Forschungsmethoden

Mitschrift der Vorlesung von
Dipl.-Psych. Ingo Totzke
im WS 06/07

Roland Pfister

Bayerische Julius-Maximilians-Universität Würzburg

Inhaltsverzeichnis

0. Vorwort

Dozent: Dipl.-Psych. Ingo Totzke

Termin: Montag, 11:00 – 12:30, Külpe Hörsaal

Klausur: 22.01.07

KVK: 30.04.07

Web: http://www.psychologie.uni-wuerzburg.de/methoden/lehre/grundstudium/skripten.php.de
User: student
PW: Fahrtauglichkeit

1. Einführung

1.1. Methoden

Die Auffassung der Psychologie als empirische Wissenschaft impliziert spezifisch methodisches Vorgehen. Nicht die Objekte, Themen und Ergebnisse eines Handelns machen eine Wissenschaft aus, sondern das „Wie“ des Handelns, also der Umgang mit wissenschaftlichen Standards (Methoden).

1.1.1. Definition: Methoden

(1) Methoden bestehen aus Regeln bzw. Systemen von Regeln, nach denen zielgerichtet gehandelt werden kann. Die Anwendung einer Methode hat dabei zumeist einen genau feststellbaren Beginn und ein genau feststellbares Ende. Methoden müssen jedoch nicht explizit als Regeln formuliert sein.
(2) Methoden enthalten Festlegungen darüber, wie die Regeln und ihre begrifflichen Bestandteile verstanden werden sollen.
(3) Methoden sind mitteilbar.
(4) Methoden haben normativen oder präskriptiven (vorschreibenden) Charakter. Die Befolgung der Regeln wird erwartet, die Verletzung einer Regel wird sanktioniert.
(5) Methoden können in einem hierarchischen Verhältnis zueinander stehen.

1.1.2. Variabilität der Methodenanwendung

Die Anwendung von Methoden ist dabei nicht als starre Abfolge eines methodenspezifischen Handlungsplans zu betrachten, der Schritt für Schritt abgearbeitet werden muss.

Methoden sind vielmehr als adaptiv, regulativ und reflexiv anzusehen:

- Adaptation: Anpassung des Handelns an spezifische Bedingungen
- Regulation: Regelmäßige Bewertung des Handelns
- Reflexion: Bewertung der methodenspezifischen Regeln

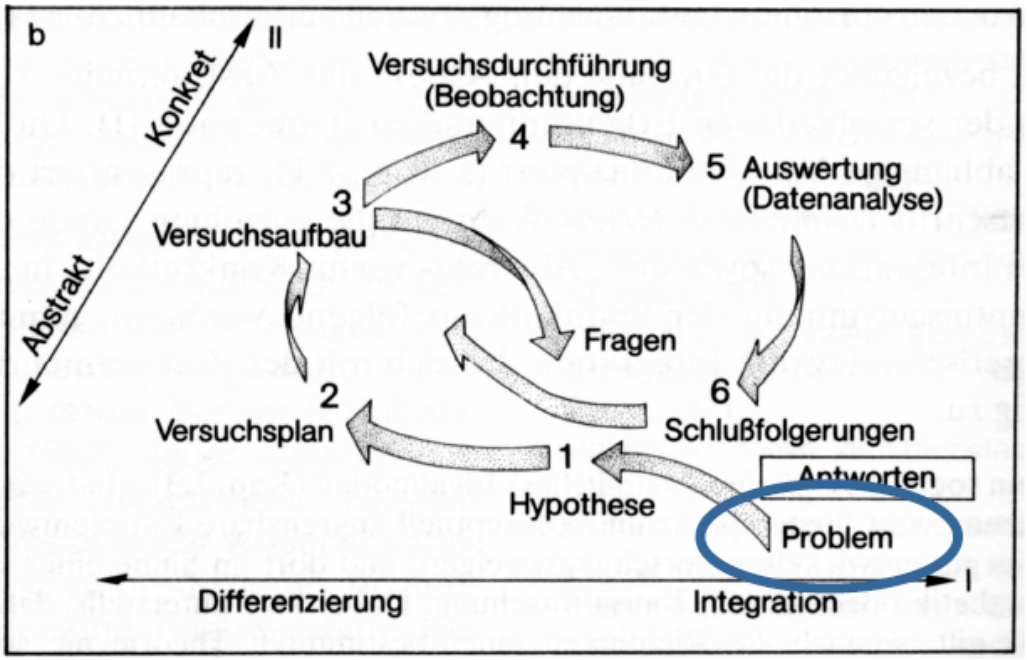
1.1.3. Zusammenfassung

Methodenanwendung ist also eine, aufgrund einer (Anwendungs-) Entscheidung erfolgende, Steuerung des zielgerichteten Handelns durch ein Regelsystem, das im jeweiligen Handlungsplan der Akteure repräsentiert und verfügbar ist. Diese methodenspezifische Handlungssteuerung ist regulativ, weitgehend adaptiv und reflexiv.

1.2. Forschung als Spiralenmodell

Sarris beschreibt den Vorgang der Forschung als Spiralenmodell, bei dem, ausgehend von einem Problem, Hypothesen generiert werden zu denen ein Versuch entworfen wird (Plan & Aufbau). Dieser Versuch liefert Daten, deren Auswertung Schlussfolgerungen bzgl. der anfänglichen Hypothese zulassen.

Im Folgenden soll auf die einzelnen Elemente des Forschungsprozesses näher eingegangen werden.



1.2.1. Problem

Das Ziel der ersten Phase ist die Problempräzisierung, also die Definition von Begriffen und die Auswahl von zu erfassenden Variablen sowie die Formulierung von Hypothesen.

In die ersten Überlegungen muss auch der Aufwand – und etwaige Kosten – mit einbezogen werden, um ein Problem später auch wirklich untersuchen zu können. Es gilt, eine Balance zwischen Machbarkeit und Belanglosigkeit zu finden.

Als initiale Problemstellung kann entweder eine Frage zur dimensionalen Analyse stehen oder primär eine semantische Analyse angestrebt werden.

1.2.1.1. Dimensionale Analyse

Die dimensionale Analyse ist vor allem bei deskriptiven Untersuchungen von Bedeutung. Es geht darum, neue Erkenntnisse zu gewinnen und damit neue Zusammenhänge zu entdecken („Licht ins Dunkel bringen“).

1.2.1.2. Semantische Analyse

Analyse des Sinns von Begriffen, die in Theorien und / oder Hypothesen verwendet werden, sowie die Verknüpfung dieser Begriffe mit realen Sachverhalten (hypothetisch).

Die semantische Analyse ist insbesondere bei theorie- und hypothesentestenden Untersuchungen von Bedeutung.

1.2.1.3. Variablen

Um Abhängigkeiten oder Zusammenhänge zwischen einer Bedingung und einem Folgeereignis zu untersuchen, müssen qualitativ und quantitativ veränderbare Variablen als Symbole für eine Menge von Merkmalsausprägungen definiert werden. Erst hierdurch lassen sich Hypothesen über die Art des Abhängigkeitsverhältnisses oder Zusammenhangs generieren.

Man unterscheidet dabei verschiedene Arten von Variablen:

Unhabhängige Variable (UV):

- Vom Versuchsleiter direkt oder indirekt verändert (durch Manipulation oder Selektion). Beispiel: UV: „Geschlecht“, Abstufungen „♀,♂“.
- Synonym: Reizvariable

Abhängige Variable (AV):

- Ereignis, das die Folge der Manipulationen der UV signalisiert
- Störeinflüsse sind wahrscheinlich
- Der Versuchsleiter hat auf die AV keinen direkten Einfluss
- Synonym: Reaktionsvariable

Wenn sich Abstufungen der UV systematisch auf die AV auswirken, also aus der Veränderung der UV eine Veränderung der AV folgt, so kann die UV als Ursache, die AV als Wirkung angesehen werden. Man kann demnach einen funktionalen Zusammenhang postulieren:

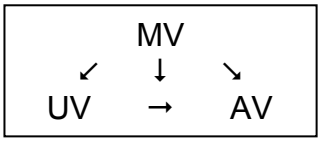
$$AV = f(UV)$$

Moderierende Variable:

Moderierende Variablen beeinflussen die Wirkung der UV auf die AV, indem sie sich entweder systematisch auf UV, AV oder das gesamte Wirkgefüge auswirken. Moderierende Variablen sind so gut wie immer vorhanden, wenn mit menschlichen Probanden gearbeitet wird.

$$AV = f(UV, \text{moderierende Variable})$$

Moderierende Variablen können sich entweder als Kontroll- oder Störvariablen manifestieren.



Kontrollvariable:

Moderierende Variablen werden zu Kontrollvariablen, wenn sie bei der Untersuchung miterhoben werden, also in eine Art unabhängige Variable umgewandelt werden.

Störvariable (SV):

Moderierende Variablen werden zu Störvariablen, wenn sie nicht beachtet oder sogar übersehen werden. Mittels bestimmter experimenteller Techniken, wie z.B. Randomisierung, lassen sich jedoch auch Störvariablen kontrollieren.

1.2.1.4. Klassifikation von Variablen

Variablen können anhand verschiedener Kriterien unterschieden werden:

- Stellenwert für die Untersuchung: UV, AV, moderierende Variable, Kontrollvariable, Störvariable
- Art der Merkmalsausprägungen: diskret (diskontinuierlich, qualitativ; dichotom vs. polytom; natürlich vs. künstlich) und stetig (quantitativ)
- Empirische Zugänglichkeit: manifeste Variable (direkt beobachtbar) vs. latente Variable (nicht beobachtbar, liegt einer manifesten als hypothetisches Konstrukt zugrunde)

1.2.2. Hypothese

„Eine wissenschaftliche Hypothese formuliert eine **Beziehung** zwischen zwei oder mehr Variablen, die für eine **bestimmte Population** vergleichbarer Objekte oder Ereignisse gelten soll.“ Mann kann zudem zwischen verschiedenen Arten von Hypothesen – inhaltlichen und statistischen – unterscheiden.

1.2.2.1. Kriterien

Um von einer wissenschaftlichen Hypothese sprechen zu können, müssen bestimmte Kriterien – Generalisierbarkeit, Konditionalsatz und Falsifizierbarkeit – erfüllt sein (für Beispiele siehe 10.1):

- Generalisierbarkeit: Eine wissenschaftliche Hypothese ist eine allgemeingültige, über den Einzelfall oder ein singuläres Ereignis hinausgehende Behauptung (All-Satz).
- Konditionalsatz: Einer wissenschaftlichen Hypothese muss zumindest implizit die Formalstruktur eines sinnvollen Konditionalsatzes (Wenn-dann, bzw. Je-desto) zugrunde liegen.
- Falsifizierbarkeit: Der Konditionalsatz muss potenziell falsifizierbar sein, d.h. es müssen Ereignisse denkbar sein, die dem Konditionalsatz widersprechen.
  Kann-Sätze sowie Existenzsätze (Es gibt…) sind demnach niemals Hypothesen, da sie weder einen Konditionalsatz implizieren, noch falsifizierbar sind.

1.2.2.2. Inhaltliche Hypothesen

Inhaltliche Hypothesen sind verbale Behauptungen über kausale / nichtkausale Beziehungen zwischen Variablen. Sie werden aus begründeten Vorannahmen, Modellen oder Theorien abgeleitet.

1.2.2.3. Statistische Hypothesen

- Zuspitzung der inhaltlichen Hypothese zu einer empirischen Vorhersage des Untersuchungsergebnisses.
- Formulierung von statistischen Aussagen bezogen auf Maße, die eine inhaltliche Aussage am besten wiedergeben.
- Definition: Annahmen über Verteilungen einer oder mehrerer Zufallsvariablen oder eines (mehrerer) Parameter dieser Verteilung.
- Statistische Hypothesen sind nicht deterministisch, sondern probabilistisch: Hypothesen sind Wahrscheinlichkeitsaussagen.

1.2.3. Versuchsplan, -aufbau und -durchführung

1.2.3.1. Operationalisierung

Die Versuchsplanung enthält immer eine Operationalisierung, also die Umsetzung der Problempräzisierung (Begriffe) in Techniken bzw. Forschungsoperationen. Die Operationalisierung enthält technische Anweisungen, wie in der Untersuchung vorzugehen ist, um Informationen zu erhalten.

Sie beinhaltet Angaben zur Gestaltung des Messinstruments (z.B. Fragebogen) sowie zu dessen Handhabung (z.B. Ort des Interviews, Reihenfolge der Fragen, Formulierungen).

Während der Operationalisierung werden einige weitere Entscheidungen getroffen, z.B. bezüglich:

- Forschungsform (z.B. Labor- vs. Feldforschung)
- Versuchsgruppen (z.B. Stichprobe, Probandenmerkmale)
- Datenquelle (z.B. Befragung, Beobachtung, objektive Verfahren)
- Versuchsplan (z.B. experimentell vs. korrelativ)

Ein Beispiel für Operationalisierung stellt eine hypothetische Studie zum Selbstwertverlust nach Hysterektomie (Gebärmutterentfernung) dar. N = 30 Versuchspersonen (Versuchsgruppen) sollen direkt nach der Operation befragt und mit einer Kontrollgruppe verglichen werden (Gestaltung und Handhabung des Messinstruments; Datenquelle). Man spricht in diesem Fall auch von einem One-Shot-Case-Study- oder Schrotschuss-Design (1 Zeitpunkt, 1 Stichprobe).

1.2.3.2. Experiment und korrelative Studien

Experimentelle Methoden und korrelative Ansätze lassen sich nicht strikt voneinander trennen und auch nicht pauschal bestimmten Teilgebieten der Psychologie zuordnen.

1.2.4. Messung: Mögliche Störfaktoren

Situation:

- Untersuchungsort (z.B. Flirtverhalten: Disko vs. Labor)
- Untersuchungszeit (z.B. Flirtverhalten: 7 Uhr morgens vs. abends)
- Atmosphäre (Leistung vs. Erleben, Technik, Ordnung, weißer Mantel)

Versuchsperson:

- Motivation („Intelligente VP“: ich glaube den Versuch verstanden zu haben und handle entsprechend; VP-Stunden: kein Interesse, „gute VP“ und Demand-Effekte: ich glaube, ich weiß, um was es in dem Versuch geht und möchte eine gute VP sein, soziale Erwünschtheit)
- Erwartung: Placebo (daher mindestens Einfachblindversuch: entweder VP oder VL weiß nicht um was es geht, welche Bedingung und welches Treatment anstehen)
- Prozesse in der VP: Aktivation, Ermüdung, Lernen, Übung

Versuchsleiter:

- Erwartung: Rosenthal-Effekt (Self-fulfilling prophecy); daher Doppelblindversuch: VP und VL wissen nicht um was es geht
- VP-VL-Interaktion: Sicherheit, Nervosität, Mann-Frau (Eiswasserversuch und attraktive vs. unattraktive VLin: Männer halten ihre Hand bei attraktiver VLin wesentlich länger im Wasser)

Kontrolle von VL-Artefakten:

- Standardisierte Instruktionen
- Konstante Untersuchungsbedingungen (z.B. Beleuchtung, Geräusche, Temperatur)

- Selbstkontrolle des VL (z.B. auf eigene Stimmungen achten und ggf. protokollieren)
- Verwendung sog. blinder Versuchsleiter
- Einhalten des zeitlichen Ablaufs
- Untersuchungsleiter soll Vorerhebung selbst durchführen
- Nachbefragung nach Beendigung des Hauptteils des Versuchs
- Aufzeichnung des gesamten Versuches per Video
- Abweichungen vom geplanten Ablauf in Untersuchungsprotokoll festhalten (z.B. mögliche Zwischenfragen der Probanden)

Die zentrale Rolle, um diese Störfaktoren zu eliminieren oder konstant zu halten, nimmt die Instruktion ein. Die Instruktion umfasst nicht nur den verbalen Anweisungsteil, sondern alle Versuchsumstände wie die Umgebung oder das Verhalten des VL.

1.2.5. Messung: Gütekriterien

Abhängig von der Art der Untersuchung und Messung gibt es eine Vielzahl von Gütekriterien. Es lassen sich jedoch 3 Hauptgütekriterien finden, die für alle Arten von Messungen gelten: Validität (Grad der Genauigkeit, das zu messen was gemessen werden soll), Reliabilität (Grad der Genauigkeit, mit dem etwas – egal was – gemessen wird) und Objektivität (Grad der Unabhängigkeit der Ergebnisse vom Untersucher).

1.2.5.1. Objektivität

Durchführungsobjektivität: Unabhängigkeit der Ergebnisse von zufälligen oder systematischen Verhaltensvariationen des Untersuchers während des Versuchs (z.B. VL-Effekte).
Auswertungsobjektivität: Unabhängigkeit der Ergebnisse von Variationen des Untersuchers während der Auswertung; insbesondere bedeutsam bei Verfahren mit vielen Freiheitsgraden wie projektive Tests oder freie Interviews.
Interpretationsobjektivität: Unabhängigkeit der Ergebnisse von der interpretierenden Person; besonders dann wichtig, wenn die Ergebnisse mehrere Schlüsse zulassen.

Alle Arten der Objektivität sind statistisch prüfbar über Korrelationen der Ergebnisse verschiedener Untersucher (z.B. Inter-Rater-Korrelation).

1.2.5.2. Reliabilität

Die klassische Testtheorie nennt 5 Axiome für eine reliable Messung:

Axiom 1: $X = T + E$
: Ein Testergebnis X setzt sich additiv zusammen aus dem wahren Wert T und einem Messfehler E.

Axiom 2: $\mu (E) = 0$
: Bei Messwiederholung kommt es zu Fehlerausgleich.

Axiom 3: $\rho (T,E) = 0$
: Wahrer Wert und Messfehler sind unabhängig voneinander; die Korrelation zwischen den beiden ist 0.

Axiom 4: $\rho (T`,E) = 0$
: Messfehler und Ausprägungsgrade anderer Merkmale sind unabhängig voneinander.

Axiom 5: ρ (E1, E2) = 0
Verschiedene Messfehler sind unabhängig voneinander.

Die Reliabilität ist also umso höher, je kleiner der zu einem Messwert X gehörende Fehleranteil E ist. Absolute Reliabilität verlangt X = T und E = 0. Bei der Mehrfachmessung zeitstabiler Merkmale müsste also gelten X1 = X2 = X3. Dieser Idealfall tritt in der Realität kaum auf, da sich Fehlereinflüsse nie ganz ausschließen lassen (technische, menschliche, situative Fehlerquellen). Daher ist es nahe liegend, die Reliabilität als Anteil der wahren Varianz an der beobachteten Varianz zu definieren.

Man unterscheidet zudem 3 Arten der Reliabilität:

Paralleltest-Reliabilität: Vergleichbare Paralleltests werden identischen Stichproben vorgegeben und deren Ergebnisse miteinander korreliert.
Retest-Reliabilität: Ein und derselbe Test wird einer Stichprobe mehrmals vorgegeben und die Ergebnisreihen miteinander korreliert.
Innere Konsistenz: Ein Test wird in zwei oder mehr Teile geteilt und die Reliabilität über Aufgabenschwierigkeit und Trennschärfekoeffizienten bestimmt (Methode der Konsistenzanalyse; Odd-even-split-half-Reliabilität).

Auch die Reliabilität ist somit ein statistisch prüfbares Kriterium.

1.2.5.3. Validität

Inhaltliche Validität:

- Genauigkeit, mit der ein zu untersuchender Inhalt (z.B. Persönlichkeitsmerkmal, Verhaltensweise) gemessen wird.
- Das gewählte Verfahren ist die optimale Möglichkeit, um Inhalt zu erfassen (Maschinenschreibtest).
- Bestimmungsmaß: Expertenrating

Konstruktvalidität:

- Genauigkeit, mit der ein zu untersuchendes Konstrukt (z.B. Eigenschaft, Fähigkeit) gemessen wird.
- Bestimmungsmaß: Expertenrating

Kriterienbezogene Validität:

- Genauigkeit, mit der ein untersuchter Aspekt mit einem unabhängig vom Test erhobenen Außenkriterium übereinstimmt.
- Bestimmungsmaß: Korrelation des Testergebnisses mit Außenkriterium

1.2.6. Interne Validität (Campbell & Stanley, 1963)

Annahme: Manipulationen der UV bedingen Veränderungen der AV. Dabei gilt es, den Einfluss von Störvariablen zu kontrollieren.

AV = f (UV, SV)

Ein Versuch ist intern valide, wenn Veränderungen der AV eindeutig auf Variationen der UV zurückzuführen sind.

Die Entscheidung, welche Veränderungen tatsächlich auf ein spezifisches Treatment zurückzuführen sind, ist in der Realität nur schwer möglich, da die interne Validität stark durch den Faktor Zeit beeinflusst wird:

- Geschichtlichkeit: Vom Untersucher unabhängig und genereller Effekt
- Entwicklung: Vom Untersucher unabhängig und spezieller Effekt
- Selektion und Messeffekte: Vom Untersucher abhängig und genereller Effekt
- Testeffekte: Vom Untersucher abhängig und spezieller Effekt

1.2.6.1. IV und Geschichtlichkeit

Die interne Validität wird durch geschichtliche Ereignisse beeinflusst, die leicht übersehen werden können, wie z.B. Kohorteneffekte oder auch „besondere Jahre“.

Ein Beispiel für ein besonderes Jahr sind die Sonntagsfahrverbote zu Zeiten der Ölkrise, die deutlich verringerte Unfallzahlen verursachten.

Ein Beispiel für Kohorteneffekte (Generationseffekte) ist die Untersuchung zur Abnahme der kognitiven Leistungsfähigkeit mit steigendem Alter, welche auch auf besondere Lebensbedingungen (II. Weltkrieg) oder auch unpassend normierte Testfragen zurückzuführen sein könnte.

1.2.6.2. IV und Entwicklung

Die Entwicklung bezieht sich auf Vorgänge während der Untersuchung. Sie ist nicht nur bei Längsschnittuntersuchungen, sondern auch bei kurzen Experimenten zu bedenken.

Beispiel I: Regelung

Starke Störungen am Anfang der Untersuchung (aktivierter Fahrradfahrer vs. langsamer Fußgänger) führen zu starken Gegenregulationen:

- Regressionseffekt B (negative Rückkopplung)
- Biologische Systeme regeln einen optimalen Systemzustand ein
- Ausgangswertgesetz von Wilder: Negative Korrelation zwischen Ausgangswert und Veränderungswert

Beispiel II: Entwicklungseffekte

Individualentwicklung: Spontanremission bei Therapie liegt bei 60%; Pbn werden müder, hungriger, lustloser
Mortalität: Stichprobe wird gesünder, je älter sie wird (Kranke sterben)

1.2.6.3. IV und Selektion und Messeffekte

Beispiel I: Regressionseffekt A

Werden Extremgruppen untersucht, tendieren die Ergebnisse bei erneuter Messung zur Mitte.

Beispiel II: Änderung der Messinstrumente

Messfühler verstellen sich, Beobachter ermüden.

1.2.6.4. IV und Testeffekte

Beispiel I: Lernen aus vorhergehender Untersuchung

Der IQ ist bei einer zweiten Untersuchung etwa 3-5 Punkte höher, Persönlichkeitstests beim zweiten Mal in Richtung stärker angepasst.

Beispiel II: Residualeffekte im Cross-Over

Die Wirkung einer Behandlung ist (trotz Cross-Over) stets durch personenbedingte Störeinflüsse verunreinigt, da Treatment 1 auch nach dem Ende von Treatment 2 nachwirken kann. Ein typisches Cross-Over-Design sieht folgendermaßen aus:

Gruppe	T1	T2
X1	A	B
X2	B	A

Beispiel III: Experimentelle Mortalität und Drop-out

Pbn scheiden systematisch aus der Versuchsplanung aus. Die 6-Monats-Katamnese von therapeutischen Interventionen wird z.B. nicht immer durchgeführt. Dabei scheiden vor allem die Klienten mit Nebenwirkungen, Rückfällen und ähnlichem aus.

1.2.7. Externe Validität

Die externe Validität beschreibt die Übertragbarkeit der Ergebnisse auf Nicht-Stichproben, also die Generalisierbarkeit der Ergebnisse. Diese ist insbesondere dann problematisch, wenn unter realen Bedingungen noch andere Faktoren eine Rolle spielen als in der Untersuchung.

Interne Validität ist eine notwendige, jedoch nicht hinreichende Bedingung für externe Validität.

1.2.7.1. EV und Störvariablen (Campbell & Stanley, 1963)

Reaktive Effekte der Experimentalsituation:

- Veränderung des Verhaltens allein durch die Situation
- z.B. Sozial erwünschte Antworten in Interviews, Reaktanz der Pbn, Demand-Effekte

Interaktion von Vortest und UV:

- Durch Vortest kann die Sensitivität der Pbn gegenüber der UV beeinflusst werden und somit das Verhalten im Haupttest verändert werden.
- z.B. Kurzinterview zur Vorauswahl einer Stichprobe, anschließend experimentelle Untersuchung

Einflüsse bei Mehrfachmessungen, z.B. sukzessive Einnahme verschiedener Medikamente.

Interaktion von Selektionseffekten und UV

- Fehler bei Selektion können zur Konfundierung der Ergebnisse mit den durch die UV bedingten Veränderungen der AV führen
- Z.B. Wirkung von Nikotin nur in Abhängigkeit vom Rauchverhalten; kogn. Training nicht nur mit intelligenteren Pbn.

1.2.8. Datenanalyse

Während der Datenanalyse werden entsprechende Verfahren zur statistischen Beschreibung der erhaltenen Rohwerte ausgewählt und durchgeführt.

Nach der initialen Datenkontrolle wird also auf Basis der Fragestellung ein angemessenes Verfahren ausgewählt, wobei auch immer darauf geachtet werden muss, ob man die zur Verfügung stehenden Verfahren auch tatsächlich durchführen kann („Verstehe ich die Cluster-Analyse?“).

Nach diesen Vorüberlegungen kann die Datenanalyse erfolgen. Sie gliedert sich dabei in vier Schritte, von denen die ersten drei eigentlich schon vor der Durchführung der Untersuchung bekannt sein müssten:

- Kodierung und Datenübertragung: Die gewonnenen Daten müssen so kodiert werden, dass sie statistisch verarbeitbar sind (z.B.: ♀ = 0, ♂ = 1) und in ein Statistikprogramm geladen werden.
- Fehlerkontrolle und evtl. Fehlerbereinigung: Nicht nur Mittelwerte und Standardabweichungen sollten hier beachtet werden, sondern es sollte auch eine Kontrolle der individuellen Rohdaten erfolgen, um beispielsweise Kodierungsfehler zu entdecken (55 auf 5-stufiger Skala).
  In diesem Schritt werden auch fehlerhafte oder fehlende Angaben (Missing Values) aus der Datenanalyse ausgeschlossen.
- Umformung und Neubildung von Variablen: Variablen müssen im Sinne der gewählten Verfahren umkodiert werden (Variablentransformation), wodurch neue Variablen, Indizes und Skalen entstehen.
- Statistische Analyse

1.2.8.1. Deskriptive Statistik

- Kennwerte der zentralen Tendenz (Lage)
  - Mittelwert (arithmetisch, geometrisch, harmonisch)
  - Median
  - Modus
- Kennwerte der Dispersion (Variablilität)
  - Standardabweichung (Varianz)
  - Bereichsmaße (Streubreite, Interquartilbereich…)
  - Informationsmaß h (Entropie)
- Kennwerte der Schiefe (Abweichung von der Symmetrie)
- Kennwerte des Exzess (Steilheit, Gipflichkeit, Kurtosis)

1.2.8.2. Signifikanztests

Der typische Signifikanztest hat vor allem zwei Funktionen:

- Screening-Prozedur: Wo lohnt es sich nachzuschauen, bzw. genauer hinzuschauen?
- Zufallskritische Absicherung: Ist das Ergebnis auch bei zufälliger Zuweisung wahrscheinlich, oder ist es sehr unwahrscheinlich (signifikant)?
- Prüfkriterien werden gebraucht, weil (psychologische) Hypothesen Wahrscheinlichkeitsaussagen sind und somit nicht durch ein einziges hypothesenkonträres Ereignis falsifiziert werden können.

1.2.9. Interpretation

In der Interpretation sollen die Befunde erklärt werden, um zu einer theoretischen Aussage zu gelangen, die auf den Ergebnissen aufbaut. Bei unerwarteten Ergebnissen sollen dabei mögliche Ursachen, wie fälschliche Annahmen, Fehler in Versuchsaufbau, -durchführung oder -auswertung, diskutiert werden. Streng genommen erscheint eine explorative Datenauswertung ebenfalls in der Diskussion, häufig wird diese jedoch am Ende des Ergebnisteils aufgeführt.

Wichtig: Die Interpretation/Diskussion stellt einen äußerst wichtigen und eigenen Schritt dar, und nicht nur eine grobe Zusammenfassung der Ergebnisse.

Häufige Mängel der Interpretation sind:

- Fehler in vorhergehenden Versuchsstadien bleiben unbemerkt bzw. unkorrigiert und werden so verschleppt.
- Dateninterpretation konzentriert sich auf die Perfektion bestimmter Versuchsstadien (Was hätte ich besser machen können?) und vernachlässigt den theoretischen Bezug.
- Keine abschließende konzeptuelle Neubewertung der Operationalisierungen von UV und AV (z.B. doch nicht nur subjektive Maße, sondern auch physiologische oder expressive Maße verwenden).
- Frage der Generalisierbarkeit (ext. Validität) wird nicht erörtert bzw. nicht in Form hypothetischer Schlussfolgerungen für neue Studien diskutiert.
- Herleitung neuer Fragestellungen orientiert sich nicht bzw. zu wenig an vorhergehenden Versuchsstadien.
- Es erfolgt keine wissenschaftliche Kommunikation (Veröffentlichung).

1.3. Zusammenfassung

Der Gedanke des Spiralenmodells nach Sarris beinhaltet, dass zu jedem Zeitpunkt der Untersuchung gefragt werden muss, was bisher geschah, welcher Schritt momentan erreicht ist und wohin man steuert. Die einzelnen Stufen des Prozesses sind also eng miteinander verzahnt, was es nötig macht, jede einzelne Stufe des Erkenntnisprozesses zu jedem Zeitpunkt im Blick zu haben.

Das Spiralenmodell beschreibt Forschung also als simultanen (nicht sukkzesiven) Prozess und lässt sich auch auf mehr als ein Experiment erweitern – im Sinne einer Theorie, die immer differenzierter untersucht wird.

2. Forschungsformen und Stichproben

2.1. Forschungsformen

2.1.1. Labor- vs. Feldforschung

Unter Laborforschung versteht man die Durchführung einer Untersuchung in einem speziell für den Zweck der Untersuchung entwickelten, also künstlichen, Milieu. Feldforschung hingegen bezeichnet die Beobachtung und Analyse natürlich auftretender, also vorgefundener, Situationen.

2.1.1.1. Laborforschung: Vor- und Nachteile

Vorteile:
- Situation und Verhalten sind leichter manipulierbar.
- Störvariablen können besser kontrolliert werden.
- Schaffung optimaler Bedingungen für die Untersuchung.

Nachteile:
- Die Umgebung ist ungewohnt und unnatürlich und kann so zu Artefakten führen (Abhilfe: Gewöhnungsphase).
- Die Personen wissen, dass sie untersucht werden und verändern so u.U. ihr Verhalten.
- Problem der Übertragbarkeit auf normales Verhalten.

2.1.1.2. Feldforschung: Vor- und Nachteile

Vorteile:
- Natürliche Umgebung
- Spontanes, normales Verhalten
- Besser übertragbar auf natürliches Verhalten
- Keine oder geringe Verfälschung durch Wissen um Studie

Nachteile:
- Störvariablen schlecht zu kontrollieren
- Manipulation von Situation und Verhalten schwierig
- Das Verhalten ist schwer zugänglich
- Die Untersuchungsbedingungen sind nicht optimal

2.1.1.3. Validität

Der allgemeine Konsens spricht der Laborforschung eine hohe interne und eine geringe externe Validität zu, der Feldforschung im Gegensatz eine hohe externe und eine niedrige interne Validität.

Es finden sich jedoch wenige systematische Vergleiche von Labor- und Feldforschung; der Konsens unterliegt zudem einer Betrachtung der Pole möglicher Labor- und Feldforschung.

Welche Forschungsform gewählt werden sollte ist u.a. abhängig vom aktuellen Erkenntnisstand. Liegen z.B. viele Laborstudien vor (interne Validität; Wirkgefüge), sollte der vermutete Zusammenhang in einer Feldstudie untersucht werden. Generell ist eine Kombination beider Methoden zu einer sog. experimentellen Felduntersuchung sinnvoll (z.B. Bortz & Braune, 1980: Manipulation d. polit. Einstellung durch Lektüre v. Tageszeitungen).

2.1.2. Web-Experimente

Das Web-Experiment ist eine Ergänzung des klassischen methodischen Repertoires von Labor- und Feldforschung. Unterschiede zur klassischen Laborforschung sind:

- Der Versuch kommt zum Probanden (inkl. den dort wirkenden SVn)
- Pbn können den Versuch jederzeit abbrechen
- Abhängig von technischer Ausstattung der Probanden
- Untersuchung zu heterogener Population und z.T. sehr großer Stichprobenumfang (n > 1000 in kurzer Zeit)

2.1.2.1. Vor- und Nachteile

Vorteile:

- Untersuchung heterogener Populationen (demographische und soziale Merkmale); nicht nur Studenten als Pbn
- Zugang zu spezifischen Populationen (z.B. Surfer)
- Hohe externe Validität: Generalisierbarkeit auf Populationen, Settings und Situationen
- Keine organisatorischen Probleme (Raum, Apparatur, VL…)
- Pbn-Motivation bestimmbar (durch High-Hurdle oder Drop-out)
- Pbn nehmen freiwillig teil
- Sehr große Stichproben (hohe statistische Power)
- Geringe Kosten (Raum, Zeit, Ausstattung, Durchführung)
- Hoher Automatisierungsgrad (Kontrolle von Vp-Vl-Effekten sowie Demand-Effekten)

Nachteile:

- Mögliche Mehrfachteilnahme der Pbn (Abhilfe: Personalisierungsitems, Überprüfung der internen Konsistenz und Zeitkonsistenz, IP-Adressen-Kontrolle geht nicht).
- V.a. mit between-Faktoren umsetzbar (Pbn bekommen nur kleine und damit kurze Teile einer größeren Untersuchung)
- Auswahlfehler bei Stichprobenzusammensetzung (Lösung: Multiple Site-Entry Technique)
- Fehlende Vp-Vl-Interaktion (Lösung: Vorversuche zu Instruktion und Material)
- Interne Validität fraglich
- Hohe Drop-Out Quote und Verweigerung der Information

2.1.2.2. Problem: Verweigerung einer Information

Hier muss zwischen Item-Non-Response, also einer Verweigerung die auf einzelne Items bezogen ist, und Unit-Non-Response, also der kompletten Verweigerung der Auskunft (Ablehnung der Teilnahme an Interview, keine Rücksendung des Fragebogens) unterschieden werden.

Die Item-Non-Responder-Quoten können relativ einfach vermindert werden, indem alle Fragen beantwortet werden müssen (Check). Die Verringerung der Unit-Non-Responder-Quote kann durch finanzielle Anreize, persönliche Fragen zu Versuchsbeginn oder auch die Vermeidung ladeaufwändiger Inhalte erreicht werden.

2.1.2.3. Problem: Drop-out

Die Drop-out-Quote lässt sich bei der Verwendung von One-Item-One-Screen-Designs oder mindestens Multipage-Designs recht einfach bestimmen. Single-Web-Page-Designs sind unter diesem Gesichtspunkt grundsätzlich zu vermeiden.

Zur Verringerung der Drop-out-Quote wurden verschiedene Techniken entwickelt, von denen die drei wichtigsten vorgestellt werden sollen (Reips, 2002):

- **High-Hurdle Technique:** Auf die Motivation negativ wirkende Informationen (lange Dauer, pers. Daten) werden möglichst konzentriert zu Versuchsbeginn dargeboten. Auf den folgenden Seiten werden Konzentrationsanforderung und Bedeutung kontinuierlich reduziert.
- **Warm-up Technique:** Ausgangspunkt dieser Überlegung ist, dass Dropouts zumeist nach wenigen Seiten auftreten, da sich Pbn zunächst im Versuch orientieren und schließlich endgültig über ihre Teilnahme entscheiden. Demnach sollte der Hauptteil des Versuchs erst nach einigen Webseiten starten; zuvor werden Instruktion und Übungsseiten (z.B. für Material und Reaktionen) dargeboten.
- **Seriousness-Check:** Abfrage der Involviertheit der teilnehmenden Pbn zu Versuchsbeginn. Bei niedrigen Involviertheits-Scores kann die Teilnahme verweigert bzw. der Datensatz nicht ausgewertet werden.

2.1.2.4. Problem: Interne Validität

In einem Web-Experiment lässt sich nicht kontrollieren, ob die Pbn angemessen auf Stimuli reagieren, was zumindest eingeschränkt über eine Erfassung von Browsertyp, Betriebssystem etc. erfasst werden kann.

Auch lässt sich nicht kontrollieren, ob die Pbn tatsächlich auf Stimuli reagieren, bzw. ob eine Reaktion wirklich von ihnen kommt oder evtl. von einer anderen Person, die z.B. den Rechner benutzt, während der eigentliche Pb diesen kurz verlässt.

2.1.3. Einzelfallforschung

Bei der Einzelfallforschung wird eine Untersuchungseinheit (Individuum, Familie, Verein…) häufig mittels nicht standardisierter Verfahren untersucht. Die Einzelfallforschung kann zum Zwecke einer detaillierten Beschreibung eines Phänomens oder aber auch zur Hypothesengenerierung eingesetzt werden.

Die erste Einzelfallstudie wird von Ebbinghaus (1885) berichtet, der sich selbst über einen langen Zeitraum sinnlose Silben einprägte (Konsonant-Vokal-Konsonant-Trigramme), um so Lernprozesse zu untersuchen.

2.1.3.1. Vorteile

- Seltene Phänomene sind beschreibbar.
- Problem der Übertragbarkeit von statistischen Gruppenkennwerten auf Einzelfälle ergibt sich nicht.
- Bei Auswahl von Einzelfällen sind Voraussetzungen z.B. einer Zufallsstichprobe nicht notwendig.

- Ergebnisse werden häufig so behandelt, als wären sie unabhängig voneinander. Einzelfallforschung ist bei Mehrfacherhebung stets abhängig und kann durch spezielle Verfahren kontrolliert werden.

2.1.3.2. Nachteile

- Problem der Replizierbarkeit der Ergebnisse zur Beschreibung einer Gesetzmäßigkeit sowohl hinsichtlich direkter Replikation (Variation der Zeit- und Personenvariable) als auch indirekter Replikation (Kombination von Setting-, Zeit-, Pbn-, VL- und Störvariablen.)
- Zusammenfassung von Einzelergebnissen (Aggregation) problematisch; lediglich bei vielen Einzelfallanalysen lassen sich Varianzanalysen mit standardisierten Zeitreihenwerten (z.B. z-Werte) durchführen.
- Geringe Verallgemeinerbarkeit der Ergebnisse auf nicht untersuchte Elemente.

2.1.4. Querschnittstudien

Trautner (1978): Zu einem bestimmten Zeitpunkt werden mehrere Stichproben von Individuen mit demselben oder einem vergleichbaren Messinstrument jeweils einmal untersucht.

Querschnittstudien werden beispielsweise in der Entwicklungspsychologie eingesetzt, um den kognitiven Entwicklungsstand verschiedener Altersstufen zu vergleichen.

2.1.4.1. Vorteile

- Kurze Durchführungsdauer der Untersuchung
- Geringer Personalaufwand und geringe Kosten
- Umfang der Stichprobe bleibt im Erhebungszeitrum konstant

2.1.4.2. Nachteile

- Unterschiede in Versuchsgruppen können durch Unterschiede zwischen Gruppen (Manipulation der UV) oder zwischen Probanden bedingt sein.
- Unabhängige Stichproben erlauben keine Aussagen zu intraindividuellen Unterschieden.
- Für unabhängige Stichproben stehen weniger effiziente statistische Verfahren zur Verfügung.
- Generalisierung der Befunde über den Zeitpunkt der Untersuchung ist streng genommen nicht erlaubt. Ein Beispiel hierfür sind politische Umfragen: die Beliebtheit von G.W. Bush schnellte nach dem 11. September drastisch nach oben, sodass eine auch nur kurz vorher erhobene Statistik nicht mehr aussagekräftig war.

2.1.5. Längsschnittstudien

Baltes (1967): Dieselbe Stichprobe von Individuen wird mehrmals zu verschiedenen Zeitpunkten mit demselben oder einem vergleichbaren Messinstrument untersucht.

Beispiele sind Untersuchungen zur Einstellungsveränderung durch Interventionsprogramme (AIDS-Kampagne) oder auch die Untersuchung der kognitiven Entwicklung über den Zeitraum des Kindesalters.

2.1.5.1. Vorteile

- Unterschiede in den Messwerten dürfen als intraindividuelle Veränderungen interpretiert werden.
- Unterschiede innerhalb der Stichprobe dürfen als interindividuelle Unterschiede interpretiert werden.
- Für die Auswertung von abhängigen Stichproben stehen effizientere statistische Verfahren zur Verfügung.

2.1.5.2. Nachteile

- Geschichtlichkeit (Anwendbarkeit derselben Methode über längeren Zeitraum bzw. in verschiedenen Altersgruppen fraglich; Einfluss geänderter Umweltbedingungen)
- Entwicklung: Mortalität und Alterung der Probanden
- Testeffekte: Lerneffekte oder reaktive Effekte
- Konzentration i.d.R. auf eine Stichprobe
- Untersuchungsverfahren sind im Verlauf der Studie nicht mehr veränderbar, ohne die Vergleichbarkeit der Ergebnisse zu gefährden.

2.1.6. Bsp. einer Längsschnittstudie: Panelforschung

Die Panelforschung ist ein Spezialfall der Längsschnittstudie. Dabei werden in bestimmten zeitlichen Abständen bei denselben Untersuchungseinheiten dieselben Merkmale erhoben.

Mit der Panelforschung lassen sich Wandlungsprozesse untersuchen – oder genauer sowohl die intraindividuellen als auch die interindividuellen Veränderungen im Lauf der Zeit. Ein berühmtes Beispiel ist das sozioökonomische Panel der BRD, in dem etwa 6000 Haushalte beobachtet und Angaben zu Erwerbstätigkeit, demographischen Inhalten etc. gesammelt werden.

Bei diesem Panel handelt es sich um ein Beobachtungspanel, es sind jedoch auch beispielsweise Experimentalpanels denkbar.

2.1.6.1. Nachteile

Die wichtigsten Panel-Effekte sind Lern- und Testeffekte. Lerneffekte sind dabei abhängig von der Anzahl der Panel-Erhebungen (sog. Wellen) und deren Frequenz.

Bedeutsame Testeffekte sind:

- Veränderung bzw. Genese von Einstellungen und Verhaltensweisen
- z.B. verändertes Kaufverhalten durch erhöhtes Preisbewusstsein

Weitere Nachteile der Panelforschung sind:

- Mortalität, also der Ausfall von Erhebungseinheiten, der häufig bis zu 60% der Ausgangsstichprobe ausmacht und zum einen durch zufällige Ausfälle (Umzug, Tod) wie auch durch systematische Ausfälle (Desinteresse) entsteht. Diese syst. Ausfälle können mit den erhobenen Merkmalen zusammenhängen – man spricht in diesem Zusammenhang auch vom Effekt der positiven Selbstauswahl.

- Selektionseffekte: Bereits in der Anwerbephase stellt sich das Problem der Verweigerungsquote, die oft über 20% liegt.
- Geschichtlichkeit: Bei Langzeit-Panels kann sich Bedeutungsumfang und -inhalt verändern, so dass die Vergleichbarkeit der Daten fraglich ist.

2.1.6.2. Lösungsvorschläge

Bildung einer sehr großen Ausgangsstichprobe, so dass bis zum Ende der Panel-Studie hinreichend viele Einheiten erhalten bleiben. Hier bleibt trotzdem das Problem der positiven Selbstauswahl erhalten, zumal der zu erwartende Aufwand bei derartig großen Stichproben immens ist.

Alternativ lassen sich ausgefallene Einheiten auch auffüllen, wodurch jedoch die Repräsentativität der Untersuchungseinheiten stark gefährdet ist. Also: Auf das Auffüllen verzichten.

Um den Effekt der positiven Selbstauswahl sowie des Stichprobenauffüllens zu vermeiden, wurden verschiedene Panel-Designs entwickelt:

- Alternierendes Panel
- Rotierendes Panel
- Geteiltes Panel

2.1.6.3. Alternierendes Panel

Bildung von Subgruppen, die abwechselnd untersucht werden. Hier ist jedoch ein relativ großer Stichprobenumfang Voraussetzung, da zusätzlich die Mortalität zu berücksichtigen ist.

	t1	t2	t3	t4	t5	t6
Gruppe 1	x		x		x	
Gruppe 2		x		x		x

2.1.6.4. Rotierendes Panel

Bildung von Subgruppen (1-3), die bei der ersten Welle alle erhoben werden. Bei zweiter Welle scheidet eine Subgruppe aus und wird durch eine neue Subgruppe ersetzt usw. Manche Gruppen werden dabei nur einmal befragt; dieses Panel-Design beinhaltet also kleine Querschnittstudien.

Der Nachteil ist, dass dieses Design sehr aufwändig ist, da bei jeder Erhebung eine neue Subgruppe gebildet werden muss.

	t1	t2	t3
Gruppe 1	x	x	x
Gruppe 2	x	x	
Gruppe 3	x		
Gruppe 4		x	x
Gruppe 5			x

2.1.6.5. Geteiltes Panel

Eine Subgruppe durchläuft alle Wellen, eine zweite Subgruppe wird nach jeder Welle durch eine neue ersetzt. Die Gruppen 2 bis 5 (n) werden also nur einmal befragt (Querschnittstudie) und dienen als Kontrollgruppen für Gruppe 1.

Nachteil: Das Design ist sehr aufwändig, da bei jeder Erhebung eine neue Subgruppe gebildet werden muss.

	t1	t2	t3	t4
Gruppe 1	x	x	x	x
Gruppe 2	x			
Gruppe 3		x		
Gruppe 4			x	
Gruppe 5				x

2.1.7. Quer- vs. Längsschnittstudien

Längsschnittstudie	Querschnittstudie
Eine Stichprobe	Verschiedene Stichproben
Zwei oder mehr Zeitpunkte	1 Zeitpunkt
Mehrmalige Untersuchung mit dem selben od. vgl. Messinstrument	Einmalige Untersuchung mit selbem od. vergleichbarem Messinstrument

Eine Längsschnittstudie kann auch als eine Summe abhängiger Messungen (within-subject-design, within-design, Innergruppendesign), eine Querschnittstudie als Summe unabhängiger Messungen (between-subjects-design, between-design, Zwischengruppendesign) bezeichnet werden.

2.1.8. Sekundäranalysen

Primäranalysen beziehen sich auf eine selbstständige Datenerhebung als wesentlichen Bestandteil des Forschungsvorhabens (klassische Experimente). Sekundäranalysen hingegen greifen auf bereits existierende Datenbestände zurück, die im Rahmen einer anderen Untersuchung (häufig zu einem anderen Zweck) erhoben wurden.

Beispiele sind Auswertungen in der Wirtschaftsforschung (BIP, demographische Daten), Metaanalysen, etwa zur Wirksamkeit verschiedener Therapieformen oder Literaturreviews.

2.1.8.1. Vorteile

- Kosteneinsparung (z.B. keine Versuchsmaterialien)
- Schnelle Verfügbarkeit, geringer Aufwand
- Nachkontrollierbarkeit

2.1.8.2. Nachteile

- Daten wurden i.d.R. für einen anderen Zweck erhoben
- Qualität der Daten hängt vom Vorgehen der Untersucher ab
- Möglicher Abweichung der Grundgesamtheit, Auswahl der Erhebungs- und Untersuchungseinheiten, Begriffsdefinitionen und Operationalisierungen vom eigenen Projekt
- Daten sind unter Umständen veraltet

2.2. Selektion: Das Problem der Stichprobe

Auch eine sorgfältig gezogene Stichprobe kann die Merkmalsverteilung einer Grundgesamtheit niemals exakt wiedergeben. Daher sind Unterschiede zwischen den an mehreren Stichproben ermittelten Verteilungskennwerten zu erwarten.

2.2.1. Definition

Eine Grundgesamtheit (Population) besteht aus allen potenziell untersuchbaren Einheiten, die ein gemeinsames Merkmal bzw. eine gemeinsame Merkmalskombination aufweisen.

Eine Stichprobe ist eine Teilmenge aller Untersuchungseinheiten, die die relevanten Eigenschaften der Grundgesamtheit möglichst gut abbilden sollte. In diesem Zusammenhang unterscheidet man auch zwischen globaler Repräsentativität, bezogen auf eine Vielzahl verschiedener Merkmale oder Merkmalskombinationen, und spezifischer Repräsentativität, bezogen auf einige wenige Merkmale oder Merkmalskombinationen.

Je besser die Stichprobe die Population repräsentiert und je größer die Stichprobe ist, desto präziser treffen die durch sie erhaltenen Aussagen auf die Population zu.

Man muss zwischen zufallsgesteuerter und nicht-zufallsgesteuerter Stichprobe unterscheiden.

2.2.2. Zufallsgesteuerte Stichproben

Zu den zufallsgesteuerten Stichproben gehören die reine Zufallsstichprobe sowie Klumpen- und geschichtete (stratifizierte) Stichprobe. Zufallsgesteuerte Stichproben sind dabei in jedem Fall nicht-zufallsgesteuerten vorzuziehen.

2.2.2.1. Zufallsstichprobe

Grundprinzip: Jedes Element der Grundgesamtheit kann mit gleicher Wahrscheinlichkeit in die Stichprobe aufgenommen werden. Die Aufnahme ist dabei unabhängig von weiteren Elementen. Hierzu muss eine zufällige Auswahl von Untersuchungseinheiten aus einer Grundgesamtheit erfolgen, was sich annäherungsweise über Listen des Einwohnermeldeamtes gewährleisten lässt.

Die Zufallsstichprobe sollte dann verwendet werden, wenn über ein relevantes Untersuchungsmerkmal praktisch nichts bekannt ist.

Ein Problem der Zufallsstichprobe sind mögliche systematische Fehler im Auswahlverfahren – auf den Listen des Einwohnermeldeamtes stehen beispielsweise nicht immer alle Einwohner der Stadt (z.B. Erstwohnsitz von Studenten).

2.2.2.2. Klumpenstichprobe

Bei einer Klumpenstichprobe wird auf mehrere, zufällig ausgewählte Teilmengen zurückgegriffen, die bereits vorgruppiert sind. Anschließend werden alle Einheiten dieser Teilmengen untersucht.

Beispiele sind Konsumentenbefragungen oder eine Untersuchung der Abstraktionsfähigkeit von Alkoholikern in verschiedenen Kliniken.

Die Untersuchung eines einzelnen Klumpens (z.B. einer Schulklasse) wird als Ad-hoc-Stichprobe bezeichnet.

Generell sollte die Klumpenstichprobe vor allem aus ökonomischer Sicht gewählt werden, da die Generalisierbarkeit nicht zwingend gegeben ist und von der Homogenität (Ähnlichkeit der Einheiten) eines Klumpens abhängt (Klumpen-Effekt).

2.2.2.3. Geschichtete (stratifizierte) Stichprobe

Stichproben werden hinsichtlich einer Drittvariable zusammengestellt, die als moderierende Variable bezüglich des Untersuchungsmerkmals bekannt ist oder angenommen wird. Innerhalb einer Schicht soll per Zufall oder nach dem Klumpenverfahren vorgegangen werden.

Spiegelt die prozentuale Verteilung der Schichtungsmerkmale der Stichprobe die Verteilung in der Grundgesamtheit wider, so spricht man auch von einer proportional geschichteten Stichprobe.

Ein Beispiel ist die Berücksichtigung des Jahreseinkommens bei der Untersuchung von Konsumgewohnheiten.

Die geschichtete Stichprobe kann nur dann verwendet werden, wenn bereits Kenntnisse zu moderierenden Einflüssen auf das Untersuchungsmerkmal vorliegen. Zudem muss das Schichtungsmerkmal nicht nur mit dem Untersuchungsmerkmal korrelieren, sondern zugleich auch erhebbar sein (evtl. ist die Abstraktionsfähigkeit ein guter Prädiktor für die Leistung in einem bestimmten Test, aber nicht sinnvoll messbar).

Probleme: Nicht die Anzahl der geschichteten Merkmale bestimmt die Repräsentativität der Stichprobe, sondern die Relevanz der Merkmale. Zudem erfolgt eine Explosion der Schichtanzahl bei mehreren Schichtungsvariablen.

2.2.2.4. Sonderfall: Mehrstufige Stichprobe

Bei der mehrstufigen Stichprobe werden Klumpenstichproben oder geschichtete Stichproben ausgewählt. Danach erfolgt eine stichprobenartige Untersuchung mehrer Klumpen bzw. Schichten.

Die mehrstufige Stichprobe kommt dann zum Einsatz, wenn die zu untersuchenden Klumpen oder Schichten zu groß sind (z.B. Pisa-Test).

2.2.3. Nicht-zufallsgesteuerte Stichproben

Nicht-zufallsgesteuerte Stichproben beziehen sich auf die Auswahl der Stichprobe anhand vorher definierter Kriterien, wie beispielsweise:

- Befragung fotogener Passanten in der Einkaufspassage während der Rush-Hour (willkürliche Auswahl anhand subjektiver Kriterien)
- Einbezug von typischen Konsumenten bei Produktentwicklung (Auswahl typischer Fälle; Bsp.: BMW-Fahrer)
- Konzentration auf besonders dominante Elemente der Grundgesamtheit (z.B. Einkommensmillionäre; Auswahl nach Konzentrationsprinzip)
- Quoten-Stichproben (s.u.)

2.2.3.1. Quoten-Stichproben

Die Untersucher (Interviewer) müssen vorgegebene Quoten (=festgelegte Häufigkeitsverteilungen relevanter Merkmale) erfüllen, die Auswahl innerhalb dieser Quoten bleibt i.d.R. dem Untersucher überlassen und erfolgt daher häufig nach dem Verfügbarkeitsprinzip.

Quoten-Stichproben werden beispielsweise in der Umfrageforschung verwendet.

Probleme: Nur die prozentuale Aufteilung der Quotierungsmerkmale wird betrachtet, i.d.R. allerdings nicht die Aufteilung von Merkmalskombinationen (Arzt zwischen 40-50 mit 4 Kindern). Auch die Quotenerfüllung nach Verfügbarkeitsprinzip ist problematisch (Vernachlässigung höherer Stockwerke).

2.2.4. Stichproben und Repräsentativität

Stichprobenkennwerte repräsentieren den jeweiligen Populationsparameter immer nur mit einer bestimmten Wahrscheinlichkeit, eine wirklich repräsentative Stichprobe gibt es im Grunde genommen also nicht.

Um zu prüfen, inwiefern empirische Kennwerte adäquate Schätzwerte für Populationsparameter sind, ist die Bestimmung des Konfidenzintervalls sinnvoll.

Das Konfidenzintervall ist der Bereich eines Merkmals, in dem sich z.B. 95% aller möglichen Populationsparameter befinden, die den Stichprobekennwert erzeugt haben können. Je größer die Stichprobe ist, desto kleiner ist das Konfidenzintervall.

Vor der Durchführung einer Untersuchung sollte entschieden werden, wie viele Personen benötigt werden, um Aussagen mit der gewünschten Genauigkeit machen zu können. Eine Verkleinerung des Konfidenzintervalls geht mit einer Quadrierung des benötigten Stichprobenumfangs einher (Halbierung des KI → 4-facher Stichprobenumfang)

3.Versuchsplanung I

3.1. Idee der Versuchsplanung

3.1.1. Beziehung zwischen UVn und AVn

Werden in einem Experiment mögliche Störgrößen kontrolliert oder eliminiert, so verändern die Abstufungen der UV systematisch die AV – die UV kann also als Ursache, die AV als Wirkung angesehen werden:

AV = f(UV)

Der Kausalzusammenhang zwischen UV und AV kann also geprüft werden, indem Situationen betrachtet werden, die sich nur durch die Ausprägung der UV unterscheiden. Verändert sich die AV, dann können nur die Unterschiede in der UV dafür verantwortlich sein.

3.1.1.1. Versuchsplan

Wird mehr als eine UV verwendet, wie im unten stehenden Beispiel, so wird dies normalerweise durch den Ausdruck „AxB-Plan“ beschrieben, wobei A und B jeweils signalisieren, wie viele Ausprägungen die UV hat. Ein 2x2-Plan besteht also aus 2 UVn mit je zwei Ausprägungen und lässt sich in einer Vierfeldertafel beschreiben. Ein 3x2-Plan beschreibt 2 UVn, von denen eine 3, die andere 2 Ausprägungen aufweist.

↓ UV2 / UV1 →	+	-
+		
-		

2x2-Plan

↓ UV2 / UV1 →	-1	0	+1
+			
-			

3x2-Plan

Bemerkung: Statt von UVn kann auch von Faktoren gesprochen werden (ein 2-faktorieller Versuchsplan beinhaltet also 2 UVn). Die AV kann auch als Variate bezeichnet werden (univariat = eine AV).

3.1.1.2. Hypothetisches Beispiel

Angenommen, man will die Auswirkung der Trommelfrequenz auf die Lautstärke eines Geräusches untersuchen, das entsteht, wenn man einem anderen auf den Rücken trommelt.

UV 1: Trommelfrequenz (gering vs. hoch)
UV 2: Lungenvolumen (klein vs. groß)

Aus den UVn ergibt sich ein 2x2-Plan mit folgenden 4 Versuchsgruppen:

- Geringe Frequenz – kleine Lunge
- Hohe Frequenz – kleine Lunge
- Geringe Frequenz – große Lunge
- Hohe Frequenz – große Lunge

3.1.2. Varianzarten der AV

Die erhaltenen Ergebnisse des hypothetischen Experiments lassen sich auf drei Quellen zurückführen: die experimentelle Variation (UV1: niedrige vs. hohe Frequenz), stabile Personenmerkmale (UV2: Lungenvolumen), sowie durch zufällige Personen- und Situationsmerkmale (Variabilität) beeinflusst.

In diesem Zusammenhang spricht man auch von Primärvarianz (durch experimentelle Variation), Sekundärvarianz (durch stabile Personenmerkmale) und Fehlvarianz (Variabilität). Der Grundgedanke des Experiments lässt sich in diesem Zusammenhang umformulieren: Es gilt die Primärvarianz möglichst hoch und die Fehlvarianz möglichst niedrig zu halten (interne Validität).

3.1.2.1. Definition

Primärvarianz (PV) bezieht sich auf die systematische Variation der Messwerte und ist zurückzuführen auf die Variation der UV. Die Primärvarianz resultiert also aus einer Kausalbeziehung von UV und AV.
Sekundärvarianz bezieht sich ebenfalls auf eine systematische Variation der Messwerte und ist auf die Variation identifizierbarer Störvariablen (Kontrollvariablen) zurückzuführen.
Fehlervarianz (FV) entsteht durch die unsystematische, also zufällige, Variation der Messwerte (Zufallsfehler). Sie ist weder auf den Einfluss der Variation der UV, noch auf den Einfluss identifizierbarer Störvariablen zurückzuführen.

3.1.2.2. Beispiel

Primärvarianz: Unterschiede zwischen den beiden Gruppen aufgrund der Trommelfrequenz (hohe vs. niedrige).
Sekundärvarianz: Unterschiede zwischen den beiden Untergruppen in den beiden Versuchsgruppen (kleines vs. großes Lungenvolumen).
Fehlervarianz: Unterschiede innerhalb der vier Untergruppen.

3.1.2.3. Hierarchisches Modell

Sarris versucht der Aufteilung der Gesamtvarianz eine hierarchische Struktur zu geben. Sie lässt sich zunächst aufteilen in Primärvarianz (Behandlungsvarianz) und Fehlervarianz I (alle andere inter- und intraindividuelle Varianz). Diese Fehlervarianz I wiederum lässt sich in Sekundärvarianz (systematische Fehler) und Fehlervarianz II (Zufallsfehler, Rest) untergliedern.

3.1.2.4. PV und FV

Wie bereits erwähnt, kann die Wirkung der unabhängigen Variablen nur dann festgestellt werden, wenn die Primärvarianz der abhängigen Variable deutlich größer ist als deren Fehlervarianz:

$$\frac{PV}{FV} > 1$$

3.1.3. Die Logik der statistischen Prüfung

Es stellt sich nun die Frage, wie „deutlich größer“ interpretiert werden soll. Zu diesem Zweck muss die gefundene Abweichung in ein Wahrscheinlichkeitsmodell übertragen werden. Es muss also eine Verteilung von Ereignissen erstellt werden, die unter Annahme der Nullhypothese zu erwarten ist.

Nun erfolgt eine Entscheidung des Untersuchers bzgl. des Fehlers erster Art. Dabei stellt sich die Frage, welche Gruppenunterschiede zufällig auftreten könnten, also wie wahrscheinlich das empirische Ereignis unter der Annahme der H_0 ist. Ist diese Wahrscheinlichkeit geringer als das α-Niveau (Ablehnungsbereich), so kann die H_0 verworfen werden.

3.1.3.1. α-Niveaus und Hypothesen

Bei der Anwendung eines α-Niveaus in einem statistischen Test muss grundsätzlich beachtet werden, ob man auf eine gerichtete oder ungerichtete Hypothese prüft.

3.1.3.2. Interpretation von α-Niveaus

1%	Hochsignifikant
5%	Signifikant
5-10%	marginalsignifikant
[10-20%	Tendenz]

3.2.Varianzanalyse: Wirkungen der UV

Die Varianzanalyse greift den bereits erwähnten Grundgedanken PV/FV>1 auf und setzt diesen statistisch um.

3.2.1. Grundidee der Varianzanalyse

Die Idee hinter der Varianzanalyse ist es also, zu zeigen, wie viel Variation der AV durch die UV erzeugt wird (PV). Hierzu wird die Gesamtvarianz in PV und FV aufgeteilt und deren Teile miteinander verglichen. Ein statistisches Modell hilft bei der Entscheidung, ab wann das „Größer“ bedeutsam ist.

3.2.1.1. Definitionen

Man unterscheidet bezüglich der Quellen der Primärvarianz mehrere Arten von Wirkungen – Haupt- und Wechselwirkungen. Eine Hauptwirkung (HW) bezieht sich auf Unterschiede bzgl. der verschiedenen Stufen einer UV, die Wechselwirkung (WW) bezieht sich auf die gemeinsame Wirkung von mehreren UVn. Wechselwirkungen beziehen sich also auf systematische Variationen, die nicht auf additive Effekte der UVn zurückgeführt werden können.

In einem einfaktoriellen Versuchsplan gibt es also nur eine HW. In einem zweifaktoriellen Versuchsplan gibt es die HW der UV1, die HW der UV2 sowie die WW der beiden UVn. Diese Wechselwirkungen können nur von relativ wenigen statistischen Verfahren nachgewiesen werden.

Bemerkung: Statt HW kann auch synonym von Haupteffekt (HE), statt WW auch von Interaktion gesprochen werden.

3.2.1.2. Hypothesen

Bei einfaktoriellen Versuchsplänen besagt die H_0 grundsätzlich, dass die Experimentalgruppen aus Populationen mit dem gleichen Mittelwert kommen, die H_1 hingegen, dass sich mindestens 2 Mittelwerte unterscheiden.

Bei zweifaktoriellen Versuchsplänen werden die Hypothesen bzgl. der Hauptwirkungen analog zu einfaktoriellen Versuchsplänen gebildet. Zusätzlich werden jedoch Hypothesen hinsichtlich der WW formuliert: H_0 = Die Zellenmittelwerte setzen sich additiv aus den Haupteffekten zusammen, H_1 = die Zellenmittelwerte setzen sich nicht additiv aus den Haupteffekten zusammen.

3.2.1.3. Quadratsummen

Zusätzlich zu den Hypothesen muss noch ein Maß für die Veränderungen in der AV gesucht werden. Die Differenz zwischen Versuchsgruppen kann hier nicht mehr herangezogen werden, wenn mehr als zwei Stufen der UV bzw. mehr als zwei UVn untersucht werden.

Von daher werden (Abweichungs-)Quadratsummen als Maß der Unterschiedlichkeit verwendet:
QS Total: Wie unterschiedlich sind die Personen, die ich untersucht habe (Gesamtvarianz)?
QS HWen, QS WWen: Wie unterschiedlich sind die Gruppen unter den Stufen der UV (Primärvarianz)?
QS Fehler: Wie unterschiedlich sind die Personen noch, wenn die Unterschiede die durch die UVn entstanden sind, abgezogen wurden (Fehlervarianz)?
Die Sekundärvarianz wird in der Varianzanalyse nicht betrachtet.

3.2.2. Additives Modell der Varianzanalyse

3.2.2.1. Quadratsummen

Eine wichtige Voraussetzung der Varianzanalyse ist, dass sich die einzelnen Abweichungsquadratsummen additiv verhalten. So kann QS Total als Maß der Unterschiedlichkeit von Stichproben in einzelne Komponenten zerlegt werden in QS_i (HW des Faktors i), QS_{ij} (WW der Faktoren i und j) sowie QS_{Fehler}:

1-faktoriell: $QS_{Tot} = QS_A + QS_{Fehler}$
2-faktoriell: $QS_{Tot} = QS_A + QS_B + QS_{AxB} + QS_{Fehler}$
3-faktoriell: $QS_{Tot} = QS_A + QS_B + QS_C + QS_{AxB} + QS_{AxC} + QS_{BxC} + QS_{AxBxC} + QS_{Fehler}$

3.2.2.2. Messwerte

Auch der Messwert, den eine Person liefert, setzt sich aus verschiedenen Einflussgrößen zusammen. Im zweifaktoriellen Beispiel:

$$X_{ijk} = G_{...} + A_{i..} + B_{.j.} + AB_{ij.} + E_{ijk}$$

Mit X_{ijk} = Messwert der Person, $G_{...}$ = Typischer Wert der untersuchten Stichprobe, $A_{i..}$ = Einfluss der Stufe i der ersten UV, $B_{.j.}$ = Einfluss der Stufe j der zweiten UV, $AB_{ij.}$ = Einfluss der Kombination der jeweiligen Stufe von UV1 und UV2 und E_{ijk} = Typischer Wert der Person („Fehler“).

3.2.3. Berechnung…

3.2.3.1. …der Quadratsummen

QS Total – auf Basis der Anfangstafel: $\sum \left(x_{ij} - m_{ges} \right)^2$

QS Fehler – auf Basis der letzten Tafel: $\sum x_{ij}^2$

QS HW und WW – erwartete Effekte in Tafel einsetzten: $\sum x_{ij}^2$

3.2.3.2. …der Freiheitsgrade

Auch die Freiheitsgrade (df) verhalten sich additiv, von daher kann man sich die häufig aufwändige Berechnung der Freiheitsgrade von QS Fehler ersparen, wenn man die restlichen Freiheitsgrade berechnen kann.

QS A (entsteht aus k Werten, der Zufall darf k-1 wählen): $df_A = k-1$
QS B (entsteht aus l Werten, der Zufall darf l-1 wählen): $df_B = l-1$
QS WW (Multiplikation der df der Haupteffekte): $df_{AxB} = (k-1)x(l-1)$
QS Fehler: $df_{Fehler} = df_{Total}$ - Summe(restliche Freiheitsgrade)
QS Total (bei Stpn der Größe n können nur (n-1) Abweichungen vom Mittelwert variiert werden): $df_{Total} = n-1$

3.2.3.3. …der mittleren Quadratsummen

Beim Vergleich von Variationsmaßen muss berücksichtigt werden, durch wie viele Quellen die Variation insgesamt entstehen kann. Daher werden die Quadratsummen an der Anzahl der Freiheitsgrade normiert. Man spricht dann von mittleren Quadratsummen (MQ).

$$MQ = \frac{QS}{df}$$

Diese normierten Variationsmaße können nun miteinander verglichen werden, wodurch ein F-Bruch entsteht:

$$F = \frac{MQ_{A/B/AxB}}{MQ_{Fehler}}$$

3.2.4. Statistisches Modell

Zusammenfassend besteht die Varianzanalyse also aus folgenden Teilen:

- Erzeugen einer Verteilung von PV/FV unter der Nullhypothese.
- Nullhypothese: UV erzeugt keine große Variation der AV.
- Wenn empirisches Verhältnis PV/FV in der Verteilung hinreichend unwahrscheinlich ist, dann ist das Modell nicht gut.
- Folge: Ablehnung der Nullhypothese; UV hat gewirkt.
- Was heißt hierbei unwahrscheinlich? → Festlegung durch α-Risiko

3.2.4.1. Prüfung der Effekte

Im zweifaktoriellen Beispiel liegen drei Arten von Primärvarianz vor, erzeugt durch UV1 (HW1), UV2 (HW2) sowie der Zusammenwirkung der UVn (WW).

Nun werden die mittleren Quadratsummen (MQ) miteinander verglichen: MQ_A mit MQ_{Fehler}, MQ_B mit MQ_{Fehler} sowie MQ_{AxB} mit MQ_{Fehler}.

3.2.4.2. Tafel der Varianzanalyse

Q.d.V	QS	df	MQ	F	p
Geschlecht	108	1	108	108	
Alkohol	432	1	432	432	
WW	48	1	48	48	
Fehler	8	8	1		
Total	596	11			

Mit Q.d.V. = Quelle der Variation, QS = Quadratsumme, df = Anzahl der Freiheitsgrade, MQ = Mittlere Quadrate (MQ = QS/df), F = F-Bruch (MQ Effekt / MQ Fehler), p = Wahrscheinlichkeit des F-Wertes unter Annahme der H_0.

3.2.5. Interpretation von Haupt- & Wechselwirkungen

3.2.5.1. Wann ist ein F-Bruch groß?

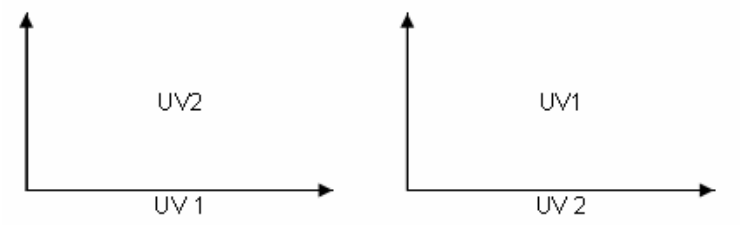
$$F = \frac{MQ_{Effekt}}{MQ_{Fehler}}$$

Ist der F-Wert kleiner 1, so liegt mit Sicherheit keine Wirkung vor. Für F > 1 liegt möglicherweise eine Wirkung vor. Die Wahrscheinlichkeit eines bestimmten F-Wertes lässt sich in einer F-Wertetabelle nachschlagen, sodass auf Basis eines gesetzten α-Niveaus entschieden werden kann, ob die Nullhypothese verworfen werden soll.

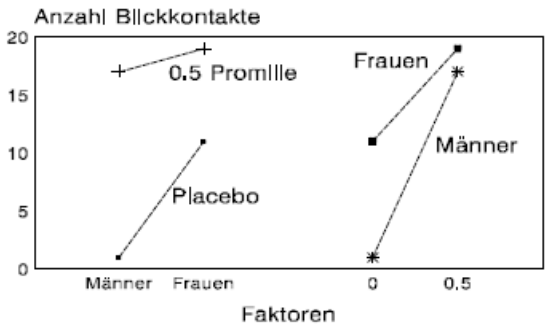
3.2.5.2. Interpretation

Solange die WW nicht signifikant wird, können alle Hauptwirkungen interpretiert werden. Wenn die WW signifikant wird, so hängt die Interpretation der HWen von der Art der WW ab (natürlich vorausgesetzt die HWen sind selbst signifikant).

Die Frage ist hierbei, ob sich die Richtung der HW durch die WW ändert. Wenn nein darf die HW interpretiert werden, wenn ja dann nicht. Die Interpretation ist dadurch nur möglich, indem man sich die Ergebnisse in einem Graphen oder zumindest einer Tabelle veranschaulicht:



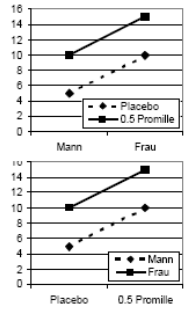
bzw.



3.2.5.3. Arten von Wechselwirkungen

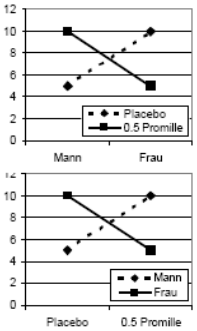
Ordinale Wechselwirkung (beide Graphen schneiden sich nicht und verlaufen annähernd parallel):

- Beide Hauptwirkungen dürfen interpretiert werden. Die Reaktion auf die UVn ist zwar unterschiedlich stark aber in die gleiche Richtung
- Wechselwirkung darf interpretiert werden.



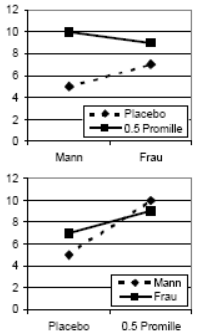
Disordinale Wechselwirkung (beide Graphen schneiden sich):

- Keine der beiden Hauptwirkungen darf interpretiert werden: die Wechselwirkung beeinflusst die Richtung beider Haupteffekte.
- Nur die Wechselwirkung darf interpretiert werden.



Semi-disordinale oder **hybride** Wechselwirkung (ein Geradenpaar schneidet sich, das andere verläuft quasi parallel):

- Eine Hauptwirkung (diejenige mit parallelen Geraden, also ohne Verzerrung des Haupteffekts durch die Wechselwirkung) darf interpretiert werden, die andere nicht.
- Die Wechselwirkung darf interpretiert werden.



Bei Signifikanz der Wechselwirkung darf diese also immer interpretiert werden, die Hauptwirkungen nur in Abhängigkeit von der Richtung der Wechselwirkung.

Bemerkung: Wenn eine semi-disordinale Wechselwirkung schon nahe an eine disordinale WW grenzt – die Geraden also deutlich unparallel sind – so liegt ein nicht geklärter Grenzfall vor, in dem theoretisch von beidem ausgegangen werden kann, jedoch ohne theoretische Bestätigung.

4. Versuchsplanung II

4.1. Vorexperimentelle Versuchspläne

4.1.1. Schrotschuss-Design

Bei einer One-Shot-Case-Study (Schrotschuss-Design) erfolgt eine einmalige Nachher-Messung an einer einzelnen Versuchsgruppe, der ein bestimmtes Treatment verabreicht wurde. Darstellung nach Sarris:

	Versuchs-gruppe	Vorher-Messung	Treatment X	Nachher-Messung
V	1	-	X	$\overline{Y}_{nach}$

Der große Vorteil des Schrotschuss-Designs ist, dass sich die Untersuchung mit geringstmöglichem Aufwand durchführen lässt. Allerdings weißt dieses Design auch erhebliche Nachteile auf, die die interne Validität stark gefährden:

- Fehlende experimentelle Kontrolle
- Keine Vergleichsmöglichkeiten der Untersuchungsbedingungen
- Aufgrund irreführender Plausibilität der Ergebnisse: Gefahr der missbräuchlichen Anerkennung dieses Designs

Daher finden sich in der Fachliteratur auch keine ernstzunehmenden Arbeiten, die auf diesem Design beruhen.

4.1.2. Einfache Vorher-Nachher-Messung

Eine einfache Vorher-Nachher-Messung (Prä-Post-Design) verläuft analog zur Schrotschuss-Untersuchung, es wird jedoch zusätzlich eine Ausgangsmessung eingeführt (Baseline). Auch hier können gefundene Differenzen nicht eindeutig auf die Behandlung zurückgeführt werden (interne Validität).

	Versuchs-gruppe	Vorher-Messung	Treatment X	Nachher-Messung
V	1	$\overline{Y}_{vor}$	X	$\overline{Y}_{nach}$

4.1.2.1. Vorteile:

- Interindividuelle Verhaltensvariabilitäten untersuchbar
- Zumindest Vergleich möglich, d.h. die Frage nach Veränderung der AV ist untersuchbar.

4.1.2.2. Nachteile

- Müdigkeits- oder Gewöhnungseffekte können für Ergebnis verantwortlich sein.
- Testeffekte aufgrund zweimaliger Testung.
- Fehlen eines Doppelblindversuchs, d.h. reaktive Verhaltensweisen von Pb und VL nicht kontrollierbar.

4.1.3. Statischer Gruppenvergleich

Beim statischen Gruppenvergleich werden zwei oder mehrere experimentell behandelte Gruppen miteinander verglichen. Die Gruppen werden jedoch nicht mittels einer Zufallsbildung zusammengestellt, sondern es werden bereits existierende, vorgegebene (=statische) Gruppen untersucht.

	Versuchs-gruppe	Vorher-Messung	Treatment X	Nachher-Messung
V	1	-	X_1	$\overline{Y_{1\,nach}}$
	2	-	X_2	$\overline{Y_{2\,nach}}$
	3	-	$X_3(X_0)$	$\overline{Y_{3\,nach}}$

Der statische Gruppenvergleich sollte dann verwendet werden, wenn eine Randomisierung nicht möglich ist (z.B. Pisa-Studie, Therapievergleichsstudie).

4.1.3.1. Vorteile

- Zumindest Vergleich zwischen verschiedenen Versuchsbedingungen möglich, also die Untersuchung der Veränderung der AV.

4.1.3.2. Nachteile

- Gleichheit der Versuchsgruppen ist nicht gewährleistet.
- Reifungseffekte werden nicht kontrolliert.

4.1.4. Bewertung

Bei allen vorexperimentellen Designs sind Effekte auf Grund der Zeit ein großer Störfaktor hinsichtlich der internen Validität, inklusive Geschichtlichkeit, Entwicklung, Selektion und Messeffekte sowie Testeffekte (vgl. 1.2.6).

4.1.4.1. Vorgehen und Störvariablen

In jedem der genannten Designs erfolgt eine explizierte Einführung einer experimentellen Bedingung. Allerdings erfolgt praktische keine Kontrolle von Störvariablen. Die Untersuchungsbefunde können also durch Störvariablen verzerrt sein.

Die Ergebnisse solcher Versuche sind demnach prinzipiell mehrdeutig und es kann nicht über die Gültigkeit von Alternativerklärungen entschieden werden.

4.1.4.2. Eignung

Vorexperimentelle Designs eignen sich vor allem für Pilotstudien mit dem Ziel der Hypothesengenerierung und der Entwicklung eines adäquaten Versuchsdesigns.

4.2. Das Experiment

Die bereits aufgeführten Kritikpunkte an vorexperimentellen Designs, lassen sich über ein kontrolliertes Experiment ausschalten.

4.2.1. Schematische Darstellung

Population	↱ → → →	EG	$\xrightarrow{UV1}$	Messung (AV)
	↕ *Random*			
	↳→ → →	KG	$\xrightarrow{UV0}$	Messung (AV)

Nach Sarris:

	Versuchs-gruppe	Vorher-Messung	Treatment X	Nachher-Messung
R	E	-	X_E	$Y_{E\ nach}$
	K	-	(X_K)	$Y_{K\ nach}$

4.2.2. Definition des Experiments

„Unter einem Experiment versteht man einen **systematischen Beobachtungsvorgang**, aufgrund dessen der Untersucher das jeweils interessierende Phänomen planmäßig erzeugt sowie variiert (**Manipulation**) und dabei gleichzeitig systematische und/oder unsystematische Störfaktoren durch hierfür geeignete Techniken ausschaltet bzw. kontrolliert (**Kontrolle**).“ (Sarris, 1990)

4.2.2.1. Hauptmerkmale

- Datengewinnung über systematische Beobachtung (AV). Das Experiment ist also ein nicht-iteratives Verfahren (iterativ = unsystemat.).
- Experimenteller Eingriff: Manipulation einer UV.
- Ausschalten bzw. Kontrolle von Störvariablen: Sicherstellen, dass nur UV Veränderungen der AV bewirkt (experimentelle, statistische und versuchsplanerische Kontrolltechniken): interne Validität.

4.2.2.2. Die Logik des Experiments

Das Ziel des Experiments ist die Verifizierung einer Kausalursache. Die UV (Zeitpunkt 1) soll also kausal verantwortlich für die Veränderung der AV (Zeitpunkt 2) sein.

In der Zeit zwischen T1 und T2 geschehen jedoch viele Dinge, es muss also sichergestellt werden, dass nur die UV wirkt. Dies kann dadurch geschehen, dass Situationen hergestellt werden, die sich nur in der Ausprägung der UV unterscheiden. Verändert sich dann die AV, können die Ursachen hierfür nur die Unterschiede in der UV gewesen sein.

Dies lässt sich durch eine systematische Manipulation der UV sowie der Kontrolle von Störvariablen erreichen.

4.3. Das Max-Kon-Min-Prinzip

Das MAX-KON-MIN-Prinzip geht auf Kerlinger (1973) zurück und fasst den Grundgedanken des Experiments schematisch zusammen:

- MAXimiere die Primärvarianz: Wähle die Stufen der UV so, dass möglichst große Unterschiede in der AV zwischen den Gruppen entstehen, die diese Stufen erhalten.
- KONtrolliere die Sekundärvarianz: Sorge dafür, dass bekannte Störvariablen in allen Gruppen gleich wirken (interne Validität) und bestimme deren Einfluss, d.h. die Varianz, die sie erzeugen.
- MINimiere die Fehlervarianz: Vermeide Fehler auf Seiten der Versuchssituation (Konstanthalten der Bedingungen), der Datenerfassung (Beobachter-Reliabilität, Messinstrumente) und der Datenverarbeitung (doppelte Eingabe).

Im Folgenden sollen verschiedene Kontrolltechniken zur Umsetzung des Max-Kon-Min-Prinzips betrachtet werden.

4.3.1. MAXimiere die Primärvarianz

Wähle die Stufen der UV so, dass möglichst große Unterschiede in der AV zwischen den Gruppen entstehen, die diese Stufen erhalten. Das Ziel ist es also, die Effekte der UV durch die Versuchsplanung möglichst maximal zum Vorschein zu bringen.

Kontrolltechniken:

- Wahl von extremen experimentellen Bedingungen (Extremgruppenverfahren).
- Wahl von mehrfaktoriellen Designs (≥ 2 UVn)
- Wahl von mehreren experimentellen Bedingungen (>2 Stufen)

Die Anzahl der zu wählenden Stufen der UVn hängt vom vermuteten Zusammenhang von UV und AV ab. Ein linearer Zusammenhang (je...desto, wenn…dann) kann schon mit 2 Stufen erfasst werden, ein quadratischer (U-förmiger) Zusammenhang verlangt nach mindestens 3 Stufen und ein kubischer (S-förmiger) Zusammenhang mindestens 4.

4.3.2. KONtrolliere die Sekundärvarianz

Sorge dafür, dass bekannte Störvariablen in allen Gruppen gleich wirken (interne Validität) und bestimme deren Einfluss, d.h. die Varianz, die sie erzeugen.

Das Ziel der Kontrolle ist es also, Effekte von Nicht-UVn, die als Störvariablen einen systematischen Einfluss haben können, bestmöglich unter Kontrolle zu halten.

Man unterscheidet experimentelle Kontrolltechniken – wie Abschirmung, Eliminierung und Konstanthaltung – und statistische Kontrolltechniken – wie allgemeine statistische Kontrolle und kovarianzanalytische Kontrolle.

4.3.2.1. Experimentelle Kontrolltechniken

Experimentelle Kontrolltechniken werden vor der Versuchsdurchführung geplant und während dieser umgesetzt.

Abschirmung bezeichnet eine Beschränkung möglicher Störeffekte, **Eliminierung** hingegen eine vollständige Ausschaltung dieser (z.B. in Laborexperimenten). Zwischen Abschirmung und Eliminierung besteht also nur ein gradueller Unterschied.
Beispiel „Lärm“: Schließen des Fensters (Abschirmung) vs. lärmisolierter Raum (Eliminierung).

Konstanthaltung bezeichnet die Gleichhaltung von Störvariablen unter verschiedenen Versuchsbedingungen mit dem Ziel, die Effekte auf die AV für alle Bedingungen gleich stark einzustellen und so die Sekundärvarianz gering zu halten.

4.3.2.2. Statistische Kontrolltechniken

Die statistische Kontrolle setzt erst nach der Datenerhebung ein. Dabei sollte man bestimmte Aspekte beachten:

(1) Zunächst die Verteilung der Rohwerte in einer graphischen Darstellung (Rohwert-Plot) betrachten, um einen Gesamteindruck über die Form und Art der Datenverteilung zu erhalten (vgl. 4.3.2.3). Dabei sollten nicht nur die Rohwerte der Gruppen, sondern auch individuelle Rohwerte kurz überflogen werden.
(2) Die individuelle Rohdatenanalyse macht vor allem deshalb Sinn, weil Mittelwerte nicht immer die besten Repräsentanten einer Stichprobe sein müssen (z.B. bei einer bimodalen Verteilung).
(3) Kein Mittelwert ohne Dispersionsmaß (SD, KI, SE). Der Standardfehler ist sinnvoll zur Bestimmung der praktischen Signifikanz von Mittelwertsunterschieden.
(4) Überprüfung der statistischen Ausgangswerte (Baseline) bei Vorher-Nachher-Versuchsplänen.
(5) Kovarianzanalytische Kontrolle (Kovarianzanalyse): Betrachtung von Effekten auf die AV, die nicht auf die UV zurückzuführen sind. Ziel: Bereinigung der Werte der AV bzgl. der Effekte der Störvariablen.

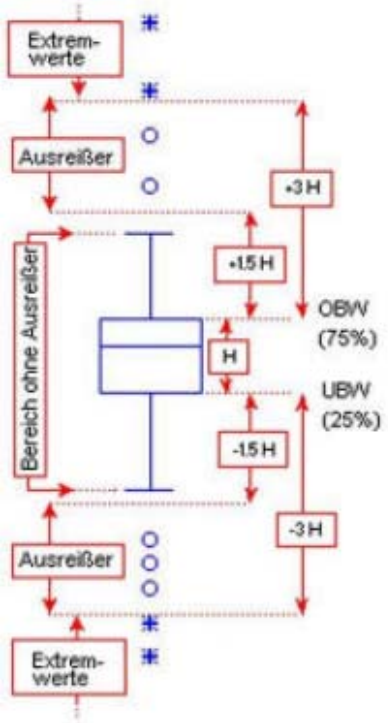
4.3.2.3. Exkurs: Boxplots

Boxplots werden von vielen Statistikprogrammen nach einem standardisierten System erstellt.

Zunächst werden 25%- und 75%-Quantile erfasst, welche den Körper der Box bilden. Darin werden Median (Strich) und Mittelwert (Punkt) angetragen.

Der Bereich ohne Ausreißer reicht auf beiden Seiten der Box maximal 1,5x deren Höhe, jedoch nur bis zum letzten empirisch gefundenen Wert.

Bemerkung: Liegt der md in der Mitte der Box kann eine Normalverteilung angenommen werden, liegt er nahe am Rand so liegt vermutlich eine andere Verteilungsform vor.



4.3.3. MINimiere die Fehlervarianz

Vermeide Fehler auf Seiten der Versuchssituation (Konstanthalten der Bedingungen), der Datenerfassung (Beobachter-Reliabilität, Messinstrumente) und der Datenverarbeitung (doppelte Eingabe).

Bemerkung: Doppelte Eingabe lässt sich über die Ausgabe von Häufigkeiten kontrollieren (Werte wie 7 oder 11 auf einer 6-stufigen Skala sollten nicht zwingend vorkommen).

Das Ziel der folgenden Techniken ist es, die Auswirkungen von unbekannten Störvariabeln so klein wie möglich zu halten.

Kontrolltechniken:

- Randomisierung
- Blockbildung
- Wiederholungsmessung

Diese drei Techniken werden auch als versuchsplanerische Kontrolltechniken bezeichnet, da jede von ihnen einen eigenen Versuchsplan hervorbringt.

4.3.3.1. Randomisierung

Unter Randomisierung versteht man die zufällige Zuweisung der Pbn zu den Gruppen und der Gruppen zu den Versuchsbedingungen. Im Mittel sollten die Gruppen also vergleichbar sein und keine systematischen Unterschiede aufweisen (Erwartungswertgleicheit).

Durch Randomisierung erstellte Gruppen sollten sich also hinsichtlich der Ausgangsbedingungen und Ausgangsmesswerte prinzipiell vergleichen lassen. Das Ziel der Erwartungswertgleichheit wird jedoch nicht nur durch die Vermeidung systematischer Unterschiede erreicht, sondern auch durch die Kontrolle der interindividuellen Varianz (Fehlvarianz) durch Mittelung.

Randomisierung sollte dann verwendet werden, wenn eine Vielzahl möglicher Störvariablen kontrolliert werden soll, über deren Effekt nichts Genaueres bekannt ist. Sie ist allerdings nur dann effektiv, wenn die Stichproben hinreichend groß sind; bei kleinen Stichproben (n<=10) ist eine gleiche Zusammensetzung der Versuchsgruppen unwahrscheinlich (dann besser Blockversuchspläne oder Wiederholungsmessungen).

Hieraus ergibt sich ein großes Problem, da schon bei einem simplen 2x2-Versuchsplan (4 Zellen) 40 Probanden untersucht werden müssten. Bei der Umsetzung des MAX-Prinzips (z.B. in einem 3x3x3-Plan) würde man schon 270 Probanden benötigen.

Bemerkung: Randomisierung kann über Münzwürfe, Würfel, Zufallszahlengeneratoren oder auch Tabellen für Zufallszahlen realisiert werden.

4.3.3.2. Blockbildung

Blockbildung bezeichnet die Umwandlung möglicher Störvariablen, die einen Einfluss auf die AV haben könnten (also mit ihr korrelieren), in eine UV, wie beispielsweise jede Form von Organismusvariablen (Alter, Intelligenz).

Das Ziel der Parallelisierung ist v. a. die Kontrolle der interindividuellen (Fehler-)Varianz. Sie lässt sich zudem bei kleinen Stichproben anwenden.

Der Grundgedanke des Vorgehens ist also die Zuordnung der Pbn zu den Versuchsbedingungen aufgrund der Merkmale, die man als potentielle Einflussgröße auf die AV erwartet.

Vorgehen:

(1) Auswahl von Pbn, die sich hinsichtlich des Parallelisierungsmerkmals gleichen.
(2) Aufstellen einer Rangreihe (bezogen auf die Ausprägung des Parallelisierungsmerkmals).
(3) Bildung von Blöcken von Pbn mit jeweils benachbarten Rangplätzen (statistische Zwillinge): Pbn eines Blocks sind sich hinsichtlich des Parallelisierungsmerkmals ähnlicher als Pbn aus unterschiedlichen Blöcken.
(4) Zuordnung der Pbn eines Blocks zu Versuchsbedingungen (per Zufall).

Nicht alle Merkmale eigenen sich zur Parallelisierung. Hier sollten nur solche verwendet werden, die eine mittlere Korrelation mit der AV aufweisen. Korrelieren die beiden Größen deutlich schwächer, so ist das Merkmal irrelevant, korrelieren sie stärker so sind sie praktisch identisch.

4.3.3.3. Wiederholungsmessung

Der Grundgedanke der Wiederholungsmessung ist die Eliminierung von interindividuellen Unterschieden zwischen Bedingungen aufgrund von Mehrfachmessung. Es soll also auch hier die interindividuelle Varianz kontrolliert werden.

Dafür werden alle Pbn unter sämtlichen Versuchsbedingungen untersucht (Within-Subject-Design). Es sind also keine expliziten Kenntnisse über Personenvariablen, die mit der AV korrelieren notwendig. Zudem ist die Versuchsdurchführung auf Basis der Wiederholungsmessung sehr ökonomisch.

Ein großer Nachteil der Wiederholungsmessung sind Zeiteffekte (Lernen etc.).

4.3.3.4. Bewertung der Kontrolltechniken

Bewertung der Technik nach Kriterium:

A) Kontrolle der Personenvariablen
B) Reproduzierbarkeit in anderen Untersuchungen (externe Validität)
C) Chance, vorhandene Effekte zu entdecken
D) Zahl der Voraussetzungen bei der statistischen Überprüfung
E) Sparsamkeit an Versuchspersonen (für Blocken/Schichten sind zusätzliche Vortests nötig)
F) Einflüsse aus mehrfacher Applikation von Behandlungen (Lernen, Ermüdung, Carry Over)

	A	B	C	D	E	F
Messwiederholung	+++	---	+++	+++	+++	+++
Blocken	++	-	++	+	0	0
Schichten	+	0	+	+	0	0
Randomisation	0	++	0	---	-	0

Bemerkung – Unterschied zwischen Blocken und Schichten: Es gibt zwei mögliche Unterscheidungen für die Techniken der Blockbildung und der Schichtung, einmal hinsichtlich des Zeitpunktes und einmal hinsichtlich der Anzahl der Probanden pro Zelle des Versuchsplans:

- Bei der Blockbildung wird nach dem Vortest je 1 Proband pro Block je einer Bedingung zugewiesen. Bei der Schicht können mehrere Pbn einer Schicht einer Bedingung zugeordnet werden.
- Blöcke werden vor der Datenerhebung gebildet, Schichten im Nachhinein.

4.3.4. Überblick: Kontrolltechniken

Experimentelle (instrumentelle Kontrolltechniken):

- Anwendung bereits vor der Datenerhebung
- Anwendung apparativer Techniken
- z.B. Abschirmung, Eliminierung, Konstanthaltung

Versuchsplanerische Kontrolltechniken:

- Anwendung vor der Datenerhebung
- Anwendung bestimmter Versuchsplanungsstrategien
- ➔ Randomisierung, Parallelisierung, Wiederholungsmessung

Statistische Kontrolltechniken:

- Anwendung erst nach der Datenerhebung
- z.B. allgemeine statistische Kontrolle, kovarianzanalytische Kontrolle

4.4. Problemkreise Experiment

Die Aussagekraft eines Experiments hängt von drei Faktoren ab:

(1) Ist es wirklich die UV, die die Veränderungen der AV verursacht (Design des Experiments, interne Validität)?

(2) Sind die Veränderungen der AV bedeutsam, d.h. größer als zufällige Schwankungen (Planung: Max-Kon-Min-Prinzip; Prüfung: Inferenzstatistik)

(3) Für wen gelten die Ergebnisse meines Versuchs, bzw. inwieweit kann ich die Ergebnisse verallgemeinern (Operationalisierung, externe Validität)?

5. Versuchsplanung III

5.1. Klassifikation von Versuchsplänen

Versuchspläne lassen sich durch verschiedene Merkmale klassifizieren:

- Anzahl der untersuchten Stichproben
  - Ein-, zwei- vs. Mehrstichproben-Plan
  - Placebo vs. Alkohol; Placebo vs. wenig vs. viel Alkohol
- Anzahl der unabhängigen Variablen
  - Einfaktorieller vs. Mehrfaktorieller Plan
- Anzahl der abhängigen Variablen
  - Univariater vs. Multivariater Plan
- Werden dieselben, ähnliche oder verschiedene Pbn unter den Stufen der UV untersucht?
  - Abhängige Gruppen vs. Blockplan vs. unabhängige Gruppen

5.2. Experimentelle Designs

Experimentelle Designs erlauben kausale Schlüsse, da die systematische Manipulation relevanter Variablen sowie die Kontrolle von Störvariablen für eine hohe interne Validität sorgen (AV = f(UV, ~~SV~~)).

Im Bereich der experimentellen Designs unterscheidet man Versuchspläne mit Zufallsgruppenbildung, Versuchspläne mit wiederholter Messung sowie Blockversuchspläne. Jeder dieser Pläne beruht auf einer anderen Kontrolltechnik, die im Rahmen des Max-Kon-Min-Prinzips besprochen wurden (vgl. 4.3.3).

Anmerkung: Bei einfaktoriellen Zufallsgruppenplänen werden nach unten Versuchsgruppen und nach rechts die Zeit angetragen.

5.2.1. Versuchspläne mit Zufallsgruppenbildung

Bei diesen Plänen erfolgt eine zufällige Zuweisung der Pbn zu Versuchsgruppen und danach eine zufällige Zuweisung der Versuchsgruppen zu den einzelnen Bedingungen (Randomisierung). Dadurch wird das Prinzip der Erwartungswertgleichheit realisiert, was eine prinzipielle Vergleichbarkeit der Ausgangsbedingungen und Ausgangsmesswerte ermöglicht. Zunächst sollen Zweistichprobenpläne und anschließend Dreistichprobenpläne betrachtet werden.

5.2.1.1. Zufallsgruppenplan ohne Vortest (2 Stpn)

	Versuchsgruppe	Vorher-Messung	Treatment X	Nachher-Messung
R	E	-	X_E	$Y_{E\ nach}$
	K	-	(X_K)	$Y_{K\ nach}$

Bei diesem Beispiel handelt es sich um einen sehr einfachen und ökonomischen Versuchsplan.

5.2.1.2. Zufallsgruppenplan mit Vortest (2 Stpn)

	Versuchs-gruppe	Vorher-Messung	Treatment X	Nachher-Messung
R	E	$Y_{E\ vor}$	X_E	$Y_{E\ nach}$
	K	$Y_{K\ vor}$	(X_K)	$Y_{K\ nach}$

Mit diesem Versuchsplan (Vorher-Nachher-Messung / Prä-Post-Design) werden zusätzliche Informationen durch die Vorher-Messung gewonnen. Hierdurch lassen sich interindividuelle Messwertdifferenzen und Ausgangslageunterschiede kontrollieren.

Das größte Problem der Vorher-Nachher-Messung ist die Reaktivität – also die Tatsache, dass die Vorher-Messung die Wirkung des Treatments beeinflussen oder überlagern kann. Dieses Problem lässt sich über einen Zufallsgruppenplan mit teilweisem Vortest umgehen, z.B. dem Solomon-Dreigruppen-Versuchsplan.

5.2.1.3. Zufallsgruppenlan mit teilweisem Vortest (2 Stpn)

	Versuchs-gruppe	Vorher-Messung	Treatment X	Nachher-Messung
R	1	$Y_{1\ vor}$	X_1	$Y_{1\ nach}$
	2	$Y_{2\ vor}$	X_2	$Y_{2\ nach}$
	3	-	X_1	$Y_{3\ nach}$

Dieser Solomon-Dreigruppen-Versuchsplan zählt per Definition zu den 2-Stichproben-Plänen. Der große Vorteil gegenüber dem Zufallsgruppenplan mit Vortest besteht in der Abschätzbarkeit möglicher Effekte des Vortests auf die Wirkung des Treatments. Unterscheiden sich $Y_{1\ nach}$ und $Y_{3\ nach}$, so ist dieser Unterschied auf Reaktivitätseffekte zurückzuführen.

Der Solomon-Dreigruppen-Versuchsplan wird allerdings relativ selten verwendet.

5.2.1.4. Einfaktorieller Mehrstichprobenplan ohne Vortest

	Versuchs-gruppe	Vorher-Messung	Treatment X	Nachher-Messung
R	1	-	X_1	$Y_{1\ nach}$
	2	-	X_2	$Y_{2\ nach}$
	3	-	X_3	$Y_{3\ nach}$

Dieser Plan stellt lediglich eine Verallgemeinerung der Zweistichprobenpläne auf drei oder mehr Versuchsgruppen dar.

5.2.1.5. Zweifaktorieller Zufallsgruppenplan (4+ Stpn)

R

	Faktor A („Stroop-Bedingung")	*Faktor B* („Verzögerte akustische Rückmeldung")	
		B_1 „ohne„	B_2 „mit"
R	A_1 „Farbwörter lesen"	$\overline{Y}_{11}$ Gruppe 1	$\overline{Y}_{12}$ Gruppe 2
	A_2 „Farben benennen"	$\overline{Y}$ Gruppe 3	$\overline{Y}$ Gruppe 4

AV: Bearbeitungszeit in Sekunden ($\overline{Y}$)

Bemerkung: Bei verzögerter akustischer Rückmeldung (alles was man sagt, bekommt man erst ca. 0,5 s später über Kopfhörer rückgemeldet) tritt der Lee-Effekt auf: die Sprache wird immer langsamer.

5.2.1.6. Vorteile Zufallsgruppenpläne

Im Mittel ist eine Gleichheit der Merkmale in Versuchsgruppen zu erwarten, was zu einer hohen internen Validität führt. Durch einen zusätzlichen Vortest lassen sich auch interindividuelle Messwertdifferenzen und Ausgangslageunterschiede kontrollieren.

Mehrstichprobenpläne sind dabei Zweistichprobenplänen vorzuziehen. Sie besitzen sowohl eine höhere interne Validität (Max-Prinzip: breitere Aussagen möglich) als auch eine höhere externe Validität (sachrepräsentativere Aussagen möglich: in der Realität gibt es normalerweise auch nicht entweder 1,0 Promille oder nüchtern).

Auch erlauben multifaktorielle Versuchspläne, im Gegensatz zu einfaktoriellen, Aussagen über Haupt- und Wechselwirkungen zwischen untersuchten Variablen zu machen.

5.2.1.7. Nachteile Zufallsgruppenpläne

Bei kleinen Stichproben (je Gruppe n<=10) ist eine gleiche Zusammensetzung der Versuchsgruppen statistisch unwahrscheinlich. In diesem Fall sind Blockpläne oder Wiederholungsmessungen aussagekräftiger.

Bei Mehrstichprobenversuchsplänen steigt die Anzahl der Versuchsgruppen mit der Anzahl der Faktoren stark an – zudem sind Interaktionen bei drei- und mehrfaktoriellen Plänen kaum interpretierbar

5.2.2. Versuchspläne mit wiederholter Messung

Bei diesen Designs wird eine Versuchsgruppe zu verschiedenen Zeitpunkten untersucht. Auch bei diesen Plänen muss zwischen Zweistichproben- und Mehrstichprobenversuchsplänen sowie einfaktoriellen, zweifaktoriellen und mehrfaktoriellen Versuchsplänen unterschieden werden.

Anmerkung: Bei Messwiederholungsplänen werden die einzelnen Probanden nach unten, die Zeit (W = Wiederholungsmessung) nach rechts angetragen.

W

Pb-Nr.	*Faktor A* („Abstraktionsgrad“)	
	A_1 „konkret“	A_2 „abstrakt“
1	$Y_{1.1}$	$Y_{1.2}$
2	$Y_{2.1}$	$Y_{2.2}$
3	$Y_{3.1}$	$Y_{3.2}$
.	.	.
.	.	.
.	.	.
N	$Y_{N.1}$	$Y_{N.2}$
	$\overline{Y}_1$	$\overline{Y}_2$

AV: Anzahl der Bilddarbietungen bis zur richtigen Silbenzuordnung ($\overline{Y}$)

5.2.2.1. Vorteile

Aufgrund der geringen benötigten Probandenzahl handelt es sich hier um ein sehr ökonomisches Design. Zudem tritt eine geringere interindividuelle Varianz als bei Einfachmessungen auf, die Wirksamkeit experimenteller Effekte (Primärvarianz) ist also leichter nachweisbar.

5.2.2.2. Nachteil

Probleme entstehen, wenn sog. Carry-over-Effekte (Übertragungseffekte) auftreten, also evtl. ein Teil des Effekts bei Y2 auf Effekte des Treatments A1 zurückzuführen ist.

Dieses Problem lässt sich entweder durch die Wahl eines hinreichend großen Zeitabstandes oder aber auch über die komplette Ausbalancierung der Reihenfolge der Versuchsbedingungen lösen. Wenn keine komplette Ausbalancierung vorliegt, handelt es sich um ein quasiexperimentelles Design.

Bemerkung: Eine Ausbalancierung bei zwei Versuchsgruppen und zwei Zeitpunkten wird auch als Crossover-Design bezeichnet.

5.2.3. Blockversuchspläne

Versuchspläne mit parallelisierten Gruppen stellen eine Kombination aus Designs der Zufallsgruppenbildung und der Wiederholungsmessung dar. Der Unterschied von Zwei- und Mehrstichprobenplänen bezieht sich hier auf die Unterscheidung von zwei oder mehreren parallelisierten Versuchsgruppen.

5.2.3.1. Vergleich mit anderen Designs

Wie bereits erwähnt stellt der Blockversuchsplan eine Kombination der beiden anderen experimentellen Designs dar. Die Mehrfachmessung ist dadurch gegeben, dass die Bildung von Blöcken über einen Vortest erfolgt, die Zuweisung der parallelisierten Probanden zu den Gruppen geschieht jedoch durch Zufall.

5.2.3.2. Vorteile

Die Nachteile beider anderen Versuchspläne werden kompensiert. Auch bei kleinen Stichproben ist eine homogene Zusammensetzung wahrscheinlich, und auch Übertragungseffekte können nicht auftreten.

5.2.3.3. Nachteile

Vortestvariablen, die einerseits erhebbar sind und andererseits entsprechend stark mit der AV korrelieren (.4 - .6) sind schwer aufzufinden. Zudem ist mit den Blockversuchsplänen ein höherer Versuchsaufwand verbunden, der vor allem dann zu einem Problem werden kann, wenn Probanden mit bestimmten Vortestwerten ausfallen.

5.2.4. Mischversuchspläne

Mischversuchspläne sind zwei- oder mehrfaktorielle Designs, bei denen die Faktoren verschiedenen Design-Haupttypen entsprechen:

- Zufallsgruppenfaktor (R = Randomisierung)
- Faktor mit wiederholter Messung (W = Wiederholung)
- Blockfaktor (O = Block; O = Organismusvariable)

Die Symbolfolge (z.B. RO-Mischdesign) wird zur Charakterisierung der Versuchspläne verwendet. Mischversuchspläne lassen dabei alle möglichen Faktorenkombinationen zu und sind somit äußerst flexibel an die jeweilige inhaltliche Fragestellung anpassbar.

Design 4.4 Dreifaktorieller Mischversuchsplan mit Zufallsgruppenbildung für Faktor A ($p = 2$) und wiederholten Messungen auf den Trendfaktoren B ($q = 3$) und C ($r = 4$), einem Organismusfaktor: *Design RWO – 2 × 3 × 4.*

$W_{rr}O_{tr}$

Faktor A („Gehalt“)	C_1 *Faktor B* („appar. Bürokomf.“) B_1	B_2	B_3	*Faktor C* („Alter“) C_2 *Faktor B* („appar. Bürokomf.“) B_1	B_2	B_3	...	...	C_4 *Faktor B* („appar. Bürokomf.“) B_1	B_2	B_3
A_1	$\overline{Y}_{111}$ Gruppe 1	$\overline{Y}_{112}$ Gruppe 1	$\overline{Y}_{113}$ Gruppe 1	$\overline{Y}_{121}$ Gruppe 2	$\overline{Y}_{122}$ Gruppe 2	$\overline{Y}_{123}$ Gruppe 2	...	...	$\overline{Y}_{141}$ Gruppe 4	$\overline{Y}_{142}$ Gruppe 4	$\overline{Y}_{143}$ Gruppe 4
R A_2	$\overline{Y}_{211}$ Gruppe 5	$\overline{Y}_{212}$ Gruppe 5	$\overline{Y}_{213}$ Gruppe 5	$\overline{Y}_{221}$ Gruppe 6	$\overline{Y}_{222}$ Gruppe 6	$\overline{Y}_{223}$ Gruppe 6	...	...	$\overline{Y}_{241}$ Gruppe 8	$\overline{Y}_{242}$ Gruppe 8	$\overline{Y}_{243}$ Gruppe 8

AV: Schreibleistungen im Büro ($\overline{Y}$)

5.2.5. Zusammenfassung

Versuchspläne mit Zufallsgruppenbildung: Zufällige Zuweisung der Pbn zu Versuchsgruppen, danach zufällige Zuweisung der Versuchsgruppen zu den Bedingungen.

Versuchspläne mit wiederholter Messung: Untersuchung einer Versuchsgruppe zu verschiedenen Messzeitpunkten.

Blockversuchspläne: Kombination aus Designs der Zufallsgruppenbildung und der Wiederholungsmessung.

Mischversuchspläne: Eine beliebige Kombination aus den drei reinen Designs.

Es lassen sich einige Richtlinien finden, wann welcher Plan angebracht ist:

- Wenn der Zeitverlauf interessiert: Mischversuchsplan (z.B. RW).
- Wenn Patienten untersucht werden, die alle behandelt werden Müssen: Messwiederholung (abhängiger Plan im Crossover-Design).
- Wenn hoher Aufwand bei der Probandengewinnung zu erwarten ist: Abhängiger Plan.
- Wenn Testeffekte zu erwarten sind: Unabhängiger Plan.

- Wenn Wirkungen in verschiedenen Verhaltens- und Erlebensbereichen erwartet werden: Multivariater Plan.

5.3. Quasi-experimentelle Designs

Auch bei quasi-experimentellen Designs erfolgt eine systematische Manipulation relevanter Variablen, jedoch erfolgt keine vollständige Kontrolle von Störvariablen. Beispiele für quasi-experimentelle Designs sind Zeitreihenversuchspläne mit einer Gruppe oder mit statischen Gruppen, Einzelfallversuchspläne sowie Versuchspläne mit unvollständiger Ausbalancierung.

Quasi-Experimente sind nur dann empfehlenswert, wenn kein reines Experiment durchgeführt werden kann, da ihre interne Validität immer fraglich ist (IV und der Faktor Zeit).

5.3.1.1. Zeitreihenversuchspläne

Bei einem Zeitreihendesign werden die Versuchsgruppen wiederholt untersucht und behandelt, z.B. ABAB-Plan (mit A = Kontrollbedingung, B = Experimentalbedingung).

Attraktiv ist die Erweiterung auf mehrere verschiedene Gruppen: Mehrgruppen-Zeitreihendesign (mit vorgegebenen statischen Gruppen), die von Mehrgruppen-Zeitreihendesigns mit Zufallsgruppenbildung als experimentelle Designs abgegrenzt werden müssen.

5.3.1.2. Versuchspläne mit unvollständiger Ausbalancierung (bei Wiederholungsmessungen)

Wenn eine Konfundierung von UV und der gewählten Darbietungsabfolge zu erwarten ist, eine vollständige Ausbalancierung der Reihenfolge jedoch nicht möglich ist, kann eine Wiederholungsmessung auch unvollständig durchgeführt werden und zählt dann zu den quasi-experimentellen Designs.

Eine Abart dieses Designs ist das lateinische Quadrat.

5.3.1.3. Bewertung

Vorteile:

- Zeitreihenversuchspläne: Untersuchung von Prozessen
- Einzelfallversuchspläne: Brückenschlag zwischen Allgemeiner und Differentieller Psychologie.
- Ausbalancierungspläne: Bestmögliche Kontrolle der Bedingungsabfolge.

Nachteile:

- Generell: Probleme des Faktors Zeit (geringere IV).
- Einzelfallversuchspläne: z.T. fehlende inferenzstatistische Verfahren; Problem der Verallgemeinerbarkeit.

5.4. Vorexperimentelle Designs

Vorexperimentelle („ungültige“) Designs – also Schrotschuss-Design, einfache Vorher-Nachher-Messung und statischer Gruppenvergleich – wurden bereits unter 4.1 besprochen und sollen hier nicht nochmals aufgeführt werden.

5.5. Übersicht Versuchspläne

Experimentelle Designs:

- kausaltheoretische Vorhersage vorhanden
- systematische Manipulation relevanter Variablen
- Kontrolle von Störfaktoren, die die Interpretierbarkeit und Gültigkeit der Ergebnisse beeinträchtigen könnten.
- systematische Beobachtung

Quasi-experimentelle Designs:

- systematische Manipulation relevanter Variablen
- keine Kontrolle von Störvariablen

Vorexperimentelle Designs:

- explizite Einführung einer experimentellen Bedingung, aber keine Manipulation
- keine Kontrolle von Störvariablen

6.Versuchsplanung IV

Im Folgenden sollen weitere Versuchsdesigns, namentlich Ex post facto- und korrelative Designs behandelt werden

6.1. Ex post facto-Designs

Bei Ex post facto-Designs wird versucht aus nicht manipulierten (bzw. nicht manipulierbaren) Variablen einen Kausalzusammenhang abzuleiten. Es findet also keine aktive Manipulation der UV statt, sondern eine Gegenüberstellung von bereits vorgefundenen Ausprägungen (z.B. natürliche „Treatments“).

6.1.1. Beispiel

Redelmeier und Tibshirani (1997) untersuchten, ob das Nutzen eines Mobiltelefons im Fahrzeug die Unfallrate erhöht. Hierzu verwendeten sie die Daten von Polizeiberichten oder Versicherungsdaten für größere Unfälle.

In diese wurde beispielsweise versucht, soziodemographische Pseudo-UVn einzugliedern (keine aktive Manipulation), die ein kausales Wirkgefüge zeigen.

6.1.2. Bewertung

Da keine aktive Manipulation durch den Untersucher vorliegt ist die Wirkung des Treatments weniger stark abgegrenzt. Auch ist davon auszugehen, dass sich die Probandengruppen in mehr Merkmalen als der Ausprägung der UVn unterscheiden – es kann also keine Erwartungswertgleichheit der Versuchsgruppen angenommen werden und man muss mit systematischen Fehlern rechnen. Die Daten sind also streng genommen nur korrelativ zu interpretieren.

Der große Vorteil der Ex post facto-Designs ist die leichte und ressourcenschonende Durchführung sowie die Möglichkeit auch Fragen zu untersuchen, die aus ethischen Gründen nicht experimentell angegangen werden können.

6.2. Exkurs: Forschungsethik

Ein Beispiel für die Notwendigkeit ethischer Prinzipien in der Forschung stellt das Milgram-Experiment (1963) dar.

Die wichtigsten Richtlinien in der psychologischen Forschung sind:

1. Wäge Kosten und Nutzen gegeneinander ab.
2. Übernimm persönliche Verantwortung.
3. Informiere den Teilnehmer und schließe mit ihm eine Übereinkunft (die Pbn müssen wissen, dass sie untersucht werden).
4. Sei offen und ehrlich (Cover-Storys müssen nach Beendigung des Versuchs aufgelöst werden).
5. Arbeite mit freiwilligen Versuchspersonen zusammen (keine Studenten im Rahmen des Studiums oder Arbeitslose, die das Geld brauchen).
6. Nutze Versuchspersonen nicht aus.
7. Schütze die Teilnehmer vor Schaden (vgl. Milgram: psychische Folgen)
8. Kläre adäquat auf.
9. Schließe negative Folgen für die Teilnehmer aus.
10. Bewahre Vertraulichkeit.

6.3. Korrelative Designs

Die einzig mögliche Aussage von korrelativen Daten ist die Art und Intensität des gemeinsamen Variierens zweier oder mehrerer Merkmale. Im Hinblick auf gerichtete Fragestellungen interessiert vor allem, ob eine Korrelation gleichsinnig oder gegensinnig (positiv oder negativ) ist.

Kausale Wirkungsmodelle können nur bei Vorliegen von untersuchungstechnischen Vorkehrungen oder inhaltlichen Überlegungen getroffen werden. Dabei muss vor allem die Anzahl kausaler Alternativerklärungen eingeschränkt sein.

Großer Vorteil der korrelativen Ansätze ist der geringe untersuchungstechnische Aufwand, z.B. im Sinne von Kontrolltechniken.

6.3.1. Übersicht: Korrelationen

Merkmal x: Merkmal y↓	Intervall-skala	Ordinal-skala	Künstliche Dichotomie	Natürliche Dichotomie	Nominal-Skala
Intervall-skala	Produkt-Moment-K.	Rang-korrelation	Biseriale Korrelation	Punktbise-riale Kor.	Kontingenz-koeffizient
Ordinal-skala		Rang-korrelation	Biseriale Rangkor.	Biseriale Rangkor.	Kontingenz-koeffizient
Künstliche Dichotomie			Tetracho-rische Kor.	Phi-Koeffizient	Kontingenz-koeffizient
Natürliche Dichotomie				Phi-Koeffizient	Kontingenz-koeffizient
Nominal-skala					Kontingenz-koeffizient

6.3.2. Bivariate Fragestellungen

Bivariate Fragestellungen (2 AVn) beziehen sich auf den Zusammenhang von zwei Merkmalen X und Y (evtl. gerichtet). Statistisch wird ein Korrelationskoeffizient errechnet (deskriptiv) und über einen Signifikanztest geprüft (interferenzstatistisch).

Wichtig ist hierbei, dass nicht jedem Untersuchungsobjekt zwei Merkmale zugewiesen werden müssen (Lernaufwand eines Studenten und dessen Leistung in einer Klausur), sondern es muss lediglich eine eindeutige Zuordnung von Messwertpaaren erfolgen (Gewicht von Hundebesitzer und Hund).

6.3.3. Multivariate Fragestellungen

Bei multivariaten Fragestellungen werden Zusammenhänge zwischen 3 und mehr AVn untersucht, wobei entweder partielle oder multiple Zusammenhänge interessieren.

6.3.3.1. Partielle Zusammenhänge

Partielle Korrelationen sind eine Bereinigung des Zusammenhangs zwischen X und Y um dritte Variablen Z, die sowohl X als auch Y (bzw. den Zusammenhang zwischen X und Y) beeinflussen.

Ziel dieser Untersuchungen ist also das statistische Ausschalten einer Kontroll- oder Störvariable und nicht das untersuchungstechnische Ausschalten über Kontrolltechniken.

6.3.3.2. Multiple Zusammenhänge

Bei einer Studie zu multiplen Korrelationen sollen die Beziehungen zwischen einem Merkmalskomplex mit den Merkmalen X1, X2,…,.Xn und einem Merkmal Y untersucht werden. Falls die Richtung eines möglichen kausalen Zusammenhangs begründbar ist handelt es sich hierbei um Prädiktorvariablen und Kriteriumsvariable (Bsp.: Pfadanalyse).

6.3.3.3. Beispiele

Beispiel partielle Korrelation: Zusammenhang zwischen Prüfungsleistung und beruflichem Erfolg, wenn Prüfungsangst herauspartialisiert wird.

Beispiel multiple Korrelation: Zusammenhang zwischen Anzahl der Trainingsstunden, Konzentration beim Training, Intensität des Trainings etc. auf die sportliche Leistung.

6.3.3.4. Meehl'sches Paradoxon (1950)

Sind drei Ereignisse paarweise unabhängig, wobei jedoch aus jeweils zwei Ereignissen eine perfekte Vorhersage des dritten möglich ist, so kann das Meehl'sche Paradoxon auftreten: Werden in diesem Fall nur zwei der drei Merkmale betrachtet, können scheinbar offensichtliche Zusammenhänge leicht übersehen werden.

Statistisch: Werden multivariate Hypothesen in Form bivariater Hypothesen geprüft, so können entscheidende Informationen verloren gehen.

6.3.4. Das Problem der Stichprobe

Bei korrelativen Studien muss immer mit systematischen Fehlern aufgrund der Stichproben gerechnet werden. Beispiele sind:

- Nullkorrelation aufgrund nichtrepräsentativer Stichprobe
- Extremgruppen: Überschätzung des Zusammenhangs in der GG
- Gegenläufige Trends in Stichproben und Grundgesamtheit
- Scheinkorrelation durch Ausreißer

Von daher sollten bei korrelativen Studien Scatter-Plots der Rohwerte eingehend geprüft werden.

6.3.5. Weitere korrelative Ansätze

- Korrelative Versuchsanordnungen: Nicht manipulative Ansätze
- Korrelative Messmethodologie: Korrelationen zur Bestimmung der Reliabilität, Validität oder Objektivität (Testtheorie)
- Korrelativ-statistische Kontrolle: Kovarianzanalyse zur Kontrolle von bekannten Einflüssen auf UVn.
- Korrelationen als Alternative für kennwerte der zentralen Tendenz, z.B. Rho (Prodkut-Moment-Korrelation) als Alternative für Mittelwerte.

6.4. Forschungshypothesen

Zum Abschluss der Versuchsplanung sollen die behandelten Typen von Hypothesen zusammengefasst werden – also Unterschiedshypothesen, Veränderungshypothesen und Zusammenhangshypothesen.

6.4.1. Unterschiedshypothesen

Unterschiedshypothesen befassen sich meist mit Unterschieden in Maßen der zentralen Tendenz (m, md, mo). Beispiele hierfür sind:

- Maßnahme (Treatment) hat Einfluss auf die AV.
- Zwei Maßnahmen A1 und A2 unterscheiden sich in Wirkung auf AV.
- Zwei Populationen unterscheiden sich in Bezug auf AV.
- Mehrere Treatments unterscheiden sich in Bezug auf AV.
- Zwischen zwei UVn besteht eine Interaktion.

6.4.2. Veränderungshypothesen

- Treatment übt veränderte Wirkung auf AV aus.
- Treatment verändert AV in Population A stärker als in Population B.
- Veränderung der AV hängt von Drittvariablen ab.
- Intervention führt zu sprunghafter Änderung einer Zeitreihe.

6.4.3. Zusammenhangshypothesen

- Zwischen zwei Merkmalen X und Y besteht ein Zusammenhang.
- Zwischen zwei Merkmalen X und Y besteht auch dann ein Zusammenhang, wenn der Einfluss eines dritten Merkmals Z außer Acht gelassen wird (Partialkorrelation).
- Zwischen mehreren Prädiktorvariablen (X1, X2,…, Xn) und einer (Y) oder mehreren (Y1, Y2,…, Yn) Kriteriumsvariablen besteht ein Zusammenhang (multiple Korrelation).
- Zusammenhänge zwischen vielen untersuchten Variablen lassen sich auf wenige hypothetisch festgelegte Faktoren zurückführen.

7. Datenquellen I: Befragung

7.1. Was ist Befragung?

Im Sinne des Alltagsverständnisses ließe sich Befragung als verbale Kommunikation zwischen Personen definieren.

7.1.1. Wissenschaftliche Befragung

Eine wissenschaftliche Befragung ist definiert als Informationsfluss zwischen Personen (ohne Fokussierung auf den verbalen Kanal), der einer systematischen Vorbereitung und Durchführung unterliegt.

Die Ergebnisse einer Befragungssituation sind dabei abhängig von:

- Sozialem Vorgang, d.h. Wechselwirkungen zwischen Personen
- Zielgerichtetheit der Befragung
- Verwendeten Hilfsmitteln (z.B. Telefoninterview) und Bedingungen der unmittelbaren räumlichen Umwelt (z.B. Ruhe vs. Stress)
- Normative Orientierung (Ausbildung von Verhaltenserwartungen)

Die zwei Kennzeichen der methodischen Kontrolle in einer wissenschaftlichen Befragung sind die Zielgerichtetheit und die systematische Durchführung.

Anmerkung: In einem Fragebogen sollte das Prinzip der Extraspektion verwirklicht werden, sprich die Beschreibung der objektiven Außenreize. Also kein „Was denken sie, wie Person X abschneiden wird?“ (Introspektion) sondern „Wie wird Person X abschneiden?“ (Extraspektion).

7.1.2. Einsatz

Die wissenschaftliche Befragung wird zur Überprüfung theoretischer Zusammenhänge eingesetzt. Dabei müssen Merkmale der befragten Person bei der Gestaltung des Befragungsinstruments berücksichtigt werden (z.B. Alter, Geschlecht, Bildungsstand).

Eine Befragung erfolgt dabei praktisch immer an einer Stichprobe aus der Gesamtpopulation. Unerlässlich ist es, für jeden Probanden möglichst gleiche Kontextbedingungen zu schaffen (gleiche Befragungsinstrumente, Bedingungen der unmittelbaren räumlichen Umwelt etc.). Es muss also jeder einzelne Befragungsschritt kontrolliert ablaufen.

7.1.3. Der Interviewee

Der Befragte wird häufig auch als Interviewee bezeichnet. Ihm muss klar sein,

- über welchen Gegenstand er berichten soll (Ob und wie ist der Gegenstand beim Befragten repräsentiert?).*
- welches Sprachsystem er verwenden soll (Welche Eigenschaften besitzt die verwendete Skala?).
- mit welcher Intention (Urteilshaltung) er berichten soll; z.B. sachorientierte Beschreibung vs. wertorientierte Stellungnahme. Dies kann durch die Instruktion verdeutlicht werden.

*) Anmerkung – Non-Attitutes: Pbn geben auch dann Antworten, wenn sie vom Inhalt der Frage keine Ahnung haben können (z.B. Wie bewerten Sie den Trade-Act zwischen Island und Dänemark?).

7.2. Klassifikation von Befragung

Eine Befragung lässt sich hinsichtlich des Ausmaßes der Standardisierung, des Autoritätsanspruchs des Interviewers, der Art des Kontakts, der Anzahl der befragten Personen, der Anzahl der Interviewer sowie der Funktion des Interviews klassifizieren.

7.2.1. Ausmaß der Standardisierung

Man unterscheidet strukturierte, halb-strukturierte und unstrukturierte Befragungen. Standardisierung bedeutet dabei die Vorgabe der Abfolge und des Wortlauts der Fragen.

Eine standardisierte Befragung eignet sich vor allem für umgrenzte Themengebiete und Themenbereiche, für die bereits Vorwissen existiert.

Standardisierung bedeutet dabei nicht, dass bestimmte Antwortalternativen vorgegeben werden; sowohl geschlossene als auch offene Fragen können in einem strukturierten Interview verwendet werden.

7.2.1.1. Geschlossene vs. offene Fragen

Geschlossene Fragen sollten verwendet werden, wenn Antworten auf bestimmte Gebiete begrenzt werden sollen und davon ausgegangen werden kann, dass die Probanden alle Alternativen verstehen.

Offene Fragen bieten sich bei stetigen Merkmalen (Alter) an, oder wenn die Antworten vorher nicht absehbar sind.

7.2.1.2. Mittelweg: Halbstandardisierte Befragung

Bei einer halbstandardisierten (halb-strukturierten) Befragung liegt ein Interviewer-Leitfaden vor, der die Art und Inhalte der Befragung nicht vollkommen festlegt.

Teilweise werden auch bestimmte Fragen vorgegeben, die in einer beliebigen Reihenfolge abgearbeitet werden können.

7.2.2. Autoritätsanspruch des Interviewers

Man unterscheidet weiche, neutrale und harte Interviews.

Ein **weiches Interview** basiert auf Prinzipien der Gesprächspsychotherapie von Rogers: nicht-direktiv, empathisch, wertschätzend und selbstkongruent. Das Ziel dieses Interviews ist es, reichhaltige und aufrichtige Antworten zu bekommen.

Ein **neutrales Interview** betont die informationssuchende Funktion der Befragung. Fragender und Befragter werden als gleichwertige Partner angesehen.

Bei einem **harten Interview** nimmt der Fragende eine autoritär-aggressive Haltung ein: häufiges Anzweifeln der Antworten, schnelles Aufeinanderfolgen der Fragen. Ziel eines harten Interviews ist das Überrennen von Abwehrmechanismen der Befragten.

Ein Beispiel für ein hartes Interview stellen die Untersuchungen von McKinsey (50er) dar, der u.a. die Masturbationspraktiken in Amerika untersuchen wollte, und die prüde Haltung seiner Probanden mit einem sehr einschüchternden, harten Interview überwinden wollte.

7.2.3. Art des Kontakts

Ein Interview kann direkt/persönlich, telefonisch oder schriftlich erfolgen (klassische Varianten).

Persönliche Befragung (Face-to-Face-Interview): Auch komplexe, persönliche bzw. die Privatsphäre betreffende Inhalte können thematisiert werden. Diese Art der Befragung ist jedoch mit einem enormen Aufwand verbunden.

Telefonische Befragung (Telefoninterview): Das Telefoninterview kann schnell und preiswert (bei geringer Verweigerungsquote) durchgeführt werden und wird vom Befragten als anonymer und weniger bedrängend erlebt als die persönliche Befragung. Allerdings können hier nur vergleichsweise einfache Frageinhalte verwendet werden.

Schriftliche Befragung (Paper-and-Pencil): Die schriftliche Befragung ist äußerst kostspielig und nimmt eine unkontrollierte Erhebungssituation in Kauf. Zudem zeigt sich hier eine äußerst heterogene Rücklaufquote.

7.2.3.1. Neue Befragungstechniken

Für die drei klassischen Formen des Kontakts wurden computerunterstützte Techniken entwickelt (wichtig sind v. a. CAPI, CATI, EMS und CASI):

- Computerunterstützte persönliche Befragung (Computer Assisted Personal Interviewing; CAPI)
- Computerunterstützte telefonische Befragung (Computer Assisted Telephone Interviewing; CATI)
- Computerunterstützte schriftliche Befragung
    - Computer Assisted Self Interviewing (CASI)
    - Computerized Self-Administered Questionnaire (CSAQ)
        - Electronic Mail Survey (EMS)
        - Disk by Mail (DBM; veraltet)
- Touchtone Data Entry (Eingabe über Telefontastatur; veraltet) bzw. Voice Recognition (veraltet)
- Fax Surveys

7.2.3.2. Vergleich der Befragungstechniken

Klassisch Computerunterstützt	Persönlich CAPI	Telefonisch CATI	Schriftlich EMS	 CASI
Modus	akust/vis	akustisch	visuell	akust/vis
Reihenfolge	sequenziell	sequenziell	simultan	sequenziell
Zeitdruck	hoch	möglich	nein	gering
Zusatzerklärungen	sehr mögl.	möglich	unmöglich	evtl. mögl.
Interviewermerkmale	sehr wichtig	wichtig	nein	nein
Subjekt. Vertraulichk.	sehr gering	gering	hoch	sehr hoch
Externe Ablenkung	???	???	???	???/gering

Anmerkung: Unter subjektiver Vertraulichkeit bezeichnet man den Grad, zu dem die Probanden glauben, dass ihre Daten vertraulich behandelt werden (nach Schwarz et al., 1991).

7.2.4. Anzahl der befragten Personen

Hier wird zwischen Einzelinterview und Gruppeninterview unterschieden.

Einzelbefragung: Sollte vor allem dann eingesetzt werden, wenn der jeweilige Themenbereich das individuelle Eingreifen des Fragenden nötig machen kann (z.B. Gebiete ohne Vorwissen), oder wenn Gruppeneffekte erwartet werden (z.B. Leistungsdruck, sozialer Druck).

Gruppenbefragung: Befragte machen Angaben auf einem Antwortbogen. Dadurch entstehen nur geringe Kosten und es kann eine einheitliche Befragungssituation zumindest für die Mitglieder der Gruppe geschaffen werden.

7.2.5. Anzahl der Interviewer

Es wird zwischen einem Interviewer, Tandem und Hearing unterschieden.

Ein Interviewer: Aus ökonomischer Sicht die geschickteste Lösung.

Tandem-Interview: Zwei Interviewer befragen einen Interviewee. Üblicherweise führt dabei einer der beiden Protokoll oder kontrolliert die Fragen des anderen. Das Tandem-Interview ist bei anspruchsvollen Befragungssituationen sinnvoll, beispielsweise bei der Erfragung des Wissens von Experten.

Hearing / Board-Interviews: Mehr als zwei Personen befragen einen oder mehrere Kandidaten. Hierbei bietet sich die Möglichkeit zur gegenseitigen Ergänzung der Interviewer, die Situation wird vom Befragten allerdings als belastend wahrgenommen.

7.2.6. Funktion des Interviews

Ein Interview kann **informationsermittelnd** sein (Erfassung von Fakten, Zeugeninterviews, Panel-Befragungen), oder aber auch **informationsvermittelnd** (z.B. Beratungsgespräch).

7.3. Allgemeines psychologisches Grundmodell

Die Antwort einer Person wird als Folge von Aspekten der Frage (z.B. Wortlaut, Reihenfolge), Merkmalen des Befragten (z.B. Motivation, Kompetenz) sowie des Kontexts der Befragungssituation (z.B. Art und Zweck des Interviews) beeinflusst.

Da sich sowohl Situation, Frage als auch der Befragte selbst auf die Antwort auswirken, sind bei der Befragung extrem viele Fehlerquellen vorhanden. Im Folgenden sollen die wichtigsten Einflussfaktoren näher betrachtet werden.

7.4. Einflussfaktoren auf die Antwort

7.4.1. Aspekte der Frage

Hier sind folgende Aspekte zu berücksichtigen:

- Reihenfolge der Fragen
- Formulierung der Fragen
- Formatierung der Antwortskala
- Kategorienanzahl, Mittelkategorien, Verankerung, Balancierung und optische Gestaltung

7.4.1.1. Reihenfolge der Fragen

Bei gemeinsamem Kontext von aufeinander folgenden Fragen wird versucht, die zweite Frage mit anderen Informationen zu beantworten (gemäß den Regeln der Kommunikation).

Man spricht hierbei von Subtraktionseffekten (Given-new Contract): Antworten auf Fragen mit identischem Inhalt korrelieren nur mit 0.16, wenn sie in gemeinsamem Kontext gestellt werden, mit 0.55, wenn sie in getrenntem Kontext gestellt werden.

Strack et al. (1989) stellten beispielsweise die Fragen „Wie zufrieden sind Sie mit Ihrem Dating?“ und „Wie zufrieden sind sie mit ihrem Leben allgemein?“ direkt aufeinander folgend bzw. getrennt voneinander. Werden sie als Folge dargeboten, so erfolgt die Antwort auf die zweite Frage ohne Bezugnahme auf Informationen, die bei der ersten Antwort berichtet wurden.

7.4.1.2. Formatierung der Fragen

Semantisch eindeutige Informationen können anders interpretiert werden (z.B. erlauben vs. verbieten).

Auf die Frage „Sollte die USA öffentliche Angriffe auf die Demokratie erlauben?“ antworten 74% der Probanden mit Ja, auf die Frage „sollten die USA…verbieten?“ jedoch 54%.

Trotz semantischer Gleichheit scheint es hier also einen gefühlten Unterschied zu geben: Verneinung führt nicht immer zum Gegenteil des Originalsatzes (Schumann & Presser, 1981).

Diese Befunde müssen vor allem beim Umpolen von Fragen berücksichtigt werden.

7.4.1.3. Formatierung der Antwortskala

Die Ausgestaltung der Antwortskala beeinflusst die Interpretation der Frage (z.B. -5,…0,…+5 vs. 0…10). Dabei wird z.B. intuitiv zwischen Fragen nach Alltagsproblemen (Daily Hassles) und allgemeinen Grundproblemen unterschieden. Auch sind hier die Verankerungsheuristik sowie die abweichende Deutung negativer Skalenwerte zu berücksichtigen.

7.4.1.4. Skalenvarianten

Eine Skala kann unipolar (schwach…stark extravertiert) oder bipolar (extravertiert…introvertiert) sein und es existiert eine Vielzahl unterschiedlicher Skalen, beispielsweise: Nummeralskala, Verbalskala, Kombinierte Nummeral-/Verbalskala, Symbolskala, graphische Skala (visuelle Analogskala) oder Standardskala.

- Nummeralskala
  - Verwendung negativer Skalenwerte umstritten
  - Können Urteile in Zahlen ausgedrückt werden? (Abstraktheit)
  - Anfälliger für Urteilseffekte als Verbalskalen
  - Durch verbale Verankerung präziser
- Verbalskala
  - Durch verbale Bezeichnung u.U. unpräzise
  - Äquidistanz der Kategorien nicht immer sichergestellt

  - Weniger anfällig für Urteilseffekte als Nummeralskalen
- Symbolskala (Bsp.: Thermometerskala; Kunin/Smiley-Skala)
  - Für Kinder und untrainierte Probanden geeignet
- Graphische Skala (z.B. visuelle Analogskala)
  - Eine Linie von ca. 5 cm Länge soll unterteilt werde (evtl. mit Verankerung)
  - Hoher Auswertungsaufwand (z.B. über künstliche Einführung von Kategorien)
  - Anfangs: hohe Unsicherheit der Probanden
  - Später: höhere Motivation der Befragten, Antwortabgabe leichter und schneller als bei Nummeralskala
  - Feinere Abstufungen der Urteile möglich
  - Entspricht Intervallniveau
  - Geringe Erinnerungseffekte: Befragte können angegebene Position nur schwer erinnern
- Standardskala (durch Beispiele verankerte Skala; Bsp.: Allgemeine Skala zur zentralen Aktiviertheit)
  - Hoher Entwicklungsaufwand
  - Problematisch bei besonders großen/kleinen Streuungen innerhalb der Stichprobe
  - Plastische Wirkung für Befragten

7.4.1.5. Empfehlung I: Verankerung?

Es bietet sich an, unteren und oberen Pol zu kennzeichnen und das dazwischen liegende Intervall in eine fünfstufige Kategorienskala zu unterteilen (Häufigkeitsskala von nie bis immer, Intensitätsskala von nicht merkbar bis unerträglich, Bewertungsskala von völlig ungenügend bis optimal).

7.4.1.6. Empfehlung II: Kategorienzahl?

Die Anzahl der Kategorien sollte bei einer einzelnen Ratingskala 9 +/- 2 betragen, bei einer Itembatterie 5 +/- 2. Manche Autoren empfehlen auch generell 5 Kategorien und schlagen nur bei akademischen / studentischen Stichproben eine größere Kategorienzahl vor.

Die Anzahl ist des Weiteren von verschiedenen Merkmalen abhängig:

- Differenziertheit des Messgegenstandes
- Differenzierungsfähigkeit des Urteilers
- Messsetting (Formulierungsaufwand der Antwort steigt mit der Anzahl der Kategorien)
- Messintention
- Größe des Effekts (kleiner erwarteter Effekt = viele Kategorien)

7.4.1.7. Empfehlung III: Mittelkategorie?

Mittelkategorien (= ungerade Anzahl von Kategorien) werden häufig verwendet, wenn sie explizit vorgegeben werden, vor allem bei schwachen Einstellungen. Die Unterscheidung von neutraler Position, Unwissenheit und Bequemlichkeit ist dabei nur schwer möglich, was durch eine Ausweichkategorie (optisch getrennt; „weiß nicht“) teilweise kompensiert werden kann.

Ohne Mittelkategorien (= gerade Anzahl von Kategorien) wird die Entscheidung eines Probanden erzwungen, wodurch ein kleiner Effekt leichter gefunden wird. Es besteht jedoch auch die Gefahr häufiger Missings bzw. der Überschreitung des Wissensstandes der Probanden.

7.4.1.8. Empfehlung IV: Balancierung und opt. Gestaltung

Die Anzahl der positiven und negativen Kategorien sollte gleich sein (balancierte Skala).

Die einzelnen Kategorien sollten räumlich voneinander getrennt sein, da sonst die Gefahr von Kreuzen zwischen zwei Kästchen besteht; eine horizontale Anordnung sollte aus ökonomischen Überlegungen vorgezogen werden.

7.4.2. Merkmale des Befragten

Die Antwort wird maßgeblich durch Motivation und Kompetenz des Befragten beeinflusst.

7.4.2.1. Kompetenz

Die Kompetenz der Probanden muss bei Überlegungen zur Zielgruppe der Befragung berücksichtigt werden.

7.4.2.2. Motivation

- Demand-Effekte (guter Proband)
- Self-Disclosure: fehlende Bereitschaft, Angaben zu machen
- Impression-Management: Motive zur Selbstdarstellung und Streben nach Konsistenz („Ach, das Item kam doch schon mal, was habe ich denn auf Seite 1 angekreuzt?“)
- Soziale Erwünschtheit, sowohl als Trait als auch als State (keine negativen Angaben über Ausländer bzw. Frauen)

7.4.3. Kontext der Befragungssituation

Sowohl Art der Befragung, Zweck der Befragung als auch Merkmale des Interviewers wirken sich auf die Antwort aus.

7.4.3.1. Zweck der Befragung

Der Zweck der Befragung (Täuschung!) beeinflusst die Bereitschaft, an der Untersuchung teilzunehmen und die Ausführlichkeit und Inhalte der Antworten. [Psychologisches Institut vs. IZVW; 5 → 6]. Beispiele sind:

- Befragung als wahrgenommener Täuschungsversuch durch VL: Ablehnung der Teilnahme
- Befragung als wahrgenommenes Bürgerreferendum: Vertreten extremer Positionen
- Befragung als wahrgenommenes intimes Gespräch: Fragen des Pb nach Meinung das VL.

7.4.3.2. Merkmale des Interviewers

Es existiert ein starker Einfluss der wahrgenommenen Merkmale des Interviewers (Alter, Geschlecht, Rasse) v. a. auf merkmalsrelevante Antworten.

7.5. Ausfälle bei Befragungen

7.5.1. Item-Non-Responder

Ursachen für eine Nicht-Antwort bei einzelnen Items können sein:

- Verweigerung der Auskunft
- Nicht-Informiertheit
- Meinungslosigkeit
- Unentschlossenheit

Item-Non-Response tritt vor allem bei sehr persönlichen, intimen Fragen auf. Besondere Klassen von Probanden sind hierbei besonders betroffen: unsichere Personen, ältere Menschen und Personen mit geringem Sozialstatus.

7.5.2. Unit-Non-Responder

Unit-Non-Responder (komplette Verweigerung der Auskunft) lassen sich durch Auffüllen der Stichprobe oder schon anfänglich hinreichend große Stichproben kontrollieren. Non-Responder unterscheiden sich dabei aber möglicherweise systematisch von Respondern, wodurch die Ergebnisse der Befragung möglicherweise verfälscht werden.

Im Interview handelt es sich bei Unit-Non-Respondern vornehmlich um alte Menschen, Frauen, und Personen mit geringer Schulbildung.

In der schriftlichen Befragung sind vor allem Personen mit geringer Schulbildung, geringerer Intelligenz, geringem Interesse am Forschungsthema oder fehlender Beziehung zum Untersucher betroffen.

7.5.3. Verweigerungsquoten

Bei persönlicher Befragung und telefonischem Interview betragen die Verweigerungsquoten etwa 10%. Bei der schriftlichen Befragung liegen sie hingegen zwischen 10% und 90%, über den jeweils zu erwartenden Wert ist dabei keine Aussage möglich. Zudem sind die Antworten später antwortender Personen meist unzuverlässiger.

Bei computerunterstützten Techniken sind die Verweigerungsquoten etwa gleich, die Rücklaufgeschwindigkeit ist dagegen wesentlich höher.

7.5.3.1. Rücklaufquoten für schriftliche Befragung

Hohe Rücklaufquoten finden sich bei:

- Stichproben, die Umgang mit schriftlichen Texten gewohnt sind
- Aktuellen, interessanten Themen
- Ansprechender Gestaltung (Frageformulierung, Layout, persönliches Anschreiben)
- Vorherigen Ankündigungsschreiben (2x so hoch) oder kurzen Anrufen (3x so hoch)
- Angabe eines Rücksendedatums (Deadline)
- Anreiz (Belohnung, Gewinnspiel)

7.5.3.2. Rücklaufstatistik

Die Verwertbarkeit der Ergebnisse der schriftlichen Befragung hängt nicht von der Höhe des Rücklaufs ab sondern von der Zusammensetzung der Stichprobe der Responder.

Möglichkeiten zur qualitativen Kontrolle von Rückläufen sind Gewichtungsprozeduren bei Über-/Unterrepräsentation einzelner Merkmale der Stichprobe im Vergleich zur Grundgesamtheit oder eine gezielte telefonische, schriftliche oder persönliche Nachbefragung der Non-Responder.

8. Datenquellen II: Beobachtung

8.1. Was ist Beobachtung

Beobachtung ist die grundlegende Methode der Datengewinnung in den empirischen Wissenschaften. Sie besteht aus dem Sammeln von Erfahrung im nichtkommunikativen Prozess mit Hilfe sämtlicher Wahrnehmungshilfen.

Beobachtung umfasst verschiedene Methoden, wie beispielsweise das Ablesen von Skalen, Auswerten von Fragebögen, Beobachten von Verhalten oder das Ablesen von Testergebnissen (z.B. Reaktionszeiten). Dabei muss immer zwischen unsystematischer, naiver **Alltagsbeobachtung** und **wissenschaftlicher Beobachtung** unterschieden werden. Letztere zeichnet sich durch **Zielgerichtetheit** und **methodische Kontrolle** aus (vgl. Befragung).

Die Zielgerichtetheit ergibt sich als direkte Konsequenz aus der beschränkten Informationsverarbeitungskapazität des Beobachters. Dieses Kriterium impliziert, dass der Beobachter eine Theorie über den Beobachtungsgegenstand hat.

Die methodische Kontrolle umfasst den Kontext der Beobachtung (Wo wird beobachtet?), das Beobachterverhalten (Wahrnehmung als aktiver Prozess) sowie das Speichern der Beobachtung (Zugriff auf die Ergebnisse muss jederzeit mögl. sein: Forderung nach der Entwicklung von Kategoriensystemen).

8.2.Beobachtungssysteme: Kodierung

Grundlegend wird zwischen Verbal- und Nominalsystemen unterschieden.

8.2.1. Verbalsysteme

Verbalsysteme sind möglichst genaue, freie verbale Beschreibungen von Verhaltensweisen. Da nichts vorgegeben ist, sind Verbalsysteme umfassend aber untereinander auch wenig vergleichbar.

8.2.2. Nominalsysteme

Nominalsysteme beruhen auf der Codierung von Verhaltensweisen nach einem vorgegebenen Schema über einen Katalog möglicher Verhaltensweisen, die möglichst genau definiert und beschrieben sind. Für jede Verhaltensweise muss ein Zeichen (Code) festgelegt werden.

Werden Nominalsysteme verwendet, um die Beobachtung zu strukturieren, so findet also schon vor der Beobachtung eine Kategorisierung von Verhalten statt. Innerhalb der Nominalsysteme unterscheidet man zwischen Zeichensystemen und Kategoriensystemen.

8.2.2.1. Zeichensysteme

Die Zeichen eines Zeichensystems schließen sich nicht gegenseitig aus, d.h. mehrere Zeichen können pro Beobachtungseinheit verwendet werden. Sie sind normalerweise nicht vollständig; für manche Beobachtungseinheiten ist also keine Codierung möglich.

Ein Beispiel ist das Facial Action Coding System (FACS) von Paul Ekman, bei dem die Mimik hinsichtlich verschiedener Merkmale wie Form der Augenbraue, Stellung der Augen, Form des Mundes etc. beschrieben wird.

Zeichensysteme kommen dabei mit einer geringen Zahl von Beobachtungskategorien aus, jedoch besteht die Gefahr einer Überlastung des Beobachters.

8.2.2.2. Kategoriensysteme

Die Kategorien eines Kategoriensystems schließen sich gegenseitig aus, d.h. pro Beobachtungseinheit wird nur ein Zeichen (Kategorie) verwendet. Über Kategoriensysteme ist jedes Verhalten codierbar (pro Beobachtungseinheit kann eine Kategorie vergeben werden), u.U. zumindest über eine Restkategorie.

Im Vergleich zu Zeichensystemen ergibt sich jedoch evtl. eine deutlich erhöhte Kategorienzahl. Ohne Hilfsmittel (wie eine Videokamera, Tonband) sollten dabei maximal 30 Kategorien verwendet werden, mit Hilfsmitteln gibt es keine Obergrenze.

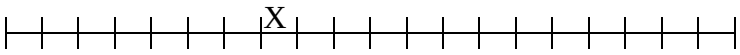
Auf dem Beobachtungsprotokoll sollten immer die Rahmenbedingungen (z.B. Zeit, Geschlecht, Beobachter) und die exakte Kodierung der Kategorien vorgegeben werden, evtl. zusätzlich ein Ablaufplan.

8.3. Quantifizierung des Verhaltens

Die Frage nach der Ausprägung eines bestimmten Verhaltens lässt sich durch bestimmte Beobachtungseinheiten angehen. Wichtig sind hierbei vor allem Time-Sampling und Event-Sampling.

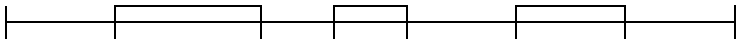
8.3.1. Time-Sampling (Zeitstichprobe)

Die Beobachtungseinheit ist ein festes Zeitintervall; pro darin liegendem Zeitintervall wird dabei codiert, welches Verhalten aufgetreten ist. Das Time-Sampling liefert also annähernde Informationen über Häufigkeit und Dauer eines Verhaltens, da nur ja/nein-kodiert wird, ob ein Verhalten auftritt, nicht jedoch wie häufig oder intensiv. Über die Wahl der einzelnen Intervalle lässt sich jedoch grob eingreifen.



8.3.2. Event-Sampling (Ereignisstichprobe)

Die Beobachtungseinheit ist eine Verhaltensweise, von der Beginn und Ende festgehalten werden. Das Event-Sampling liefert also exakte Informationen über Häufigkeit und Dauer eines Verhaltens. Hierbei handelt es sich um eine präzise Aufgabe, sodass das jeweilige Verhalten immer nur bei einer Person zur gleichen Zeit betrachtet werden kann. Zudem kann Event-Sampling nur bei längeren Verhaltensweisen verwendet werden.



8.3.3. Beobachtungseinheit: Empfehlungen

Da das Time-Sampling nur annähernde Informationen über Häufigkeit und Dauer des Verhaltens liefert sollte das Zeitintervall sinnvoll in Abhängigkeit von der Dauer der Verhaltensweise, die beobachtet werden sollen, festgelegt werden.

Das Event-Sampling sollte nur auf wenige Verhaltensweisen beschränkt oder ganz vermieden werden.

8.3.4. Erweiterung: Ratingverfahren

Bisher wurden Angaben zu Häufigkeit und Dauer eines Verhaltens gemacht. Über Ratingskalen kann jedoch zusätzlich die Intensität eines Verhaltens erfasst werden.

8.4. Fehler und Güte bei Beobachtungen

8.4.1. Beobachterfehler

- Überschreitung der Grenzen der Leistungsfähigkeit: Ermüdung, Langeweile, Aufmerksamkeitsschwankung, Überlastung (Vigilanzproblematik vs. Überlastung).
- Unklarheit über Ziel der Beobachtung: Beobachter muss selbst eine Auswahl bzgl. des zu beobachtenden Verhaltens treffen. Durch diese Interpretation ist die Objektivität der Beobachtung gefährdet.
- Unklare Definition der Kategorien: Der Beobachter muss jede Kategorie individuell präzisieren (evtl. Beispielinhalte).
- Mangelndes Training der Beobachter: Mangelnde Beherrschung des Kategoriensystems oder Abweichung des Beobachterverhaltens von geplanten Verhalten.

8.4.2. Erwartungseffekte (generell vs. speziell)

8.4.2.1. Generelle Effekte

Der Rosenthal- oder Pygmalion-Effekt (Self-Fulfilling Prophecy) bezieht sich auf die unbeabsichtigte Beeinflussung des Pbn durch verbales und nonverbales Verhalten des Beobachters.

8.4.2.2. Spezielle Effekte

Unter **zentraler Tendenz** (Tendenz zu Mitte) versteht man die häufigere Verwendung von mittleren Kategorien bei Beobachtung und Befragung.

Die **Milde-Tendenz** ist ein ähnliches Phänomen und bezeichnet die systematische Verzerrung der Kategorien in Richtung „geringerer Extremität“ (v.a. bei sozialer Erwünschtheit).

Der **Primacy-Recency-Effekt** tritt vor allem bei Aufzeichnungen nach Ende der Beobachtung auf: Nur Verhaltensweisen am Anfang und Ende des Beobachtungszeitraumes werden erinnert. Primacy bezieht sich dabei v.a. auf Inhalte des Langzeitgedächtnisses, Recency auf Inhalte des KZG.

Halo-Effekt: z.B. unzulässige Generalisierung von beobachteten Verhaltensweise auf erwartete Persönlichkeitsmerkmale.

8.4.3. Verbesserung der Beobachterleistung

Eine Verbesserung der Beobachterleistung kann nur dann erfolgen, wenn an der entsprechenden Stelle des Beobachtungsprozesses (vor, während oder nach der Beobachtung) eingegriffen wird. Die Optimierung ist dabei nicht nur auf den Beobachter selbst, sondern auch auf das Kategoriensystem bezogen.

8.4.3.1. Vor der Durchführung: Beobachtertraining

- Verbesserung der Beobachterleistung (Benchmark: Feedback über Werte anderer Beobachter).
- Angleichung des Hintergrundes von Beobachtern (Hintergrundwissen schaffen).
- Verbesserung des Ratertrainings (z.B. Übungsmaterial, Regelspezifikation).

8.4.3.2. Während der Durchführung der Beobachtung

- Einfachere Informationsverarbeitung (keine Interpretationen)
- Veränderung der Skalenbeschreibung (kombinierte Verbal- und Nummeralskalen).
- Verwendung von Beispielen („ist gemeint“ und „ist nicht gemeint“).
- „Merkmal für Merkmal“ (anstelle eines Globalurteils)

8.4.3.3. Nach der Durchführung: Auswerteprozeduren

- Ausschluss von Beobachtern und/oder Beobachtungsgegenständen
- Verwendung von Mittelwerten über mehrere Rater (bei hinreichender Beobachterzahl).
- Verwendung von zusammengefassten Werten (anstelle Werten für Einzelkategorien).

8.4.4. Reliabilität der Beobachtung

8.4.4.1. Retest-Reliabilität

Die Retest-Reliabilität entsteht aus einem intraindividuellen Vergleich, und gibt also die Übereinstimmung bei wiederholter Durchführung (Stabilität und Konsistenz) an.

Zu beachten ist hierbei, dass nicht zweimal die Beobachtungssituation hergestellt wird, sondern die Beobachtung selbst: Das Video wird ein zweites Mal angesehen; nur die Güte der Beobachtung selbst interessiert.

8.4.4.2. Interraterreliabilität

Über den interindividuellen Vergleich bezieht sich die Interraterreliabilität auf die Übereinstimmung verschiedener Beobachter. Das Grundproblem hierbei ist, dass Beobachter und Beobachtungssystem nicht zu trennen sind: Eine hohe Reliabilität ergibt sich, wenn zwei Beobachter perfekt kodieren, aber auch dann, wenn zwei Beobachter gleichermaßen falsch kodieren.

Für die Interraterreliabilität lassen sich Reliabilitätskoeffizienten bestimmen. Dies kann sowohl auf Basis intervallskalierten Ratingskalen geschehen, als auch wenn die Beobachtung auf einem nominalskalierten Kategoriensystem beruht.

8.4.5. Reliabilitätskoeffizienten bei Kategorien

8.4.5.1. Kappa

Das Kappa-Maß spiegelt eine zufallskorrigierte Übereinstimmung der Beobachter wider und berechnet sich folgendermaßen:

$$Kappa(\kappa) = \frac{P_{beobachtet} - P_{Zufall}}{1 - P_{Zufall}}$$

$P_{beobachtet}$ steht dabei für die prozentuale Übereinstimmung in der Beobachtung, P_{Zufall} für die erwartete Übereinstimmung, wenn ein zufallsgesteuerter Entscheidungsprozess der Beobachter vorliegen würde.

Der Wertebereich von Kappa liegt zwischen -1 und +1, wobei ein Kappa von 0 für eine Übereinstimmung auf dem Zufallsniveau steht. Werte < 0 sprechen für eine schlechtere Reliabilität als ein zufallsgesteuerter Entscheidungsprozess. Kappa-Werte $> 50\%$ gelten als zufriedenstellende Übereinstimmung, Kappa-Werte $> 70\%$ als gute Übereinstimmung.

Sprachlich bedeutet ein Kappa von .70 eine 70%ige Übereinstimmung der Beobachter, inhaltlich eine zufallskorrigierte Beobachterübereinstimmung, d.h. die Beobachter sind 70% besser als der Zufall.

8.4.5.2. 2 Beobachter, 2 Kategorien

		Auftreten nein	ja
Zeitpunkt	t1	2	0
	t2	2	0
	t3	1	1
	t4	0	2
	t5	2	0
	Σ	7	3

$P_{beobachtet}$ (prozentuale Übereinstimmung) = 4/5 = 80%

P_{Zufall}: Was wäre wenn zufällig vergleichbar oft Ja/Nein geraten worden wäre: Quadrate der rel. Randh'keiten addieren.
$P_{Zufall} = (7/10)^2 + (3/10)^2 = 0{,}58$

$$Kappa(\kappa) = \frac{P_{beobachtet} - P_{Zufall}}{1 - P_{Zufall}} = \frac{0{,}80 - 0{,}58}{1 - 0{,}58} = 0{,}52$$

➔ zufrieden stellende Übereinstimmung

8.4.5.3. 2 Beobachter, mehrere Kategorien

		Kategorie A	B	C
Zeitpunkt	t1	2	0	0
	t2	0	0	2
	t3	1	1	0
	t4	0	2	0
	t5	2	0	0
	Σ	5	3	2

$P_{beobachtet}$ = 4/5 = 80%

P_{Zufall}: = $(5/10)^2 + (3/10)^2 + (2/10)^2$ = 0,38

$$Kappa(\kappa) = \frac{P_{beobachtet} - P_{Zufall}}{1 - P_{Zufall}} = \frac{0{,}80 - 0{,}38}{1 - 0{,}38} = 0{,}67$$

➔ zufrieden stellende Übereinstimmung

8.4.5.4. Mehrere Beobachter, mehrere Kategorien

		Kategorie A	B	C
Zeitpunkt	t1	2	0	2
	t2	0	0	4
	t3	1	1	2
	t4	4	0	0
	t5	0	1	3
	Σ	7	2	11

P_{Zufall}: = $(7/20)^2$ + $(2/20)^2$ + $(11/20)^2$ = 0,435

$P_{beobachtet}$: Tabelle mit allen mögl. Beobachterkombinationen anlegen.

Bei vier Beobachtern P, Q, R und S (wie in der obigen Tabelle) ergibt sich folgendes Übereinstimmungsmuster:

		Kombination PQ	PR	PS	QR	QS	RS	Σ
Zeitpunkt	t1	X					X	2
	t2	X	X	X	X	X	X	6
	t3	X						1
	t4	X	X	X	X	X	X	6
	t5	X	X		X			3
							Σ	18
*: m =	[Σ(Zeilensummen)]/ Anzahl der Zeitpunkte						M	3,6
*: P_{beob} =	m / Anzahl der Beobachterpaare						$P_{beobachtet}$	0,60

$$Kappa(\kappa) = \frac{P_{beobachtet} - P_{Zufall}}{1 - P_{Zufall}} = \frac{0{,}60 - 0{,}435}{1 - 0{,}435} = 0{,}292$$

➔ keine zufrieden stellende Übereinstimmung

8.5. Selbst und Fremdbeobachtung

8.5.1. Selbstbeobachtung

8.5.1.1. Probleme

Da der Beobachter weiß, was beobachtet werden soll kann das Problem der **Reaktivität** entstehen, also der Veränderung des Beobachtungsgegenstandes durch die Beobachtung selbst.

Da der Beobachter zudem gleichzeitig Tätigkeit und Beobachtung ausführt, können die Grenzen seiner **Verarbeitungskapazität** überschritten werden. Eine Lösung ist das Training von lautem Denken (Thinking Aloud Technique), welche in Würzburg entwickelt wurde (s.u.).

Bei retrospektiver Beobachtung treten **Erinnerungseffekte** (z.B. Primacy-Recency) auf. Auch sind Auslassungen und Verzerrungen durch emotionale Einflüsse (Mood congruent memory) von Bedeutung.

Im Rahmen der Selbstbeobachtung sind auch nicht alle Prozesse **beobachtbar**, wie beispielsweise automatische Prozesse oder während des Schlafens.

Zudem sind die Ergebnisse der Beobachtung nicht **nachprüfbar**, es ist also keine Aussage über die Güte der Beobachtung möglich.

8.5.1.2. Anwendung

Die Selbstbeobachtung sollte aufgrund der Vielzahl von Problemen nur zur Hypothesengenerierung, nicht jedoch zur Hypothesenprüfung verwendet werden.

8.5.1.3. Exkurs: Würzburger Schule

Das Prinzip der Introspektion stellte einen Meilenstein der Forschungsmethodik und einen Gegenpol zur Black-Box-Metapher dar. In der Würzburger Schule wurde dabei u.a. die Technik des Lauten Denkens (Thinking Aloud Technique) entwickelt, die das Berichten über eigene Gedanken derartig automatisiert, dass es nicht mehr mit den Prozessen des Denkens interferiert.

8.5.2. Fremdbeobachtung

Im Folgenden sollen Aspekte der Fremdbeobachtung beschrieben werden. Hierbei unterscheidet man:

- Natürlich vs. künstlich
- Wissentlich vs. unwissentlich (offen vs. verdeckt)
- Teilnehmend vs. nicht-teilnehmend
- Direkt vs. indirekt
- Vermittelt vs. unvermittelt

8.5.2.1. Natürlichkeit

Diese Unterscheidung betrifft die Frage, ob Verhalten in einer natürlichen Umgebung und bei spontanem Auftreten betrachtet wird oder unter speziell dafür hergestellten Bedingungen provoziert wird. Dieser Aspekt zielt also auf die Unterscheidung von Labor- und Feldforschung ab.

Vorteile der Feldforschung sind:

- Natürliche Umgebung
- Spontanes, „normales“ Verhalten
- Besser übertragbar auf natürliches Verhalten
- Keine oder geringe Verfälschung durch Wissen um Studie

Nachteile hingegen sind:

- Störvariablen schlecht zu kontrollieren
- Manipulation von Situation und Verhalten schwierig
- Das Verhalten ist schwer zugänglich
- Die Untersuchungsbedingungen sind nicht optimal.

Feldstudien weisen also eine geringe interne Validität, dafür jedoch eine hohe externe Validität auf.

Neben der Unterscheidung von Feld- und Laborforschung können noch 3 weitere Aspekte der Natürlichkeit abgegrenzt werden:

- Beobachtung mit vs. ohne Manipulation von Variablen (Problem: Ausnutzen einer natürlichen Variation von Variablen beeinträchtigt die interne Validität).

- Beobachtung mit vs. ohne Instruktion der zu beobachtenden Personen (Problem: Ohne Instruktion tritt Verhalten u.U. zufällig erst sehr spät oder gar nicht auf).
- Beobachtung mit vs. ohne Manipulation am beobachteten System, z.B. Markierung von Tieren im Rudel oder Blickbewegungskamera (Problem: Proband bemerkt nichts von Manipulation, das soziale Umfeld möglicherweise schon.).

8.5.2.2. Wissentlichkeit

Die Unterscheidung von wissentlich und unwissentlich (offen vs. verdeckt) bezieht sich sowohl auf das Wissen der Probanden, dass sie beobachtet werden, als auch auf das Wissen, was beobachtet wird.

Beispiele für unwissentliche Beobachtungen sind Wartesituationen vor dem „eigentlichen“ Versuch oder die Beobachtung über Videokameras.

Ziel einer unwissentlichen Beobachtung ist es dabei, reaktive Effekte (Wissenseffekte) zu vermeiden.

Eingrenzen lassen sich diese Effekte beispielsweise über die Einführung einer Gewöhnungsphase, die Täuschung der Pbn über die interessierenden Variablen des Verhaltens über eine Cover-Story oder auch den Einsatz von Aufzeichnungsgeräten (Video) statt menschlichen Beobachtern.

Anmerkung – Toilettenverhalten von Männern: Steht man am Pissoir Schulter an Schulter so geht es langsamer und schlechter (weniger Druck und insg. weniger). Auch wird wenn möglich immer ein Pissoir freigehalten.

8.5.2.3. Teilnahme an Erhebung

Die Unterscheidung von teilnehmend und nicht-teilnehmend bezieht sich auf den Einfluss des Beobachters bzw. die Interaktion von Beobachter und beobachteter Person.

Probleme können bei teilnehmender Beobachtung einerseits hinsichtlich der Überstrapazierung der Verarbeitungskapazität des Beobachters und andererseits hinsichtlich der Beeinflussung des Beobachtungsgegenstandes durch den Beobachter entstehen.

Andererseits sind teilnehmende Beobachtungen häufig die einzige Möglichkeit, den Beobachtungsgegenstand auch tatsächlich beobachten zu können, z.B. bei Hierarchien in Gefängnissen.

8.5.2.4. Direktheit

Wird das Verhalten selbst beobachtet, so spricht man von einer direkten Beobachtung, bei der reaktive Effekte auftreten können. Werden jedoch Spuren oder Auswirkungen des Verhaltens betrachtet, so liegt eine indirekte (non-reaktive) Beobachtung vor.

Kennzeichen der non-reaktiven Beobachtung sind:

- Nicht das Verhalten selbst, sondern Spuren oder Auswirkungen des Verhaltens werden beobachtet.
- Die Versuchsperson weiß in der Regel nicht, dass sie Daten produziert hat (keine VP-VL-Interaktion).

- Häufig ist keine Individualzuweisung der Daten möglich, sondern lediglich prozentuale Angaben.
- Der Zugriff auf die Daten verändert diese nicht (non-reaktiv).

Beispiele für non-reaktive Beobachtungen sind physische Spuren (z.B. abgetretene Teppiche als Gütezeichen von Bildern in Museen), Graffiti, Schilderdichte, Hausordnungen, Archive, Statistiken oder Presseartikel.

Auch sind provozierte Verhaltensweisen denkbar, wie beispielsweise die Lost-Letter-Technique (Rücklaufquote von Briefen mit unterschiedlichen Adressaten) oder die Wrong-Number-Technique. Bei der letzteren wird im Hotel angerufen und nach einer bestimmten Person gefragt. Dabei wird manipuliert, ob diese fiktive Person einen weißen oder afroamerikanischen Vornamen trägt (Rassenvorurteile). Als AV wird die Hilfsbereitschaft auf eine nachfolgende Frage verwendet.

Die Interpretation der Daten von non-reaktiven Beobachtungen verlangt jedoch eine genaue Verhaltenstheorie (Welches Verhalten erzeugt die Daten?), und eine Stichprobentheorie (Wer kann die Daten erzeugt haben?).

Insgesamt werden 4 Grade der Direktheit unterschieden: echt non-reaktiv (fehlende Verbindung), einseitige Verbindung, einseitig verdeckt sowie vollständig offen:

- Fehlende Verbindung: Nur Spuren existieren; die VP generiert ein Datum, der VL greift darauf zurück. Im Nachhinein ist keine Individualzuweisung möglich.
- Einseitige Verbindung („aufgeweicht“ non-reaktiv): Wie oben, allerdings hat der VL Einfluss auf die VP, wobei diese nicht weiß, dass sie beobachtet wird. Beispiele sind provozierte Daten (Lost-Letter- und Wrong-Number-Technique).
- Einseitig verdeckt: Der Versuch hat für die VP eine andere Bedeutung als für den VL (Cover-Story).
- Vollständig offen: Reaktives Messverfahren; die übliche Beobachtung.

8.5.2.5. Mittel der Erhebung

Bei einer vermittelten (vs. unvermittelten) Beobachtung wird das Verhalten, das beobachtet werden soll in irgendeiner Form gespeichert (Audio oder Video). Zusätzlich stellt sich die Frage, ob die Daten jederzeit zugänglich sind oder nicht.

Vorteile der vermittelten Beobachtung:

- Beliebige Abrufbarkeit des beobachteten Geschehens
- Unbegrenzte Speichermöglichkeiten
- Wieder- bzw. Weiterverwendbarkeit der gespeicherten Daten (die Beständigkeit einer CD beträgt dabei 10 Jahre, die einer DVD hingegen nur etwa 1 Jahr).

„Die Nachteile der vermittelten Beobachtung ergeben sich aus den Eigenschaften des Aufzeichnungsgerätes.“ Jedes Gerät kann schließlich nur eine Untermenge der vorhandenen Variablen aufzeichnen (z.B. Tonband nur akustische Signale).

Generell gilt: Wenn vermittelt beobachtet wird, sollte zusätzlich zum Aufzeichnungsgerät mindestens 1 Beobachter eingesetzt werden; wird dagegen unvermittelt erhoben, so sollten mindestens zwei Beobachter eingesetzt werden.

8.6. Problemkreise

Bei einer Beobachtung gibt es eine Vielzahl von Fehlerquellen, aber damit auch Potential für eine gute Beobachtung. Eine gute Beobachtung wird dabei u.a. von folgenden Gesichtspunkten ausgemacht:

- Definition des Beobachtungsgegenstandes
  ➔ Welches Verhalten ist eigentlich interessant bzw. entspricht der Fragestellung?
- Erstellung und Überprüfung eines Beobachtungssystems
  ➔ Übersetzung des Beobachtungsobjekts in ein Zeichen
- Entscheidung für ein Quantifizierungsverfahren
  ➔ Wie soll Häufigkeit, Dauer und Intensität bestimmt werden? Time- vs. Event-sampling? Evtl. zusätzliche Ratingskala?
- Auswahl der Beobachtungssituation
- Training der Beobachter (Ziel: möglichst objektive Beobachtung)
- Durchführung der Beobachtung
- Überprüfung der Güte der Beobachtung (Reliabilität über Kappa)

9. Datenquellen III: Apparative Techniken

Apparative Techniken umfassen zum eine psychophysiologische Messinstrumente und zum anderen Apparate zur Verhaltensmessung (z.B. Skinner-Box, Reaktionszeiten).

9.1. Grundannahme

„Die Herstellung kausaler Beziehungen zwischen Gehirn, Körper und Verhalten erfordert die simultane Beeinflussung von physiologischen und psychologischen Variablen." (Birbaumer & Schmidt, 2002, S. 486)

Zur Untersuchung dieser Beziehung wurden drei unterschiedliche Forschungsstrategien entwickelt:

- Manipulation des physiologischen Substrats, Verhalten als UV
  Insbesondere in der biologischen Psychologie im Tierversuch. Untersuchung von Zusammenhängen zwischen Hirn**strukturen** und Verhalten
- Manipulation des Verhaltens, Änderungen des Substrates als AV
  Insbesondere in der Psychophysiologie im Humanbereich. Zusammenhänge zwischen Hirn**prozessen** und Verhalten.
- Interaktiver Ansatz (Hybridansatz, Kreismodell)

9.2. Psychophysiologische Methoden

9.2.1. Biosignale

Der Gegenstand physiologischer Methoden sind meist elektrische Indikatoren (Biosignale). Biosignale können alle physikalisch messbaren, kontinuierlich oder nahezu kontinuierlich registrierbaren Körperfunktionen sein. Man unterscheidet:

- **Direkte bioelektrische Signale** (resultieren aus Aktivität von Spannungsgeneratoren im Körperinneren, die mit elektrischen Änderungen einhergehen, wie z.B. Herzschlag, Hirnaktivität)
- **Indirekte bioelektrische Signale** (z.B. Hautleitfähigkeit)
- **Nicht-elektronische Biosignale** (z.B. Blutdruck, Atemfrequenz, Magenmotalität (Kontraktion), Temperatur; lassen sich in bioelektrische Signale umwandeln).

9.2.1.1. Charakteristika von Biosignalen

Biosignale werden über die Ausprägungen von Amplitude, Frequenz und Wellenform definiert. Je nach Fragestellung kann auch nur eine Ausprägung von Interesse sein (z.B. Frequenz von R-Zacken im EKG für Herzfrequenz).

9.2.1.2. Übersicht Biosignale

Physiologisches System	Messmethoden	Messfühler
Kardiovaskuläres System	Elektrokardiogramm (EKG)	Elektroden
	Atemtiefe	Atemgürtel (Brust- und

	Atemzeitwerte	Bauchatmung!)
	Blutdruckmessung	Blutdruckmessgerät (Manschette drückt Vene zusammen)
	Plethysmographie: Messung der peripheren Durchblutung (Auch: Messung der sexuellen Erregung)	Fotowiderstand oder -transistor: Lichtquelle scheint direkt durch Finger / Wasserverdrängung
Motorisch-muskuläres System	Elektromyographie (EMG)	Oberflächenelektroden
Haut- und Schweißdrüsen	Hautwiderstand (alt) Hautleitfähigkeit (neu)	Oberflächenelektroden
Zentralnervöser Bereich	Elektroencephalographie (EEG)	Oberflächenelektroden
Sensorisches System	Augenbewegungen: Elektrookulographie (EOG)	Oberflächenelektroden (rechts, links, über und unter dem Auge)

9.2.1.3. EEG

Die Orte der EEG-Ableitung werden nach dem Jasperschen 10/20 System festgelegt. Die Strecke von Nasion zu Inion wird als 100% definiert und die Elektroden in Schritten von zunächst 10% und schließlich 20% der Gesamtstrecke angebracht (Kappe erleichtert anbringen).

Bei jeder Messung von elektrischen Signalen können Artefakte auftreten. Ein Beispiel ist das 50-Hz-Netzbrumm, welches durch elektrische Einstreuung entsteht (z.B. wenn das EEG in der Nähe einer Spannungsquelle (Steckdose) betrieben wird). Das Signal muss also aufbereitet werden (Filterung).

9.2.2. Typische Messanordnung

Die typische Messanordnung physiologischer Messungen besteht aus einer Messquelle (Biosignalen), Messfühlern (Elektroden, Wandlern), Apparaten zur Signalverarbeitung (Filterung und Verstärkung), sowie zur Aufzeichnung, Darstellung und Speicherung.

9.2.2.1. Messfühler: Elektroden

Man unterscheidet bipolare Ableitungen (EMG) und unipolare Ableitungen (Puls im EKG und EEG). Zudem muss zwischen Oberflächenelektroden und Subdermalelektroden unterschieden werden.

9.2.2.2. Messfühler: Wandler

Nichtelektrische Biosignale müssen über Wandler in solche umgewandelt werden. Beispiele sind mechanische Äußerungen oder Druckänderung, wie etwa bei Atemgürteln, Thermofühlern oder dem Plethysmographen.

Analog-Digital-Wandler wandeln ein wert- und zeitkontinuierliches analoges Signal in ein digitales Signal (Ziffern oder Zahlen). Man unterscheidet dabei zeitkontinuierliche, wertdiskrete und zeit- und wertdiskrete Zahlenfolgen.

9.2.2.3. Signalverarbeitung: Verstärkung

Ausgangsspannungen bioelektrischer Signale liegen im Bereich von μV – 100 mV. Daher ist eine Verstärkung (um den Faktor 10^3 - 10^6) notwendig.

9.2.2.4. Signalverarbeitung: Kontrolle und Filterung

Problematisch sind Störgrößen (Artefakte), die entweder gefiltert oder kontrolliert werden müssen. Kontrollmöglichkeiten sind:

- Abschirmung des Raumes bzw. des Probanden
- Verwendung abgeschirmter Leitungen
- Vorverstärkung nahe am Ableitort (um Übertragungsartefakte zu entkräften)
- Prinzip der Differenzverstärkung
- Verwendung einer Masse-Elektrode
- Elektronische Filter

Die Differenzverstärkung beruht auf den unterschiedlichen Ausbreitungsgeschwindigkeiten von Störgrößen (Ausbreitung im elektrischen Feld) und Biosignalen (langsame Ausbreitung im Körper). Da Störsignale also fast gleichphasig an beiden Elektroden ankommen, werden diese über Differenzverstärkung weniger bedeutsam. Trotzdem muss noch eine zusätzliche Filterung erfolgen.

Filter wirken nicht nach einem 0-1-Prinzip, die Filterung ist hingegen ein kontinuierlicher Prozess. Ein Tiefpass-Filter bewirkt eine Dämpfung bei hohen Frequenzen und kann somit Rauschen unterdrücken (gängigste Filterart). Ein Hochpass-Filter bewirkt eine Dämpfung tiefer Frequenzen.

Neben Hochpass- und Tiefpass-Filtern unterscheidet man noch Bandpass- (Selektive Passage ausgewählter Frequenzbänder) und Bandsperr-Filter (selektives Herausfiltern spezifischer Frequenzbänder (z.B. 50-Hz-Netzburmm)).

Als Filtergrenze wird dabei immer der Wert ermittelt, bei dem noch 70,7% der Ausgangsamplitude (50% der Energie) vorhanden sind.

9.2.2.5. Aufzeichnung und Darstellung: Eichung

Je größer die Anzahl der Verarbeitungsstufen vom Biosignal bis hin zur endgültigen (nummerischen) Verarbeitung, desto wichtiger ist die Eichung der gesamten Messung.

Unter Eichung versteht man die Messung eines bekannten Referenzsignals, welches durch die gesamte Anordnung geschickt wird. Eichung ist also die künstliche Herstellung eines Bezugspunktes.

9.2.3. Messprobleme

9.2.3.1. Artefakte

Unter Artefakt versteht man ein aufgefangenes Signal, welches anderen Ursprungs ist als das zu messende Biosignal. Man unterscheidet Artefakte physiologischer Herkunft, Bewegungsartefakte sowie Artefakte durch externe elektrische Einstreuung.

Artefakte beziehen sich also immer auf Probleme innerhalb der Technik.

- **Artefakte physiologischer Herkunft**
  Potenzialschwankungen und Signalstörungen von begleitenden physiologischen Prozessen; Lösung: Bessere Elektroden bzw. elektronische Komponenten. Beispiele: Augenbewegungen im EEG.
- **Bewegungsartefakte**
  Bsp.: Kopfbewegung im EEG. Lösung: Optimale Platzierung der Elektroden.
- **Artefakte durch externe elektrische Einstreuung**
  Lösung: Bessere elektronische Komponenten, Filterung und Verstärkung; Beispiel: 50-Hz-Netzburmm.

9.2.3.2. Spezifitätsproblematik

Im Gegensatz zu Artefakten bezieht sich die Spezifitätsproblematik auf den Menschen als Quelle der Biosignale. Man unterscheidet individualspezifische, stimulusspezifische und motivationsspezifische Reaktionen.

- **Individualspezifische Reaktion**
  Personen reagieren auf physiologischer Seite unabhängig vom Stimulus in einer für sie typischen Reaktionsweise (Janke).
- **Stimulusspezifische Reaktion**
  Alle Individuen reagieren auf einen Stimulus in ähnlicher Weise.
- **Motivationsspezifische Reaktion**
  Unter einem bestimmten Motivationszustand reagieren alle Personen in ähnlicher Weise.

Bei biopsychologischen Untersuchungen sind diese drei Anteile immer zu berücksichtigen.

9.2.3.3. Ausgangswertproblematik

Das Ausgangswertgesetz von Wilder (1931) besagt: „Je stärker vegetative Organe aktiviert sind, desto stärker ist ihre Ansprechbarkeit auf hemmende Reize und desto schwächer ist ihre Ansprechbarkeit auf aktivierende Reize."

Es gibt also eine negative Korrelation zwischen Ausgangswert und Veränderungswert. Auch der Regressionseffekt B fällt unter die Ausgangswertproblematik. Veränderungswerte enthalten also immer einen systematischen Fehler in Abhängigkeit von ihrem Ausgangswert.

Statistisch lassen sich diese Probleme folgendermaßen kontrollieren:

- Differenzbildung (Behandlung - Baseline)
- Prozentuale Veränderung ((Behandlung - Baseline)/Baseline)
- Kovarianzanalytische Methoden
- Regressionsanalytische Methoden

9.2.3.4. Innere und äußere Variablen

Bei allen physiologischen Messungen ist immer zu beachten, dass sowohl innere Variablen wie Motivation, Stimmung aber auch Lebensalter und Geschlecht als auch äußere Variabeln wie Tageszeit, Raumtemperatur, relative Luftfeuchtigkeit etc. auf die Versuchsperson einwirken.

Es gibt daher Empfehlungen zur Einrichtung eines Labors, wie z.B. angenehme Farbtöne (Pastell), Fenster, angenehme Einrichtung (inkl. Teppich) sowie (falls irgend möglich) eine Tür.

9.3. Verhaltensmessung

- weggelassen -

10. Ingos Klausurtipps I

10.1. Einführung

- Beispiele für Hypothesenkriterien:
  - „Bei starkem Zigarettenkonsum kann es zu einem Herzinfarkt kommen“ $(G \quad K \quad \bar{F})$
  - „Es gibt Kinder die niemals weinen.“ $(\bar{G} \quad \bar{K} \quad \bar{F})$
  - „Frauen sind kreativer als Männer.“ $(G \quad K \quad F)$
  - „Mit zunehmender Müdigkeit sinkt die Konzentrationsfähigkeit.“ $(G \quad K \quad F)$
- Validität
  - Regressionseffekt B (negative Rückkopplung)
  - Campbell & Stanley (1963,1970): Externe Validität und SVn; „Interne Validität ist eine notwendige, aber nicht hinreichende Voraussetzung für externe Validität.
  - Interne Validität: AV = f(UV, SV); Störeinflüsse ausschalten
  - Externe Validität = Generalisierbarkeit
- Spiralenmodell nach Sarris: simultanes, nicht sukzessives Modell

10.2. Forschungsformen und Stichproben

- Folie 10: Vor- und Nachteile des Webexperiments
- Folie 37: Grundgesamtheit und Stichprobe

10.3. Versuchsplanung I

- Varianzarten: Primär-, Sekundär-, Fehlvarianz (Definition; Folie 11)
- Varianzanalyse generell
- Interpretation des F-Bruchs (Folie 42)
- Beim Grenzfall von semi-disordinalen zu disordinalen WWen auch hinschreiben, dass es sich um einen Grenzfall handelt.

10.4. Versuchsplanung II

- Folie 15 (Hauptmerkmale des Experiments: systematische Beobachtung, Manipulation und Kontrolle)
- Max-Kon-Min-Prinzip (MAX, Kontrolltechniken, v.a. Folie 25 und 29)
- Problemkreise, Zusammenfassung (Folie 38 und 39)
- MIN-Kontrolltechniken

10.5. Versuchsplanung III

- IV und Faktor Zeit (schon wieder)

10.6. Versuchsplanung IV

- Übersicht Korrelationen
- Meehl’sches Paradoxon (1950)

10.7. Datenquellen I

- Klassifikation von Befragung (Folie 6)
- Folie 43, 44: Wer verweigert Antworten? Stereotype raushängen lassen: unsichere und alte Personen, Pbn mit niedrigem Sozialstatus
- Folie 45: Verweigerungsquoten (ca. 10% außer bei schriftlicher Befragung; dort heterogen)

10.8. Datenquellen II

- Definition von Beobachtung (Folie 3)
- Quantifizierung der Beobachtung (Folie 8)
- Reliabilität einer Beobachtung: Folie 14+ (!!!!)
- Berechnung von Kappa bei mehreren Beobachtern und mehreren Kategorien
- Tafeln für Reliabilitätskoeffizienten selbst aufstellen können
- Problemkreise (Folie 47)

10.9.Datenquellen III

- nichts

11. Ingos Klausurtipps II

Ingos Weisheiten aus der Krüger-Vertretungsstunde.

11.1. Empirisches Vorgehen

11.1.1. Vorbemerkungen

11.1.1.1. Spiralenmodell nach Sarris

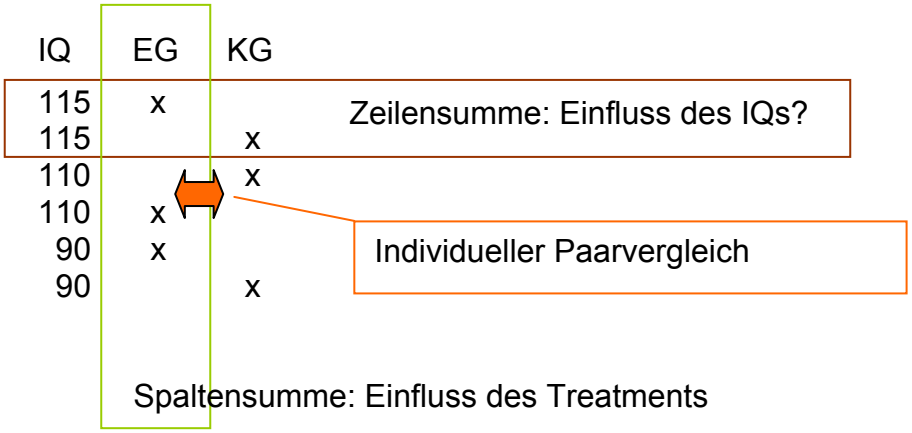
Das Spiralenmodell nach Sarris beschreibt den Forschungsvorgang als simultanen Prozess: alle Stufen sind eng miteinander verzahnt. Auch: „Wo komme ich her und wo will ich hin?“ (Totzke, 2002, 2003, 2004, 2005, 2006, 2007).

11.1.1.2. Experiment und Korrelation

Experimentelle und korrelative Ansätze sind nicht spezifischen Gebieten der Psychologie zuzuordnen – beide können überall zu finden sein. Auch lassen sich praktisch alle Fragestellungen auf beide Arten untersuchen.

Beispiel: „Beeinflusst Intelligenz die Lernleistung?“ – korrelatives Vorgehen einfach: IQ messen, Lernleistung messen, korrelieren.
Experimentell? Hier bietet sich ein Blockplan an: Blocken nach Intelligenz, EG mit Lernstrategie, KG ohne Lernstrategie lernen lassen.

Vorteile des Blockplans:



➔ Nicht nur Gruppenvergleiche (Zeilen, Spalten) sind möglich, sondern auch **Einzelvergleiche** (vergleich innerhalb eines statistischen Zwillings)! Es ist also auch eine Auswertung auf Individualebene möglich.

11.1.1.3. Forschungshypothesen

Die Unterscheidung von experimentellem und korrelativem Vorgehen findet sich auch in Form der Hypothesen wider: Unterschiedshypothesen und Veränderungshypothesen werden auch als experimentelle Hypothesen bezeichnet, Zusammenhangshypothesen hingegen sind korrelativen Designs zuzuordnen.

Auch die verschiedenen Hypothesen lassen sich ineinander überführen.

11.1.2. Datenauswertung

Der typische Fehler hinsichtlich der Datenauswertung ist, sofort nach der Datenerhebung mit dem Rechnen (Prüfen) zu beginnen. Zwischen den beiden Vorgängen liegen mindestens 2 Stufen:

1.) Kodierung und Datenübertragung (Datenaufbereitung, z.B. Scores bestimmter Fragen umpolen etc.)
2.) Fehlerkontrolle und evtl. Fehlerbereinigung (99 auf Skala von 1-6 ist etwas seltsam: Häufigkeiten ausgeben lassen)
3.) Neukodierung (Summenscores, Mittelwerte…)
4.) Statistische Analyse

Ein weiterer typischer Fehler ist es, den Test in den Vordergrund zu stellen. Wichtiger ist jedoch die eigentliche Forschungsfrage, auf deren Basis das entsprechende statistische Verfahren auszuwählen ist. Auch sollte hierbei beachtet werden, ob das jeweilige Verfahren verstanden ist.

➔ "Die Datenanalyse sollte der in der Planung festgelegten Weise entsprechen.".

11.1.2.1. Fehler nach Sarris

„Welche Fehler werden bei der statistischen Datenauswertung nach Sarris normalerweise gemacht?" [1. bezieht sich auf die Folie zu Boxplots, 2.-5. beziehen sich auf die Folie „Statistische Kontrolltechniken II" – F 27 in Versuchsplanung II]

1.) Die Verteilung der Messwerte wird nicht beachtet (Boxplots geben erste Hinweise auf Symmetrie, Schiefe, Streubreite…)
Liegt der Median (q_2) beispielsweise nicht in der Mitte des Intervalls [q_1, q_3] liegt sicher keine symmetrische Verteilung vor und damit auch keine Normalverteilung: keine ANOVA. Und selbst wenn der Median in der Mitte dieses Intervalls liegt, ist über die Verteilungsform nicht viel gesagt (Stichwort: bimodale Verteilung).
2.) Individuelle Rohdatenanalyse: Verteilung nicht betrachtet (zumindest Boxplots)
3.) SE nicht berücksichtigt: Der SE (SD/Wurzel n) gibt Auskunft über die Streuung der Mittelwerte.
4.) Überprüfung der Ausgangswerte bei Vorher-Nachher-Messung nicht beachtet (beispielsweise: prozentuale Veränderung, Differenzen). Ein Beispiel für einen Störeinfluss ist der Regressionseffekt B.
5.) Kovarianzanalytische Kontrolle: Herauspartialisieren von potentiellen Einflussfaktoren.

11.1.2.2. Mängel der Interpretation

Die entsprechende Folie trägt genau diesen Titel. Wichtige Punkte sind:

- Verschleppung von Fehlern
- Konzentration auf die Perfektion einer bestimmten Phase
- Keine wissenschaftliche Kommunikation

11.2. Max-Kon-Min-Prinzip

Die folgenden Angaben sollen die Umsetzung des Max-Kon-Min-Prinzips von Kerlinger (1973) für die Frage „Macht Alkohol cool und lustig?“ demonstrieren.

UV1: 0 Promille 1 Promille 2 Promille (>2 Stufen)
UV2: Erwartung von Alkohol vs. Nicht (mehrfaktoriell)

Kontrolle von Störvariablen: Geschlecht, Körpergewicht und -größe, Zeit

Blockplan nach Extraversion / Cool&Lustig-Fragebogen

AV1 Subjektives Rating
AV2 Verhaltensbeobachtung (inkl. Baselinemessung)

IV: sehr hoch
EV: nicht ganz so toll: Erwartungsvariable sinnvoll? Alkoholkonsum als sozialer Prozess.

Die Frage nach IV / EV darf dabei nicht auf die Umsetzung des Max-Kon-Min-Prinzips bezogen werden: Streng genommen haben die beiden nichts miteinander zu tun!

Die Frage nach der Umsetzung des M-K-M zielt darauf ab, einen existierenden Effekt zu zeigen, die Frage nach der IV / EV jedoch darauf, ob tatsächlich ein Effekt vorliegt. Forschungslogisch sind dies zwei völlig unterschiedliche Bereiche.

Die Beeinflussung der EV kann auch über Zeiteffekte erfolgen (Folie „EV und der Faktor Zeit“):

- Reaktive Effekte der Experimentalsituation
- Interaktion von Vortest und UV
- Einflüsse bei Mehrfachmessungen
- Interaktion von Selektionseffekte und UV

11.3. Grundprinzipien inferenzstatistischer Datenanalyse

„Warum gibt es so viele Tests?“: Für jede Kombination aus Art des Versuchsplans und Datenniveau gibt es einen eigenen Test, der auf genau diese Situation zugeschnitten ist.

Tests können sich auf qualitative oder quantitative (Zentrale Tendenz, Dispersion, Verteilungsform) Unterschiedshypothesen oder Zusammenhangshypothesen beziehen. Vorberg und Blankenberger (1999) haben hierfür einen Entscheidungsbaum aufgestellt.

Die Grundlegende Unterscheidung wird zwischen Unterschieds- und Zusammenhangshypothesen getroffen.

11.3.1. Unterschiedshypothsen

Es muss zwischen qualitativen und quantitativen Unterschiedshypothesen unterschieden werden. Quantitative Unterschiedshypothesen werden über Tests zur zentralen Tendenz, Dispersion oder Verteilungsform geprüft.

Hier müssen auch immer die Anzahl der untersuchten Stichproben sowie die Skalenart (intervall, ordinal: quantitativ; nominal: qualitativ) berücksichtigt werden.

11.3.1.1. + quantitativ + Mittelwerte

Parametrische Tests:

- 1-2 Stp: t- und z-Test (GG-Varianz unbekannt / bekannt)
- $>$ 2 Stp: ANOVA

Nicht-parametrische Tests:

- 1 Stp (abhängig): Vorzeichentest
- 2 Stp: Wilcoxon-Test, U-Test

11.3.1.2. + quantitativ + Dispersion (Varianzen)

Kein System. Wenn Normalverteilung angenommen werden kann: Chi^2 (1 Stp), Varianzen bei zwei unabhängigen Stpn: F-Test…

11.3.1.3. + quantitativ + Verteilungsform

Diskrete Variablen: Chi^2
Stetige Variablen: Kolmogorov-Smirnoff-Test (KS-Test)

11.3.1.4. + qualitative

Kein wirkliches System, meistens aber Chi^2, bei einer Stp auch Binomialtest möglich.

11.3.2. Zusammenhangshypothesen

Die Wahl des Testverfahrens ist hier in besonderem Maße von der Datenqualität abhängig (siehe 6.3.1).

11.4. Vordiplomsfragen

11.4.1. Nummer 1

„Was sind zufallsgesteuerte bzw. nicht-zufallsgesteuerte Stichproben. Zu jeder ein Beispiel.“ Erwartungshorizont: 3 Zeilen.

11.4.2. Nummer 2

„Diskussionen über IV … zeitbedingt sein?“

- Max-Kon-Min-Prinzip spielt bei IV keine Rolle
- Wird die EV von Zeit beeinflusst? Ja! 4 Möglichkeiten; siehe „EV und Störeinflüsse“
- Welche Art von Versuchsplänen wird beeinflusst? Alle.

11.4.3. Nummer 3

„Hat die Auswahl der Pbn einen Einfluss auf die IV / EV?“ Ja, auf beide. EV: Repräsentativität, IV: Selektionseffekte.

11.4.4. Nummer 4

„In einem Artikel lesen sie: „Bei der vorliegenden Studie mit univariatem 2x3-Mischversuchsplan…“ Liegen abhängige oder unabhängige Messungen vor?“

Kann man nicht sagen! Ein Mischversuchsplan besteht mindestens aus 2 verschiedenen Designs aus Randomisierung (R), Messwiederholung (W) und Blockbildung (O).

R = unabhängige Gruppen, W = abhängige Gruppen, O = Mischung (keine Aussage möglich). In einem 2x3-Plan (2 UVn) könnten also beispielsweise abhängige und unabhängige oder abhängige und weder/noch-Gruppen enthalten sein. Insgesamt lässt sich diese Information also nicht ermitteln.

11.4.5. Nummer 5

ANOVA-Kennwerte: $F_{HWA}(2,24) = 0,7$; $F_{HWB}(1,24) = 1,24$; $F_{WW}(2,24) = 3,6$.

„Wie viele Probanden haben teilgenommen?“

Q.d.V.	df
HW A	p-1 = 2
HW B	q-1 = 1
WW	(p-1)(q-1) = 2
Fehler	24 (Nennerfreiheitsgrade)
Total	N-1 = Summe der Restlichen

In der Vorliegenden Studie haben also N = 30 Pbn teilgenommen. Einfacher: df(Fehler) = N - (pq) (= Gesamtanzahl der Probanden - Anzahl der untersuchten Gruppen).

„Welche Aussagen dürfen getroffen werden?“

HW A darf sicher nicht interpretiert werden ($F < 1$), HW B und WW vielleicht. Dies muss jedoch über das jeweilige kritische Quantil sowie den Einfluss der WW auf die Richtung von HW B bestimmt werden.

12. Die KVK

12.1. Bemerkungen

- Es muss zwischen inhaltlichen und statistischen Hypothesen unterschieden werden. Inhaltliche Hypothesen stellen die sprachliche Formulierung eines Zusammenhangs dar. Statistische Hypothesen hingegen sind die Zuspitzung der inhaltlichen Aussage auf einen statistischen Kennwert.
- Zudem muss zwischen Zusammenhangs- und Unterschiedshypothesen unterschieden werden. Statistisch gesehen benötigen erstere ein Zusammenhangsmaß (r) und letztere eine Hypothese bzgl. dem Unterschied von Lageparametern (m, md, mo)
- Kein Mittelwert ohne Dispersionsmaß (SE, KI und v.a. SD)! Auch die graphische Veranschaulichung dieser Maße ist Pflicht.

12.2. Das Internet

- Wikipedia nicht als Quelle verwenden
- Kurze Begriffsdefinitionen finden sich entweder im Dorsch oder bei www.psychologie.de bzw.:
- http://lotse.uni-muenster.de/psychologie

12.3. Kritik

- Versuchsplan klassifizieren (inkl. schematischer Darstellung)
- Rest: Siehe Blatt