was es zu vermeiden gilt…



xkcd

Corona-Pandemie als natürliches Experiment?

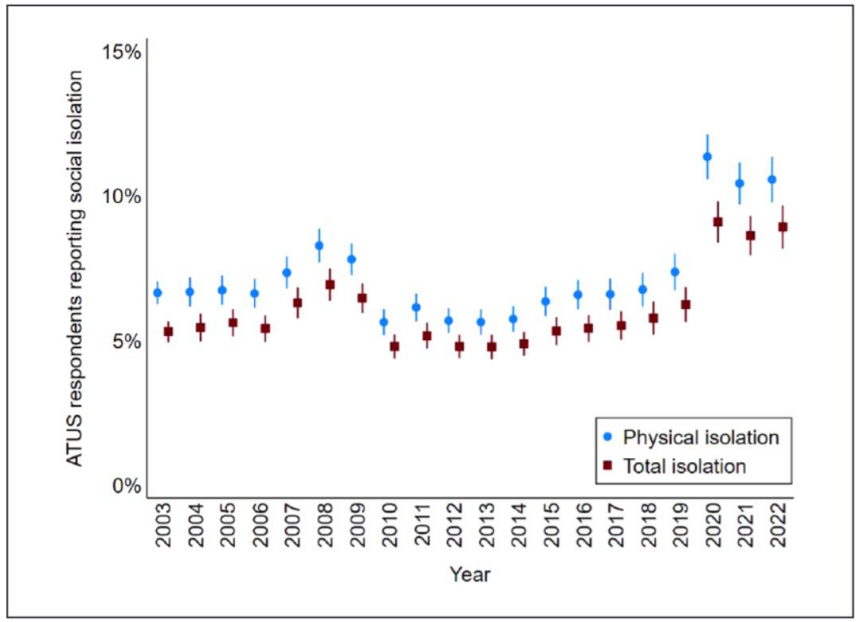


Figure 1. Percentage of American Time Use Survey (ATUS) respondents who were socially isolated within a 24-hour period from 2003 to 2022.
Source: American Time Use Survey (https://www.bls.gov/tus). As shown, there was an increase during the coronavirus pandemic (2020–2022).
Note: Blue circles represent respondents who reported zero face-to-face interactions during their diary days (i.e., physical isolation). Red squares represent respondents who reported zero face-to-face interactions and zero telecommunications during their diary days (i.e., total isolation). All estimates are nationally representative weighted averages.

"Social Isolation in America? A 20-Year Snapshot"

- Outcome: soziale Isolation
- Trenddaten (ATUS)

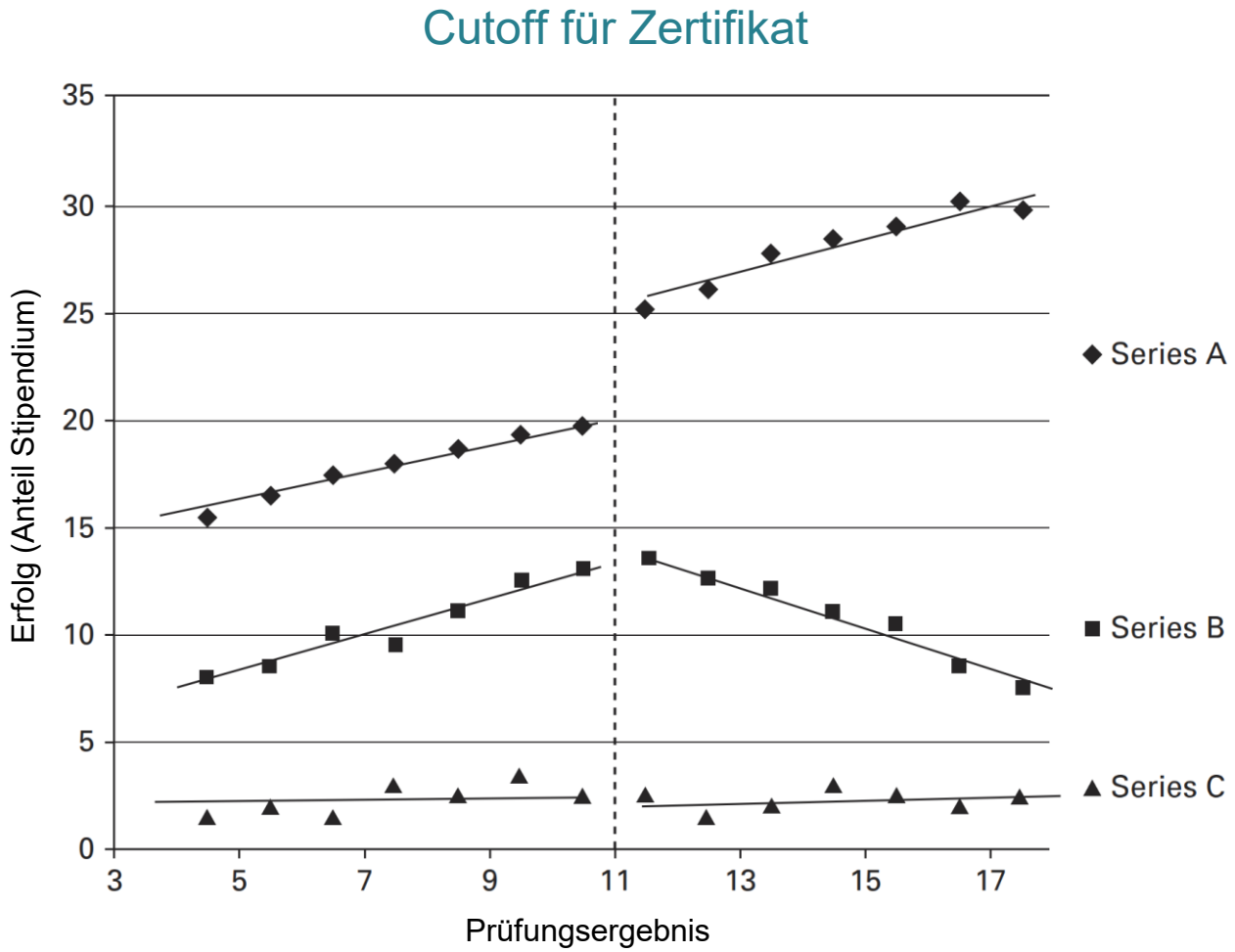
→ Was fällt Ihnen auf?

→ Was könnte 2007/2008 losgewesen sein?

→ Allgemein: Natürliches Experiment?

Roth (2024)

Natürliches Experiment: Ja/Nein?



- Natürliches Experiment? Unter welchen Bedingungen/Annahmen?
- Serie B? Probleme?
- Welche „*Neighbourhood*“ (Bandbreite um Cutoff) ist für die Identifikation des Kausaleffekts sinnvoll?

Breit: keine „as-if“ Randomisierung

Schmal: geringe Power (kleines *n*)

Modifiziert von Dunning 2012: S. 66

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

The Research Proposal

Short research proposals „due“ 31.01., 23:59

- Optimal ~ 2 pages (w/o literature)
  - Sketch your idea but try to answer the big questions. What is your:
    - Topic and research goal
    - Research question
    - Theoretical argument & hypotheses
    - Research design & method
- Everything can be in German or in English
- Your ideas do not yet have to be complete and perfect – it’s a draft to get feedback!

The Research Proposal: Outline

- Title
- Introduction
  - Topic & relevance, research question: what effect do you want to identify, why and how?
- Theory & literature overview
  - State of the art in the literature – what theories are used?, what has(n't) been shown so far?
  - Why are existing research designs likely to be inadequate for valid answers to your question?
  - Question(s), theoretical approach, state of research, hypotheses/qualitative expectations
- Research design, methods, data: Why is your design optimal to answer the question?
- Short summary and discussion of the project's overall importance and limitations
- Literature

Title

- Working title (preliminary?)
- Main title (can be an eye-catcher but no overkill!) – Subtitle
- Name, matr. no., study code, field of study, type of work

Common mistakes:

- Too long or too complex
- Too imprecise
- No clear definition & thus demarcation of the project → the goal has to be clear!

Topic and Research Context

- Has the character of an introduction
  - Places the topic in a research field or research context
- Outlines the central thesis/question
  - Why is what I am doing exciting/relevant (socially and scientifically)?
    [See chapter 1 on relevant research questions]
  - The central question can also be whether results replicate when making important improvement to an existing research design!

Common mistakes:

- No clear delimitation of the topic; too broad; a realistic delimitation is absolutely necessary!
- No clear research question
- No citations → arguments have to be based on existing knowledge

Research Question

- Clear and precise statement of your research question
  - Breakdown into sub-questions + hypotheses; but less is more: one question/hypothesis is enough!
- What is my goal?
  - Precise delimitation of your own question that specifically targets an aspect that has not yet been addressed in the literature (which may be limitations of the research designs/methods used by other authors):
    What exact effect do you want to identify? (e.g., direct or indirect effect)
  - Draw a DAG that is illustrating the expected causal structure and summarizes the theoretical discussion (start your proposal immediately with the research question, but use this to make the research question even more precise after you have discussed the state of research, and start with proposing your research design)

Common mistakes:

- Not asking a well-defined specific question: If I am not able to ask a precise question, I will not be able to find a precise answer
- Unrealistic, poorly defined questions and objectives
- Too future-oriented, no reference to the literature

State of Research

- Concise description of the state of the art research on the topic
  - Are there controversies? Do different theory-based approaches exist? Do I join one of them?
  - Do existing designs show limitations that might jeopardize their conclusions?
    (You might even discuss only one specific study you want to replicate and extend in detail)
    - For example: are there possible confounders not adequately considered by existing literature (e.g. due to using observational data or experiments with threats to the internal validity)
- Structuring the literature not merely listing them is required!
  - Focus on relevant literature for your research project
    - How does my project relate to the state of research?
    - How are the works cited relevant to answering my question?

Common mistakes:

- Not adequately informed/read
- No position on current discussions/controversies (outdated literature) or research designs / methods
- Too concise/too detailed
- No citations

Research Design / Methods

- How do I plan to answer my question or achieve my goal?
- Which partial steps do I undertake? Which specific design and methods? Why these?
  - What sources, what type of data: this is the core part for this seminar! Why is your design best suited to answer your research question / to fill in the identified research gap?
    How does it help to increase the internal and external validity?
    How and why do you deal with unavoidable trade-offs?
  - Short sketch: How will data be processed, what statistical analysis procedures will be used?
  - Remaining threats to the validity should be transparently discussed in the Discussion section

Common mistakes:

- No clear research design or operationalization strategy
- Methods / data that do not fit the research question
- Hypotheses and data have little to do with each other

Dos and Don’ts

- Do not start paragraphs with “xy said z”
  - Structure your paragraphs based on arguments and not base on texts or authors
- Do not *only* write what others have done but explain also how you use others’ results in your study
- Do not include “empty” sentences
  - E.g. “The article provides another insight to the phenomenon migration relevant to my study.”
- Do not use fancy vocabulary → write precise but simple to help readers understand and follow → keep sentences short and use active voice!
- Do not misuse the terms “representative” and “significant”
- Justified texts (Blocksatz) looks better

Summary – Research Proposal Structure

1) Title
   Should capture the question in a concise and preferably exciting way

2) Introduction (~10% of the proposal)
   Topic, relevance, “gap”(what do we know and what not) , research goal/question, research design → needs to refer to scientific literature

3) Theory & literature overview (~40%)
   Literature overview to the theoretic concepts in the field, research question and theoretical argument, hypotheses/expectations, what have other’s found, and possible threats to the validity of the used research designs? → NO GENERAL DISCUSSIONS that do not connect to one’s own research goal

4) Research design and methods (~40%)
   Research design, operationalisation, data, methods (analytical strategy): why is exactly this design the best approach to answer your research question and fill in the identified research gap

5) Discussion (summary and importance) (~10%)
   Discussion of the importance of this research project and how it is relevant for our understanding of the phenomenon in question in light of what we already know; possible limitations

(3 Theory & literature overview are often combined in one chapter but can be separated)

10 Steps to“find” a research question and a matching research design I

1. Find an interesting **topic**
   - **Work and society during the economic crisis**
   - **Unemployment during the economic crisis**
   - **Unemployment and job matching in the context of recessions**

narrow down the topic and be as specific about the terms as possible

2. **Literature** research (also used for 1)
3. Identify the **gap** – what questions have been answered for now?
4. Formulate a scientific **RQ**
5. Continue **reading** (could lead you back to point 3!)
   - Specify your question (use the five Ws (Who, What, When, Where, Why)

10 Steps to“find” a research question and a matching research design II

6. **Define** all the necessary **terms** in your question
   (could lead you back to point 4!)
7. Consider how to **measure** the terms in your RQ
   (could lead you back to point 4!)
8. Propose your **Hypotheses** (could lead you back to point 4-7!)
   1. What/Which relationships do I want to explain?
   2. What are my expectation?
   3. Do not state too many Hypotheses
9. Suggest an **empirical method** – how to test your hypotheses (could lead you back to point 4-8!)
10. Find appropriate **data** for your research project

Reconsider whether you can really answer your RQ testing these specific hypotheses, using the empirical method with the data suggested

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Total Survey Error

Zum Einstieg

Empirie wird zur reinen Glaubenssache

Was stimmt und richtig ist, wird dann von einem Gegenstand der belegbaren Empirie zu einer Glaubenssache. Das wird die Repräsentativität und die Ergebnisse der Umfrageforschung Stück für Stück entwerten. Sie werden nach und nach eingereiht in die

CIVEY, YOUGOV UND CO

Wie Umfragen unsere Meinung beeinflussen

VON THOMAS PERRY am 30. April 2019

Politiker, Medien und Leser orientieren sich immer häufiger auch an Online-Umfragen von Civey, YouGov und Co. Doch die Qualität der Online-Methoden ist zweifelhaft. Was nach messbarer, repräsentativer Beteiligung aussieht, schadet unserer Demokratie

e Nachfrage von Medien
gsforscher wehren sich
ie Methoden der Online-
s gefährlich Stimmung

ag nach dem Foto erscheint,
beiden Spieler sollten nicht
nt – aber nach Ansicht vieler

falsch – und so zum Auslöser für eine Debatte über die Qualität von Meinungsumfragen geworden ist. Am Dienstag dieser Woche berät der Presserat darüber, ob die Umfrage so hätte erscheinen dürfen.

Wozu (trotz aller Kritik) Surveys?

- „If you love surveys or sausages, you should not watch either being made" (J. Kochevar)
- Für viele Forschungszwecke sind Surveys aber essenziell
  - Beispiele?
  - Messung von Einstellungen, Meinungen und „beliefs" als Motive für Handlungen
  - Verteilung von Handlungen, Eigenschaften und damit verbundene Ungleichheiten
  - …
- Viele Aspekte in prozessproduzierten Daten nicht erfasst
  - Daher ist z.B. auch der Zensus eine wichtige Quelle amtlicher Statistik
- Wichtig aber auch hier: Daten, die Forschungszwecke möglichst gut erfüllen
  - Möglichst „valide" Deskriptionen oder Kausalanalysen
  - Möglichkeit angestrebter Verallgemeinerungen
  - Durchführbar mit gegebenen (Zeit-)Ressourcen / hinreichend schnell und günstig

Total Survey Error (TSE)

- TSE = Measurement Errors + Representation Errors
  - Antworten spiegeln Merkmale d. Befragten nicht akkurat wider
  - Sample an Befragten bildet Grundgesamtheit (interessierende Population) nicht adäquat ab
- Alle Fehler können die interne & externe Validität bedrohen
  - Etwa da Konzepte oder Gruppen nicht richtig operationalisiert und gemessen werden
  - Selbstselektionen statt Zufallsauswahlen vorliegen
  - …
- Das TSE Konzept berücksichtigt neben möglichen Fehlern auch Ressourcen / Kosten
  - Damit Effizienz verschiedener Ansätze
  - Etwa: Lohnen Incentives im Hinblick auf genauere Schätzungen?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Measurement

Representation

- Schritte bei Durchführung Surveys und mögliche Fehlerquellen
- In diesem Kapitel: Fokus auf Measurement
- (Für Representation: s. Kapitel 4 zu Stichproben)
- Was für Fehlerquellen gibt es bei Measurement? Beispiele?

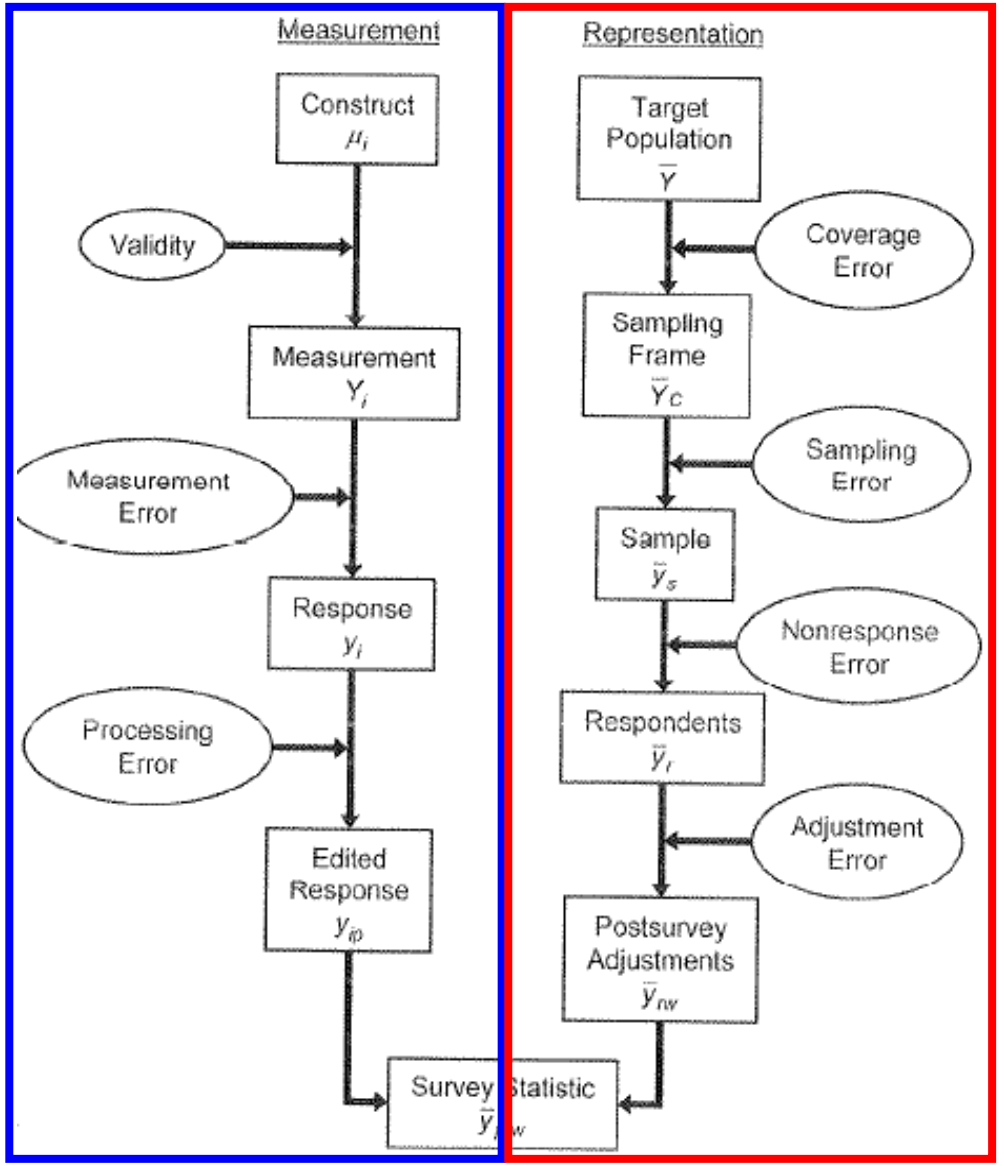
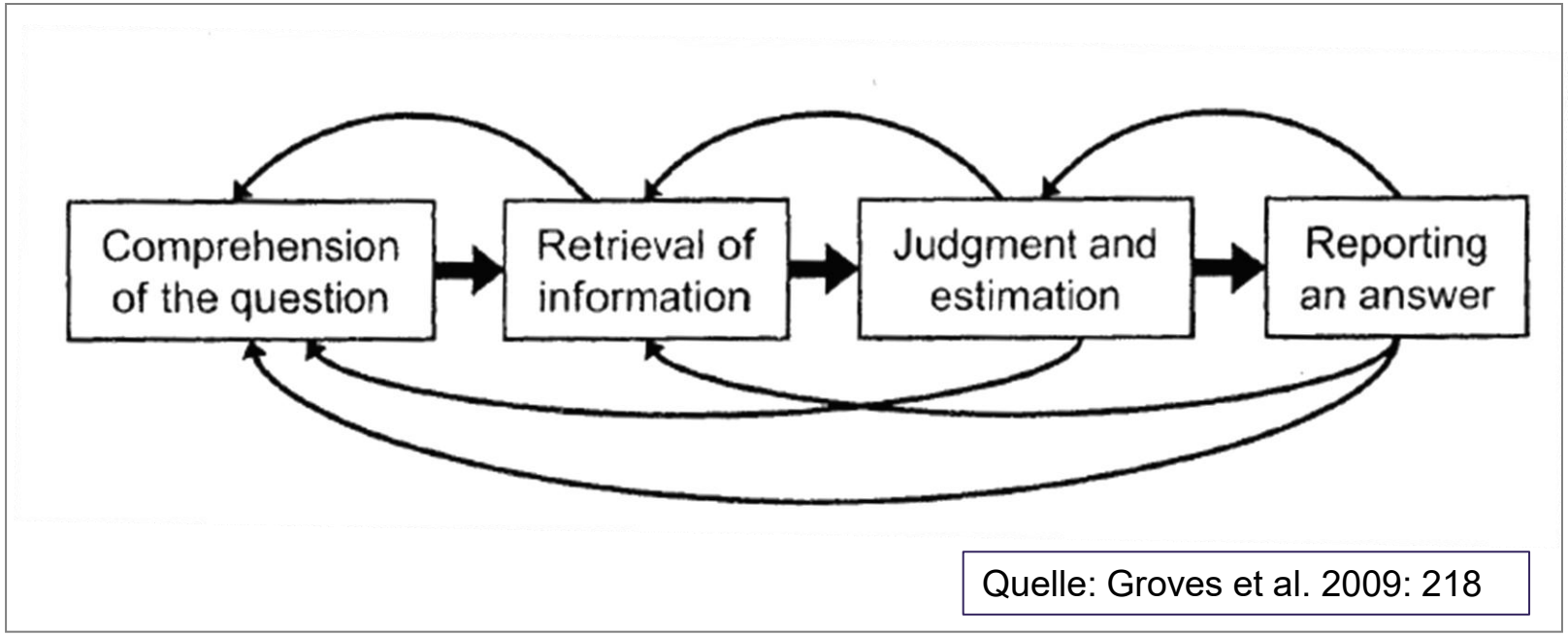


Figure 2.5 Survey lifecycle from a quality perspective.

Messfehler

- Definition: „Discrepancies between the true answer to a question and the answer that finds its way into the final database“ (Groves et al. 2009, S. 225)
- Fehler möglich durch Befragte, Interviewer*innen und weitere am Datenproduktionsprozess Beteiligte (z.B. Vercodende): dadurch auf oft stark „mode“-spezifisch
- U.a. da mehrere aktive Leistungen erforderlich sind



Quelle: Groves et al. 2009: 218

Erklärungsansätze und ableitbare Hypothesen

- Kognitionspsychologische Leistung
  - Verstehen, Erinnerungsleistung etc. als komplexe Leistung
  - Fällt leichter bei z.B. guten kognitiven Fähigkeiten; hohem Interesse an Befragung/Themen; “Ankern” die Erinnerung erleichten; weniger weit zurückliegenden Ereignissen; bereits geformten Meinungen etc.
  - Damit Erklärung Anker-Effekte, Reihenfolge-Effekte etc.
- Handlungsentscheidung mit Kosten und Nutzen
  - Valide Antwort wahrscheinlicher bei geringen Kosten (z.B. viel Zeit, geringe Sanktionsgefahr) und hohem Nutzen (z.B. starkem Interesse an Meinungsäußerungen, Wissenschaft)
  - „Satisficing“ statt „Optimizing“
  - Erklärung sozialer Erwünschtheit, Responsesets (Antwortmuster), Akquieszenz (Zustimmungstendenz) und andere „Shortcuts“

Nicht nur Befragte sind relevant!

"Lest you forget, Brimmer, you were hired to do market research. These reports of yours are becoming more and more autobiographical."

For those interested see also:
For and Against National Service | Yes, Prime Minister
https://www.youtube.com/watch?v=ahgjEjJkZks

New Yorker, XXX

Übungsaufgabe

- In einer Befragung des IfS sollten im Jahr 2016 mit dem Konzept der „gruppenbezogenen Menschenfeindlichkeit“ u.a. feindselige Einstellungen gegenüber Muslimen erhoben werden.

1. Beschreiben Sie kurz, wie muslimenfeindliche Einstellungen in der Befragung gemessen wurden.
2. Nennen Sie kurz ein Beispiel für jeden der im Total Survey Error unter „Measurement“ genannten möglichen Fehlerquellen (Groves et al. 2009, S. 48; idealerweise bezogen auf das Beispiel der Messung muslimenfeindlicher Einstellungen).
3. Möglichkeiten, das Vorliegen solcher Effekte empirisch zu prüfen? Wie würden Sie dazu vorgehen?
4. Oftmals werden monetäre Incentives eingesetzt, um die Befragungsteilnahme zu erhöhen. Unter welchen Bedingungen reduzieren Incentives vermutlich den Nonresponse-Bias? Können sie den Bias möglicherweise auch verstärken? Diskutieren Sie das idealerweise kurz am vorliegenden Beispiel.

Lösungsansätze: Messung

- Messung und Kritik

29.	Im Zusammenhang mit der Diskussion über Zuwanderung und Integration würde uns auch interessieren, inwieweit Sie den folgenden Aussagen zum **Islam** zustimmen.					
		Stimme gar nicht zu	**Stimme eher nicht zu**	**Teils/ teils**	**Stimme eher zu**	**Stimme voll und ganz zu**
	Die muslimische Kultur passt gut nach Deutschland.	☐	☐	☐	☐	☐
	Die Sitten und Bräuche des Islam sind mir nicht geheuer.	☐	☐	☐	☐	☐
	Es gibt zu viele Muslime in Deutschland.	☐	☐	☐	☐	☐
	Die Verschleierung von Frauen im Islam ist frauenfeindlich.	☐	☐	☐	☐	☐

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Lösungsansätze

29.	Im Zusammenhang mit der Diskussion über Zuwanderung und Integration würde uns auch interessieren, inwieweit Sie den folgenden Aussagen zum **Islam** zustimmen.					
		Stimme gar nicht zu	Stimme eher nicht zu	Teils/ teils	Stimme eher zu	Stimme voll und ganz zu
	Die muslimische Kultur passt gut nach Deutschland.	☐	☐	☐	☐	☐
	Die Sitten und Bräuche des Islam sind mir nicht geheuer.	☐	☐	☐	☐	☐
	Es gibt zu viele Muslime in Deutschland.	☐	☐	☐	☐	☐
	Die Verschleierung von Frauen im Islam ist frauenfeindlich.	☐	☐	☐	☐	☐

- Messung und Kritik
- Validity: Zielkonzept ↔ Frage
  - Messung von Fremddimensionen (etwa Sorgen um den sozialen Frieden oder die Integration von Migrantinnen; Einstellungen zu Geschlechterrollen)
  - Einige Items messen ggf. auch primär Xenophobie?
  - „Muslimische Kultur“? Etliche Konzepte sind sehr vage
- Measurement Error: wahrer Wert bei Befragten ↔ Antwort
  - Etwa Verzerrungen durch soziale Erwünschtheit
  - Response-Sets, Akquieszenz
- Processing Error: Antwort ↔ Codierung im Datensatz
  - Fehlkodierungen von Antworten
  - Fehler bei Datenaufbereitung etc.

Lösungsansätze: Testung, Incentives

- Möglichkeiten der empirischen Testung? U.a.:
  - Korrelation mit „Fremddimensionen“, kognitive Interviews etc.
  - Variation der Anonymität um Effekte sozialer Erwünschtheit zu testen
  - Prüfung der Daten auf Fehler und auffällige Muster (z.B. Ankreuzen der Mittelkategorie, geringe Antwortzeiten, inkonsistente Antwortmuster, …)
- Wirkung von Incentives?
  - Können Non-response error reduzieren, wenn sie gegenüber einer Zufallsstichprobe unterrepräsentierte Gruppen eher zu Antworten motivieren und dabei ins. Selektionsbias reduzieren
  - Können Sampling Error aber auch erhöhen, falls Überrepräsentation/Selektionsbias steigt

  ➢ Inhaltliche Hypothesen hilfreich – wer wird motiviert, Zusammenhang mit Y

  - (Zu beachten sind auch Effekte auf das weiteren Antwortverhalten)

Appendix

- https://www.youtube.com/watch?v=ahgjEjJkZks

Die Bundesheer-Umfrage

Startseite › Österreich

Bundesheer | Mehrheit findet Grundwehrdienst zu kurz

Die Dauer des Grundwehrdiens … Landesverteidigung … ıa
geworden. Hier die Ergebnisse

07.57 Uhr, 04. Dezember 2019

Mehrheit der Österreicher für längeren Grundwehrdienst

Mehrheit der Bevölkerung findet Grundwehrdienst zu kurz

Weitere Milizübungen in der Dauer von zwei Monaten werden als notwendig erachtet

4. Dezember 2019, 07:00 1.372 Postings

https://www.derstandard.at/story/2000111857083
https://www.diepresse.com/5732966/mehrheit-der-osterreicher-fur-langeren-grundwehrdienst
https://www.kleinezeitung.at/oesterreich/5732944/Bundesheer_Mehrheit-findet-Grundwehrdienst-zu-kurz

....

Die Bundesheer-Umfrage

Startseite › Österreich

Bundesheer | Mehrheit findet Grundwehrdienst zu kurz

Die Dauer des Grundwehrdiens… Landesverteidigung
geworden. I…
07.57 Uhr, 0…

POLITIK INLAND

04.12.2019

Heer: Ex-FPÖ-Minister Kunasek sieht in Umfrage Bestätigung

Steirischer FPÖ-Chef: Reduktion des Wehrdienstes auf sechs Monate war "klare Fehlentscheidung"

Weitere Milizübungen in der Dauer von zwei Monaten werden als notwendig erachtet

4. Dezember 2019, 07:00 1.372 Postings

https://www.derstandard.at/story/2000111857083
https://www.diepresse.com/5732966/mehrheit-der-osterreicher-fur-langeren-grundwehrdienst
https://www.kleinezeitung.at/oesterreich/5732944/Bundesheer_Mehrheit-findet-Grundwehrdienst-zu-kurz

....

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Eine genauere Betrachtung

Wahrscheinlichkeit von Bedrohungen

Im Folgenden finden Sie eine Reihe von Ereignissen, die die Sicherheit Österreichs bedrohen könnten. Was glauben Sie, wie wahrscheinlich sind diese Ereignisse bzw. wie wahrscheinlich könnten sie eintreten? (Befragung November 2019)
Negative Auswirkungen des Klimawandels
Natur– oder technische Katastrophen in Österreich
Cyberangriffe in Österreich
Massenmigration nach Österreich
Neutralitätsverletzung des österr. Luftraums
Vereinzelte Terroranschläge in Österreich
Blackout in Österreich
Pandemie
Groß angelegter Terrorangriff auf Österreich
Hybrider Konflikt

Auswirkungen von Bedrohungen

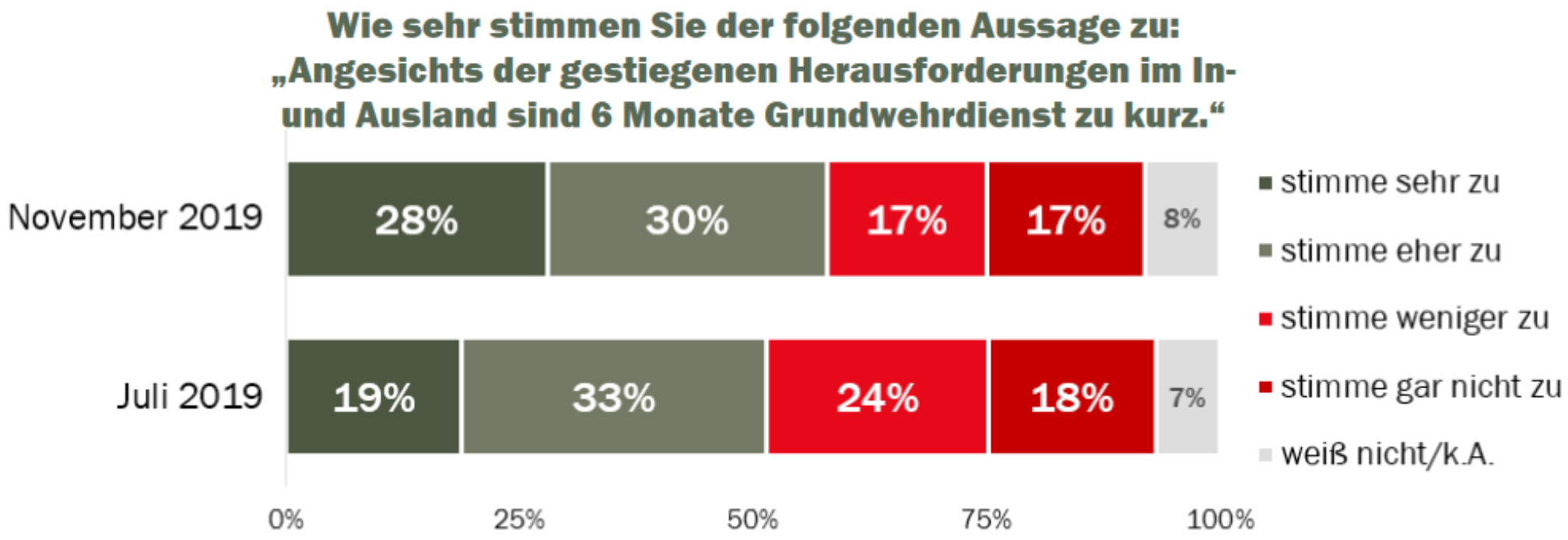
Falls die folgenden Ereignisse tatsächlich eintreten sollten. Wie schätzen Sie die möglichen Auswirkungen dieser Ereignisse auf Österreich ein? (Befragung November 2019)
Negative Auswirkungen des Klimawandels
Natur– oder technische Katastrophen in Österreich
Cyberangriffe in Österreich
Massenmigration nach Österreich
Neutralitätsverletzung des österr. Luftraums
Vereinzelte Terroranschläge in Österreich
Blackout in Österreich
Pandemie
Groß angelegter Terrorangriff auf Österreich
Hybrider Konflikt

Bedeutsamkeit von Bedrohungen

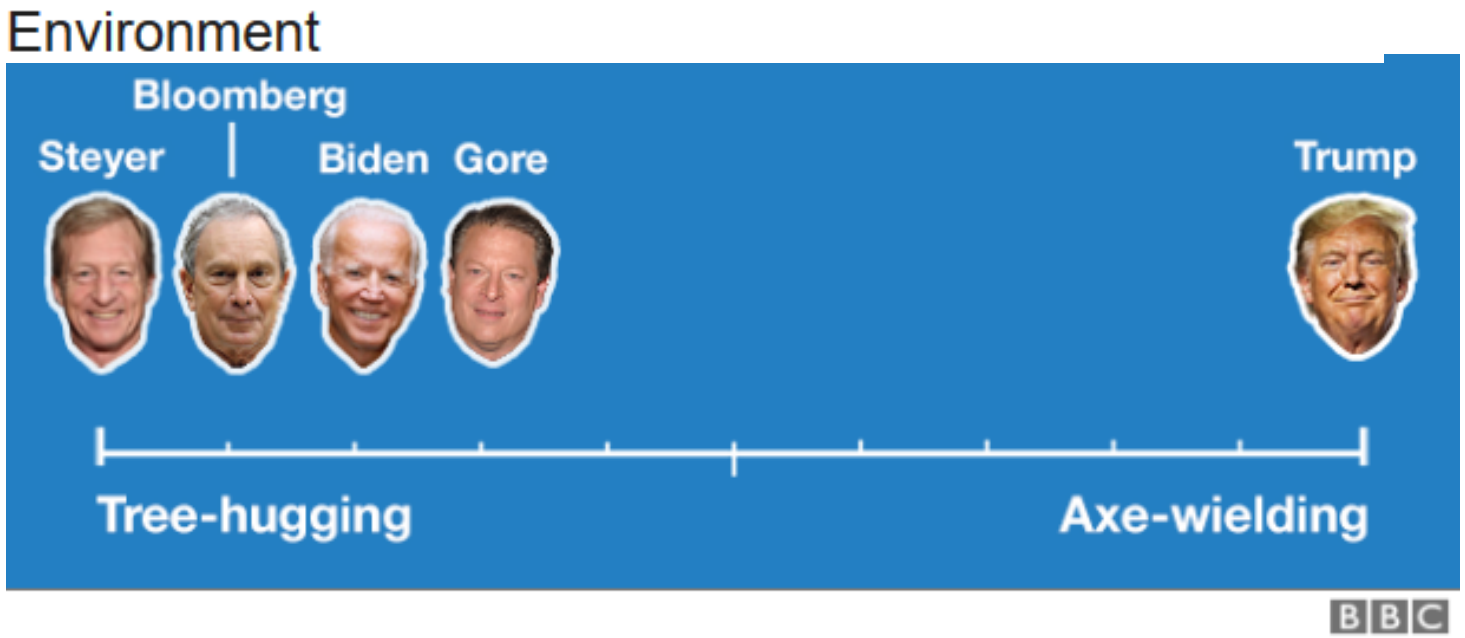
Geben Sie uns bitte nun Ihre Einschätzung, ob diese Ereignisse in Zukunft bedeutsamer für die Sicherheit Österreichs werden, etwa gleichbleiben oder weniger bedeutsam werden? (Befragung November 2019)
Negative Auswirkungen des Klimawandels
Natur– oder technische Katastrophen in Österreich
Cyberangriffe in Österreich
Massenmigration nach Österreich
Neutralitätsverletzung des österr. Luftraums
Vereinzelte Terroranschläge in Österreich
Blackout in Österreich
Pandemie
Groß angelegter Terrorangriff auf Österreich
Hybrider Konflikt

Sollte das österreichische Bundesheer für diese vielfältigen Aufgaben besser vorbereitet werden oder ist das eher nicht erforderlich?	auf jeden Fall	eher schon	teils/ teils	eher nicht	sicher nicht	weiß nicht	keine Antwort
Gesamt November 2019	45	28	18	4	2	2	0

Eine genauere Betrachtung



Examples for clear, easy to understand, unidimensional, and neutral endlabels



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Total Error, Replikationen, Publication Bias und Meta-Analysen

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Übungsaufgabe

1. Welche Form der Unsicherheit wird mit Konfidenzintervallen und Standardfehlern abgebildet?
   Wie hängt diese mit der Art von Stichprobenziehungen zusammen (Größe, etc.)?
2. Was gibt es für „Non-Standard-Errors? Erläutern Sie mindestens zwei Fehler anhand eines konkreten, selbstgewählten Forschungsbeispiels. Gehen Sie zudem darauf ein, wie diese mithilfe von Replikationen (welcher Art) aufspürbar sind.
3. Beschreiben Sie kurz, was Meta-Analysen sind
   (Ziel, worin besteht das Forschungsdesign bzw. die besondere Methodik).
4. Eine Fehlerquelle, die speziell in Meta-Analysen aufgedeckt wird, ist „Publication Bias“.
   Was ist damit gemeint? Wie lässt sich der Bias entdecken?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Übungsaufgabe Nr. B3

Philipp Häußler

„ Teilaufgabe 2: Was gibt es für „Non-Standard-Errors? Erläutern Sie mindestens zwei Fehler anhand eines konkreten, selbstgewählten Forschungsbeispiels. Gehen Sie zudem darauf ein, wie diese mithilfe von Replikationen (welcher Art) aufspürbar sind.“

„Non-Standard-Errors“

- Werden nicht über den Standardfehler erfasst
- Verschiedene Formen
  - „Coverage Errors“
    - Auftreten von Fehlern durch unvollständige oder veraltete Listen
  - „Non-Response Errors“
    - Fehler durch Fehlen von Antworten von Personen(-gruppen)
  - Messfehler
  - Publication Bias?
- Folge: nicht zufällige Stichprobenzusammensetzung führt zu systematisch verzerrten Ergebnissen
  - Problem: Verzerrung der Stichprobe kann nicht erfasst werden

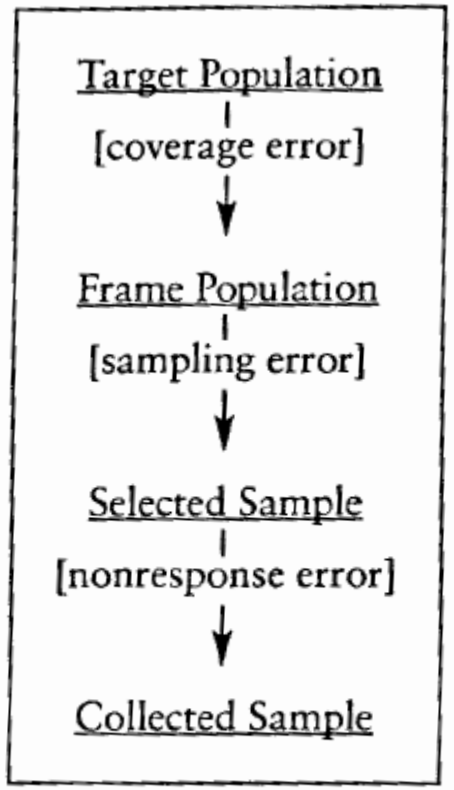


Figure 4.1. Sources of Exclusion Error in Data Collection

Firebaugh (2008)

„Non-Response Error“

- Befragung von Restaurantbesitzer:innen zum Stand der Umsetzung aufwendiger Hygienemaßnahme während der Coronapandemie
  - Unterdurchschnittliche Rücklaufquote mit hoher Umsetzungsquote
  - Verzerrung, da sich hauptsächlich Gastwirt:innen zurückmeldeten, die die Maßnahmen tatsächlich umgesetzt haben
- Lösung: Interne Replikation
  - Wiederholung der Studie mit Ziehung einer Stichprobe aus den „Nicht-Respondenten“ und anschließendem Vergleich

„Coverage Error“

- Messung der Zufriedenheit von Mitarbeitenden in DAX-Konzernen
  - Erfassung anhand der unternehmensinternen Mitarbeiterbefragungen
  - Verzerrung, da befristet angestellte Personen in solchen Umfragen häufig nicht berücksichtigt werden
- Lösung: Externe Replikation
  - Vergleich mit weiterer Studie, die in Stichproben auch Zeitarbeitskräfte berücksichtigt

Literatur

- Firebaugh, G. (2008). *Seven Rules for Social Research*. Princeton University Press.

„Total Error“: Standard und Non-Standard

- Standard: Konfidenzintervalle, Standarderrors: quantifizieren Zufallsfehler durch Verwendung einer Stichprobe statt Grundgesamtheit (s. Kap. 4); sinkt mit Stichprobengröße
- Warum reicht das nicht? Was sind „Non-Standard“ Errors?
- Etliche weitere Quellen für zufällige und systematische Fehler (Bias)
  - Verzerrte Stichproben (coverage und nonresponse error)
    - – Selektion nach Y verursacht Bias (s. Kap. 4)
  - Messfehler
  - Coding Errors
  - Selektives Publizieren
  - …

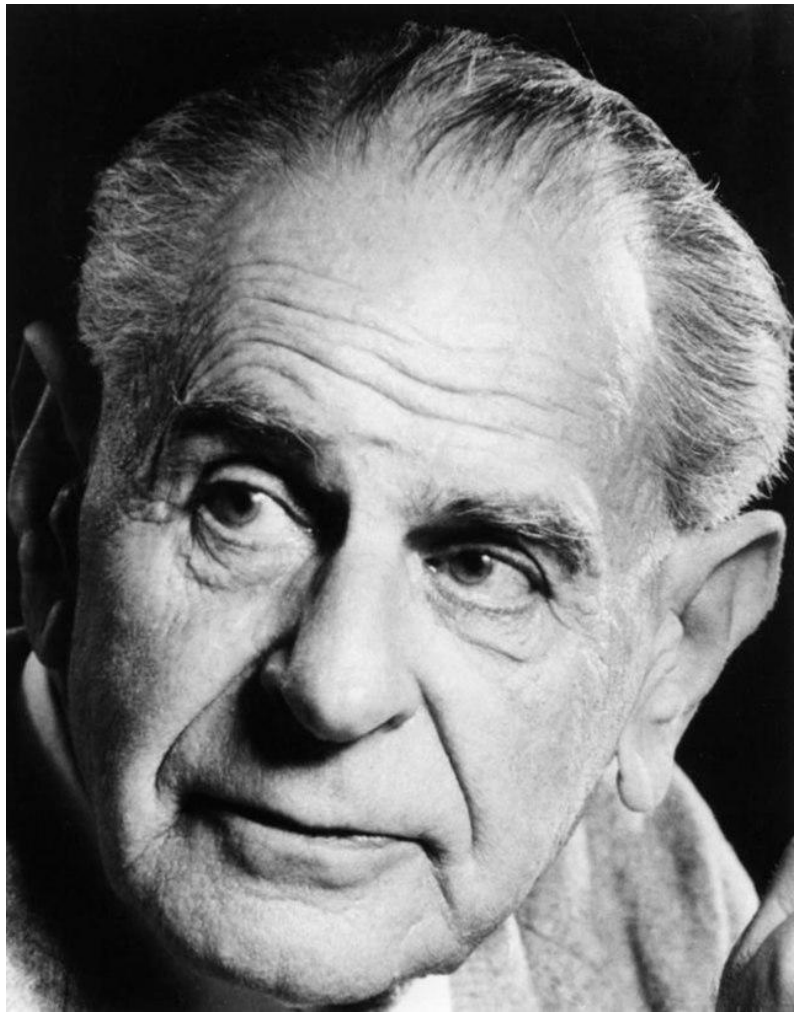
Weil Standardfehler diese nicht-propabilistischen Fehler nicht berücksichtigen unterschätzen konventionelle KIs die Unsicherheit in den analytischen Resultaten

- Möglichkeiten der Aufdeckung und Quantifizierung?
  - U.a. durch Wiederholungen (Replikationen)
  - Mit dem gleichen oder einem anderen Design?
    - ➢ Beides ist wichtig!

Gleich: Aufdeckung von Fehlern / Bedrohungen interner Validität;
Anderes Design: Aufdeckung von Moderatoren / Bedrohung externer Validität

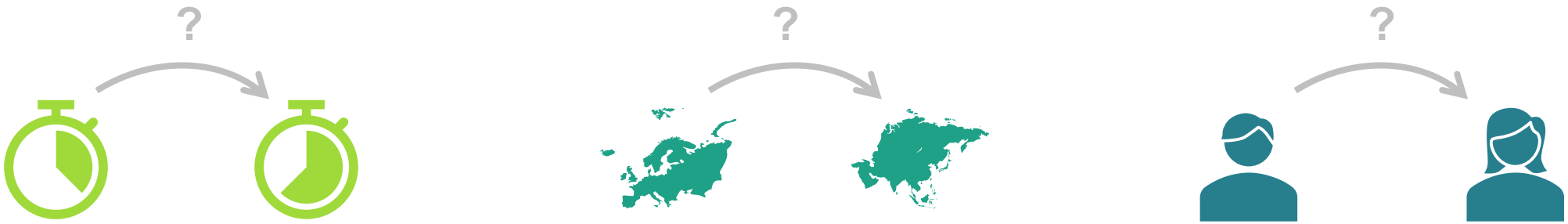
Wissen durch Wiederholung…

„Only by such **repetitions** can we convince ourselves that we are not dealing with a mere isolated ‘coincidence,’ but with events that, because of their **regularity and reproducibility**, are in principle intersubjectively testable.” (Popper 1959: 45)



…und die Rolle von Replikationen

"A replication experiment to demonstrate that the same findings can be obtained in any other place by any other researcher is […] proof that the experiment reflects **knowledge that can be separated from the specific circumstances** (such as time, place, or persons) under which it was gained." (Schmidt 2009: 90)



Zweifel an…

Interner Validität

→ Korrektheit?

Neutrinos Travel Faster Than Light, According to One Experiment

Others doubt the mind-boggling claim, which would overturn Einstein's theory of special relativity

22 SEP 2011 · BY ADRIAN CHO

Once Again, Physicists Debunk Faster-Than-Light Neutrinos

Five different groups agree that the elusive particles obey Einstein's speed limit after all

8 JUN 2012 · BY ADRIAN CHO

Externer Validität

→ Verallgemeinerbarkeit?

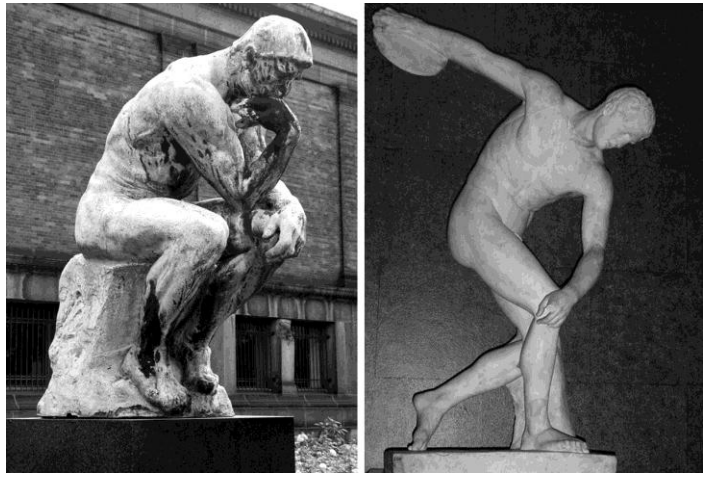
WEIRD

e	d	n	i	e
s	u	d	c	m
t	c	u	h	o
e	a	s		c
r	t	t		r
n	e	r		a
	d	i		t
		a		i
		l		c
		i		
		z		
		e		
		d		

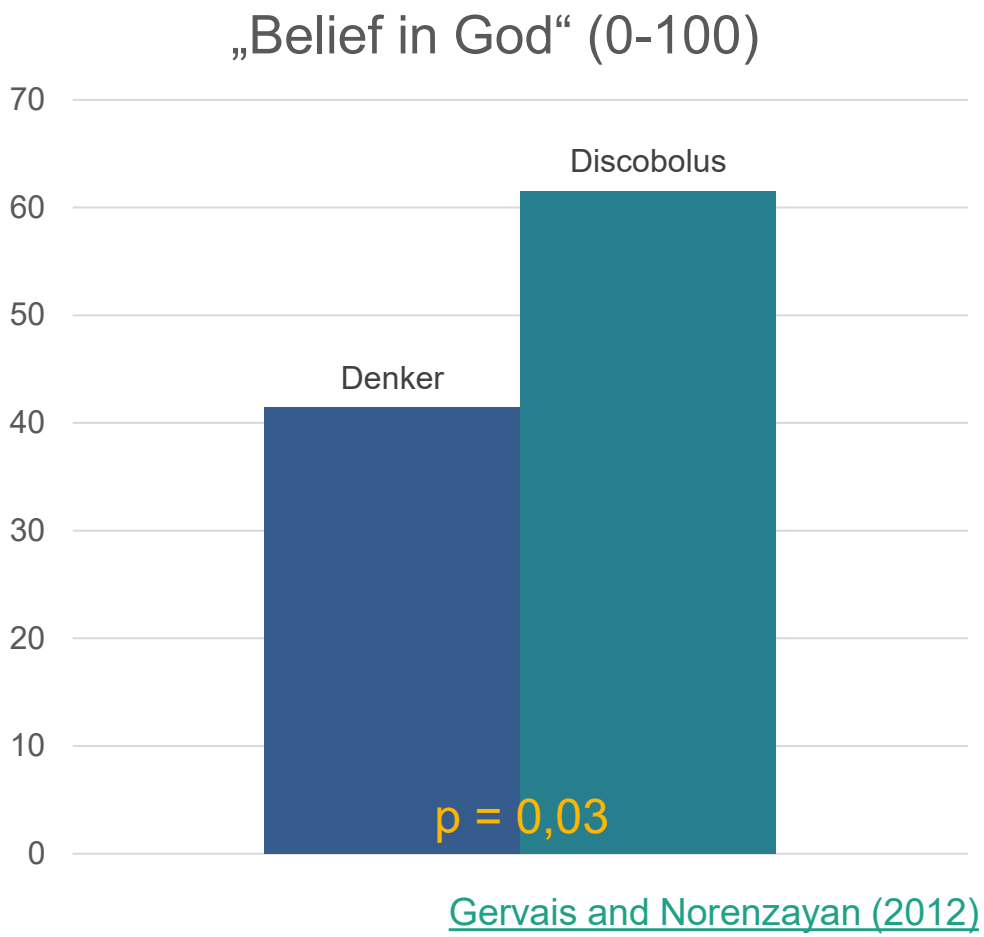
„Too good to be true“ Ergebnisse

Beispiel 2: Atheismus

Führt „Priming“ mit analytischen Motiven zu religiösem Skeptizismus?



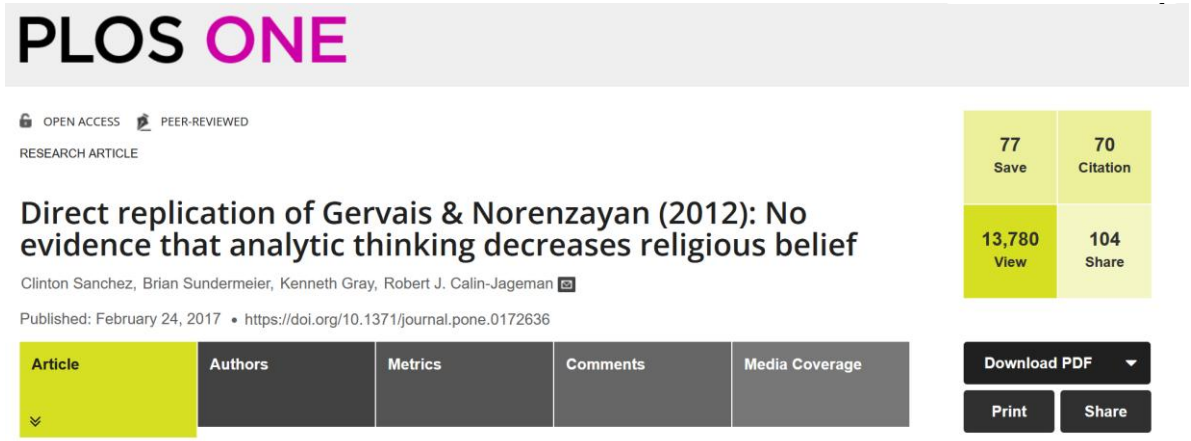
→ Plausibilität der Signifikanz/Effektstärke?



Gervais and Norenzayan (2012)

„Too good to be true“ Ergebnisse

Beispiel 2: Atheismus


PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Direct replication of Gervais & Norenzayan (2012): No evidence that analytic thinking decreases religious belief

Clinton Sanchez, Brian Sundermeier, Kenneth Gray, Robert J. Calin-Jageman

Published: February 24, 2017 • https://doi.org/10.1371/journal.pone.0172636

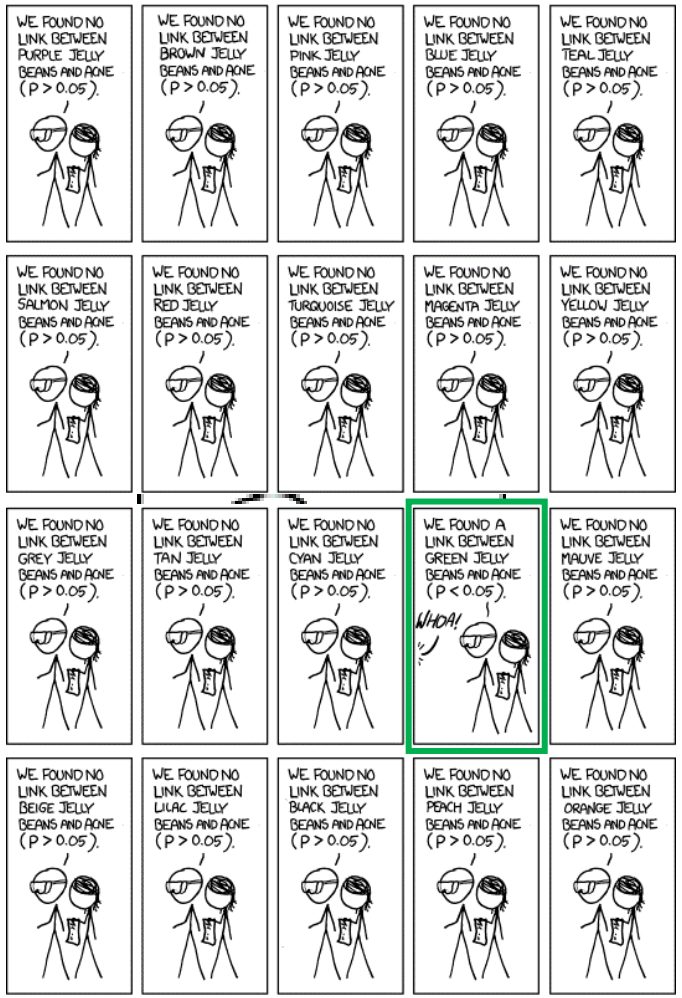
77 Save | 70 Citation | 13,780 View | 104 Share

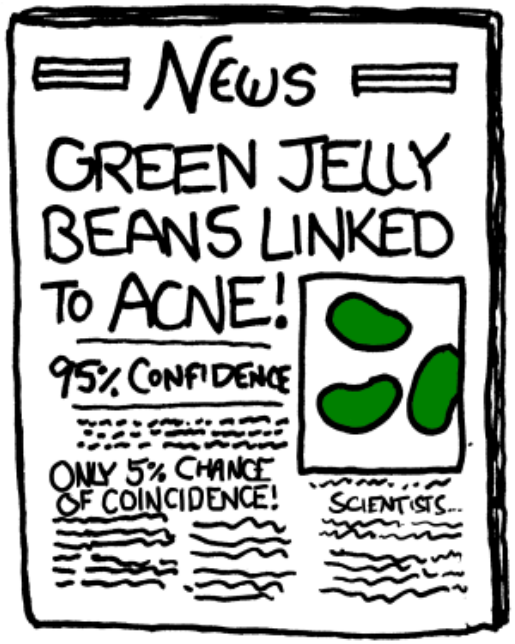
Article | Authors | Metrics | Comments | Media Coverage

Download PDF | Print | Share

“We conducted a precise, large, multi-site pre-registered replication of one of these experiments. We observed little to no effect of the experimental manipulation on religious belief (d = 0.07 in the wrong direction, 95% CI[-0.12, 0.25], N = 941). The original finding does not seem to provide reliable or valid evidence that analytic thinking causes a decrease in religious belief.”

Publication Bias: Selektives Publizieren signifikanter und/oder hypothesen-konformer Ergebnisse





XKCD (modifiziert)

Übungsaufgabe Nr. B3

Ozan Aykaç

„ Teilaufgabe 2: Was gibt es für „Non-Standard-Errors?
Erläutern Sie mindestens zwei Fehler anhand eines konkreten, selbstgewählten Forschungsbeispiels.
Gehen Sie zudem darauf ein, wie diese mithilfe von Replikationen (welcher Art) aufspürbar sind.“

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Kernergebnisse 3 & 4

Meta-Analysen

- Forschungsdesign der **sekundären Evidenzsynthese**
- Analyseeinheit: **Studien / Effektschätzer**, nicht Individuen

Ziele:

- Präzisere Gesamtschätzung
- Analyse zwischenstudialer Heterogenität
- Meta-Analyse = **formalisierte externe Replikation**

Publication Bias & p-Hacking

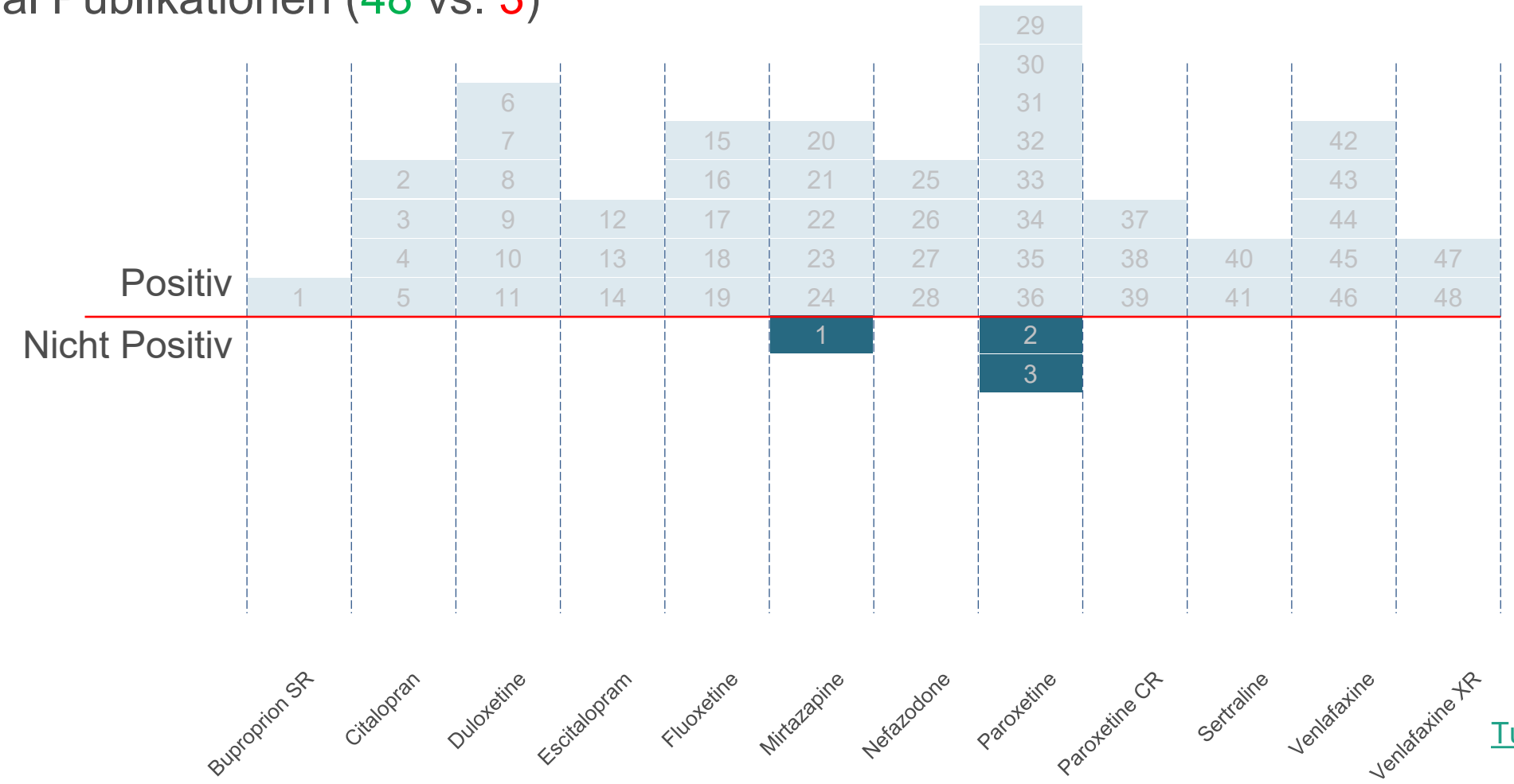
- Selektion auf Signifikanz verzerrt die publizierte Literatur
- p-Hacking produziert Signifikanz durch flexible Analyseentscheidungen
- Meta-Analysen können dadurch **false precision** erzeugen

Diagnoseinstrumente

- **Funnel Plots**: Asymmetrie als Hinweis auf Signifikanzselektion
- **Egger-Test**: formale Prüfung der Asymmetrie
- **p-Wert-Verteilungen**: Häufung knapp unter 0,05 als Hinweis auf p-Hacking
- **Kohortenvergleiche / Registerdaten**: direkte Evidenz für Publication Bias

Publication Bias & Spin: Antidepressiva

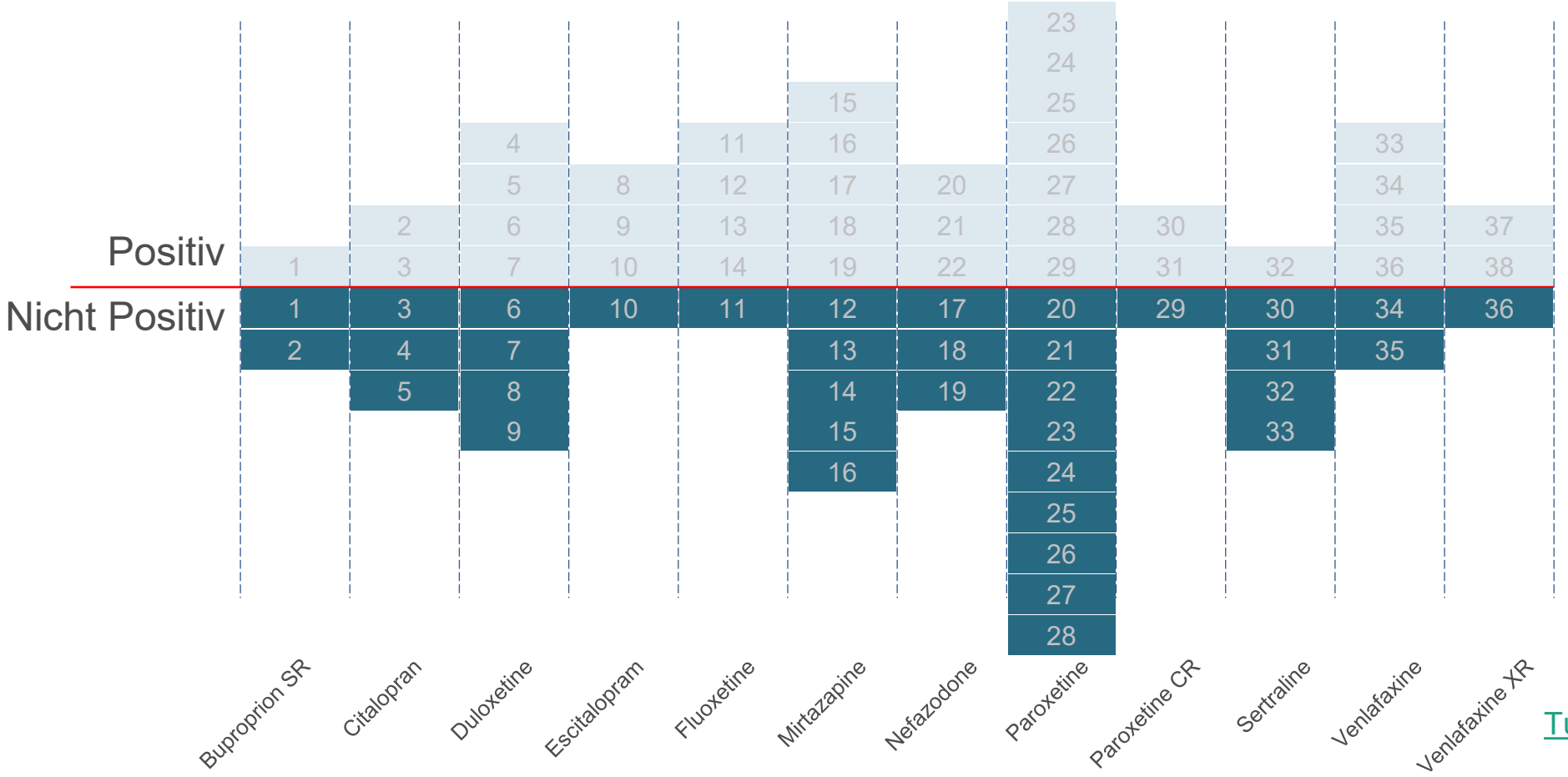
Journal Publikationen (48 vs. 3)



Turner et al. (2008)

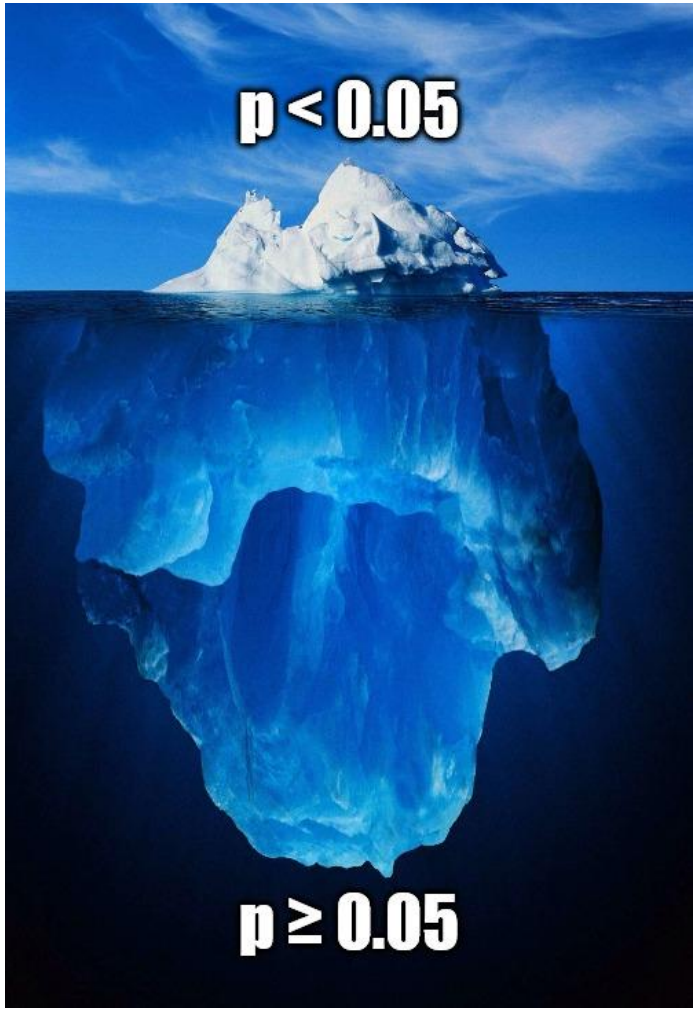
Publication Bias & Spin: Antidepressiva

Food and Drug Administration (38 vs. 36)



Turner et al. (2008)

Eisberg voraus!

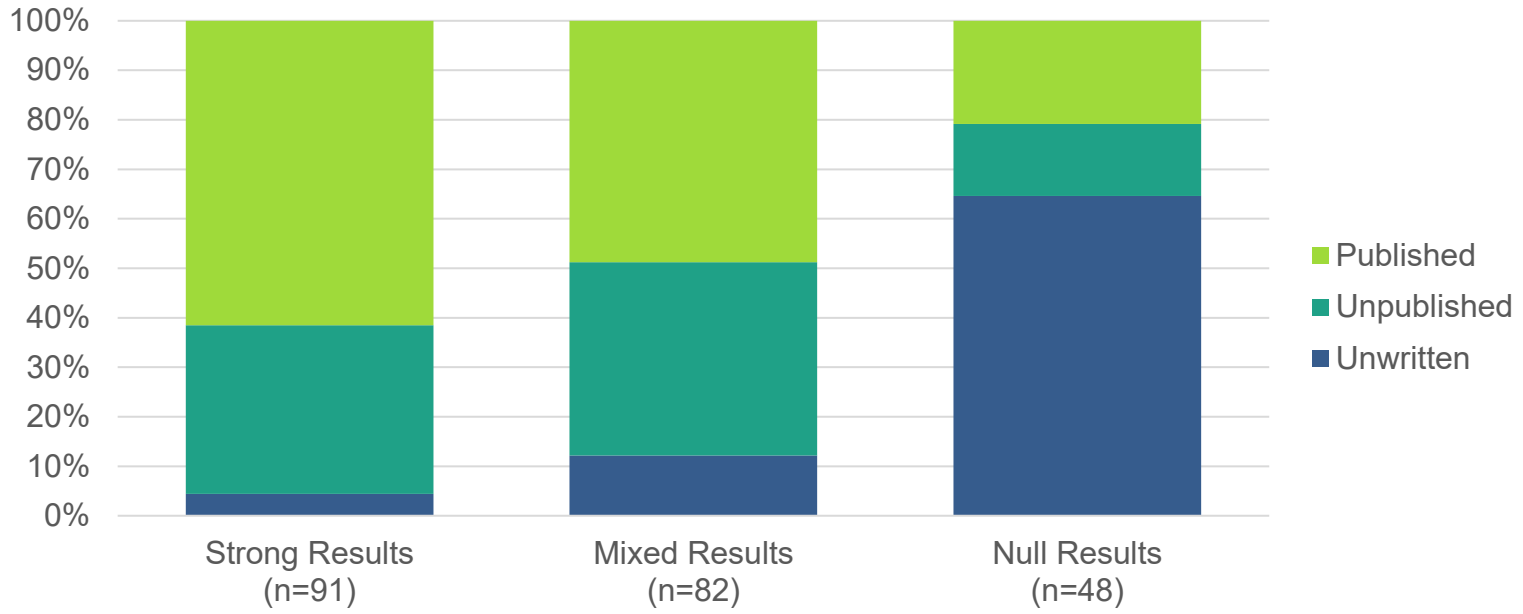


Imgflip

Der „File Drawer“ in den Sozialwissenschaften

Publikationsstatus von Experimenten aus dem TESS-Förderprogramm?

(TESS = Time Sharing Experiments in the Social Sciences)



→ “Null-Ergebnisse erblicken selten das Tageslicht”

Franco et al. (2014), Mervis (2014)

Gründe für das „Scheitern“ von Replikationen

- **Zufall**
  „Glück“ in der Originalstudie (95% Konfidenzintervall: 5% Alpha-Fehler)
- **Publication Bias** (s. vorherige Folien)
  Selektion nach Signifikanz führt zu vielen „falsch positiven“ Ergebnissen in der Literatur
- **Manipulation**
  Bewusst/Unbewusst (z.B. „optional stopping“, Outlier, fehlerhaft kodierte Daten, …)
- **Mangelnde externe Validität**
  Insbesondere „Scope Conditions“ und Effektheterogenität
- **Fehler**
  z.B. bei Replikation läuft etwas schief

Ein nützliches Werkzeug: Meta-Analysen

- Quantitativ-statistische Zusammenfassung eines Forschungsstands

- **Effizienz** (finanziell & statistisch!)
- **Objektivität** (kein Cherry-Picking)
- **Aber:** „Garbage in, garbage out“
  + weitere potenzielle Probleme

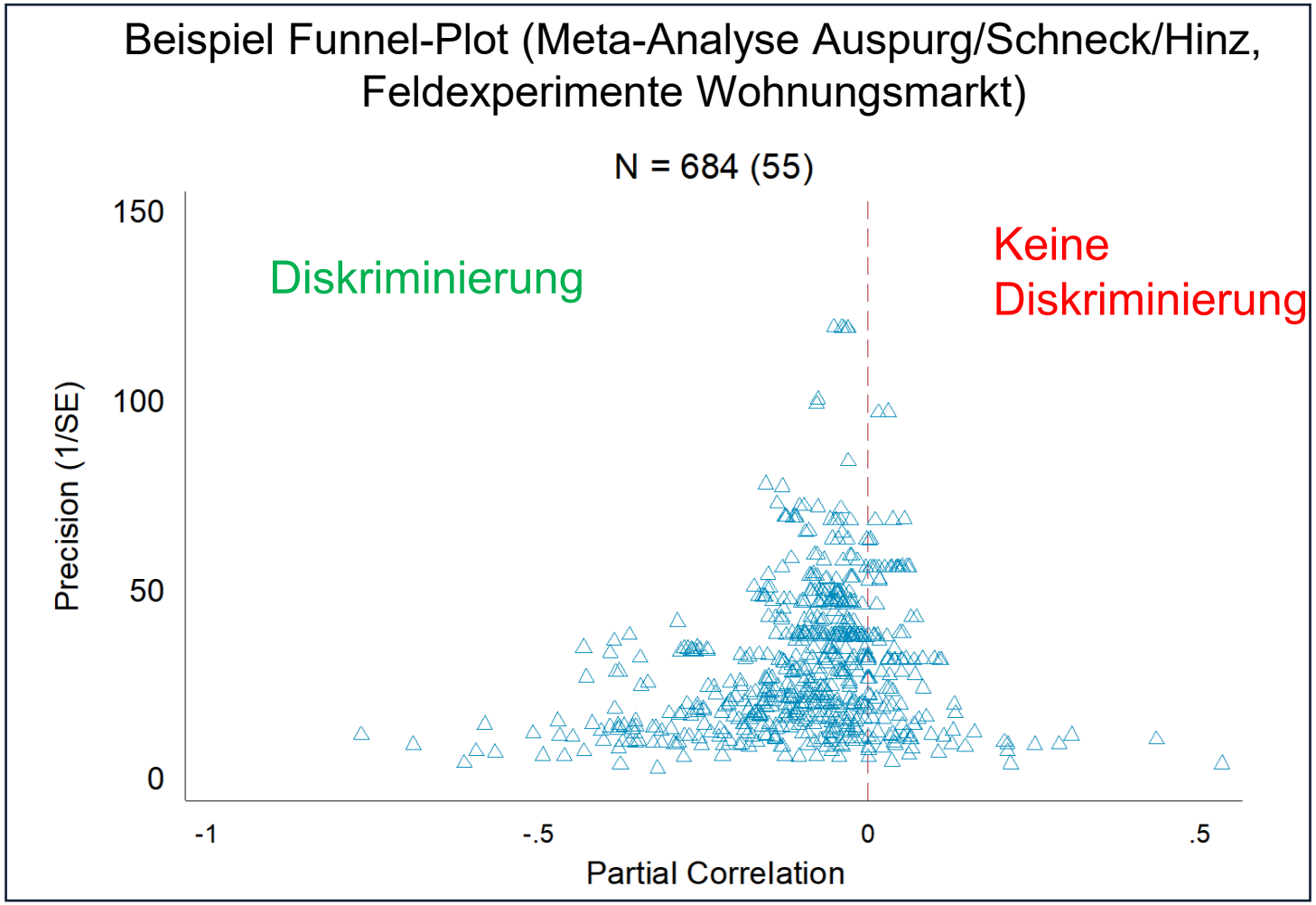
Schematisches Vorgehen:

1. **Systematische Literaturrecherche**
   (z.B. basierend auf PRISMA)
2. **Codierung** der Primärliteratur:
   - Effektstärken
   - Statistische Parameter (p, n,...)
   - Sonstige Studienmerkmale
     (z.B. Setting, Methode, etc.)
3. **Berechnung** eines mittleren Effekts
4. **Ggf. weitere Analysen**

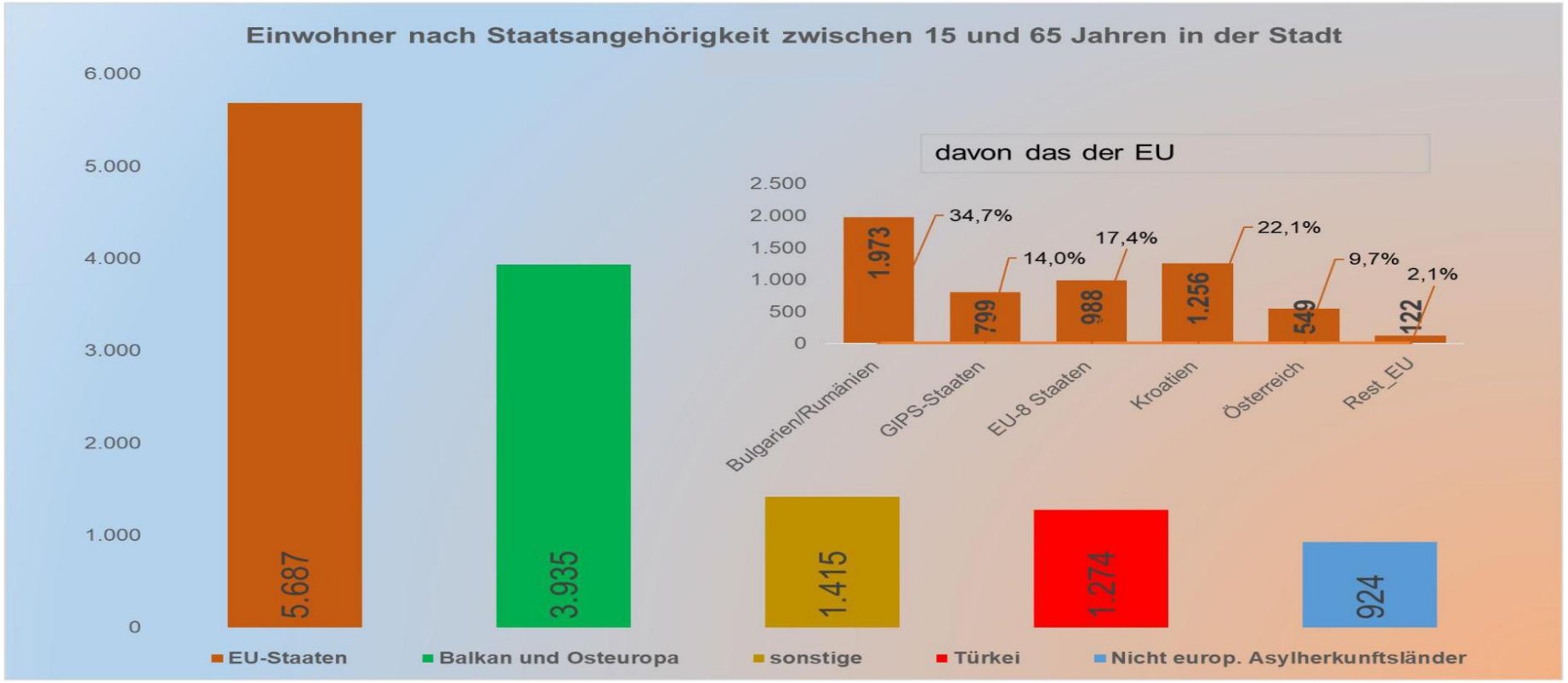
See also https://www.prisma-statement.org/

Aufdeckung von Publication Bias in Meta-Analysen

- Zusammenhang Effektstärke und Unsicherheit?
  - Bei unsicheren Ergebnissen (kleine Stichproben) per se starke Zufallsschwankung
  - Damit auch zufällig Extremergebnisse, die von wahrem Effekt abweichen
  - Werden diese „herausgepickt“, gibt es eine negative Korrelation zwischen Stichprobengröße und Effektstärke
  - Etwa grafisch erkennbar in „Funnel Plots“



See you in 2026



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Prozessproduzierte „Big Data“

„Big Data“ – Hope

- Die klassischen 3 V‘s
  - Volume (Datenmasse)
  - Variety (verschiedene, kombinierte Datenquellen)
  - Velocity (schnelle und dichte Datenerfassung)
- Als „big“ Daten gelten Daten mit
  - Vielen Beobachtungen („long“) und/oder
  - vielen Variablen („wide“)
  - (egal ob digital oder nicht)

- Insb. auch viele digitale „bit data“
  - Digitales Leben (z.B. Social Media)
  - Digitale Verhaltensspuren (z.B. Internethandel)
- Vorteile prozessproduzierter Daten, u.a. Dynamiken in „Echtzeit“

Beispiel: Nachfrage nach Corona-Schnelltests

Eigene Recherche mit Google Trend-Daten

Beispiel: Setzt Corona der Kopflaus ein Ende?

Eigene Recherche mit Google Trend-Daten

Beispiel aus Ungleichheitsforschung

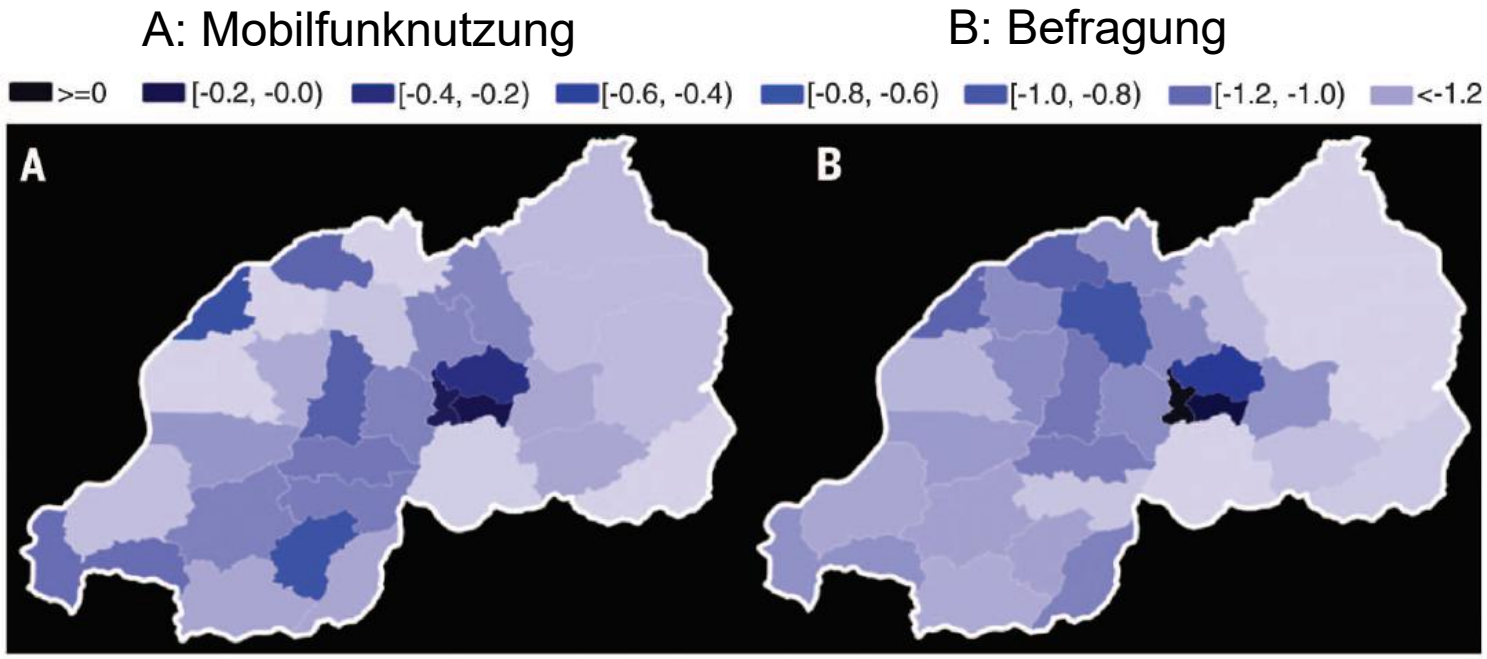
- Problem:
  - Befragungen werden gerade in ärmeren Ländern nur sporadisch durchgeführt
  - Bevölkerungsbefragungen sind teuer und langsam
- Methodik von Blumenstock, Cadamuro and On (2015):
  - Digitale Mobilfunkdaten von 1,5 Mio Nutzern in Ruanda
    - Häufigkeit Gespräche, soziale Netzwerke
    - Bewegungsprofile
  - Befragung von Mobilfunknutzer*innen (mit wenigen Teilnehmenden) zu ihrem Besitz
    - Z.B. im Hinblick auf Fahrzeuge
    - Abfrage Erlaubnis zum *Linkage* mit den Mobilfunkdaten
  - Amtliche Daten zur Soziodemographie
    - Gewonnen durch eine frühere Befragung

Schätzung regionale Vermögens-verteilung

Validierung der Methodik

Beispiel aus Ungleichheitsforschung

- Ergebnis: Mit Mobilfunkdaten kann die Vermögensverteilung „gut“ abgebildet werden

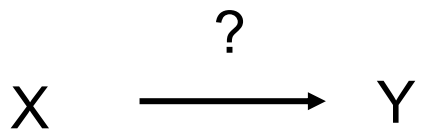


- Mobilfunkdaten erlauben kleinräumigere und raschere Schätzungen (einfache Updates möglich)

Blumenstock et al. 2015: 1075

Or Hype?

- „Big“ ist kein Wert an sich!
  - „a well defined sample of even a small size can tell us more than millions of poorly defined cases“ (Square 1988)
- Oft wichtiger: Das *Wie* und *Warum*
  - Forschung dient Beantwortung einer konkreten Forschungsfrage; oftmals Ursache-Wirkung (s. erste Sitzungen)



- Gute Forschungsdesigns ermöglichen gute Antworten
  - Isolierung des interessierenden Effekts
  - Ausschluss von alternativen Erklärungen
- „Better data, not more data are helpful“ (Salganik 2018)
  - Einschätzung mit den in den Grundlagen (Kap. A) besprochenen Gütekriterien
  - Oftmals fehlen wichtige Variablen, u./o. die beobachteten Einheiten sind selektiv

Beispiel Amazon

- Bis 1997 stellte Amazon noch Buchrezensenten ein, um Kunden Buchempfehlungen zu geben
- Neue Idee: Kunden auf Grundlage ihrer individuellen Kaufpräferenzen bestimmte Bücher empfehlen
- 2 Ansätze:
  - Analysieren, welche Kunden was kaufen und ähnlichen Kunden ähnliche Produkte empfehlen
  - Einfache Kaufkorrelationen zwischen Produkten berechnen
- “Knowing what, not why, is good enough.”

- Da diese Ansätze jedoch im Wesentlichen naiv sind, können Anomalien zu Problemen führen:

→ Warum Fragen bleiben also nach wie vor essenziell (zumindest um Fehler von Algorithmen verstehen zu können)

https://vikramoberoi.com/posts/an-internship-working-on-customers-who-bought-this-also-bought-at-amazon-16-years-ago/
Mayer-Schönberger & Cukier (2013): Big Data

Grundsätzlich: Datentypen mit Vor- und Nachteilen

Custommades

- David (Michelangelo): ein Kunstobjekt wird als solches hergestellt
- Bei Daten: Produktion von Daten für einen definierten Zweck
  (z.B. gezielte Datenerhebung für Sozialforschung)

Readymades

- Fountain (Marcel Duchamp): Pissoir wird, so wie es ist, als Kunstobjekt umgewidmet
- Bei Daten: Verwendung von Daten über eigentlichen Zweck hinaus
  (z.B. Telefondaten zur Schätzung Vermögensverteilung)

Quelle: Salganik 2017: 7

Übungsaufgabe

- Big Data werden u.a. dazu genutzt, um Erklärungen für Homophilie zu trennen (Präferenzen vs. Gelegenheitsstrukturen). Dazu werden insbesondere Online-Partnermärkte erforscht.

1. Was sind **Vorteile der Nutzung dieser „big data“** gegenüber Alternativen wie Surveys?
2. **Mögliche Bedrohungen von interner und externer Validität?** Fachbegriffe; welche Verzerrungen oder Uncertainty sind zu erwarten: Wie beeinflusst das die Identifikation des Treatment-Effekts und/oder seine Verallgemeinerbarkeit? Idealerweise Visualisierung.
   (s. dazu z.B. https://www.sociologicalscience.com/download/volume-2/january/comment-rejoinder/SocSci_v2_20to31.pdf)
3. Sehen Sie für die vorliegende oder ähnliche Studien **Lösungsmöglichkeiten?** Etwa durch andere Auswertungen, andere Schlussfolgerungen? Was lernt man für künftige Studien?
4. Ist das von Ihnen diskutierte **Problem auch bei anderen Studien mit digitalen Daten zu erwarten**? Diskussion an einem Beispiel.

Example: Anderson et al. 2013: Political Ideology and Racial Preferences in Online Dating

- Goal?
- Data?
- Stated Preferences & Revealed Preferences ?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Übungsaufgabe Nr. B4

Camila Pinto Moncada

Teilaufgabe 1: „Was sind für dieses Forschungsziel Vorteile der Nutzung dieser „big data“ gegenüber Alternativen wie Surveys? Erläutern Sie dies mit wenigen Sätzen.“

Big Data vs. (post-match) surveys

- Vier Vorteile von Big Data (in Bezug auf die Studie von Anderson et al., 2014):
  1. **Nicht reaktiv:** Die Teilnehmende agieren in einer natürlichen Setting und sind sich nicht bewusst, dass sie beobachtet werden, was die Verzerrung durch social desirability bias erheblich reduziert.
  2. **Wirklichkeitsnah:** Es spiegelt wider, was tatsächlich passiert, und nicht selbst gemeldete Informationen (in diesem “self-reported preferences“).
  3. **Always-on Datenerfassung:** Es ermöglicht, das Verhalten in einem frühen Stadium zu untersuchen und Konflikte zwischen der tatsächlichen präferenzbasierten Homophilie und neuen Präferenzen, die infolge der post-match Konvergenz angenommen wurden, zu vermeiden.
  4. **Kosten:** Die Kosten können erheblich gesenkt werden und die Probe habt eine größere Reichweite.
- **ABER** trotz dieser Vorteile gibt es auch beachtliche Limitationen.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Übungsaufgabe Nr. B4

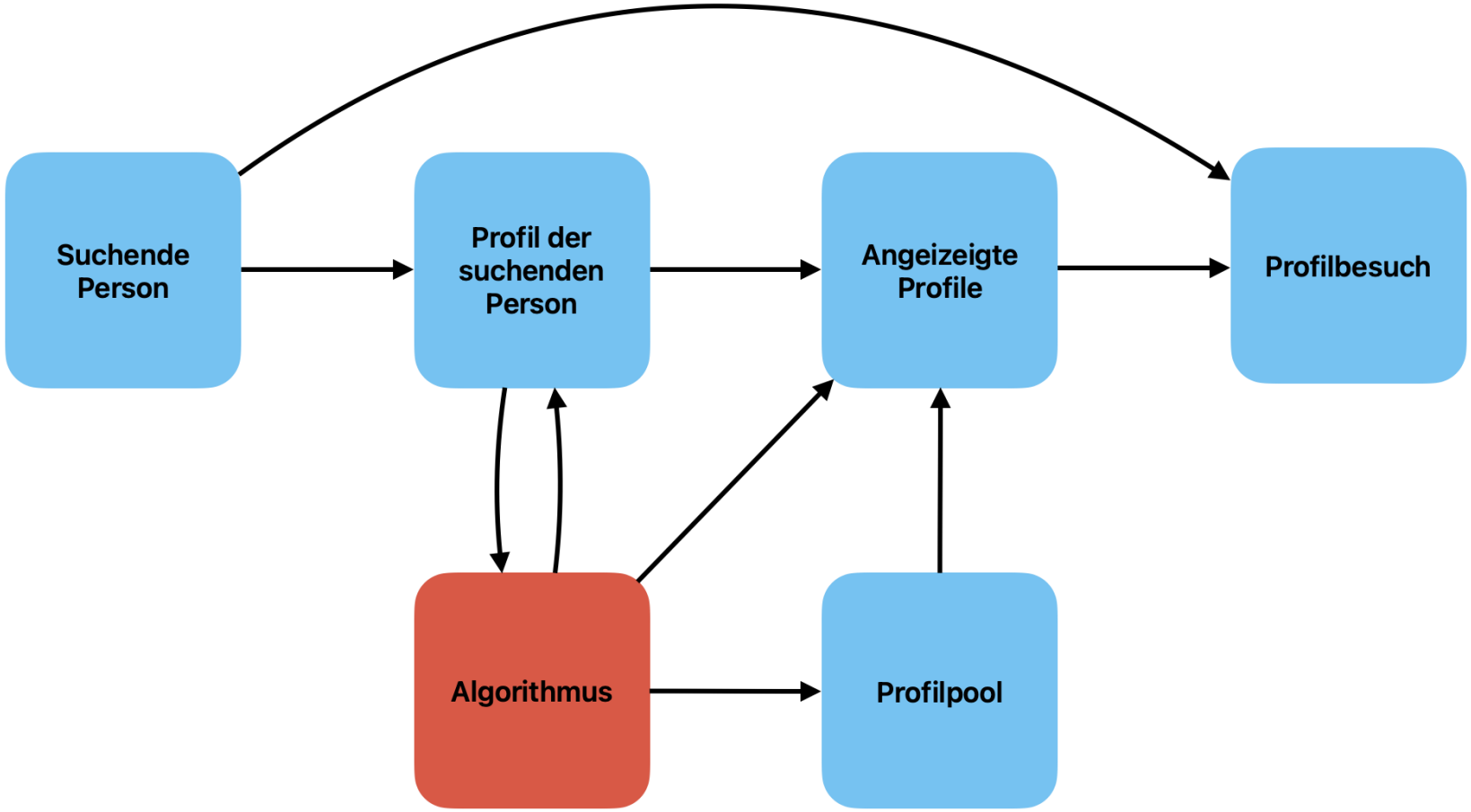
Paul Burlon

„2. Bedrohung der internen Validität“

Bedrohung der internen Validität

- Strategische Anreize sich besser darzustellen
  - Nicht zufällige Messfehler
- Algorithmic Confounding
  - Treatment wird durch den Algorithmus beeinflusst
  - Algorithmus ist unbekannt
- Performative Algorithmic Confounding
  - Self-fulfilling Prophecy
- Sample Selection
  - Kürzung des Samples aufgrund von arbiträren Maßstäben

DAG


Suchende Person
Profil der suchenden Person
Angeizeigte Profile
Profilbesuch
Algorithmus
Profilpool

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Übungsaufgabe Nr. B4

Natalia Velić

„Aufgabennummer 2: Bedrohung der externen Validität“

Zusammenfassung der Ergebnisse zur Teilaufgabe 2

- Studie von Anderson et al. (2014) enthält keine näheren Informationen bezüglich der Größe und Zusammensetzung der Nutzer der Online-Dating-Website

  → lediglich deskriptive Beschreibung der Stichprobenpopulation

  → essenzielle Unterschiede ggf. nicht erkennbar, was die Interpretation der erzielten Ergebnisse beeinflussen kann

- Sample besteht ausschließlich aus Nutzern, die:
  - Im Zeitraum von Oktober-November 2009 aktiv waren
  - Weitestgehend vollständige Profilangaben machten (u. a. Alter, Geschlecht, ethnische Zugehörigkeit, Partnerpräferenz)
  - Lediglich „White“ und „Black“, d. h. „Asian“ und „Hispanics“ blieben in Analyse unberücksichtigt

Zusammenfassung der Ergebnisse zur Teilaufgabe 2

- Websites durch sich ständig verändernde Online-Mitgliederzahl gekennzeichnet und Nutzer unterscheiden sich im Aktivitätenstatus

  → Zeitversetzte Profilaufrufe zweier Nutzer sollten nicht berücksichtigt werden

- Vielzahl an Gründen, weshalb „*[…] preferences might be narrower during this time (e.g., because users are particularly concerned about family approval) or more open (e.g., because users feel particularly lonely)*“ (Lewis 2015, S. 21).

- Folgen und Gefahr:

  → Stichprobenpopulation kann sich *systematisch* von Nutzern/Gesamtpopulation unterscheiden

  → Gefahr: ***Selection Bias***

Zusammenfassung der Ergebnisse zur Teilaufgabe 2

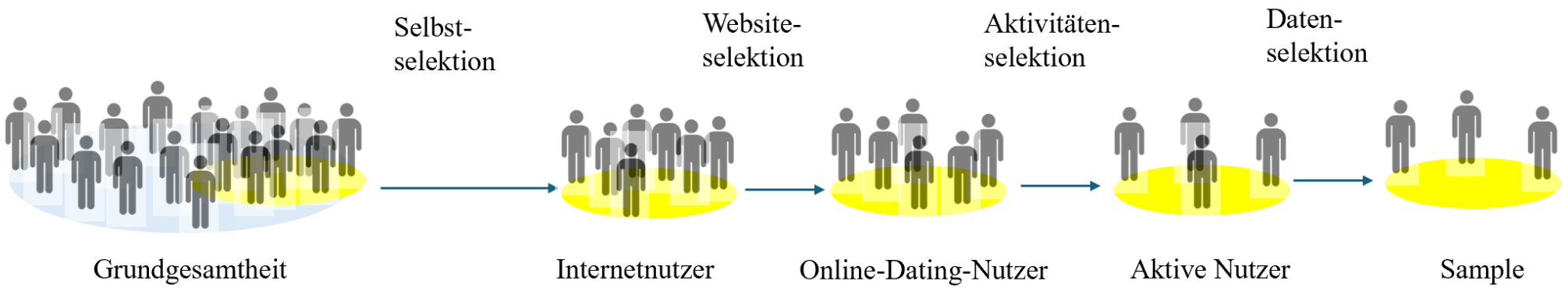


Abbildung 1: Stichprobenskizze. Eigene Darstellung.

10 Charakteristika von Big Data (Salganik 2018)

Vorteile

- Big
- Always-on
- Nichtreaktiv

Nachteile

- Incomplete
- Inaccessible
- Nonrepresentative
- Drifting
- Algorithmically Confounded
- Dirty
- Sensitive

10 Charakteristika von Big Data

+ Großer Datenumfang (big)

- Beobachtung seltener Fälle
- Präzisere Schätzer, Möglichkeit kleine Effekte zu entdecken
- Subgruppenanalysen, Effektheterogenität
  - Etwa: Hohes *N* erlaubt kleinräumigere Analysen

10 Charakteristika von Big Data

+ Stetig aktualisierend (always-on)

- Beobachtung unerwarteter Ereignisse möglich, etwa natürliche Experimente
- Beobachtungen in Echtzeit, damit „Nowcasting“ möglich (z.B. Grippe-Epidemien)
- Beobachtung dynamischer Prozesse
  - Herkömmliche Erhebungen (insb. Befragungen) erfassen nur bestimmte Zeitpunkte, etwa Ein-Jahresrhythmus; damit nur ungenaue Beobachtung von Veränderungen
  - Beispiel: Genauere Beobachtung von Segregationsprozessen, Bewegungsprofilen, Netzwerken: Wer hält sich wann mit wem wo auf?

10 Charakteristika von Big Data

+ Nichtreaktiv

- Zentrales Problem der Sozialforschung: Wissen um Teilnahme an Studien kann Verhalten ändern
- Verhaltensspuren haben dieses Problem nicht in dem Umfang
  - Jedoch sind Verzerrungen durch soziale Erwünschtheit immer noch möglich!
  - Beispiele?

- Allerdings ethische Probleme!

10 Charakteristika von Big Data

– Fehlende Variablen (incomplete)

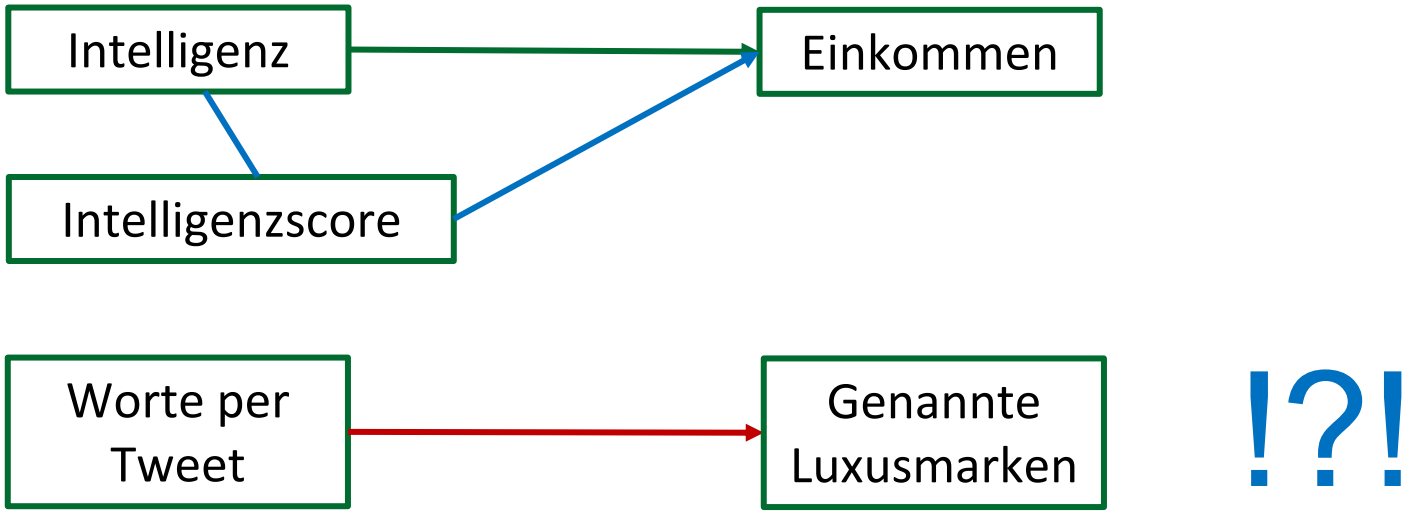
- Gerade bei ready-made Daten sind oft nicht alle interessierenden Variablen verfügbar
- Die verfügbaren Variablen messen oft nicht ganz die theoretisch interessierende Größe (geringe Konstruktvalidität + „long inferential leap“)
  - Z.B.: Facebook-“Freunde“ sind nicht unbedingt Freunde
  - Hilfreiche Methode: ersetzen sie im Kernresultat das theoretische Konstrukt mit der gemessenen Variable – erscheint die Interpretation noch immer schlüssig?
    - Konstrukt: Intelligentere Personen bekommen höhere Einkommen.
    - Gemessen: Personen, die längere Wörter auf Twitter benutzen erwähnen häufiger Luxusmarken

➢Validierungen mit anderen Datenquellen sind wichtig!

10 Charakteristika von Big Data

Beispiel (Salganik 2017S. 24f.)

Forschungsfrage: Haben intelligentere Menschen ein höheres Einkommen?
(wie es etwa die Humankapitaltheorie vorhersagt)



10 Charakteristika von Big Data

– Oft kein Datenzugang (Inaccessible)

- Fehlender Datenzugang kann unterschiedliche Gründe haben
  (z.B. Datenschutz, kommerzielle Interessen)
- Daten werden wenn, dann oftmals nur selektiv frei gegeben
  (z.B. nur „positive“ Daten zu Unternehmen)
- Oft eingeschränkte Nachnutzung, damit geringe Replizierbarkeit und Nachprüfbarkeit von Analysen (Reanalysen, erweiterte Replikationen)

10 Charakteristika von Big Data

– Nicht-“repräsentativ“

- Oft nur selektive Daten, z.B.
  - Nur selektive Auswahl von Freundschaften auf Facebook
  - Online-Immobilienanzeigen bilden nicht auf den gesamten Immobilienmarkt ab
- Aber Vergleiche innerhalb des Samples können durchaus aufschlussreich für Mechanismen (Kausalanalysen) sein. Beispiele?
- Für Extrapolationen über Sample hinaus sind Theorie + Empirie erforderlich: Gibt es Moderatoren? (s. Kapitel zu externer Validität)

10 Charakteristika von Big Data

– Messung nicht zeitkonstant (drifting)

- If you want to measure change, don‘t change the measurement“ (Fischer 2011)
- Erhebungen digitaler Daten (wie z.B. social media Daten) sind oft besonders stark von Veränderungen der Messung betroffen
  - Veränderung der Komposition der Stichprobe (**population drift**)
  - Veränderung im Nutzungsverhalten (**behavioural drift**)
  - Änderungen in den (technischen) Rahmenbedingungen für Verhaltensspuren (**system drift**)

➢Welche Beispiele fallen Ihnen hierzu ein?

10 Charakteristika von Big Data

– Messung abhängig von Plattform bzw. Datenprovider (algorythmic confounding)

- (Soziologische) Theorie beeinflusst oftmals die Daten (performativity)
- U.a. dadurch drohen „Artefakte“: durch Algorithmen verursachte Phänomene
- Beispiel?
  - Facebook schlägt Freunde von Freunden vor (Transitivität) - das Feststellen von Transitivität in den Daten ist daher artifiziell, und kein Beleg für die Thesen Granovetters

➢ Zur Abhilfe ist gutes Wissen über den Datengenerierungs-Prozess wichtig!

- Wer ist erfasst, wer fehlt
- Inwieweit ist Verhalten „natürlich“

10 Charakteristika von Big Data

– Verzerrt (dirty)

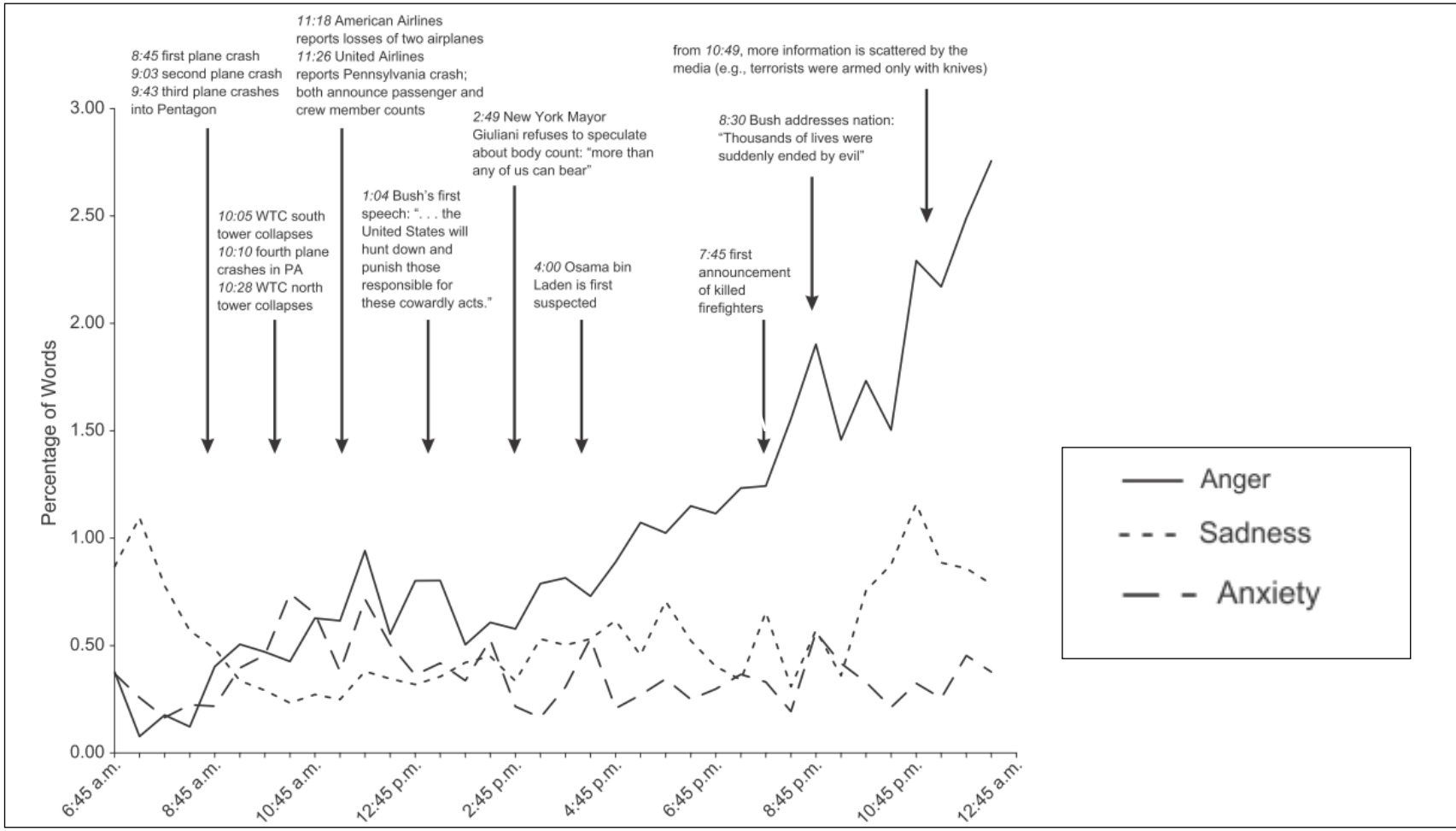
- Daten können Elemente enthalten, die nicht zur Population gehören (Over-Coverage)
- Diese können den interessierenden Effekt verzerren
- Beispiele: spam, bots, hoaxes
- Zudem andere mögliche systematischer Messfehler

- Wiederum: Wissen um Datengenerierungsprozess ist wichtig!

10 Charakteristika von Big Data

Beispiel: „The Emotional Timeline of September 11“

- Kodierung von Textmessages (Pager) Daten in 3 emotionale Kategorien (Anxiety, Sadness, Anger)

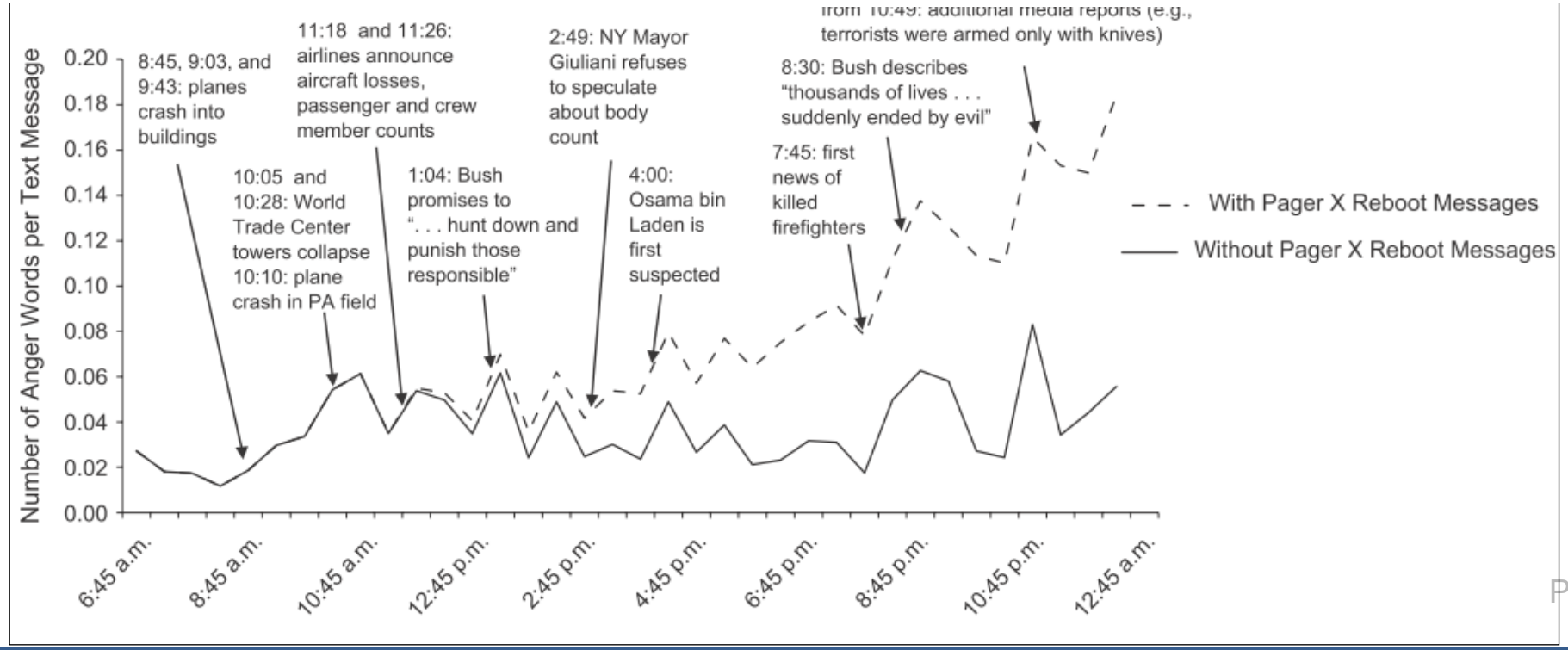


Back, Küfner and Egloff 2010: 1418

10 Charakteristika von Big Data

Problem: Zunahme von „Anger“ ist reines Artefakt:

- Ein Pager hat ein technisches Problem und sendet ständig „Reboot NT machine [name] in cabinet [name] at [location]:CRITICAL:[date and time]. “
- „Critical“ wurde als „Anger“ kodiert
- Ausschluss dieses Pagers reduziert *N* von Kategorie „Anger“ deutlich



Pury 2011: 836

10 Charakteristika von Big Data

– Sensible Daten (sensitive)

- Es gibt etliche ethische Probleme
  - U.a. Verletzungen der Privatheit, keine informierte Einwilligung
- Forscher/innen können Daten anonymisieren
  - Allerdings ggf. re-anonymisierbar
- Es fehlen oft noch Standards und/oder praktische Richtlinien!

➢Abhilfe: Orientierung an Ethik-Richtlinien; etwa der Akademie für Soziologie

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

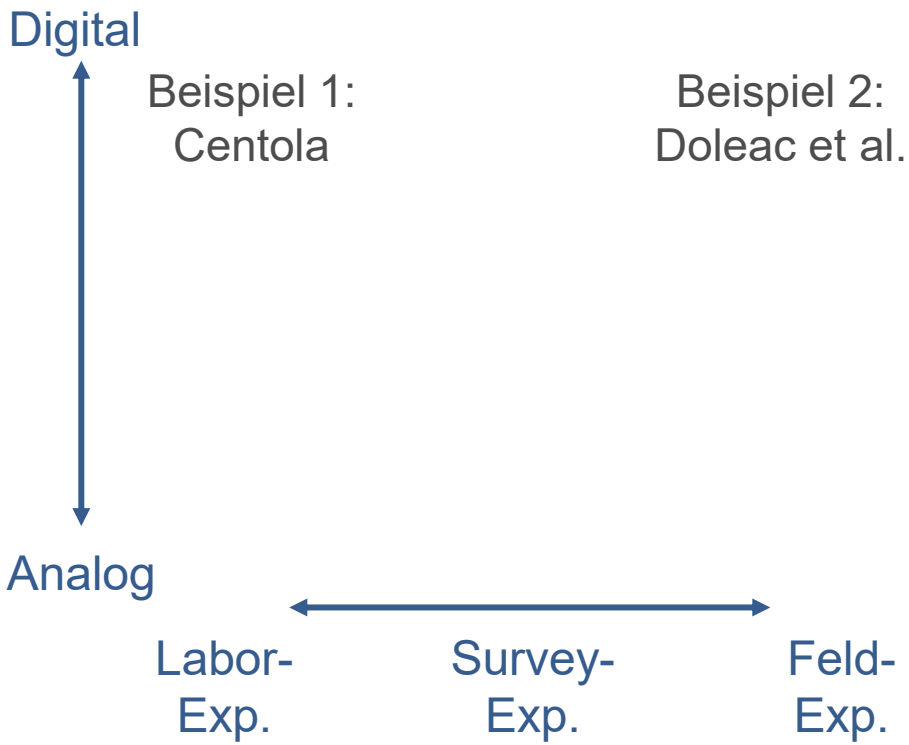
Digitale Feldexperimente

Von Lab zu Field – unterschiedlich viel „digital“

„Digital“ = Nutzung digitaler Infrastruktur

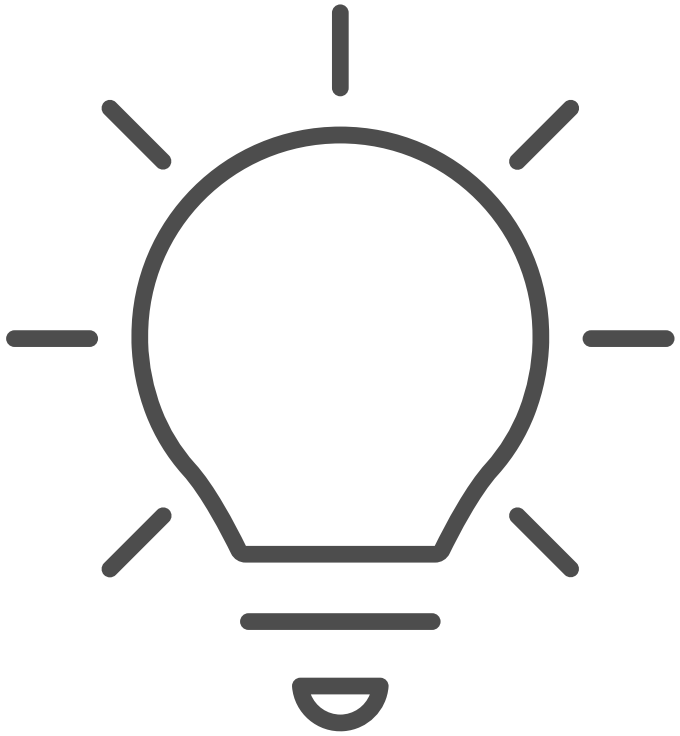
- Rekrutierung
- Randomisierung
- Treatmentsetzung/Outcome-Messung
- „Big-Data“
- „Wearables“
- Digitale Messgeräte (Strom, Verkehr etc.)

„In other words, **digital experiments are not just online experiments“** (Salganik 2018: 155)



Eigene Darstellung nach Salganik 2018, S. 152

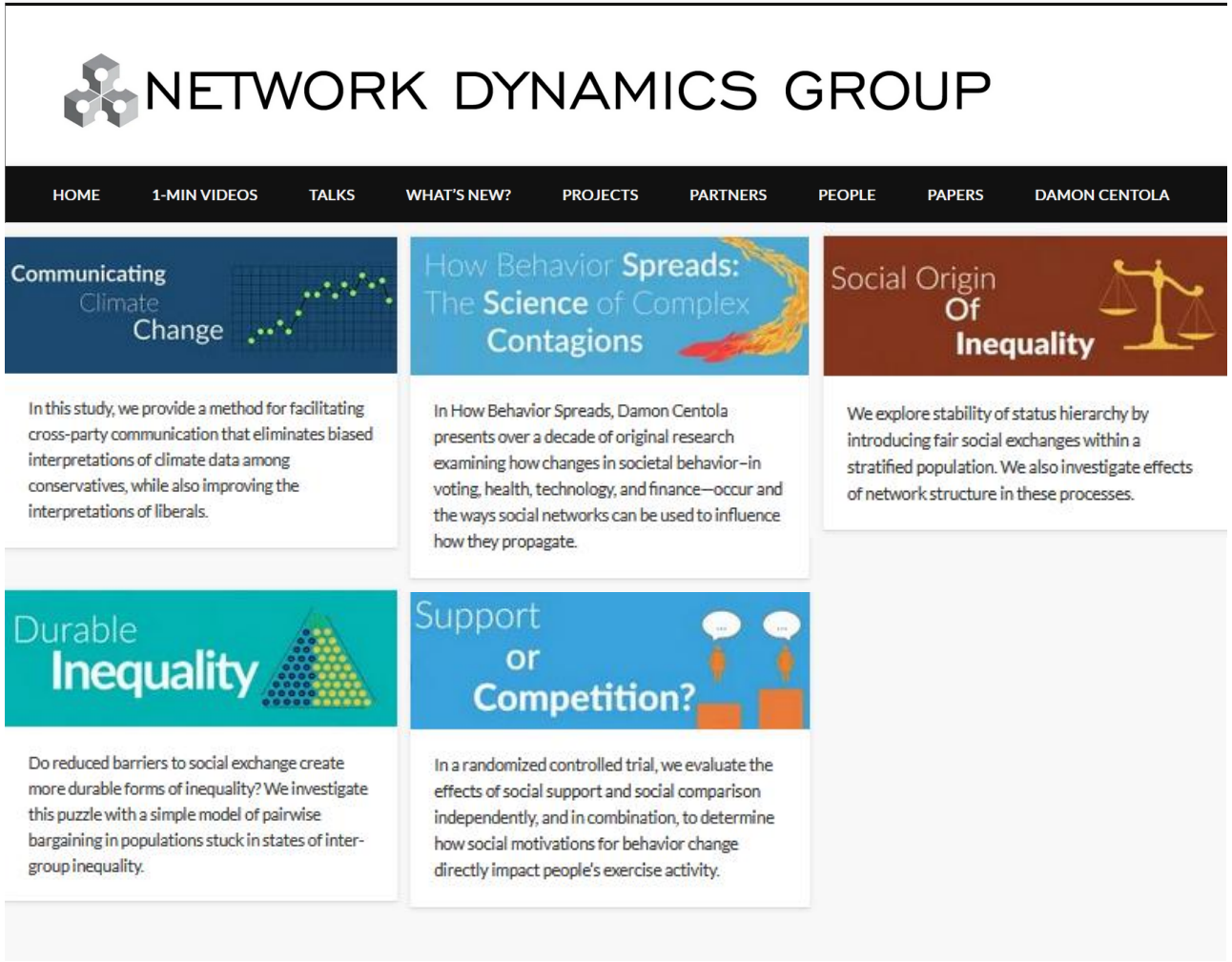
Vorteile digitaler Designs



- Grundsätzlich: „Klassische“ Experimentallogik
- Plus: Skalierbarkeit („Massenexperimente“)
  - + *Quantitativ*: Höhere statistische Power
  - + *Qualitativ:* Messung von Effektheterogenität
  - = Versprechen: Mehr Evidenz dafür, welcher Mechanismus Ergebnisse (nicht) erzeugt
- Innovationspotenzial v.a. durch digitale Feldexperimente
  - „combine control and realism at scale“
- Außerdem: z.T. Pre-Treatment Information

Beispiel 1: Digitales Laborexperiment

- Experimente von Damon Centola
- Beispiel: Can social influence reduce bias in the interpretation of scientific information?



Beispiel 2: Digitales Feldexperiment

- Verhalten (+) ohne soziale Erwünschtheit (++)
- Mehrfaktorielles Design (3x3x2):
  - Hautfarbe/Tattoos
  - Preis
  - Textqualität
- Testung auf Märkten mit unterschiedlicher Konkurrenz, Segregation, Kriminalität, etc.
- Dadurch: Aufschluss über Mechanismen, Bedingungen, Stärke von Diskriminierung
  - Etwa: Zahlungsbereitschaft/Verluste

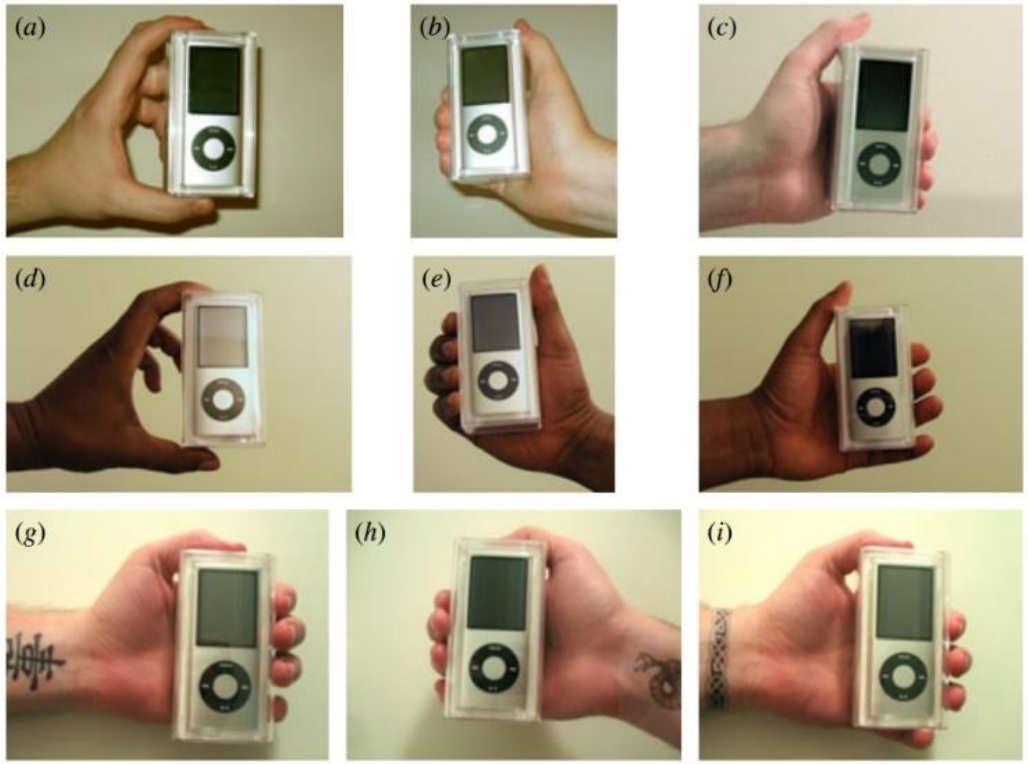
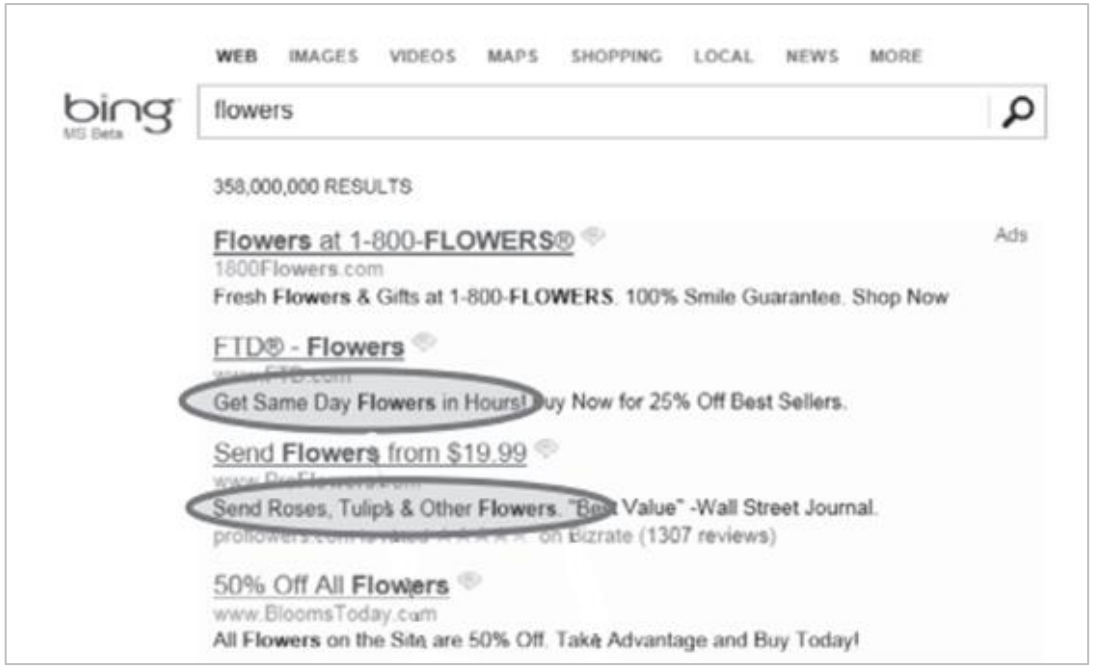
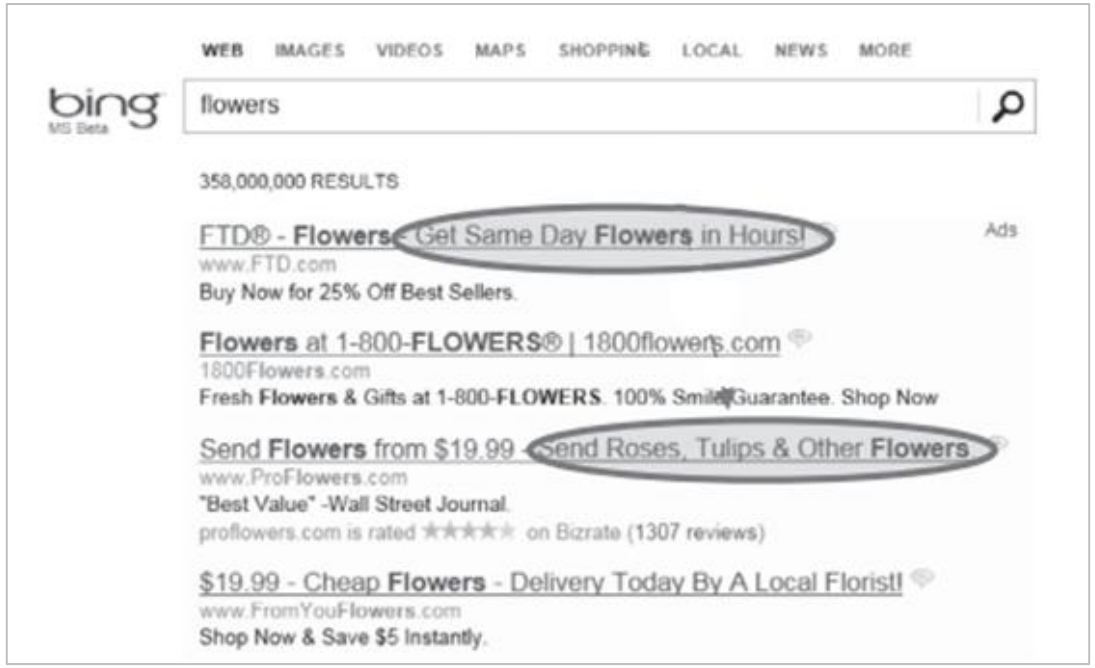


Figure 4.13: Hands used in the experiment of Doleac and Stein (2013). iPods were sold by sellers with different characteristics to measure discrimination in an online marketplace. Reproduced courtesy of John Wiley and Sons from Doleac and Stein (2013), figure 1.

Weitere (kommerzielle) Beispiele



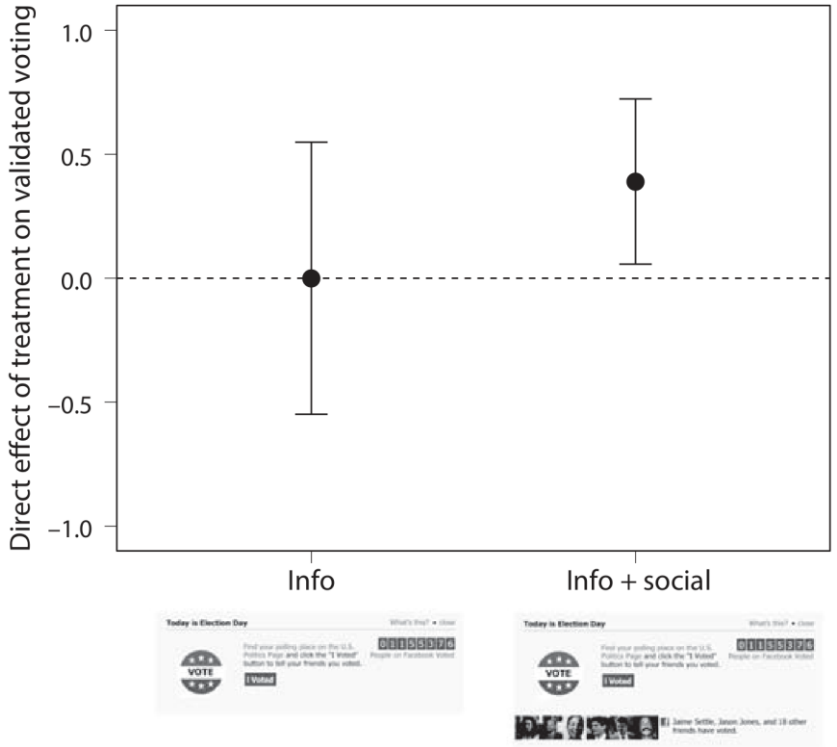


"Bing's revenue increased by a whopping 12%, which at the time translated to over $100M annually in the US alone" (Kohavi et al. 2020: 3f)

Aber: Wissenschaftlicher Erkenntnisgewinn? Ethik? Datenschutz?

Weitere (kommerzielle) Beispiele

„Today Microsoft and several other leading companies – including Amazon, Booking.com, Facebook, and Google – each conduct **more than ten thousand online controlled experiments annually**, which individually engage millions of users.“ (Thomke 2020: 83)



Do you want more?

Yes please.

No only one experiment.

Salganik 2018: 187

The $2.1 Billion McDonald's Machine: https://www.youtube.com/watch?v=BKX6EhDrgqQ

Übungsaufgabe

- Suchen Sie sich ein Feld-Experiment aus, das im Internet durchgeführt wurde oder ein anderes Experiment, das prozessproduzierte, digitale Beobachtungsdaten nutzt.

1. Was ist die **Fragestellung?** Wie wird die **Identifikation des Treatmenteffekts** sichergestellt? Zeichnen Sie dazu auch einen **DAG**.
2. Könnte man den Treatmenteffekt/Mechanismus **ähnlich auch ohne digitale Daten untersuchen?** Beschreiben Sie kurz ein alternatives offline-Design.
3. Was sind besondere **Vorteile, das Experiment online durchzuführen?**
4. **Bedrohungen der Validität**, wenn das Experiment online durchgeführt wird? Können die Autoren in dem von Ihnen betrachteten Beispiel diese Bedrohungen überzeugend ausräumen? Mindestens ein Kritikpunkt, der nicht angeführt oder überzeugend ausgeräumt wird.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

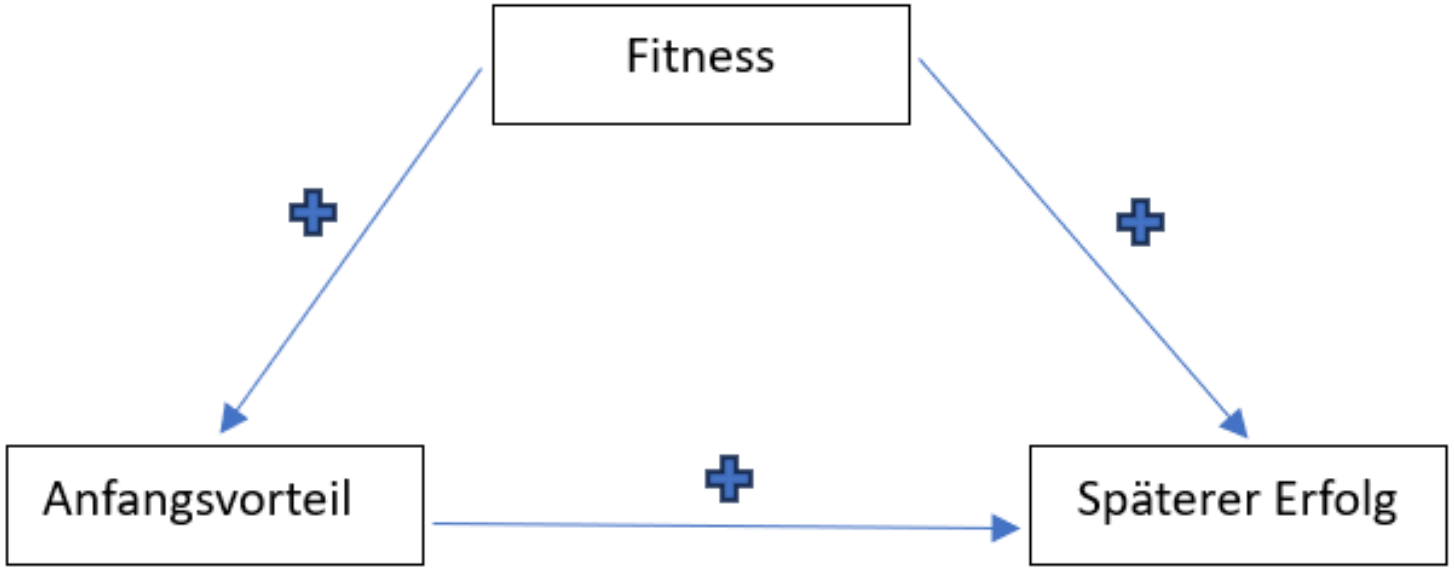
Übungsaufgabe Nr. 06

Laura Zehner

Aufgabe 1:

Was ist die Fragestellung? Wie wird die Identifikation des Treatmenteffekts sichergestellt? Zeichnen Sie dazu auch einen DAG.

- divergierende Erfolgsverläufen zwischen ähnlichen Individuen
- Forschungsfrage: Lösen kleine zufällige Anfangsvorteile durch positive Rückmeldungen selbstverstärkende Prozesse aus, die sich zu solchen starken Differenzen im Erfolg kumulieren (‚success breeds success'–Hypothese)
- Inwieweit beeinflusst die Größe von frühem Erfolg die Stärke der divergierenden Erfolgsentwicklungen?
- vergleichbare Ausgangsbedingungen
- mögliche Konfundierung durch unbeobachtete Dimensionen von ‚Fitness'
- durch randomisierte, experimentelle Zuweisung des Anfangsvorteils ist unbeobachtete ‚Fitness' nicht systematisch mit dem Treatment korreliert und der beobachtete Unterschied im späteren Erfolg kann kausal interpretiert werden


Fitness
Anfangsvorteil
Späterer Erfolg

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Übungsaufgabe Nr. B5

Lea Kreppold

„Teilaufgabe 4 - Validität“

Interne Validität: Identifikation & verbleibende Unsicherheiten

- **Zentrale Bedrohung:** Unbeobachtete Heterogenität (z.B. Qualität, Motivation, soziale Netzwerke) *(van de Rijt et al. 2014: 6935)*
- **Design-Lösung**: Randomisierte Vergabe früher Erfolgssignale → kein systematisches Confounding zwischen Treatment und Outcome
- **Identifikationslogik**: Randomisierung schließt Backdor-Pfade & kausale Schätzung einer ATE
- **Weitere adressierte Alternative**: Früher Erfolg als bloßes Qualitätssignal → geschwächt durch künstliche, zufällige Setzung des Erfolgssignals *(ebd.: 6935f.)*

→ Hohe interne Validität hinsichtlich des kausalen Effekts früher Erfolge

Validitätskritik und externe Validität

- **Nicht vollständig ausgeräumt** → mögliche indirekte algorithmische Effekte (Plattformen strukturieren Sichtbarkeit nicht neutral)
- **Zentrale offene Frage:** Ist der Effekt somit rein sozial vermittelt oder teilweise durch plattformspezifische Mechanismen verstärkt?
- **Externe Validität – Stärke**: Replikation identischer Interventionen (vier Plattformen mit unterschiedlichen Funktionslogiken)
- **Externe Validität – Einschränkung**: Generalisierbarkeit primär auf digitale Plattformkontexte <-> begrenzte Übertragbarkeit auf nicht-digitale soziale Felder

→ **Gesamtbewertung**: interne Validität überzeugend, externe Validität kontextgebunden

Field experiments of sucess-breeds-sucess dynamics

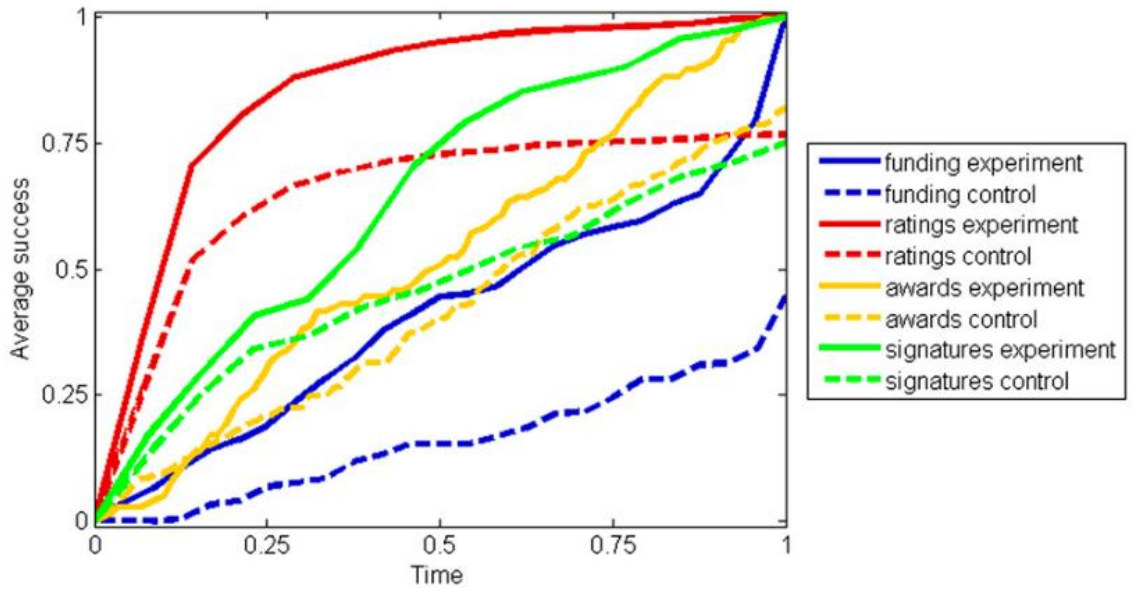


Fig. 2. The success-breeds-success effect over time. The curves represent running numbers of donations (blue), positive ratings (red), awards (yellow), and campaign signatures (green) in the experimental condition (solid lines) and the control condition (dashed lines). The horizontal axis is normalized so that 0 marks the time of experimental intervention, and 1 marks the end of the observation period. The vertical axis is normalized so that for each system a value of 1 equals the maximum across time and conditions.

Van de Rijt et al. (2014)

Vor- und Nachteile von Online-Experimenten

Vorteile

- Wenig Fehler
  - Computer misst automatisch („always on“)
- Geringe Kosten (→ large *N*)
  - Viele Manipulationen (ggf. mehrfaktoriell)
  - Effektheterogenität studierbar
- „Pre-Treatment Information“
  - Oft prozessproduzierte Daten
  - Damit Effektmoderatoren studierbar
- Längere Beobachtungszeiträume

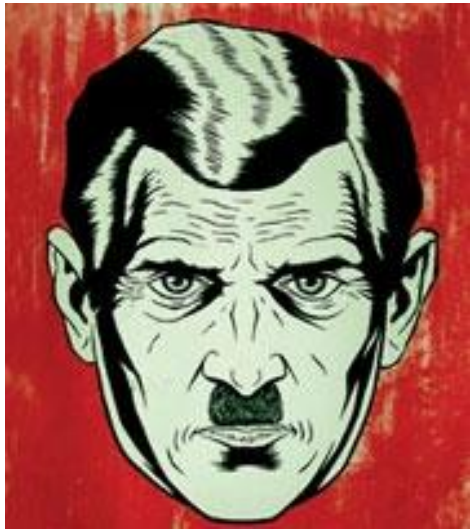
Nachteile

- Wenig Kontrolle (v.a. auf Plattformen)
  - Messung, Sample, Daten, Publikation
- **Algorithmic Confounding:**
  „Effect driven by system-specific dynamics“
  - Tweet-Längen, Rangordnungen, etc. z.T. endogen von Outcomes abhängig
  - Schwer bis gar nicht beobachtbar
- Eigene Plattform: Aufwand/Künstlichkeit
- Ethik: Informed Consent? Ethikvotum?

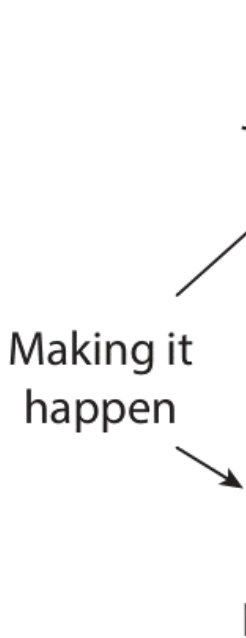
Auf einen Blick – mit Empfehlungen für Praxis

- Mehr interne Validität: „jein“
  - Mehr Kontrolle über Randomisierung
  - Weniger über Messung theoretischer Konstrukte
- Mehr externe Validität: i.d.R. „ja“
  - Mehr Treatments und Settings
  - Nicht nur „ob“, sondern auch „warum“/“wie“
    - Scope-Conditions, Replizierbarkeit
  - Damit: Fortentwicklung Theorien und effektiverer Interventionen
- Aber: Big Brother…

- Praktische Empfehlungen:
  - Kostenminimierung („enjoyable“ Experimente mit automatischer Messung)
  - Plattform als Partner suchen
  - Ethik abwägen, Ethikvotum einholen (!)



„Making it happen“ – eine Typologie

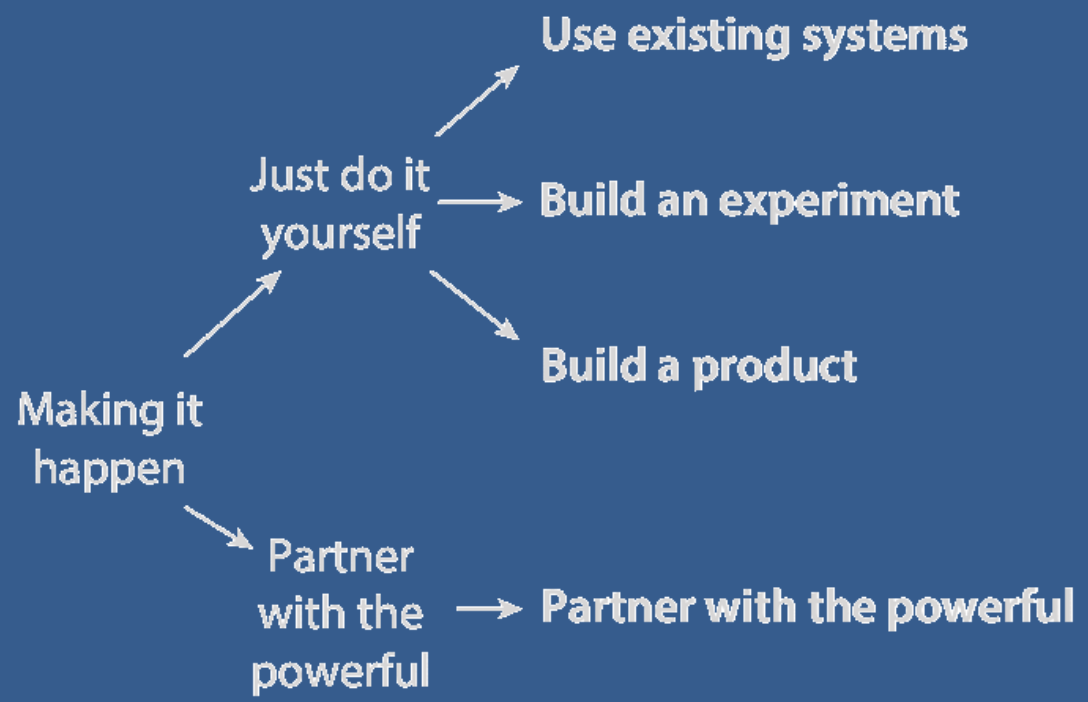


	Cost	Control	Realism	Ethics
Use existing systems	Low	Low	High	Potentially complex
Build an experiment	Medium	High	Medium	Relatively easy
Build a product	High	High	High	Relatively easy
Partner with the powerful	Low	Medium	High	Potentially complex

Salganik 2018: 174

Diskussionsfrage

Sie möchten Diskriminierung (z.B. nach Ethnie) auf Online-Partnermärkten untersuchen.



→ Welche Implementierung wählen Sie? Warum?

→ Welche konkreten Nachteile bzw. Probleme ihrer Implementierung fallen Ihnen ein?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Räumliche und georeferenzierte Daten

Räumliche bzw. Geodaten

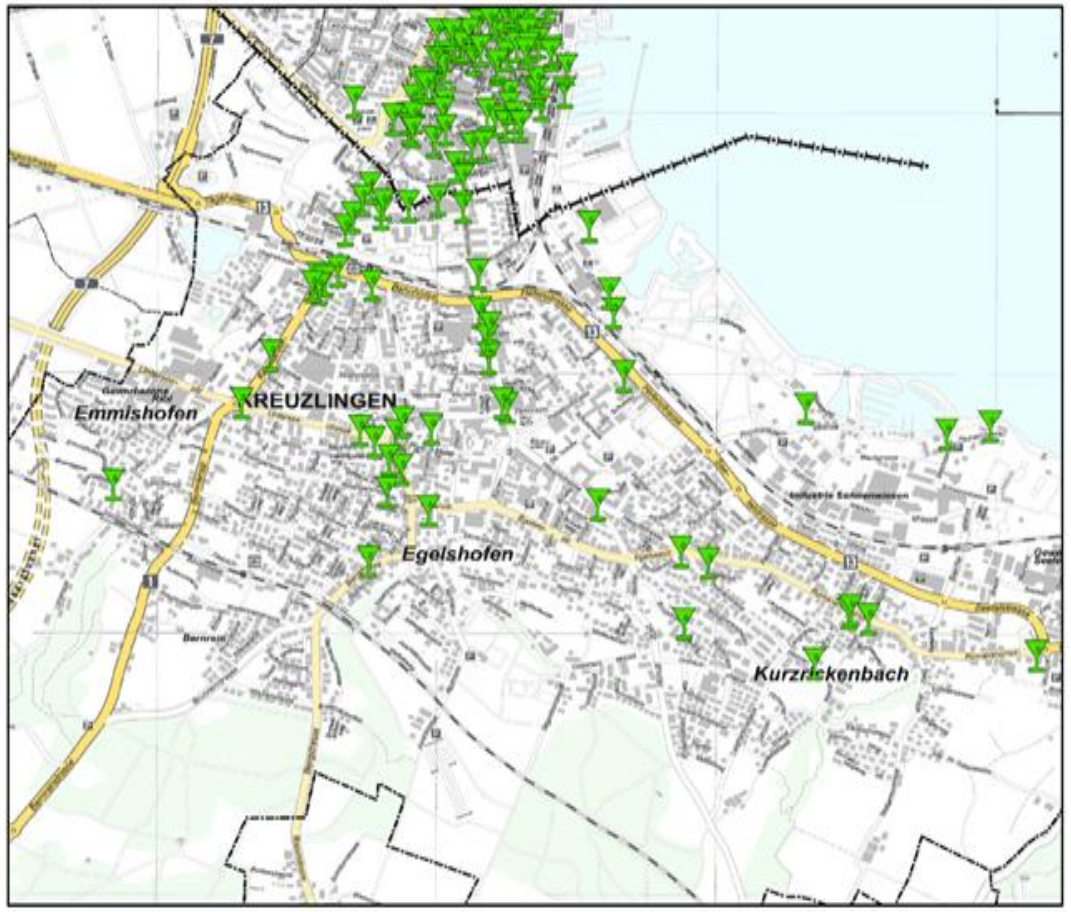
- Räumliche Daten sind durch Geokoordinaten eindeutig im Raum verortbar
- Damit sind relationale Datenbanken erstellbar
  - Information über den Ort: Koordinaten
  - Verlinkung mit an diesem Ort gegebenen Merkmalen

Latitude	Longitude	Ort	Strasse	Name	Sauberkeit	Bier
47.652502	9.166325	Kreuzlingen	Hauptstr. 34	American Blue Bar	3	5.5

- Beispiel: https://www.mapdevelopers.com/geocode_tool.php
- Anwendung der üblichen Gütekriterien
  - Objektivität, Reliabilität, Validität

Arten von Geodaten: Punktdaten

- Beispiel: Koordinaten von Kneipen
- **Use case?**
  - **Drinking Alone: Local Socio-Cultural Degradation and Radical Right Support**
  - I show that individuals living in districts that experience one additional community pub closure (relative to the total number of pubs per district) are more likely to support UKIP than any other party by 4.3 percentage points.



(Bolet 2021: Drinking Alone: Local Socio-Cultural Degradation and Radical Right Support—The Case of British Pub Closures)

Arten von Geodaten: Vektoren/Polygone

- Beispiel: Nachbarschaften, Straßennetz

Arten von Geodaten: Rasterdaten

- Beispiel: Geräuschlevels in db

Weitere Begrifflichkeiten

- Geodaten
  - Daten beziehen sich auf eine bestimmte Position im Raum (auf der Erde), z.B. Längen- und Breitengrade
- GIS (Geo-Informations-System)
  - Werkzeuge um mit Geodaten zu arbeiten, etwa sie zu analysieren
- Geocoding
  - Prozess des Matchings von Einheiten mit Geokoordinaten
  - Werkzeuge: ArcGIS, Google-Maps, Ados in Stata, etc.
- Georeferenzierte Daten
  - Analyseeinheiten, die mit Geodaten verbunden wurden
  - Einheiten sind z.B. Befragte in Surveys, Organisationen, Länder

Interessierende Variablen

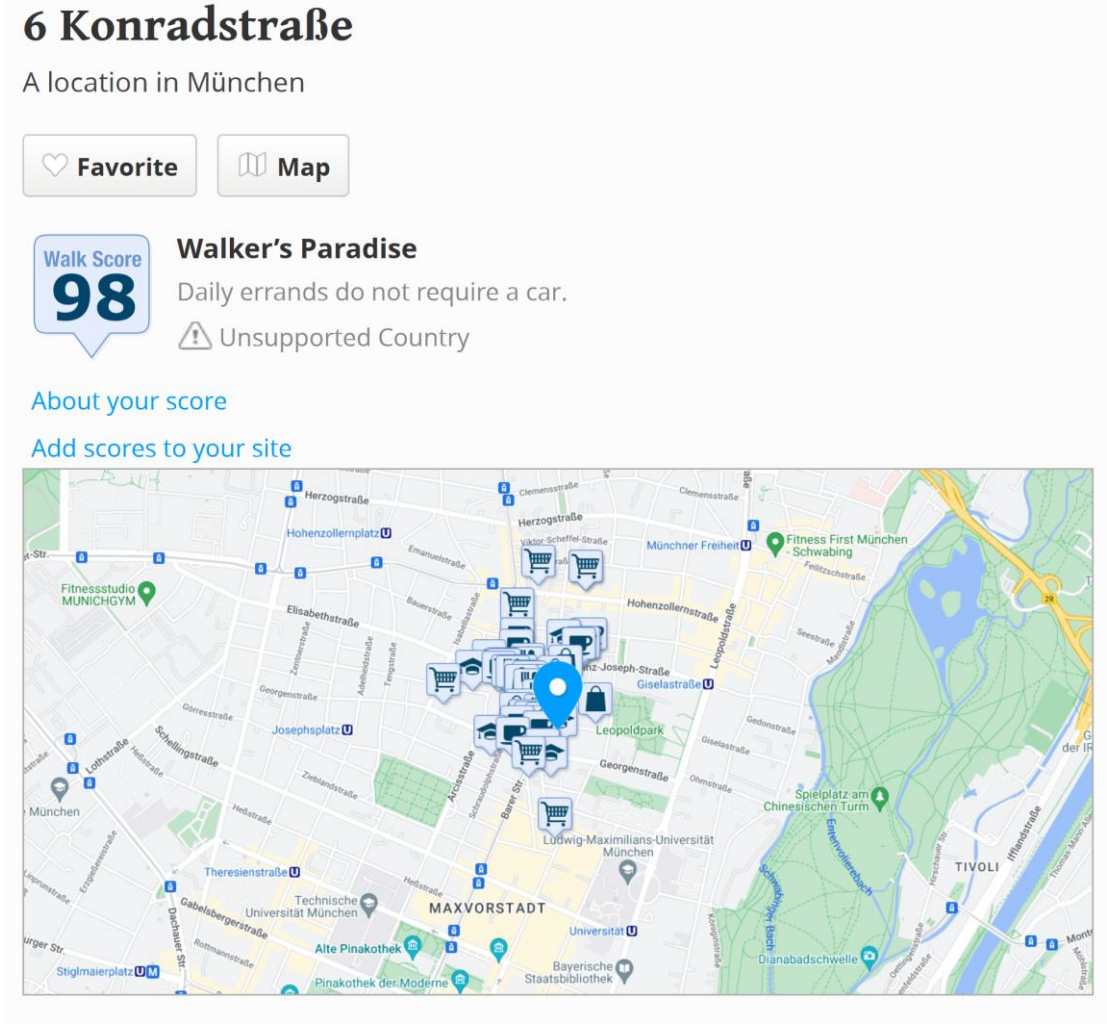
- Merkmale räumlicher Einheiten
  - Z.B. Ressourcen, Ausländeranteile
  - Untersch. Definitionsmöglichkeiten
    - Z.B. Berücksichtigung von Straßen-/Bahntrassen
    - „Nearest neighbors“ / Ellipsen um Punktdaten
- Entfernung zu anderen Einheiten: Luftlinie oder Wegstrecken
  - Z.B. Pendelstrecken, Entfernung zum Stadtzentrum
- Kombinationen von Dimensionen: Indizes, z.B. Walkability

What makes a neighborhood walkable?

- **A center:** Walkable neighborhoods have a center, whether it's a main street or a public space.
- **People:** Enough people for businesses to flourish and for public transit to run frequently.
- **Mixed income, mixed use:** Affordable housing located near businesses.
- **Parks and public space:** Plenty of public places to gather and play.
- **Pedestrian design:** Buildings are close to the street, parking lots are relegated to the back.
- **Schools and workplaces:** Close enough that most residents can walk from their homes.
- **Complete streets:** Streets designed for bicyclists, pedestrians, and transit.

Beispiel „Walk Scores“

- www.walkscore.com
  (bislang allerdings für Deutschland nur bedingt empfehlenswert)



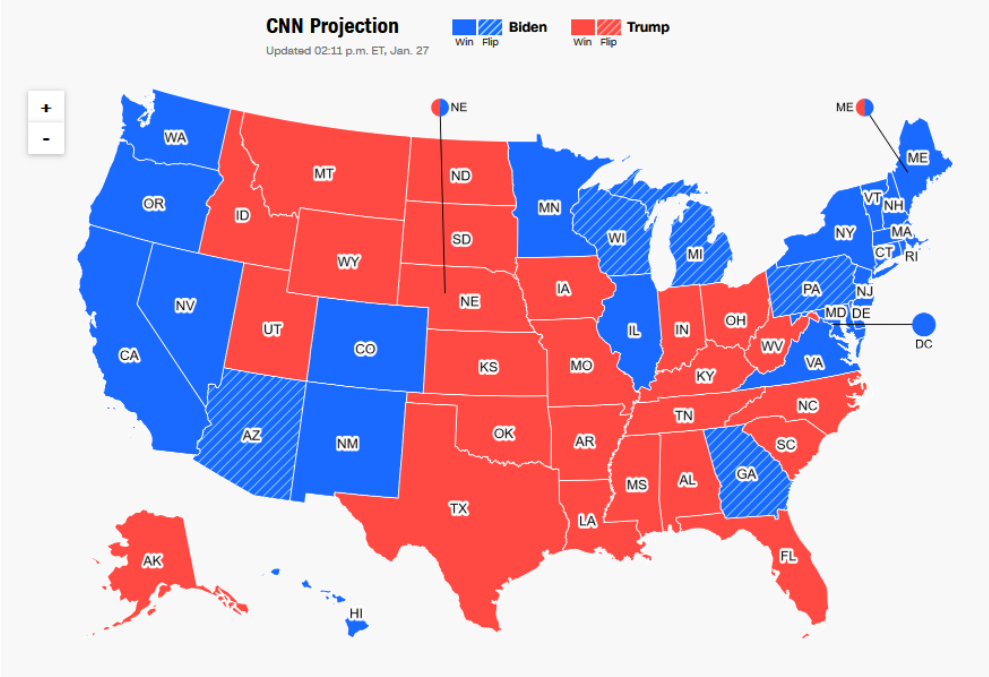
Wozu räumliche Daten?

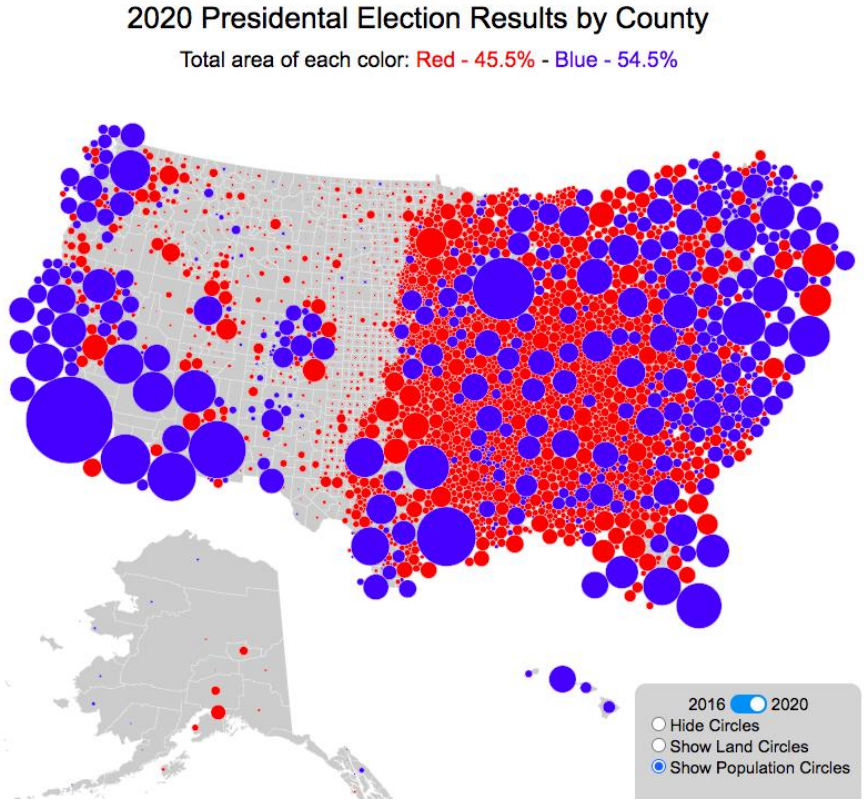
- Illustration und Exploration
- Analyse räumlicher Zusammenhänge; auch unabhängig von (willkürlichen) Gebietseinheiten!
  - (Nachbarschafts-)Effekte: Macht das Wohnumfeld einen Unterschied?
    - Begegnungsmöglichkeiten
    - Einfluss Entfernungen und Wegstrecken
    - Verfügbare Ressourcen: Jobs, Kinderbetreuung etc.
    - *Exposure* mit z.B. Umweltbelastungen oder Umweltgütern
  - Welches Umfeld ist (für was) bedeutsam?
    - Z.B. Nahumfeld versus größere Pendelregionen
  - Neue Möglichkeiten bei kausaler Inferenz
    - Z.B. Distanzen als Instrument
- Geodaten sind oft zuverlässiger als subjektive Einschätzungen durch Befragte oder Interviewende

Beispiel: The Ghost Map (John Snow 1855)

Aber Vorsicht bei visuellen Darstellungen!

- „All Maps of Parameter Estimates are Misleading“ (Gelman/Price 1999)





CNN
https://engaging-data.com/county-electoral-map-land-vs-population

Übungsaufgabe

Zur Prüfung, ob ethnische Diversität in Nachbarschaften ethnische Diskriminierung beeinflusst, werden zuweilen Daten von E-Mail-Korrespondenztest mit Kontextmerkmalen angereichert.

1. **Kann man davon ausgehen, dass die Kontaktthese stimmt, wenn** man **ein geringeres Ausmaß an Diskriminierung in Landkreisen** findet, **in denen der Ausländeranteil hoch ist?**
2. **Probleme** davon, **regionale Einheiten wie Landkreise oder Stadtviertel für die Messung der Kontextvariable** (ethnische Diversität) **zu verwenden?** Was wäre ggf. ein besseres Vorgehen?
3. Fallen Ihnen **weitere Bedrohungen der internen oder externen Validität** ein? Diskutieren Sie mindestens eine weitere mögliche Bedrohung.
4. Könnte man hier **ggf. auch von einem natürlichen Experiment profitieren**, um den Effekt von ethnischer Diversität oder vermehrten Konflikten/Konkurrenz zu prüfen?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Übungsaufgabe Nr. **6**

Gleb Belous

„Aufgabennummer 1 /Gilt die Kontaktthese in dem Forschungsdesign? “

Forschungsfrage & Ansatz

- Fragestellung: Zusammenhang zwischen ethnischer Diversität und Diskriminierung
- E-Mail-Korrespondenztests: Diskriminierung durch Wohnungsanbieter/Arbeitgeber
- Kontextvariable: amtlich gemessener Ausländeranteil im Landkreis
- Beobachtung: Unterschiede in Diskriminierung je nach Ausländeranteil

Bewertung / Validität

- **Ergebnis:** Keine eindeutige Verbindung zwischen Ausländeranteil und Diskriminierung messbar
- **Mechanismen unklar:**
  - Kontakt → hypothetisch weniger Diskriminierung
  - Konkurrenz → hypothetisch mehr Diskriminierung
- **Limitationen:** Kausalität unklar, Ausländeranteil ≠ tatsächlicher Kontakt

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Übungsaufgabe Nr. 1

Trang Nguyen

„Kann man davon ausgehen, dass die Kontaktthese stimmt, wenn man ein geringeres Ausmaß an Diskriminierung durch Wohnungsanbieter oder Arbeitgeber in Landkreisen findet, in denen der Ausländeranteil hoch ist? Was wären ggf. Einwände gegen die Validität dieser Schlussfolgerung?“

Kontaktthese - Probleme

Kontaktthese:
Häufiger Kontakt zwischen Ingroup und Outgroup baut Vorurteile und Stereotype ab
→ Zusammenhang zwischen ethnischer Diversität und ethnischer Diskriminierung erklären

Selbstselektion:

- Menschen mit vielen Vorurteilen ziehen eher in Gegenden mit geringem Ausländeranteil
- Menschen mit weniger Vorurteilen ziehen eher in Gegenden mit hohem Ausländeranteil

→ Weniger Diskriminierung könnte an den Menschen liegen, nicht am Kontakt

Kontaktthese - Probleme

Stadt-Land Unterschied

In Städten:

- höherer Ausländeranteil
- oft höheres Bildungsniveau
- oft liberalere Einstellungen

→ Weniger Diskriminierung könnte an Bildung oder Urbanität liegen, nicht am Kontakt mit Ausländer:innen

Ausländeranteil = Kontakt ?

- Landkreise sind oft sehr heterogen
- Menschen können trotz hoher Diversität im Stadtviertel, in ethnisch homogenen Nachbarschaften leben

→ Hoher Ausländeranteil bedeutet nicht automatisch viel persönlicher Kontakt

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Übungsaufgabe Nr. B6

Martin Kurzenberger

„Räumliche und georeferenzierte Daten“

Kontaktthese & Probleme regionaler Messgrößen

1. Gültigkeit der Kontaktthese

- Geodaten allein → keine direkten Aussagen über tatsächliche soziale Kontakte
- Diskriminierung kann unabhängig vom Kontext auftreten (z. B. „Hanna & Ismail" → gleiche Bewerbungen, trotzdem deutliche Benachteiligung)
- Regionale Unterschiede häufig durch Drittvariablen erklärbar:
  - Urbanität / räumliche Struktur
  - Marktbedingungen (Wohnungsmarkt, Arbeitgeber)
  - Bevölkerungszusammensetzung
- Geodaten bilden Netzwerke & Interaktionen nur unvollständig ab
- **Fazit:** Kontaktthese kann mit reinen Geo-Daten nicht bestätigt werden

2. Probleme regionaler Einheiten (Landkreise, Stadtviertel)

- Administrative Grenzen ≠ tatsächliche Lebenswelten.
- Regionen oft intern sehr heterogen → Verzerrung (Makroebene vs. Mikroebene).
- Unterschiedliche Datensätze → unterschiedliche Ergebnisse
- Vergleichbarkeit zwischen Regionen schwierig (Größe, Struktur, Dichte)
- GIS bietet bessere Alternativen:
  - Pufferzonen (Buffers)
  - Rasterdaten, Distanzmaße
  - flexible, theorienahe Kontexträume

Validitätsbedrohungen & natürliche Experimente

3. Weitere Validitätsbedrohungen

- **Räumliche Autokorrelation:**
  - Nahe Beobachtungen ähneln sich stärker → Verletzung der Unabhängigkeitsannahme
- **Messfehler / Datenquellenprobleme:**
  - Raster vs. Vektordaten → teils stark abweichende Ergebnisse
  - Ungenaue Geokodierung
- **Externe Validität:**
  - Ergebnisse oft nur für untersuchten Raum gültig → begrenzte Generalisierbarkeit

4. Nutzen natürlicher Experimente

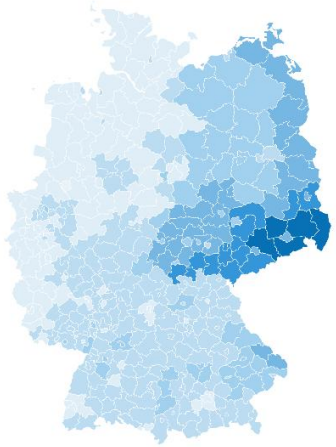
- Ermöglichen kausalere Schlussfolgerungen als Querschnittsvergleiche
- Geeignet für exogene Veränderungen in:
  - Diversität
  - Infrastruktur / Grenzen / Umweltbedingungen
- Kombination mit GIS ideal, um:
  - räumlich präzise und flexibel abzugrenzen
  - Interaktionsräume theoriegerecht zu definieren
  - Abhängigkeiten & Heterogenität kontrollieren zu können
- **Fazit:** Natürliche Experimente + GIS → starke methodische Ergänzung

ÜA 10 – Lösungsansätze

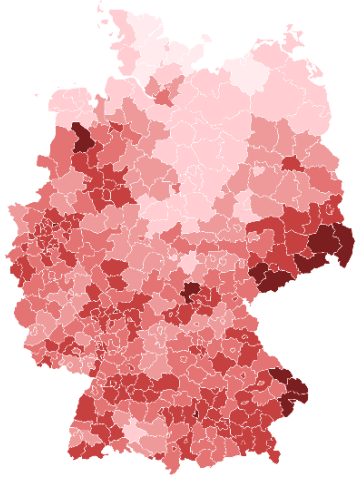
1. Hinreichende Evidenz für die Kontaktthese?
   - Problem des **ökologischen Fehlschlusses:**
     von Zusammenhängen auf der Makroebene kann nicht einfach auf die entsprechenden Zusammenhänge auf der Mikroebene geschlossen werden
   - Kontakt ist auf der Mikroebene anzusiedeln, eine hohe ethnische Diversität resultiert nicht unbedingt in mehr Kontakt

Beispiel eines ökologischen Fehlschlusses

AfD Wählerstimmen
Bundestagswahl 2017



Covid-19 Fallzahlen



- Durch Möglichkeit des ökologischen Fehlschlusses nicht unbedingt kausal!

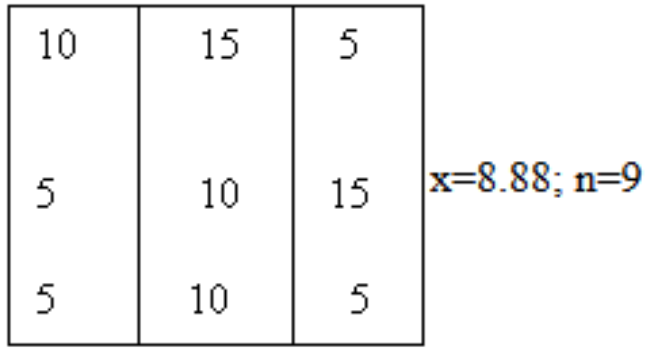
Quelle: Tagesspiegel

ÜA 10 – Lösungsansätze

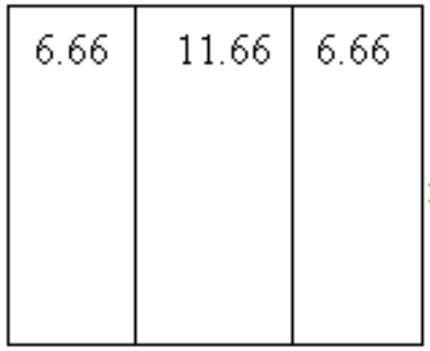
2. Probleme von regionalen Einheiten wie Landkreise oder Stadtviertel?
   - Zu große Gebiete bzw. lebensferne Gebietsgrenzen
   - Alternativ: Kleinräumigere Daten: Häuserblock, 1km² Raster; oder Einheiten, die stärker „Nachbarschaften“ messen
   - Vektordaten bzw. Punktdaten

Einfluss der räumlichen Klassifizierung

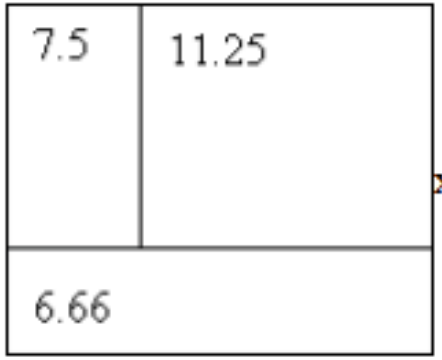
- **Scaling effects:**
  Ergebnisse abhängig von *N* Gebietseinheiten (Feingliedrigkeit; s. auch *Checkerboard Problem* nächste Folie)
- **Zoning Effects:**
  Ergebnisse abhängig davon, wie Gebiets-einheiten gebildet werden



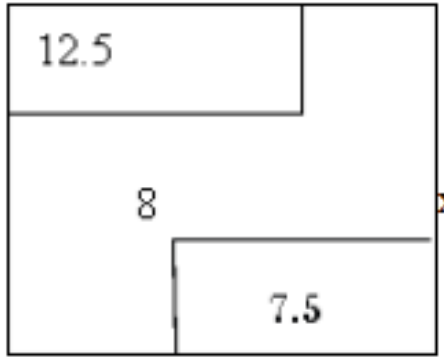
x=8.88; n=9



x=8.33; n=3



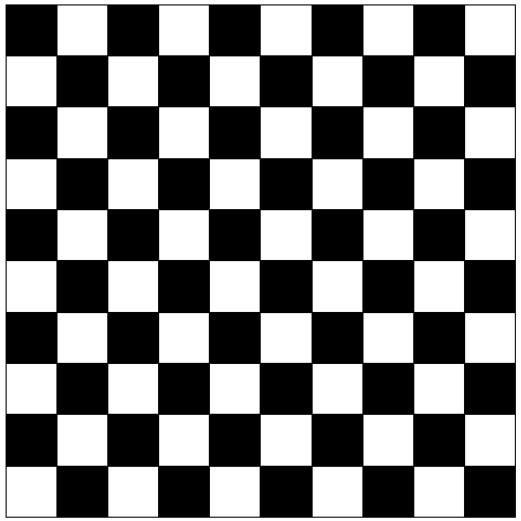
x=8.47; n=3

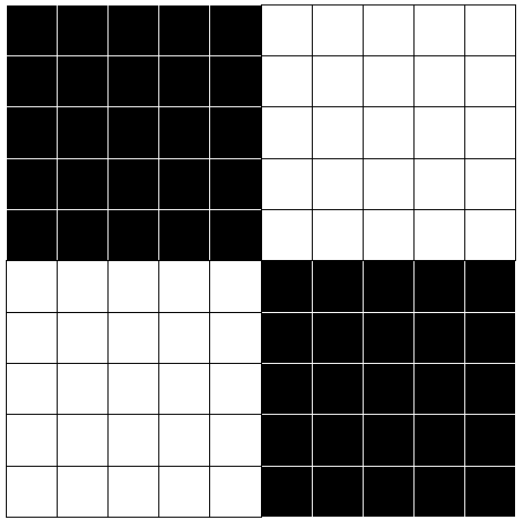


x=9.33; n=3

Einfluss der räumlichen Klassifizierung

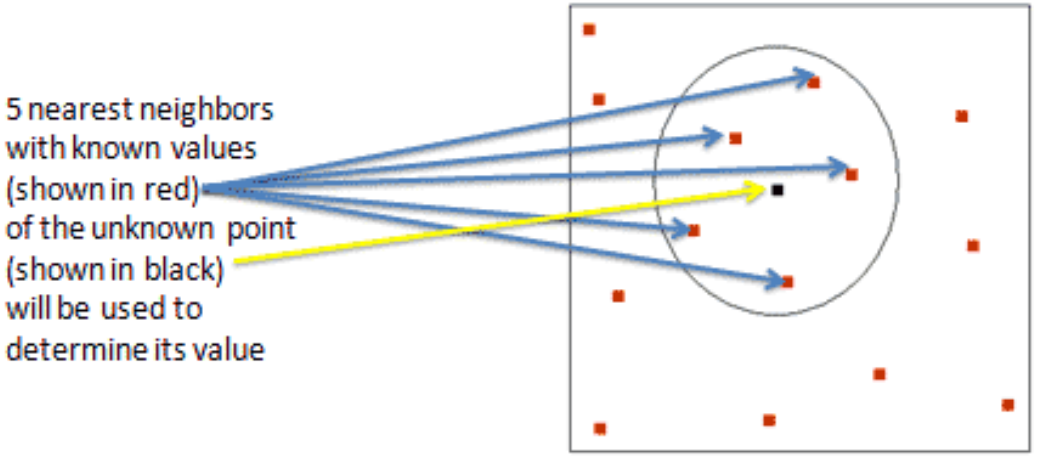
- **Checkerboard Problem**: Aggregationen können Muster auf der Mikroebene verdecken und zu falschen Schlüssen verleiten





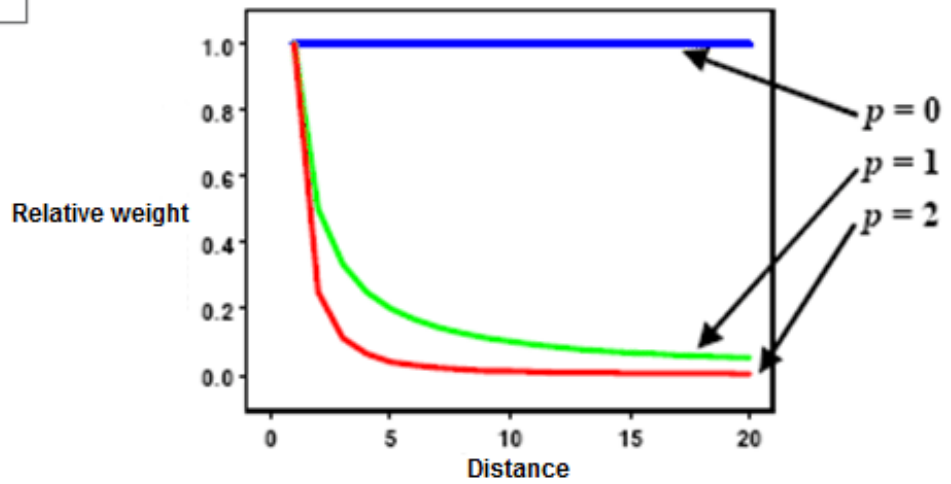
Alternativ: Räumliche Analysen mit Punktdaten

- Definition individueller Nachbarschaften und Interpolationen



Räumliche Interpolation als Dichte-Analyse oder als gewichteter Mittelwert der Umgebung; z.B.

- *Inverse Distance Weighted*-Interpolation
- Grafik links: 3 mögliche Gewichte: $\frac{1}{Distance^p}$



ÜA 10: 3. Weitere Bedrohungen Validität?

Interne Validität

- Mögliche Konfundierung durch andere Kontextmerkmale (z.B. Leerstand, GDP)
- Datenaktualität (insb. 1km² Raster basiert auf Zensusdaten von 2011)
- Unterschiedliche Datenqualität (und Definitionen) bei versch. Datenprovidern
  - Z.B. häufig nur Verfügbarkeit Staatsangehörigkeit
  - Gerade bei kommerziellen Anbietern besteht wenig Transparenz und ist die Datengüte ggf. gering
- Problematische Definition von „Kontakt“
- Scaling/Zoning

Externe Validität

- Selektive Datenverfügbarkeit: insb. besser in urbanen Räumen

Abschließend: Tools, ergänzende Hinweise

- Gängige Softwaretools:
  - ArcGIS
  - Stata (insb. einige nützliche ados), R
- Datenorganisation: Oft Überlagerung von Punktdaten und Kartenschichten („Layer“: im Vektor-, Punkt- oder Rasterformat
  - Etwa Punktdaten zu Wohnadressen mit Kartenschichten zu Straßennetzwerken, Umweltbelastungen, Standorten von Windrädern
- Statistische Analysen sollten räumliche Clusterungen beachten
  - Fehlende Unabhängigkeit (*spatial autocorrelations*)
  - Ohne Modellierung drohen verzerrte Ergebnisse (Koeffizienten, Standardfehler)
- S. Grundlagentext für weitere (Literatur-) Hinweise!
  - Für shapefiles siehe Eurostat GISCO
  - How to map?
    - R: Kieran Healy https://visualizingsociety.com/
    - Stata: Asjad Naqvi https://medium.com/the-stata-guide