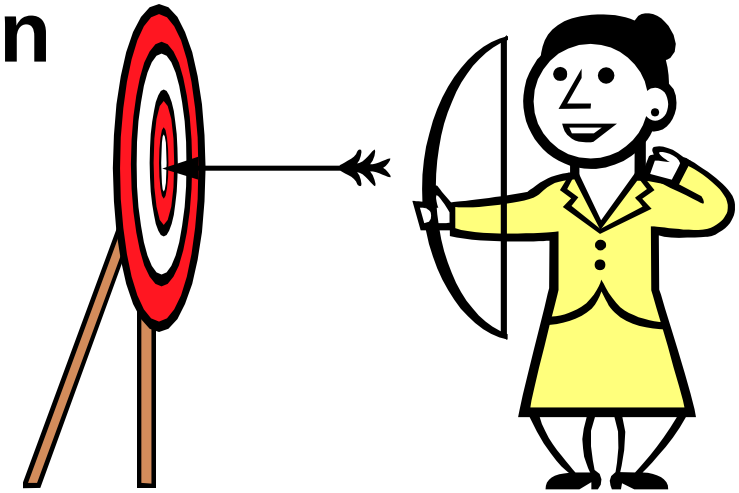
# »Wirkungsmessung im Kontext von Evaluationen – Möglichkeiten und Grenzen in der Praxis«

**Weiterbildungsseminar S3**
im Rahmen der 19. DeGEval-Jahrestagung
21. September 2016, Salzburg

**Prof. Dr. Alexandra Caspari** caspari@fb4.fra-uas.de
**Fachbereich 4** Soziale Arbeit und Gesundheit

# Zielsetzung

**TeilnehmerInnen kennen die verschiedenen Forschungsdesigns, deren Vor- und Nachteile und können ein Untersuchungsdesign selbst entwerfen**



# Wie sicher ist kausaler Einfluss?

- Ja häufiger eine Person Horrorfilme konsumiert, desto häufiger neigt sie zu Aggressivität/eigener Gewaltanwendung
  → *Einfluss Mediengewalt?*
- Arbeitslose sind häufiger krank als Nicht-Arbeitslose
  → *Folge der Arbeitslosigkeit?*
- Kranke Personen, die ein Medikament erhalten, fühlen sich nach zwei Wochen deutlich besser
  → *Einfluss des Medikaments?*
- Frauen erhalten (in Dt) nur 78% des Lohns von Männern
  → *Folge von Lohndiskriminierung?*
- allgemein: Personen, die an einer Maßnahme teilnehmen, sind/geht es anschließend „besser“
  → *Folge der Maßnahme?*

➔ Häufig wird auf Basis eines nachweisbaren ***Zusammenhangs*** ein *kausaler (ursächlicher) Schluss* gezogen, d.h. eine ***Wirkung*** von Mediengewalt, Arbeitslosigkeit, Maßnahmen abgeleitet

# Was sind Wirkungen?

- Veränderungen nach Beendigung einer Maßnahme

  → Veränderungen, die sowohl auf Maßnahme als auch beliebige Anzahl anderer Einflüsse zurückzuführen sind

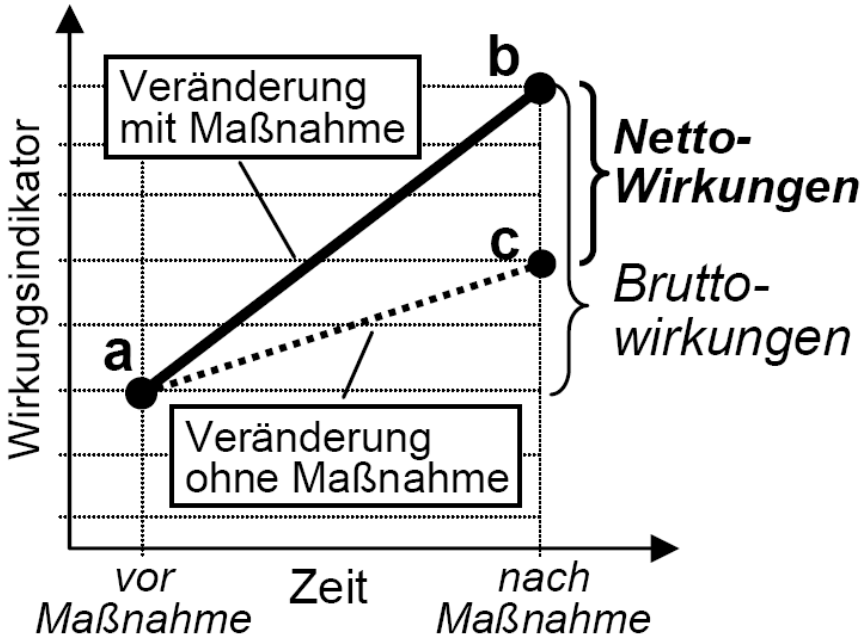
  = ***Bruttowirkungen*** (Differenz b-a)

- Veränderungen, die *allein* auf die durchgeführte Maßnahme zurückzuführen sind

  → isolierter Anteil an insgesamt auftretenden Veränderungen, die nicht beobachtbar gewesen wären, wenn Maßnahme nicht durchgeführt worden wäre

  = ***Nettowirkungen*** *oder* ***Projektwirkung*** (Differenz b-c)

  **= *kausaler Effekt***




➔ Zielerreichung ≠ Nettowirkung

# Nettowirkungen/Kausaler Effekt

- Ein vom Auftreten eines kausal wirksamen Faktors $T$ (Maßnahme) abhängiger kausaler Effekt $\delta_i$ (Wirkung) ist die Differenz zw. dem Ereignis $Y_1^i$, das bei Auftreten von $T\ (T=1)$ realisiert wird, und dem alternativen Ereignis $Y_0^i$, das ohne $T\ (T=0)$ eintreten würde:

$$\delta_i = Y_1^i\ (X_i, T=1) - Y_0^i\ (X_i, T=0) = Y_1^i - Y_0^i$$

- Wirkungen sind nicht *direkt* beobachtbar:
  - Ereignis $Y^i$ nur für $T=1\ (Y_1^i)$ *oder* $T=0\ (Y_0^i)$ beobachtbar
  - für Teilnehmer einer Maßnahme $(X_i, T=1)$ ist Ergebnis $Y_0^i\ (X_i, T=0)$ *nicht beobachtbar* (= ***das Kontrafaktische***)
- Wirkungen werden anhand *durchschnittlicher* Werte *empirisch erschlossen:* $\hat{\delta} = \overline{Y}_1 - \overline{Y}_0$
- Vergleich Ereignis bei Zielgruppe (ZG) und *hypothetischem Ereignis*, das ohne Maßnahme eingetreten wären

# Nettowirkungen/Kausale Effekte

Wie können Nettowirkungen bzw. kausale Effekte analysiert werden?

- notwendig für Ableitung eines *kausalen* Zusammenhangs:
  - es muss Zusammenhang zwischen 2 Variablen X und Y bestehen
  - Ursache X muss Wirkung Y zeitlich vorausgehen
  - Zusammenhang zwischen X und Y darf nicht durch andere Einflüsse Z bedingt sein
    (andere Erklärungen des Ursache-Wirkungs-Zusammenhang müssen eindeutig ausgeschlossen werden)

→ Einflüsse Z können am eindeutigsten ausgeschlossen werden, wenn außer X und Y alle Bedingungen konstant bleiben *(=„Knackpunkt“!)*

➔ **Experiment** am besten geeignet:
Prinzip der Bedingungskontrolle durch Einführung von ***Untersuchungsgruppe (UG)*** *und* ***Kontrollgruppe (KG)***
➔ ***Kontrollgruppe*** = *hypothetische Veränderungen*, die ohne Maßnahme eingetreten wären *(Kontrafaktische)*

# Zum Experiment allgemein

- ***Experiment*** bezeichnet Untersuchungen, die Aussage über *Kausalzusammenhang* zweier Variablen ermöglichen
- Aspekte eines Experiments:
  - Unterscheidung der zwei Variablen $X$ und $Y$ in
    → ***unabhängige Variable*** (UV, erklärende = $X$) und
    → ***abhängige Variablen*** (AV, zu erklärende = $Y$)
  - UV muss AV zeitlich vorausgehen (Sequenz UV→AV)
  - Daten von mind. zwei Probandengruppen werden verglichen
- Vorgehensweise:
  - Einteilung der Probanden in zwei Gruppen:
    → ***Untersuchungsgruppe*** (UG) &
    → ***Kontrollgruppe*** (KG)
    Kontrollgruppe = *hypothetische* Veränderungen,
    die ohne Maßnahme eingetreten wären *(=Kontrafaktische)*
  - durch Forscher/in kontrollierte Manipulation des „Stimulus“
    d.h. der unabhängigen Variable (UV)

# Typen von Experimenten

- **„Echte“ Experimente / randomized controlled trial (RCT)**
  - ***(Labor-) Experiment:***
    Einteilung in Untersuchungsgruppe (UG) und Kontrollgruppe (KG) ***randomisiert*** (→ neutralisiert personengebundene Störgrößen)
    Randbedingungen bekannt/kontrollierbar
  - ***Feldexperiment:***
    Experiment in „natürlichem“, vorhandenen Setting
    dennoch *Logik* des klassischen Experimentes (Randomisierung)
- **Quasi-Experiment:**
  - orientiert an Experimental-*Logik,* aber nicht alle Bedingungen des „echten“ Experiments erfüllt (keine randomisierte KG)

  → *Vergleichsgruppe* (VG) anstatt KG!
  - wird aufgrund vorhandener Eigenschaften (Alter, Geschlecht, …) (re-)konstruiert (=nicht randomisiert)
  - ermöglicht keine vollständige Kontrolle von Drittvariablen

# Beispiel Experiment

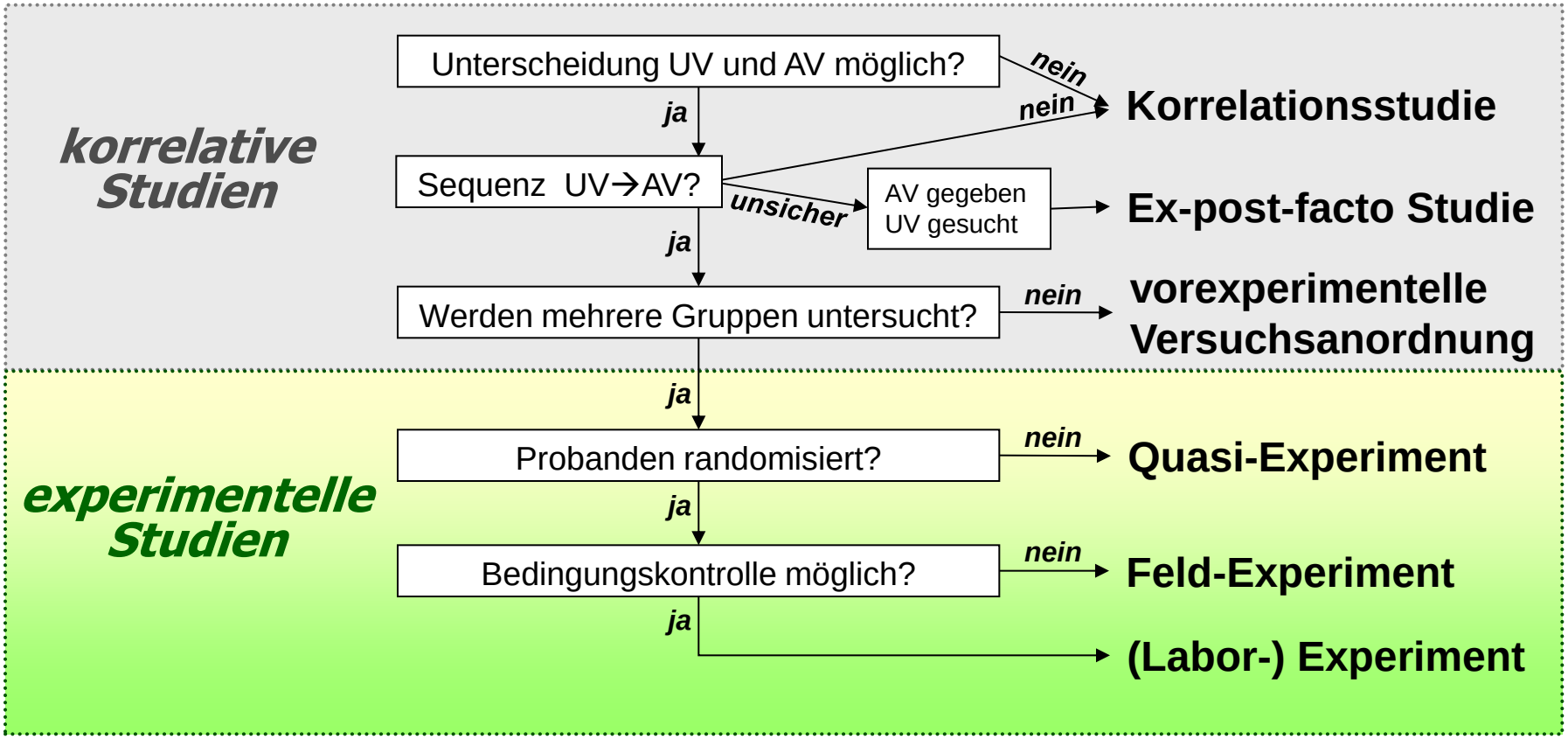
- Konsum Horrorfilme (UV) → Aggressivität (AV)?
  - Einteilung der Probanden in zwei Gruppen: UG & KG
    → durch Randomisierung werden „Drittvariablen“ neutralisiert
  - UG erhält „Stimulus“: Horrorfilm $(T=1)$
  - KG erhält keinen „Stimulus“ $(T=0)$
    → *hypothetische* Veränderungen, die ohne Maßnahme eingetreten wären *(=Kontrafaktische)*
  - anschließend wird Aggressivität in beiden Gruppen gemessen
  - Unterschiede zw. Gruppen im *durchschnittlichen* aggressiven Verhalten kann eindeutig auf Horrorfilm zurückgeführt werden
    $\hat{\delta} = \bar{Y}_1 - \bar{Y}_0$ (=kausale Attribution)
    = **Nettowirkungen / kausaler Effekt**

→ Umsetzung mit Hilfe angemessener ***Forschungsdesigns/Untersuchungsdesigns***

# Forschungsdesigns

- auch: Untersuchungsdesign, Untersuchungs-/Versuchsanordnung, Versuchsplan
- beschreibt, wie Fragestellung untersucht werden soll
- legt fest, *wer/welche Personen  wann/wie oft* untersucht werden
- entscheidend für Aussagekraft der Untersuchungsergebnisse
- zwei grundsätzliche Untersuchungsansätze:
  - korrelative Studien
  - experimentelle Studien

# Typen von Untersuchungsansätzen




UV: unabhängige (erklärende) Variable
AV: abhängige (zu erklärende) Variable

nach: Musahl/Schwennen 2000, in Anlehnung an Hager 1987

# Experimentelle (& Vorexperimentelle) Designs

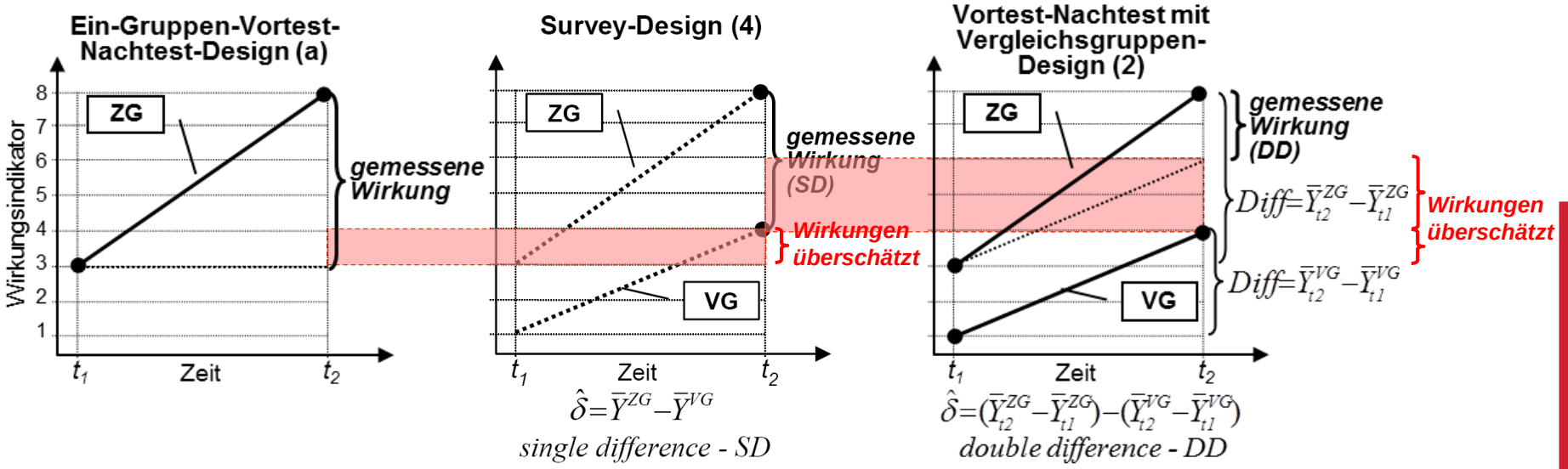
QUALITÄT

| | Design | Vorher-Messung $t_1$ (Baseline) | Stimulus | Nachher-Messung $t_2$ (Survey) |
|---|---|---|---|---|
| ***Experimentelle Versuchsanordnung/"randomised controlled trial" (RCT):*** | | | | |
| (1) | Kontrollgruppen-Design | $ZG_{t1}$<br>$KG_{t1}$ | X<br>– | $ZG_{t2}$<br>$KG_{t2}$ |
| ***Quasi-experimentelle Versuchsachordnung:*** | | | | |
| (2) | Vortest-Nachtest mit Vergleichsgruppen-Design | $ZG_{t1}$<br>$VG_{t1}$ | X<br>– | $ZG_{t2}$<br>$VG_{t2}$ |
| (3) | Vortest-Nachtest mit Nachtest Vergleichsgruppen-Design | $ZG_{t1}$ | X<br>– | $ZG_{t2}$<br>$VG_{t2}$ |
| (4) | Survey-Design | | X<br>– | $ZG_{t2}$<br>$VG_{t2}$ |
| ***Vorexperimentelle/Nicht-experimentelle Versuchsachordnung:*** | | | | |
| (a) | Ein-Gruppen-Vortest-Nachtest-Design | $ZG_{t1}$ | X | $ZG_{t2}$ |
| (b) | Ein-Gruppen-Nachtest-Design | | X | $ZG_{t2}$ |

ZG: Zielgruppe, KG: Kontrollgruppe (randomisiert), VG: Vergleichsgruppe (nicht randomisiert)
*t:* Zeitpunkt (erste, zweite Datenerhebung/Messung), X: Stiumulus (Projekt/Maßnahme)

# Unterschiede in den geschätzten Wirkungen

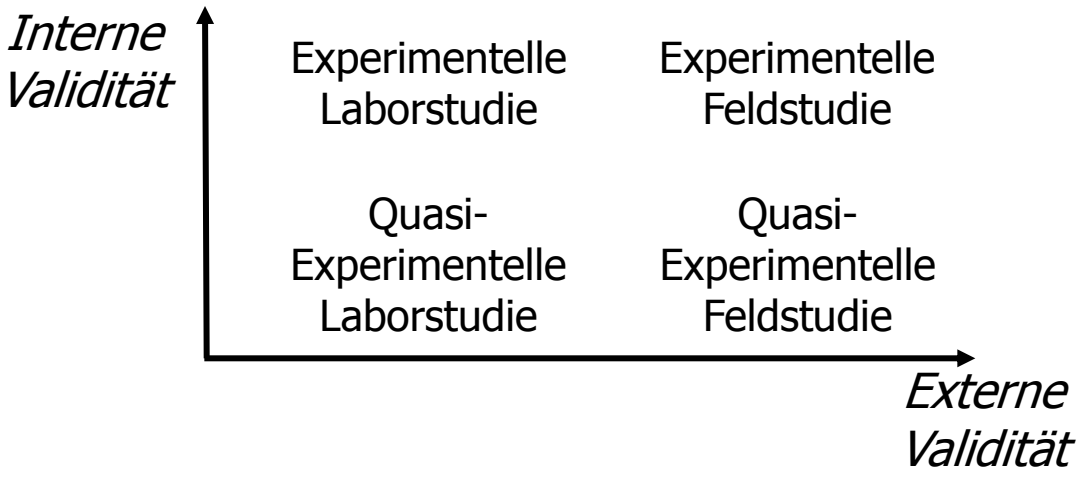
- Vorexperimentelle Versuchsanordnung
  - ohne Kontrollgruppe (kontrafaktische Situation)
  - Randbedingungen unkontrolliert
- Quasi-experimentelles Design
  - mit Vergleichsgruppe, aber nur Nachher-Messung
- Quasi-experimentelles Design
  - mit Vergleichsgruppe und Vorher-Nachher-Messung




● Datenerhebung, *ZG:* Zielgruppe, *VG:* Vergleichsgruppe, *t:* Zeit (erste, zweite Datenerhebung), *Diff*: Differenz

# Validität experimenteller Designs

- **Validität**: Gültigkeit/Belastbarkeit eines aufgezeigten Ursache-Wirkungs-Zusammenhangs
- **Interne Validität:** gemessene Veränderungen sind *eindeutig* auf *Veränderungen der UV* (Maßnahme) *zurückzuführen*
- **Externe Validität:** Ergebnis ist auf andere Populationen, Situationen und Zeitpunkte *generalisierbar*




- Interne Validität gefährdet durch Störfaktoren „THIS MESS“
- Externe Validität gefährdet durch Störfaktoren „UTOS“

# Umsetzungsmöglichkeiten

*Erfahrungen aus der Praxis:*

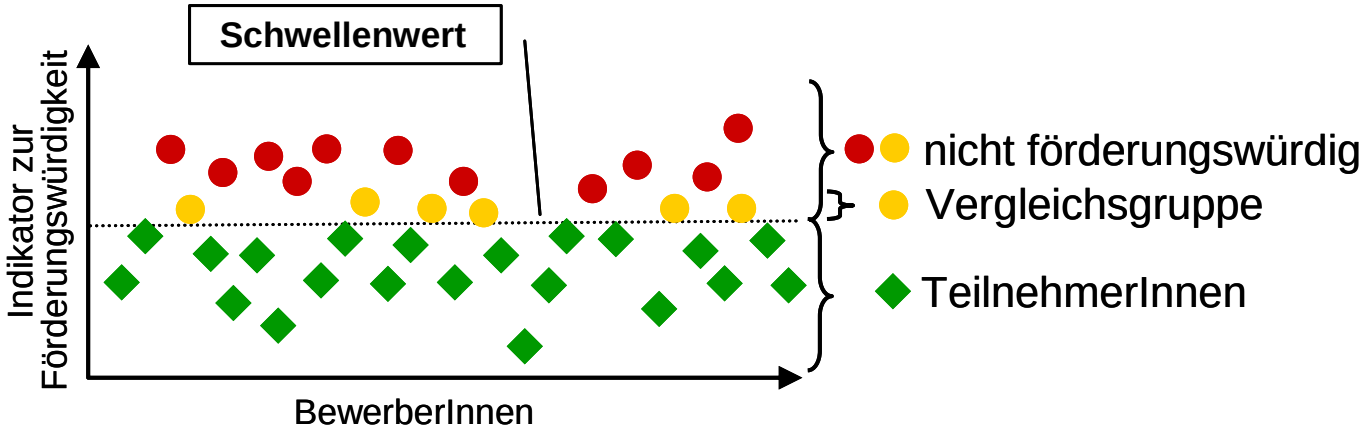
- angemessene Designs werden oft als unnötig anspruchsvoll oder aufgrund ethischer Vorbehalte abgelehnt
- realistische Wege, wie experimentelle oder quasi-experimentelle Designs in der Praxis angewandt werden können:
  - Matching on Observables
  - Regression Discontinuity
  - Propensity Score Matching (PSM)
  - Pipeline Approach
  - Multiple Comparison Group Design

# Matching on Observables

- quasi-experimentelles Design:
  - bewusste Auswahl anhand gleicher charakteristischer Merkmale (relevanter Drittvariablen) der ZG
    z.B. Alter, Zugang zu Service, Typ & Qualität Haus, ökonomische Situation, zentral/abseits gelegen, etc.
  - VG wird aus Personen, Dörfern, Regionen, Bezirken gebildet, die höchste Übereinstimmung in Eigenschaften mit ZG aufweisen
  - nicht beobachtbare Merkmale („unobservables“), z.B. Motivation, schwer zu berücksichtigen
- Konstruktion einer VG für Nachher-Messung im Rahmen einer Evaluation möglich ($t_2$)
  → „nur“ single-difference (SD) möglich
- oder bereits bei Planung
  → auch Vorher-Messung ($t_1$) → double-difference (DD) möglich

# Regression Discontinuity

- quasi-experimentelles Design: Konstruktion VG für Vorher- & Nachher-Messung
  - wenn Teilnahme an Maßnahme an bestimmte Voraussetzung mit gesetztem *Schwellenwert* gebunden, z.B. Einkommen, Alter, Testergebnis (Sprachtest), Leistung (Schule, Hochschule) etc.
  - wenn Erfüllung der Voraussetzung vorab überprüft wird
- VG = Personen, die Schwellenwert nur *knapp* nicht erreicht haben, aber sehr ähnliche Charakteristika wie ZG aufweisen




→ double-difference (DD) möglich

# Propensity Score Matching (PSM)

- quasi-experimentelles Design:
  - Konstruktion VG für Vorher- & Nachher-Messung
  - wenn Daten aus allgemeinen Surveys mit interessierenden Fragen zu Zeitpunkt $t_1$ und $t_2$ existieren
  - anhand charakteristischer Merkmale werden „Ähnlichkeitsindices“ geschätzt (berechnet)
  - auf Basis dieser „Ähnlichkeitsindices“ wird für jede Einheit der ZG eine (oder mehrere) „passende“ Einheiten aus dem Survey für VG ausgewählt, die sich bzgl. der Merkmale nicht von der ZG-Einheit unterscheidet („statistischer Zwilling“)

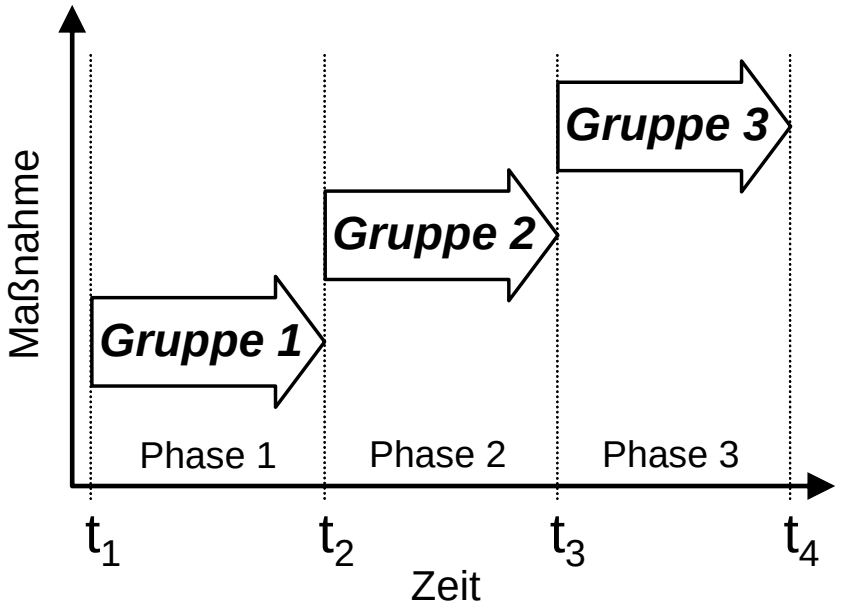
→ „Qualität“ der VG ~ KG

→ double-difference (DD) möglich

# Pipeline Verfahren

- experimentelles Design: KG für Vorher- & Nachher-Messung
  - wenn größeres Programm mit langer Laufzeit in *mehreren Phasen zeitversetzt* implementiert wird (Schulen, Schulklassen, Städte, Stadtteile, Dörfer, Regionen)
  - wenn *keine bewusste Entscheidung* darüber, warum Klassen, Stadtteile, Dörfer etc. an der ersten Phase, andere erst später teilnehmen sollen („randomized phasing in“)
- Einheiten, die erst an der 2. & 3. Phase teilnehmen = KG für Personen der 1. Phase

→ double-difference (DD) möglich



# Wichtige Anmerkungen

- Internationale Diskussion um Wirkungsmessung bezieht sich nur auf *kleinen Ausschnitt* im Kontext einer Evaluation
  → Frage, wie *eindeutige Wirkungszuschreibung* (kausale Attribution) methodisch realisiert werden kann
- Wirkungsmessung ≠ Wirkungsevaluation!
- Wirkungsmessung *notwendig*, jedoch *nicht hinreichend*!
  → nur „Untersuchung“, *ob* Maßnahme wirkt *oder nicht*
  → Frage nach *Warum* bleibt unbeantwortet „Black Box“
- *aussagekräftige* Wirkungsevaluationen benötigen ebenso:
  → Ursache-Wirkungs-Hypothesen (LogFrames incl. TOCs)
  → qualitative Methoden: Methodenmix & Triangulation

**Nur dann zeigen Wirkungsevaluationen evidenzbasierte Handlungsoptionen für die Implementation zukünftiger (Politik-) Maßnahmen auf**

# Reflexion

**TeilnehmerInnen kennen die verschiedenen Forschungsdesigns, deren Vor- und Nachteile und können ein Untersuchungsdesign selbst entwerfen**