

Vorlesung: Empirische Bildungsforschung

Sitzung 6: Evaluationsparadigma

Dr. Edgar Treischl
Last update: 2022-04-25

This presentation is licensed under a CC-BY-NC 4.0 license. You may copy, distribute, and use the slides in your own work, as long as you give attribution to the original author on each slide that you use. Commercial use of the contents of these slides is not allowed.



Ablauf

1. Evaluationsparadigma der EBF
2. Kausalität in der Bildungsforschung
3. Experimentelle Bildungsforschung

1. Evaluationsparadigma der EBF

Reformen des Bildungssystems

Problem:

Oftmals keine wissenschaftliche Begleitung der Reformen; Wenig Wissen über die Wirkung der Reformen!



Bild: dpa



Bild: Bürgerbewegung Pro NRW



Wirkungsevaluation

Literaturauszug: Treischl & Wolbring (2020: Kapitel 2):

“Bei Evaluationen geht es um den Wert einer Sache bzw. genauer um dessen Bestimmung, wobei das zu bewertende Objekt als Evaluationsgegenstand (Evaluandum) bezeichnet wird.”

“Gegenstand einer Evaluation können dabei Produkte, Prozesse, Veranstaltungen, Interventionen oder ganze Maßnahmenbündel sein, die mehr oder weniger systematisch bewertet werden.”

“Wie Kromey (2001, S. 106) erläutert, meint Evaluation im alltagssprachlichen Sinne jedwede Form der mehr oder weniger systematischen Beurteilung:”

“Irgend etwas wird von irgend jemandem nach irgendwelchen Kriterien in irgendeiner Weise bewertet. Derselbe Sachverhalt kann – wie die Alltagserfahrung – von verschiedenen Individuen sehr unterschiedlich bis gegensätzlich eingeschätzt und beurteilt werden (Kromey 2001, S. 106).”

Wissenschaftliche Evaluation

- Systematische Sammlung empirischer Evidenzen
- Systematik und präzise Dokumentation
- Intersubjektive Nachprüfbarkeit
- Identifikation von Ursache und Wirkung

“Das wesentliche Ziel vieler Evaluationen entspricht dieser engeren Definition des Evaluationsbegriffs, nämlich Antworten auf Fragen der kausalen Inferenz zu geben.”

“Die Erfassung und Bewertung von Wirkungen und ihre kausale Ursachenzuschreibung sind die zentralen Aufgaben von Evaluation“ (Stockmann 2006, S. 106; Hervorhebung im Orig.).”

Funktionen von Evaluationen

(siehe Treischl & Wolbring 2020:21-24)

- Erkenntnisfunktion
- Kontrollfunktion
- Transparenz und Dialogfunktion
- Legitimationsfunktion

Evaluationsstandards

- Systematische Untersuchung
- Kompetenz
- Integrität und Ehrlichkeit
- Respektvoller Umgang mit Menschen
- Verantwortung gegenüber der öffentlichen Wohlfahrt

2. Kausalität in der Bildungsforschung

Schul-Klischees im Faktencheck

Kleine Klassen lernen besser - oder doch nicht?

Oh Gott, so viele Schüler! Lehrer beklagen oft, ihre Klassen seien zu groß, Eltern stimmen gern mit ein. Wie viel ist dran an der gefühlten Wahrheit, die Klassen würden immer größer? Und nützen kleinere Lerngruppen tatsächlich dem Unterricht?

Von Heike Sonnberger



Evaluation anhand von Beobachtungsdaten

Effekt der Klassengröße auf die Schulleistung

Schulklasse A: > als 25 Kinder

Faktisch



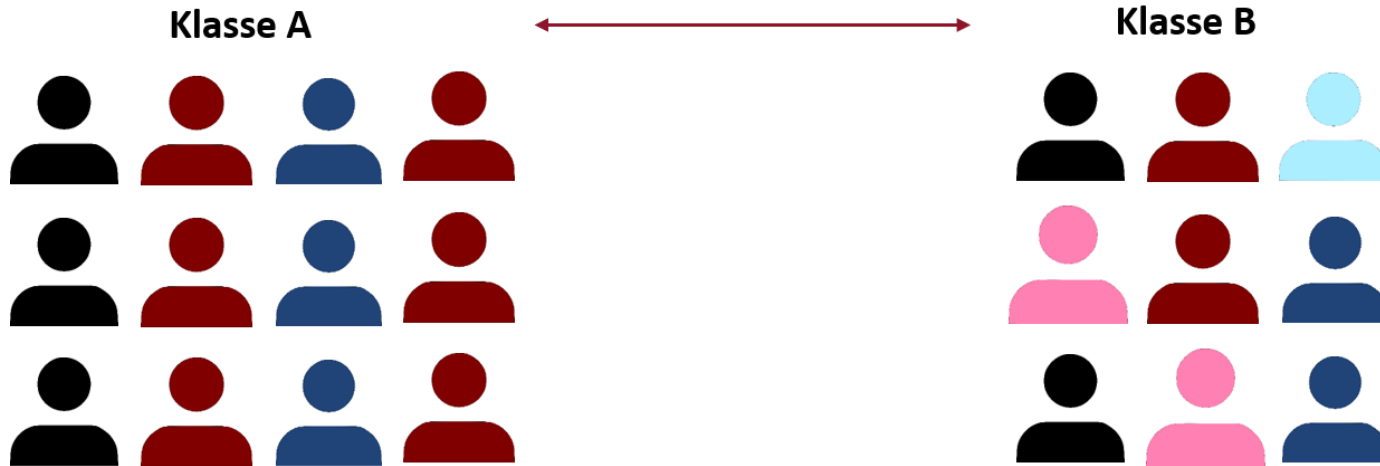
Schulklasse B: max. 15 Kinder



Kontra-
faktisch

Bessere
schulische
Leistung?

Messung kausaler Effekte



Problem: Unbeobachtete Heterogenität zwischen Klasse A und B.

ADD APPLES, BUT DON'T ADD, DIVIDE, OR COMPARE APPLES AND ORANGES! 🧙

The Fundamental Problem of Causal Inference: I

Auf der Individualebene

Große Klasse



Edgar

Faktisch

Kleine Klasse



Kontrafaktische Frage:

Wenn Edgar eine kleine Schulklasse besucht hätte, wäre seine Noten besser?

Kontra-
faktisch



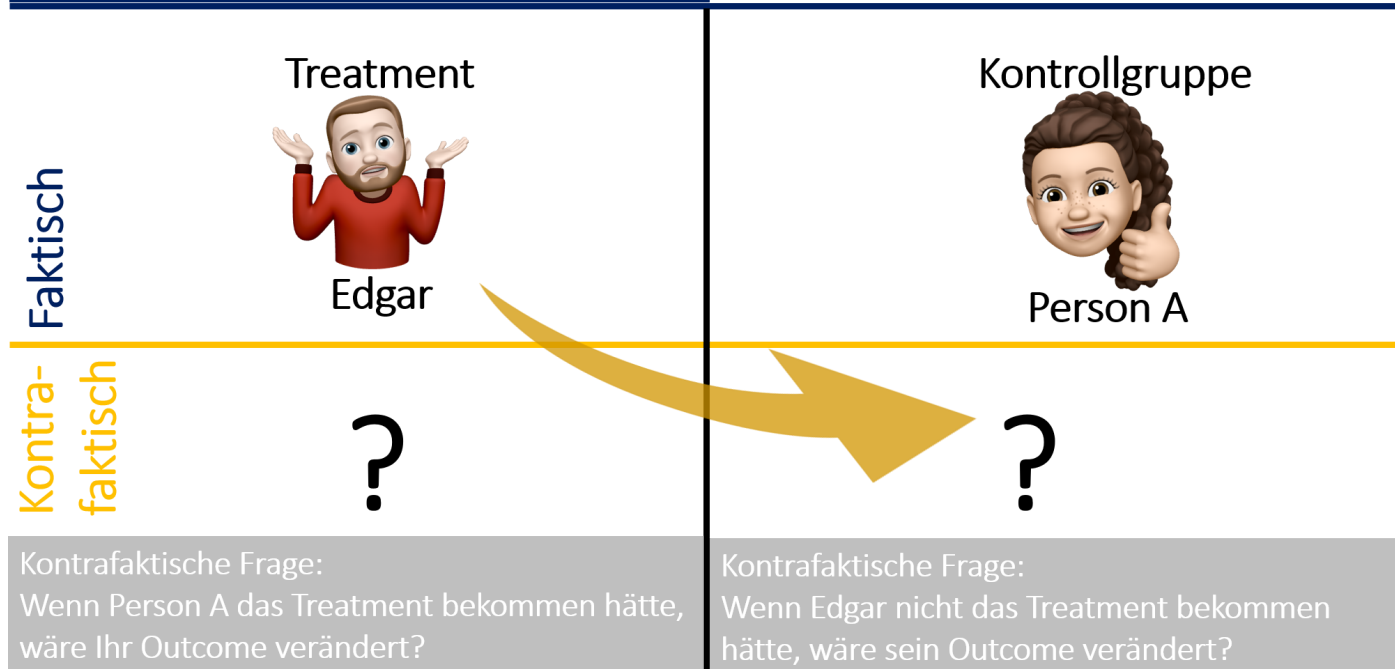
Wenn Edgar eine große Klasse besucht hätte, wären seine Noten schlechter?



Edgar

The Fundamental Problem of Causal Inference: II

Auf der Individualebene



Etwas formaler (1):

- Sei D eine Dummy-Variable (d.h. eine binär kodierte Variable), welche das Auftreten bzw. Wegbleiben eines Stimulus (Ursache) anzeigt.
- Der kausale Effekt δ ist definiert als Differenz zwischen Y^1 , dem Zustand der Zielvariable Y nach Auftreten ($D = 1$) der Ursache, und Y^0 , dem Zustand der Zielvariable ohne Auftreten ($D = 0$) der Ursache

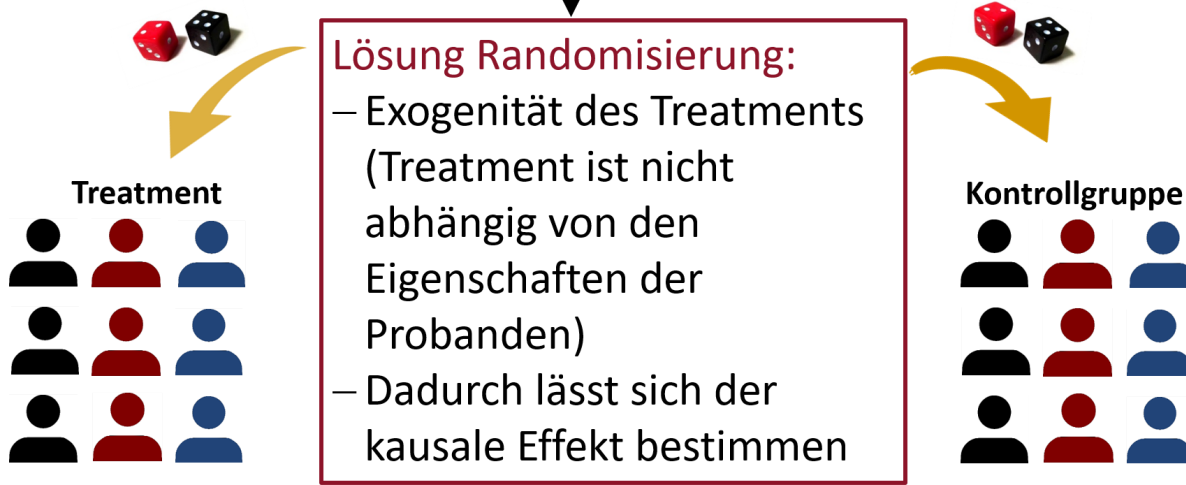
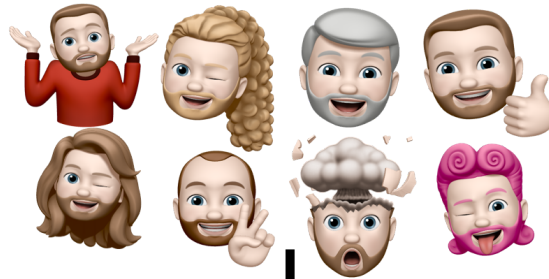
$$- \delta = Y^1 - Y^0$$

Etwas formaler (2):

- Eine Person kann nicht gleichzeitig ein Treatment bekommen und in der Kontrollgruppe sein.
- Ergo: Messung von Kausalität auf der Mikroebene des Individuums ist damit selbst im Experiment ausgeschlossen.
- Problem: Wir sind auf der Suche nach einer kontrafaktischen Annäherungen eines (unbeobachtetes) Ergebnisses: Welchen Schulerfolg hätte Person i gehabt, wenn die Person die Klasse B anstelle der Klasse A besucht hätte?

Kausaler Effekt: Messung des Average Treatment Effects

Auf der
Aggregats-
ebene



Dies korrespondiert mit dem **Prinzip von Compare Like with Like**: “Durch die Randomisierung ist bei hinreichend großen Gruppen von einer Gleichverteilung in der Kontroll- und Versuchsgruppe (Bsp: ca. 50% Männer/Frauen) in allen möglichen Drittvariablen auszugehen.”

“Randomisierung umgeht somit das Problem der unbeobachteten Heterogenität (Gruppen unterscheiden sich ohne dass wir es beobachtet haben) von Gruppen. Die Randomisierung hilft uns auch bei methodischen Problemen (wie bspw. das Problem von Scheinkorrelationen).”

Hat die Randomisierung funktioniert?

Table 4.1 Case 3—Balance between Treatment and Comparison Villages at Baseline

| Household characteristics | Treatment villages (N = 2964) | Comparison villages (N = 2664) | Difference | t-stat |
|--|-------------------------------|--------------------------------|------------|--------|
| Health expenditures (\$ yearly per capita) | 14.48 | 14.57 | -0.09 | -0.39 |
| Head of household's age (years) | 41.6 | 42.3 | -0.7 | -1.2 |
| Spouse's age (years) | 36.8 | 36.8 | 0.0 | 0.38 |
| Head of household's education (years) | 2.9 | 2.8 | 0.1 | 2.16* |
| Spouse's education (years) | 2.7 | 2.6 | 0.1 | 0.006 |
| Head of household is female = 1 | 0.07 | 0.07 | -0.0 | -0.66 |
| Indigenous = 1 | 0.42 | 0.42 | 0.0 | 0.21 |
| Number of household members | 5.7 | 5.7 | 0.0 | 1.21 |
| Has bathroom = 1 | 0.57 | 0.56 | 0.01 | 1.04 |
| Hectares of land | 1.67 | 1.71 | -0.04 | -1.35 |
| Distance to hospital (km) | 109 | 106 | 3 | 1.02 |

3. Experimentelle Bildungsforschung

- Zufallsbedingte Aufteilung (Randomisierung) auf Kontroll- und Versuchsgruppe (Treatment)
- Zufällige Aufteilung auf beide Gruppen kontrolliert für alle Drittvariablen (Compare Like with Like)
- Manipulation: Vermutete Ursache (= Stimulus) in der Versuchsgruppe
- Manipulation: Keine Ursache (= kein Stimulus) in der Kontrollgruppe
- Werte in der Versuchsgruppe die über der Schwelle der Zufallsvarianz (t-Test) liegen werden auf den Stimulus zurückgeführt und als kausaler Effekt interpretiert.

Experimentelle Forschungsdesigns: Beispiel Weiterbildungsmaßnahme

| Zuweisung | X | Messung t_1 |
|----------------|----------------|----------------|
| R ¹ | Kontrollgruppe | O ² |
| R | Versuchsgruppe | O |

[1] Randomisierung

[2] Messung (Observation)

Quasi-experimentelles Design

- Zuordnung auf Kontroll- und Versuchsgruppe erfolgt nicht durch Randomisierung
- Quasi-Experimente: Stimulus wird gegeben, Zuordnung erfolgt selbst durch die Forschungsobjekte
- Evaluationsforschung basiert oftmals auf quasi-experimenteller Versuchsanordnungen (Probleme?)
- Notation Quasi-Experiment:

| Zuweisung | X | Messung t_1 |
|-----------|----------------|---------------|
| | Kontrollgruppe | O |
| | Versuchsgruppe | O |

“Experiments in the social sciences? What should we and what should we not do. The Little-Albert-Experiment.”

Kaum experimentelle Forschungsdesigns in der EBF

Häufige Bedenken von Stakeholdern:

1. Experimente sind häufig nicht ethisch vertretbar.
2. Experimente sind nicht praktisch umsetzbar sind.
3. Experimente sind künstlich aufgrund der Künstlichkeit des Settings.
4. Experimente basieren auf einer geringen Fallzahl und die Ergebnisse sind nicht auf andere Kontexte übertragbar.

Einwände zum Teil nicht zutreffend, weil:

1. Ethikboard müssen Verstöße gegen ethische Einwände einlegen.
2. Experimente sind auch in der Bildungsforschung möglich.
3. Künstlichkeit betrifft in erster Instanz Laborexperimente.
4. Geringe Fallzahl ist kein allgemeines Problem experimenteller Forschung.

"REPRÄSENTATIVITÄT"

Beschreibung/ Deskriptive Frage:

Wie viele Jugendliche in Deutschland mit niedriger sozialer Herkunft studieren?

Erklärung/ Kausale Frage:

Zeigen Kinder in kleinen Schulklassen die besseren schulischen Leistungen? X: Klassengröße => Y: Noten.

Random Samples:

Für Populationsschätzer (Anzahl an Jugendlichen) und den Rückschluss auf eine Grundgesamtheit (Jugendliche in D) brauchen wird eine probabilistische Stichprobe (bspw. Zufallsstichprobe).

Random Assignment: Zum Testen von Zusammenhangshypothesen ist keine Zufallsstichprobe notwendig, vielmehr hat die Randomisierung von Probanden (Random assignment) Vorteile gegenüber Umfragen.

Bildungsforschung braucht beide Ansätze um die gesteckten Ziele zu verfolgen!

Statt dessen müssen wir uns über die interne und die externe Validität unserer Evaluationsbefunde Gedanken machen! Keine Ahnung was interne oder externe Validität ist? 😊 Ein kurzer Auszug aus der Literatur zur Vertiefung (Treischl & Wolbring 2020, Kapitel: 6).

“Der Begriff der internen Validität zielt darauf ab, inwiefern ein Effekt tatsächlich kausal auf das Treatment zurückführbar ist.”

“Anders ausgedrückt, bemisst die interne Validität damit, inwieweit mögliche Störfaktoren durch das gewählte Forschungsdesign ausgeschlossen werden oder ob nicht doch alternative Erklärungen einen Befund hervorgebracht haben könnten.”

“Im Kern beziehen sich Fragen nach der internen Validität also auf die Belastbarkeit von Wirkungsaussagen im Zuge einer Evaluation.”

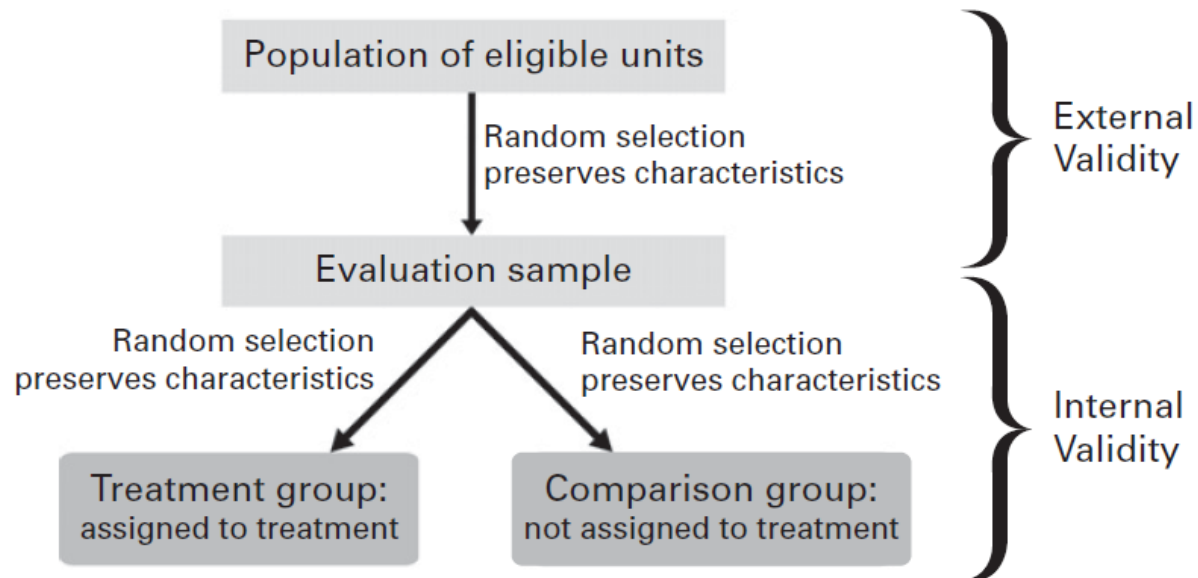
“Neben Fragen der internen Validität stellt sich auch bei Evaluationen die Frage nach der Generalisierbarkeit, also der externen Validität empirischer Befunde. Dabei gilt es vier Dimensionen der Generalisierbarkeit zu berücksichtigen.”

UTOS: Dimensionen der Generalisierbarkeit nach Cronbach (1982)

- Versuchspersonen (Units)
- Interventionen (Treatments)
- Messungen (Observations)
- Kontexte (Settings)

Zusammenspiel interne und externe Validität

Figure 4.2 Random Sampling and Randomized Assignment of Treatment



Source: Authors.

Zusammenfassung

- EBF ist Teil eines Evaluationsparadigmas (Reformen, Unterrichtsmethoden etc.)
- Wirkungsevaluation = Kausalanalyse
- Vorteil von Randomisierung (Eliminierung von Störvariablen)
- Bedenken gegen experimentelle Forschungsdesigns in der EBF
- Güte experimentelle Forschungsdesigns: Interne und externe Validität