Forschungsmethoden und Evaluation I / Modul 5

Einführungsskript:

WIE SOLL EINE UNTERSUCHUNG DURCHGEFÜHRT WERDEN?

Untersuchungsplanung muss **äußerst sorgfältig** durchgeführt werden. Prüfung einer wissenschaftlichen Hypothese erfolgt im Rahmen einer empirischen Untersuchung durch eine Struktur, die sie zusammensetzt aus

- ***unabhängigen Variablen*** (UV) -> im **Wenn-Teil der Hypothese** und
- ***abhängigen Variablen*** (AV) -> im **Dann-Teil der Hypothese**

BEISPIEL: **Hypothese**: „Wenn SchülerInnen am Förderunterricht teilnehmen, dann erzielen sie bessere Leistungen als SchülerInnen, die nicht am Förderunterricht teilnehmen."

UV: Teilnahme am Förderunterricht
(2 Abstufungen: teilgenommen vs. nicht teilgenommen)

AV: Leistung der SchülerInnen (Indikatoren sind z.B. Leistungstests)

Untersuchungsplan:

- Schülerinnen werden den ***2 Versuchsbedingungen*** (Förderkurs ja Förderkurs nein)
- ***randomisiert*** (um systematische Störfaktoren [z.B. Lehrereffekte] auszubalancieren) zugeteilt.
- Vor dem Kurs Leistungsmessung, dann Förderkurs, nach dem Kurs wieder Leistungmessung (= ***Kontrollgruppenplan mit Pre- und Posttest***).
- ***Versuchsgruppe*** nimmt am Kurs teil; ***Kontrollgruppe*** nimmt am Kurs nicht teil (hier KEINE Auswirkungen des Förderkurses!).

ABER:

- Es können zwar **Aussagen über die Effektivität des Förderkurses** (d.h. über seine Wirkungen) gemacht werden, **NICHT aber darüber, WAS am Förderkurs leistungssteigernd gewirkt hat** (z.B. es könnt z.B. auch das motivierende Lehrerverhalten gewesen sein, nicht die Inhalte des Kurses!).
- Es kann **nicht ausgeschlossen** werden, dass nicht auch **andere Maßnahmen denselben Effekt** gehabt haben könnten (z.B. Spielen anstatt Förderkurs).

Fazit:

- **Generalisierung** von Aussagen auf Basis von empirischen Untersuchungen **muss sehr präzise erfolgen**;
- Generalisierung **hängt von der Wahl des Versuchsplans ab**.

BEISPIEL: **Posterpräsentation**

Kontrollgruppen-Plan mit Pre- und Posttest:

- alle eingeladenen Personen werden VOR der Veranstaltung (= **Pretest**) nach ihrem Wissen über psychologische Forschung gefragt
- und dann noch einmal NACH der Veranstaltung (= **Posttest**).
- **Kontrollgruppe** = Personen, die die Veranstaltung nicht besuchen;
- **Versuchsgruppe** = Personen, die die Veranstaltung besuchen.

Nachteil: ***Randomisierung unmöglich*** (man kann ja die eingeladenen Personen nicht zufällig in Teilnehmer und Nicht-Teilnehmer einteilen...). Entscheidung zur Teilnahme an der Veranstaltung erfolgte nicht zufällig -> Aussagekraft der Ergebnisse wären bei so einem Versuchsplan eingeschränkt!

[keine randomisierte Zuteilung möglich -> ***quasi-experimentelles Design***]

Aufwand wäre bei so einem Design sehr hoch gewesen im Vergleich zum Ertrag -> wenig Effizienz!

Daher: ***One-Shot-Case-Study*** : es wurden nur EINMAL [nach der Veranstaltung] und nur in EINER Gruppe [= VG] Daten erhoben.

Bortz, J. & Döring, N (2002). Forschungsmethoden und Evaluation. Berlin: Springer. S 59f.

HATTE DAS MEISTERTRAINING EINEN EFFEKT?

One – Shot Case Study:

Firma möchte Führungsqualitäten des Meisters einer Abteilung verbessern -> schickt ihn in ein Trainingsprogramm. Danach überprüft Firma Betriebsklima, Arbeitszufriedenheit, Produktivität der entsprechenden Abteilung (= Evaluationsstudie). Fragebogen zeigen, dass es nichts zu beanstanden gibt.

Design: **T -> M**
T = Treatment (Trainingsprogramm)
M = Messung (Mitarbeiterbefragung)

ABER: **Ergebnis ist kausal nicht interpretierbar** (Tatsache, dass es nach dem Training nichts zu beanstanden gab, kann nicht auf das Meistertraining zurückgeführt werden, weil es ja auch schon vorher nichts zu beanstanden gegeben haben könnte [Lage vor dem Training wurde ja nicht erhoben!]).

Mit einer One – Case Shot Study sind Veränderungen NICHT feststellbar!

Ein – Gruppen – Pretest – Posttest – Design:

wäre besser gewesen, weil man damit Veränderungen feststellen kann (-> **Messung vor dem Treatment + Messung nach dem Treatment** => Vergleich!)

Design: **M_1 -> T -> M_2**
M_1 = Pretest
M_2 = Posttest

ABER: **kein zwingender Schluss möglich**, dass das Treatment die Veränderungen bewirkt hat! **Veränderung könnte auch zurückzuführen sein auf**:

- zwischenzeitliche ***Einflüsse, die unabhängig vom Treatment*** sind (z.B. Lohnerhöhung)
- ***Weiterentwicklung der Untersuchungsteilnehmer***, die unabhängig vom Treatment ist (z.B. bessere Vertrautmachung mit den Aufgaben)
- ***Pretest könnte zu Verhaltensänderung geführt haben*** (z.B. weil Untersuchungsteilnehmer dadurch auf bestimmte Probleme aufmerksam wurden)
- Tatsache, dass das ***gemessene Verhalten von Haus aus starken Schwankungen unterliegt*** (z.B. saisonale Schwankungen)
- Messungen können sich aus statistisch-formalen Gründen verändern (= ***Regressionseffekt***: bei wiederholten Messungen tendieren Extremwerte zur Mitte)

Ex Post Facto – Plan:

Behandelte Gruppe (VG; = Abteilung MIT Meistertraining) wird mit einer nicht – behandelten, nicht – äquivalenten Kontrollgruppe (= Abteilung OHNE Meistertraining) verglichen. [nicht – äquivalent, d.h. KG wird natürlich angetroffen, kommt NICHT durch Randomisierung zustande]

Design: T -> M_1 Vergleichende **Messung erst NACH dem Treatment**
M_2

ABER: **schlechte Interpretierbarkeit** der Ergebnisse, weil **Unterschiede** zwischen den beiden Gruppen ja auch **schon VOR dem Treatment bestanden haben könnten**.

Kontrollgruppenplan mit Pre- und Posttest:

wiederholte Messungen bei BEIDEN Gruppen -> zuverlässigere Interpretation!

Design: M_{11} -> T -> M_{12}
M_{21} ⟶ M_{22}

M11, M12 = Pretest + Posttest in der VG
M21, M22 = Pretest + Posttest in der KG

Bei Unterschieden zwischen den beiden Gruppen, **erkennt man an der Kontrollgruppe, ob dieser Unterschied auf das Treatment zurückzuführen sein könnte**.

ABER: **Unterschiede sind KEIN sicherer Beleg für die kausale Wirksamkeit des Treatments**, denn es könnte auch sein, dass z.B. das Alter als Störvariable wirkt [z.B. jüngere Mitarbeiter reagieren auf den neuen Führungsstil des Meisters mit mehr Leistung, ältere Mitarbeiter dagegen nicht...]

Fink, A. (1995): Design von Programm - Evaluationen

BEISPIEL 1: **Evaluations-Design mit Messungen zu verschiedenen Zeitpunkten an ein und derselben Gruppe**

Ein neues Spanisch-Schreibprogramm soll in allen 5. Klassen der insgesamt 12 Grundschulen eines Bezirks evaluiert werden. Wenn es effektiv ist, soll es das bisher verwendete Programm ersetzen. Evaluationsfrage: Verbessern sich die Kenntnisse der Schüler durch dieses neue Programm? Voraussetzung für die Teilnahme: Schüler müssen der Schulstufe entsprechend (oder besser) Spanisch lesen können. Standard: statistisch signifikante Verbesserung der Schüler. Testung 1 Monat vor Programmbeginn und nach 1 Jahr).

[**Hypothese**: Schüler können nach dem Programm besser Spanisch schreiben als mit dem alten Programm
UV: Programm
AV: Kenntnisse der Schüler beim Schreiben von Spanisch
VG: Schüler der 5. Klassen aller 12 Schulen
Design: Ein-Gruppen-Plan mit Pre- und Posttest

BEISPIEL 2: **Evaluations-Design mit Messungen zu verschiedenen Zeitpunkten an zufällig ausgewählter Gruppe**

wie oben, **ABER**: Testung nur in 6 der insgesamt 12 Schulen (zufällige Auswahl der Schulen)

VG: Schüler der 5. Klassen von 6 zufällig ausgewählten Schulen
Design: Ein-Gruppen-Plan mit Pre- und Posttest

BEISPIEL 3: **Evaluations-Design mit Randomisierung, zufällig ausgewählten Gruppen und Messungen zu verschiedenen Zeitpunkten**

Wie oben, **ABER**: die 6 teilnehmenden Schulen werden zufällig ausgewählt. Dann werden per Zufall 3 davon ermittelt, die nach dem alten Programm unterrichtet werden, während die anderen 3 nach dem neuen Programm unterrichtet werden.

DAHER: **VG**: 3 Schulen MIT neuem Programm
KG: 3 Schulen OHNE neues Programm
Design: Kontrollgruppenplan mit Pre- und Posttest

Diese Beispiele sollen **3 grundlegende Evaluationsmethoden** illustrieren:

BEISPIEL 1: Evaluationsdesign ohne Randomisierung, 2-malige Messung an ein- und derselben Gruppe zur Feststellung der Veränderung

BEISPIEL 2: Evaluationsdesign mit randomisierter Auswahl der zu testenden Gruppe, 2-malige Messung an dieser Gruppe zur Feststellung der Veränderung

BEISPIEL 3: Evaluationsdesign mit randomisierter Auswahl der zu testenden Gruppe und randomisierter Zuteilung zu VG und KG, 2 malige Messung an beiden Gruppen.

Bevor man sich das **Design für eine Evaluationsstudie überlegt**, sollte man mindestens **6 Fragen** überlegen:

1) Wie lautet die Evaluationsfrage und welche Standards sollen angewandt werden?
2) Was sind die UV?
3) Welches sind Einschluss- und Ausschlusskriterien?
4) Soll es eine KG geben? Wenn KG, dann was sind ihre Charakteristika?
5) Wann sollen die Messungen vorgenommen werden?
6) Wie oft sollen Messungen vorgenommen werden?

AUSWAHL EINES EVALUATIONS-DESIGNS: 6 FRAGEN

1) Evaluationsfrage und Standards:

Standards sind die **Richtlinie für ein Minimaldesign:**

- Wenn die Standards **Veränderungen über die Zeit hinweg** berücksichtigen sollen, so braucht man ein Design, in dem **Messungen zu 2 verschiedenen Zeitpunkten** durchgeführt werden müssen (vorher / nachher).
- Die **Einbeziehung einer KG** in ein Evaluationsdesign **stärkt immer die Aussagen, die über die Veränderungen** gemacht werden (z.B. Verbesserungen der Klassen, die neues Programm durchlaufen haben – keine Verbesserungen der Klassen, die neues Programm nicht durchlaufen haben. Ohne KG könnten Verbesserungen ja auch auf natürlichen Reifeprozess der Kinder zurückgeführt werden)

2) **Unabhängige Variable(n)**:

= **experimentelle bzw. Prädiktor-Variablen**. Ihre **systematische Variation wirkt sich auf die AV aus**. Sie sind **vom Programm selbst unabhängig** und Teil des Untersuchungsdesigns.

unabhängige Variablen bei Evaluationsstudien:

- Teilnahme am Programm (VG und KG), sozioökonomischer Status, Bildungsstand der Probanden und dergl.
- sind meist sehr leicht zu finden, weil die Kategorien selbsterklärend sind, z.B. Geschlecht, Teilnahme am Programm
- schwieriger ist es bei UVs wie Alter, sozioökonomischer Status der Eltern, komplexen Problemen.
- wichtig ist immer, die Fachliteratur zu beachten, um Variablen entsprechend dem Forschungsstand auswählen zu können.

abhängige Variablen bei Evaluationsstudien:

- Ergebnisse des zu evaluierenden Programms

1. Schritt: ***Herstellen von Beziehungen zwischen Evaluationsfragen, Standards, UVs und Ergebnissen***:

BEISPIEL:

Evaluationsfrage: Wie unterscheiden sich VG und KG hinsichtlich ihrer sozialen Aktivitäten?
Standard: Statistisch signifikanter Unterschied in den sozialen Aktivitäten, mehr soziale Aktivitäten in VG
UV: Teilnahme an VG vs. Teilnahme an KG
AV: soziale Aktivitäten
Hypothese: Teilnahme an der VG bewirkt mehr und bessere soziale Aktivitäten als Teilnahme an KG

BEISPIEL:

Evaluationsfrage: Wie unterscheiden sich Kinder der Unterschicht, die Advance-Start-Programm mitgemacht haben von Kindern der Unterschicht, die herkömmliches Programm zur Förderung von Lesen und Schreiben mitgemacht haben?
Standard: statistisch signifikanter Unterschied zwischen neuem und herkömmlichem Programm zugunsten des neuen Programms
UV: Teilnahme am neuen Programm
AV: Lese- und Schreib-Fähigkeiten
Hypothese: Teilnahme am Programm steigert die Fähigkeit zum Lesen und Schreiben

2. Schritt: ***Evaluationsfrage, Standards, UVs, Ergebnisse und Design***:

BEISPIEL:

Evaluationsfrage: In welchem Ausmaß hat sich die Lebensqualität der Kinder durch die Programmteilnahme verbessert?

Standard: statistisch signifikanter und praktisch bedeutsamer Unterschied in der Lebensqualität zwischen teilnehmenden und nicht-teilnehmenden Mädchen und Buben mit unterschiedlich komplexen Problemen.

UV: Teilnahme am Programm, Geschlecht, Alter, Probleme

AV: Lebensqualität

Design: 2 Gruppen -> VG + KG, weiter unterteilt in Knaben und Mädchen und noch weiter unterteilt in sehr große Probleme / durchschnittliche Probleme / leichte Probleme [=> 12 Gruppen-Design]

3) **Einschluss- und Ausschlusskriterien**:

Einschlusskriterien (= eligibility criteria):

- **trennen** potentielle Teilnehmer von Nicht-Teilnehmern,

- bilden die **Basis für Schlüsse** darauf, für welche Gruppen eine Teilnahme am Programm wahrscheinlich vorteilhaft sein wird.

BEISPIEL: **Evaluation soll Effektivität eines Programms erheben, das Schülern Gelegenheit zu außerschulischer Computer-Praxis bietet**. Geeignet für Schüler von 10-12 Jahren.

ABER: Programm erfordert, dass **Schüler an bestimmten Ort fahren müssen**, um teilnehmen zu können -> jetzt ist es nur mehr für jene Schüler geeignet, die zwischen 10 – 12 Jahre alt sind UND über die Möglichkeit, dorthin zu kommen verfügen. Diese Kinder können sich nun sehr von ersteren unterscheiden, weil sie weit mehr Anreiz brauchen, um an diesem Programm teilzunehmen [sie müssen ja dort hin fahren!]

- Einschlusskriterien **ergeben sich** bei einer Evaluation **primär oft aus praktischen Überlegungen, dazu gehören**:
  - geografische und zeitliche Nähe,
  - demografische und körperliche Charakteristika.

BEISPIEL: **Einschlusskriterien**

- für ein kreatives Schreibprogramm:
  - Schulen mit mindestens 5 Lehrern, die teilnehmen wollen
  - Kinder, die schon mindestens 1 Gedicht geschrieben haben

Ausschlusskriterien:

= Kriterien, die bestimmte **potentielle Teilnehmer von einer Teilnahme ausschließen**, weil **ihre Teilnahme** den Ablauf **der Evaluation stören und die Daten verzerren** würde.

BEISPIEL: **Ausschlusskriterien**

- für ein Schreibprogramm
  - Schulen, an denen das Programm weniger als 1x pro Woche durchgeführt wird
- für ein außerschulisches Beratungsprogramm:
  - Kinder mit ernsthaften Störungen wie Autistmus und Schizophrenie, die die Programmregeln nicht einhalten könnten

Einschluss- und Ausschlusskriterien ergeben sich aus den Evaluationsfragen.

BEISPIEL: **Einschluss- und Ausschlusskriterien und Evaluationsfrage**

Evaluationsfrage: Haben Schüler ihre Schreibfähigkeit im Spanischen verbessert?

Frage: Welche Merkmale müssen Schüler aufweisen, die in Evaluation einbezogen werden, um zu einem fairen Urteil über das Programm zu kommen? Für welche Schülergruppen sollen die Ergebnisse relevant sein? Wann und wo kann ich solche Familien für eine Teilnahme gewinnen?

- **Einschlusskriterien**:
  - Schüler muss in einer der Schulen sein, die das Programm durchgeführt haben
  - Schüler muss Fähigkeit spanisch zu lesen haben, die der Schulstufe entspricht

- **Ausschlusskriterien**:
  - Schüler mit Lehrern, die weniger als 1 Jahr Praxis haben
  - Schüler mit Durchschnittsnote unter 3
  - Schüler, die im letzten Jahr durchschnittlich 5 oder mehr Tage pro Monat gefehlt haben

4) **Kontrollgruppe ja oder nein**:

Untersuchungsteilnehmer = Personen, die an der Evaluation teilnehmen, egal ob sie das Treatment erhalten haben oder nicht.

Typisches Evaluationsdesign umfasst:
- ***VG***: nimmt am neuen Programm teil
- ***KG***: nimmt an herkömmlichem Programm teil

Manche Evaluationen haben mehr als 1 VG und mehr als 1 KG.

BEISPIEL: **Neues Schreibprogramm für 5. Klasse soll evaluiert werden, dabei Vergleich von 2 verschiedenen Instruktionsstilen**:
- computer-unterstützt (Versuchsbedingung A)
- in Kleingruppe (Versuchsbedingung B)

Dazu 2 KG:
- Unterricht durch einen Lehrer und mit Diskussionen (Kontrollbedingung 1)
- selbst-gesteuertes Lernen in Lernmodulen (Kontrollbedingung 2)

Das ergibt ein **4 – Gruppen – Design**

Oft wird zuwenig Wert darauf gelegt, was mit der KG passiert. Meist besteht das Programm der KG aus dem herkömmlichen Programm oder keinem.

KG muss in ihrer Zusammensetzung mit der VG vergleichbar sein (d.h. Verhalten, Wissensstand, Einstellungen, demografische Charakteristika, usw.). Die in VG und KG verwendeten Programme können sich unterscheiden in Zielen, Methoden, Settings, Dauer, Durchführung und Ressourcen.

5) **Zeitpunkt der Messungen**:

Messungen können **vor, während und nach** (unmittelbar danach oder später) **dem Programm** durchgeführt werden.

Pre-Messungen:

- werden **durchgeführt z.B.**
  - um Teilnehmer am Programm auszuwählen,
  - zur Erhebung des Bedarfs nach dem Programm,
  - zur Erhebung der Base-Line.
- **Pre-Tests sind dann von Vorteil, wenn**
  - sie nicht den Postmessungen zu ähnlich sind
  - sie nicht zeitlich zu nahe an den Postmessungen durchgeführt werden
  - sie die Aufmerksamkeit der Teilnehmer auf das Programm richten

BEISPIEL: **Pre-Test, um Gruppen auszuwählen**
Die Spanisch-Kenntnisse von Schülern werden vor einem Lernprogramm erhoben, um diejenigen herauszufinden, die dem Einschlusskriterium (Spanischkenntnisse mindestens 5. Lernjahr) entsprechen

Post-Tests

- liefern Beweise für die **Wirksamkeit eines Programms**.
- Sie **umfassen** Leistungstests, Interviews, Telefoninterviews, Fragebogen, Beobachtungen etc.

BEISPIEL: **Post-Test / Schreibprogramm**
Durchschnittliche Ergebnisse eines Schreibtests von VG und KG werden nach Teilnahme an Programm verglichen.

6) **Häufigkeit der Messungen**:

- Die **Anzahl der Messung hängt ab von**:
  - ***Zeit, die es dauert, bis Effekte des Programms wirksam werden*** (hängt ab von den erwarteten Ergebnissen des Programms, seiner Wirksamkeit und den Kosten).

    BEISPIEL: **Auswirkungen eines Schreibprogramms** auf die Leistungen der Schüler am Ende des Schuljahrs, in dem das Programm durchgeführt wurde.

    ABER:

    Auswirkungen eines Programms für gesundheits-förderliches Verhalten von Jugendlichen am Ende des Programms zu messen, wäre sinnlos, weil solche Veränderungen länger dauern.
  - ***Dauer des Programms und der Evaluation***
  - ***vorhandenen Ressourcen*** für die Messung
- **Je mehr Zeit** man **zwischen Programm und Messung** / Beobachtung verstreichen lässt, **umso mehr** Gelegenheit haben **externe Effekte, um das Ergebnis zu verzerren** (z.B. Änderungen in der Schulorganisation, Entwicklungsfortschritt, etc.)
- **Jährliche Messung, um Langzeiteffekt eines Programms abzuchecken**, kann angezeigt sein, ABER: **Effekt von Programmen hält selten länger als 3-5 Jahr**e. Um die Möglichkeit einer kontinuierlichen Evaluation zu haben, ist es sinnvoll, ein Datensammelsystem zu errichten, das kontinuierlich upgedated wird.

DESIGNS FÜR PROGRAMM – EVALUATIONEN

Prospektive Evaluation:

- **Bewertung einer Programmkonzeption**.
- Sammlung von Daten **von Beginn der Evaluation an** [bzw. schon vorher]. Prospektive Evaluation ist jene Form, bei der es die **meiste Kontrolle** gibt.

BEISPIEL: **Evaluation eines Schreibprogramms** mit der Frage: Haben die 5. Klassler durch die Teilnahme ihre Fähigkeiten verbessert. Messung: 1 Monat vor Programmstart + 1 Jahr nach dem Programm. Vergleich der Testleistungen vorher / nachher

Retrospektive Evaluation:

- = **summative Evaluation**; sind **„historische" Analysen**;
- hier ist **sehr viel dem Zufall überlassen**, um den Ansprüchen einer sorgfältigen Untersuchung der Charakteristika und Meriten des Programms genügen zu können.

BEISPIEL: **Retrospektive Studie**: Leistungstest für 5-Klassler in allen Schulen des Bezirks während der vergangenen 5 Jahre. Danach Untersuchung der Testergebnisse und Teilung in Personen mit hohen und niedrigen Scores. Dann Vergleich der beiden Gruppen darauf, ob es einen Zusammenhang zwischen dem Score und der Teilnahme an einem neuen Programm gibt.

ABER: Kann man daraus wirklich auf die Effektivität des Programms schließen? (vielleicht waren die High-Scorer von Haus aus an besser gewesen...)

- **Positive Ergebnisse** können aber **auch auf Charakteristika der Teilnehmer an der Maßnahme zurückgehen**, müssen **nicht notwendigerweise Auswirkungen eines Programms** sein (vielleicht waren bessere Teilnehmer einfach besser motiviert?).
- **Kontrolle** der Programminhalte, der Auswahl und Zuweisung zu Gruppen, der Datenerfassung ist **wichtige Voraussetzung für die Validität der Ergebnisse der Evaluation**.

KLASSIFIKATION DER PROGRAMM – EVALUATIONS – DESIGNS:

I. **Experimente**:

A. Teilnehmer werden zufällig den Gruppen zugewiesen werden (= Randomisierung, ***echte Experimente***)

B. Teilnehmer werden NICHT zufällig den Gruppen zugewiesen (= ***Quasi-Experimente***)

C. ***Self-Controls***: erfordern Pre- und Posttests, heißen „Längsschnitt-" oder „***Vorher – Nachher - Designs***"

D. ***Historische Kontrollen***: Verwendung von Daten, die an VPn aus anderen Studien erhoben wurden

II. **Beobachtungs- oder deskriptive Designs**:

A. ***Cross – Sections***: liefern deskriptive Daten zu EINEM Zeitpunkt [„Querschnitt-Untersuchung"]

B. ***Trends***: liefern Daten über eine allgemeine Population (z.B. Psychologie-Studenten)

C. ***Kohorten***: liefern Daten über eine spezifische Population (z.B. Maturajahrgang 1973)

D. ***Case Controls oder Matched Controls***: retrospektive Studien, die in die Vergangenheit zurückgehen, um ein gegenwärtiges Phänomen zu erklären

I. Experimentelle Evaluationen:

a) Echtes Experiment:

Zunächst wird eine Gruppe potentieller Teilnehmer identifiziert (z.B. Schüler der 5. Klassen AHS). Dann **randomisierte Zuweisung zu VG** (macht neues Programm) **und KG** (macht altes Programm). **Ergebnisse werden in Bezug auf ein relevantes Kriterium verglichen** (z.B. Leistung). Dieses Design heißt auch „randomized trial", „randomized controlled trial", „control trial", „true experiment".

double – blind experiment:	weder Evaluator noch Teilnehmer wissen, wer KG und wer VG ist
blinded trial:	Teilnehmer wissen nicht, welcher Gruppe sie angehören; Evaluator weiß es schon

Oft ist es logistisch oder ethisch nicht vertretbar, blinde Designs zu verwenden (könnte zu Bias führen). **Bias entsteht** auch, wenn **Programm nicht für alle**

gleich durchgeführt wird (z.B. 2 Lehrer unterrichten zwar gleiche Inhalte, aber mit unterschiedlichen Unterrichtsstilen).

Experimentelles Design lässt als einziges **kausale Schlüsse** zu. Um **Generalisierung** zu ermöglichen, experimentelles Design mit **demselben Programm an vielen Orten, mit vielen verschiedenen Teilnehmern und über viele Jahre hinweg** durchführen.

Experimentelles Design wäre das ideale, aber viele ***Faktoren sprechen gegen den Einsatz dieses Designs***:

- Mangel an Ressourcen
- Mangel an VPn, um Gruppen zu bilden
- zu wenig Geld und zu wenig Zeit, Design durchzuführen
- Daten werden oft rasch benötigt, zu kurze Zeit für die Evaluation
- ethische und andere Probleme
- bei sehr großen Evaluationen ist es sehr schwierig in allen Gruppen exakt gleiche Bedingungen für alle Personen bei Durchführung des Programms und Messung der Ergebnisse herzustellen

b) **Quasi – Experiment**:

Aufstellung einer VG und einer KG OHNE Randomisierung. Ist **leichter anzuwenden wie expertimentelles Design**. Experimentelles Design ist oft auch gar nicht sinnvoll (z.B. Evaluation eines Lehrer-Trainings: Jedes Jahr nehmen 3 Lehrer aus 6 Schulen teil; Dauer = 3 Jahre. Am Ende haben 54 Lehrer teilgenommen -> Randomisierung hier nicht sinnvoll, weil zu kleine Zahlen...)

Bias bei Quasi-experimentellen Designs:

- **Teilnehmer – Bias (Membership bias):**
  existiert schon in den bestehenden Gruppen [also in der Teil-Population...], (z.B. Schüler an der Wenzgasse vs. Schüler an der Diefenbachgasse), z.B. unterschiedliche Vorkenntnisse, Einstellungen, etc.

- **Nonresponse – Bias**:
  tritt auf, wenn Teilnehmer eingeladen werden, freiwillig an einer Evaluation teilzunehmen. Diejenigen, die teilnehmen, könnten sich von denen, die nicht teilnehmen, grundsätzlich unterscheiden (z.B. größere Offenheit gegenüber Neuem, mehr Zeit,...)

c) **Ein – Gruppen – Pretest – Posttest Design**:

Teilnehmer werden mit sich selbst verglichen; heißen auch „Vorher – Nachher Designs“.

BEISPIEL: **Evaluation eines Lehrer-Trainings**
Testung zu 3 Zeitpunkten:
- am Anfang des Jahres zur Erhebung der Vorkenntnisse
- unmittelbar nach Kursende (wie viel wurde gelernt?)
- 2 Jahre später (wie lange hat der Effekt des Trainings angehalten?

Designs mit nur 1 Gruppe sind sehr biasanfällig. Man kann auch nicht darauf schließen, dass die beobachteten Effekte auf das Training zurückzuführen sind.

Mögliche Bias:

- **Hawthorne – Effekt**:
  = Tendenz, sich aus Enthusiasmus über das neue Programm anders (bzw. besser) zu verhalten als gewöhnlich.
- Physische, emotionale, intellektuelle **Reife der Teilnehmer**
- **äußere Umstände**, **frühere Situationen**

BEISPIEL: Teilnehmer an Lehrerkurs erwerben neue Fähigkeiten und behalten diese bei. Das kann mit dem Kurs zusammenhängen, der so motivierend war wegen der Teilnahme an einem experimentellen Programm. Während des Jahres erhalten die Teilnehmer einige anregende Vorlesungen durch einen Gastprofessor. Ob das verbesserte Lehrerverhalten der Teilnehmer jetzt mit dem Programm oder dem Gastprofessor kann nicht gesagt werden.

Die **Güte von Self-Control Designs hängt auch ab von der Anzahl und den Zeitpunkten der Messungen**. Es kann vorkommen, dass ein Programm für ineffektiv gehalten wird, nur weil die Daten zu früh erhoben wurden, also zu einem Zeitpunkt, zu dem die erhofften Ergebnisse nicht gar nicht eingetreten sein können.

d) **Evaluationen mit „historischen“ Kontrollen: Verwendung von Normen**

Solche Evaluationen **stützen sich auf früher erhobene Daten aus einer anderen Quelle**. Historische Kontrollen umfassen **etablierte Normen** (z.B. Körpergröße, Gewicht, Blutdruck, usw.) und **Ergebnisse aus standardisierten Tests** (z.B. Lesetests und SATs). Das können Daten von Personen sein, die an einem anderen Programm teilgenommen haben oder auch an diesem Programm, aber in einem anderen Setting.

Die **Verwendung historischer Daten ist bequem**, **ABER: Bias, weil es keine Vergleichsmöglichkeit** gibt zwischen den VPn, an denen diese Daten erhoben wurden, und den VPn der aktuellen Evaluation.

II. **Beobachtungs- oder deskriptive Evaluationen**

a) **Überblick – Designs (Survey Designs):**

Evaluatoren benutzen solche Designs:
- um Baseline – Informationen über VG und KG zu erhalten,
- um die Programmentwicklung zu überwachen und
- als Datenquelle für die Untersuchung der Einflüsse der Umgebung auf das Programm.

Solche Designs erlauben dem Evaluator die **Untersuchung eines Querschnitts durch eine oder mehrere Gruppen zu einem bestimmten Zeitpunkt** [= ***Querschnitt-Untersuchung***]

BEISPIEL: **Workshop für Lehrer, die Schüler über AIDS unterrichten sollen**.
Untersuchung der TN an 10 solchen Workshops, die im Schulbezirk angeboten wurden. Teilnahme am Workshop war freiwillig und billig. Daten werden benutzt, um eine bezirksweite öffentliche Kampagne zu starten, um für künftige Workshops die höchst-möglichen Teilnehmerzahlen zu erreichen.

b) **Evaluation von Kohorten**:

Kohorte = Gruppe von Personen mit gemeinsamem Merkmal, die über lange Zeit hinweg Teil dieser Gruppe bleiben [z.B. Maturajahrgang 1973]; werden in Längsschnittstudien untersucht.

Bei Evaluationen besteht eine **Kohorte aus Teilnehmern an einem Programm**, die **über lange Zeit hinweg beobachtet** werden, um festzustellen, wie lange sich de Effekte des Programms gehalten haben und in welcher Weise das Programm zukünftiges Verhalten beeinflusst hat.

BEISPIEL: **Programm, das Studenten zum Lehrberuf motivieren soll**.
Teilnehmende Studenten-Kohorten werden später daraufhin untersucht, wie viele tatsächlich Lehrer geworden sind.

Kritik:

- Kohorten – Untersuchungen sind **teuer**, weil es **Langzeitstudien** sind
- **Bias aufgrund der Selektion** [d.h. diejenigen, die freiwillig teilnehmen, werden sich wahrscheinlich sehr von denen, die nicht teilnehmen unterscheiden]
- Bias aufgrund des **natürlichen Abgangs** [d.h. Person kann zu späterem Zeitpunkt nicht mehr getestet werden, weil sie schon gestorben ist...]

c) **Kombinationen: Das Salomon – 4 - Gruppen – Design**

= ein **sehr mächtiges Gruppen – Design**; Teilnehmer werden in 4 Gruppen randomisiert -> ist daher ein **echtes experimentelles Design**

BEISPIEL: **Programm zur Verbesserung der Geschichte-Kenntnisse von Schülern**

Teilnehmer werden ***per Zufall in 4 Gruppen*** geteilt:
Gruppe 1: Pretest – Programm – Posttest
Gruppe 2: Pretest – Posttest
Gruppe 3: Programm – Posttest
Gruppe 4: Posttest

Folgende ***Resultate*** kann die Evaluation ergeben:

1) Kenntnisse der Gruppe 1 beim Posttest sollen höher sein als beim Pretest
2) Kenntniszuwachs sollte in Gruppe 1 höher sein als in Gruppe 2
3) Posttest der Gruppe 3 sollte höheren Kenntnisstand ergeben als Pretest der Gruppe 2
4) Posttest der Gruppe 3 sollte höheren Kenntnisstand ergeben als Posttest der Gruppe 4

INTERNE UND EXTERNE VALIDITÄT

Design =

- ***internal valide***,
  wenn es sicherstellt, dass der Evaluator darauf vertrauen kann, dass ein **Programm in einem spezifischen experimentellen Fall effektiv** ist.

- ***extern valide***,
  wenn der Evaluator zeigen kann, dass die **Resultate des Programms auf Teilnehmer an anderen Orten und zu anderen Zeiten anwendbar** sind.

Interne Validität

= essentiell -> **wenn nicht gegeben**: es kann nicht gesagt werden, ob der Effekt mit dem Programm, einem anderen Faktor oder einem Bias zusammenhängt!

BEISPIEL: **Evaluation eines 5 – Jahres Programms zur Gesundheits-erziehung von Gymnasiasten**: Studenten könnten sich geistig und seelisch entwickeln, was die Programmeffekte beeinflussen könnte (= ***Reifung***)

Andere Risiken für die interne Validität:

- ***Geschichte***:

  Historische Ereignisse können einen **Bias in den Resultaten** einer Evaluation bewirken.

  BEISPIEL: **Landesweite Kampagne, um Mädchen vom Rauchen abzuhalten**.
  Während des Programms läuft eine beliebte TV-Serie mit einer nicht -rauchenden Protagonistin -> Evaluator kann nicht ausschließen, dass diese Serie einen Bias in seinen Ergebnissen erzeugt…

- ***Instrumentation***:

  Wenn die **Messinstrumente**, die verwendet werden, **nicht angemessen** sind, kann der Evaluator keine genauen Ergebnisse erzielen.

BEISPIEL: **Vorher – Nachher – Design**:
Posttest vor dem Pretest spricht irrtümlicherweise für das Ergebnis des Programms. Untrainierte und nachlässige Beobachter oder Tester könnten das Ergebnis zugunsten des Programms beeinflussen, während trainierte und strenge Tester das Ergebnis zuungunsten des Programms beeinflussen könnten.

- ***Abnutzung***:

  Manchmal sind die Personen, an denen evaluiert wird, aufgrund des **Drop – Outs** nicht mehr dieselben.

Risiken für die externe Validität:

- **Rekrutierung und Zuteilung der VPn zu den Gruppen**.

  BEISPIEL: Personen könnten sich in einem Experiment anders benehmen als sonst, weil sie wissen, dass sie an einem Experiment teilnehmen (= **Hawthorne – Effekt**)

- In der **Testsituation** könnten Personen sich bestimmter Verhaltensweisen erst bewusst werden und diese verstärkt / vermindert einsetzen

Vogt, W.P. (1999): Glossar

1) **A – B – A – B Design**:

Design, bei dem abwechselnd Baseline – Erhebung einer Variable und Messung dieser Variable nach dem Treatment durchgeführt wird. Wird verwendet für Einzel- oder Ein – Gruppen Experimente mit 1 Treatment und keiner KG.
A = Baseline – Erhebung, B = Treatment; auch möglich = A – B oder A – B – A.
Solche Designs werden meist angewendet, wenn es aus ethischen Gründen nicht möglich ist, eine KG vom Treatment auszuschließen (z.B. bei Therapie).

BEISPIEL: **4 wöchige Untersuchung**:
Woche 1: kein Treatment, Baseline – Erhebung
Woche 2: Treatment
Woche 3: Treatment oder keines, 2. Baseline – Erhebung
Woche 4: Treatment

2) **Vorher – Nachher – Design**:

Design, bei dem ein Pre- und ein Posttest durchgeführt wird (vgl. A – B – A – B Design)

3) **Between – Subjects – Design**:

Design, bei dem verschiedene Personen verglichen werden -> jeder Score kommt von einer anderen Person. Gegensatz dazu = Within – Subjects – Design (hier Vergleich derselben Person zu verschiedenen Zeiten oder unter verschiedenen Treatments)

4) **<u>Block – Design</u>**:

Design, in dem Personen in Kategorien oder „Blocks“ eingeteilt werden. Diese Blöcke werden dann als experimentelle oder Analyseeinheiten verwendet (vgl. ranomisiertes Block – Design)

5) **<u>Fall – Studie</u>**:

Sammlung und Analyse von Daten eines individuellen Falles, um ein breiter angelegtes Phänomen zu untersuchen. Annahme, dass dieser Fall auf irgendeine Art typisch für das Phänomen ist. Fall kann sein: Person, Stadt, Ereignis, Gesellschaft, usw.

- **Vorteil einer Fall – Studie**: erlaubt sehr intensive Analyse von spezifischen empirischen Details.
- **Nachteil**: Generalisierung der Ergebnisse ist kaum möglich.

BEISPIEL: **Fallstudie über die Eignung eines Kandidaten für ein politisches Amt**

6) **<u>Sequentielle Kohorten – Untersuchung</u>**:

Kombination von Querschnitt- und Längsschnitt – Design, wird üblicherweise angewendet, um Generationseffekte oder Kohorteneffekte aufzuspüren

7) **<u>Kohorten – Studie</u>**:

Untersuchung einer bestimmten Gruppe (Kohorte) über die Zeit hinweg, aber nicht notwendigerweise auch Untersuchung der einzelnen Individuen innerhalb der Gruppe. Gegensatz zur Panel – Studie, wo dieselben Individuen zu jedem Zeitpunkt untersucht werden

BEISPIEL: Ziehung von **Zufallsstichprobe aus der Kohorte der 1975 Geborenen** in den ersten drei Jahren, in denen sie wahlberechtigt waren -> es ist sehr unwahrscheinlich, dass hier in allen 3 Gruppen dieselben Personen untersucht werden (in jeder Gruppe sind andere Mitglieder der Kohorte enthalten)

8) **<u>Kontrolle</u>**:

Die Effekte einer Variable (= Kontrollvariable) statistisch abziehen, um zu sehen, wie der Zusammenhang ohne sie aussieht [= Herauspartialisieren]

BEISPIEL: Bei **Untersuchung über Durchschnittseinkommen von verschiedenen ethnischen Gruppen** wollen wir z.B. den Bildungsstand konstant halten. In diesem Fall könnten wir die Effekte der ethnischen Zugehörigkeit unabhängig von den unterschieden im Bildungsniveau zwischen den Gruppen messen. Das könnte wichtig sein, wenn wir zeigen wollen, wie groß die Einkommensunterschiede zwischen ethnischen Gruppen mit gleichem Bildungsstand sind.

9) **Kontrollgruppe**:

Experimentelle Gruppe, die kein Treatment erhält, um Auswirkungen Treatment ja / nein vergleichen zu können (vgl. Experimentelle Gruppe)

10) **Kontrollvariable**:

Externe Variable, die man nicht untersuchen möchte, die aber Einfluss auf das Untersuchungsergebnis haben könnte -> muss kontrolliert (d.h. konstant gehalten) werden. Heißt auch **Kovariate**.

11) **Korrelierte Gruppen – Designs**:

Design, bei dem ein Teil der Varianz der abhängigen Variable dadurch zustande kommt, dass es eine Korrelation zwischen den Gruppen gibt (oder zwischen den Gruppenscores). Meist gebräuchliche Form dieses Designs = Vorher – Nachher – Studie

BEISPIEL: 5. – Klassler bekommen einen Vokabeltest vorgelegt (= Pretest). ½ von ihnen erhält dann ein spezielles Vokabeltraining, die andere ½ ein herkömmliches Vokabeltraining. Am Ende des Semesters 2. Vokabeltest (= Posttest). AV = Ergebnis des Posttests. Viele der Unterschiede in den Scores der Schüler (= Varianz) können aus ihren Pretest – Ergebnissen erklärt werden: z.B. Schüler mit großem Wortschatz VOR dem Experiment haben noch immer größeren Wortschatz. Unabhängig vom erhaltenen Treatment werden die Resultate der Schüler im Pretest und im Posttest korrelieren (manche sogar hoch).

12) **Crossed – Factor – Design**:

= übliche Art, wie 2 oder mehr Faktoren in einem faktoriellen Design miteinander kombiniert werden. Tritt jeder Level eines Faktors gemeinsam mit einem Level des anderen Faktors auf, so sind sie vollkommen gekreuzt. Das Gegenteil davon ist „nested“ (vgl. Nested – Design)

13) **Cross – Sectional – Study**:

Studie, die nur zu EINEM Zeitpunkt durchgeführt wird, d.h. es wird ein Teil der Population zu einem bestimmten Zeitpunkt untersucht. Vgl. dazu Panel – Studie, Längsschnitt – Studie, sequentielles Kohorten – Design. Cross – Sectional Studien liefern nur indirekte Hinweise über die Effekte und müssen mit größter Vorsicht behandelt werden, wenn man daraus Schlüsse über Veränderungen ziehen will.

BEISPIEL: Eine Cross – Sectional – Untersuchung zeigt, dass Personen zwischen 60 – 65 Jahren eher rassistische Vorurteile haben als Personen mit 20 – 25 Jahren. Das bedeutet NICHT, dass junge Menschen generell weniger rassistische Vorurteile haben und es heißt auch nicht, dass die untersuchten Personen weniger Vorurteile hatten, als sie jünger waren!

14) **Deskriptive Forschung**:

= Forschungsarbeit, die Phänomene so beschreibt, wie sie existieren. Steht normalerweise im Gegensatz zu experimenteller Forschung, bei der die Umgebung kontrolliert wird und die Personen verschiedene Treatments bekommen. ABER: Auch in deskriptiver Forschung darf Inferenzstatistik verwendet werden (nicht bloß Deskriptivstatistik!)

15) **Design**:

= Untersuchungsplan, der einer Studie zugrunde liegt.

16) **Doppel – Blind – Verfahren**:

= Verfahren, um Bias in einem Experiment zu reduzieren, indem sichergestellt wird, dass weder Untersucher noch Untersuchter wissen, ob sie in der VG oder der KG sind.

BEISPIEL: **Studie zur Überprüfung der Wirksamkeit verschiedener Kopfweh – Tabletten**:

80 Kopfweh – Patienten werden zufällig 4 Versuchsbedingungen zugeteilt. Gruppe A bekommt Aspirin, Gruppe B bekommt Ibuprofen, Gruppe C bekommt Acetaminophen, Gruppe D bekommt ein Placebo. Alle Pillen schauen gleich aus. Wenn Patienten über Kopfweh klagen, bekommen sie das für ihre Gruppe vorgesehene Medikament und werden danach nach ihrem Befinden befragt. Bei echtem Doppel – Blind – Versuch wissen Auswerter nur, dass z.B. blaue Pulver besser gewirkt haben als rote Pulver, aber sie wissen nicht, um was es sich dabei gehandelt hat.

17) **Experiment**:

In empirischer Forschung angewendet; Untersucher hat Kontrolle über die Bedingungen, unter denen die Untersuchung stattfindet, und über bestimmte Aspekte der UV. Randomisierte Zuteilung der VPn zu VG und KG

BEISPIEL: kein Experiment = Befragung von Kinobesuchern nach Kinobesuch über Einstellungsänderung aufgrund des gesehenen Films
eher ein Experiment = VL wählt Film aus, randomisierte Zuteilung der VP zu VG und KG (sehen unterschiedliche Filme), dann Befragung über Einstellungen

18) **Experimentelles Design**:

= Planung und Durchführung eines Experiments. Vorteil: hohe interne Validität: Hier kann man sicherer als bei allen anderen Designs die Veränderungen auf die Variationen der UV zurückführen. Nachteil: geringe externe Validität: Ergebnisse können nicht auf Situationen außerhalb des Labors anwendbar sein.

19) **Versuchsgruppe**:

Gruppe, die in einem Experiment ein Treatment erhält. Gesammelte Daten aus der VG werden verglichen mit denen der KG (haben kein Treatment erhalten) oder mit anderer VG (haben ein anderes Treatment erhalten)

20) **Ex – post – facto – Design**:

- = jede Untersuchung, die bereits erhobene Daten verwendet, anstatt selbst Daten zu erheben. Ursachen werden nach ihren Auswirkung erforscht.

- = jede nicht – experimentelle Untersuchung, die durchgeführt wird, nachdem die zu untersuchenden Bedingungen aufgetreten sind (z.B. Untersuchung, in der es nur einen Posttest, aber keinen Pretest gibt). Manchmal versucht man, den fehlenden Pretest oder die Baseline – Erhebung dadurch zu kompensieren, dass Parallelisierung der VPn erfolgt, oder dass Variablen, die das Ergebnis beeinflussen könnten, so gut wie möglich kontrolliert werden.

21) **Faktorielles Design**:

Design mit 2 oder mehr-kategoriellen UVs (= Faktoren), von denen jeder in 2 oder mehreren Stufen untersucht wird. Ziel der faktoriellen Designs ist es, die Wechselwirkungen der Faktoren zu untersuchen. Wenn die Treatments einander nicht beeinflussen, so können die kombinierten Effekte dadurch erhoben werden, dass sie getrennt untersucht werden und dann gemeinsam mit den separierten Effekten.

BEISPIEL: **Studie über die Auswirkungen der Pubertät** (verschiedene Stadien der Pubertät) **und Drogenkonsum** (3 verschiedene Dosen). Hier könnte es eine Wechselwirkung zwischen Pubertät und Drogenkonsum geben, wenn z.B. in einem Stadium der Pubertät eine Droge als Tranquilizer wirkte, in einem anderen Stadium aber als Stimulans.

22) **Feldexperiment**:

Experiment in einer natürlichen Umgebung (im Feld), nicht im Labor. Experimentelle Bedingungen werden nicht vom Untersucher hergestellt, die UV kann aber variiert werden. BEISPIEL für ein Feld: **Schulklassen**

23) **Feldforschung**:

Untersuchung, die im realem Setting durchgeführt wird, nicht im Labor. Untersucher kann die UV weder herstellen noch manipulieren, sondern nur beobachten.

24) **<u>unabhängige Stichproben</u>**:

Gruppen von Untersuchungsobjekten, die voneinander unabhängig sind, d.h. die Messungen in einer Gruppe haben keine Auswirkung auf die Messungen in einer anderen Gruppe

25) **<u>Unabhängige Variable</u>**:

= die angenommene Ursache in einer Untersuchung; ist auch eine Variable, die verwendet werden kann, um die Ausprägungen anderer Variablen zu erklären oder vorherzusagen. Variable, die vom Untersucher manipuliert werden kann, um die Auswirkungen dieser Variation auf die abhängige Variable zu untersuchen. Wird manchmal NUR in Zusammenhang mit einem Experiment verwendet.

26) **<u>Interne Validität</u>**:

Ausmaß, in dem die Ergebnisse einer Untersuchung (normalerweise eines Experiments) den Treatments zugeschrieben werden können und nicht irgendwelchen Fehlern im Design. Interne Validität = der Grad, in dem man valide Schlüsse über die Kausaleffekte der einen Variable auf die andere ziehen kann. Hängt davon ab, wie gut externale Variablen kontrolliert werden können.

27) **Labor – Forschung**:

Jede Art von Untersuchung, bei der die Untersuchungsobjekte isoliert werden, um Umgebungseinflüsse auszuschalten. Oft synonym mit „Labor-Experiment“ gebraucht.
- **Vorteil**: keine unerwünschten Umgebungseinflüsse auf die VP; Fokussierung auf die interessierenden UVs
- **Nachteil**: Ergebnisse sind schwer zu generalisieren auf Situationen außerhalb des Labors, weil sich Personen im Labor selten so verhalten wie im richtigen Leben. **Hohe interne Validität, aber geringe externe Validität**.

28) **<u>Lateinisches Quadrat</u>**:

Methode, um in einem Within – Subject – Design Personen zu Versuchsgruppen zuzuordnen. Heißt so, weil die Treatments mit lateinischen Buchstaben (nicht mit griechischen) bezeichnet werden. Hauptziel = Vermeidung von Positionseffekten beim Durchlaufen der Versuchsbedingungen

BEISPIEL: A, B, C, D = Treatments. Es gibt 4 VP und 4 Abfolgen der Treatments. Wichtig: Lateinisches Quadrat MUSS ein Quadrat sein, d.h. **Anzahl der Zeilen und Spalten muss gleich sein**. Außerdem muss die **Anzahl der VPn gleich sein oder ein Vielfaches der Anzahl der Treatments**. D.h. hier: 4 Treatments -> 4, 8, 12, 16VPn.
Reihenfolge der Treatments für VP1: ABCD
Reihenfolge der Treatments für VP2: BCDA
Reihenfolge der Treatments für VP3: CDAB
Reihenfolge der Treatments für VP4: DABC

29) **Manipulierte Variable**:

= andere Bezeichnung für UV, heißt auch „Treatment – Variable“. Vgl. Prädiktor – Variable.

Merke:
- Experimentelle Forschung arbeitet mit manipulierten Variablen,
- Korrelationsforschung dagegen nicht

30) **Matched Pairs** [= Parallelisierung]:

Design, in dem die VP nach bestimmtem, relevantem Merkmal zu Paaren geordnet werden. Wenn die VP parallelisiert sind, wird ein Teil des Paars randomisiert der VG oder KG zugeteilt, der andere Teil kommt in die jeweils andere Gruppe. Ohne Randomisierung wird Parallelisierung nicht als gute Methode angesehen. Heißt auch „Subject Matching“

BEISPIEL: Professoren wollen die **Effektivität von 2 verschiedenen Schulbüchern** testen. Ordnen die Schüler von 2 Klassen nach den Noten zu Paaren -> eine Klasse benutzt Buch A, die andere Buch B.

31) **Natürliches Experiment**:

Untersuchung einer Situation, die im wirklichen Leben vorkommt, d.h. ohne Manipulation durch den Untersucher. Variablen treten dabei in natürlicher Weise so auf, dass sie Merkmale von Kontroll- und Versuchsgruppen haben.

Gegensatz = **artifizielles Experiment**: ist KEIN Labor-Experiment, sondern eine Umgebung, die künstlich von allen jenen Variablen befreit wurde, die den Forscher nicht interessieren. Vgl. natürliches Setting, Beobachtungsforschung, Korrelationsforschung, Quasi – Experiment.

BEISPIELE: Sonnenfinsternis bietet Astronomen die Gelegenheit, die Sonne in einer Situation zu beobachten, die sie selber experimentell nicht herstellen können.
Vergleiche der Anzahl der Kariesfälle in einer Stadt mit viel Fluor im Wasser und Stadt mit wenig Fluor im Wasser.
Vergleich des Anwachsens des Wortschatzes von Kindern während des Schuljahres und während der Ferien.

32) **Nested Design**:

faktorielles Design, in dem die Abstufungen nur in EINEM Faktor auftreten, diesen Abstufungen des einen Faktors entspricht nur EINE Stufe des zweiten Faktors. Gegenteil davon sind: gekreuzte Faktoren

33) **Nicht – orthogonales Design**:

faktorielles Desing, in dem Zellen (oder Treatments) ungleiche Anzahl haben, d.h. Designs mit ungleich vielen Häufigkeiten. Heißen auch „***unbalancierte Designs***“

34) **Panel – Studie**:

Längsschnittuntersuchung an ein- und derselben Gruppe von Untersuchungsobjekten (d.h. am Panel). Bei einer Panel – Untersuchung werden dieselben Personen zu verschiedenen Zeiten untersucht, bei einer Kohorten – Untersuchung dagegen wird die Stichprobe aus derselben Gruppe aber zu unterschiedlichen Zeiten gezogen. Panel – Untersuchung wird deshalb manchmal als „echte“ Längsschnittuntersuchung bezeichnet, während die Kohorten – Untersuchung nur eine Annäherung an eine Längsschnittuntersuchung darstellt.

BEISPIEL: Untersuchung an Schülern einer Schulklasse, die später während des Studiums noch einmal untersucht wird und dann auch noch 50 Jahre später…

35) **Posttest**

= Test oder Messung nach einem experimentellen Treatment. Vgl. Pretest

36) **Praktische Signifikanz**:

= Eine Untersuchung ist von praktischer Signifikanz, wenn sie praktische Bedeutung hat. Gegensatz dazu = reine statistische Signifikanz. Für praktische Signifikanz gibt es keine Signifikanztests. Ob etwas praktisch signifikant ist, ist Ansichtssache. In der Forschung spricht man oft von geringer praktischer Signifikanz, wenn das Ergebnis zu groß ist, um es dem Zufall zuzuschreiben, aber nicht groß genug, um von Nutzen zu sein. Manchmal kann die praktische Signifikanz aber groß sein, auch wenn sie hoch ist.

BEISPIEL: **Statistischer Zusammenhang zwischen Ausmaß an Körperübungen** und langem Leben ist sehr bescheiden, aber er ist praktisch signifikant, weil:

- langes Leben den Menschen wichtig ist
- wir direkte Kontrolle über das Ausmaß der Körperübungen haben (bei anderen Variablen, die langes Leben bewirken haben wir das nicht)

37) **Prädiktive Forschung**:

= Forschung, deren Ziel es ist, den Wert einer Variablen aus den Werten einer oder mehrerer anderer Variablen vorherzusagen (aber NICHT zu erklären!). Wird normalerweise als Gegensatz zu erklärender Forschung betrachtet (Ziel = Verstehen der Ursachen von Beziehungen). Vielfach wird von einer Dichotomie experimentelle Forschung (kann etwas erklären) und Korrelationsforschung (kann nichts erklären, sondern nur Zusammenhänge feststellen) gesprochen.

38) **Qualitativ**:

- Bei Bezugnahme auf Variablen heißt qualitativ soviel wie kategoriell oder nominal

- Bei Bezugnahme auf Forschung bedeutet qualitativ, dass es hier um Untersuchungen geht, bei denen es schwierig ist, etwas zu quantifizieren, also zu messen (z.B. Kunstgeschichte).

Bezeichnung qualitative Forschung wird auf alles, was nicht – quantitative Forschung ist, angewendet. Der Streit qualitativ vs. quantitativ ist aber oft überzogen... Es ist schwierig, quantitative Elemente in qualitativer Forschung zu vermeiden und umgekehrt

39) **Quantitativ**:

bezeichnet Variablen oder Forschung, bei der mit numerischem Material gearbeitet wird.

40) **Quasi-Experiment**:

Forschungsdesign, um Studien im Feld oder in realer Umgebung, wo der Untersucher eventuell einige UVs variieren kann, wo es aber nicht möglich ist, Personen in VG und KG zu randomisieren. Viele der Methoden für Quasi-Expermente wurden in Evaluationsprojekten entwickelt.

41) **Randomisierte Block – Designs**:

Forschungsdesign, in dem Untersuchungsobjekte in einer Variable, die der Untersucher kontrollieren will, parallelisiert werden. Die Objekte werden in Gruppen eingeteilt (= Blöcke), die Anzahl der Gruppen entspricht der Anzahl der Treatments. Die Personen in jedem Block werden randomisiert den verschiedenen Versuchsbedingungen zugeordnet. Vgl. Lateinisches Quadrat; Varianzanalyse für Messwiederholungen.

BEISPIEL: **Studie über die Effektivität von 4 Methoden des Statistik-unterrichts**. 80 VPn sollen in 4 VG mit jeweils 20 Personen eingeteilt werden. Bei Verwendung eines randomisierten Block – Designs wird den Personen ein Test zur Erhebung ihrer Statistik – Kenntnisse vorgegeben. Die besten 4 im Test sind im 1. Block, die nächsten 4 im 2. Block, usw. bis zum 20. Block. Die 4 Personen werden jeweils einem Block per Zufall zugeteilt, und zwar eine Person pro Versuchsbedingung. Die Blöcke dienen dem Ausgleich der Varianz innerhalb jeder Versuchsgruppe, indem so gewährleistet ist, dass in jeder VG Personen mit ähnlichem Vorwissen über Statistik sind.

42) **Design mit Messwiederholung**:

Forschungsdesign, in dem die AV 2 oder mehrmals gemessen werden. Meist werden den VPn mehrere Treatments gegeben und nach jedem gemessen, anstatt für jede Abstufung des Treatments verschiedene VPn heranzuziehen. Das bedeutet, dass jede VP mit sich selbst verglichen wird.

43) **Forschungsdesign**:

= Plan für die Durchführung einer Untersuchung, um die höchst mögliche Validität der Ergebnisse zu erhalten. Wird meist als „Design“ bezeichnet. Beim Planen einer Untersuchung entwirft man ein Set von Regeln, wie man Daten erheben und interpretieren wird.

44) **Forschungsstrategie**:

Genereller Plan zur Durchführung einer Forschungsarbeit. Beispiele für Forschungsstrategien sind experimentelle, Korrelations-, Längsschnitt- und kombinierte Methoden. Üblicherweise wird zwischen Forschungsstrategie und Forschungsdesign unterschieden -> Forschungsdesign = spezifischerer und detaillierterer Plan zur Durchführung einer Untersuchung.

45) **Überblick (Survey):**

Forschungsdesign, in dem eine Stichprobe aus einer Population gezogen und untersucht wird, um Schlüsse über die Population ziehen zu können. Design wird oft als Gegensatz zum echten Experiment angesehen, in dem die Untersuchungsobjekte den Versuchsbedingungen oder Treatments zufällig zugewiesen werden.

46) **Treatment**:

= in einem Experiment das, was die VL mit den Personen der VG machen, nicht aber mit den Personen der KG. Treatment ist somit eine UV. Wird außerdem oft verwendet als weit gefasste Bezeichnung jeglicher Prädiktor - Variable.

BEISPIEL: **Studie über den Effekt von Verkehrsunfällen auf Geschwindigkeitsbeschränkungen** -> Treatment wäre hier die Geschwindigkeitsbegrenzung.

47) **Unbalancierte Designs**:

So nennt man faktorielle Designs, wenn es ungleiche Zahlen für die Beobachtungen verschiedener Faktoren gibt oder wenn die Zellen ungleiche Anzahlen von VP enthalten. Heißen auch „nicht – orthogonale faktorielle Designs“.

48) **Within – Subject – Designs**:

= eine Vorher – Nachher – Untersuchung oder eine Untersuchung derselben VP, die verschiedene Treatments erhalten [d.h. eine VP bekommt alle vorgesehenen Treatments bzw. durchläuft alle Versuchsbedingungen]. Forschungsdesigns, bei dem innerhalb derselben Gruppe Pre- und Posttest durchgeführt wird; Design OHNE Kontrollgruppe.

Überblick über die Evaluations-Designs (Vor- und Nachteile):

Design:	***Vorteile:***	***Nachteile:***
echtes Experiment:	es kann festgestellt werden, zu welchem Grad das Programm das Ergebnis kausal bewirkt	logistisch und methodisch schwierig in der Durchführung
Quasi – Experiment:	keine Randomisierung, daher leichter durchführbar als ein echtes Experiment	jede Menge Bias kann auftreten, wegen: - nicht jeder hat gleiche Chance teilzunehmen - systematische Unterschiede zwischen TN - systematische Unterschiede zwischen VG und nicht teilnehmenden Gruppen
Self – Control oder Pre- und Posttest:	logistisch ziemlich leicht durchführbar liefert stichprobenabhängige Daten	Es muss sichergestellt werden, dass die Messungen zeitgerecht durchgeführt werden. Ohne KG kann Evaluator nicht sagen, ob anscheinend experimentelle Effekte in der VG auch in der KG feststellbar wären oder nicht.
Historische Kontrollen oder Normen:	leicht durchzuführen, keine Belästigung von VPn	Es muss sichergestellt werden, dass „normative“ Daten zum Vergleich herangezogen werden können
Querschnitt – Untersuchungen (Überblicks-Designs):	liefert Baseline - Infos über die Teilnehmer an der Evaluation und deskriptive Information über das Programm, sein Setting und seine Ressourcen	Liefert ein statisches Bild von Teilnehmern und Programm zu EINEM Zeitpunkt

Design:	***Vorteile:***	***Nachteile:***
Kohorten – Untersuchungen:	Liefert Längsschnitt-Information oder Follow – Up – Information	Kann sehr teuer sein, weil meist (sehr) Langzeitstudie. Teilnehmer können sich über die Zeit hinweg unterschiedlich entwickeln; viele Ausfälle
Salomon – 4 – Gruppen:	strenges Design, das kausale Schlüsse ermöglicht. Schützt gegen die Effekte des Pretests auf den Posttest	Man braucht genügend Teilnehmer, um 4 Gruppen bilden zu können; sehr teuer in der Durchführung.