Versuchsplanung

Begleittext zur Übung „Versuchsplanung“

Teil A
Von der Fragestellung zur empirisch prüfbaren Hypothese

J. Fahrenberg, C. Klein, M. Peper und P. Zimmermann

Inhaltsverzeichnis

Psychologisches Institut der Universität Freiburg
Niemensstr. 10
79085 Freiburg i.Br.

Redaktion und Layout: U. Probst

Stand 4/2000

Teil A: Grundlagen der Versuchsplanung

1. Typische Untersuchungsstrategien. Experimentelle Versuchspläne. Feldstudien. Allgemeine Prinzipien und Schritte der Versuchs-(Untersuchungs-)Planung

1.1 Experimentelle Methodik

Methodenlehre (Methodik, Methodologie) bezeichnet die Lehre von den Wegen wissenschaftlicher Erkenntnis und den Verfahrensweisen der Wissenschaften. Sie enthält Anleitungen zum planmäßigen wissenschaftlichen Vorgehen ("Methode", d.h. "der richtige Weg", als Voraussetzung für richtige wissenschaftliche Erkenntnis).

In der Methodenlehre der Psychologie sind die verschiedenen Ursprünge und Traditionen des Faches gegenwärtig: die geisteswissenschaftliche, die biologisch-naturwissenschaftliche, die verhaltenswissenschaftliche und die sozialwissenschaftliche Orientierung. So ist ein **Methodenpluralismus** wie in keiner anderen Humanwissenschaft entstanden: Hermeneutik und phänomenologische Reduktion; Deutung und Interpretation; Introspektion, Verhaltensbeobachtung und physiologische Messung; psychologischer Test und Experiment; Systemanalyse, Simulation und Modellierung; Diagnostik, Beratung und Intervention; Vorhersage und Evaluation u.a.

Gemeinsam ist diesen Methodentypen die Auffassung der **Psychologie als Erfahrungswissenschaft** und – vielleicht – eine allgemeine Vorstellung von **Wissenschaftlichkeit** (d.h. nach Stegmüller: das Bemühen um sprachliche Klarheit; die Möglichkeit der Kontrolle durch andere Wissenschaftler; Begründungen durch rationale Argumente). Als Kontrast zu den **wissenschaftlichen** Methoden der Erkenntnisgewinnung nennt Kerlinger (1973) die Methode der "tenacity" (Zähigkeit/Anhänglichkeit z. B. an eine bestimmte Ideologie), die Methode der Autorität (z. B. Expertenurteile) und die Methode der 'Intuition' (auch 'a priori'-Methode: dasjenige, was mit unserer eigenen Vernunft übereinstimmt und keiner nachfolgenden Bestätigung durch empirische Beobachtungen bedarf, wird als 'wahr' angesehen).

Der Begriff der Empirie wird uneinheitlich verwendet: **Empirie im engeren Sinn** als äußere, **intersubjektiv prüfbare (öffentliche) Erfahrung** und **Empirie im weiteren Sinn** auch als innere Erfahrung, die zwar grundsätzlich privat ist, aber wenigstens ausschnittweise (mit methodischem Training und durch Kontrollen) als reflektierte Introspektion/Selbstbeobachtung zugänglich gemacht werden kann. Dieser Gegensatz von äußerer Erfahrung (durch Beobachtung des Verhaltens anderer Menschen) und innerer Erfahrung eigener Bewußtseinsinhalte, Gefühle, Subjektivität, Intentionalität wird immer wieder anregen, neue Verfahren zu entwickeln. Die Methoden sollen dem interessierenden Phänomen **adäquat** sein und zugleich **möglichst gut kontrollierbar** sein, d. h. Fehler vermeidend. Die Fragen, **welche** Methode adäquat ist und welches methodische Training verlangt werden müssen, können u. U. schwierige Diskussionen auslösen.

Vielfach wird eine psychologische Fragestellung nicht mit einer einzigen Methode befriedigend zu beantworten sein, so daß Methodenkombinationen verwendet werden müssen. Dies gilt für die Datenerhebung (etwa die Erfassung verschiedener Facetten eines psychologischen Konstrukts, u.a. durch Beobachtungsmethoden, Interview, Fragebogenskalen), und für die allgemeine Untersuchungskonzeption, in welcher einander ergänzende Strategien (z. B. beschreibende Einzelfallstudie, Experiment, Feldstudie) komplementär genutzt werden können.

Die Methodik der Psychologie steht im Zusammenhang mit wissenschaftstheoretischen Grundsätzen und deshalb ist die Auswahl bestimmter Methoden in Forschung und Praxis nicht unabhängig von den wissenschaftstheoretischen Überzeugungen, z. T. auch philosophischen Vorentscheidungen. Im Unterschied zu einem biologisch-naturwissenschaftlichen Experiment wird eine psychologische Untersuchung in der Regel durch differentiell-psychologische (u. a. Einstellungen und Motivation) und sozialpsychologische Bedingungen (soziale Rollen und Interaktion) der Beteiligten kompliziert. Die teilnehmenden Personen sind sich der Untersuchung bewußt, d.h. sie werden eigene Hypothesen, Erwartungen, Kausalannahmen (Attributionsprozesse) und viele Schemata einführen, so daß hier spezielle Kontrollen notwendig sind. Die gesamte Versuchsplanung muß sich schließlich an den berufsethischen Grundsätzen der informierten Zustimmung und des Datenschutzes orientieren. Die Übung Versuchsplanung baut hier auf die entsprechenden Kapitel in der Vorlesung "**Geschichte, Wissenschaftstheorie und Berufsethik der Psychologie**" im 1. Semester auf (siehe Skriptum).

Mit dem folgenden Schema soll an die unerläßlichen, aber z. T. schwierigen Abgrenzungen wissenschaftlicher und unwissenschaftlicher Erkenntnis erinnert werden:

Experiment (Kausalforschung)	**Anschauung und Reduktion**	**Überzeugungssysteme**
Deduktiv-nomologische DN, induktiv-statistische IS Erklärungen und statistische Analysen	von Phänomenen (Bewußtseinsinhalte, Texte, Werke) in kritisch-reflektierter Interpretation	Naive/implizite Schemata, „Intuitionen“ und Kausaldeutungen (Attributionen)
Erklären	**Verstehen (Deuten)**	**Spekulieren**

1.2 Allgemeine Untersuchungsstrategien und Versuchsplanung

Unter **Versuchsplanung** im weiteren Sinn ist die Festlegung aller zur Durchführung und Auswertung einer empirischen Untersuchung, sei es ein experimenteller Versuch oder eine Erhebung, notwendigen Maßnahmen gemeint (einschl. Schaffung der finanziellen, organisatorischen, technischen und personellen Voraussetzungen; Prüfung ethischer und gesetzlicher Belange; Festlegung der Untersuchungsprozedur und statistischer Analysenkonzepte, siehe auch Bock, 1991). Im **engeren Sinne** bezieht sich die Versuchsplanung auf die experimentellen und quasiexperimentellen Versuche und deren statistische Auswertung. Außer den **experimentellen** Untersuchungen, die wegen ihrer hochgradigen Kontrolliertheit als allgemeines Vorbild der Methodik – zumindest in didaktischer Hinsicht – gelten, gibt es andere, in vielen Bereichen unverzichtbare Strategien.

Typologie von Untersuchungsstrategien

Isaac und Michael (1974, S. 13 siehe Tabelle 1.1) haben neun typische Strategien unterschieden: (1) Historische Studien, (2) Beschreibende Studien, (3) Entwicklungsstudien (Zeitreihen, Verlaufsstudien), (4) Einzelfall- und Feldstudien, (5) Korrelationsstudien, (6) Kausalanalytische oder "ex post facto"-Untersuchungen, (7) Echte Experimente, (8) Quasi-experimentelle Untersuchungen, (9) Aktionsforschung. In der Übung Versuchsplanung wird vorwiegend die **experimentelle** und **quasiexperimentelle Untersuchungsmethodik** behandelt.

Tabelle 1.1: Neun grundlegende Forschungsansätze nach Isaac & Michael (1974, p. 13f.)

	Strategie	Ziel	Beispiele
1)	Historische Studien	Systematische und objektive (?) Rekonstruktion vergangener Ereignisse durch Sammlung, Bewertung und Prüfung von Dokumenten und Zeitzeugnissen.	
2)	Beschreibende Studien	Systematische und exakte Beschreibung der Fakten und Merkmale einer vorgegebenen Population oder eines Untersuchungsfeldes.	Umfragen über Wahl- und Konsumverhalten. Erhebung über Rückfallhäufigkeit nach Strafvollzug oder Drogentherapie. Epidemiologische Untersuchungen über das Auftreten bestimmter Krankheiten.
3)	Entwicklungsstudien (Zeitreihen, Verlaufsstudien)	Untersuchungen von Mustern oder Abfolgen der Reifung oder der Veränderung im zeitlichen Verlauf.	Längsschnittuntersuchung, welche die Art und den Verlauf von Veränderungen an einer Stichprobe von Kindern wiederholt auf verschiedenen Entwicklungsstufen erfaßt. Therapieverlaufsuntersuchungen.
4)	Einzelfall- und Feldstudien	Intensive Untersuchung des Hintergrunds, des aktuellen Zustandes und der Interaktion mit der Umwelt einer vorgegebenen sozialen Einheit, eines Individuums, einer Gruppe, Institution oder Gemeinde.	Untersuchungen von Verhaltens- und Kommunikationsstrukturen in Rand- und Subgruppen. In die Tiefe gehende psychologische Analyse eines verhaltensauffälligen Kindes oder Therapiefalles. Anthropologische Untersuchungen spezieller Kulturkreise.
5)	Korrelationsstudien	Untersuchungen über das Ausmaß der Kovariation zwischen zwei oder mehreren Merkmalen.	Zusammenhang von Familienmerkmalen und Verhaltensmerkmalen von Kindern und Jugendlichen. Zusammenhang von Testdaten mit bestimmten Erfolgskriterien (Schule, Beruf, Therapieerfolg). Faktorenanalytische Untersuchung verschiedener Persönlichkeitstests.
6)	Kausal-analytische oder "Ex post facto"-Untersuchungen	Die Untersuchung möglicher Ursache - Wirkungszusammenhänge durch Beobachtung bestehender Folgen und Rückwärtssuche nach kausalen Faktoren anhand konkreter Daten.	Suche nach charakteristischen Merkmalen bei Personen mit hoher Unfallhäufigkeit im Beruf oder Verkehr ("Unfall-Typ"). Untersuchung der Bedingungsfaktoren jugendlicher Delinquenz.
7)	Echte Experimente	Erfassung von Ursache-Wirkungszusammenhängen durch systematische und willkürliche Veränderung der Ausgangsbedingungen und exakte Erfassung der Folgezustände durch den Vergleich von Experimental- und Kontrollgruppen.	Effektivitätsuntersuchungen mehrerer Lehr- und Behandlungsmethoden. Die Wirkung von Belastungsfaktoren auf die Leistung. Die Wirkung von Pharmaka und Drogen auf das Verhalten.
8)	Quasi-experimentelle Untersuchungen	Annäherung an die echten experimentellen Bedingungen, bei denen jedoch nicht alle relevanten Variablen kontrolliert oder systematisch variiert werden können mit entsprechenden Einschränkungen in der internen und externen Validität (meist anwendungsorientierte Forschung).	Effektivitätsvergleich zweier Lehrmethoden anhand zweier (oder mehrerer) Schulklassen. Feldexperimente über die Wirkung von Aktionen zur gesundheitlichen Aufklärung.
9)	Aktionsforschung	Untersuchungen zur Entwicklung neuer Ansätze oder Problembewältigungsstrategien mit direkter Anwendung auf die Praxis und mit direkter Rückkoppelung durch die Betroffenen.	Verbesserung der Betriebs- oder Organisationsstruktur einer bestehenden Einrichtung. Trainingsprogramm für Sozialarbeiter für den Umgang mit Kindern aus Problemgruppen.

1.3 Versuchspläne: Experiment und Feldstudie

1.3.1 Experimentelle Versuchspläne

Das Experiment gilt seit Beginn der Aufklärung (F. Bacon, G. Galilei u. a.) als Vorbild der Kausalforschung, d.h. für die Absicht, die entscheidenden, notwendigen und hinreichenden Ursachen der betreffenden Sachverhalte zu entdecken. Typische experimentelle Versuchspläne wurden bereits von F. Bacon und von John Stuart Mill (Logic, 1843) systematisch beschrieben. Vor allem die "Differenzmethode", d.h. das Kontrollgruppenexperiment, und die "Konkomitanzmethode", d.h. die stufenweise Änderung der Einflußgrößen bzw. die Korrelationsrechnung, sind in diesem Jahrhundert wesentlich ausgebaut worden. Für diese Versuchspläne wurde die inzwischen als "klassisch" bezeichneten Statistik entwickelt, wobei an erster Stelle R.A. Fisher (The design of experiments, 1935) zu nennen ist.

Das **naturwissenschaftliche Experiment** hat eine besondere Überzeugungskraft, welche auf der prägnanten Forschungslogik der Kausalforschung beruht: Wenn die Manipulation einer unabhängigen Variablen **UV** regelmäßig einen Effekt auf eine abhängige Variable **AV** hat, so sind wir bald überzeugt, daß hier eine gesetzmäßige Erklärung existiert, obwohl diese Verifikation im Sinne der Induktionslogik letztlich nicht zwingend ist. Höchstens die **Falsifikation**, daß ein behaupteter Effekt aufgrund der UV-Manipulation eben gerade nicht eintritt, könnte völlig überzeugen.

Die Begriffsbestimmung des psychologischen Experimentes orientiert sich zunächst am naturwissenschaftlichen Experiment, das auch in der Psychologie als Vorbild diente. Es bleibt aber zu prüfen, welche Besonderheiten in psychologischen Experimenten zu beachten sind. Unter **Experiment** ist zu verstehen:

(1) eine **systematische Beobacht**ung im Gegensatz zur Gelegenheitsbeobachtung, d.h. Datenerhebung unter einer definierten Fragestellung und nach bestimmten Ordnungsgesichtspunkten
(2) eine **planmäßige Bedingungsvariation**, d.h. die Experimentatoren greifen in ein bestimmtes System von Beziehungen ein, isolieren und verändern aktiv einen bestimmten Parameter (sog. UV) und prüfen die Auswirkung auf den Vorgang, an dem das spezielle Interesse besteht (sog. AV).

Ein **Experiment** sieht die absichtliche Auslösung eines Vorgangs vor, wobei eine planmäßige Variation von Bedingungen zum Zwecke der systematischen Beobachtung und Bedingungsanalyse vorgenommen wird (vgl. Wilhelm Wundt: (1) **absichtliche Auslösung** bzw. "Willkürbarkeit", (2) **Bedingungsvariation**, (3) **Wiederholbarkeit**). Experiment und auch Quasiexperiment sind relativ gut kontrollierte Untersuchungsstrategien, die als Vorbild für weniger gut kontrollierbare Untersuchungsansätze gelten können. Andererseits lassen sich grundsätzliche Methodenprobleme und typische Fehlerquellen am Beispiel des Experiments gut analysieren.

Wesentliche Kennzeichen des psychologischen Experiments sind also - kurzgefaßt:

- **Randomisierung**, d. h. zufällige Zuweisung von Personen zu Experimentalbedingungen bzw. Kontrollbedingungen (treatments),
- **aktive Änderung (Manipulation) von Bedingungen**,
- mehrere Aspekte der **Kontrolle**,

Diese Merkmale werden in den folgenden Kapiteln zur Versuchsplanung ausführlich behandelt. Einführend sind bereits zu nennen

(1) das **MAX-MIN-KON-Prinzip**: **Max**imiere die experimentell interessierende Varianz, **min**imiere die Fehlervarianz und **kon**trolliere die unwesentliche Varianz und
(2) das Begriffspaar **interne Validität – externe Validität**, d. h. die Präzision der Bedingungskontrolle und die Generalisierbarkeit (siehe folgende Kapitel).

Es gibt verschiedene **Einteilungsgesichtspunkte** für Experimente:
(1) Erkundungsexperiment (pilot study, exploratives Experiment) vs. Entscheidungsexperiment ('experimentum crucis');
(2) echtes Experiment vs. Quasi-Experiment, d.h. tatsächliche Manipulation der UV und aktive Zufallszuweisung der Personen zu den Bedingungen oder nachträgliche (ex post facto) Analysen, in denen die nicht möglich gewesene Randomisierung durch statistische Kontrollen und Auswahlprozeduren näherungsweise ersetzt werden soll (Untersuchung von Attributvariablen, d. h. der "natürlichen Variiertheit" statt eine experimentelle Variation durchzuführen; praktische und/oder ethische Unmöglichkeit einer Zufallszuweisung zur Experimentalbedingung);
(3) Einteilung nach Versuchsplänen und Datenanalyse-Strategien (s. Abschnitt über Datenanalyse).

Die experimentalpsychologische Forschung stellt Anforderungen, welche in vielen Bereichen bzw. für viele Fragestellungen der Psychologie, die interessant und praktisch wichtig sind, höchstens näherungsweise oder überhaupt nicht zu erfüllen sind: Randomisierung, Standardisierung der Untersuchung, Kontrollen, außerdem im Prinzip auch Zufallsstichproben aus der Population. In der Entwicklungs- und Sozialpsychologie, in der Arbeitspsychologie und Pädagogischen Psychologie sind oft andere Strategien erforderlich, auch die Aufklärung der Entstehung psychischer Krankheiten ist durch Laborexperimente nicht möglich. Als Beispiel anderer Forschungsstrategien werden hier die Feldstudien, welche unter dem Gesichtspunkt der ökologischen Validität wichtig sind, hervorgehoben.

1.3.2 Feldstudien

Feldstudie und Laborexperiment als alternative Formen psychologischer Forschung werden seit langem diskutiert, etwa unter den Gesichtspunkten der **Lebensnähe** der Feld- und Aktionsforschung (Lewin, 1927), der **ökologischen Validität** und **repräsentativen Versuchsplanung** (Brunswik, 1956), der **internen und externen Validität** (Campbell, 1957; Campbell & Stanley, 1963), der **sozioökologischen Einheiten** (Barker, 1968), der **naturalistischen Beobachtung** (Willems & Raush, 1969), **der Partialisierung des Lebenszusammenhangs** (Holzkamp, 1973).

Es ist in der Methodenlehre der Psychologie üblich, bestimmte Vorteile und Nachteile eines typischen Laborexperiments im Gegensatz zu einer typischen Feldstudie hervorzuheben. So werden dem Laborexperiment die Vorzüge der aktiven Kontrollierbarkeit zugeschrieben: die Möglichkeit zur präzisen Prüfung von Gesetzeshypothesen aufgrund zweckmäßiger Isolierung von Bedingungen und rigoroser Kontrolle der allgemeinen Versuchsbedingungen, insbesondere durch Randomisierung der Ausgangsbedingungen. Als relative Nachteile oder sogar grundsätzliche Mängel werden die "Künstlichkeit" der Untersuchungsbedingungen, die möglichen Artefakte aufgrund der Einstellungen und Erwartungen von Versuchspersonen und Versuchsleitern, der Zweifel an der Generalisierbarkeit und die deswegen geringe praktische Relevanz der Ergebnisse genannt. Demgegenüber sei mit der Natürlichkeit der Untersuchungsbedingungen im Feld zugleich eine größere praktische Gültigkeit von Feldstudien gegeben. Als typische Nachteile gelten (1) die häufig vorkommenden, multiplen (konfundierten, zusammenhängenden) Effekte, (2) die Schwierigkeiten oder die Unmöglichkeit der für eindeutige Schlußfolgerungen notwendigen Randomisierung der Ausgangsbedingungen

und die wahrscheinlich aus vielfältigen Gründen oft herabgesetzte Zuverlässigkeit der Datenerhebung unter Feldbedingungen.

Diese Argumentation bewegt sich deutlich im Rahmen des von Campbell und Stanley (1963) ausgearbeiteten Schemas interner gegenüber externer Validität, d.h. einerseits Aussagenpräzision aufgrund von methodischen Kontrollen und andererseits Aussagengeneralisierbarkeit aufgrund ökologisch-repräsentativer, gültiger Auswahl der Bedingungsvariation. Es gibt jedoch durchaus Feldexperimente mit einer Zufallszuweisung zu Behandlungen, und es gibt auch Zwischenstufen von Laborexperiment und Feldstudie.

Patry (1982) nennt als Absicht von Feldforschung "*Aussagen darüber zu machen, wie sich der Mensch in seiner sozialen und materiellen Umwelt verhält, auch wenn er nicht Gegenstand einer Untersuchung ist, was er tut, wenn kein Versuchsleiter ihn direkt oder indirekt beeinflußt, und was ihn veranlaßt, es zu tun*" (S. 17). Im Unterschied zur künstlichen Bedingungsvariation im Labor geschieht die Variation im Feld natürlich, von sich aus. Der Begriff "natürlich" ist jedoch mehrdeutig, weil auch mögliche Bewertungen einer Bedingung als "naturgemäß" oder "nicht vom Menschen gemacht" hineinspielen. Einige Autoren bezeichnen diese Alltagsbedingungen als "naturalistisch", andere Autoren nennen auch die den natürlichen Settings nachgebildeten (analogen, simulierten, in-vivo-Testbedingungen) Settings naturalistisch. Zur Charakterisierung von Feldforschung hat Tunnel (1977) drei weitgehend unabhängige Dimensionen der Natürlichkeit definiert: **natürliches Verhalten** aus dem Repertoire des Individuums, **natürlicher Kontext**, in welchem sich das Individuum befindet, und **natürliche Bedingungsänderung**, welche auch ohne die Anwesenheit des Untersuchers als diskretes Ereignis eintreten kann. Patry (1982) gelangt zu einem Schema mit 16 Varianten, indem er diese Einteilung noch durch die beiden Gesichtspunkte erweitert, ob den Probanden bekannt ist, daß sie untersucht werden, und ob sie ggf. die Untersuchungshypothese kennen.

Diese Labor-Feld-Diskussion hat sich zu sehr an den Gesichtspunkten der internen und externen Validität orientiert, statt der **Strukturähnlichkeit von Kontexten** genauer nachzugehen. Für die Bewertung der Forschungsergebnisse bleibt diese Strukturähnlichkeit eine wesentliche Frage. "*Die **Lebensnähe** des Experiments ist nicht in der quantitativen Übereinstimmung mit der Wirklichkeit zu suchen, sondern entscheidend ist, ob beide Male wirklich der gleiche **Geschehenstypus** vorliegt. Handelt es sich nämlich um Geschehnisse gleicher Struktur, so ist innerhalb breiter Bereiche ein Schluß ... zulässig*". (Lewin, 1927, S. 419).

Aus wissenschaftstheoretischer Sicht besteht keine prinzipielle Trennung von Problemen der **internen** und Problemen der **externen** Validität, so daß auch kein Zweistufenplan von Laborexperiment und empirischer Generalisierbarkeitstudie aufgestellt werden muß. Die Frage der **Kontextspezifität** von Erklärungshypothesen gehört bereits in die anfängliche Theorieexplikation und die zugehörige Diskussion der Operationalisierungsentscheidungen (Gadenne, 1976; Westmeyer, 1982). In diesem Zusammenhang unterscheidet Gadenne (1976) drei Aspekte der **Repräsentativität**: (1) Repräsentativität von Geschehenstypen der Untersuchung im Hinblick auf die gemeinten theoretischen Sachverhalte, (2) Repräsentativität von Geschehenstypen der Untersuchung im Hinblick auf die Anwendungssituationen bzw. die Übertragbarkeit, und (3) die stichprobentechnische Repräsentativität der untersuchten Personengruppe. Angesichts der Vielfalt möglicher Strategien und Pläne ist es sicher mißverständlich, von Feldstudien zu reden, ohne die wesentlichen Kennzeichen zu nennen. Es wäre auch viel zu einfach, die Vorteile der Laborforschung als Nachteile der Feldforschung herauszustellen und umgekehrt. Wahrscheinlich gibt es kein Methodenproblem, keine typische Artefaktquelle oder Bedrohung der inneren und äußeren Gültigkeit, welche nicht innerhalb und außerhalb des Labors eine Entsprechung hätte.

Zum psychologischen Begriff der Situation
Der psychologische Begriff der **Situation** (bzw. 'Kontext' oder 'Setting') ist vieldeutig, weil sowohl die objektiven Merkmale als auch die erlebnismäßige Interpretation (und ggf. der in der Untersuchung intendierte Aufforderungscharakter bzw. die vom Untersucher beabsichtigte Wirkung) gemeint sein können.

Kontext: die Gesamtheit der relevanten Bedingungen (Umwelt, Milieu, Rahmenbedingungen des Geschehens) mit bestimmten kontextuellen Variablen einschließlich ambienter Umweltparameter, z. B. Lärmpegel, Helligkeit, Temperatur.
Setting: ein primär räumlich und durch Gegenstände und Anordnungen objektiv beschreibbarer Kontext (Aufenthaltsort), z. B. das Setting einer Wohnung, mit bestimmten Tätigkeiten.
Behavior Setting (Barker, 1968): ein Setting mit typischem Verhaltensprogramm („Aufforderungscharakter"), z. B. ein Hörsaal oder ein Restaurant.
Situation: ist wesentlich durch die subjektive und erlebnismäßige Beschreibung eines Settings bestimmt und schließt Handlungen ein. (Situation = individuell bewertetes Setting mit entsprechend großer inter- und intraindividueller Variation).

Ambulantes Monitoring/Ambulantes Assessment
Ambulantes Monitoring: die Überwachung von frei beweglichen Personen unter Alltagsbedingungen, z. B. 24-Stunden-Registrierung von Blutdruck oder EKG, im Gegensatz zur stationären Überwachung von Patienten, z. B. auf der Intensivstation nach Operationen.
Ambulantes Assessment (Feldpsychodiagnostik, Pawlik, 1988): die systematische Erfassung psychologischer, physiologischer und kontextueller Daten unter alltäglichen Bedingungen (am Arbeitsplatz, in der Schule, in der Freizeit) für einen bestimmten Zweck (siehe Fahrenberg & Myrtek, 1996).

1.3.3 Vergleich zwischen Laborexperiment und Feldstudie

In der Tabelle 1.2 sind typische Merkmale von Laborexperiment und Feldstudie aufgeführt. Diese lassen sich nicht allein unter dem Aspekt interne/externe Validität bewerten.

Tabelle 1.2: Vorzüge und Nachteile von Laborexperiment und Feldstudie

Laborexperiment	Feldstudie
Präzision der Hypothesenprüfung Zuverlässigkeit der Schlußfolgerungen aufgrund von Isolation und aktiver Manipulation der UV, Kontrolle der Fehlervarianz	**Ökologische Validität** (Lebensnähe, Praxisbezug) aufgrund naturalistischer Beobachtungen und repräsentativer Versuchsplanung; Generalisierbarkeit
→ relativ höhere interne Validität	**→ relativ höhere externe Validität**
Künstlichkeit der Untersuchungsbedingungen, u. U. geringe Relevanz für Teilnehmer	**Multiple Effekte,** deren Komplexität schwer kontrollierbar ist
Effektstärken aus ethischen und praktischen Gründen beschränkt	Höhere, realistische **Effektstärken** möglich

Statt die Vorzüge und Nachteile gegeneinander auszuspielen, kommt es auf die Entwicklung von neuen Strategien, u. a. von Labor-Feld-Kombinationen, von Feldexperimenten (d. h. mit Randomisierung) u. a. **innovativen Strategien** an.

Auf dem Gebiet der Psychologie kann das Laborexperiment als **methodisches Ideal** aus mehreren Gründen keine allgemeine Gültigkeit beanspruchen. Als besonders prägnanter Methodentyp ist jedoch das Experiment in der Forschung und im Unterricht besonders geeignet, die **Unvollkommenheiten, Schwächen und Fehlerquellen** auch der anderen Methodentypen erkennen und diskutieren zu können.

1.4 Statistische Versuchsplanung

Das Ziel der statistischen Versuchsplanung ist es, sicherzustellen, daß die für die Fragestellung relevante Information vollständig und in auswertbarer Form gewonnen und dokumentiert wird; Aufwendung und Nutzen sollen in einem vertretbaren Verhältnis stehen; die Risiken von möglichen Fehlentscheidungen sollen kontrolliert werden können; eine vorgegebene Mindestgenauigkeit soll erreicht werden.

Fünf Schritte der allgemeinen Versuchsplanung sind:

(1) **die Präzisierung der Fragestellung**

Hierfür sind drei Fragen zu beantworten: (1) Es ist zu klären, welche Merkmale im Versuch beobachtet werden sollen; welche Merkmale sind Zielmerkmale, welche sind wichtige Einflußgrößen bzw. Störgrößen; die Erfassung bzw. Skalierung dieser Merkmale ist festzulegen. Das Ziel des Versuchs und die Festlegung der Beobachtungseinheiten und die zu prüfenden Hypothesen sollten vollständig beschrieben werden. (2) Eindeutige Ein- und Ausschlußkriterien für Beobachtungseinheiten sollen formuliert werden (z. B. Zufallsauswahl von Beobachtungseinheiten). (3) Festlegung von Kriterien, anhand derer die Güte und Sicherheit von Aussagen beurteilt werden kann (vgl. alpha-, beta-Fehler, Konfidenzintervalle, Toleranzintervalle etc.).

(2) **die Wahl des statistischen Modells**

Hier sind das Skalenniveau der Merkmale (diskrete, ordinale, metrische Skalierung), Information zum Verteilungstyp, Varianzinhomogenitäten, Gruppierungen, Abhängigkeiten u. a. Aspekte zu berücksichtigen.

(3) **die Konstruktion oder Auswahl der geeigneten Versuchsanordnung**

Versuchspläne (Designs) sind Schemata, die Struktur, Umfang und Ablauf von Versuchen beschreiben: Faktorenstufen oder Werte der Einflußvariablen (können vom Versuchsleiter festgelegt werden), die Ausschaltung von Störfaktoren oder die Untersuchung von Faktoren mit fest vorgegebenen Stufen. Die verschiedenen Versuchspläne werden später ausführlich dargestellt.

(4) **die Auswahl der Auswertungsmethode**

Die Auswertungsmethode hängt mit der Fragestellung, dem statistischen Modell und dem Versuchplan zusammen.

(5) **die Festlegung des Versuchsumfangs**

Der Versuchsumfang ist entscheidend für Güte und Sicherheit der Aussagen. Die statistische Planung des Umfangs erfordert Vorinformationen, u.a. über die Variabilität des Merkmals. Gleichzeitig müssen aber auch praktische Möglichkeiten und Aufwand berücksichtigt werden.

Ausgewählte Literatur

Bortz, J. (1995). *Forschungsmethoden und Evaluation (2. Aufl.)*. Berlin: Springer. (1. Aufl.; *Lehrbuch der empirischen Forschung für Sozialwissenschaftler*, 1984).

Fahrenberg, J. & Myrtek, M. (Eds.) (1996). Ambulatory assessment. Computer-assisted psychological and psychophysiological methods in monitoring and field studies. Seattle, WA: Hogrefe Publ.

Huber, O. (1987). *Das psychologische Experiment: Eine Einführung.* Bern: Huber.

Isaac, S. & Michael, W.B. (1974). *Handbook in research and evaluation.* San Diego, Ca.: Knapp.

Kerlinger, F.N. (1978/79). *Grundlagen der Sozialwissenschaften.* Weinhein: Beltz (besser die amerikanische Ausgabe, 4. Aufl. 2000).

Pawlik, K. (1988). "Naturalistische" Daten für Psychodiagnostik: Zur Methodik psychodiagnostischer Felderhebungen. *Zeitschrift für Differentielle und Diagnostische Psychologie, 9*, 169-181.

Pedhazur, E.J. & Schmelkin, L.P. (1991). *Measurement, design, and analysis. An integrated approach*. Hillsdale, N.J.: Lawrence Erlbaum.

Sarris, V. (1990). *Methodologische Grundlagen der Experimentalpsychologie*. (2 Bände). München: Reinhardt.

Schnell, R., Hill, P.B. & Esser, E. (1993). *Methoden der empirischen Sozialforschung* (4. Aufl.). München: Oldenbourg.

Hinweis
In diesem Skript werden nicht alle speziellen Literaturhinweise vollständig zitiert. Wer an einer Vertiefung interessiert ist, kann sich an die Dozenten wenden, um ausführliche Hinweise zu erhalten.

2. Fragestellungen und Operationalisierungen. Taxonomie von Methoden der Datenerhebung. Messung und Skalierung.

2.1 Fragestellungen

Wissenschaftliche Fragestellungen ergeben sich aus einem grundlagenwissenschaftlichen Engagement oder aus dem Wunsch, praktische Probleme besser lösen zu können, jedenfalls aus einem Erkenntnisinteresse. **Fragestellungen** haben die noch relativ allgemein gehaltene, in den Einzelheiten oft noch vage Formulierung eines **Problems**:

- besteht ein Zusammenhang von ...
- besteht ein Unterschied zwischen ...
- verändert sich ein Sachverhalt in Abhängigkeit von einer Bedingungsänderung?
- ist eine Vorhersage dieser Verhaltensweise (z. B. der schulischen oder beruflichen Leistung, Veränderung der Symptomatik ...) möglich?

Außer den grundsätzlich beantwortbaren Fragestellungen gibt es solche, die (1) empirisch nicht beantwortbar sind oder (2) empirisch nicht mehr oder noch nicht entscheidbar sind.

Durch Bezug auf theoretisches Wissen, insbesondere auf den aktuellen Stand der international zugänglichen Fachliteratur, kann aus der Fragestellung (dem Problem) eine prägnante Hypothese in inhaltlicher und in statistischer Schreibweise entwickelt werden. **Die Hypothese ist eine ausformulierte Antwort auf die Fragestellung**; die Hypothese ist so genau ausgeführt, daß sie empirisch entscheidbar ist, d.h. an der Erfahrung scheitern kann. Diese Logik der Forschung und der Hypothesenprüfung führen zu schwierigen erkenntnis- und wissenschaftstheoretischen Überlegungen, deren Annahmengefüge nicht bei jeder wissenschaftlichen Untersuchung mitgeteilt werden kann (oder müßte, weil diese Probleme den methodenkritisch Informierten geläufig sind). In der "scientific community" existieren viele Konventionen und "common sense" Argumente, die jedoch – wie überdauernde Diskussionen lehren – nicht allgemein verbindlich sind (Logik und Syllogistik ausgenommen).

Die Entwicklung prüfbarer Hypothesen aus Fragestellungen verlangt **Definitionen** (Abgrenzungen) und **Explikationen** (Überführung eines mehr oder weniger unexakten Begriffs in ein exaktes Konzept), vor allem **operationale Definitionen** des Gemeinten, d.h. genaue Angaben, mit welchen Operationen gearbeitet werden soll. Diese Operationalisierungen sind nicht etwa einfache Auswahlentscheidungen über "Methoden", sondern es sind eminent theoretische Entscheidungen und Begründungen, weshalb gerade diese Methode der "richtige Weg" ist. Die Operationalisierungsentscheidungen sollten gerechtfertigt werden können (Adäquatheitsbedingungen von Phänomen und Methode, Konstruktexplikation – siehe Abschnitt über Operationalisierungsfehler und Skript der Vorlesung "Wissenschaftstheorie").

Wegen des bekannten Pluralismus psychologischer Theorienbildungen besteht gewöhnlich ein großer Spielraum. Nur bei sog. Entscheidungsexperimenten oder bei Replikationen eigener und fremder Untersuchungen sind die Operationalisierungen festgelegt. Möglichst gleichartige Wiederholungen (cross-laboratory replication) sind in der Psychologie leider sehr selten; sie scheinen als weniger prestigeträchtig oder nützlich zu gelten (siehe Schweizer, 1989). Zu vielen wichtigen Fragestellungen bestehen deshalb überdauernde Diskrepanzen der publizierten Befunde („inconsistent findings“) mangels Standardisierung der Methodik!

Auch in der wissenschaftstheoretischen Grundlagendiskussion gibt es – abgesehen von der geisteswissenschaftlichen Sicht (Hermeneutik und Phänomenologie siehe u.a. Danner, 1994) – zahlreiche konkurrierende Auffassungen, welche wahrscheinlich Auswirkungen auf die

Methodenlehre, speziell auf die Formulierung von Forschungsprogrammen und deren Bewertung haben werden (s. Breuer, 1988; Chalmers, 1986; Gadenne, 1994a, 1994b).

Die meisten sog. "Theorien" der psychologischen Literatur sind – wie Analysen u.a. von Madsen und Rekonstruktionsversuche in den letzten Jahren, u.a. von Westermann, Westmeyer zeigen – weit von der Strukturierung der Theorien im naturwissenschaftlichen Bereich entfernt. Bei kritischer Würdigung entsprechen viele psychologische Theorienbildungen eher einem im geisteswissenschaftlichen Bereich verbreiteten Stil: es handelt sich oft um spekulative Ideen, gedankliche Entwürfe, Erklärungsskizzen, Wie-ist-es-möglich, daß-Erklärungen (zur Typologie der Erklärungsarten und zur Formalisierung und Bewertung von Theorien siehe Stegmüller, 1970; sowie Gadenne, 1994a und 1994b). Zu dieser Einschätzung passen auch die Beobachtungen: In der Fachwelt wird relativ geringes Gewicht auf die Sicherung von Befunden durch **Replikation seitens anderer Forscher** gelegt und die intensive Suche nach den Gründen von Diskrepanzen ist selten anzutreffen.

In den folgenden Kapiteln wird die **empirisch-analytische Forschungsstrategie** dargestellt. Die wichtigsten Schritte sind:

- theoretische Auffassungen des interessierenden psychischen Prozesses (Phänomen, Sachverhalt);
- Repräsentation (Abbildung) dieses Prozesses in Strukturmodellen bzw. Modellierungen und vor allem in **empirisch faßbaren Merkmalen** (latente vs. **manifeste Variablen** (synonym: Observable, Beobachtungsprädikat, Referent, Indikator).
- allgemeine Datentheorie, d.h. Meßtheorie, Skalentheorie, Testtheorie, Fehlertheorie;
- Operationalisierungen für unabhängige Variablen (UV), abhängige Variablen (AV), Kontrollvariablen bzw. Kovariablen (KV);
- **Instrumentierung** im weiteren Sinn mit möglichst präzisen Angaben (Setting, "Paradigma", Versuchsaufbau, Geräte, verwendetes Stimulusmaterial, Instruktionen, Aufgabe, Test usw.);
- Hypothesenformulierung;
- Datenerfassung;
- Stichproben- und Auswahl-Verfahren;
- Versuchsplanung;
- Datenanalyse;
- Hypothesenprüfung (Entscheidung);
- kritische Diskussion (Evaluation).

Einige dieser Schritte werden in der Übung Versuchsplanung erläutert, andere in den Praktika, in den Übungen zur Statistik und in den theoriebezogenen Lehrveranstaltungen. Ein Leitfaden, der für die Praktika II, III und IV gedacht ist, wurde in dieses Skriptum aufgenommen (siehe Anhang 1).

2.2 Probleme der Operationalisierung

Die Auswahl einer manifesten Variablen, welche ein latentes Konstrukt "repräsentieren" soll, wird als Operationalisierung bezeichnet (siehe Carnaps Unterscheidung von Theorie-Sprache und Beobachtungssprache). Damit wird von der "operationalen Definition" Gebrauch gemacht, womit ein Konstrukt durch Angabe der "repräsentierenden" Variable(n) und der zugehörigen Meßvorschriften definiert wird. In diesem Sinne erfolgen Definitionen in der Psychologie oft als Konstruktdefinitionen durch die Selektion von Variablen sowie durch die Verwendung von bestimmten Untersuchungsparadigmen. Da solche Selektionen oft per fiat (so sei es!) erfolgen, drohen Operationalisierungsfehler oder zumindest Mißverständnisse. Die Angemessenheit/Zulässigkeit bestimmter Operationalisierungen wird in Untersuchungen zur Konstruktvalidität geprüft.

Typische Operationalisierungsfehler sind
(1) Fehler der **unzureichenden Referenz** (ein Konstrukt hat mehr Bedeutungskomponenten als durch die verwendete(n) Variable(n) erfaßt sind);
(2) **Mißspezifikation** (die verwendete Variable gehört nicht zu dem Zielkonstrukt);
(3) **Überschußbedeutung** (die verwendete Variable trägt auch Bedeutungskomponenten anderer Konstrukte, die dann fälschlicherweise dem Zielkonstrukt zugeschrieben werden).

Nach R. B. Cattell ist besonders der bivariat arbeitende Experimentator (Pawlow-Wundt-Tradition) der Gefahr von Operationalisierungsfehlern ausgesetzt, weniger dagegen der multivariat und differentiell orientierte Forscher (Galton-Spearman-Thurstone-Tradition), weil dieser **multiple Operationalisierungen** verwendet. Kompetente Psychologen/innen werden in vielen Fällen eine Methodenkombination auswählen, vor allem wenn es um verhältnismäßig breite theoretische Begriffe (Angst, Intelligenz u.a.) geht oder wenn es auf riskante, folgenreiche Entscheidungen ankommt.

Eine besonders anspruchsvolle Strategie der Operationalisierung stellt das **Drei Ebenen-(Drei-Systeme-) Konzept** des Assessment dar: introspektiv-verbale, behaviorale und physiologische Daten sind zu kombinieren (triple response measurement, TRM-Strategie). Empirisch zeigt sich allerdings oft Divergenz ("response fractionation") statt Konvergenz solcher Indikatoren, so daß nicht nur das Konstrukt ("Angst" war das Beispiel im Praktikum I) fragwürdig wird, sondern auch schwerwiegende Fehleinschätzungen bei univariater Vorgehensweise unterstellt werden müssen. Die sehr kritischen Ergebnisse solcher Operationalisierungsstudien sind eine Herausforderung ("multimodale Diagnostik als Standard der klinischen Psychologie" – siehe Seidenstücker und Baumann).

Für die Theoretiker und für die Verhaltenstherapeuten bedeutet es gleichermaßen eine schwierige Herausforderung, wenn z. B. das Angst**gefühl** (subjekt-verbale Ebene), das ängstliche Vermeidungs**verhalten** (behaviorale Ebene) und die vegetativ-endokrine Angst**erregung** (physiologische Ebene) weder zu Beginn, noch im Prozeß oder am Ende einer Therapie konvergent sind. Kann aber die Therapie als beendet angesehen werden, wenn nur auf subjektiver Ebene ein Erfolg erzielt wurde, jedoch nicht hinsichtlich behavioraler und physiologischer Aspekte der Angst? (siehe Barlow u. a.). Da auf jeder der sog. Ebenen – genau genommen – viele relativ unabhängig funktionierende Subsysteme angenommen werden müssen, ist die Situation noch komplizierter. Der globale Begriff „Angst“ könnte sehr irreführend sein.

Der Multitrait-Multimethod-Ansatz MTMM von Campbell und Fiske (1959) ist ein Verfahren, die empirische Konvergenz bzw. Divergenz von multiplen Operationalisierungen eines Konstrukts zu prüfen. Wenn z. B. in Hamiltons Untersuchung drei theoretisch unterschiedene Eigenschaftskonstrukte (Selbstachtung, Dominanz, Aufgeschlossenheit) mit verschiedenen unabhängigen Methodentypen (Persönlichkeitsfragebogen, Selbsteinstufung, Fremdeinstufung durch Bekannte) erfaßt werden, dann müßten bei erfolgreicher multipler Operationalisierung die Indikatoren **eines** Konstrukts über verschiedene Methodentypen substantiell korrelieren (**konvergente Validität**), jedoch nicht mit den Indikatoren der anderen Konstrukte (**diskriminante Validität**). Das Multitrait-Multimethod Schema (Abbildung 2.1) verdeutlicht die konvergente und die diskriminante Validität sowie die Reliabilität. Es gibt in der Literatur nur sehr wenige überzeugende MTMM Analysen mit mehreren Methodentypen (z.B. Drei-Ebenen-Analysen). In solchen Operationalisierungsstudien hat sich fast regelmäßig gezeigt, daß als einheitlich angenommene Konstrukte eher als Anordnung von relativ unabhängigen Sub-Konstrukten aufzufassen sind. Die befriedigende Konvergenz multipler Indikatoren ist eher die Ausnahme, Divergenzen (bzw. unerwartet niedrige Korrelationen) sind häufig. Deshalb hat u.a. Fiske sehr viel **genauere operationale Definitionen von**

Subkonstrukten und speziellen construct-operation-units verlangt (siehe Multimodale Diagnostik, Themenheft Diagnostica, 1987)

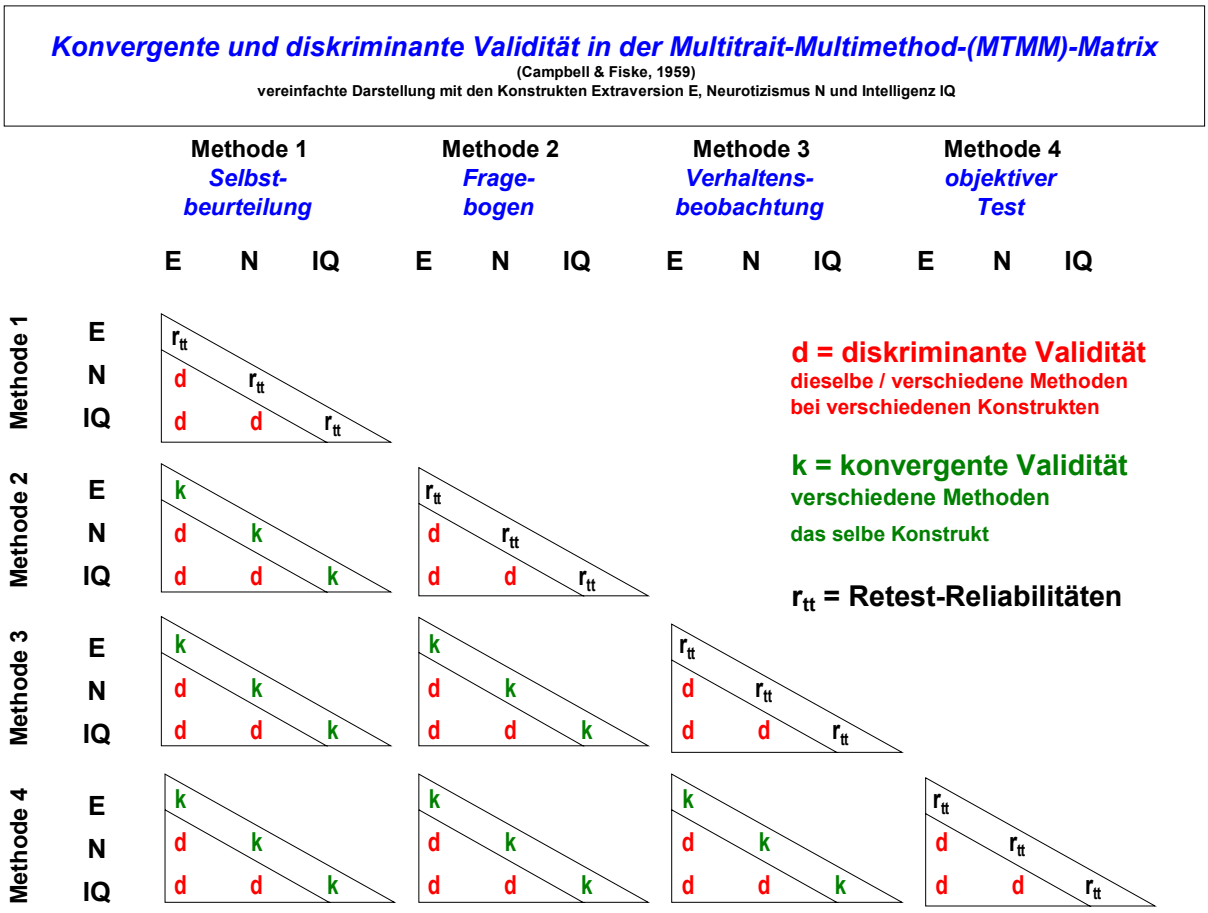


Abbildung 2.1: Multitrait-Multimethod-Matrix

Das **Linsenmodell von Brunswik** (1956) bezieht sich auf die repräsentative Auswahl von Variablen. Wenn es z. B. um statistische Vorhersagen des Verhaltens aus bestimmten Testbefunden geht, dann soll zwischen dem Satz der Prädiktorvariablen und dem Satz der Kriterienvariablen eine **symmetrische Beziehung** (Linsendarstellung) bestehen, d.h. die Breite und Güte der Prädiktoren/Kriterien sollten sich entsprechen. Das Konzept von **vier Datenboxen** (Prädiktoren, experimentelles Treatment, nicht-experimentelles Treatment und Kriterien) wurde von Wittmann (1985) in Anlehnung an Brunswik und Cattell entwickelt, um die notwendigen Präzisierungen von Assessmentstrategien und Validitäts- und Reliabilitätsaspekten zu erreichen.

Das Verfahren der **Kovarianzzerlegung** ermöglicht die Aufgliederung der Beziehungen in der viermodalen Datenbox (Personen x Variablen x Situationen x Replikationen der Situationen). Die korrelativen Beziehungen können unter verschiedenen Aspekten statistisch analysiert werden.

Aggregation ist die – meist additive, d. h. ungewichtete – Zusammenfassung von Elementen
über **Zeitpunkte** (Meßwiederholungen),
über **Situationen** (Settings, Untersuchungsbedingungen),
über **Items** (Konstrukt-Facetten, Verhaltensweisen),

und andere Dimensionen der Datenbox bzw. als **mehrdimensionales Aggregat**, in der Absicht, einen valideren (repräsentativen, symmetrischen Index) auf Prädiktor – oder auf Kriterienseite zu erhalten. Das Verfahren kann pragmatisch **kriteriumsorientiert** („Indexmessung") oder theoretisch **konstruktorientiert** (Konstruktoperationalisierung) sein. Aggregationen können eventuell auch über mehrere Untersuchungen vorgenommen werden (siehe Metaanalyse).

Bereits die Antwort auf ein typisches Fragebogen-Item z. b. „ich bin häufig angespannt" liefert ein kompliziertes Aggregat (Zeitpunkte, Situationen, Symptome, Facetten). Der Intelligenzquotient einer Person wird durch Aggregation über Items und Aufgabengruppen, d. h. Klassen von Inhalten und Operationen mit Meßwiederholungen, gewonnen.

Das Prinzip der **Reliabilitäts-Steigerung durch Verlängerung des Tests**, d. h. Hinzufügen relativ homogener Items, geht bereits auf Spearman & Brown zurück. Die neuere Diskussion über Aggregation wurde u. a. durch Fishbein und Aijzen (multiple act-Ansatz) und durch die Mischel-Epstein-Kontroverse über die angebliche Validitätsgrenze bei r = 0.30 angeregt.

Ein **asymmetrisches Aggregationsniveau** läge dann vor, wenn z. B. der Testwert eines Persönlichkeitsfragebogens für „Extraversion" als Prädiktor herangezogen wird, um die an einem bestimmten Tag beobachtbare Geselligkeit und Unternehmungslust vorherzusagen. Der Testwert E als Index einer überdauernden Persönlichkeitseigenschaft wird durch Auskünfte des Individuums (d. h. zeitliche und inhaltliche Aggregation bestimmter Aspekte vieler Erfahrungen und Gewohnheiten) und durch rechnerische Aggregation über viele Items gewonnen. Die Verhaltensbeobachtung des Kriteriums bezieht sich dagegen nur auf einen kurzen Zeitraum. Korrelationsforschung dieser Art kann nur geringe Beziehungen aufzeigen, doch muß dies nicht an der mangelnden Validität der Fragebogenskala liegen, wie Mischel glaubte, sondern kann auch an der mangelnden Symmetrie liegen, wie Epstein zeigte, als er durch Aggregation über viele Tage, Meßzeitpunkte und Variablen ein vergleichbares Aggregationsniveau herstellte (die Mischel-Epstein-Kontroverse ist ein wichtiges Thema der Differentiellen Psychologie). In der Forschung und Praxis mangelt es auch heute noch oft an angemessenen Strategien der Aggregation von Daten bzw. selbstkritischen Überlegungen hinsichtlich der Repräsentativität der Designs im Sinne von Brunswik (siehe auch Nübling, Schweizer, Wittmann). Diese Fragen müssen – im Vergleich zu vielen Detailfragen der statistischen Analysenkonzepte – als ungleich wichtiger angesehen werden.

Die Überlegungen zur Operationalisierung sollen zu einer genaueren **Konstruktexplikation** führen bzw. zwingen und schließlich die theoretisch und empirisch begründete Identifizierung der zu untersuchenden Variablen (UV, AV, KV) im Kontext eines Annahmengefüges erreichen.

Operationalisierung: Theoretisch und empirisch begründete Festlegung
Welcher Index (welche Variable) „repräsentiert" das gemeinte (latente) theoretische Konstrukt? Welche Meßvorschriften und Untersuchungsanordnungen gehören dazu? Welche Klärungen müssen erfolgen, um hier Mißverständnisse zu vermeiden? Ist eine multiple Operationalisierung möglich? Was ergibt die Kontrolle der konvergenten und diskriminanten Validität im MTMM-Ansatz? Welche Ergänzungen sind erforderlich, um die Symmetrie von Prädiktor und Kriterium zu verbessern?

2.3 Taxonomie von Methoden der Datenerhebung

Die Taxonomie (Ordnungssystem) der Datenquellen inter- und intra-individueller Differenzen ist in neuerer Zeit vor allem von R.B. Cattell entwickelt worden. **Informationsquellen** sind:

L **(Lebens-)Daten** aus Einstufungen des Verhaltens L(R) oder Beobachtungen L(O); auch als Behavior Rating (BR) und Behavior Observation (BO) bezeichnet;

Q **(Questionnaire-)Daten** aus Selbstbeurteilungen Q' oder aus standardisierten Fragebogen Q stammend;

T **(Test-)Daten** aus standardisierten Labor- oder Testbedingungen, welche **objektive, d.h. von den Erwartungen und Einstellungen des/der Untersuchten unabhängige Daten** liefern.

In Cattells **Kovariationsschema** sind die drei Modalitäten (1) Personen, (2) Variablen und (3) Situationen (oder Beobachtungszeitpunkte) kombiniert, um – je nach Blickrichtung in diesem Datenwürfel – verschiedene Datenbeziehungen und Datenerhebungspläne deutlich zu machen (siehe Abb. 2.2)

Datenbox und Korrelationstechniken (nach Cattell, 1957)

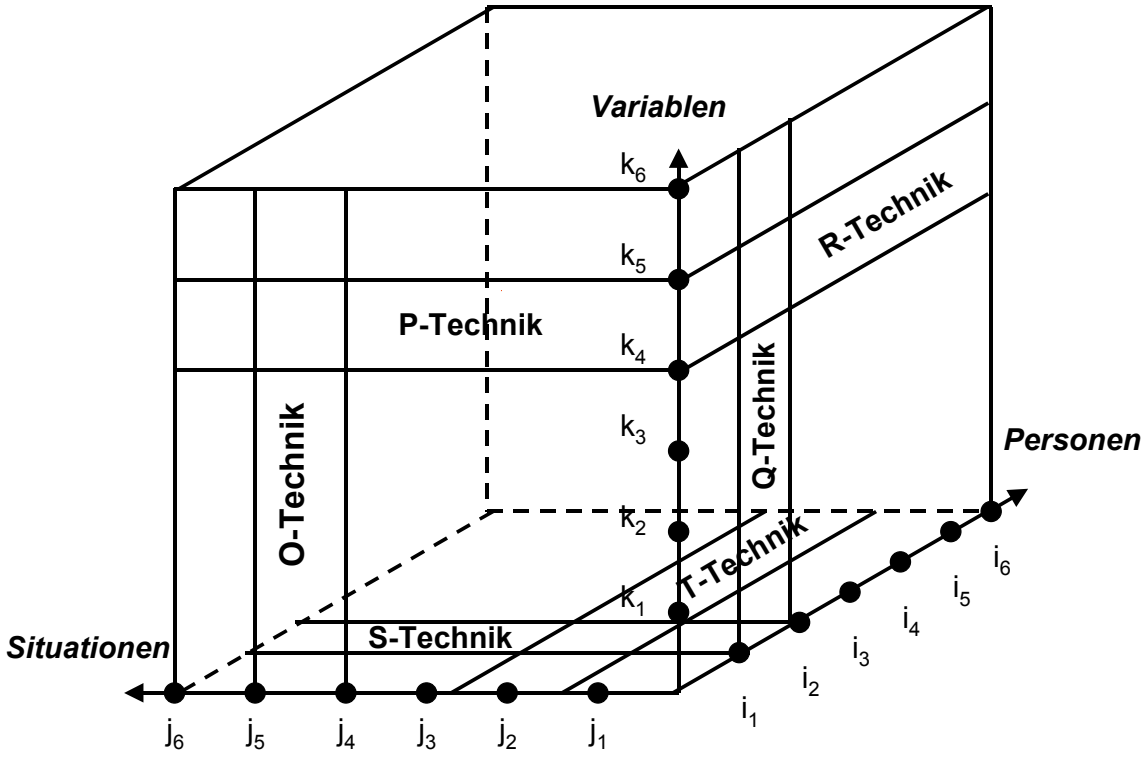


Abbildung 2.2: Datenbox (Kovariationsschema) und Korrelations-(Faktorenanalyse-)Techniken

Aus Cattells Datenbox lassen sich sechs Techniken ableiten (siehe Abb. 2.1). Cattel hat später die Datenbox zur Basic Data Relation Matrix BDRM erweitert, indem fünf Modalitäten und bei jeder Modalität noch ein zeitlicher Aspekt, z.B. die Person (der Organismus) und der aktuelle Zustand der Person (des Organismus) berücksichtigt werden (siehe Abb. 2.2). Dementsprechend ergeben sich zahlreiche mögliche Ordnungsansätze und Versuchspläne der **multivariaten Forschung.**

Tabelle 2.1: Techniken der Korrelation und Faktorenanalyse nach Cattell

Technik	konstant gehalten	interkorreliert und faktorisiert	erhaltene Faktoren sind
R (trait)	Zeitpunkt (Situation)	Variablen	**Eigenschaftsfaktoren**
Q	Zeitpunkt	Personen	**Typenfaktoren**
P (state)	Versuchsperson	Variablen	**Zustandsfaktoren**
O	Versuchsperson	Situationen	Situationsfaktoren
T	Variable	Situationen	Situationsfaktoren
S	Variable	Person	Typenfaktoren

Tabelle 2.2: Basic Data Relation Matrix BDRM oder Data-Box mit 10 Gesichtspunkten (Cattell, 1966)

Modalität	Phase
(1) Organismus (Person)	(6) Zustände des Organismus
(2) Fokaler Reiz	(7) Varianten des fokalen Reizes
(3) Reizhintergrund	(8) Phasen des Reizhintergrundes
(4) Verhaltensweise	(9) Stadien der Verhaltensweise
(5) Beobachter	(10) Zustände des Beobachters

Methoden der Datenerhebung

Die hier in Anlehnung an Fiske, Seidenstücker, Baumann u.a. Autoren entwickelte **Taxonomie von Methoden der Datenerhebung** unterscheidet Datenebenen, Aufgaben für Proband, Beobachter und Untersucher, Reaktivität, Herkunft des Index (vom Proband, vom Beobachter), häufige Fehlerquellen und andere Aspekte (siehe Tabelle 2.3).

Im Praktikum I diente das Phänomen "Angst" als Beispiel für die Operationalisierungsdiskussion. Durch die Operationalisierung wird eine Entscheidung getroffen, welche Ausschnitte (Komponenten) des Phänomens überhaupt berücksichtigt werden, und in welchen Ausschnitten die empirische Auseinandersetzung bzw. Anwendung einer theoretischen Konzeption überhaupt erfolgt. Dieser Entscheidungsprozeß und seine fachliche Rechtfertigung sind mindestens so wichtig wie die Berücksichtigung der konventionellen Gütekriterien. Als **Gütekriterien** sind zu nennen: (1) **Validitätsnachweis** einer speziellen Methode, (2) **Reliabilität**, (3) **Objektivität** und weitere, praktisch wichtige Gesichtspunkte wie (4) **Ökonomi**e, d.h. notwendiger Aufwand an Training, Zeit und Mitteln für eine Methode, (5) **Bekanntheitsgrad** einer Methode in bisheriger Forschung und Literatur, (6) **Vergleichsmöglichkeiten** mit anderen Studien, Norm- und Vergleichs-Werten und gewiß auch (7) die **Zumutbarkeit, Akzeptanz und Plausibilität** (aus der Sicht des Untersuchten "face validity") einer speziellen Methode für eine bestimmte Aufgabenstellung (siehe auch Gütekriterien psychologischer Tests).

Die Erhebung und Verwendung empirischer Daten für psychologische Fragestellungen kann, wie bereits betont, in schwierige methodologische und wissenschaftstheoretische Diskussionen führen (Methodologie = Lehre von den Methoden) führen.

2.4 Mathematisierung, Messung, Skalierung

Ist das, was rational und klar zu denken ist, grundsätzlich in einem mathematischen Kalkül zu formulieren oder gibt es bestimmte Grenzen?

"Wer die höchste Gewißheit der Mathematik schmäht, nährt seinen Geist von Verwirrung und wird den sophistischen Lehren, die immer nur auf Wortstreitigkeiten hinauslaufen, niemals Einhalt gebieten können" (Leonardo da Vinci).
"Messen was meßbar ist, und meßbar machen, was noch nicht meßbar ist" (Galilei zugeschrieben).
"Ich behaupte aber, daß in jeder besonderen Naturlehre nur so viel eigentliche Wissenschaft angetroffen werden könne, als darin Mathematik vorkommt" (Kant).
"Wir sind überzeugt, daß letztlich eine befriedigende Erklärung von Gedanken und Verhalten in einer Sprache gegeben werden kann, die wie die Sprache der Physik ist, d.h. in mathematischen Begriffen" (Braitenberg, 1992).

Können diese Auffassungen auch für die Bewußtseins- und für die Verhaltenspsychologie gelten? Ist der "Mensch meßbar"? Oder ist die Gegenüberstellung von "qualitativ" vs. "quantitativ" nur ein Pseudogegensatz, weil es in jeder Qualität Abstufungen, eben das quantitative Mehr oder Minder gibt, welches mit Quantoren-Namen belegt wird (siehe Graumann, Stegmüller)? Wahrscheinlich verbirgt sich hinter dem **Gegensatz quantitativ – qualitativ** oft der **Gegensatz eindeutig – mehrdeutig**, d. h. ein Problem der größeren oder geringeren bzw. fehlenden objektiven Prüfbarkeit bzw. Konvergenz der Interpretation. "Qualitative" Psychologie ist ein auch fachpolitisch zu verstehendes Modewort. Wenn empirische Aussagen formuliert werden, dann sind diese in der Regel auch quantitativ abgestuft, zumindest im Sinne einer vorhanden – nicht-vorhanden und größer-kleiner Beziehung (siehe unten: Nominal- und Ordinal Skala).

Messen heißt Abbildung eines empirischen auf ein numerisches relationales System, wobei eine eindeutige Abbildung nur dann erreicht ist, wenn Isomorphie, also Strukturgleichheit besteht (Tarski). Die Abbildungsfunktion des empirischen in ein numerisches System wäre dann die Skala (siehe Orth, 1983).

Weitere Definitionen von Messung:
Campbell (1938): *„Messung ist die Zuordnung von Zahlzeichen zur Darstellung von Eigenschaften materieller Systeme, die keine Zahl sind, aufgrund der diese Eigenschaften beherrschenden Gesetze."*
Stevens (1959): *„Zuordnung von Zahlen zu Objekten oder Ereignissen entsprechend einer Regel – irgendeiner Regel."*
Orth (1974): *„Messung ist die Bestimmung der Ausprägung der Eigenschaft eines Dinges. Das Messen erfolgt durch die Zuordnung von Zahlen zu Dingen, die Träger der zu messenden Eigenschaft sind, und beruht auf einer homomorphen Abbildung eines empirischen Relativs in ein numerisches Relativ."*

Seit Stevens (1946) wurden in der Methodenlehre der Psychologie **vier Skalentypen** unterschieden: Nominal-, Ordinal-, Intervall- und Verhältnisskala. Außerdem können noch die logarithmische Intervallskala und die absolute Skala hervorgehoben werden. Diese Skalentypen sind durch die jeweils zulässigen Transformationen gekennzeichnet – vereinfacht gesagt:

Nominalskala: Meßwerte nur zur Benennung eines empirischen Objekts;
Ordinalskala: monoton steigende Funktion, d.h. die Rangordnung von Objekten bleibt invariant (größer-kleiner-Relation);
Intervallskala: positiv lineare Transformationen, d.h. die Verhältnisse von Intervallen bleiben invariant;
logarithmische Intervallskala, d.h. Potenztransformationen sind zulässig (durch log. Transformation in Intervallskala überführbar);

Verhältnisskala: Ähnlichkeitstransformationen sind zulässig, d.h. die Verhältnisse von Skalenwerten bleiben invariant;
absolute Skala: Identitätstransformationen (v = v) lassen die Skalenwerte selbst invariant – die eindeutigste Form.

In der Psychophysik sind Skalierungsverfahren entwickelt worden, z. B. durch Paarvergleich, durch auf- und absteigende Verfahren, durch sog. Herstellungsverfahren usw.
Thurstone-Skala: gleich erscheinende Intervalle, durch Paarvergleich entwickelt;
Guttman-Skala: kumulative Skala homogener Items steigender Schwierigkeit;
Likert-Typ-Skala: Summation über die Itembeantwortungen, z. B. eine Anzahl fünfstufiger Items einer Einstellungsskala.

Skalierungsmodelle postulieren einen spezifischen Zusammenhang zwischen Skalenwerten und psychologischen Objekten oder Beobachtungsdaten. Skalierungsmodelle haben sowohl psychologisch-theoretische als auch meßtheoretische Anteile. Psychologisch-theoretische Anteile, weil sie Hypothesen über den psychologischen Gegenstandsbereich enthalten, z. B. aufgrund welcher Operationen eine Person zwei Reize vergleicht. Meßtheoretische Anteile, weil Skalierungsmodelle die Repräsentation von psychologischen Objekten und ihren Relationen im numerischen Relativ spezifizieren.

Semantische Modelle für "innere Realität". Die innere Realität eines Menschen – Wahrnehmung, Denken, Fühlen – ist einer direkten Beobachtung durch andere nicht zugänglich. Sie muß daher in einem anderen Medium mitgeteilt werden, was eine Transformation der **Bedeutung** bzw. **Semantik** der inneren Realität erfordert. Medien für solche Abbilder können Gegenstände (Kunstwerke), Ausdrucksverhalten (Mimik etc.), Alltagssprache, dichterische Sprache, Wissenschaftssprache, **numerische Systeme (Meßmodelle)**, nicht-numerische formale Systeme sein. Bei der Wahl eines Mediums ist die kritische Reflexion vonnöten, welche semantischen Strukturen jeweils abbildungstreu sind. Numerische Systeme haben einen hohen Organisationsgrad und große Flexibilität für verschiedenste Abbildungsvorschriften.

Repräsentationstheorie der Messung (nach Gigerenzer, 1981). Definitionen: Ein System, welches eine Menge numerischer Objekte (Zahlen, z. B. natürliche, ganze, rationale, reelle oder Vektoren) und mindestens eine numerische Relation zwischen diesen Objekten (z. B. =, ·, +) enthält, heißt **numerisches Relativ (System)**. Ein System, welches aus mindestens einer Menge von empirischen Objekten (z. B. Personen, Institutionen, psychischen Eigenschaften) und aus mindestens einer empirischen Relation (z. B. Gleichheit bzw. Ununterscheidbarkeit, Dominanz bzw. Präferenz, Ähnlichkeit) besteht, heißt **empirisches Relativ (System)**. Die Zuordnung jedes Objekts einer Menge A zu genau einem Objekt der Menge B heißt **eindeutige Abbildung**. Die Rück-Zuordnung einer eindeutigen Abbildung auf genau das Ausgangsobjekt der Menge A heißt **ein-eindeutige Abbildung**. Die eindeutige Abbildung von Objekten inklusive der Relationen zwischen ihnen heißt **homomorphe Abbildung (Homomorphismus)**. Eine ein-eindeutige homomorphe Abbildung heißt **Isomorphismus** bzw. Strukturgleichheit. Eine homomorphe Abbildung eines empirischen Systems in ein numerisches System heißt **Messung**.

Das **Repräsentationstheorem** erläutert, ob ein empirisches Relativ auf ein geeignet gewähltes numerisches Relativ homomorph abgebildet werden kann (ist Messung überhaupt möglich?). Dabei werden Axiome als notwendige und hinreichende Bedingungen einer Repräsentation aufgestellt und überprüft.

Das **Eindeutigkeitsproblem** fragt, welche Klassen von Abbildungsfunktionen zulässig sind, welche dieselbe homomorphe Abbildung erzeugen. Dies führt zum Konzept des **Skalenni-**

veaus und der **zulässigen Transformationen** innerhalb des numerischen Relativs. Das **Bedeutsamkeitsproblem** betrifft die Frage, welche numerischen Operationen und Statistiken auf welchem Skalenniveau sinnvoll (in bezug auf die semantische Transformation) sind. Skalenniveaus gehören aber **nicht** zu den statistischen Voraussetzungen statistischer Verfahren ("the numbers do not know where they came from").

2.5 Meß- und Skalierungsprobleme

Offensichtlich genügen nur wenige der gängigen Quantifizierungen in der psychologischen Forschung dem prägnanten Begriff der Messung (siehe Definition von Orth oder Gigerenzer) oder der Konzeption von Verhältnisskalen (oft nicht einmal Intervallskalen). Statt einer repräsentierenden, abbildenden Messung handelt es sich um eine **Indexerfassung** (index vs. representational measurement nach Dawes), die einem anderen Prinzip zu folgen scheint. Dies läßt sich am Beispiel einer Einstellungsskala zeigen: "Psychotherapie als normale Krankenkassen-Leistung dringend einzuführen": - 3 lehne entschieden ab 0 unentschieden + 3 stimme völlig zu.

Hier wird kein empirisches Relativ isomorph und intern konsistent abgebildet, sondern nur **ein Index gewonnen, dessen Nützlichkeit aus der Korrelation mit den interessierenden Kriterien, d.h. aus der Prädiktorleistung**, bestimmt werden könnte.

In psychologischen Untersuchungen treten viele weitere Methodenprobleme auf, welche auf Prinzipien der allgemeinen Datentheorie und Meßtheorie zurückverweisen. Durch Meß- und Testwiederholungen werden funktionell, z. T. auch rechnerisch voneinander **abhängige Daten** gewonnen. Als **Autokorrelation** wird die korrelative Beziehung aufeinander folgender Beobachtungen bezeichnet. Gewöhnungseffekte, Einstellungsänderungen usw. können zu erheblichen Effekten und korrelierten Fehlern führen. Solche Abhängigkeiten ziehen Probleme für valide Operationalisierungen und für statistische Inferenzen nach sich (Möbus & Nagl, 1983).

Die Messung einer Veränderung, d. h. der Zunahme oder der Abnahme eines quantitativ erfaßten Merkmals, scheint zunächst eine einfache Angelegenheit zu sein. Biometrisch ist jedoch dies „measurement of change“ eine komplizierte Angelegenheit. Bei der **Definition von geeigneten Veränderungsmaßen (Reaktionswerten) müssen funktionelle, statistische (sog. Fehlermodelle) und rechnerische Abhängigkeiten bedacht werden.** In einigen Funktionsbereichen bestehen **Ausgangswert-Beziehungen**, d.h. systematische Beziehungen zwischen dem Ausgangsniveau (baseline, Ruhewert) und dem Betrag der Veränderung, außerdem gibt es sog. **Deckeneffekte und Bodeneffekte**. Außer den Differenzen (Belastungswert minus Ausgangs-(Ruhe)wert, synomym mit „Differenz zwischen Prä – zu Post – Messung“) werden auch kovarianzanalytisch gewonnene Residuen (Auspartialisierung von Ausgangswertunterschieden) sowie spezielle Reaktionswerte aufgrund bestimmter Meßmodelle (sog. "wahre Werte" unter Berücksichtigung der Reliabilität aufgrund einer wiederholten Messung des Ausgangswertes) verwendet. Diese Abhängigkeiten erschweren die Effektbeurteilung. **Veränderungsmessungen sind jedoch für viele psychologische Fragestellungen notwendig (Prozeßforschung).**

Formulierung von Adjektivskalen und Items

Für viele psychologische Untersuchungen werden Fragebogen, Adjektivskalen oder andere „Items“ zur Selbst- und Fremdeinstufung von Befindlichkeit, Verhaltensmerkmalen u. a. psychologischen Aspekten benötigt. Solche Items bzw. Skalen werden, sofern nicht Standardverfahren verfügbar sind, oft ad hoc, d. h. ohne gründliche Methodenentwicklung und Vorprüfung entworfen.

Bei der sprachlichen Formulierung sind aber verschiedene Gesichtspunkte zu bedenken: umgangssprachlich verständlich, semantisch möglichst eindeutig, d. h. unter Vermeidung von Fremdwörtern, von komplizierter Grammatik, doppelter Verneinung. Hier helfen u. U. gründliche Formulierungsversuche und Diskussion in einer Gruppe. Wie kann das Gemeinte am besten ausgedrückt werden? Häufig gibt es Diskussionen um die Graduierungen, d. h. die Anzahl der Stufen und die Benennung bzw. Verankerung der Stufen.

Anzahl der Stufen. Die Meinungen sind geteilt, wie viele Stufen sinnvoll sind. Die Antwort hängt von dem Iteminhalt, von der Formulierung (unipolar: gutgelaunt ... nicht gut gelaunt, bipolar: gut gelaunt ... schlecht gelaunt), von der Population u. a. Aspekten ab. Bei Studierenden kann durchaus auch an mehr als 7 Stufen, sogar 11 oder 13 Stufen, bzw. grundsätzlich an die noch feiner abstufbare, 100 mm lange „Visual Analogue Scale“ gedacht werden. Die numerische Skala „perceived exertion“ von Borg hat sogar 21 Stufen. Sowohl die Skala mit ungerader Anzahl von Stufen (viele Personen wollen eine mittlere Stufe) als auch die Skala mit gerader Anzahl (die Probanden werden zur Entscheidung gezwungen) haben Argumente für sich.

Graduierung. Hinsichtlich der Quantoren gibt es eine Anzahl von methodischen und empirischen Untersuchungen. Rohrmann (1978) plädiert für die folgenden Abstufungen, gibt jedoch auch Hinweise für feinere Graduierungen:

Häufigkeitsskala
nie – selten – gelegentlich – oft – immer
Intensitätsskala
(gar) nicht – wenig – mittelmäßig – überwiegend (oder ziemlich, annähernd) – völlig
Wahrscheinlichkeitsskala
keinesfalls – wahrscheinlich nicht – vielleicht – ziemlich wahrscheinlich – sicher
Bewertung von Aussagen
stimmt nicht – wenig – mittelmäßig – ziemlich – sehr oder
gar nicht – wenig – teils-teils – ziemlich – völlig zutreffend

Schwarz und Scheuring (1992) nehmen an, daß die Formulierung der Skala implizit bereits Urteilsheuristiken nahelegt. Die mittlere Skalenstufe könnte u. U. als „normaler“ Wert der Antwort interpretiert werden.

Weitere Aspekte. Es gibt empirische Befunde, daß rückblickende Einstufungen sehr fragwürdig sind. Es kann systematische Verzerrungen geben (sog. recall-error, hindsight bias, negative Retrospektionseffekte). Deswegen kann es zweckmäßig sein, computer-unterstützte aktuelle Datenerhebungen vorzusehen (z. B. das Programm MONITOR von G. Brügner, 1998). Bei ungewöhnlichen und schwierigen Themen oder heterogenen Populationen, aber auch sonst kann es sinnvoll sein, **Meta-Fragen** zu stellen: Fragen nach der Verständlichkeit, nach der (oft zu großen) Anzahl der Items und der allgemeinen Akzeptanz für diese Methode. Zur graphischen Gestaltung von Items und Adjektivskalen gibt es viele Varianten. Von Ausschmückungen, redundanten oder ablenkenden Zusätzen, z. B. den Gesichtern mit mimischen Varianten (Smiley) ist abzuraten. Die Kombination von Zahlen (Noten), + und – Zeichen sowie verbalen Graduierungen wirkt überladen. Klarheit und Übersichtlichkeit sowie der „Aufforderungscharakter“ des Layouts, an der individuell zutreffenden Stelle das Kreuzchen zu machen, sind hier wichtig.

2.6 Von der Fragestellung zur Hypothese

Auf dem Wege von einer interessanten wissenschaftlichen Fragestellung zu einer empirisch prüfbaren Hypothese sind mehrere wichtige Vorentscheidungen zu treffen. Die Operationalisierungen haben ja nicht nur formale und durchführungspraktische Aspekte, sondern enthalten

die fundamentalen theoretischen Entscheidungen, **wie** die theoretischen Begriffe und Fragen mit den empirisch faßbaren Daten verknüpft werden sollen. Diese Entscheidungen erfordern gute Kenntnisse der speziellen Publikationen zu dem betreffenden Arbeitsgebiet (internationaler „Stand der Forschung“), Kenntnisse der möglichen Operationalisierungen, Training in deren Anwendung und methodische Reflektion zur Begründung dieser Entscheidungen. Relativ zu anderen Aspekten der Untersuchungsplanung scheinen diese Vorentscheidungen oft zu wenig reflektiert und diskutiert bzw. in Voruntersuchungen und Operationalisierungsstudien zu selten geprüft zu werden.

Ausgewählte Literatur

Breuer, F. (1988). *Wissenschaftstheorie für Psychologen* (4. Aufl.). Münster: Aschendorff.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, **56**, 81-105.

Cattell, R.B. (1966). The principles of experimental design and analysis in relation to theory building. In R.B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 19-66). Chicago: Rand McNally.

Chalmers, A.F. (1986). *Wege der Wissenschaft*. Berlin: Springer.

Danner, H. (1994). *Methoden geisteswissenschaftlicher Pädagogik* (3. Aufl.). München: Reinhardt.

Erdfelder, E. (1994). Erzeugung und Verwendung empirischer Daten. In T. Herrmann & W.H. Tack (Hrsg.) *Enzyklopädie der Psychologie*. *Themenbereich B. Methodologie und Methoden. Serie I Forschungsmethoden der Psychologie. Band 1 Methodologische Grundlagen der Psychologie* (S. 47-97). Göttingen: Hogrefe.

Erdfelder, E. & Bredenkamp, J. (1994). Hypothesenprüfung. In T. Herrmann & W.H. Tack (Hrsg.) *Enzyklopädie der Psychologie*. *Themenbereich B. Methodologie und Methoden. Serie I Forschungsmethoden der Psychologie. Band 1 Methodologische Grundlagen der Psychologie* (S. 604-648). Göttingen: Hogrefe.

Feger, H. & Bredenkamp, J. (Hrsg.) (1983). *Datenerhebung. Enzyklopädie der Psychologie*. *Themenbereich B. Methodologie und Methoden. Serie 1. Forschungsmethoden der Psychologie. Band 2. Datenerhebung.* Göttingen: Hogrefe.

Gadenne, V. (1994a). Theorien. In T. Herrmann & W.H. Tack (Hrsg.) *Enzyklopädie der Psychologie*. *Themenbereich B. Methodologie und Methoden. Serie I Forschungsmethoden der Psychologie. Band 1 Methodologische Grundlagen der Psychologie* (S. 295-342). Göttingen: Hogrefe.

Gadenne, V. (1994b). *Theoriebewertung. In T. Herrmann & W.H. Tack (Hrsg.) Enzyklopädie der Psychologie*. *Themenbereich B. Methodologie und Methoden. Serie I Forschungsmethoden der Psychologie. Band 1 Methodologische Grundlagen der Psychologie* (S. 388-427). Göttingen: Hogrefe.

Gigerenzer, G. (1981). *Messung und Modellbildung in der Psychologie*. München: Reinhardt. (UTB-Taschenbuch 1047).

Groffmann, K. J. & Michel, L. (Hrsg.) (1983). *Enzyklopädie der Psychologie*. *Themenbereich B. Methodologie und Methoden. Serie 2. Psychologische Diagnostik. Band 1-4.* Göttingen: Hogrefe.

Hager, W. & Westermann, R. (1983). Planung und Auswertung von Experimenten. In J. Bredenkamp & H. Feger (Hrsg.). *Enzyklopädie der Psychologie*. *Themenbereich B. Methodologie und Methoden. Serie I Forschungsmethoden der Psychologie. Band 5 Hypothesenprüfung* (S. 24-238). Göttingen: Hogrefe.

Hussy, W. & Möller, H. (1994). Hypothesen. In T. Herrmann & W.H. Tack (Hrsg.) *Enzyklopädie der Psychologie*. *Themenbereich B. Methodologie und Methoden. Serie I Forschungsmethoden der Psychologie. Band 1 Methodologische Grundlagen der Psychologie* (S. 475-507). Göttingen: Hogrefe.

Möbus, C. & Nagl, W. (1983). Messung, Analyse und Prognose von Veränderungen. In J. Bredenkamp & H. Feger (Hrsg.) *Enzyklopädie der Psychologie. Themenbereich B. Methodologie und Methoden. Serie 1 Forschungsmethoden der Psychologie. Band 5 Hypothesenprüfung.* (S. 239-470). Göttingen: Hogrefe.

Nübling, R. (1991). *Psychotherapiemotivation und Krankheitskonzept. Zur Evaluation psychosomatischer Heilverfahren.* Phil. Diss. Universität Freiburg. Frankfurt a. M.: VAS Verlag.

Orth, B. (1983). Grundlagen des Messens. In H. Feger & J. Bredenkamp (Hrsg.). *Enzyklopädie der Psychologie*. *Themenbereich B. Methodologie und Methoden. Serie I Forschungsmethoden der Psychologie. Band 3 Messen und Testen* (S. 136-180). Göttingen: Hogrefe.

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*, **9**, 222-245.

Schwarz, N. & Scheuring, B. (1992). Selbstberichtete Verhaltens- und Symptomhäufigkeiten: Was Befragte aus Antwortvorgaben des Fragebogens lernen. *Zeitschrift für Klinische Psychologie*, **21**, 197-208.

Schweizer, K. (1989). Eine Analyse der Konzepte, Bedingungen und Zielsetzungen von Replikationen. *Archiv für Psychologie*, **141**, 85-97.

Stegmüller, W. (1970 ff). *Probleme und Ergebnisse der Wissenschaftstheorie und Analytischen Philosophie*. Berlin: Springer. - Insbesondere Bd. IV, St. Teil A, Einleitung S. 1-95 (Inhalt der Wissenschaftstheorie, Wertproblem, Wahrscheinlichkeit, Induktion). Bd. I St. Teil 1 Kap. I S. 72-153 (Begriff der Erklärung).

Themenheft "Multimodale Diagnostik". (1987). *Diagnostica*, **33**, (Heft 3).

3. Auswahlentscheidungen und Stichprobentechnik. Prüfung von Theorien/Hypothesen

3.1 Auswahlentscheidungen

Psychologische Forschung strebt – soweit es sich nicht um idiographisch orientierte Einzelfallstudien handelt – Gesetzeshypothesen (im Sinne statistischer Analysen und Begründungen) mit allgemeiner Geltung für eine bestimmte Population von Individuen an. Da die Vollerhebung einer Population nur selten möglich sein wird, sind Schlußfolgerungen (Inferenzen, Repräsentationsschlüsse) notwendig. Als Stichprobe (sample) wird ein systematisch gewonnene Auswahl der Population (Grundgesamtheit) bezeichnet; falls das Auswahlprinzip nicht genau bekannt ist, sprechen wir von "Gelegenheitsstichprobe" oder besser nur von "Gruppe", um diese Unsicherheit auszudrücken.

Auswahlprinzipien: Variablen, Settings, Meßwiederholungen
Beim Begriff Stichprobe wird gewöhnlich an eine Personen-Stichprobe gedacht, doch sind in der Datenbox auch andere Stichproben möglich: Stichprobe von Variablen, falls eine Grundgesamtheit (Universum der Variablen, z. B. Intelligenzaufgaben oder Stimmungs-Items) angenommen werden kann; Stichprobe von Settings bzw. Situationen (Reizbedingungen), wobei hier die Population nicht ohne weiteres zu bestimmen ist; Stichprobe von Meßwiederholungen, d. h. Terminen bzw. Ereignissen (time- and event-sampling). Der Begriff „Stichprobe“ bleibt aber in diesem Zusammenhang fragwürdig, weil die Grundgesamtheit z. B. der Settings nicht ohne weiteres definiert werden kann. Deshalb ist besser von **Auswahlstrategien hinsichtlich der Datenbox** zu sprechen.

Die Auswahlentscheidungen sollen in engem Zusammenhang mit den Überlegungen zur Operationalisierung der theoretischen Konstrukte stehen und müssen unter den Gesichtspunkten der **internen und externen Validität** des Versuchsplans – für die betreffende Fragestellung – als adäquat (optimal) gerechtfertigt werden. Wichtige Auswahlprinzipien sind u. a. **repäsentatives Design** (Brunswiks Linse), **Datenbox** (Cattell, Wittmann), **konvergente – diskriminate Validität** (Campbell & Fiske), Zuverlässigkeit der Befunde, Stabilität bzw. Replizierbarkeit. Maßgeblich werden außerdem der praktisch mögliche Aufwand, verfügbare Kompetenzen, bereits vorhandene Instrumente, Zumutbarkeit und Ökonomie sein.

Begründung der Auswahlentscheidungen?
Personen bzw. Gruppe von Personen
Variablen UV AV KV
Settings bzw. experimentelle Paradigmen
Termine bzw. Meßwiederholungen
Repräsentatives Design/ Symmetrieüberlegungen?
Interne und externe Validität
Ökonomie u. a. Gesichtspunkte

3.2 Stichprobentechnik (Personen)

Die folgende Übersicht folgt Mayntz, Holm und Hübner (1969) und Kerlinger (1973) zur Stichprobenmethodik von Personen. **Jede Stichprobe ist nur ein fragwürdiger Ersatz für die wünschenswerte Totalerhebung.**

Zufallsstichprobe (random sample)
Die Zufallsstichprobe ist die logisch zwingende Voraussetzung für die Generalisierung auf die Population (Repräsentationsschluß). Zur Ziehung einer Zufallsstichprobe muß **jede Einheit der Grundgesamtheit dieselbe Chance haben, in die Auswahl aufgenommen zu**

werden. Diese Chancengleichheit ist nur selten uneingeschränkt herzustellen, denn die Grundgesamtheit müßte physisch oder symbolisch gegenwärtig und manipulierbar sein (Mischung, Auswahltechniken). In gleich großen Zufallsstichproben aus einer Grundgesamtheit werden die Mittelwerte quantitativer Merkmale X eine angenäherte Normalverteilung ergeben, deren Maximum beim wahren Wert μ in der Grundgesamtheit liegt. Als Streuungsmaß dient die Standardabweichung der Mittelwerte aller Stichproben $\sigma_{\bar{X}}$. (Für Alternativmerkmale p, q gilt Entsprechendes). Mit Hilfe dieser $\sigma_{\bar{X}}$ lassen sich Vertrauensintervalle ausdrücken, z.B. 95 % aller gezogenen Stichproben werden einen Mittelwert X aufweisen, der im Intervall von $\mu \pm 1{,}96\ \sigma_{\bar{X}}$ liegt (= mit 95 %iger Sicherheit).

Statt von den bekannten Merkmalen in der Grundgesamtheit auf die zu ziehende Stichprobe zu schließen, wird der Empiriker in der Regel gerade aus den an Stichproben ermittelten Werten auf die unbekannten Werte (Parameter) der Grundgesamtheit zu schließen versuchen (Parameterschätzung durch sog. Repräsentationsschluß). Als Schätzwerte verwendet man also die in der gezogenen Stichprobe gefundenen Werte unter der Voraussetzung, daß diese Stichprobe eine hinreichende Größe (Daumenregel mindestens n = 30) aufweist. Der geschätzte Mittelwert und seine geschätzte Standardabweichung lassen dann eine Aussage zu, in welchem Vertrauensintervall mit welcher Sicherheit der weiterhin unbekannte Wert μ liegen wird. Damit ist der Abbild- bzw. Auswahlfehler bestimmt.

Die für eine statistische Aussage dieser Art notwendige Stichprobengröße hängt ab: 1. vom gewünschten Sicherheitsniveau (z. B. 95 %, 99 %), 2. von der Breite des Vertrauensintervalls, das noch als tragbar angesehen wird, 3. von der Größe der Grundgesamtheit: je höher das Signifikanzniveau, je kleiner das Vertrauensintervall und je größer die Grundgesamtheit, um so größer muß die Stichprobe sein.

Sonderformen der Zufallsstichprobe

(a) **Geschichtete Zufallsstichprobe** dient dazu, für bestimmte Merkmale die Zufallsstreuung zu reduzieren oder auszuschalten.

Proportionale Schichtung. Die Schichtung erfolgt nach dem Merkmal, dessen Parameter gesucht werden bzw. nach einem möglichst eng verbundenen Ersatzmerkmal (z. B. sozialer Status durch Einkommen oder Schulbildung). Dabei müssen die Größe der Grundgesamtheit und der Schichten bekannt sein (z. B. amtliche Statistik). Aus jeder der Schichten wird dann eine Zufallsstichprobe gezogen, z. B. 1 % jeder Schicht, so daß die resultierende Gesamtstichprobe hinsichtlich des Schichtungsmerkmals völlig repräsentativ für die Grundgesamtheit ist – eine direkte Zufallsauswahl hätte vermutlich einen größeren Auswahlfehler. Die **geschichtete Stichprobe kann deswegen bei gleicher Güte kleiner sein als die normale Zufallsstichprobe**.

Disproportionale Schichtung. Falls das Schichtungsmerkmal zu starken Unterschieden der Schicht-Größe führen würde, könnten diese Teil-Stichproben gleich groß gemacht werden (z.B. gleich viel Patienten in jeder Diagnosegruppe).

Mehrdimensionale Schichtung nach mehreren Merkmalen ist dann sinnvoll, wenn – wie in den anderen Fällen – eine deutliche Abnahme der internen Streuung der einzelnen Schicht (im Vergleich zur Streuung zwischen den Schichten) und damit eine deutliche Einengung des Vertrauensintervalls der Gesamtstichprobe geschieht.

(b) **Mehrstufen- und Klumpenauswahl** sind angezeigt, wenn die Grundgesamtheiten für ein Auswahlverfahren nicht vollständig zugänglich sind. So kann z. B. die Bevölkerung einer Stadt nach Häuserblocks mit ungefähr gleich großer Bevölkerungszahl eingeteilt werden.

Eine Zufallsstichprobe dieser Blocks wird gezogen und von diesen Blocks („Klumpen“ oder „Cluster“) werden nun **alle** Bewohner in die Stichprobe einbezogen. In dieser zweiten Stufe könnte aber auch eine erneute Zufallsstichprobe gezogen werden (zweistufige Zufallsauswahl). Je größer die Zahl der Stufen, desto größer auch das Vertrauensintervall.

Quotaverfahren ist ein Annäherungsverfahren an die Zufallsauswahl, indem die als **wesentlich angesehenen Merkmale** der Grundgesamtheit quotiert, d.h. systematisch repräsentiert werden. Wir gehen von Merkmalen (Alter, Geschlecht usw.), deren Verteilung in der Grundgesamtheit bekannt ist, aus und hoffen dann, daß auch die anderen Merkmale mehr oder weniger repräsentativ vertreten sein werden. Dies wird für solche Untersuchungsmerkmale gelten, die mit den gewählten Quotenmerkmalen relativ eng zusammenhängen. Der Interviewer wählt die Zielpersonen selbst aus, diese müssen jedoch in bestimmte Zellen des Quotenschemas passen: Keine Zufallsauswahl, sondern Konstruktion mit einer Überprüfungsmöglichkeit anhand nicht quotierter, aber in der Verteilung bekannter Merkmale. **Diese Gütemerkmale sind hier mehr oder minder überzeugende Behauptungen, also etwas anderes als der mathematisch-statistische Repräsentationsschluß tatsächlicher Zufallsauswahl**.

Fehlerquellen
Die Quotenauswahl hat spezielle Fehlerquellen (Interviewer-Bias, Auffüllen bei Nicht-Antreffen oder Verweigerung, u.U. eine zu indirekte Beziehung zwischen Quoten- und Untersuchungsmerkmalen), während die eigentlichen Zufallsauswahl-Verfahren sehr oft unrealistisch oder nur mit sehr großem Aufwand und Datenschutzproblemen lösbar wären. Hier ist der Untersucher oft zu Zugeständnissen gezwungen, welche allerdings manchmal verschwiegen werden.

Viel zu wenig diskutiert wird in den Lehrbüchern auch das Problem sog. Verweigerer, das gerade bei psychologischen Fragestellungen, wo diese Einstellung bzw. Verhaltensweise psychologisch bedeutsam sein könnte, zu schwerwiegenden Verzerrungen der Repräsentativität führen wird. Nicht selten wird z. B. eine Rücklaufquote von 70 oder 80 % des ausgegebenen Materials als üblich oder gut bezeichnet. Oft wird es sich dennoch um eine massive Verzerrung (einen bias) handeln (Myrtek, 1987, z. B. beschrieb, daß sich die 10 % der Patienten, welche den Katamnese-Fragebogen über die Rehabilitation nach Herzinfarkt **nicht** ausfüllten, bereits im FPI-Eingangstest in mehreren Skalen signifikant von den übrigen 90 % unterschieden hatten). Das Verweigerer-Problem ist aus naheliegenden Gründen schwer zu untersuchen.

Leider beruht psychologische Forschung bis heute ganz überwiegend auf **unrepräsentativen Gruppen** ("Gelegenheitsstichproben") wie Schüler(innen), Studenten(innen) oder sogar nur Psychologie-Studenten(innen) sowie kleinen, in sich oft extrem heterogenen Patientengruppen, so daß bei Generalisierungsversuchen größte Vorsicht angebracht ist. **Repräsentationsschlüsse ohne Zufallsstichproben sind grundsätzlich fragwürdig und unzuverlässig**.

3.3 Bestimmung des Stichprobenumfangs

Nach der notwendigen Anzahl von Vpn (bzw. Beobachtungs- oder Untersuchungseinheiten) für jede Untersuchungsbedingung (Symbol: n) oder für die gesamte Untersuchung (Symbol: N) wird oft gefragt. Faustregeln wie n = 20 für jede Zelle der ANOVA oder N = 100 ("well over a hundred") für eine multiple Regression (bzw. ca. 30 Vpn pro Prädiktor) stammen aus der Erfahrung, wirken jedoch willkürlich und sind keinesfalls auf verschiedene Untersuchungsvorhaben generalisierbar.

Die Bestimmung der "optimalen" Anzahl von Personen (siehe Hager & Westermann 1983, S. 170) bewegt sich zwischen:

- Wähle eine möglichst (sehr) große Stichprobe, etwa N > 1000! (statistische Optimalität nach zentralem Grenzwertsatz und Gesetz der großen Zahl)
- Wähle eine möglichst (sehr) kleine Stichprobe, etwa N = 5! (Kostenüberlegungen).

Für einen optimalen Versuchsplan muß hier ein Kompromiß gefunden werden. Angesichts der großen Zahl unreproduzierbarer „Resultate" darf auf ein kritisches Anspruchsniveau nicht verzichtet werden.

Grundsätzlich ist zu unterscheiden, ob es sich (1) um die (Intervall-) Schätzung eines Parameters, (2) um die Prüfung von statistischen Hypothesen oder (3) um die Lösung von Auswahlproblemen (siehe psychologisches Assessment) handelt. Im Hinblick auf die **Prüfung wissenschaftlicher Hypothesen** stellen Hager und Westermann (1983, S. 171) als notwendige Bedingung fest: „*Wähle die Größe der Stichprobe derart, daß bei vorgegebenem experimentellen Mindesteffekt die Prüfung der statistischen Hypothese bei festgelegten geringen Maximalwerten für die Wahrscheinlichkeiten von Fehlern 1. und 2. Art erfolgen kann!*"
Es gibt verschiedene Ansätze und Prinzipien der Bestimmung des Stichprobenumfangs. Am bekanntesten ist der (Poweranalyse-) Teststärkenanalyse-Ansatz von Cohen bzw. Cohen & Cohen. Der (die) Planer(in) und Auswerter(in) entscheiden sich, welchen (kleinen oder größeren) Effekt sie erwarten bzw. als psychologisch bedeutsam ansehen wollen, und können dann aufgrund der Teststärkenanalyse aus Tabellen ablesen, welche Personenzahlen für die verschiedenen Typen von Versuchsplänen nötig sind (Cohen, 1977).

Poweranalyse
Die Effektstärke im **univariaten Fall** bei zwei unabhängigen Stichproben ist nach Cohen (1988) gleich dem Unterschied der Mittelwerte, der auf die gemeinsame Populations-Standardabweichung standardisiert ist $d = (\mu_1 - \mu_2) / \sigma_{gesamt}$. Cohen hat hier als Daumenregel vorgeschlagen, ein d um 0.20 als kleine, ein d um 0.50 als mittlere und ein d > 0.80 als große Effektstärke zu bezeichnen.

Tabelle 3.1: Indices für Effektstärken und deren Werte für kleine, mittlere und große Effekte (Cohen, 1992). ES = Effektstärke in der Population.

		Effect size		
Test	ES index	Small	Medium	Large
1. m_A vs. m_B for independent means	$d = (m_A - m_B) / \sigma$	.20	.50	.80
2. Significance of product-moment *r*	*r*	.10	.30	.50
3. r_A vs. r_B for independent r	$q = z_A - z_B$ where z = Fisher's z	.10	.30	.50
4. *P* = .5 and the sign test	g = P - .50	.05	.15	.25
5. P_A vs. P_B for independent proportions	$h = \Phi_A - \Phi_B$ where Φ arcsine transformation	.20	.50	.80
6. Chi-square for goodness of fit and contingency		.10	.30	.50
7. One-way analysis of variance		.10	.25	.40
8. Multiple and multiple partial correlation		.02	.15	.35

Tabelle 3.2: Stichprobengröße für kleine, mittlere und große Effektstärken für ein Signifikanzniveau von α = .01, .05 und .10 und Power = .80. ES = Effektstärke in der Population, Sm = kleine, Med = mittlere, Lg = große Effektstärken, Dif = Differenz. Die Nummerierung bezieht sich auf die Tests in der anderen Tabelle (aus Cohen, 1992).

		α = .01			α =.05			α =.10		
	Test	Sm	Med	Lg	Sm	Med	Lg	Sm	Med	Lg
1.	Mean dif	586	95	38	393	64	26	310	50	20
2.	Sig *r*	1.163	125	41	783	85	28	617	68	22
3.	*r* dif	2.339	263	96	1.573	177	66	1.240	140	52
4.	*P* = .5	1.165	127	44	783	85	30	616	67	23
5.	P dif	584	93	36	392	63	25	309	49	19
6.	χ^2									
	1 *df*	1.168	130	38	785	87	26	618	69	25
	2 *df*	1.388	154	56	964	107	39	771	86	31
	3 *df*	1.546	172	62	1.090	121	44	880	98	35
	4 *df*	1.675	186	67	1.194	133	48	968	108	39
	5 *df*	1.787	199	71	1.293	143	51	1.045	116	42
	6 *df*	1.887	210	75	1.362	151	54	1.113	124	45
7.	ANOVA									
	2 g^a	586	95	38	393	64	26	310	50	20
	3 g^a	464	76	30	322	52	21	258	41	17
	4 g^a	388	63	25	274	45	18	221	36	15
	5 g^a	336	55	22	240	39	16	193	32	13
	6 g^a	299	49	20	215	35	14	174	28	12
	7 g^a	271	44	18	195	32	13	159	26	11
8.	Mult *R*									
	2 k^b	698	97	45	481	67	30			
	3 k^b	780	108	50	547	76	34			
	4 k^b	841	118	55	599	84	38			
	5 k^b	901	126	59	645	91	42			
	6 k^b	953	134	63	686	97	45			
	7 k^b	998	141	66	726	102	48			
	8 k^b	1.039	147	69	757	107	50			

[a] Anzahl der Gruppen. [b] Anzahl unabhängiger Variablen.

Beispiele

(1) Um eine mittelgroße Differenz (d = .50) zwischen zwei unabhängigen Stichproben (siehe Tabelle 3.1, Zeile 1) mit einem α = .05 entdecken zu können, sind N = 64 in jeder Gruppe notwendig (siehe Tabelle 3.2, Spalte 5).

(2) Um einen großen Effekt einer Korrelation r ≥ .50 (Tabelle 3.1, Zeile 2) mit Signifikanzniveau α = .01 feststellen zu können, ist eine Gruppengröße von N = 41 notwendig (Tabelle 3.2, Spalte 3)

(3) Ein kleiner Effekt wird erst bei relativ hohem N deutlich, insbesondere wenn ein anspruchsvolles Signifikanzniveau (α = .01) gewählt wird.

3.4 Effektstärken oder Signifikanzen?

Sollen in wissenschaftlichen Arbeiten Effektstärken (ES bzw. d) oder Signifikanzen (p-Werte) berichtet werden? Diese Frage hat insbesondere seit Cohens Publikationen viele Diskussionen ausgelöst, wobei oft das Schema der Null-Hypothesen-Prüfung kritisiert und die wichtigere Information der Effektstärke betont wurde (z. B. Greenwald et al., 1996; Wittmann, 1985).

Wissenschaftliche Publikationen sollten Ergebnisse enthalten, welche durch andere qualifizierte Untersucher zuverlässig reproduzierbar sind. Fisher (1951, p. 14): „*To demonstrate that a natural phenomenon is experimentally demonstrable, we need not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment [that] will rarely fail to give us a statistically significant result.*"

Die Kritik an der statistischen Null-Hypothesen-Prüfung lautet:
(1) Psychologische Null-Hypothesen sind sozusagen immer falsch (unerwünscht), weswegen deren Testen **uninformativ** ist.
(2) Die Forscher möchten im Prinzip nicht einen p-Wert für die Wahrscheinlichkeit ihrer Ergebnisse relativ zur Null-Hypothese wissen, sondern primär eine Auskunft über die **Größe** von Behandlungseffekten oder die Stärke von Zusammenhängen zwischen Variablen.
(3) Die Zurückweisung der Null-Hypothese unterstützt die Annahme einer Alternativ-Hypothese, aber eine Nicht-Zurückweisung ist **nicht in symmetrischer** Weise als Unterstützung der Null-Hypothese zu interpretieren (vereinfacht gesagt: Eine Null-Hypothese ist nicht zu beweisen).

Argumente für die Prüfung der Null-Hypothese sind:
(1) Es gibt Fragestellungen, zu denen eine dichotone Antwort statt einer ES gewünscht wird, z. B.: Ist diese Behandlung einem Placebo überlegen? Sind eineiige Zwillinge einander ähnlicher als zweieiige? Ist die prädiktive Validität dieses Leistungstests hinreichend, um in die Testbatterie für die Auswahl von Bewerbern aufgenommen zu werden?
(2) Die p-Werte bilden einen allgemein verständlichen Index für den Vergleich verschiedener Prüfstatistiken wie t, F, r, χ^2 u. a.
(3) Trotz aller Vorbehalte (siehe oben) gibt der p-Wert doch einen ersten, elementaren Hinweis auf die Demonstrierbarkeit bzw. mögliche Wiederholbarkeit des Ergebnisses.

Greenwald et al. (1996) empfehlen:
(1) p-Werte nicht als < oder > berichten, sondern genau, d. h. p = ...
(2) p ≅ .050 als interessante, aber noch nicht überzeugende Unterstützung für ein einzelnes Resultat der Null-Hypothesen-Prüfung ansehen.
(3) p ≅ .005 als Indikator für die Demonstrierbarkeit eines einzelnen Resultats ansehen.
(4) p-Werte für alle wichtigen Hypothesen-Prüfungen berichten.
(5) Alle notwendigen Informationen mitteilen, damit Sekundäranalysen (Metaanalysen siehe unten) möglich sind.

Wahrscheinlich ist es auf den meisten Gebieten der anwendungsorientierten Psychologie und bei Prädiktor-Kriterien-Beziehungen sehr **viel informativer**, die **Effektstärken zu berichten und zu interpretieren**. (Dies schließt ja die Mitteilung der p-Werte nicht aus!) Effektstärken und p-Wert sind zwei verschiedene Akzentsetzungen statistischer Analysen und Begründungen, die sich ergänzen. Dasselbe Problem wird auch unter den Stichworten „klinische" und statistische" Signifikanz diskutiert.

3.5 Prüfung von Theorien/Hypothesen

Eine zentrale Aufgabe des wissenschaftlichen Arbeitens ist die Prüfung von Theorien. Der *naive Falsifikationismus* geht dabei so vor: Aus einer Theorie T wird eine empirische Folgerung E abgeleitet. Tritt E tatsächlich ein, gilt T als bewährt. Tritt E nicht ein, gilt T als falsifiziert. Diese einfache Strategie wird jedoch den realen Verhältnissen nicht gerecht, (1) weil es nur für Hypothesen einer bestimmten logischen Struktur, d.h. für die unbeschränkt univer-

sellen Hypothesen, zutrifft, und (2) weil das Hintergrundwissen (die Voraussetzungen, Vorannahmen), das mit in die Ableitung von E eingeht, unberücksichtigt bleibt.

Unbeschränkt universelle Hypothesen postulieren eine Anwendbarkeit auf alle Fälle, räumlich-zeitlich unbeschränkt. Streng genommen ist die Diskussion um Probleme der Prüfbarkeit von Theorien nur auf diesen Fall zugeschnitten. Folgende Begriffe sind auseinanderzuhalten:
Bewährung: T bewährt sich an E relativ zu dem Hintergrundwissen A genau dann, wenn E aus T und A logisch ableitbar ist, aber nicht aus A allein, und wenn A und T logisch verträglich sind.
Entkräftigung: T wird durch E relativ zu A entkräftet, wenn sich eine Negation von E ergibt.
Indifferenz: T ist indifferent gegenüber E relativ zu A, wenn T sich an E weder bewährt noch durch E entkräftet wird relativ zu A.
Falsifikation: T **und** A werden durch E falsifiziert.

Folgerungen:
- Eine Theorie, die bestens bewährt ist und die den strengsten Prüfungen widerstanden hat, ist eine Theorie, für die sich bisher alle Gründe für eine Verwerfung haben ausschließen lassen.
- Eine Theorie nicht verwerfen ist nicht gleichbedeutend mit "eine Theorie akzeptieren"!

Hypothesen (Gesetzes-Annahmen) sind vorweggenommene Antworten auf die Fragestellung. Mit der grundsätzlichen Unmöglichkeit der empirischen Verifikation von Hypothesen und mit der vergleichsweise höheren Überzeugungskraft von Falsifikationen hat sich die neuere Wissenschaftstheorie eingehend befaßt.

So ist nach strukturalistischer Wissenschaftskonzeption (Sneed, Stegmüller u. a.) **eine Theorie keine irgendwie falsifizierbare Entität**, sondern ein hierarchisiertes Netz mit einem Fundamentalgesetz, aus dem die Spezialgesetze der Theorieelemente hervorgehen. Es ist möglich, durch konsequente Anwendungen des Falsifikationsprinzips **spezielle Anwendungen** der Theorie (aber nicht die Theorie an sich) als wissenschaftlich unhaltbar zurückzuweisen. Die erfolgreichen und die erfolglosen Anwendungsversuche beantworten auch die Frage nach der adäquaten Operationalisierung. Über die Funktion der Hypothese im Forschungsprozeß, über die Beurteilung der Widerspruchsfreiheit und Operationalisierbarkeit von Hypothesen und über die Beziehungen von inhaltlichen Hypothesen und statistischen Formulierungen (einschließlich der Vorhersagen) siehe u.a. Erdfelder und Bredenkamp (1994) sowie Hussy und Möller (1994).

Zusammenfassung
Im **Übergang von der allgemeinen Fragestellung zur Formulierung einer empirisch entscheidbaren Hypothese** (spezieller Anwendung einer Theorie aufgrund von Operationalisierungen) sind also zu entscheiden:
- Präzisierung der Fragestellung unter Bezug auf die benutzte(n) Theorie(n)
- Identifizierung der zu untersuchenden Variablen
- Identifizierung desjenigen Spezialgesetzes der Theorie, das eine Antwort auf die Fragestellung ermöglicht
- Spezifizierung des empirischen Systems, auf das die Theorie bzw. das Spezialgesetz angewendet werden soll (einschließlich der Variablen der Untersuchung mit präzisen Operationalisierungen (UV, AV, KV, Instrumentierung) und einschließlich der Auswahl (Sampling) der zu untersuchenden Personen
- Formulierung der auf das Spezialgesetz und das empirische System bezogenen empirischen Hypothese

- Überlegungen zur statistischen Prüfung der inhaltlichen Vorhersagen: Zuordnung von statistischen zu den inhaltlichen Vorhersagen
- Umsetzung der statistischen Vorhersagen in direkt testbare statistische Null- und Alternativhypothesen.

Als Ergebnis der Operationalisierungn und Auswahlentscheidungen sollen die **Variablen der Untersuchung** eindeutig festgelegt sein, wobei zu unterscheiden sind:

- **Unabhängige** Variablen;
- **Abhängige** Variablen;
- **Kovariablen** bzw. Drittvariablen („Mitveränderliche“, die u. U. wichtig sind);
- **Störvariablen** (welche die Fehlervarianz erhöhen);
- **Konfundierte** Variablen (aus gemeinsamer Quelle stammend und deswegen grundsätzlich nicht voneinder zu isolieren);
- **Aktive (manipulierte)** Variablen (d. h. vom Untersucher veränderbar bzw. durch Zuweisung der Personen zu den Stufen des Treatments (Randomisierung) kontrollierbar;
- **Attributive** (nur „gemessene“) Variablen (d. h. feste Eigenschaften wie Alter, Geschlecht, Persönlichkeitsmerkmale, die unbeeinflußbar sind):
- **Prädiktorvariablen** und **Kriteriumsvariablen**.

Die endgültige Formulierung der inhaltlichen und statistischen Hypothesen muß in engem Bezug auf den gewählten Forschungs- und Versuchsplan vorgenommen werden. Hinsichtlich der sprachlichen und begrifflichen Genauigkeit der Hypothese und hinsichtlich der formalen Seite (Parameter-Schreibweise) sollte ein hohes Anspruchsniveau bestehen. Anschließend, d.h. **vor** Beginn der Datenerhebung müssen die Hypothesen protokolliert und ("im Panzerschrank") geschützt werden, damit keine nachträgliche Modifikation (Umdeutung, Immunisierung) angesichts "unerwünschter" bzw. "negativer" Ergebnisse geschehen kann.

Literaturhinweise

Bortz, J. (1995). *Forschungsmethoden und Evaluation* (2. Aufl.). Berlin: Springer. *(*1. Aufl.; *Lehrbuch der empirischen Forschung für Sozialwissenschaftler*, 1984) .

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, **112**, 155-159.

Greenwald, A.G., Gonzalez, R., Harris, R.J. & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology*, **33**, 175-183.

Hager, W. & Westermann, R. (1983). Planung und Auswertung von Experimenten. In J. Bredenkamp & H. Feger (Hrsg.) *Enzyklopädie der Psychologie*. *Themenbereich B. Methodologie und Methoden. Serie 1 Forschungsmethoden der Psychologie. Band 5 Hypothesenprüfung* (S. 24-238). Göttingen: Hogrefe.

Kerlinger, F.N. (1978/79). *Grundlagen der Sozialwissenschaften*. Weinheim: Beltz (besser die amerikanische Ausgabe, 4. Aufl. 2000).

Mayntz, R., Holm, K. & Hübner, P. (1969). *Einführung in die Methoden der empirischen Soziologie*. Köln: Westdeutscher Verlag.

Wittmann, W.W. (1985). *Evaluationsforschung. Aufgaben, Probleme und Anwendungen*. Berlin: Springer.

4. Testtheorie und Testkonstruktion. Assessment.

4.1 Definitionen

Die Theorie psychologischer Tests geht von den folgenden allgemeinen Annahmen aus:

(1) Personen unterscheiden sich in der quantitativen Ausprägung von Eigenschaften (Persönlichkeitsmerkmalen, Fähigkeiten);
(2) diese Eigenschaften sind nicht direkt beobachtbar und meßbar, sondern **theoretische** Begriffe (**latente** Eigenschaften, Konstrukte);
(3) es gibt jedoch einen indirekten Zugang über das beobachtbare (manifeste) Verhalten. Als psychologischer Test wird ein Verfahren bezeichnet, welches in standardisierter Form Verhaltensstichproben provoziert und damit in empirisch nachprüfbarer Weise einen Rückschluß von Testwerten (Indizes) auf individuelle Unterschiede in bestimmten Eigenschaften ermöglicht.

In der Regel ist davon auszugehen, daß einem Test eine entsprechende Theorie (Intelligenztheorie, Persönlichkeitstheorie, Einstellungstheorie usw.) zugrunde liegt. Gemäß einer solchen Theorie werden Items entworfen, die als Indikator für das intendierte Konstrukt (Intelligenzkomponente, Persönlichkeitsmerkmal, Einstellung usw.) dienen können. Die Brauchbarkeit dieser Items zur Erfassung eines Konstrukts wird durch die Itemselektion und die Überprüfung der Gütekriterien des Tests gewährleistet.

Lienert: „*Ein Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung*." (1967, S. 7). Der Begriff „Test" ist allerdings mehrdeutig: Test-Material, Test als Prüfungsverfahren, Test-**Ergebnis** (auch statistischer Test im Sinne von Signifikanzprüfung).

Taxonomie psychologischer Tests und Übersichten

Es gibt verschiedene Einteilungsgesichtspunkte für die weit über 5000 publizierten Tests (siehe "Mental Measurement Yearbook"). Brickenkamp (1997) unterscheidet zwischen:

Leistungstests	Entwicklungstests, Intelligenztests, Schulleistungstests, speziellen Funktions- und Eignungstests
Persönlichkeitstests	Einstellungs- und Interessentests, Persönlichkeitsfragebogen, projektiven Tests, objektive Persönlichkeitstests
Klinischen Tests	Fragebogen zu körperlichen und psychischen Störungen, Funktionsprüfungen
Verhaltenstests	standardisierte Bedingungen für Verhaltensbeobachtungen und Verhaltensmessungen

Welche Tests im deutschsprachigen Bereich verfügbar sind, ist aus den umfangreichen Übersichten (mit zahlreichen Kurzbeschreibungen der Verfahren) von Brickenkamp (1997) und Westhof (1993) sowie der Literaturbank PSYNDEXplus with Test Finder (enthält u.a. Tests 1945-1999) zu entnehmen. Interessant sind die relativen Häufigkeiten, mit denen bestimmte Tests für verschiedene Aufgaben in den Praxisfeldern eingesetzt werden. Die Umfragen von Schorr (1995) und Steck (1997) ergaben, daß Intelligenztests und Persönlichkeitsfragebogen dominieren, aber auch projektive Tests (TAT, Rorschach) immer noch häufig verwendet werden.

Ein Test kann nur dann den Anspruch eines psychologischen Diagnosesystems erfüllen, wenn die Güte des Tests gewährleistet ist. Diese mißt sich an den Gütekriterien für Tests. Zu diesen gehören: die **Objektivität**, die **Reliabilität** und die **Validität**.

4.2 Qualitätssicherung und Testgüte-Kriterien

Wegen der breiten Anwendung psychologischer Tests sind von berufsständischen Organisationen wie der American Psychological Association APA schon 1954 und 1974 und vom Testkuratorium der Föderation Deutscher Psychologenvereinigung (1986) Gütekriterien definiert worden (siehe auch Häcker, Leutner & Amelang, 1998). Diese beziehen sich auf die wissenschaftliche Testgrundlage, Testdurchführung, Testverwertung, Testevaluation und äußere Testgestaltung. Eine wichtige Funktion haben hier auch die kritischen Testrezensionen (z. B. der 25 wichtigsten Verfahren, Kubinger, 1997). Die allgemeine Forderung nach Qualitätssicherung hat zu weiteren Initiativen des Testkuratoriums geführt. Eine schwierige Diskussion ergibt sich, wenn nach der notwendigen Kompetenz für Anwendung, Auswertung und Interpretation von psychologischen Tests gefragt wird. Für welche Schritte dieser Methodik werden qualifizierte Diplom-Psychologen(innen) benötigt, für welche Schritte ist dies nicht erforderlich? Einerseits haben nicht mehr alle Diplomierten eine hinreichende Ausbildung in psychologischer Diagnostik, andererseits sind viele Testverfahren von der Eingabe bis zur Auswertung (z. T. sogar bis zum schriftlichen Befund) computerisiert, d. h. durch Software-Systeme realisiert worden.

Testgüte-Kriterien

Die Qualität eines wissenschaftlich entwickelten Tests ist unter bestimmten Gesichtspunkten zu beurteilen und auf diese Weise von Pseudo-Tests (Spielen, Illustrierten-Tests) abzugrenzen:

Objektivität

Die **Objektivität** meint, daß das Untersuchungsergebnis unabhängig ist von
(1) Untersucher und Untersuchungssituation („Durchführungsobjektivität“),
(2) Registrierung und Auswertung („Auswertungsobjektivität“) und
(3) Interpret („Interpretationsobjektivität“).
Mit anderen Worten: Verschiedene Untersucher kämen zum selben Testergebnis. Durch die standardisierte Durchführung und Auswertung eines Tests soll dessen Objektivität gesichert werden.

Reliabilität

Die Reliabilität (auch „Zuverlässigkeit“) meint die Genauigkeit, mit der ein Test ein reproduzierbares Ergebnis liefert (instrumentelle Präzision), unabhängig davon, ob es dem entspricht, was der Test zu messen vorgibt. Die Objektivität ist eine Voraussetzung für die Reliabilität. Es müssen **verschiedene Reliabilitätsformen** unterschieden werden:
(1) Die Reproduzierbarkeit des Ergebnisses zu verschiedenen Zeitpunkten („Retest-Reliabilität“)
(2) Die Vergleichbarkeit der Ergebnisse von Parallelformen („Paralleltest-Reliabilität“)
(3) Die Präzision des Test als solchem („Splithalf-Reliabilität“, „innere Konsistenz“)

Da bei den unterschiedlichen Durchführungen (einmalige Testdurchführung, Testdurchführungen mit Paralleltest oder zu verschiedenen Zeitpunkten) unterschiedliche Fehler wirksam werden, liefern die verschiedenen Formen der Reliabilitätsbestimmung unterschiedliche Werte.

Die Reliabilität ist eine notwendige, aber nicht hinreichende Bedingung für die Validität des Testverfahrens.

Validität
Die Validität (auch „Gültigkeit“) eines Tests stellt das zentrale Gütekriterium dar: Mißt der Test das Merkmal oder Konstrukt, das er zu messen vorgibt? Für diesen Nachweis ist das Testergebnis mit externen Kriterien zu vergleichen, wobei unterschiedliche Kriterien herangezogen werden können:
(1)überzeugt das Testergebnis auch Experten („**Inhaltsvalidität**“, „content validity“),
(2)korreliert das Testergebnis mit gleichzeitig erhobenen Kriterien („**Übereinstimmungsvalidität**“, „concurrent validity“),
(3)korreliert das Testergebnis mit einem später auftretendem Kriterien („**Vorhersagevalidität**“, „predictive validity“),
(4)stimmt das Testergebnis mit dem theoretischen Anspruch des Tests überein („**Konstruktvalidität**“)?

Bei der Übereinstimmungsvalidität und der Vorhersagevalidität interessieren praktische Aspekte. Das externe Kriterium wirft jedoch bei diesen Validitätsbestimmungen seinerseits Probleme auf: Auch im Hinblick auf das Kriterium ist die Frage nach Zuverlässigkeit und Gültigkeit zu stellen: Validität des Validitätskriteriums? Gleichzeitig ist zu beachten, daß die so bestimmten korrelativen Zusammenhänge von Testergebnis und Kriterium maßgeblich von der jeweils untersuchten Population abhängig sind; damit wird der Gültigkeitsnachweis entsprechend eingeschränkt.

Zur theoretischen Rechtfertigung eines Tests ist eine sorgfältige Überprüfung der Konstruktvalidität erforderlich. Zur **Konstruktvalidierung** sind aus der Theorie Hypothesen über die Zusammenhänge des durch den Test erfaßten Konstrukts mit anderen Verhaltensmanifestationen abzuleiten und empirisch zu überprüfen. Dies ist nur selten der Fall, weil solche theoretisch-deduktiven Vorhersagen aus vielen Gründen schwierig sind. Gelegentlich wird auch eine inhaltlich befriedigend ausgefallene Faktorenanalyse eines neuen Tests als "Konstruktvalidierung" bezeichnet. Dies ist jedoch nur eine interne Analyse und keine kriterienbezogene Konstruktvalidierung im engeren Sinn.

Validität	**Objektivität der**	**Reliabilität**
Inhaltsvalidität	Durchführung	Paralleltest-R
Übereinstimmungsvalidität	Registrierung	Retest-R
Vorhersagevalidität	Auswertung	Halbierungs-R
Konstruktvalidität	Interpretation	Konsistenz

Normierung
Für den praktischen Einsatz des Tests kommt der Forderung nach einer ausreichenden Normierung eine große Bedeutung zu. Da die Beurteilung der Merkmalsausprägung einer Person in der Regel eine relative Aussage in Bezug auf die Merkmalsverteilung in der Population darstellt (z. B. als IQ, Standardwert, Prozentrang oder ähnliches, siehe Abb. 4.1) ist die gründliche Normierung eine wichtige Voraussetzung für die praktische Anwendung eines Tests. Bei der Normierung sind relevante Subpopulationen hinsichtlich Alter, Geschlecht, Schulbildung usw. zu berücksichtigen und gegebenenfalls müssen für die verschiedenen Subpopulationen getrennte Normen erstellt werden.

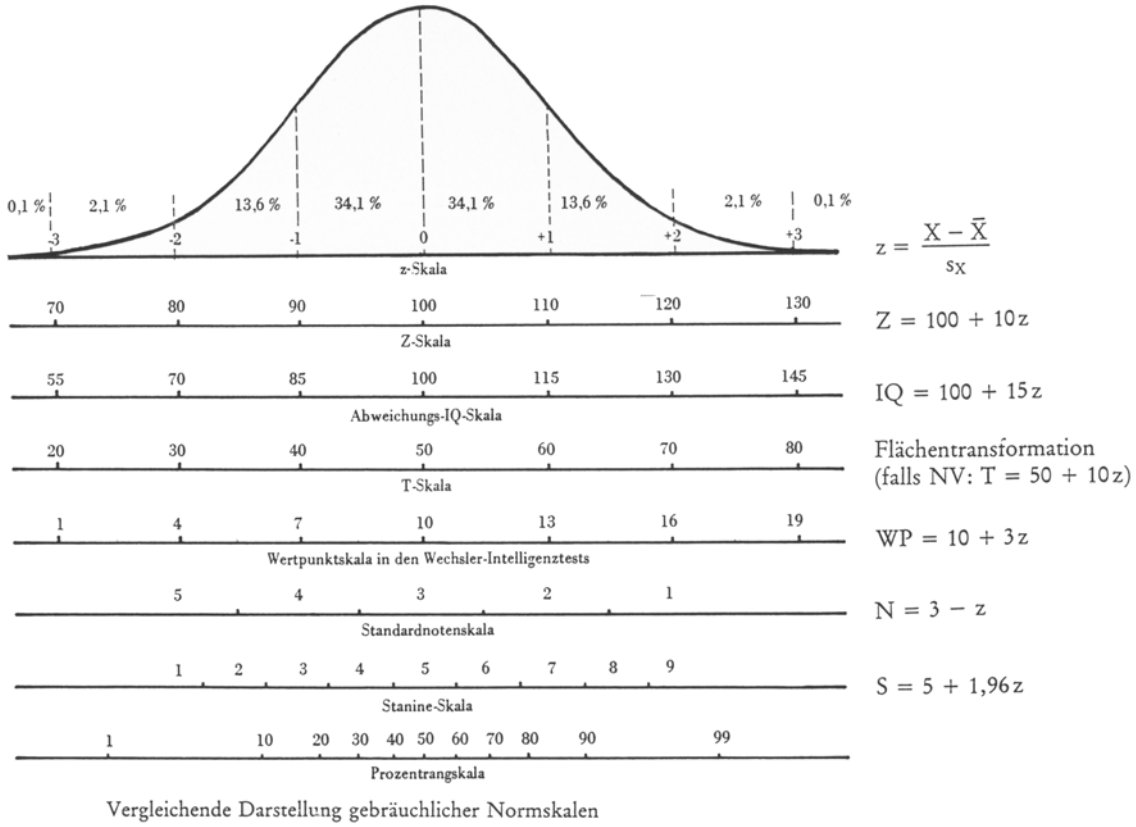


Abbildung 4.1: Vergleichende Darstellung gebräuchlicher Merkmale

Nebengütekriterien
Die sogenannten Nebengütekriterien beinhalten Aspekte, die insbesondere für den praktischen Einsatz der Tests relevant sind, dazu gehören auch Fragen des Zeitaufwands und der Kosten für einen Test (**Testökonomie**), die **Fairness des Tests** (spezielle Gruppen, z. B. Männer/Frauen, ethnische Minderheiten, werden bei der Schätzung der Kriteriumsvalidität nicht systematisch diskriminiert) sowie die Plausibilität (**Augenschein-Validität**), d. h. der von Probanden akzeptierte Sinn der Untersuchung, wichtig.

Manual
Das Manual (Testhandbuch, Testanweisung) sollte alle wichtigen Informationen enthalten, die für eine Beurteilung der Güte des Test notwendig sind. So sollte es Angaben zum Anspruch bzw. zur Zielsetzung des Verfahrens und zu dessen Gütekriterien enthalten, alle notwendigen Angaben für eine standardisierte Durchführung (u.a. Instruktion, Material, Testzeit), sowie alle Angaben zur Testauswertung und zu den Normwerten.

4.3 Testtheorie

Es existieren eine umfangreiche und z. T. mathematisch-statistisch anspruchsvoll ausgearbeitete Theorie psychologischer Tests und methodisch differenzierte Verfahren zur Konstruktion und Bewährungskontrolle von Tests. Viele dieser Prinzipien gelten auch für die anderen Verfahren psychologischer Datenerhebung.

Von den beiden wichtigsten Konzeptionen psychologischer Tests (Meßmodellen), der klassischen Testtheorie und der probabilistischen Testtheorie ("Rasch-Skalierung"), wird hier nur erstere dargestellt und letztere kurz besprochen.

Jedes Meßmodell muß bestimmte Annahmen machen, auch wenn diese nur näherungsweise oder nur unter bestimmten Rahmenbedingungen zutreffen. Grundsätzlich wird die äquivalente Wiederholbarkeit postuliert, d. h. jede Messung ist beliebig wiederholbar – im Bereich der Psychologie sicher eine fragwürdige Annahme, wenn Phänomene wie methodenbedingte Reaktivität, Gewöhnung, Vertrautheit mit der Aufgabe, Lernen, Ausbildung von Strategien und subjektiven Theorien, bedacht werden.

Die Annahmen der klassischen Testtheorie:

Grundlage der klassischen Testtheorie ist die Annahme, daß sich jeder gemessene Wert X_i aus einem „wahren" Wert t_i und einer Fehlerkomponente e_i (u. U. durch verschiedene Fehler) zusammensetzt:

$$X_i = t_i + e$$

Das Reliabilitätskonzept wird durch die **drei Grundannahmen der klassischen Testtheorie** begründet. Diese besagen:

(1) $$\mu_e = 0$$

in Worten: Das Mittel aus allen Fehlereinflüssen ist null [bzw. der Erwartungswert für alle Fehlereinflüsse ist null]

(2) $$\rho_{e1,\,e2} = 0$$

in Worten: Verschiedene Fehler sind unkorreliert

(3) $$\rho_{et} = 0$$

in Worten: Der Fehler ist nicht korreliert mit dem „wahren" Wert

Aus diesen Grundannahmen ergibt sich für die Varianz des Meßwertes X_i, daß diese sich additiv aus den Varianzen von „wahrem" Wert t_i und Fehler e_i zusammensetzt:

$$\sigma_X^2 = \sigma_t^2 + \sigma_e^2$$

Die Reliabilität:

Die Reliabilität r_{tt} (Zuverlässigkeit) eines Tests bzw. einer Messung wird statistisch definiert als das Verhältnis von wahrer Varianz σ_t^2 zu Testvarianz σ_X^2:

$$r_{tt} = \frac{\sigma_t^2}{\sigma_X^2}$$

Die Reliabilität r_{tt} entspricht dem Quadrat der Korrelation von Testwert und wahrem Wert ρ_{Xt}^2. Praktisch ergibt sich die Reliabilität als Korrelation zwischen zwei Testaufnahmen X_{1i} und X_{2i}, d.h. $r_{tt} = r_{X1, X2}$ unter dem Postulat, daß es sich um "Parallelmessungen" handelt.

Methoden der Reliabilitätsbestimmung
Nach dem Modell der klassischen Testtheorie gilt, wie oben festgestellt: $X_i = t_i + e_i$

Es ist davon auszugehen, daß ein Testwert X_i nicht allein durch einen einzigen Fehler beeinflußt wird, sondern durch verschiedene Fehler, z. B. solche, die aus der Meßungenauigkeit des Test selbst stammen und solche, die auf die Vpn (z. B. Motivation, Wachheit), auf die Untersuchungssituation oder den Untersuchungszeitpunkt zurückzuführen sind. Dementsprechend liefern unterschiedliche Formen der Reliabilitätsbestimmung unterschiedliche Werte für r_{tt}. Zusätzlich ist zu bedenken, daß solche Reliabilitätsbestimmungen populationsabhängig sind; von besonderer Bedeutung ist hier z. B. eine Einengung der Streubreite („restriction of range“).

Formen der Reliabilitätsbestimmung (siehe oben):

(1) Retest-Reliabilität
(2) Paralleltest-Reliabilität
(3) Halbierungs-Reliabilität („Splithalf-Reliabilität“)
(4) Innere Konsistenz

Die **Retest-Reliabilität** bestimmt sich aus der Korrelation der Testwerte, die zu zwei unterschiedlichen Zeitpunkten aufgenommen wurden. In sie fließen Fehler aufgrund der Befindlichkeit der Versuchsperson, der Versuchssituation und des Zeitpunkts ein. Darüber hinaus werden systematische Fehlereinflüsse wirksam, die nicht der Zuverlässigkeit des Tests anzulasten sind: dazu gehören Erinnerungseffekte (an das Testmaterial, die Items) und insbesondere Veränderungen des Merkmals. Insofern kann für zeitlich instabile Merkmale (z. B. Zustands-Merkmale wie die Stimmung) keine Retest-Reliabilität bestimmt werden.

Die **Paralleltest-Reliabilität** läßt sich dann bestimmen, wenn von einem Testverfahren mehrere Parallelformen vorliegen. Die Paralleltest-Reliabilität entspricht der Korrelation der Testwerte der Parallelformen. In diese Reliabilitätsbestimmung fließen die Ungenauigkeiten der einzelnen Tests, die fehlende Äquivalenz der Testformen und, falls die Paralletests zu unterschiedlichen Zeitpunkten aufgenommen wurden, die gleichen Fehlerfaktoren wie bei der Retest-Reliabilität ein.

Zur Bestimmung der **Halbierungs-Reliabilität** („Splithalf-Reliabilität“) wird der Itempool des Tests in zwei Hälften aufgeteilt, und die aus den Testhälften berechneten Testwerte werden korreliert. Die Reliabilität des gesamten Tests wird nach einer Korrekturformel (Spearman-Brown Formel) aus der Korrelation der Testhälften geschätzt. In der Regel wird eine gerade/ungerade Aufteilung der Items („odd-even“) für die Erstellung der Testhälften vorgenommen.

Die **innere Konsistenz** eines Tests ergibt sich aus der Äquivalenz der einzelnen Items eines Tests. Wenn die Items das gleiche Merkmal messen und die Spezifität der Items (im faktorenanalytischen Sinn) gering ist, d. h. die Items hohe Interkorrelationen aufweisen, hat der Test eine hohe Meßgenauigkeit. Das Maß der inneren Konsistenz wird meist durch den Koeffizienten α (sog. Cronbach α) angegeben. Es besteht eine enge Beziehung zwischen der inneren Konsistenz eines Test und der Halbierungs-Reliabilität.

Standardmeßfehler
Die Zuverlässigkeit einer Messung hat eine große praktische Bedeutung: Sie bestimmt letztlich, wie stark die gemessenen Werte um den vermuteten wahren Wert streuen. Das Reliabilitätsmaß erlaubt insofern für den Einzelfall anzugeben, in welchen Intervall (Konfidenzintervall) der wahre Wert liegen wird. Die fehlerbedingte Streuung verschiedener Messungen bei einem Individuum wird als **Standardmeßfehler** bezeichnet.

Die Reliabilität r_{tt} gibt den Anteil der zuverlässigen Varianz an (siehe oben) und der Ausdruck $(1 - r_{tt})$ dementsprechend den Anteil der Fehlervarianz an der Messung. Der Standardmeßfehler ergibt sich somit aus der Formel:

$$s_e = s_X * \sqrt{1 - r_{tt}}$$

Beispiel: IQ-Test ($\mu_{IQ} = 100$, $\sigma_{IQ} = 10$, $r_{tt} = .84$); der individuelle Testwert sei 120. Der Standardfehler ergibt sich:

$$s_e = s_X * \sqrt{1 - r_{tt}} = 10 * \sqrt{1 - .84} = 4{,}0$$

Mit diesem Standardmeßfehler kann nun das Vertrauensintervall berechnet werden, indem mit einer Wahrscheinlichkeit von z.B. 95% der tatsächliche Meßwert für die Person mit dem IQ 120 liegen wird:

$$X_i - z_{95\%} * s_e \leq \text{wahrer Wert} \leq X_i + z_{95\%} * s_e$$
$$120 - 1.96 * 4.0 \leq \text{wahrer Wert} \leq 120 + 1.96 * 4.0$$
$$112 \leq \text{wahrer Wert} \leq 128 \quad \text{(gerundet)}$$

Der Standardmeßfehler bietet auch die Grundlage für den Vergleich verschiedener Einzelleistungen einer Person (sog. kritischen Differenzen) oder den Vergleich zwischen den Leistungen verschiedener Personen sowie für die Bestimmung der Reliabilität von Differenzen zweier wiederholter Messungen (Differenzwerte).

Die **Reliabilität** stellt nur **einen** wichtigen Aspekt der Test- und Meßmethodik dar; in den Lehrbüchern sind die Darstellungen der Reliabilitätsaspekte häufig überwertig im Vergleich zu den Validitätsaspekten. Validitätsfragen sind zwar deutlich schwieriger zu behandeln, doch sind diese **Validitätsnachweise theoretisch und praktisch wichtiger** für die diagnostische Nützlichkeit eines Tests. Die operationale Definition eines psychologischen Merkmals durch einen Test kann nur bei einer entsprechenden Validitätsüberprüfung als gesichert gelten.

Reliabilität und Validität
Die Reliabilität ist eine notwendige, aber nicht hinreichende Bedingung für die Validität eines Tests. Es gilt, daß die erklärbare Varianz (d.h. das Quadrat der Validität r_{xy}^2) nicht größer sein kann als die zuverlässige Varianz (Reliabilität r_{tt}). Es gilt also:

$$r_{xy}^2 \leq r_{tt}$$

r_{xy} = Validitätskoeffizient r_{tt} = Reliabilitätskoeffizient X = Testwert Y = Kriterium

Der Zusammenhang von Reliabilität und Validität gestaltet sich im Hinblick auf die Vorhersage von konkreten Kriterien (im Rahmen der Übereinstimmungsvalidität oder Vorhersagevalidität, z. B. Schul- oder Berufserfolg) deutlich komplexer. Dadurch, daß bei der Erfassung der meisten praktischen Kriterien eine Fülle von unterschiedlichen Variablen wirksam werden (z. B. verschiedene Intelligenzkomponenten, Motivationsaspekte, Persönlichkeitsmerkmale, Merkmale des Sozialverhaltens), lassen sich durch einen Test in der Regel nur eine oder wenige Facetten des Kriteriums vorhersagen. Dieser Tatbestand kann zu einer paradoxen Bezie-

hung von Reliabilität und Validität führen (siehe Abbildung 4.2): Ein homogener und damit reliabler Test kann nur eine Facette des Kriteriums vorhersagen, der Validitätskoeffizient ist demnach gering; ein heterogener und damit wenig reliabler Test erfaßt u. U. mehrere Facetten des Kriteriums und erbringt damit einen höheren Validitätskoeffizienten.

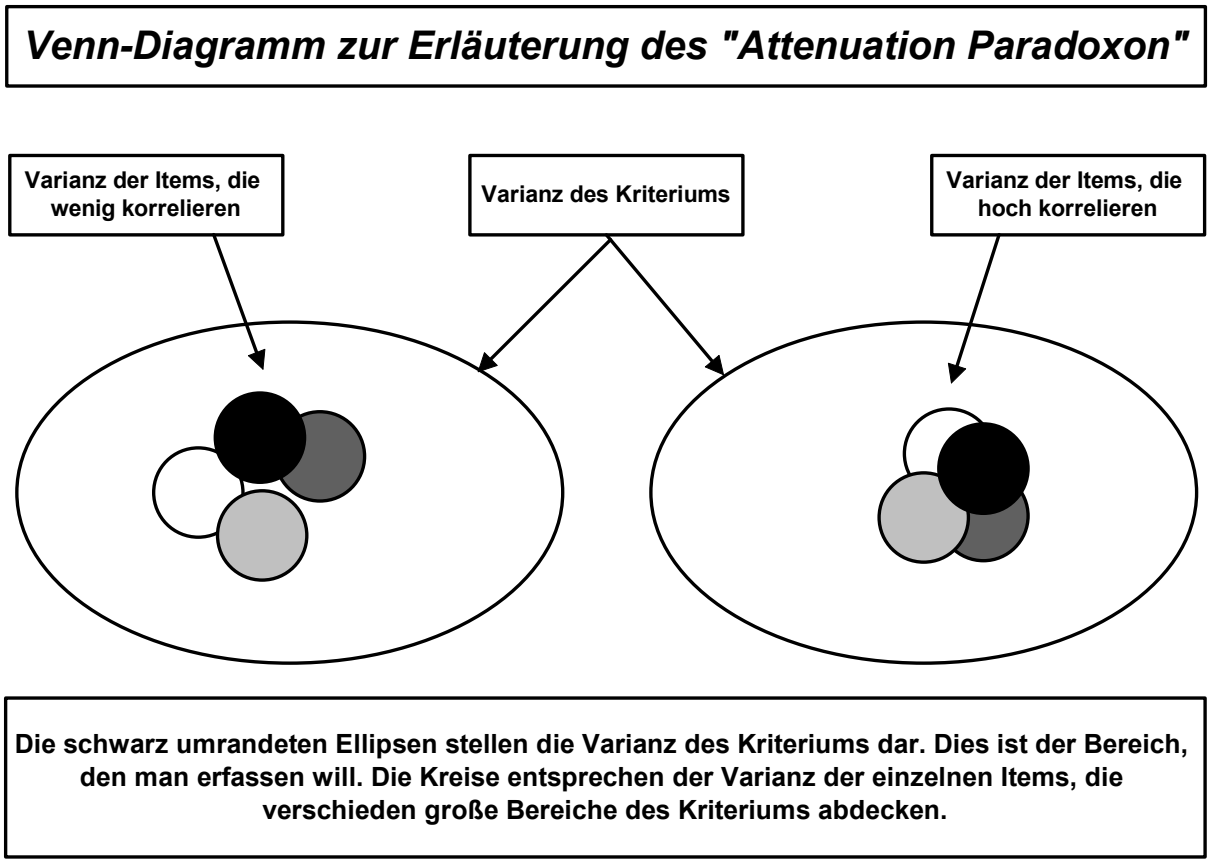


Abbildung 4.2: Paradoxe Beziehung zwischen Konsistenz-Reliabilität und Validität

Die optimale Lösung dieses Problems besteht darin, das Kriterium über eine multiple Regression mit mehreren homogenen Tests vorherzusagen, wobei diese Tests jeweils unterschiedliche Facetten des Kriteriums vorhersagen („inkrementelle Validität“).

4.4 Testkonstruktion

Die Testkonstruktion geht gewöhnlich in den folgenden Schritten vor sich:

Testplanung	Zielsetzung, Vergleich mit bisherigen Tests, Festlegung des gewünschten Geltungs- und Gültigkeitsbereichs;
Testentwurf	Entwicklung der Items und Instruktionen;
Testdurchführung	Datenerhebung in einer Stichprobe der Ziel-Population (Geltungsbereich);
Itemanalyse	Berechnung von Itemschwierigkeiten, Trennschärfeindizes, Verteilung der Testwerte, Reliabilitätskoeffizienten, Validitätskoeffizienten u. a. Statistiken (z. B. Faktorenanalysen des Itempools) zwecks Itemselektion und Revision des Testentwurfs;
Testendform	Normierung der revidierten Testform an einer für den Geltungsbereich repräsentativen Stichprobe und Erhebung weiterer Validitätshinweise für den behaupteten Gültigkeitsbereich der Tests;
Publikation	Testanweisung (Manual) und Testmaterial.

Besonders wichtige Schritte sind hier (1) die psychologisch gut begründete, sprachlich bzw. inhaltlich genaue Formulierung der Items (Aufgaben) und der Instruktion, (2) die empirisch begründete Revision des Testentwurfs aufgrund von Itemanalysen, und (3) die Validierungsstudien, welche im Prinzip nie abschließbar sind, sondern dem Erfolgs- und Forschungsstand folgen müssen. Nur die Itemanalyse soll hier noch angesprochen werden.

Itemanalyse und Itemselektion
Nach dem Entwurf der Items ist deren Brauchbarkeit für den endgültigen Test empirisch zu überprüfen. Dazu werden die Items zu einer Testvorform zusammengestellt, die einer entsprechenden Stichprobe zur Beantwortung vorgelegt wird.
Die Itemanalyse anhand derer über die Eignung eines Items beurteilt wird, kann nach unterschiedlichen methodischen Verfahren erfolgen:
(1) Klassische Itemanalyse und Itemselektion nach Trennschärfe und Itemschwierigkeit
(2) Faktorenanalyse

Grundlage der **Itemanalyse** sind die Kennwerte von Itemschwierigkeit und Trennschärfe:
Die **Itemschwierigkeit** ist bei dichotomen Items (z. B. richtig/falsch Items) gleich dem prozentualen Anteil der richtigen Lösungen; bei mehrstufigen Antwortmöglichkeiten – je nach Polung – gleich dem Mittelwert aus den Punktwerten der Itemantworten.

Die **Trennschärfe** eines Items wird über die Korrelation der Itemantworten mit dem Gesamttestwert (als Summe aller übrigen Itemwerte, jedoch ohne das betreffende Item, sog. „part-whole“ Korrektur) bestimmt. Brauchbare Items sollten eine mittlere Schwierigkeit und eine hohe Trennschärfe aufweisen. Mit einer Auswahl von Items mit hoher Trennschärfe ist auch die innere Konsistenz des künftigen Tests gesichert.

Die Itemselektion mittels **Faktorenanalyse** dient der Konstruktion von möglichst homogenen Test(-Skalen). Hier wird mit den vorliegenden Daten eine Faktorenextraktion (in der Regel nach der Hauptkomponentenanalyse) durchgeführt. Die Notwenigkeit einer Faktorenrotation entscheidet sich nach der Absicht der Konstruktion. Ausgewählt werden Items mit hohen Faktorenladungen auf einem Faktor und möglichst geringen Ladungen auf Nebenfaktoren. Es besteht ein enger Zusammenhang von Faktorenladung und Trennschärfe des Items. An der Ladungshöhe und Kommunalität sind u. U. besonders geeignete „Markier“-Items des betreffenden Faktors zu erkennen. Bei relativ heterogenen Items ist eine faktorenanalytisch gestützte Strukturierung in mehrere Skalen (Dimensionen) vorzunehmen.

Die Eignung eines Items zur Messung des Konstrukts ist anhand von statistischen Kennwerten, natürlich unter Berücksichtigung der inhaltlichen Gesichtspunkte (Konstrukt-Explikation), zu beurteilen. Die Testkonstruktion ist ein Optimierungsprozeß, wobei u. U. in mehreren Revisionsschritten Kompromisse zwischen teils zusammenhängenden, teils widersprüchlichen Maßstäben (siehe oben) erreicht werden müssen.

Bei der Itemselektion sind die folgenden Ziele zu beachten:
- Die Testwerte sollten, um gute Differenzierungen zu leisten, möglichst normal verteilt sein, aber auch in den vorkommenden Extrembereichen noch diskriminieren.
- Der Test sollte möglichst kurz und reliabel sein, jedoch nicht auf Kosten der empirisch erreichbaren Validität und des Entscheidungsnutzens hinsichtlich wichtiger Kriterien in der Praxis.

In anschließenden Validierungsstudien sollte plausibel aufgezeigt werden, daß erhaltene Testwerte mit externen Kriterien (gegenwärtigen oder künftigen) in einem theoretisch und praktisch produktiven Zusammenhang stehen. Aus der formalen Testkonstruktion, d.h. allein auf-

grund der internen Itemanalyse, kann im Prinzip ein sehr reliabler Test entstehen, welcher praktisch unbrauchbar und überflüssig ist.

4.5 Probabilistische Testtheorie

Aus meßtheoretischer Sicht ist die klassische Testtheorie in mehrfacher Hinsicht unbefriedigend. Die Teststatistiken (Trennschärfe, Reliabilität usw.) beziehen sich als Korrelationskoeffizienten immer auf bestimmte (Sub-)Populationen – als ob sozusagen die Länge des Metermaßes jeweils von der Gruppe der untersuchten Objekte beeinflußt würde. Die Annahme unkorrelierter Fehler ist unrealistisch. Es ist schwierig, die meßmethodisch – für das Konzept paralleler Messungen – wichtige Homogenität von Items klar zu definieren.

Aus der Kritik an der älteren („klassischen") Testtheorie wurden neue Ansätze entwickelt, die meist als probabilistische Testtheorien zusammengefaßt werden. Diese Bezeichnung rührt daher, daß in dieser Konzeption nicht die direkte Messung eines Merkmals behauptet wird, sondern nur Aussagen über die **Auftretenswahrscheinlichkeit** einer bestimmten Testantwort gemacht werden. Eine beobachtete Testantwort (Itemwert) ist lediglich Indikator für die latente Eigenschaft (latent trait), deren Messung der Test dient.

Am bekanntesten ist das Meßmodell von Rasch bzw. die sog. **Rasch-Skalierung**. Ein wichtiges Konzept hierbei ist die Item-Charakteristik-Kurve (ICC), welche in einer Funktion (wie eine Kennlinie) den Zusammenhang zwischen latenter und beobachteter Variable, d. h. Fähigkeit und Itemwert, erfaßt. Der beobachtete Wert eines „idealen" Items sollte (1) mit steigender Fähigkeit zunehmen, (2) diese Funktion sollte monoton verlaufen, aber (3) nicht linear, damit verschiedene Ausprägungen der Fähigkeit mit guter Differenzierung (logistische Funktion) – unterschieden werden können. Die Rasch-Skalierung dient der Prüfung der Items und der Selektion möglichst geeigneter (rasch-homogener) Items, welche diesen Anforderungen nahe kommen.

Auf Modellannahmen, Modelltests und spezielle Aspekte dieses Modells wird hier nicht eingegangen. Es gibt bisher nur sehr wenige Tests, welche nach diesem Meßmodell konstruiert wurden. Offensichtlich werden dafür größere Personenzahlen und inhaltlich sowie formal verhältnismäßig homogene Items benötigt. Dies ist am ehesten bei eng definierten Fähigkeitstests, jedoch nicht bei Persönlichkeits-Fragebogen u. a. Tests zu erwarten. Aus theoretischer Sicht haben probabilistische Konzeptionen wesentliche Vorzüge, in der Praxis dominieren die meßtheoretisch schlechter fundierten, „klassisch" konstruierten Tests.

4.6 Assessment

Methodisch stehen die Testtheorie und Testkonstruktion in dem größeren Rahmen der allgemeinen Methodenlehre der Psychologie; sie bilden einen wesentlichen Teil des Assessment von Personen. In Abbildung 4.3 ist der diagnostische Entscheidungsprozeß dargestellt.

Diagnose (griechisch „hindurch erkennen") heißt in der Medizin Erkennen und Benennen der Krankheit aufgrund von Symptomen; Differentialdiagnose meint die Unterscheidung ähnlicher Krankheitsbilder. Psychologische Diagnostik (Psychodiagnostik) hat eine viel breitere und unschärfere Bedeutung: Anwendung von psychologischen Methoden zur Beschreibung individueller Unterschiede, z. B. Charakterisierung und Begutachtung von Personen. Da das Erkennen von Krankheiten (d. h. die Zuordnung von Individuen zu nosologischen Einheiten, um aus dem akkumulierten Wissen über Ätiologie, typischen Verlauf und wirksame Therapien, die optimale Behandlung abzuleiten) in den vielen Praxisfeldern der Psychologie relativ selten ist oder nicht vorkommt, kann die medizinische Terminologie stören. In der Differentiellen Psychologie und Persönlichkeitsforschung wird zunehmend der Begriff „**Assessment**" verwendet, inzwischen ist auch der Begriff „**Assessment Center**" in der Personalpsychologie (A&O Bereich) geläufig. Assessment vermeidet die Konnotationen von „Diagnostik" (bzw.

Krankheit) und von „Messung“ und wird oft in einer prägnanteren Bedeutung als Psychodiagnostik verwendet.

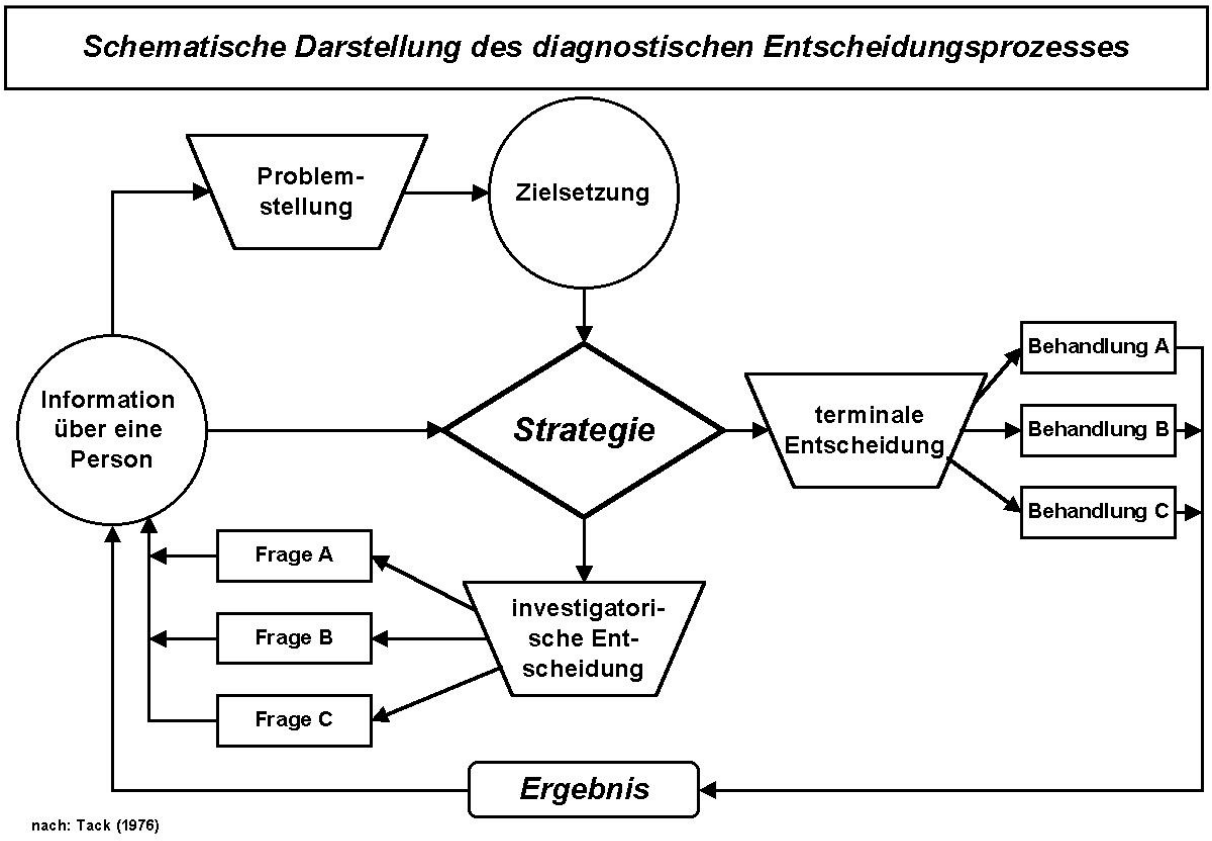


Abbildung 4.3: Der diagnostische Entscheidungsprozeß

Ein typisches Assessmentproblem lautet: Soll die begrenzte Zahl von Studienplätzen in der Medizin aufgrund von Abiturnoten, Zufallsprinzip, Eignungstests oder Bewerbungsgesprächen vergeben werden? Welche Prädiktoren und Kriterien führen hier zu einer möglichst rationalen Entscheidung? Psychologisches Assessment soll Gruppierung, Auswahl und Zuweisung von Personen ermöglichen, beruflich zu bestimmten Aufgaben, pädagogisch zu bestimmten Programmen oder klinisch zu bestimmten Behandlungen (siehe Vorlesung Differentielle Psychologie und Persönlichkeitsforschung im 3. Fachsemester).

Assessment ist die Erfassung von psychologischen Merkmalen nach bestimmten methodischen Prinzipien zu einem praktischen Zweck, welcher eine rationale Entscheidung verlangt. Oft handelt es sich um **Prädiktoren-Kriterien-Beziehungen**, wobei auch der **Aufwand und der Entscheidungsnutzen** dieser Urteilsprozesse bewertet werden. Das bis heute unübertroffene Lehrbuch stammt von J. S. Wiggins (1973): Personality and prediction. – Principles of personality assessment. Carver und Scheier (1988) haben ihr Lehrbuch der Persönlichkeitspsychologie so angelegt, daß die enge theoretische und methodische Beziehung zwischen Persönlichkeitstheorie und spezieller Methodik (z. B. Murray/TAT, Kelly/REP, Eysenck/EPQ) deutlich werden: **Persönlichkeitstheorie als Assessmenttheorie** und **Assessmentstrategie als empirische Interpretation (Operationalisierung) der Persönlichkeitstheorie**. Dieses moderne Denken in Assessmentstratetegien ist leider auch in neueren deutschen Lehrbüchern der Differentiellen Psychologie und Persönlichkeitsforschung kaum repräsentiert.

Assessmentstrategien sind Pläne, die festlegen, welches Konstrukt mit welchem Untersuchungs- und Auswertungs-Konzept erfaßt werden soll. Im weiteren Sinn gehören zum Assessment auch die Organisation (z. B. Assessment Center) und vor allem die empirische Evaluation (Alternativstrategien, Aufwand, Entscheidungsnutzen, Akzeptanz?). Die Assessmenttheorie, welche die Assessmentstrategien und die anderen methodischen Prinzipien systematisch verbindet, ist die Brücke zwischen der differentiellen Psychologie (und der Persönlichkeitsforschung) und der angewandten Psychologie in Praxisfeldern wie Schule und Weiterbildung, A&O, Klinische und Gesundheitspsychologie. Viele Inhalte der Assessmenttheorie werden, ebenso wie ausgewählte Untersuchungsmethoden und Tests im Fach „Psychologische Diagnostik", im Hauptstudium sowie in den speziellen Praxisschwerpunkten unterrichtet (siehe auch Übung „Assessmentstrategien und Testkonstruktion"). Andererseits werden sowohl in der Übung „Versuchsplanung" wie auch in der Vorlesung „Differentielle Psychologie und Persönlichkeitsforschung" Prinzipien und ausgewählte Beispiele behandelt, um diese Grundlegung für die praktische Psychologie herauszuarbeiten.

4.7 Vorhersage

Die Vorhersage des individuellen Verhaltens ist als eine zentrale Aufgabe der differentiellen Psychologie anzusehen, und das psychologische Assessment hat häufig solche direkten oder indirekten Prädiktionen zum Ziel. Es wäre ein Mißverständnis, eine deterministische Auffassung zu unterstellen, das menschliche Verhalten sei kausal eng bestimmt und festgelegt. Angesichts der multiplen Bedingungen des Verhaltens („Kausalnetze") kann die Vorhersage nur **probabilistisch** sein, d. h. als **Wahrscheinlichkeit des Auftretens des betreffenden Verhaltens** („Erwartungswahrscheinlichkeit"). Grundsätzlich an der – partiellen und bedingten – Vorhersagbarkeit des Verhaltens zu zweifeln, würde weitgehend auf eine Negation der empirischen Psychologie und der Berufspraxis hinauslaufen bzw. diese einem Versuch- und Irrtums-Prinzip oder dem Zufallsprinzip unterstellen.

Verhaltensvorhersage durch psychologisches Assessment ist zweifellos extrem schwierig. Zum Vergleich sei an die Unsicherheiten der Wetterprognosen trotz größtem meteorologischen Forschungsaufwand und umfangreicher Datenerhebung erinnert. – Wieviel anspruchsvoller ist die psychologische Vorhersage! Implizit ist die Annahme, daß Verhalten vorhersagbar ist, in vielen Anwendungsfeldern und bei allen Mittel-Ziel-Analysen und Interventionen vorhanden: diese Maßnahme wird wahrscheinlich jenes Ergebnis (Veränderung, Entwicklung, Verbesserung der Symptomatik usw.) haben.

Die Beziehungen zwischen Prädiktoren und Kriterien sind in der Regel als **Regressions-Korrelations-Problem** zu formulieren. Außer den speziellen statistischen Aspekten sind hier mehrere Prinzipien zur Optimierung der Vorhersage zu berücksichtigen (siehe auch Amelang & Zielinski, 1994), u. a.

- **Symmetrie-Beziehungen** (repräsentatives Design),
- **Multimodale Diagnostik** (u. U. auch Multitrait-Multimethod-Kontrollen),
- **Aggregation** (über Zeitpunkte, Items, u. U. mehrdimensionale Aggregate),
- **Dimensionierung** (Faktorenanalyse zur Datenreduktion),
- **Dekompensation** (Zerlegung von heterogenen Varianzen),
- **Suppression** (Bindung irrelevanter Prädiktorenvarianz, d. h. Unterdrückung unerwünschter Varianzanteile, z. B. Zusammenstellung positiv und negativ alterskorrelierter Items, falls das Alter kontrolliert werden soll),
- **Moderation** (Berücksichtigung einer Kovariablen, welche nicht-lineare Prädiktor-Kriterien-Beziehungen bedingt, z. B. differentielle Vorhersagbarkeit von Subgruppen).

Meßtheorie und Skalenprobleme spielen hier oft eine geringere Rolle, denn im Sinne des „index measurement“ wird pragmatisch eine Maximierung der Effektstärken der Vorhersage bzw. des multiplen R^2 angestrebt.

Gefährdungen dieser Regressions-Korrelations-Konzepte ergeben sich vor allem aus der **Kapitalisiering des** Zufalls (capitalization of chance) und aus **mangelnder Validität des Kriteriums**.

Trotz der Empfehlung der statistischen Minderungskorrektur des R^2 und der Daumenregel (für jeden weiteren Prädiktor weitere 30 Vpn) bzw. trotz der Möglichkeit gründlicher Poweranalysen, werden häufig Regressions-„Erfolge“ mitgeteilt, die durch zahlreiche Prädiktoren und durch Probieren erzielt wurden. Kritische Gegenmaßnahmen sind: unabhängige Replikation der gesamten Untersuchung (Kreuzvalidierung), zumindest eine Kontrolle durch Vergleich der zufallshalbierten Teil-Stichproben, um die **Robustheit der Regression** zu prüfen (präzisere Techniken sind hier die sog. Jackknife-Technik und die sog. Bootstrap-Methode).

In den Anwendungsfeldern wird es oft besonders schwierig sein, das Kriterium eindeutig zu definieren: das Kriterium einer richtigen Personalempfehlung oder Schulberatung, der Therapie- oder Rehabilitations-Erfolg usw. Eine gelegentlich zu hörende Forderung lautet, auf die **Abgrenzung, Anreicherung und Operationalisierung des Kriteriums** ebenso viel Mühe zu verwenden wie auf die Prädiktoren. Hierbei sind oft Auswahlentscheidungen und Bewertungen erforderlich (Erfolg der Rehabilitation?), welche nicht nur das Wissen von Experten, sondern auch die Bewertungen durch die Betroffenen, durch die Institutionen und die Finanzierungsstellen verlangen. Es gibt für diesen Zweck Modelle der **sozialen Urteilsbildung**, um den Prozeß von Information, Evaluation und Kompromißfindung transparent zu machen.

4.8 Entscheidungsfehler und Entscheidungsnutzen

Nur ein Aspekt der Assessmentmethodik ist hier hervorzuheben. Das psychologische Assessment soll eine rationale, empirisch begründete Entscheidung ermöglichen. Deshalb ist immer nach dem **Entscheidungsnutzen** zu fragen. Wieviel besser ist eine Entscheidung aufgrund eines Assessments (Tests, Interview usw.) im Vergleich zu einer Entscheidung ohne solche Hilfen, d.h. nach Zufallsprinzip oder subjektiver "Intuition"?

Schema für Klassifikation (Vorhersage) und Entscheidungsnutzen

		Zuordnung aufgrund des Prädiktors	
		richtige Diagnose krank	falsche Diagnose gesund
tatsächliche Zuordnung	krank	wahre Positive	falsche Negative
	gesund	falsche Positive	wahre Negative

Betrachtung der Risiken der Urteilsbildung und Analyse des Entscheidungsnutzens einer diagnostischen Maßnahme (im Vergleich zu den Kosten) nach Cronbach

Sensitivität der Zuordnung:
Anteil der richtig diagnostizierten Kranken in der Gruppe der Kranken

Spezifität der Zuordnung:
Anteil der richtig diagnostizierten Gesunden in der Gruppe der Gesunden

Prädiktiver Wert einer positiven Zuordnung:
Anteil der richtig diagnostizierten Kranken an allen als krank diagnostizierten Personen

Prädiktiver Wert einer negativen Zuordnung:
Anteil der richtig diagnostizierten Gesunden in der Gruppe der als gesund diagnostizierten Personen

Abbildung 4.4: Vorhersage und Entscheidungsfehler

Entscheidungen und Entscheidungsfehler. Grundlegend ist das Vierfelder-Schema, das bei vielen Entscheidungen in Medizin und Psychologie, bei Diagnose und bei Behandlungen und Maßnahmen gilt: Zuordnung aufgrund (1) des Prädiktors und (2) tatsächlicher Zugehörigkeit. Möglich wären bei der Diagnose krank/gesund (mit den jeweiligen **Grundraten** als Erwartungswert): wahre positive, wahre negative, falsche positive und falsche negative Zuordnung. Dieses Schema definiert auch die Risiken und die **Sensitivität** der Zuordnungsregel (Anteil der richtig diagnostizierten Kranken in der Gruppe der Kranken) und die **Spezifität** der Zuordnungsregel (Anteil der richtig diagnostizierten Gesunden in der Gruppe der Gesunden). Die Begriffe sind in der Medizin üblich, außerdem die Begriffe „**falsche Positive**“ und „**falsche Negative**“.

Die **Sensitivität** eines medizinischen Tests bezieht sich also auf die Fähigkeit Personen mit der fraglichen Krankheit vollständig herauszufiltern; die **Spezifität** dagegen die Fähigkeit, **ausschließlich** Personen mit der fraglichen Krankheit zu erfassen. Beide stehen meist in einem umgekehrten Verhältnis zueinander, d. h. je spezifischer ein Test ist, desto **un**vollständiger ist die Erfassung und umgekehrt.

Entscheidungsnutzen. In der Personalpsychologie und in anderen Bereichen praktischer Psychologie stellt sich die nüchterne Frage, ob sich die Kosten für psychologische Untersuchungen lohnen – im Vergleich zur psychologielosen Praxis. So war die zeitweilig sehr deutliche Abnahme der Testpsychologie eine Folge der pragmatischen Überlegungen, wofür nutzt die Information über den IQ mehr als das Wissen über den Schulabschluß? Was nutzt ein Persönlichkeitsgutachten, wenn es um die konkrete Therapie eines Verhaltensproblems geht? Diesen kritischen Fragen kann nur durch empirischen Nachweis aufgrund von Evaluationen (Bewährungskontrollen) begegnet werden.

Das Konzept des Entscheidungsnutzens ist u. a. von Cronbach & Gleser (1965) im Rahmen der Zuordnungs- und Klassifikationsstrategien (diagnostische Urteilsbildung) entwickelt worden. Der Gesamtnutzen ergibt sich aus der Differenz des Nutzens von Einzelkomponenten (Informationsnutzen, Behandlungsnutzen, Ergebnisnutzen) und dem Kostenaufwand. **Nutzen** ist hier oft materiell, aber nicht ausschließlich materiell, sondern in „equal units of satisfaction" gemeint, d. h. auch in komplexen **Nutzenschätzungen durch Experten und Betroffene** formulierbar. Das Entscheidungskalkül ist oft weniger rational als es zunächst aussieht, da die symmetrisch notwendige **Kalkulation des Schadens** kaum gelingen wird. Der Nutzen und der Schaden dieser diagnostischen Entscheidungen sind zwar real, aber nicht ohne weiteres quantitativ zu bemessen. Eine faire Bewertung würde hier einen Prozeß sozialer Urteilsbildung aller Beteiligten erfordern.

Die Vielfalt der methodischen Prinzipien wurde hier nur einführend skizziert, um deutlich zu machen, wie viel Erfahrung und Kompetenz heute für die Entwicklung guter Assessmentstrategien erforderlich sind.

Weiterführende Literatur

Amelang, M. & Zielinski, W. (1994). *Psychologische Diagnostik und Intervention*. Berlin: Springer.

Brickenkamp, R. (1997). *Handbuch psychologischer und pädagogischer Tests*. (2. Aufl.) Göttingen: Hogrefe.

Föderation Deutscher Psychologenvereinigungen (1986). Testkuratorium. *Psychologische Rundschau*, **37**, 162-165.

Häcker, H., Leutner, D. & Amelang, M. (Hrsg.) (1998). *Standards für pädagogisches und psychologisches Testen*. Göttingen: Hogrefe, Bern: Huber.

Krauth, J. (1995). *Testkonstruktion und Testtheorie*. Weinheim: Beltz.

Kubinger, K.D. (1997). Editorial zum Themenheft „Testrezensionen: 25 einschlägige Verfahren". *Zeitschrift für Differentielle und Diagnostische Psychologie*, **18**, 1-3.

Lienert, G.A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. (6. Aufl.). Weinheim: Psychologische Verlags-Union.

Michel, L. & Conrad, W. (1982). Theoretische Grundlagen psychometrischer Tests. In: K.J. Groffmann & L. Michel (Hrsg.) *Enzyklopädie der Psychologie. Themenbereich B. Methodologie und Methoden. Serie II. Psychologische Diagnostik. Band 1.Grundlagen psychologischer Diagnostik (S. 1-129)*. Göttingen: Hogrefe.

Schmid, H. (1992). *Psychologische Tests: Theorie und Konstruktion*. Bern: Huber.

Schorr, A. (1995). Stand und Perspektiven diagnostischer Verfahren in der Praxis. Ergebnisse einer repräsentativen Befragung westdeutscher Psychologen. *Diagnostica*, **41**, 3-20.

Steck, P. (1997). Psychologische Testverfahren in der Praxis. *Diagnostica*, **43**, 267-284.

Westhoff, G. (Hrsg.) (1993). *Handbuch psychosozialer Meßinstrumente: ein Kompendium für epidemiologische und klinische Forschung zu chronischer Krankheit*. Göttingen: Hogrefe.

Wiggins, J.S. (1973). *Personality and prediction: Principles of personality assessment.* Reading; MA.: Addison-Wesley.

Wittmann, W.W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J.R. Nesselroade & R.B. Cattell (Eds.) *Handbook of multivariate experimental psychology* (pp. 505-560). New York: Plenum.

5. Differentiell- und sozialpsychologische Aspekte. Validität von Untersuchungen. Evaluation, Metaanalyse, Kommunikation. Wissenschaftliche und ethische Qualitätsmerkmale des Forschungsprozesses.

5.1 Differentiell- und sozialpsychologische Aspekte

Humanpsychologische Experimente unterscheiden sich von tierexperimentellen bzw. naturwissenschaftlichen Arbeiten u. a. durch:

(1) Teilnahme-Motivation (Freiwilligkeit, Verweigerung)
(2) Bewußtseinsprozesse der Vp und Informationsverarbeitung, spezielle Motive bzw. Einstellungen,
(3) soziale Interaktion Vp/Vl (Kommunikation, Rollenübernahme)
(4) spezielle ethische Fragen bzw. Konventionen über Versuche am Menschen.

Ein einfaches S-R-Modell ist inadäquat. Statt der einfachen Abfolge

Erkennen der Instruktion ↘
Reaktionen
Erkennen der Reizbedingungen ↗

sind mehrere differentiell- und sozialpsychologische Aspekte des experimentellen Geschehens hervorzuheben. Es sind wesentliche Kennzeichen humanpsychologischer Experimente, die je nach Perspektive als zu kontrollierende Störeffekte oder als – an sich – interessante psychologische Phänomene zu sehen sind. Beiträge stammen u.a. von Lewin, Rosenthal, Orne, Holzkamp, Campbell & Stanley. Die folgende Übersicht stützt sich u.a. auf Gniech (1976) und Maschewsky (1977).

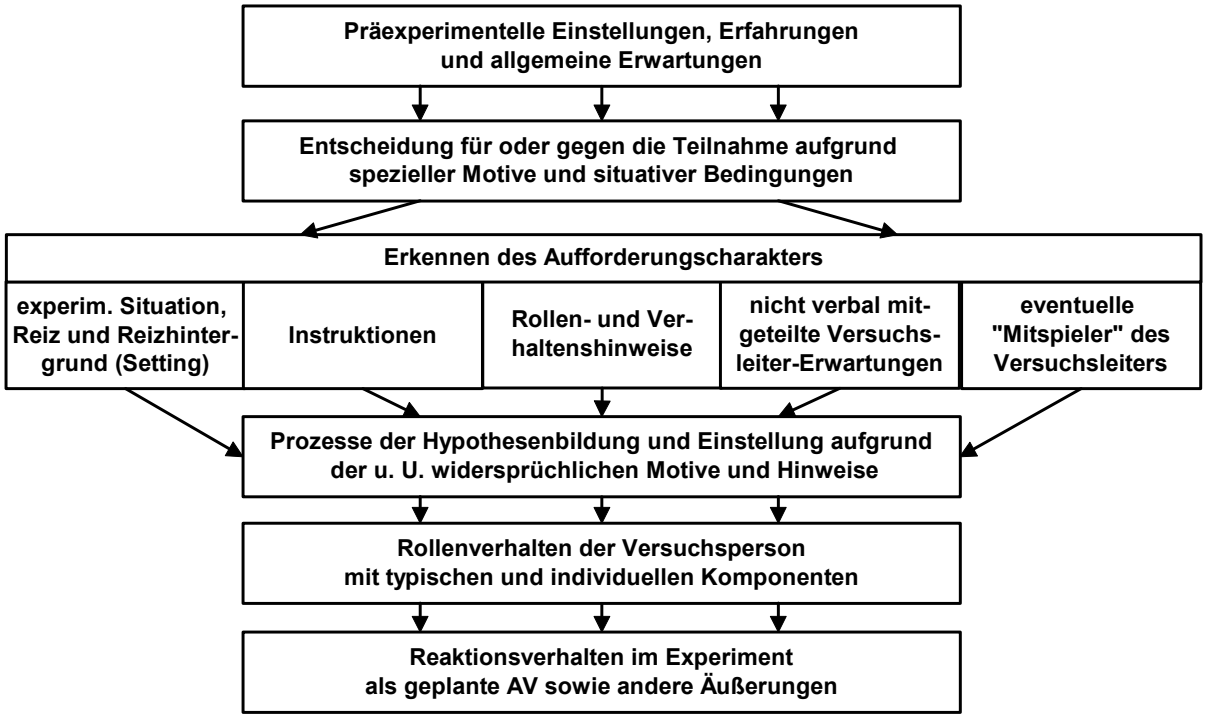


Abbildung 5.1: Übersicht über differentiell- und sozialpsychologische Aspekte

Als häufige **Teilnahme-Motive** gelten Neugier, Gefälligkeit, Teilnahmehonorare u.a. Belohnungen, Sachzwänge, Bereitschaft, Forschung zu unterstützen. Rosenthal und Rosnow faßten amerikanische Studien zur Psychologie der freiwilligen Vp zusammen: höherer Bildungs- und Sozial-Status, Streben nach Anerkennung, geringere Ausprägung der autoritären Einstellung, eventuell auch geselliger, unkonventioneller, Abwechslung suchend. Es wird vermutet, daß bei Versuchen mit klinisch-psychologischer, psychopharmakologischer und parapsychologischer Thematik verhaltensauffällige Personen überrepräsentiert sind. Bei den Motiven für das Verhalten während des Experiments nennt Gniech positive, wie Bedürfnis nach sozialer Anerkennung oder Bewertungsangst, welche zu Anpassungsverhalten führen, und negative, wie psychologische Abwehr und Reaktanz ("Gegenhandeln"), welche zu Vermeidung und verstecktem oder offenem Widerstand führen. Reaktanz kann durch zu hohe Anforderungen an die Vp, subjektiv als Nötigung erlebte Instruktionen oder als Folge von Täuschungsmanövern ausgelöst werden.

Unter **Aufforderungscharakter** (demand characteristic) wird seit Lewin die Eigenschaft einer Situation, Handlungen nahezulegen bzw. auszulösen, verstanden. Der Aufforderungscharakter wird nicht vom Individuum auf die Situation projiziert bzw. ihr zugeschrieben, existiert aber auch nicht autonom, sondern ergibt sich in Wechselwirkung von Feldsituation und Bedürfnisspannungen/Befindlichkeiten.

Orne sieht die experimentelle Situation als Problemlösesituation, in welcher die u.U. verunsicherte Vp sich zu orientieren versucht und dabei Hinweisreize für die geforderte Leistung und für die eigene Rolle und Motivation aufnimmt, um die "wahre" Absicht des Vl und des Experiments zu erfassen. So werden demand characteristics der Instruktion, der experimentellen Situation (setting), des Versuchsablaufs, der Vl-Erwartungen und des evtl. beteiligten Mitspielers des Vl unterschieden. Spezifische Vorerfahrungen, Argwohn und bestimmte Persönlichkeitsmerkmale bedingen u.U. eine intensivere Suche nach demand characteristics bzw. entsprechen Hinweisreizen (cues).

Informationsverarbeitung im Sinne der Instruktion wird von den Vpn erwartet; die Aktivität der Vpn und damit auch ihre Einflußnahme auf die Untersuchungsergebnisse können u.U. weit über dieses geplante Rollenverhalten hinausgehen: "die Versuchsperson als Wissenschaftler" – die Versuchsperson als Anhänger eines Überzeugungssystems, z. B. bestimmter Ansichten der "folk psychology", common sense psychology (etwa der populären Stress-Theorie) oder spezieller psychotherapeutisch-weltanschaulicher Systeme.

Vor allem aus den Theorien über kognitive Prozesse und aus sozialpsychologischen Theorien sind wichtige Gesichtspunkte und bestimmte **Anregungen für Kontrollen** abzuleiten:

- Hypothesentheorie der Wahrnehmung im Hinblick auf Personenwahrnehmung und Situationswahrnehmung;
- Erfolg und Mißerfolg usw. in der Untersuchungssituation;
- Attributionstheorie im Hinblick auf die Zuschreibung von Verhaltensursachen ("Kausaldeutungen", "naive Verhaltenstheorien");
- Informationsverarbeitung im Hinblick auf Wissenserwerb, Ausbildung von Schemata (Stereotypien, Überzeugungs-(belief-)systemen), implizite (naive) Konzepte, Urteils-(Entscheidungs-)prozesse; eigene Hypothesen über Zweck und Ergebnis der Untersuchung;
- Impression-Management-Theorie im Hinblick auf die Steuerung des Eindrucks auf andere Menschen (Selbstdarstellung, verschiedene Formen der assertiven (sich behauptenden) oder defensiven Selbstpräsentation);
- die Theorie psychologischer Reaktanzphänomene bei Einengung der Handlungsfreiheit (u.a. Widerstand, Abwehr, verringerte Compliance, "aus dem Felde gehen").

Typisierung von Vp-Einflüssen (Effekten), Vp-Rollen
Über die Beschreibung des **Hawthorne-Effekts** und **des Verhaltens nach sozialer Erwünschtheit** hinaus haben Orne u.a. Autoren versucht, speziellere **Rollentypen** hervorzuheben, die heuristische Bedeutung haben, u.a.:
(a) die "gute" Vp ist motiviert, die Instruktion genau zu befolgen und die Hypothesen des Vl zu erraten, um sie ihm bestätigen zu helfen,
(b) die "um ihre Bewertung besorgte" Vp versucht einen möglichst guten Eindruck zu machen und ihr Reaktionsverhalten unter diesem Ziel zu beeinflussen,
(c) die "ehrliche" Vp bemüht sich um eine sehr genaue Befolgung der Instruktion und enthält sich jeder Hypothesenbildung,
(d) die "negativistische" Vp wehrt sich gegen die Ausforschung, versucht die vermutete Vl-Hypothese zu widerlegen oder deren Beantwortung zu sabotieren.

Versuchsleiter-Effekte (experimenter bias)
Mit Rosenthal lassen sich unterscheiden:
Vl-Effekte (experimenter effects), d.h. Effekte der Person des (der) Vl, die unabhängig von seinen (ihren) Erwartungen sind:
(a) bio-soziale Effekte (Geschlecht, Alter, Rasse, Aussehen),
(b) psycho-soziale Effekte (Persönlichkeitsmerkmale, Verhaltensstil), Freundlichkeit, Wärme - Kühle, Dominanz,
(c) situative Effekte (Aufdringlichkeit, verunsichernde Wirkung)

Vl-Erwartungs-Effekte (experimenter expectancy effects), d.h. Effekte aufgrund der Erwartungen des (der) Vl:
(a) Beeinflussung der Vp (durch besondere Handhabung der Instruktion, verbale Konditionierungsstrategien, paraverbale Signale, z. B. Betonung, non-verbale Signale, z. B. Mimik),
(b) hypothesenkonforme Fehler (bei Datenbeobachtung, Datenaufzeichnung, Datenverarbeitung).

Besonders bekannt sind die Effekte aufgrund der eigenen Erwartung des (der) Vl über die Ergebnisse der Untersuchung (als Beispiel: Beurteilung einer Serie von Gesichtern hinsichtlich erfolgreich - nicht erfolgreich), doch sind diese Effekte nur schwer zu reproduzieren, so daß in der neueren Literatur oft von einer Überschätzung der Rosenthal-Effekte gesprochen wird – der Vl-Effekt selbst ein Rosenthal-Effekt?

Reaktivität
Unerwünschte Effekte, die durch die Versuchsteilnehmer oder durch die Untersucher selbst herbeigeführt wurden, schränken die innere Validität der Untersuchung ein und werden als Fehler und Artefakte bezeichnet. **Methodenbedingte Effekte**, welche durch das spezielle Verfahren der Datenerhebung bedingt sind, also nicht durch das Merkmal selbst, werden allgemein als **reaktive** Effekte bezeichnet. Psychologische Verfahren sind mit wenigen Ausnahmen (Analyse von Texten und Werken, objektiven Spuren des Verhaltens) reaktiv, allerdings in unterschiedlichem, auch von den jeweiligen Umständen abhängigem Maß. Die Anwendung der Methode (z. B. die Introspektion oder das Interview) beeinflußt das Phänomen. Dies kann auch für physiologische Messung, z. B. die Blutdruckmessung, zutreffen. Generell kann hier eine Unschärferelation (analog der von Heisenberg beschriebenen Unschärferelation in der Quantenphysik) behauptet werden.

Kontrollmethoden
Wenn also die Vpn eines Experiments keine beliebig austauschbare Elemente einer Population, sondern erlebende, hypothesenbildende und interagierende Individuen sind, dann müssen

Experimentalpsychologen differentielle und sozialpsychologische Aspekte berücksichtigen, sei es durch Erhebung wesentlicher Kontrollvariablen (bzw. Kovariablen) oder durch zusätzliche Kontrollmaßnahmen.

(1) **Kontrollen der Situation** (nach Orne u.a.) postexperimentelle Befragung zur Exploration der von der Vp wahrgenommenen und sie beeinflussenden demand characteristics (Versuchserleben); Nicht-Experiment (gedankliche Vorstellung und Schilderung); Rollenspiel oder Simulation (als-ob-Experiment, Vl unwissend).

(2) **Kontrolle der Vl-Effekte**
Blind- und Doppelblind-Versuche; Manipulation des Vl durch Induktion einer Hypothese; Einführung einer Erwartungskontroll-Gruppe; Kontrolle des Vl-Verhaltens durch Training und/oder Beobachtung (durch Vpn, durch Experten).

(3) **Kontrolle der Vp-Effekte**
Weitgehende Vorinformation und Offenheit; Eingewöhnung und Vorbereitung (Prätest); Alternative: entweder keine spezielle Information über H_1 oder Doppelblind-Versuch; Kontrolle durch Null-Behandlung; Erwartungskontroll-Gruppe (Wartelisten-Gruppe); Kontrolle durch alleinige Nachher-Beobachtung (Schätzung des Prätest-Effektes, siehe Solomon-4-Gruppen Plan).

(4) Weiterentwicklung des „**Ambulanten Assessment**“ unter **naturalistischen Bedingungen** und der sog. Aktionsforschungs-Methodologie (Feldstudien) als Alternative der herkömmlichen Forschungspläne, sofern dies für bestimmte Fragestellungen zweckmäßig ist.

Sarris (1992, S. 251) nennt als typische Mängel bei der Versuchsdurchführung:

- Die jeweils besondere Versuchsleiter – Versuchsteilnehmer – Dynamik bei Untersuchung von verschiedenen Bezugsgruppen wird nicht hinreichend beachtet.
- Die Instruktion sowie die gesamte Versuchsdurchführung sind nicht zuvor erprobt worden.
- Der (die) Vl hat das eigene Experiment nicht im Selbstversuch kennengelernt und ist mit den verschiedenen reaktiven Effekten, die vorkommen können, nicht vertraut.
- Es erfolgt keine sorgfältige Exploration im Anschluß an die Untersuchung.

"Pflege der Versuchspersonen"
Zur Motivierung der Vpn und zur Optimierung der Vl-Vp-Interaktion empfiehlt sich die "Pflege" der Versuchspersonen: bei Anwerbung, fairer Vorab-Information, Begrüßung und Anwärm-(Adaptions-) phase, Instruktion und Durchführung, abschließender Information über Absichten und Ergebnisse, Verabschiedung (siehe auch Huber, 1987, mit Erläuterungen zur Rolle des Vl: "Vl als Gastgeber").

Die **Zumutbarkeit der Untersuchungsbedingungen** (u.a. Dauer, Art der Aufgaben) bzw. deren Akzeptanz aus Sicht der Vp sind wichtige Aspekte der Untersuchungsplanung. Die Vp soll sich in der Untersuchungssituation möglichst wohl fühlen, und die Abmachungen sollen eingehalten werden – nach dem **Leitprinzip der Austauschgerechtigkeit**.

Die Untersucher müssen außerdem über wichtige **berufsethische und rechtliche Prinzipien** informiert sein:

- das Prinzip der informierten Zustimmung (Freiwilligkeit der Teilnahme nach Aufklärung über wesentliche Aspekte; Möglichkeit, jederzeit ohne Nachteile abbrechen zu können);
- Datenschutzbestimmungen.

(Siehe auch Vorlesung zu "Wissenschaftstheorie, Geschichte und Berufsethik der Psychologie" sowie Schuler, 1980).

Literaturhinweise

Frey, D. & Irle, M. (Hrsg) (1984). *Theorien der Sozialpsychologie*. Band I. Kognitive Theorien (2. Aufl.) Bern: Huber

Frey, D. & Irle, M. (Hrsg) (1985). *Theorien der Sozialpsychologie*. Band III. Motivations- und Informationsverarbeitungstheorien. Bern: Huber.

Gniech, G. (1976). *Störeffekte in psychologischen Experimenten*. Stuttgart: Kohlhammer.

Huber, O. (1987). *Das psychologische Experiment: Eine Einführung*. Bern: Huber.

Maschewsky, W. (1977). *Das Experiment in der Psychologie*. Frankfurt a. M.: Campus.

Rosenthal, R. & Rosnow, R.L. (Eds.) (1969). *Artifact in behavioral research*. New York: Academic Press.

Sarris, V. (1992). *Methodologische Grundlagen der Experimentalpsychologie*. Bd. 2. München: Reinhardt.

Schuler, H. (1980). *Ethische Probleme psychologischer Forschung*. Göttingen: Hogrefe.

5.2 Validität von Untersuchungen

Mit Validität (Gültigkeit) ist herkömmlich die Validität eines psychologischen Datenerhebungsverfahrens gemeint, insbesondere die Validität eines psychologischen Tests: Mißt der Test tatsächlich das, was er vorgibt zu messen?

Der Begriff Validität wurde auf experimentelle Versuchspläne bzw. Untersuchungspläne im allgemeinen übertragen und entsprechend erweitert (Campbell & Stanley: Aspekte der internen und externen Validität). Validität ist das Ausmaß, in dem eine bestimmte Untersuchung (Experiment) zur Prüfung einer Hypothese (Kausalhypothese) geeignet ist (siehe Hager & Westermann, 1983; Cook & Campbell, 1979):

Variablenvalidität der Untersuchung

- Übersetzung der theoretischen Begriffe in beobachtbare Variablen bzw. Fehler dieser Operationalisierung;
- **interne Validität**
- Randomisierung (Person/Treatment) als Voraussetzung strenger Hypothesenprüfung bzw. Gefährdungen der internen Validität bei mangelhafter Randomisierung;
- **Populations- und Situationsvalidität** (externe Validität)
- Generalisierbarkeit der Hypothesenprüfung (Untersuchungsbefunde) auf andere Personen, Settings, ggf. auch andere Operationalisierungen der Konstrukte;
- **statistische Validität**

Adäquatheit der statistischen Auswertungsmethoden

- Adäquatheit des Signifikanztests für die zu prüfenden wissenschaftlichen Hypothesen bzw. typische Fehler; Effektgrößen und Wahl des Stichprobenumfangs u.a.

Variablenvalidität

Den theoretischen Begriffen der Hypothese müssen bestimmte empirisch beobachtbare Variablen als **Operationalisierungen** (Realisierungen) zugeordnet werden, wobei verschiedene Störfaktoren zu beachten sind. Eine Hypothese kann um so strenger geprüft werden (Hager & Westermann, 1983, S. 46):

- je eindeutiger den theoretischen Begriffen der Hypothese empirische Variablen ("Operationalisierungen") zugeordnet sind,
- je mehr Operationalisierungen der theoretischen Begriffe berücksichtigt werden und je unterschiedlicher diese Operationalisierungen sind,
- je mehr die berücksichtigten Ausprägungen der UV den möglichen Ausprägungen der entsprechenden theoretischen Variablen entsprechen,

- je eher das Skalenniveau der empirischen Variablen der Struktur der zugehörigen theoretischen Begriffe entspricht,
- je weniger bei den berücksichtigten Operationalisierungen andere theoretische Variablen mit den Begriffen der Hypothese konfundiert sind.

Interne Validität
Die strenge Prüfung einer Kausalhypothese verlangt ein Experiment, d.h. **Randomisierung** der Zuordnung von Personen (Untersuchungseinheiten) und Bedingungen bzw. Randomisierung der Reihenfolge, in der Personen unter diesen Bedingungen beobachtet werden. Fehlt diese Randomisierung, handelt es sich um Quasi-Experimente (Campbell & Stanley), um Korrelationsstudien oder andere Forschungsansätze. Wenn gezeigt werden kann, daß in einem quasi-experimentellen Versuchsplan bestimmte Störfaktoren **keine** Rolle spielen, nähert sich dieser Forschungsansatz dem Experiment an. Vorbild der strengen Prüfung bleibt aber das Experiment.

Auch im Experiment gibt es **Gefährdungen der internen Validität**, weil nicht alle wesentlichen Bedingungen und Kontexte konstant gehalten werden können. Störfaktoren der internen Validität können dazu führen, daß nicht nur die experimentelle UV, sondern andere Bedingungen ebenfalls zu unterschiedlichen Werten auf der AV führen. Hager und Westermann (1983) unterscheiden die Variation personaler und situationaler Merkmale, Störfaktoren bei Meßwiederholung sowie bei der interindividuellen und bei der intraindividuellen Bedingungsvariation.

Einprägsam ist die folgende **Aufzählung THIS MESS** (siehe Cook & Campbell, 1979), welche experimentelle und quasi-experimentelle Versuchspläne betrifft und dabei auch die Sequenzeffekte (Erst- Zweit-Messung) hervorhebt:
Testing: Testeffekte, d.h. Auswirkung einer Testdurchführung (Datenerhebung) auf die folgende Durchführung, z. B. durch Vertrautheit mit der Aufgabe, Lern- bzw. Übungseffekte, Sensibilisierungseffekte usw.
History: Zwischenzeitliches Geschehen, d.h. besondere Ereignisse, die zwischen erster und zweiter Durchführung eintreten, z. B. relevante politische Tagesereignisse oder individuelle Erlebnisse.
Instrumentation: Veränderung in der Beschaffenheit der Untersuchungsinstrumente, Veränderung der Kriterien bei Beobachtung und Bewertung von Variablen, z. B. besseres Training oder Einstellungsänderung auf der Seite der Beobachter/Auswerter, oder Wechsel des Vl.
Statistical Regression: "Regression zur Mitte" bei einer zweiten Durchführung aufgrund der nicht-perfekten Reliabilität der Meßwerte - besonders deutlich, wenn Vpn mit relativ extremer Merkmalsausprägung (Kontrastgruppen) untersucht wurden.
Maturation: "Reifung", d.h. psychologische und physiologische Veränderungen der Vpn, welche unabhängig von der experimentellen Bedingungsvariation eintreten, z. B. Vpn sind bei der zweiten Untersuchung müder, hungriger, älter usw.
Experimental Mortality: in den Gruppen fallen unterschiedlich viele Vpn aus, z. B. bei längeren Untersuchungen oder Behandlungen, u.U. reaktiv in Abhängigkeit von der Art der Behandlung.
Selection: Auswahlverzerrung, d.h. unterschiedliche Zusammensetzung der Gruppen, so daß Ausgangsbedingungen, Reaktivität und Behandlungseffekte unterschiedlich sein können oder nicht-äquivalente Kontrollgruppen verwendet werden.
Selection-Maturation-Interaction: Wechselwirkung von Selektion und Reifung, d.h. Auswahl von Gruppen mit unterschiedlichen "Reifeprozessen", z. B. Untersuchung zu unterschiedlichen Tageszeiten oder bei unterschiedlichem Entwicklungsalter (z. B. bei gleichaltrigen Jungen und Mädchen).

Populations- und Situationsvalidität (externe Validität)
Psychologische Hypothesen beanspruchen – zumindest anfänglich – Gültigkeit für alle Menschen ("unbegrenzte Allsätze"). Die empirische Prüfung geschieht jedoch an einer ausgewählten Gruppe von Personen (z. B. Studierenden der Psychologie) unter bestimmten raumzeitlichen Bedingungskonstellationen (Situationen, Settings, u.a. Labor – Feld). Deshalb sind Überlegungen und empirische Prüfungen zur Generalisierbarkeit der Ergebnisse hinsichtlich **anderer Personengruppen**, hinsichtlich **anderer Settings**, ggf. auch **anderer Operationalisierungen** (z. B. andere Beobachtungs-/Testmethoden für dasselbe Konstrukt) (und **anderer Versuchsleiter**?) notwendig.

Eine Untersuchung zur Prüfung einer Kausalhypothese ist um so strenger:
- je weniger sie sich auf bestimmte Untermengen der Population,
- je weniger sie sich auf bestimmte zeitliche, räumliche und situationale Umstände aus dem Geltungsbereich der Hypothesen beschränkt (Hager & Westermann, 1983).

Es ist üblich, aus Theorien und Hypothesen, die sich unter Laborbedingungen mit relativ hoher interner Validität bewährt haben, sog. technologische Prognosen für verschiedene Anwendungsfelder außerhalb des Labors, d.h. komplexe Settings, in denen auch andere Variablen zu berücksichtigen wären, abzuleiten. Die strenge Prüfung einer Hypothese verlangt jedoch – neben der Replikation in anderen Laboratorien – auch ihre Überprüfung in anderen Settings bzw. die wechselseitige Ergänzung von Laborexperiment, Quasi-Experiment und Feldstudie bzw. Feldexperiment.

5.3 Evaluation wissenschaftlicher Ergebnisse

5.3.1 Evaluation und Replikation
Für die Evaluation und Kommunikation wissenschaftlicher Ergebnisse, d.h. von neuen Theorien und Methoden oder von Hypothesenprüfungen, haben sich in der "scientific community" bestimmte Formen herausgebildet.

Die **Evaluation (kritische Bewertung, Bewährungskontrolle)** einer wissenschaftlichen Arbeit setzt zweierlei voraus: (1) alle wichtigen Einzelheiten müssen mitgeteilt sein und (2) die Beurteiler(innen) müssen fachlich qualifiziert sein (Stegmüller: Überprüfbarkeit durch qualifizierte Beobachter als ein Kriterium der Wissenschaftlichkeit).

Die Validität einer Untersuchung kann am überzeugendsten durch **Replikation** gezeigt werden (Neuliep, 1991; Schweizer, 1989). Die "identische" Wiederholung einer Untersuchung führt zur:
- Replikation, d.h. Bestätigung und weiterer Bewährung,
- Falsifikation der Theorie für das betreffende Anwendungsfeld, und anschließend wahrscheinlich zu Vermutungen, daß die Bedingungen wider Erwarten nicht hinreichend "identisch" waren.

Replikationen durch den Autor oder – die wichtigeren – unabhängigen Replikationen durch ein anderes Labor (cross-laboratory replication) bzw. eine andere Untersuchergruppe sind in der Psychologie – im Unterschied zu Naturwissenschaften – leider selten: ein Zeichen von Kreativität? autistisch-undisziplinliertem Denken? Beliebigkeit der Operationalisierungen und Designs? Desinteresse an der Sicherung von Sachverhalten?

Wegen der methodischen Heterogenität der oft zahlreichen Untersuchungen zu wichtigen Fragestellungen, z. B. zur Bewährungskontrolle von Psychotherapie oder zur Wirksamkeit bestimmter Medikamente, sind Bilanzierungen des Forschungsstands notwendig. Die zu-

sammenfassende Bewertung einer Reihe empirischer Untersuchungen zu einer bestimmten Fragestellung kann als

- Identifikation hervorragender und vorbildlicher Untersuchungen durch kritischen Vergleich
- Sammelreferat (als narrative Übersicht oder integratives Review)

durchgeführt werden, doch sind die Maßstäbe bzw. Evaluationskriterien schlecht objektivierbar. Deshalb wurde in der Psychologie und in der Medizin die Methodik der formalisierten **Metaanalyse** entwickelt.

5.3.2 Meta-Analyse

Notwendigkeit integrativer Forschungsmethoden

Die Publikationen zu Fachthemen werden immer spezieller und immer zahlreicher. Es stellt sich die Frage, wie die stetig wachsenden Fülle an Literatur zu bewältigen ist.

Ein Nachteil des konventionellen **integrativen** Reviews besteht in der Subjektivität und Intransparenz der Selektionskriterien für zitierte/diskutierte Primärstudien. Die Kriterien, die zur Auswahl von Einzelarbeiten für die genauere Darstellung oder Diskussion im Review angewandt werden, werden oft nicht mitgeteilt. Das „Gesamtfazit" beruht zudem vielfach auf einem subjektiven „Gesamteindruck" der dargestellten Literatur.

Auch für die Zusammenfassung von Ergebnissen in integrativen Reviews müssen **Selektions- und Bewertungskriterien** explizit dargestellt werden (Objektivitätskriterium wissenschaftlichen Arbeitens).

Merkmale der Metanalyse

Cooper (1982) hat vorgeschlagen, die statistischen Methoden, die auch sonst bei der Analyse und Interpretation von empirischen Daten üblich sind, auf den integrativen Review-Prozess zu übertragen. Damit werden die Ergebnisse der zu integrierenden Untersuchungen zu „Roh-Daten" des (quantitativen) integrativen Reviews.

1. **Primäranalyse** ist die (statistische) Analyse der Daten voneinander unabhängiger Untersuchungen verschiedener Forscher (Originalarbeiten);
2. **Sekundäranalyse** ist die Analyse der Daten einer Primäranalyse mit verbesserten statistischen Methoden oder die Analyse dieser Daten unter einer neuen Fragestellung.;
3. **Meta-Analyse** ist die Analyse von Analysen im Sinne der statistischen Analyse einer größeren Sammlung statistischer Ergebnisse (Glass, 1976, S. 3), z. B. als quantitative Kumulation und Analyse der deskriptiven Statistiken der Primäranalysen (Hunter et al., 1982).

Es gibt nicht **die** Meta-Analyse, sondern „Meta-Analyse" bezeichnet ein ganzes Bündel von Maßnahmen zur Erhöhung der Gültigkeit von integrativen Aussagen, nämlich: Regeln zur Erhöhung der Objektivität der Meta-Analyse, statistische Techniken zur quantitativen Beschreibung und Integration von Effekten, statistische Techniken zur Absicherung der Aussagen gegen Zufallsfehler und Regeln zur Interpretation der erhaltenen Befunde.

Komponenten der Meta-Analyse

Die Meta-Analyse wird als Bestandteil eines integrativen Review-Prozesses verstanden, der nach Cooper (1982) folgende Schritte umfassen soll: Problemformulierung, Sammlung der Ergebnisse, **Analyse und Interpretation der Ergebnisse (→ Meta-Analyse i.e.S.)** und Präsentation der Ergebnisse.

Problemformulierung: Ausgangspunkt jedes integrativen Reviews, also auch des quantitativen, ist eine Fragestellung oder ein Problem, zu der/dem es bereits eine Reihe an Ergebnissen aus Primäranalysen gibt.
Sammlung der Ergebnisse: Wichtiger Bestandteil eines integrativen Reviews ist die möglichst vollständige Sammlung der relevanten Untersuchungen zur Frage- oder Problemstellung des Reviews. Solche Ergebnisse können gefunden werden in Büchern, Zeitschriften, Projekt-(Forschungs-) berichten, Diplomarbeiten, Dissertationen und Habilitationen.

Informationen über diese Quellen erhält man z. B. über (a) Bücher: (1) Wellek (1965): Gesamtverzeichnis der deutschsprachigen psychologischen Literatur der Jahre 1942 bis 1960, (2) Dambauer (1972 ff): Bibliographie der deutschsprachigen psychologischen Literatur, (3) Reinert (1976 ff): Bibliographie deutschsprachiger psychologischer Dissertationen, (4) Psychological Abstracts (1927 ff; mit „Thesaurus of Psychological Index Terms", Cumulated Subject Index to Psychological Abstracts" und „Author Index to Psychological Index"), (b) Datenbanken: (1) American Psychological Association (1967 ff): PsycINFO, (2) Zentralstelle für Psychologische Information und Dokumentation an der Universität Trier (1981 ff): PSYNDEX, (3) Institute for Scientific Information (1973 ff): SOCIALSCI, (4) American Psychological Association (1974 ff): PSYCLIT und (5) National Library of Medicine (1966 ff): MEDLINE

Bewertung der Ergebnisse
Die Gültigkeit der Meta-Analyse hängt von der Gültigkeit (Validität) der Ergebnisse der zu integrierenden Studien ab. Hierbei werden unterschiedliche Validitätsaspekte berücksichtigt:

Validitätsaspekt	**Erläuterung**
Interne Validität	Sind die gefundenen Effekte auf die AV eindeutig auf die Variation der UV zurückführbar ?
Externe Validität	Inwieweit lassen sich die gefundenen Ergebnisse verallgemeinern ?
Variablenvalidität	Wurde das interessierende hypothetische Konstrukt angemessen operationalisiert ?
Statistische Validität	Entspricht die formulierte statistische Hypothese der inhaltlichen (Forschungs-) hypothese ? Wird die UV innerhalb der Untersuchung in standardisierter Weise vorgegeben ? Weist der statistische Test eine ausreichende Teststärke auf ? Werden die statistischen Voraussetzungen geprüft ? Werden die Signifikanztests richtig interpretiert ? Wird der gefundene Effekt richtig interpretiert ?

Anmerkung: In diesem Stadium des integrativen Review-Prozesses können bzw. sollten Studien von methodisch minderer Qualität ausgeschlossen werden, wobei die Auswahl und Strenge der Maßstäbe sicher einen Einfluß auf die Tendenz der Ergebnisse haben wird.

Analyse der Ergebnisse
Die Meta-Analyse i.e.S. ist nicht ein, sondern eine Gruppe von statistischen Auswertungsverfahren. Im nachfolgenden sollen einige ausgewählte Beispiele vorgestellt werden.

Box-Counting ist das vergleichende Auszählen von Studienergebnissen (in einem bestimmten Untersuchungsbereich) mit signifikantem versus nicht-signifikantem Ergebnis. **Nachteile des Box-Countings**: Hoher Fehler 2. Art (fälschliches Nicht-Entdecken von vorhandenen Effekten). Brower und Owen (1973): viele Arbeiten sind hinsichtlich ihrer Stichprobengröße

so angelegt, daß allenfalls Effekte mit großer Effektstärke „signifikant“ werden können, d.h. Box-Counting führt zu vielen „misses“ und viele Reviews sind zu pessimistisch.

Empirische Vergleiche des Box-Countings mit differenzierteren Analysen ergaben folgende Ergebnisse:

Box-Counting	**Differenziertere Analysen**
15 : 25 Arbeiten finden keinen signifikanten Zusammenhang zwischen „indirektem Unterrichtsstil des Lehrers“ und Maßen der „Schülerleistung“ (Dunkin & Biddle, 1974)	Es besteht ein signifikanter Zusammenhang zwischen den beiden Variablen (Gage, 1979), bei insgesamt mittlerer Effektstärke (Fricke, 1977)
20 : 32 Arbeiten finden keinen signifikanten Zusammenhang zwischen vorstrukturierten Lernhilfen durch den/die Lehrer/in und dem Lern- und Behaltenserfolg (Barnes & Clawson, 1975)	Es besteht ein kleiner Effekt der vorstrukturierten Lernhilfen auf den Lern- und Behaltenserfolg (Luiten et al., 1980).

Summierung von Test-Statistiken

1. Fall: Liegen aus n unabhängigen Untersuchungen n unabhängige Wahrscheinlichkeiten $p(\text{Daten} \mid H_0)$ vor, so ist der folgende Wert annähernd normalverteilt:

$$a = -2 \sum_{i=1}^{n} \log_e p_i$$

Fischer-Methode, „Adding of logs“ angemessen für kleine Studienzahlen ($n \leq 5$)

2. Fall: Liegen aus n unabhängigen Studien n t-Werte vor, so können diese t-Werte nach folgender Formel zu einem integrativen Z-Wert aufsummiert werden:

$$Z = \frac{\sum_{i=1}^{n} t_i}{\text{sqrt}(df_i / (df_i - 2)}$$

(sqrt=Quadratwurzel)

3. Fall: Liegen aus n unabhängigen Studien n z-Werte vor so können diese — gewichtet mit der Anzahl Freiheitsgrade — gemittelt werden:

$$Z = \frac{df_1 z_1 + df_2 z_2 + \ldots + df_n z_n}{\text{sqrt}(df_1^2 + df_2^2 + \ldots + df_n^2)}$$

(sqrt=Quadratwurzel)

Anmerkung

Die dargestellten Verfahren bilden nur einen kleinen Ausschnitt aus dem Methodenrepertoire der Meta-Analyse. Die Metaanalyse besteht in der Definition von Effekten und Effektstärken, d.h. dem standardisierten Mittelwertunterschied zwischen Experimental-(E) und

Kontrollgruppe (K) $(X_E - X_K)/S_K$ und deren Aggregation über verschiedene Untersuchungen (z. B. Smith, Glass & Miller, 1980; The benefits of psychotherapy): Mittlerer Effekt 0.8. Dies bedeutet, daß die mittlere Veränderung in einem Test, der in der Lage ist, therapeutisch bedingte Veränderungen abzubilden, bei 0.8 der zur Standardisierung verwendeten SD liegt. Ein Patient, der die Therapie mit einem Prozentrang 50 in diesem Test beginnt, erreicht beim Abschluß einen Prozentrang von 80. Als Beispiel siehe auch die Metaanalyse von Myrtek (1993) zur psychophysiologischen Persönlichkeitsforschung (siehe Skript zur Vorlesung Differentielle Psychologie und Persönlichkeitsforschung).

Evaluationsforschung, d.h. die methodisch fortgeschrittene Bewertung von empirischen Untersuchungen (als Metaanalyse) und die Bewährungskontrolle von "Programmen", z. B. in der Pädagogik, Gesundheitserziehung, Arbeitsorganisation, ist ein zunehmend wichtiges Arbeitsgebiet von methodisch geschulten Psychologen (siehe auch Wittmann, 1985). Eine herausragende, sich auch auf die Gesetzgebung („Psychotherapie-Gesetz“) auswirkende Bedeutung haben Metaanalysen hinsichtlich der **Effekte von Psychotherapie** (Grawe, Donati & Bernauer, 1994).

Begutachtungen

Die **Evaluation wissenschaftlicher Arbeiten** (sofern nicht durch Prüfungsordnungen für Diplom-, Doktor- und Habilitationsarbeiten geregelt), geschieht durch **Gutachter**: (1) Herausgeber u.a. Gutachter bei wissenschaftlichen Zeitschriften u.a. Verlagswerken, welche über Ablehnung, Revision oder Annahme eines Manuskripts entscheiden, (2) gewählte oder ernannte Gutachter (z. B. der Deutschen Forschungsgemeinschaft DFG, des Bundesministers für Forschung und Technologie BMFT), welche Anträge (Projekte, Forschergruppen, Schwerpunktprogramme, Sonderforschungsbereiche) befürworten oder ablehnen. Solche Evaluationen können, wie Fehlprognosen und offenkundige Fehlentscheidungen lehren, nicht fehlerfrei sein. Die beste Gewähr bietet wahrscheinlich ein System, das auf dem Urteil **der von den Wissenschaftlern selbst gewählten Gutachter** mit Unterstützung von spezialisierten Sondergutachtern beruht (DFG-System).

Publikationsbias

Als **Publikationsbias** wird die - begründete - Vermutung bezeichnet, daß Publikationsentscheidungen systematischen Verzerrungen unterliegen. Eine größere Chance haben wahrscheinlich:

- bereits bekannte Autoren oder Gruppen,
- positive (neue) Ergebnisse gegenüber kritischen Auseinandersetzungen und Widerlegungen (negative, kritische Ergebnisse; beibehaltene Null-Hypothesen),
- in der herrschenden Meinung ("main stream") liegende Arbeiten gegenüber Positionen von Minderheiten oder Außenseitern,
- sprachlich-stilistisch geglückte, "lesbare" Arbeiten.

Als **file-drawer-Problem** wird die - ebenfalls begründete - Vermutung bezeichnet, daß Arbeiten mit negativen Ergebnissen aus psychologischen Gründen von vorn herein eher im Schreibtisch bleiben. Die publizierten Arbeiten zu einem Thema würden demnach ein unzutreffendes Bild der tatsächlichen Untersuchungslage liefern. In Metaanalysen läßt sich schätzen, wieviele Arbeiten mit beibehaltener H_0 in den Schreibtischen liegen müßten, um eine Bilanz von z. B. 10 positiven und 4 negativen Publikationen auszugleichen bzw. zur Insignifikanz zu verschieben. Ist diese sog. **failsafe-Statistik**, d.h. die Anzahl von zusätzlichen Studien mit Null-Resultaten, welche die beobachtete mittlere Effektgröße auf einen unbedeutenden Wert reduzieren würden, unrealistisch hoch, so spricht dies für die Validität der (positiven) Bilanz.

5.4 Wissenschaftliche Kommunikation

An wissenschaftlicher Kommunikation sind nicht allein die Wissenschaftler aus sachlichen und persönlichen Motiven interessiert, sondern auch die Öffentlichkeit. Dies gilt in hohem Maße für die psychologische Forschung, die (1) im Hinblick auf praktische Anwendungen und (2) grundsätzlich als Humanwissenschaft im Grenzgebiet von Natur-, Sozial- und Geisteswissenschaften besonderes Interesse und Breitenwirkung finden sollte. (Zur Relevanzdiskussion siehe Vorlesung "Wissenschaftstheorie, Geschichte und Berufsethik der Psychologie")

Der wissenschaftlichen Kommunikation dienen Publikationen, wissenschaftliche Tagungen und Kongresse, Vortragsreisen, Ausbildungs- und Gastaufenthalte. Heutzutage dominiert die englische Sprache, so daß es für die internationale Kommunikation unerläßlich ist, in den englischsprachigen Fachzeitschriften (mit hohem "impact", d.h. hoher Zitierhäufigkeit durch andere Wissenschaftler) zu publizieren. Durch statistische Auswertung von Literaturbanken (Science Citation Index und Social Science Citation Index) können die Anzahl der Publikationen jedes(r) Wissenschaftlers(in) und das Zitiertwerden durch andere ermittelt werden. Es gibt deutsche Autoren, die (fast) nur noch englischsprachig publizieren, andere die (fast) ausschließlich deutsch publizieren und deshalb in der internationalen scientific community unbekannt bleiben und auch wenig Einfluß auf die fortschreitende Diskussion haben werden. Die auch in Deutschland geführte Auseinandersetzung über nationale Sprache und Bezugsgruppe bzw. Provinzialismus und Internationalismus scheint weitgehend abgeschlossen zu sein: **der Trend zur englischsprachigen Wissenschaft und Publikation ist unaufhaltsam**.

Im Ausmaß der (inter)nationalen wissenschaftlichen Kommunikation gibt es große interindividuelle Unterschiede, aber auch Unterschiede zwischen Teilgebieten der Psychologie. Von der wünschenswerten Norm sind die heutigen Verhältnisse noch weit entfernt. Weshalb führt eine mit "sehr gut" bewertete Diplomarbeit nicht zu einer Publikation in einer deutschen Fachzeitschrift und zu einem Tagungs-Poster? Weshalb führt eine "magna cum laude" oder "summa cum laude" bewertete Dissertation nicht zu einer Publikation in einer internationalen Fachzeitschrift und zu einem Kongreß-Vortrag? Weshalb wird eine herausragende Habilitationsschrift nicht in einem angloamerikanischen Verlag oder zumindest ausschnittweise in internationalen Zeitschriften publiziert?

Weshalb scheinen viele deutschsprachige Zeitschriften ein milderes Begutachtungssystem als angloamerikanische Zeitschriften zu haben? Standard ist dort das sog. doppelte Review-System, d.h. die unabhängige Beurteilung durch zwei kompetente Beurteiler mit oft sehr ausführlich begründeten Revisionsvorschlägen oder Ablehnungen durch den zuständigen Herausgeber.

Für die "Bringschuld" des Wissenschaftlers an die Öffentlichkeit und die Steuerzahler, d .h. Politikberatung durch Experten und populärwissenschaftliche Publikationen bzw. Vorträge, sind noch weitere Maßstäbe zu berücksichtigen.

Ausgewählte Literatur

Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation. Design and analysis issues for field settings*. Chicago: Rand McNally.

Grawe, K., Donati, R. & Bernauer, F. (1994). *Psychotherapie im Wandel. Von der Konfession zur Profession*. Göttingen: Hogrefe.

Hager, W. & Westermann, R. (1983). Planung und Auswertung von Experimenten. In J. Bredenkamp & H. Feger (Hrsg.). *Enzyklopädie der Psychologie. Themenbereich B. Methodologie und Methoden. Serie I Forschungsmethoden der Psychologie. Band 5. Hypothesenprüfung* (S. 24-238). Göttingen: Hogrefe.

Keul, A.G., Gigerenzer, G. & Stroebe, W. (1993). Wie international ist die Psychologie in Deutschland, Österreich und der Schweiz? Eine SSCI-Analyse. *Psychologische Rundschau*, **44**, 259-269.
Neuliep, J.W. (Ed.) (1991). *Replication research in the social sciences*. Newbury Park, Ca.: Sage.
Schweizer, K. (1989). Eine Analyse der Konzepte, Bedingungen und Zielsetzungen von Replikationen. *Archiv für Psychologie*, **141**, 85-97.
Wittmann, W.W. (1985). *Evaluationsforschung. Aufgaben, Probleme und Anwendungen*. Berlin: Springer.

5.5 Wissenschaftliche und ethische Qualitätsmerkmale des Forschungsprozesses

Die Beziehung von wissenschaftlichen und ethischen Qualitätsmerkmalen des Forschungsprozesses (d.h. der Durchführung der Forschung, der Datenanalyse und der Berichterstattung) wird von Rosenthal (1994) diskutiert. Forschungsvorhaben mit mangelnder Qualität (hinsichtlich Versuchsplan, Analyse und Bericht) verbrauchen Resourcen (Zeit, Geld, Aufmerksamkeit u.a. Resourcen) und sind daher nach Ansicht Rosenthals auch ethisch fragwürdig. Die drei genannten Bereiche werden im folgenden bezüglich des Zusammenhangs von wissenschaftlichen und ethischen Qualitätsmerkmalen kurz beleuchtet.

1. Durchführung der Forschung

(1) Forschungsvorhaben, die mit Unsicherheiten für die Teilnehmer verbunden sind, sind ethisch fragwürdig.
(2) Die mangelnde Qualität eines Forschungsplans kann zu ungesicherten oder ungenauen Schlußfolgerungen führen, die für die Gesellschaft von Nachteil sein können.
(3) Die folgenden Probleme können bei der Rekrutierung von Versuchspersonen oder beim Einwerben von Forschungsgeldern auftreten:
(a) "Hyperclaiming": bei Teilnehmern oder Geldgebern werden überhöhte und kaum zu erfüllende Erwartungen geweckt.
(b) "Causism": meint die Tendenz, in Anträgen oder Rekrutierungen Kausalschlüsse zu versprechen, obwohl der Versuchsplan oder die Datenlage diese nicht zulassen. Dies zeigt sich etwa an der Verwendung kausalitätsimplizierender Begriffe (z. B. 'die Wirkung von', 'die Folge/Konsequenz/Resultat von') statt adäquaterer Formulierungen (z. B. 'vorhersagbar durch' oder 'verbunden mit'). Mögliche Gründe für die Überbewertung des kausalen Erklärungswertes einer Untersuchung beruhen meist auf einer noch ungenügenden empirischen Datenlage oder sind auf das absichtliche oder unabsichtliche Bemühen zurückzuführen, die (zu erwartenden) Ergebnisse wichtig erscheinen zu lassen.

2. Datenanalyse

(1) Ethisch fragwürdig ist der Ausschluß von Daten, die der Theorie oder der eigenen Vorhersage widersprechen. Die übliche Strategie, 'Ausreißer' zu entfernen, ist zumindest dann akzeptabel, wenn dieses Vorgehen begründet ist und im Forschungsbericht festgehalten ist. Auch die Selektion von Versuchspersonen oder Variablen sollte berichtet werden.
(2) "Snooping around in data": Nach traditioneller Sicht sind statistische Analysemethoden und Randbedingungen zuvor festzulegen. Weitere Analysen oder Reanalysen der Daten werden in aller Regel als unangemessen betrachtet. Rosenthal (1994) vertritt dagegen die Auffassung, daß Reanalysen ethisch gerechtfertigt sind, wenn Daten durch erhebliche Investitionen an Zeit, Geld u.a. Resourcen entstanden sind. Unter Verwendung angemessener Kontrolltechniken (z. B. Bonferoni-Adjustierung) und Replikationsuntersuchungen sind Reanalysen von Datensätzen aber durchaus als sinnvoll einzuschätzen.
(3) Metaanalysen stellen sinnvolle Instrumente zur Klärung von Pseudokontroversen oder zur Identifikation von bereits beantworteten Fragestellungen, d.h. zur Vermeidung von unsinnigen Investitionen, dar.

3. Berichterstattung (vgl. auch Kapitel "Wissenschaftliche Kommunikation")
(1) Problematisch ist die Fehlrepräsentation von Forschungsergebnissen:
(a) Die absichtliche Fehlrepräsentation, d.h. der Bericht nicht vorhandener Daten oder die Auswahl von Daten, die der Theorie entsprechen, ist ethisch nicht vertretbar.
(b) Auch durch unbeabsichtigte Fehlrepräsentationen (Fehler bei der Datenaufzeichnung, bei der Computeranalyse oder der verwendeten statistischen Prozeduren) kann das Kosten-Nutzen Verhältnis einer Untersuchung ungünstig beeinflußt werden.
(c) Unzulässige Generalisierungen.
(2) Die Fehlrepräsentation von Forschungsleistungen (Zuweisung von Coautoren in das 'Acknowledgement' statt zur Autorenschaft) oder die intellektuelle Aneignung von wissenschaftlichen Ideen kann ebenfalls ethisch fragwürdig sein.
(3) Fehlen oder Mißlingen von Veröffentlichungen, aufgrund von:
(a) Selbstzensur (z. B. werden zuweilen eigene unveröffentliche Datensätze, insbesondere mit ungünstigen Ergebissen unterdrückt; diese sollten jedoch ebenfalls anderen zugänglich gemacht werden);
(b) externer Zensur durch Herausgeber, Reviewer, Programmkommitees: diese ist allerdings zur Bewertung der methodischen Standards einer Untersuchung erforderlich.

An dieser Stelle wird auch auf die Vorlesungen zur Berufsethik innerhalb der Vorlesung "Wissenschaftstheorie, Berufsethik und Geschichte der Psychologie" verwiesen. Das Buch von Schuler (1980) gibt eine Übersicht über "Ethische Probleme psychologischer Forschung" und maßgebliche Ethik-Konventionen. In der Vorlesung werden vor allem ethische Probleme der psychologischen Praxis behandelt.

Literatur

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science,* **5***,* 127-134.
Schuler, H. (1980). *Ethische Probleme psychologischer Forschung*. Göttingen: Hogrefe.

5.6 Verhaltensrichtlinien zur Verhinderung wissenschaftlicher Fälschungen

Die Medizinische Fakultät der Universität Freiburg hat sich aufgrund des Fälschungsskandals Herrmann/Brach entschlossen, Empfehlungen für Verhaltensrichtlinien zu entwickeln, die zu einer Verhinderung wissenschaftlicher Fälschungen beitragen sollen. Die hierzu eingesetzte Kommission der Medizinischen Fakultät hat die nachfolgend zusammengefaßten Vorschläge erarbeitet, die auch für die meisten anderen Fakultäten oder Forschungsinstitute gelten sollen (Uni Aktuell, 7, 1998).

Wesentliche Schwerpunkte des bisherigen wissenschaftlichen Arbeitens – die Organisationsform wissenschaftlicher Arbeitsgruppen, die Qualitätssicherung und Dokumentationssicherheit, Konfliktlösungsstrategien, das Problem der Koautorenschaft und die Verhaltensregeln bei Verdacht auf wiss. Fehlverhalten – wurden von der Kommission aufgegriffen und in Verhaltensrichtlinien umgesetzt. Nach Ansicht der Kommission ist die Redlichkeit des Wissenschaftlers Grundvoraussetzung für wissenschaftliche Arbeit, die durch kein Regelwerk ersetzt werden kann. Die wichtigsten Reformvorschläge sind im folgenden geringfügig gekürzt und für das Fach Psychologie adaptiert worden:

1. Gestaltung wissenschaftlicher Arbeitsgruppen
Um eine Arbeitsgruppe effizient leiten und betreuen zu können, sollte die Gruppe eine bestimmte Größe nicht überschreiten. In der Regel sollte sie aus einem habilitierten oder vergleichbar qualifizierten Gruppenleiter, ein bis drei promovierten Wissenschaftlern, ein bis drei

Doktoranden oder Diplomanden und ein bis zwei technischen Assistenten bestehen. An größeren Einrichtungen kann eine *Abteilung* aus mehreren Arbeitsgruppen bestehen. Der Arbeitsgruppenleiter definiert die Forschungsschwerpunkte, garantiert die Betreuung der Diplomanden und Doktoranden und genehmigt die Freigabe von Ergebnissen zur Veröffentlichung.

2. Qualitätssicherung
Sämtliche wissenschaftlichen Untersuchungen der Arbeitsgruppe sind vollständig zu protokollieren. Die Protokolle haben Dokumentcharakter und sind mindestens 10 Jahre aufzubewahren. Andere Unterlagen wie Datenausdrucke und Filme sollten genau gekennzeichnet und z. B. chronologisch abgelegt werden. Auch diese Dokumentationen sollten mindestens 10 Jahre aufbewahrt werden. Zur Publikation vorgesehene Untersuchungen sollten allen Mitgliedern der Arbeitsgruppe vorgestellt werden. Manuskripte sollten von Mitgliedern der eigenen Arbeitsgruppe, aber auch anderer Arbeitsgruppen kritisch durchgelesen werden. Die Angemessenheit der einzusetzenden statistischen Verfahren muß vor Beginn der Untersuchung kompetent geprüft werden. Bei ethischen Fragen gelten die Weisungen und Empfehlungen der lokalen Ethikkommission.

3. Empfehlung zur Konfliktlösung
Bei Konflikten innerhalb der Arbeitsgruppe ist zunächst der Arbeitsgruppenleiter für deren Lösung zuständig. Doktoranden sollten den Promotionsbeauftragten aufsuchen. Die Fakultät benennt einen "Ombudsman" für Konfliktfälle.

4. Autorenschaft bei wissenschaftlichen Publikationen
Nur wer wesentlich zur Fragestellung, zum Forschungsplan, zur Durchführung der Forschungsarbeiten, zur Auswertung der Ergebnisse und zum Entwurf des Manuskripts beigetragen hat, kann bei einem wissenschaftlichen Bericht Autor sein. Die Finanzierung der Untersuchungen und die Leitung der Abteilung begründen eine Autorenschaft nicht. Auf einem Formblatt, das beim Abteilungsleiter zu hinterlegen ist, wird bei der Einreichung eines Manuskripts der Anteil der einzelnen Autoren an der Arbeit ausgewiesen

5. Verfahren bei Verdacht auf wissenschaftliches Fehlverhalten
Der Abteilungsleiter und in Ausnahmefällen der Dekan sind zu informieren. Innerhalb einer vorgegebenen Frist hat der Betroffene Gelegenheit zu einer Stellungnahme. Abteilungsleiter und Dekan entscheiden darüber, ob ein förmliches Untersuchungsverfahren eingeleitet wird. Hierfür ist ein ständiger Untersuchungsausschuß der Fakultät zuständig, der aus dem Vorsitzenden, dem Dekan, drei weiteren Fakultätsmitgliedern und einem Mitglied, das nicht der Fakultät angehört, besteht. Gegebenenfalls können Fachgutachter berufen werden. Bei nachgewiesenem Fehlverhalten sind die entsprechenden Publikationen zurückzuziehen und die Kooperationspartner zu informieren.

Anhänge zum Teil A

1. Aufbau und Darstellung einer empirischen Untersuchung Allgemeine Hinweise zur Abfassung eines wissenschaftlichen Berichts

Bei der Abfassung eines wissenschaftlichen Arbeitsberichts sollen bestimmte Richtlinien eingehalten werden. Hinweise zur Abfassung wissenschaftlicher Arbeiten sind u.a. folgenden Publikationen zu entnehmen: Hager und Spies (1991), Publication Manual der American Psychological Association (1994), Richtlinien zur Manuskriptgestaltung der Deutschen Gesellschaft für Psychologie (1997). Der folgende kurze Leitfaden lehnt sich an Sarris (1992) an und wurde in einigen Punkten erweitert.

Wissenschaftliche Untersuchungen dienen dazu, explizit formulierte Hypothesen empirisch zu prüfen. Die präzise Formulierung muß **vor** der Überprüfung (Datenerhebung) erfolgen und von dieser unabhängig sein. Aus Gründen der Aufrichtigkeit bei wissenschaftlichen Arbeiten ist daher zu fordern: Nach Formulierung sollen die Hypothesen **"in den Panzerschrank"**. Anschließend erfolgt die Entscheidung anhand der empirischen Daten.

Nachdrücklich sei hier betont, daß unter Fragestellung nicht irgend eine vage formulierte Vermutung verstanden werden soll, sondern eine **präzise Frage**, die nach einem möglichst genauen Überblick über den Problemstand und gegebenenfalls Darstellung relevanter Theorien anhand der einschlägigen Literatur formuliert worden ist. Eine empirische Untersuchung ist grundsätzlich nur vor ihrem **theoretischen Hintergrund** zu beurteilen. Die genaue Darstellung dieses theoretischen Hintergrunds ist deshalb unentbehrlich.

Bestandteile eines Forschungsberichts
Ein Forschungsbericht setzt sich nach diesen Regeln aus verschiedenen Abschnitten zusammen. Die wesentlichen Abschnitte eines Berichts sind:

1. Zusammenfassung
2. Einleitung
3. Methode
4. Ergebnisse
5. Diskussion
6. Literaturverzeichnis

Zur Untergliederung wird empfohlen, **nicht mehr als 3 Ebenen** (sog. Dezimalklassifikation) zu verwenden, d.h. Kapitel 2, mit Abschnitten 2.1, 2.2 usw. sowie Unterabschnitten 2.1.1, 2.1.2 usw.

Die **Überschrift** (Titel) soll – in knapper Form – das allgemeine Problem der Untersuchung erkennen lassen. Sie soll zusätzlich, soweit möglich, die besondere Untersuchungsmethode anzeigen.

1. Zusammenfassung (Abstract)
Einem wissenschaftlichen Bericht wird stets eine prägnante Zusammenfassung vorangestellt. Sie informiert **in aller Kürze** (d.h. auf ca. 1 Seite) über die wesentlichen Inhalte aller Einzelabschnitte des Berichts:

- Darstellung des allgemeinen Problems
- Beschreibung der Untersuchungsteilnehmer (Stichprobe)
- Angaben zur Versuchsanordnung

- Hinweise zum Versuchsablauf
- Benennung des zentralen Ergebnisses
- Andeutung der wichtigsten Diskussionspunkte

2. Einleitung

Sie soll das allgemeine Problem in einem weiteren Rahmen darstellen, aus dem sich die speziellen Hypothesen und Methoden der Untersuchung stringent ableiten lassen. Bei der Abfassung der Einleitung empfiehlt sich die Abfolge:

- Einbettung des Problems, d.h. des Themas bzw. der allgemeinen Fragestellung – im Unterschied zu den speziellen Hypothesen – in den übergreifenden Zusammenhang,
- Hinweis auf die Theorien, die für das Problem relevant sind,
- Beschreibung derjenigen Untersuchungen, an die sich die Fragestellung anlehnt, d.h. Benennung des methodischen Vorgehens, der Ergebnisse und der zentralen Schlußfolgerungen in gestraffter Form,
- Kurzzusammenfassung des gegenwärtigen Wissensstands in dem jeweiligen Problembereich,
- Darstellung von Anknüpfungspunkten der eigenen Fragestellung,
- Formulierung der spezielle(n) Versuchshypothese(n), am besten in Form eines hypothetischen "experimentellen Satzes", welcher unter Verwendung von geeigneten operational definierten Termini aus einem "theoretischen Satz" hergeleitet wird.

Die Hypothesen können jedoch nur formuliert werden, wenn wichtige Entscheidungen gefallen sind:

- Design (Versuchsplan) und
- Operationalisierung der Variablen.

Da wichtige Einzelheiten erst im folgenden Methodenkapitel genau beschrieben werden, muß hier etwas vorgegriffen und mit Hinweisen ("siehe Abschnitt x.x.x)" gearbeitet werden. Die Auffassungen sind geteilt, ob das Design nicht erst unter Methoden behandelt werden soll, oder bereits in die Einleitung gehört (siehe APA-Richtlinien). Hier wird empfohlen, das Design bereits am Ende der Einleitung zu nennen und kurz zu erläutern sowie UV, AV, ggf. Kovariate und Kontrollvariablen zu definieren. Eine ausführliche Darstellung, die Begründung des Designs und aller Auswahlentscheidungen sowie die genauen Operationalisierungen folgen später unter "Methoden".

Zunächst sollte festgelegt werden, ob es sich um einen Within-Subjects-Plan, einen Between-Subjects-Plan oder um ein Mischdesign handelt. Danach folgt eine Beschreibung:

- Mehrgruppenversuchsplan (Zufallsgruppenbildung/Between-Subject-Design): unifaktorielles Design ohne Vorhermessung/mit Vorhermessung (SOLOMON-Vier-Gruppenplan) etc.
- Versuchsplan mit Meßwiederholung (Within-Subject-Design): unifaktoriell oder multifaktoriell?
- Versuchsplan mit Blockbildung ("matched samples")
- Gemischter Versuchsplan (Zufallsgruppenfaktoren, Meßwiederholungsfaktor): zwei- oder mehrfaktoriell?

Welche Variablen werden für diesen Versuchsplan als UV, AV, KV (Kovariate, Kontrollvariablen) festgelegt?

Am Ende einer Einleitung müssen die Versuchshypothesen in Form eines oder mehrerer **experimenteller Sätze** (Operationalisierung der UV und der AV) aufgestellt sein, welche auf den früheren theoretischen und/oder empirischen Befunden basieren.

3. Methoden

In diesem Abschnitt müssen alle relevanten Verfahrensdetails beschrieben werden: die **Angaben sollen ausreichen, um den Versuch in der ursprünglichen Weise wiederholen zu können**.

Vorgeschrieben nach APA sind drei grundlegende Abschnitte des Methodenteils:

- Stichprobenauswahl und – beschreibung
- Beschreibung des Versuchsmaterials und Versuchsaufbaus
- Versuchsablauf

Für die meisten Untersuchungen sind jedoch zusätzliche Informationen unabdingbar (Begründung des Versuchsdesign, Operationalisierung der UV und der AV (Realisierung des Max-Min-Kon-Prinzips) mit Überlegungen zur internen und externen Validität. Bei einem Praktikumsbericht sollte darauf geachtet werden, **alle Auswahlentscheidungen** gut zu begründen.

3.1 Teilnehmer

Hier ist die Auswahl der Teilnehmer an der Untersuchung zu beschreiben. Bei vielen empirischen Untersuchungen wird es nicht möglich sein, Zufallsstichproben einer Population auszuwählen, so daß dieser **stichprobentechnische** Aspekt entfallen muß. Die Auswahl der Untersuchungsteilnehmer muß jedoch sehr genau beschrieben werden, damit die Leser eine Möglichkeit haben, die speziellen Einschränkungen der Repräsentativität zu beurteilen.

(1) Grundlage der Auswahl der Teilnehmer (Vpn, Probanden, Patienten) ist die Population, auf die mit statistischen Aussagen geschlossen werden soll, d.h. die Teilnehmer müssen eine gewünschte Population repräsentieren (Repräsentativität der Stichprobe). Diese Population muß charakterisiert werden. Auf dieser Grundlage sind Ein- und Ausschluß-Kriterien festzulegen. Die Erfassung solcher Kriterien, u.a. durch Fragebögen oder Tests, ist zu beschreiben. Genaue Angaben sind zur Vpn-Quelle (z.B. Universität oder Nervenklinik) zu machen. Unterlagen über die Anwerbung der Vpn/Patienten, wie Register, Fragebögen, Tests, mündliche und schriftliche Einladungen sowie Zeitungsanzeigen usw. sollten dem Anhang beigefügt werden.
(2) Entsprechend der Ein- und Ausschlußkriterien und – je nach Fragestellung – zur Kontrolle potentieller Einflußfaktoren können z.B. folgende Merkmale erfaßt werden: Geschlecht, Alter (Mittelwert und Spannweite), sozioökonomischer Status, Größe, Gewicht (Übergewicht), körperliche und psychische Krankheiten, Händigkeit, Sehtüchtigkeit, IQ usw. sowie Teilnahmebedingungen (freiwillig, bezahlt, Pflicht usw.).
(3) Zur Festlegung der Stichprobengröße gibt es Konventionen (siehe Abschnitt 3.3 des Skripts) und, z.B. für das Experimentalpraktikum, reduzierte Erwartungen.

Eine grundlegende Anforderung, die in den neueren Richtlinien zunehmend betont wird, ist die **Sicherstellung der ethischen Standards** im Umgang mit den Teilnehmern einer Untersuchung.
Folgende ethische Standards (APA, 1992) müssen erfüllt werden:

(1) Einschätzung der ethischen Akzeptanz der Untersuchung. Eine Untersuchung ist ethisch akzeptabel, wenn alle weiteren Kriterien (2)-(7) erfüllt werden können. Ist dies nicht der Fall, müssen weitere Absicherungen vorgenommen werden.
(2) Von größter Bedeutung ist die Einschätzung des physischen und psychischen Risikos für die Teilnehmer. Dies kann im allgemeinen aufgrund schon existierender vergleichbarer Studien ermittelt werden.

(3) Absicherung einer ethischen Durchführung durch **mehrere Verantwortliche**. Eine Absicherung der ethischen Durchführung wird durch eine entsprechende Instruktion oder Kontrolle mehrerer Verantwortlicher realisiert.
(4) Erhalt einer eindeutigen, freiwilligen, schriftlichen **Einwilligung der Probanden** nach Vorab-Information und Versicherung der **Rechte der Vpn wie Datenschutz, Recht zur Ablehnung, Unterbrechung oder zum Abbruch des Versuchs**. Die freiwillige schriftliche Einwilligung der Vpn muß mindestens folgende Information erhalten: Beschreibung und Zweck der Studie, Art der zu erhebenden Daten, Versuchsdauer und Anforderungen, mögliche Risiken oder Vorteile für die Vpn, Freiwilligkeit und Recht auf Ablehnung, Unterbrechung oder Abbruch des Versuchs zu jeder Zeit ohne Nachteile, Vertraulichkeit der Daten (Kodierung der Namen, sichere Aufbewahrung), Name und Tel.-Nr. einer Kontaktperson (zum Informationsbezug) sowie einer zweiten Person zur Mitteilung von versuchsbedingten Beschwerden. Unterschrift vom Vl und der Vpn (mit Datum). Eine Teilnahme legal nicht unterzeichnungsfähiger Personen verlangt eine Einwilligung der Eltern oder eines gesetzlichen Vertreters als auch die Zustimmung des Kindes (Jugendlichen). Die Einwilligung sollte aktiv und nicht passiv (z.B. durch Unterlassen einer elterlichen Absage) vollzogen werden. Die passive Taktik bringt möglicherweise mehr Vpn ein, ist aber ethisch nicht vertretbar, vor allem wenn es sich dabei um Kinder handelt.
(5) **Vermeidung unnötiger und ungerechtfertigter Täuschungen der Vpn**. Problematisch ist die Vermeidung von Täuschung der Vpn bzw. Geheimhaltung von Information, die oft zur Vermeidung von Vpn-Effekten verwendet wird (Blindversuch). Kann dieses Kriterium nicht erfüllt werden, muß zur ethischen Absicherung die Notwendigkeit oder Unabdingbarkeit der Taktik für den Versuch gerechtfertigt werden
(6) Korrektur unerwünschter, aber unvermeidbarer Konsequenzen der Teilnahme. Unerwünschte Konsequenzen der Teilnahme (z.B. verklebte Hände und Haare, Ermüdung usw.) müssen korrigiert werden (z.B. durch Bereitstellung von Waschmöglichkeiten, Gewährung von Pausen usw.)
(7) **Aufklärung der Vpn nach Erhebung der Daten**. Mitteilung der wesentlichen Details des Versuchs sowie Aufdeckung und angemessene Entschuldigung (Rechtfertigung) von Täuschungen der Vpn.

3.2 Versuchsaufbau

Setting und Untersuchungsinstrumente

Die Auswahl des Settings und des Versuchsmaterials sollte auf Überlegungen zur Operationalisierungen der theoretischen Konstrukte beruhen und in Bezug auf die interne und externe Validität (z.B. Replizierbarkeit) des Versuchsplan optimal sein. Die Beschreibung des Settings sollte Angaben über Ort und Zeit bzw. Dauer eines Versuchs, zeitliche Abstände zwischen den Versuchen, Gesamtdauer der Erhebung (evtl. erst unter "Versuchsablauf"), eine detaillierte Beschreibung und eventuell Skizze der Untersuchungsräume mit der Anordnung der Möbel, Fenster, Türen, der Beleuchtungsverhältnisse usw. sowie der Position von Versuchsinstrumenten, Versuchsleiter und den im Raum befindlichen Versuchspersonen enthalten. Die Untersuchungsinstrumente bzw. Tests müssen detailliert bezüglich Größe, Marke und Modell bzw. Form, Aufbau und Publikation beschrieben werden. Ist eine Bedienung der Instrumente durch den Versuchsleiter oder die Vpn nötig oder muß sich die Vpn bestimmten Aufgaben oder Tests unterziehen, müssen genaue Instruktionen formuliert werden (Anhang).

Realisierung der uV im Experiment

Die konsequente und plangemäße Realisierung (Manipulation) der uV's kann gewährleistet werden durch Fremd- und Eigenkontrolle der Versuchsleiter sowie durch Befragung der Vpn, falls Instruktionen oder Meinungen manipuliert werden (sind die Vpn tatsächlich von den Manipulationen bzw. der vorgegebenen Information überzeugt?). Der/die Vl sollte über die experimentellen Hypothesen nicht Bescheid wissen, ausgenommen es gibt eine wirksame

Methode zur Selbsttäuschung (Blindversuch). Kennt der/die Vl die Hypothesen, empfehlen sich zur Kontrolle eines möglichen Vl-Effekts eine Nachbefragung der Vpn über das Verhalten des Vl oder Tonbandaufnahmen des Versuchs, die dann von einem Uneingeweihten ausgewertet werden. Verlangt die Manipulation der aV's bestimmte Fähigkeiten oder komplizierte Ausführungen, muß der/die Vl dafür ausgebildet und seine Kompetenz vor und während des Versuchs überprüft werden (s. oben). Kontrolle potentieller Störfaktoren durch Eliminierung (Randomisierung, Parallelisierung, Konstanthaltung), konsequente Erfassung (Fragebögen, Vortests, Protokoll) und Umwandlung zu weiteren aV's.

Adäquate Darstellung der aV
Die aVs sind ausreichend zu beschreiben und zu begründen. Beobachtungsmethoden, Tests, Fragebögen oder Instrumente zur Erfassung der aV sollten benannt und nach Autor/Hersteller, Nummer und Datum der Publikation/Herstellung sowie Form/Aufbau beschrieben werden (evtl. erst unter Versuchsmaterial und -aufbau). Eine Begründung des Meßverfahrens bezüglich psychometrischer Angemessenheit (Konstruktvalidität), Reliabilität und interner Validität ist erforderlich. Darlegung von Antwort-/ Reaktionsmöglichkeiten bzw. Festlegung des Wertebereichs.
Beschreibung der Auswertung/ Exploration der Daten (Überführung der Messungen in statistisch verwendbare Daten). Codes, Richtlinien, Auswertungsbögen und -tabellen gehören in den Anhang.

3.3 Versuchsablauf

Die Anwerbung bzw. Einladung: telefonisch, schriftlich, durch Anzeige, Zusendung von Fragebögen und Information (evtl. unter "Teilnehmer") sollte genau beschrieben werden. Kompetenzen von Vl und Assistenten sind abzustecken: Wer ist für die Vpn („Pflege“ der Vpn) bzw. die einzelnen Versuchsabschnitte zuständig?
Das Vorgehen beim Eintreffen der Vpn am Versuchsort ist festzulegen: Begrüßung, Vorab-Information (Allgemeines zur Studie, Versicherung der Rechte der Vpn), schriftliche Einwilligung der Vpn, Instruktion, Adaption (Prätest), Versuchsablauf (Kontrolle von Störfaktoren, Vermeidung von Vpn-Effekten wie z.B. psychischer Reaktanzphänomene und Vl-Effekten durch Blind- und Doppelblindversuche (Nachbefragung der Vpn (Kontrolle des Vl-Verhaltens, abschließend Aufklärung der Vpn (detaillierte Information über Absichten und Möglichkeit, die Ergebnisse zu erfahren), Nachbefragung (zur Aufdeckung von Vpn und Vl-Effekten), Verabschiedung.

3.4 Statistische Analysenkonzepte und Datenanalyse

Am Ende des Abschnitts zur Methode wird – auf der Basis der genauen Beschreibung von Teilnehmerauswahl, Versuchsaufbau und Versuchsablauf noch einmal das Design angesprochen und jetzt auch das statische Analysenkonzept präzisiert. Wie soll der Versuchsplan statistisch ausgewertet werden?

Mit der Formulierung von **Null- und Alternativhypothese** wird die theoretische Fragestellung und deren empirische Formulierung in eine statistisch entscheidbare Form gebracht. Null- und Alternativhypothese sind Formalismen im Rahmen des statistischen Entscheidungsmodells und sollten daher entsprechend formal dargestellt werden.
Im Rahmen der Festlegung von Null- und Alternativhypothese sollte über eine **ein- oder zweiteilige Fragestellung** und das **statistische Entscheidungsniveau** entschieden werden.

Zur praktischen Datenanalyse werden hier Hinweise erwartet, wie die Datenkontrolle durchgeführt werden soll: Anzahl fehlender Daten, Detektion und Ausschluß von Ausreißerwerten; Ausschluß von Teilnehmern mit genauen Angaben über die Gründe.

In einem zusätzlichen Abschnitt über (primäre) Datenanalyse ist festzuhalten, welche Auswertungsschritte bei der Datenerhebung vollzogen werden, z.B. Berechnung von Testwerten, Indizes, Summenwerten, bevor die (sekundäre) statistische Hypothesenprüfung beginnt. Die Auswertung der Daten ist vollständig, ggf. mit den verwendeten Strategien und Regeln mitzuteilen.

An dieser Stelle ist zweckmäßig zwischen dem **statistischen Entscheidungsmodell, d.h. der inferenzstatistischen Prüfung** (z.B. auf Unterschieds- oder Korrelations-Hypothesen), und zusätzlichen statistischen Analysen zu unterscheiden. Solche zusätzlichen Analysen sind:

vorausgehende Analysen, z.B. eine Itemanalyse an einem neuen Fragebogen oder eine Verteilungsanalyse der aV, um Decken- oder Bodeneffekte zu erkennen, oder **explorative (post hoc) Analysen**, in denen z.B. Beziehungen zwischen Variablen näher betrachtet und beschrieben werden, ohne daß es in speziellen Hypothesen angekündigt war.

In diesen Abschnitt gehören schließlich die Angaben zum Rechengang und die Angabe der verwendeten Computerprogramme, z.B. aus dem SPSS oder SAS (auch im Literaturverzeichnis).

4. Ergebnisse

Im Ergebnisteil werden die statistisch verarbeiteten Hauptbefunde im Hinblick auf die der Einleitung zu entnehmende(n) Fragestellung(en) bzw. Hypothese(n) dargestellt. Die individuellen Rohdaten eines Praktikumsversuchs sollten in einem Anhang (oder eventuell auf Diskette) beigefügt werden. Das Ergebniskapitel muß eine überschaubare Darstellung der Befunde enthalten, die es dem Leser ermöglicht, zu einer genauen Bewertung zu gelangen.

- Schrittweise Beschreibung der Ergebnisse, damit sie beim Lesen nachvollzogen werden können.
- Wahl einer übersichtlichen Darstellungsweise mit Tabellen und Abbildungen, wenn Verständnis und Übersichtlichkeit der Hauptbefunde dadurch gefördert werden.
- Tabellen und Abbildungen müssen ausreichend beschriftet werden, d.h. mit einer Legende versehen sein; sie sollten möglichst ohne Lektüre des Haupttexts verstanden werden können.
- Darstellung der Daten im Hinblick auf die in der Einleitung formulierten Hypothesen.
- Mitteilung aller relevanter Zahlen zur Beurteilung des Ergebnisses (z.B. Prüfgrössen, Freiheitsgrade usw.) – insbesondere sollten auch die für die statistische Entscheidung relevanten kritischen Werte mitgeteilt werden. Die Darstellung in Tabellen ist – wo möglich – zu empfehlen.
- Möglichst objektive Beschreibung der Daten und Vermeidung einer eigenen Bewertung.

Hinweise zu Tabellen und Abbildungen (Graphiken)

Die Tabellen sind fortlaufend zu numerieren, ebenso die Abbildungen. Bei **längeren** Arbeiten ist es zweckmäßig, die Nummern in jedem Kapitel neu zu beginnen (Tabelle 5.3 wäre z.B. die dritte Tabelle im fünften Kapitel).

Die Publikations-Richtlinien enthalten sehr genaue Vorschriften, wie Tabellen und Abbildungen aussehen müssen, z.B. stehen Nr. und Titel bei einer Tabelle im Kopf, bei einer Abbildung im Fuß; die Gliederung, Schrift, Striche usw. sind festgelegt. Am besten wird ein neueres Lehrbuch der Psychologie als Vorbild gewählt, um Tabellen und Abbildungen anzufertigen.

Genauso wie in dem Methodenteil eines Untersuchungsberichts sollte auch in dem Ergebnisteil eine nüchterne, aber prägnante Darstellungsweise angestrebt werden, welche die Unvoreingenommenheit (Sachlichkeit, Neutralität) des Verfassers erkennen läßt.

5. Diskussion

Enthält die Interpretation der Ergebnisse und verdeutlicht, worin der besondere wissenschaftliche Beitrag seiner Untersuchung besteht.

- Möglichst klare Aussage darüber, ob die Hauptbefunde für oder gegen die Versuchshypothesen sprechen.
- Gegebenenfalls Darlegung möglicher Gründe dafür, daß die Ergebnisse die Hypothesen nicht bzw. nur ansatzweise, d.h. tendenziell bestätigen.
- Vergleich der Ergebnisse mit denen anderer Untersuchungen.
- Formulierung eines psychologisch sinnvollen Erklärungsansatzes für die Hauptbefunde.
- Diskussion der Ergebnisse im Hinblick auf alternierende oder konkurrierende Erklärungsversuche. Welche neuen theoretischen Aspekte ergaben sich?
- Vorschläge für etwaige Verbesserungen der Versuchsanordnung im Falle einer nochmaligen Durchführung des Experiments.
- Vorschläge für weitere wichtig erscheinende Untersuchungsansätze, insoweit diese in einem konkreten und unmittelbaren Zusammenhang mit dem eigenen Experiment gesehen werden können.

Bei der Abfassung der "Diskussion" zu einem Bericht ist selbstverständlich, daß im Prinzip keine endgültigen Aussagen bezüglich komplexer psychologischer Fragestellungen zu erreichen sind: Im besten Fall führt eine gut durchdachtes und solide durchgeführtes Experiment zu relativ gut gesicherten singulären Aussagen, welche ihrerseits zu neuen Fragestellungen Anlaß geben.

Wesentliche Aspekte der rückblickenden und selbstkritischen Überlegungen sind

- die Diskussion der **inneren Validität**
- die Diskussion der **externen Validität**

In wieweit konnten die Absichten der Untersuchung verwirklicht werden und welche Vorbehalte sind zu machen: Die Gesichtspunkte THIS MESS und MAXMINCON können die Gliederung dieser Abschnitte leiten.

Selbstkritische Einschränkungen zu den Grenzen der eigenen Aussagen sind durchaus erwünscht. Eine Skizze der sich neu ergebenen bzw. unbeantworteten Fragen führt über die eigene Untersuchung hinaus. Anregungen für künftige Untersuchungen in inhaltlicher und methodischer Hinsicht runden die Diskussion ab.

6. Literaturverzeichnis

Ein wissenschaftlicher Text bezieht sich in der Regel auf bereits vorliegende Veröffentlichungen. Diese müssen **genau** zitiert werden und im Literaturverzeichnis aufgeführt werden (siehe Richtlinien).

7. Anhang

7.1 Muster bzw. Kopien der verwendeten Tests, Fragebogen, Vorlagen, Liste der Stimuli, Texte usw., falls es sich nicht um Standard-Verfahren handelt.

7.2 Instruktionen und Protokollblätter zur Datenregistrierung.

7.3 Listen der Daten (falls nicht zu umfangreich).

7.4 Kopien von Computer-Listen (Outputs) der wesentlichen statistischen Berechnungen (nur in den relevanten Ausschnitten mit reduziertem Format, d.h. nach Bearbeitung).

Literatur

American Psychological Association (1994). *Publication manual of the American Psychological Association. (4th ed.)* Washington, D.C.: American Psychological Association.

Cone, J.D. & Forster, S. (1993). *Dissertations and theses from start to finish*. Washington DC: APA.

Deutsche Gesellschaft für Psychologie (Hrsg.) (1997). *Richtlinien zur Manuskriptgestaltung*. Göttingen: Hogrefe.

Hager, W. & Spies, K. (1991). Versuchsdurchführung und Versuchsbericht. Göttingen: Hogrefe.

Sarris, V. (1992). *Methodologische Grundlagen der Experimentalpsychologie. 2: Versuchsplanung und Stadien*. München: Ernst Reinhardt Verlag.

2. Richtiges Zitieren

1. Prinzip: Eigene und fremde Ideen/Befunde sollen unterscheidbar sein (Quellenangabe).
Wörtliche Zitate werden in Anführungsstriche gesetzt und mit Autor, Jahreszahl und Seitenzahl angegeben:
"..." (Cattell, 1957, S. xx). Sinngemäße Referate von Ideen/Befunden werden nur mit Autor und Jahreszahl ohne Seitenangabe zitiert.

2. Prinzip: Es sollen nur solche Quellen zitiert werden, die selbst gelesen wurden.
Sog. Sekundärzitate sind – falls nicht zu vermeiden – mit dem Hinweis auf den **gelesenen Autor zu geben**, z. B. Wundt (1874, zitiert nach Pongratz, 1967), d.h. im Literaturverzeichnis erscheint Pongratz, L.J. (1967) *Problemgeschichte der Psychologie*. Bern: Francke Verlag.

3. Prinzip: Die Quellenangaben sollen vollständig sein.
Zu jedem Namen (Erstautor) im Text (ausgenommen die Sekundärzitate siehe 2. Prinzip) wird im Literaturverzeichnis eine Quellenangabe erwartet.

4. Prinzip: Die Quellenangaben sollen eindeutig sein.
Der Leser muß in der Lage sein, die Quelle zu identifizieren. Zu diesem Zweck gibt es spezielle Zitierrichtlinien. Die Zitierrichtlinien sind international nur zum Teil standardisiert. Es gibt u.a. Richtlinien für medizinische Fachliteratur und Richtlinien für psychologische Fachliteratur, außerdem haben einige große Verlage eigene Vorschriften, so daß Manuskripte für bestimmte Publikationszwecke umgeschrieben werden müssen.

5. Prinzip: Für Psychologen/innen sind die "Richtlinien zur Manuskriptgestaltung der Deutschen Gesellschaft für Psychologie" (1997) maßgeblich.
Beispiel für ein **Buch**
Bortz, J. (1989). *Statistik für Sozialwissenschaftler*. (3. Aufl.). Berlin: Springer.

Beispiel für ein **Buchkapitel**
Hager, W. & Westermann, R. (1983). Planung und Auswertung von Experimenten. In J. Bredenkamp & H. Feger (Hrsg.), *Enzyklopädie der Psychologie. Themenbereich B. Methodologie und Methoden. Serie I Forschungsmethoden der Psychologie. Band 5 Hypothesenprüfung* (S. 24-238). Göttingen: Hogrefe.

Beispiel für einen **Zeitschriftenaufsatz**
Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, **56**, 81-105.

Beispiel für einen **Forschungsbericht**
Wilhelm, P. (1993). *Kurze Filmszenen als Stimulusmaterial zur experimentellen Erzeugung der Grundemotionen: Angst, Ärger, Ekel, Trauer, Überraschung und Heiterkeit* (Forschungsbericht Nr. 96). Freiburg i. Br.: Universität, Psychologisches Institut.

Beispiel für eine **Diplomarbeit**
Taxis, S. (1990). *Neuropsychologische Untersuchungen nach Eingriffen am Limbischen System*. Unveröff. Dipl. Arbeit, Albert-Ludwigs-Universität, Freiburg i. Br., Psychologisches Institut.

Die Richtlinien der Deutschen Gesellschaft für Psychologie enthalten viele weitere Hinweise, wie spezielle (auch fremdsprachige) Quellen zu zitieren und anzuordnen sind, außerdem gibt es dort Richtlinien für den

- Aufbau eines Manuskripts,
- Tabellen,
- Abbildungen,
- Literaturverzeichnis.

Deutsche Gesellschaft für Psychologie. (Hrsg.) (1997). *Richtlinien zur Manuskriptgestaltung* (2. Aufl). Göttingen: Hogrefe.
American Psychological Association (1994). *Publication Manual of the American Psychological Assocition (4th ed.)*. Washington, DC: American Psychological Association.

Literaturrecherche

Literatursuche wird heute überwiegend in großen Datenbanken unternommen, doch sind die anderen Wege, zumindest für ältere Publikationen, nicht überflüssig geworden:

- Bibliografie der deutschsprachigen psychologischen Literatur,
- Psychological Abstracts,
- Sammelreferate bestimmter Fachgebiete
- Handbücher, u.a. Enzyklopädie der Psychologie
- Annual Review of Psychology
- Durchsicht relevanter Zeitschriften und Lehrbücher
- Durchsicht des betreffenden Abteils der Institutsbibliothek (siehe aushängende Systematik).

Selbständige Literaturrecherchen können am PC im Zeitschriftenraum der Institutsbibliothek im Peterhof (und von Terminals in der UB) vorgenommen werden:

PSYNDEX	Deutschsprachige psychologische Fachliteratur ab 1977
PSYCLIT	Internationale psychologische Fachliteratur ab 1974
MEDLINE	Medizinische Fachliteratur ab 1966.

3. Folienpräsentation

Die folgenden Hinweise stammen von Pohl, R. (1990). Beobachtungen und Vorschläge zur Gestaltung von Verwendung von Folien in Vorträgen. Psychologische Rundschau, 41, 155-158.

Tabelle 1 **Die häufigsten Fehler bei der Gestaltung und Verwendung von Folien**

Gestaltung von Folien:
A. Die Folie ist nicht lesbar: zu kleine Schrifttype; unleserliche Handschrift; mangelhafter Kontrast.
B. Die Folie enthält zuviele Informationen: zu kompakte Darstellung; überflüssige Informationen.
C. Grafische Gestaltungsmittel werden schlecht genutzt: fehlende "Veranschaulichung"; fehlende Farbe.
D. Der Inhalt der Folie ist unverständlich.

Verwendung von Folien:
E. Die Anzahl der Folien ist nicht angemessen: zuwenige oder gar keine Folien; zuviele Folien.
F. Die Folien passen nicht zum Vortrag: fehlende Abstimmung auf die jeweilige Situation; mangelhafte Koordinierung zwischen Folie und Vortrag.
G. Der Inhalt der Folie wird schlecht erläutert: zu kurze Darbietungszeit; verdeckte Sicht; fehlende Erklärungen; der "nervöse Finger".

Tabelle 2 **Vorschläge zur Gestaltung und Verwendung von Folien**

Gestaltung von Folien:
1. Die Schrifthöhe sollte mindestens 6 mm betragen.
2. Handschrift sollte nur ausnahmsweise und mit großer Sorgfalt verwendet werden.
3. Der Kontrast sollte maximal sein.
4. Das Layout sollte einfach und überschaubar sein.
5. Folien sollten keine überflüssigen Informationen enthalten.
6. Folien sollten "bildhaft" gestaltet sein.
7. Farbe sollte gezielt verwendet werden.
8. Folien-Inhalte sollten leicht zu verstehen sein.

Verwendung von Folien:
9. Folien sollten in ausreichendem Maße verwendet werden.
10. Die maximale Folienzahl ergibt sich aus Vortragszeit (in Minuten) geteilt durch drei. Für die Darstellung eines Experimentes sollten 7 + 2 Folien verwendet werden.
11. Vortragstext und Folien-Präsentation sollten aufeinander abgestimmt sein.
12. Folien sollten der jeweiligen Situation angepaßt und untereinander kohärent sein.
13. Folien sollten genügend lange gezeigt werden.
14. Die Sicht auf die Projektionsfläche sollte nicht verdeckt werden.
15. Folien sollten vollständig erläutert werden.
16. Besprochene Folien-Inhalte sollten markiert werden.

4. Wie gestalte ich ein gutes Poster?

Allgemeine Informationen. Grundsätzlich sollte ein Poster in sich vollständig sein und für sich selbst sprechen. Der Leser sollte sich selbständig, d.h. ohne Erläuterunen in dem Poster zurechtfinden. Der Posterautor sollte hier bei Fragen den Text erläutern. Eine Posterpräsentation erlaubt Information in kleinerem Rahmen zu diskutieren als beispielsweise ein Diavortrag. Eine Diskussion wird schwierig, wenn die Präsentation, bzw. das Poster unverständlich ist. Ein Poster soll auf einen Blick zu fassen bzw. verständlich sein, dann besteht mehr Raum für Diskussion. Beachte: der Betrachter entscheidet, wie lange er sich mit dem Poster befasst und nicht der Autor des Posters. Vor der Tagung ist zu erledigen:

1. **Planung des Posters**. Größe des Posters ca. 1,75 x 1,1 m. Für die effektive Nutzung dieses Raumes muß die Gestaltung von Abbildungen und Text nach einem strukturierten Plan erfolgen.
   - das Material sollte in Spalten angeordnet werden
   - höchstens 5 Spalten mit 29 cm breitem A4 Papier
   - Oben links mit der Einleitung beginnen, unten rechts die Schlußfolgerungen
2. **Illustrationen**. Illustrationen und Grafiken sollten auch aus der Entfernung gut sichtbar sein. Die Hauptaussagen sollten auf einen Blick erkennbar sein. Keine Grafik ohne fettgedruckte Überschrift! Die Illustration sollte mit einem Kommentar versehen werden, der auch das Ergebnis („Take home message“) beschreiben sollte. Methodische Details und Resultate (die normalerweise im Ergebnisteil oder den Schlußfolgerungen eines Aufsatzes enthalten sind), sollten notfalls kurz am Ende der Legende erläutert werden.
3. **Text**. Kurz halten und große Buchstaben sowie angemessene Zeilenabstände verwenden. Linksbündiger Textfluß, d.h. kein Blocksatz. Ganze Textpassagen sollte nicht in Groß- oder Fettbuchstaben geschrieben sein. Spiegelstriche sind für das Verständnis günstig.
4. **Titel**. Der Postertitel soll den wesentlichen Inhalt der Untersuchung sowie Autoren und Institutionszugehörigkeit (eventuell auch die Posternummer) enthalten.
5. **Layout**. Der Text kann auf farbigem Posterpapier montiert werden. Inhaltlich ähnliche Textteile könnten auf mit gleichem Hintergrund farblich unterlegt werden. Grafik nummerieren, „Abbildung xx“ ist aus Platzgründen aber nicht erforderlich.
6. Während der Tagung ist zu beachten: Fünfzehn Minuten vor dem Beginn Ihrer Sitzung sollte das Poster am vorgesehenen Platz montiert werden und dort bis zum Ende der Sitzung bleiben. Anschließend Poster entfernen. Material zur Montage eventuell mitbringen.

Quellenhinweis: Kongreßankündigung der **Society for Neuroscience**.

5. Wie gestalte ich einen effektiven Diavortrag?

Es ist allgemein bekannt, daß die Qualität von wissenschaftlichen Diavorträgen sehr unterschiedlich ausgeprägt ist. Wenig effektive Dias können dazu führen, daß die Aussagekraft des Vortrags erheblich beeinträchtigt wird. Um die Effektivität von Vorträgen zu verbessern, sollte die folgende Liste zur Gestaltung effektiver Dias beherzigt werden.

1. **Klare Zielsetzung**. Der Vortrag und jedes einzelne Dia sollte ein klares Thema und eine klare Aussage haben. Falls Zweifel bestehen, ob das Thema klar erfaßt werden kann, sollte es weggelassen werden.
2. **Unmittelbare Verständlichkeit**. Die Hauptaussage des Dias sollte sofort die Aufmerksamkeit der Zuhörer auf sich lenken und unmittelbar verständlich sein, damit sich die Zuhörer auf die Botschaft des Referenten/in konzentrieren können.
3. **Einfacher Aufbau.** Das Dia sollte einfach strukturiert aufgebaut sein, damit die Hauptaussage klar im Mittelpunkt der Aufmerksamkeit steht.
4. **Keine Präsentation unwichtiger Information**. Information, die nicht die Kernthese des Vortrags stützt, sollte im Vortrag außer acht gelassen werden und eventuell für spätere Fragen zur Verfügung stehen.
5. **Berücksichtigung der Aufnahmefähigkeit der Zuhörer.** Der Referent sollte berücksichtigen, daß es ein Limit an neuen Informationen gibt, welches das Publikum aufnehmen kann. Ansonsten riskiert der Zuhörer, daß das Publikum verwirrt wird. In einem zehnminütigen Vortrag sollte pro Dia ungefähr eine Minute referiert werden, ideal sind etwa 7 Dias. Außerdem wäre es ratsam, langsam und deutlich das jeweilige Thema vorzutragen.
6. **Vereinheitlichung und Abrundung des einzelnen Dias.** Für jedes einzelne Dia sollte das jeweilige Thema gebündelt und abgerundet dargestellt werden. Ein Diavortrag ist dann am wirkungsvollsten, wenn alle Information um ein einziges Zentralthema organisiert ist, so daß der Diavortrag eine vereinheitlichte Geschichte („Story“) erzählt.
7. **Graphisches Format.** In graphischen Darstellungen sind die qualitativen Beziehungen ausdrücklich auf Kosten von genauen numerischen Werten hervorzuheben, während für Tabellen das Gegenteil richtig ist. Qualitative Beziehungen sollten der Anschaulichkeit wegen in graphischer Form dargestellt werden (genauere tabellarische Tafeln können für weitergehende Fragen bereitgehalten werden).
8. **Vorbereitung des Materials für die mündliche Präsentation**. Umfangreiche Tabellen mit zahlreichen Reihen und Spalten sind zu vermeiden. Das Publikum möchte nicht hören, wieviel Arbeit man sich gemacht hat, sondern die Befunde und Schlußfolgerungen zum zentralen Thema des Vortrags hören.
9. **Material und experimentelle Techniken**. In einem 10 min Vortrag können keine Standardtechniken vorgestellt werden. Wenn der Vortrag nicht methodenbezogen ist, sollte ein Minimum der zur Verfügung stehenden Zeit hierauf verwendet werden. Der methodische Teil sollte so kurz wie möglich und so lang wie zum klaren Verständnis nötig gehalten werden.
10. **Lesbarkeit der Illustrationen**. Der Text sollte gut lesbar sein: speziell Zeilen und Schriftgröße (42 Zeichen breit, mind. 14 pt Schrift); gut lesbare Schriftarten, Großbuchstaben, Ränder beachten; Tabellen sollten klar und deutlich beschriftet sein; der Projektor sollte weit hinten aufgestellt werden (größeres Bild).
11. **Visueller Kontrast**. Der Kontrast zwischen Hintergrund und Schrift bzw. dem darzustellenden Gegenstand sollte möglichst gut sein.
12. **Beschriftung der Dias**. Die Reihenfolge der Dias sollten auf den Rahmen markiert sein (unten links, wenn das Dia lesbar gehalten wird), damit sie richtig eingelegt werden kön-

nen. Häufig werden unten links Punkte angebracht, welche die korrekte Lage des Dias im Schieber erkennen lassen.

13. **Integration**. Der Vortragstext sollte zu den Bildern passen und sie gut erläutern. Zu Beginn der Präsentation sollte man dem Publikum Gelegenheit geben, sich kurz auf dem Dia zu orientieren. Falls mehrfacher Bezug zu einem bestimmten Dia besteht, sollten Duplikate verwendet werden.

Klare gedankliche Abfolge. Die gedankliche Entwicklung der Ideen sollte klar und logisch nachvollziehbar zu den Ergebnissen führen. Irrelevante Nebenfragestellungen und das Festhalten an Details sollte vermieden werden. Alles was verbal oder visuell präsentiert wird, muß eine klare Funktion in Bezug auf die zentrale These des Vortrags haben.

6. "Research Readiness Checklist" (nach Cone & Foster, 1995)

	Wie gut schreiben Sie?	**ja**	**nein**
1.	Schreiben Sie im allgemeinen gut organisierte, logische und kohärente Aufsätze?	☐	☐
2.	Verwenden Sie eine korrekte Syntax und Grammatik?	☐	☐
3.	Kennen Sie das APA Format (bzw. die Zitier- und Manuskriptrichtlinien der Deutschen Gesellschaft für Psychologie) so gut, daß Sie diese Richtlinien nur gelegentlich nachschauen müssen?	☐	☐
	Besitzen Sie die notwendigen Kenntnisse in Methodenlehre?		
4.	Haben Sie Statistik I und II erfolgreich abgeschlossen?	☐	☐
5.	Haben Sie Praktikum I und II erfolgreich abgeschlossen?	☐	☐
6.	Haben Sie die Versuchsplanungsübung im Sommersemester (und eventuell ein Seminar "Testtheorie") besucht?	☐	☐
7.	Haben Sie die kritische Bewertung von Forschungstexten eingeübt?	☐	☐
8.	Hatten Sie bereits Gelegenheit (z. B. als Hilfskraft) Erfahrungen mit empirischen Forschungsvorhaben zu sammeln?	☐	☐
	Allgemeine Vorbereitung		
9.	Haben Sie mit wenigstens drei Personen über Ihre Erfahrungen mit dem Abfassen von Forschungsberichten gesprochen?	☐	☐
10.	Können Sie sich wenigstens 20 Stunden/Woche mit dem Projekt befassen?	☐	☐
11.	Ist diese Zeit für wenigstens 12-18 Monate (Praktikum III/IV = 2 Semester) verfügbar?	☐	☐
12.	Haben Sie die räumlichen Möglichkeiten, ungestört zu schreiben und Datenanalyse zu betreiben?	☐	☐
13.	Haben Sie Zugang zu adäquaten bibliographischen Quellen (z. B. Bibliotheken und Datenbanken)?	☐	☐
14.	Erhalten Sie regelmäßig Beratung/Unterstützung durch Institutsmitarbeiter oder Betreuer?	☐	☐
15.	Unterstützt Sie Ihre Familie und Angehörigen bei Ihrem Unterfangen?	☐	☐
16.	Können Sie einen Computer nutzen und haben Sie ausreichende Kenntnisse im Umgang damit?	☐	
17.	Haben Sie angemessene Fähigkeiten zum "Zeit-Mangement"?	☐	☐
18.	Haben Sie angemessene zwischenmenschliche und politische Fertigkeiten?	☐	☐
19.	Liegt Ihnen ausreichend Information bezüglich des Formates vor, welches bei der Erstellung von Forschung-/Praktikumsberichten eingehalten werden soll?	☐	☐
20.	Kennen Sie die informellen Regeln, die den erfolgreichen Abschluß eines Projektes begünstigen?	☐	☐

Nach: Cone, J.D. & Foster, S. (1993). *Dissertations and theses from start to finish*. Washington DC: APA.

6. Versuchspläne

Versuchspläne und deren statistische Auswertung in SPSS
In diesem Teil des Skripts werden unterschiedliche Versuchspläne behandelt. Dabei werden einerseits die in Kapitel 1 behandelten allgemeinen Prinzipien und Schritte der Versuchsplanung anhand unterschiedlicher Versuchspläne differenziert und präzisiert; andererseits werden die Inhalte der Kapitel 2 und 3 als bekannt und versuchsplanerisch „verwertbar“ vorausgesetzt. Für jeden einzelnen der dargestellten Versuchspläne wird das allgemeine Vorgehen, die Struktur und die Möglichkeiten uni- oder multivariater statistischer Auswertung besprochen und auf die Beschreibung der SPSS-basierten Auswertung in den Unterabschnitten von Kapitel 7 bzw. 8 verwiesen; mögliche Validitätsgefährdungen werden im Rückgriff auf Kapitel 5.3 detailliert behandelt. Unter versuchsplanerischer Perspektive wird in Kapitel 6 der Veränderungsmessung —basierend auf Meßwiederholungen— besonders viel Platz gegeben, da dieser wichtige Bereich psychologisch-empirischer Arbeit an anderen Stellen des Grundstudiums eher knapper behandelt wird. Neben speziellen Problemen der Veränderungsmessung, die z.T. unter Rückgriff auf Kapitel 4 (Testtheorie & Testkonstruktion) besprochen werden, werden die Varianzanalyse mit Meßwiederholung sowie die multivariate Varianzanalyse für Meßwiederholungsdaten in ihren statistischen Grundzügen vergleichend behandelt und nach versuchsplanerischen Kriterien bewertet.

6.1 Begriffsbestimmung

In der Literatur werden die Begriffe Forschungs- und Versuchsplan unterschiedlich verwendet. Im folgenden Abschnitt soll der Begriff Forschungsplan (FPL) in einem umfassenderen Sinne verstanden werden als der des Versuchsplans (VPL).
Ein *VPL* gibt Aufschluß über die Forschungsfrage/Hypothese, deren Ableitung aufgrund theoretischer Überlegungen oder empirischer Befunde. Er enthält Angaben über die Auswahl und Operationalisierung der Variablen und die Auswahl der Personen (z.B. der zu ziehenden Stichprobe). Außerdem sollte er Informationen über die Art der Datenerhebung, Auswertung und Interpretation enthalten. Die Darstellung von VPL erfolgt typischerweise in Exposés oder Projektanträgen. Versuchspläne (synonym "Designs") sind Bestandteile von Forschungsplänen. Ihre Funktion ist es, eine Antwort auf die Forschungsfrage zu liefern, bzw. eine Entscheidung über Annahme oder Ablehnung der aufgestellten Hypothese zu ermöglichen (Versuchsplan als Funktion der Forschungsfrage, bzw. Hypothese). Es handelt sich hierbei um bewährte, standardisierte und oft in symbolischer Form dargestellte Schemata, die festlegen, wie die UV und AV zueinander in Beziehung gesetzt werden müssen. Damit werden der Aufbau, Ablauf und (partiell) die statistische Auswertung einer Untersuchung im Sinne eines "tue dies und laß das" (Kerlinger) festgelegt. Mit dieser Festlegung ist zugleich die Bewertung einer Untersuchung in Hinblick auf ihre interne Validität verbunden, da bestimmte VPL eine eindeutigere Interpretation der Ergebnisse erlauben als andere VPL. Bei der Bewertung von VPL kommt der internen Validität eine besondere Bedeutung zu. Als intern valide soll eine Untersuchung dann bezeichnet werden, wenn sich die beobachtete Variation in der UV *eindeutig* auf eine Manipulation der AV zurückführen läßt.

6.2 Bewertung und Klassifikation von Versuchsplänen

Es gibt unterschiedliche Klassifikationen von Versuchsplänen. Der Grund hierfür könnte sein, daß in eine solche Klassifikation theoretische Annahmen eingehen, über die keine Einigkeit besteht (z.B. wie muß eine UV beschaffen sein, damit sie als Treatment/Behandlung angesehen werden kann). Ein Konsens scheint jedoch darin zu bestehen, drei unterschiedliche Kategorien von VPL zu unterscheiden: *experimentelle*, *quasi-experimentelle* und *korrelative VPL*. Die Bewertung der diesen Kategorien zugehörigen Designs wird aufgrund ihrer internen

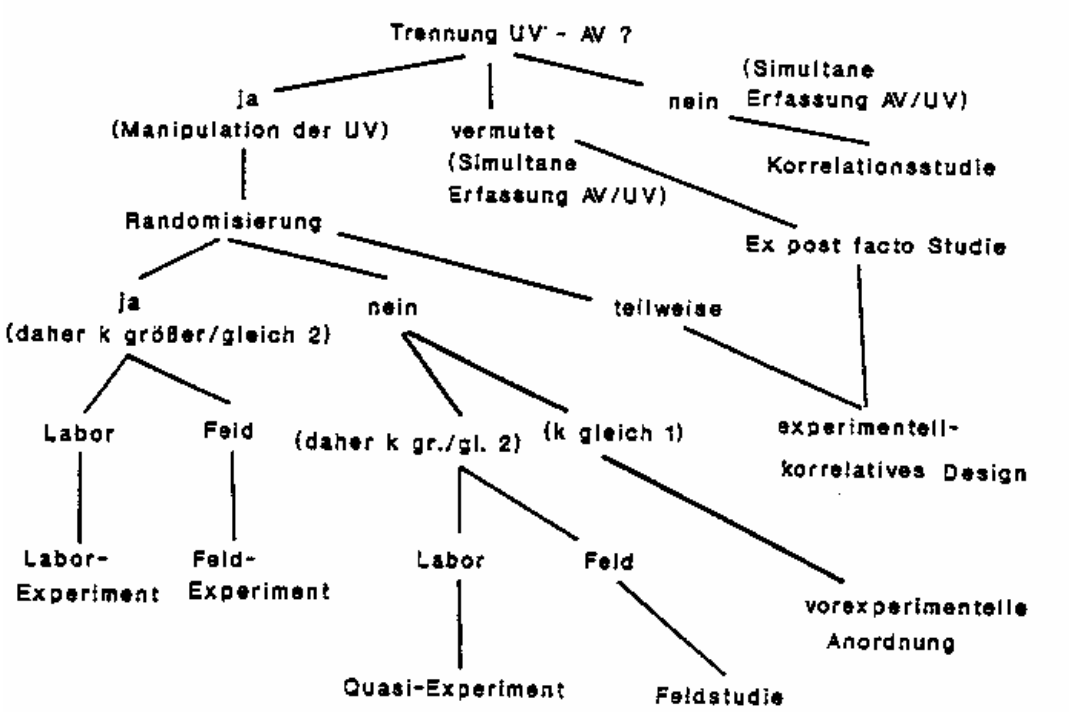
Validität vorgenommen. Hierbei ergibt sich eine Rangreihe, die von den experimentellen Designs angeführt wird.

Experiment: Ein Experiment läßt sich durch den Prozess der *Randomisierung* von den quasi-experimentellen und den korrelativen VPL abgrenzen. Unter *multipler Randomisierung* versteht man (1) die Ziehung einer Zufallsstichprobe aus der Population, (2) die zufällige Zuweisung der Vpn zu den Untersuchungsgruppen sowie (3) die zufällige Zuweisung der Untersuchungsgruppen zu den Treatments (Experimental- und Kontrollgruppe/n). Während das Ziehen einer Zufallsstichprobe auch bei Quasi-Experimenten und Korrelationsstudien möglich und wünschenswert ist (Prinzip der einfachen Randomisierung), ist die multiple Randomisierung (Nr. 1-3) nur bei experimentellen VPL gegeben. Die Möglichkeit der Randomisierung impliziert ebenfalls die willkürliche Manipulation der UV (sog. aktive Variablen wie z.B. Darbietungszeit eines Stimulus, Verstärkermenge in einem Lernexperiment usw.). Die Überprüfug kausaler Zusammenhänge kann deshalb nur im Rahmen experimenteller VPL geleistet werden. Entsprechend gilt das Experiment als die "via regia der empirischen Kausalforschung" (Sarris). Eine solche Bewertung resultiert unmittelbar aus dem Anliegen der Wissenschaft, im Prozeß der Forschung kausale Zusammenhänge aufzudecken. Die Annahme, daß A die Ursache von B ist (Wenn A, dann B), kann aber nur dann als gerechtfertigt gelten, wenn (1) A B vorausgeht, (2) andere Ursachen für das Zustandekommen von B ausgeschlossen werden können, was (3) nur durch den Nachweis, wenn Non-A, dann Non-B stringent falsifiziert, nicht aber verifiziert werden kann.

Quasi-Experiment: Als Quasi-Experiment wird ein Experiment dann bezeichnet, wenn eine multiple Randomisierung nicht möglich ist oder nicht durchgeführt wurde. Jedoch ist in den meisten Quasi-Experimenten eine mehr oder weniger willkürliche Manipulation der UV möglich. Quasi-Experimente realisieren damit ein Treatment. Ein Treatment ist durch eine definierte Bedingungsvariation gekennzeichnet, außerdem können Beginn und Ende bestimmt werden (z.B. Arbeitslosigkeit).
In den meisten Fällen werden im Rahmen von Quasi-Experimenten zwei Arten von Treatments unterschieden. Im ersten Fall handelt es sich bei den Treatments um *aktive Variablen*, wobei es jedoch nicht möglich ist, die Vpn den Treatmentstufen zufällig zuzuweisen. Deshalb muß davon ausgegangen werden, daß zwischen den Untersuchungsgruppen bereits vor Beginn der Untersuchung systematische Unterschiede bestehen (z.B. Vergleich der Effektivität zwischen zwei Lehrmethoden an zwei Parallelklassen). Im zweiten Fall handelt es sich bei den Treatments um *zugewiesene oder attributive Variablen*. Dies sind einerseits organismische Variablen (Geschlecht, Körpergewicht, Extraversion usw.), andererseits kann es sich hierbei auch um die Zugehörigkeit zu einer bestimmten Gruppe handeln (Religionszugehörigkeit, Schultyp usw.). Von einem Quasi-Experiment spricht man jedoch in der Regel nur dann, wenn es sich bei diesen attributiven Variablen *nicht* um Organismusvariablen handelt (hier ist es sinnvoller, von experimentell-korrelativen oder Mischdesigns zu sprechen s.u.). Untersucht werden vielmehr solche Variablen, die theoretisch eine randomisierte Zuordnung der Vpn zu den Faktorstufen erlauben würden (z.B. Religions- oder Klassenzugehörigkeit usw.), die Zuweisung faktisch aber nicht zufällig erfolgt(e).

Korrelationsstudien: Studien, die weder ein Experiment noch ein Quasi-Experiment darstellen, werden meist als Korrelationsstudien bezeichnet. Typischerweise liegt hier kein Treatment im eigentlichen Sinne vor. Da hier alle Variablen gleichzeitig erfaßt werden, erfolgt die Identifikation der UV stets "nachträglich aus dem bereits Geschehenen" (ex post facto). Es handelt sich deshalb lediglich um *begründete Vermutungen*. Sind solche Vermutungen nicht hinreichend zu begründen, so können lediglich Zusammenhänge (Kovariationen) konstatiert werden.

Die Durchsicht der Literatur macht jedoch deutlich, daß eine klare Trennung der drei Designtypen nur schwer durchzuhalten ist. Es gibt Grenzfälle, wo die Zuordnung eines Designs zu einer der genannten Kategorien im Ermessensspielraum des jeweiligen Autors liegt. Auch werden durch die ordinalskalierte Anordnung der Designtypen eher stetige als diskrete Übergänge nahegelegt. Die folgende Abbildung ist deshalb eher als didaktisch motivierte Orientierung zu verstehen, die eine Zuordnung eines bestimmten Designs zu den angeführten Kategorien erleichtern soll (modifiziert nach Hager, 1987).



Erläuterung: Bei experimentell-korrelativen VPL werden UV, die eine Randomisierung erlauben, mit solchen UV kombiniert, für die keine (multiple) Randomisierung möglich ist. Von einem Mischversuchsplan (nicht eingezeichnet) soll dann gesprochen werden, wenn ein VPL eine Kombination eines experimentell-korrelativen und eines quasi-experimentellen VPL darstellt. Dieser Begriff wird in der Literatur nicht einheitlich verwendet. Er ist nicht mit den sog. mixed models der Varianzanalyse zu verwechseln, bei der Faktoren mit festen und zufälligen Effekten kombiniert werden.

In der folgenden Tabelle werden diese Designs einander im Hinblick auf einige wichtige Bewertungskriterien gegenübergestellt (modifiziert nach Sarris, 1992). Das (7stufige) Rating der internen Validität wurde anhand der 13 (intern) validitätsmindernden Faktoren, die von Cook & Campbell (1979) angeführt werden, vorgenommen. Dabei wurde davon ausgegangen, daß sich die Designtypen in bezug auf Variablenvalidität (bzw. Güte der Operationalisierungen), statistischer Validität und Repäsentativität der untersuchen Personen entsprechen.

Designtyp	Multiple Randomisierung	Manipulation der UV	Kausalhypothese vorab begründet	Interne Validität
Strenges Experiment	**J**	**J**	**J**	+++
Experiment	**J**	**J**	**J**	++
Exp.-korrel. Studie	**ZT**	**ZT**	**J**	+
Quasiexperiment	**N**	**J**	**J**	+
Mischpläne	**ZT**	**ZT**	**J**	-
Ex post facto-Studie	**ZT**	**ZT**	**J**	--
Korrelationsstudie	**N**	**N**	**N**	---

Ergänzungen: Neben den bereits genannten VPL-Typen wurde eine Unterscheidung zwischen strengen experimentellen und experimentellen VPL getroffen. Besonders bei länger dauernden Experimenten (z.B. Feldexperimente, die sich über Tage und Wochen erstrecken) können validitätsmindernde Faktoren trotz Randomisierung ihre Wirkung entfalten (vgl. Cook & Campbell, 1979). So kann beispielsweise kaum davon ausgegangen werden, daß die auf eine Warteliste gesetzte Vpn einer Psychotherapiestudie (= Kontrollgruppe) während der Wartezeit nichts unternehmen, um ihren Zustand zu verbessern.

Versuchspläne *ohne* versus *mit* Meßwiederholung: Eine im Hinblick auf die statistische Auswertung relevante Unterscheidung betrifft die Frage, ob der Versuchsplan Meßwiederholung(en) umfaßt oder nicht. Die unter *6.5*-*6.7* vorgestellten Versuchspläne werden daher auch nach diesem Klassifikations-Kriterium getrennt behandelt.

Gruppenuntersuchungen versus Einzelfallstudien: Schließlich soll in diesem Skript noch zwischen Versuchsplänen zur Untersuchung von Probanden-Gruppen und Einzelfall-Plänen systematisch unterschieden werden.

6.3 Das „Max-Min-Kon"-Prinzip

Die Logik des Experiments darin besteht, die beobachtete Variation der AV auf die systematische Manipulation der UV zurückzuführen. Dies ist jedoch nur dann möglich, wenn davon ausgegangen werden kann, daß die Variation der AV nicht durch andere Bedingungen hervorgerufen wurde. Alternativhypothesen, die konkurrierend zu den eigentlichen Untersuchungshypothesen den Einfluß von *Störvariablen* postulieren, mindern die interne Validität einer Untersuchung. Als Störvariablen werden solche Variablen bezeichnet, die eine Variation der AV bedingen und nicht selbst die UV repräsentieren. Daraus folgt, daß die interne Validität einer Untersuchung unmittelbar davon abhängt, inwieweit potentielle Störvariablen kontrolliert werden können.

Die in einer Untersuchung beobachtete Variation der AV läßt sich auf die systematische Variation der UV (Primärvarianz), systematische (Sekundärvarianz) und unsystematische (Fehlervarianz) Fehlereffekte zurückführen: Das „Max-Min-Kon"-Prinzip besagt: „*Maximiere die Primärvarianz! Kontrolliere die Sekundärvarianz! Minimiere die Fehlervarianz!"*— Was bedeutet das?

Varianzquellen und das Max-Min-Kon-Prinzip	
Primärvarianz Unter Primärvarianz verstehen wir jede systematische Veränderung abhängiger Meßwerte, die allein auf die Veränderung der unabhängigen Variablen zurückzuführen ist.	Maximiere die Primärvarianz! • Wahl von Extremwerten in der UV • Wahl von optimalen Abstufungen der UV • Wahl vieler Stufen der UV
Sekundärvarianz	Kontrolliere die Sekundärvarianz!

Unter Sekundärvarianz verstehen wir alle systematischen Veränderungen der abhängigen Variablen, die auf die Wirkung von Störvariablen, nicht aber auf die Manipulation der unabhängigen Variablen zurückzuführen sind.	• Randomisierung • Parallelisierung • Konstanthaltung • Eliminierung
Fehlervarianz Fehlervarianz bezeichnet jede unsystematische Variation der abhängigen Variablen, die weder auf den Einfluß von (identifizierbaren) Störvariablen, noch auf die Manipulation der unabhängigen Variablen zurückzuführen ist. Beispiele: Meßfehler, Auswertungsfehler, individuelle Differenzen	Minimiere die Fehlervarianz! • Auswahl zuverlässiger Meßinstrumente • Geeignete Auswertungsmethoden • Konstanthaltung oder Wiederholungsmessung • Reduktion der unsystematischen Effekte individueller Differenzen • Umwandlung einer Störvariablen in weitere UV bzw. COVA • Meßwiederholung

Maximiere die Primärvarianz!

- *Wahl von Extremwerten in der UV*

Extremgruppenvergleiche können z.B. der Maximierung der Primärvarianz dienen, sind ihrerseits aber wiederum mit speziellen Problemen (statistische Regression, Generalisierbarkeit der Ergebnisse etc) behaftet.

- *Wahl von optimalen Abstufungen der UV*

Bei quantitativen UV (z.B. Größe der Itemmenge in einer Kurzzeitgedächtnisaufgabe) sollten die Stufen der UV nicht zu nahe beieinander liegen, damit Effekte auf eine AV (z.B. Behaltensleitung) deutlich werden.

Kontrolliere die Sekundärvarianz!

- *Randomisierung*

Die Randomisierung wird als *die* Kontrolltechnik schlechthin angesehen. Bis auf wenige Ausnahmefälle (s.u.) sollte sie wann immer möglich eingesetzt werden, da sie die einzige Kontrollmöglichkeit darstellt, die zumindest theoretisch *alle* Fehlerfaktoren (bzw. Randbedingungen) simultan berücksichtigt, was bei keiner anderen Kontrolltechnik der Fall ist. Im Gegensatz zu den anderen Kontrollmöglichkeiten setzt die Randomisierung kein Vorwissen über den Einfluß potentieller Störvariablen voraus. Bei der Randomisierung ist auf eine korrekte Durchführung zu achten (Verwendung von Zufallszahlen). Das Randomisierungsprinzip setzt jedoch ein ausreichend großes N voraus (Gesetz der großen Zahl). Bei weniger als 20 Vpn pro Faktorstufe bzw. Faktorstufenkombination ist deshalb die Bildung von matched samples vorzuziehen und/oder der Effekt der Randomisierung sollte in Hinblick auf die Gleichheit der (Treatment)Gruppen (= Ziel der Randomisierung) mit Hilfe eines Prä-Tests überprüft werden.

- *Parallelisierung*

Bei der Parallelisierung werden die Vpn in Hinblick auf ein oder mehrere Merkmale, die mit der AV korrelieren, in eine Rangordnung gebracht. Danach wird die so gebildete Rangreihe in Blöcke zerlegt, die jeweils soviele einander nachgeordneter Vpn wie Faktorstufen enthalten (ist die UV dreifach gestuft, so würde jeder der Blöcke aus drei Vpn bestehen). Anschließend werden die Vpn jedes Blocks zufällig den Treatmentstufen zugeordnet. Man bezeichnet dieses Vorgehen deshalb auch als die Bildung von *matched samples*. Entsprechende konzipierte Designs werden als *Blockdesigns* bezeichnet. Davon abzuheben ist die Parallelisierung nach dem Mittelwert und der Varianz. Hier wird nur dafür gesorgt, daß die Stichproben, welche den unterschiedlichen Treatmentstufen zugeordnet sind, in Hinblick auf das zu parallelisierende Merkmal den gleichen Mittelwert und die gleiche **Varianz aufweisen. Sind** Merk-

male bekannt, die mit der AV korrelieren ($r > 0.30$), ist die Bildung von matched samples der Randomisierung vorzuziehen. Dies gilt ganz besonders bei kleinen Stichproben.

- *Konstanthaltung*

Mit dieser Kontrolltechnik werden personengebundene und untersuchungsbedingte Störvariablen konstant gehalten, d.h. in ihrer Variabilität eingeschränkt. Dies geschieht bei quantitativen Variablen durch Einschränkung des Wertebereichs (beispielsweise werden nur Vpn im Alter zwischen 20 und 25 Jahren untersucht), bei qualitativen durch den Ausschluß ganzer Kategorien (wenn beispielsweise nur männliche Vpn untersucht werden). Als besonders effektiv erweist sich diese Form der Kontrolle, wenn sie sich auf solche Faktoren bezieht, die zu den von Experimentatoren hergestellten situativen Bedingungen gehören und nicht die UV im eigentlichen Sinne darstellen (z.B. VL, Instruktion, Lärmpegel, Tageszeit usw.). Solche untersuchungsbedingten Störvariablen sollten in aller Regel durch Konstanthaltung berücksichtigt werden, was bedeutet, daß sie unter allen Treatmentbedingungen gleich sind (z.B. gleicher Geräuschpegel, der/dieselbe VL usw.).

- *Elimination*

Die Begriffe Elimination und Konstanthaltung werden manchmal synonym verwendet (z.B. Kerlinger, 1979). Man kann die Elimination als ein Spezialfall der Konstanthaltung ansehen, da bei der Elimination potentielle Störvariablen konstant gehalten werden, indem man sie gänzlich ausschließt. Diese Form der Kontrolle setzt typischerweise bei untersuchungsbedingten Störvariablen an. Die Elimination von personengebundenen Störvariablen dürfte hingegen nur selten möglich sein. Untersuchungsbedingte Störvariablen können, besonders im Kontext von Laborexperimenten, durch *Abschirmung* (typisch für psychophysiologische Messungen oder im Rahmen von wahrnehmungspsychologischen Experimenten), *Standardisierung* (z.B. Verwendung von standardisiertem Stimulusmaterial) und/oder *Automatisierung* (z.B. PC-gestützte Experimente) eliminiert oder konstant gehalten werden. Ihre Kontrolle ist meist wünschenswert, da die Effektstärke der UV in solchen Laborexperimenten meist gering ist und durch unsystematisch wirkende Fehlereinflüsse leicht überlagert werden könnte. Auch die Replikation der Ergebnisse von psychophysiologischen, wahrnehmungspsychologischen und ähnlichen Untersuchungen, wird durch die drei genannten Kontrollmöglichkeiten erleichtert. Sowohl mit der Konstanthaltung als auch mit der Elimination ist ein Verlust an externer Validität verbunden.

Minimiere die Fehlervarianz!

- *Auswahl zuverlässiger Meßinstrumente*

Technisch bedingte Meßfehler tragen unmittelbar zur statistischen Fehlervarianz bei und sollten durch die Wahl zuverlässiger und geeigneter Meßinstrumente weitgehend ausgeschlossen werden.

- *geeignete Auswertungsmethoden*

Bei quantitativen Variablen sollten die Methoden zur Bildung von Kennwerten ebenfalls zuverlässig funktionieren.

- *Meßwiederholung*

Die interindividuelle Variation als Beitrag zur Fehlervarianz in Gruppenuntersuchungen kann ausgeschaltet werden, indem die gleichen Vpn mehrfach untersucht werden und als ihre eigene Kontrollen fungieren. Die Teststärke ist bei Meßwiederholungs-Plänen daher relativ hoch; gravierende Gefährungen der internen Validität *können* jedoch aus „carry over“-Effekten resultieren. Hierauf wird in einem späteren Abschnitt ausführlicher eingegangen.

- *Registrierung*

Bei vielen Experimenten und besonders bei Feldstudien ist es meist nicht möglich, Störvariablen in ausreichendem Maße zu kontrollieren. Oft können Vorfälle und Ereignisse, welche für die Untersuchung relevant sind, nur registriert und ggf. post hoc als Kontrollvariablen (KV) eingeführt werden.

- *Zufallsauswahl (einfache Randomisierung)*

Für ex post facto Designs und Korrelationsstudien sind ausreichend große, repräsentative Stichproben zu fordern. Im Hinblick auf die Populationsabhängigkeit von Korrelationskoeffizienten und die mangelnden Kontrollmöglichkeiten solcher Studien gewährleisten diese Maßnahmen ein hinreichendes Maß an interner Validität. In der psychologischen Forschung werden jedoch nur selten Zufallsstichproben aus definierten Populationen gezogen, eher erfolgt die Auswahl der Personen aufgrund von Quoten (sog. Quotenauswahl), oft sogar nur als Gelegenheitsstichprobe.

- Umwandlung einer Störvariablen in weitere UV oder eine Kovariate (Kovarianzanalyse):

Die Anwendung der COVA setzt voraus, daß eine kontinuierliche Erfassung der Störvariable(n) möglich ist und diese signifikant mit der AV korreliert (was einen linearen Zusammenhang impliziert). Die COVA bezieht sich meist auf personengebundene Störvariablen.

- Bei *experimentellen Designs* kommt ihr die Aufgabe zu, den Anteil der Fehlervarianz in der AV zu verringern.
- Bei *quasi-experimentellen Designs* dient sie außerdem dazu, prä-experimentell bestehende Gruppenunterschiede in der AV (da keine Randomisierung) aus den Post-Meßwerten der AV auszupartialisieren. Die Mittelwerte der Post-Meßwerte werden somit im Hinblick auf die Prä-Meßwerte adjustiert. Die COVA ist der Bildung von matched samples meist unterlegen (wenn AV-KV Korrelation < 0.60) und ist dieser deshalb nur vorzuziehen, wenn eine Paralellisierung nicht möglich oder unökonomisch ist.

- *Diskussion und Überprüfung von (Alternativ-) Hypothesen*

Bei ex post facto Designs und Korrelationsstudien sollten neben den eigentlichen Hypothesen unbedingt Alternativhypothesen diskutiert und ggf. mit Hilfe der Pfadanalyse oder LISREL statistisch überprüft werden. Auch wenn diese Maßnahmen keine Kontrollmöglichkeiten im eigentlichen Sinne darstellen, tragen sie doch dazu bei, die Plausibilität bestimmter Hypothesen zu prüfen.

6.4 Übersicht über spezielle Versuchspläne

Im folgenden sollen einige häufig verwendete Designs in schematisierter Form exemplarisch dargestellt werden. Aus Gründen der Übersichtlichkeit sollen auch hier nur experimentelle, quasi-experimentelle und korrelative Designs berücksichtigt werden, die Meßwiederholung(en) umfassen oder nicht. Es gibt jedoch zweifellos eine sehr viel größere Vielfalt von Forschungsplänen in der Psychologie, so daß mit den konventionellen Gliederungsgesichtspunkten und den Einteilungen aus den primär varianzanalytisch orientierten Lehrbüchern der Versuchsplanung kein repräsentatives Bild gegeben wird. Die folgenden Abschnitte fassen typische Strategien zusammen, welche für die Aufgabenstellungen der wissenschaftlichen Psychologie zweckmäßig sind. Die statistischen Analyseverfahren setzen i.a. intervall-skalierte Daten voraus.

Notation zur Darstellung der Struktur von Versuchsplänen	
R	Randomisierung
X	Treatment
~X	Kein Treatment
$\underline{X}$	Treatmentumkehr(-entzug)
Y	Erfassung abhängiger Variablen
B	Messung „*b*efore“ (*B*aseline-Phase bei mehreren Meßzeitpunkten).
A	Messung „*a*fter“ (nach Treatment bei mehreren Meßzeitpunkten.

6.5 Gruppenuntersuchungen ohne Meßwiederholung

6.5.1 Experimentelle Versuchspläne

VPL#1—Posttest Kontrollgruppenplan (mit R)

Fragestellung & Vorgehen: Aufteilung einer Gruppe von Individuen in zwei oder mehr Untergruppen (Zufallsstichproben, d.h. zufällige Zuweisung von Probanden zu Bedingungen). Das Verhalten einer Treatment-Gruppe (experimentelle Zufallsgruppe) wird mit dem einer ansonsten vergleichbaren Kontrollgruppe (ohne experimentelle Behandlung) verglichen; eine signifikante Mittelwertedifferenz zwischen beiden Gruppen wird auf den (kausalen) Einfluß der experimentellen Bedingung zurückgeführt.

Struktur:

	X	Y	Experimentalgruppe (EG)
R			
	~X	Y	Kontrollgruppe (KG)

SPSS- Auswertung:

(a) univariat: t-Test für unabhängige Stichproben, 1-faktorielle ANOVA mit 2 Stufen
(b) multivariat: Hotelling´s t-Test

Sonstige Anmerkungen: Vermeidung von Auswahlverzerrungen, d.h. ungleichmäßige Repräsentation einer Personengruppe mit bestimmten Merkmalen, durch zufallsmäßige Zuteilung (Randomisierung) der Probanden zu den exp. Bedingungen; wissenschaftslogisch sind experimentelle Versuchspläne (d.h. aktive Manipulation der UV, vollständige Randomisierung der Personen bzw. Gruppen zu Experimental- oder Kontroll-Bedingungen) überlegen. In vielen Bereichen sind sie jedoch nur eingeschränkt oder gar nicht einsetzbar; für viele Aufgabenstellungen sind sie sogar unzweckmäßig.

VPL#2—k x l-Zufallsgruppenversuchsplan

Fragestellung & Vorgehen: Unter einem multifaktoriellen Versuchsplan versteht man die Verallgemeinerung der unifaktoriellen Versuchspläne auf die Berücksichtigung des Einflusses von zwei, drei oder mehr unabhängigen Faktoren im selben Versuchsplan. Die Probanden werden randomisiert den *k x l* Zellen des Versuchsplan zugewiesen.

Struktur

		Faktor A: k=3 Stufen		
R	Faktor B: l=2 Stufen	A1	A2	A3
	B1			
	B2			

SPSS- Auswertung:

(a) univariat: 2- oder mehrfaktorielle ANOVA
(b) multivariate: 2- oder mehrfaktorielle MANOVA

VPL#3—Hierarchische Pläne

Fragestellung & Vorgehen: Bei diesem Plan sind die Stufen des Faktors B unter die Stufen des Faktors A "geschachtelt" (nested), d.h. es werden immer nur zwei Stufen (6 : 3 = 2) des Faktors B pro Faktorstufe des Faktors A realisiert. Interaktionseffekte können deshalb nicht mehr überprüft werden!

Struktur		Faktor A: k=3 Stufen					
		A1		A2		A3	
R	Faktor B: 6 Stufen	B1	B2	B3	B4	B5	B6

SPSS- Auswertung:

(a) univariate: mehrfaktorielle ANOVA
(b) Multivariat: mehrfaktorielle MANOVA

6.5.2 Quasi-experimentelle Versuchspläne

Quasiexperimentelle Versuchspläne (zu Zeitreihenversuchsplänen, Einzelfallversuchsplänen: siehe 6.6) ergänzen experimentelle Untersuchungen: wie beim Experiment sind die Versuchsbedingungen prinzipiell variierbar. Eine wesentliche Bedeutung haben allerdings Sekundärfaktoren, welche die interne Validität beeinträchtigen. Quasiexperimentelle Pläne (siehe COOK & CAMPBELL) umfassen z.B. Pläne mit nicht-äquivalenten Kontrollgruppen (z.B. Selektion u.a. Effekten); Pläne mit Replikationen bzw. Zeitreihen und ausgesetzter oder wiederholter oder umgekehrter Intervention oder verändertem Niveau nach Intervention (sog. Regressions-Diskontinuitäts-Plan); Pläne mit Kovarianzanalyse zur nachträglichen (statistischen) Kontrolle nicht-äquivalenter Gruppen.

(A) Vorexperimentelle Pläne

VPL#4—Posttest Eingruppenplan („One-shot Case Study“)			VPL#5—Posttest Kontrollgruppenplan		
Struktur	X	Y	Struktur	X	Y
				~X	Y

Die Ergebnisse dieser Pläne sind i.a. nicht „kausal“ interpretierbar.

(B) Quasi-experimentelle Versuchspläne

VPL#6—Quadratische Pläne
Fragestellung & Vorgehen: Lateinische Quadrate sind unvollständige Versuchspläne. Von den (im Beispiel) 27 Kombinationsmöglichkeiten der drei 3-stufigen Faktoren werden 9 realisiert. Jede Stufe des Faktors C ist mit jeder Stufe der Faktoren A und B genau 1x kombiniert.

Struktur				
Faktor A:		A 1	A 2	A 3
Faktor B:	B 1	C 1	C 2	C 3
	B 2	C 2	C 3	C 1
	B 3	C 3	C 1	C 2

SPSS- Auswertung:

(a) univariat: 3-faktorielle ANOVA, bei der im Design nur die Hautpeffekte spezifiziert werden
(b) multivariat: 3-faktorielle MANOVA (nur Haupteffekte)

Sonstige Anmerkungen: Das Lateinische Quadrat ist bei drei Faktoren (mit beliebig vielen Abstufungen) anwendbar. Interaktionseffekte sind im lateinischen Quadrat nicht überprüfbar. Sofern Interaktionseffekte faktisch bestehen, sind die Haupteffekte (A-C) nicht eindeutig interpretierbar.

Statistik-Exkurs: Multivariate Varianzanalyse (MANOVA) Anfang

Aus den vorangegangenen Abschnitten ging hervor, daß bei Versuchsplänen ohne Meßwiederholung und mit multivariater Datenstruktur die multivariate Varianzanalyse, MANOVA, zur statistischen Auswertung indiziert ist. Dieses Verfahren wird in Statistik II nicht behandelt und soll daher an dieser Stelle nur in den Grundzügen vorgestellt werden. Zum Verständnis der Grundzüge der MANOVA sind Kenntnisse des t-Tests, der ANOVA und ein wenig Matrixalgebra ausreichend.

(A) Es gibt Gründe für die Durchführung einer MANOVA anstelle mehrerer ANOVA.

(1) Eine Serie „fraktionierter“ univariater Tests erhöht die Typ-I-Fehlerwahrscheinlichkeit
Beispiel: Es werden 10 Tests auf dem 5%-Niveau durchgeführt. Die Wahrscheinlichkeit, keinen Typ-I-Fehler zu begehen, beträgt: (0.95)(0.95) ..... (0.95) ≈.60. Die Wahrscheinlichkeit für mindestens 1 falsche Zurückweisung der H_0 beträgt daher: 1-.60=.40 (Genaueres bei (D)).
(2) Die univariaten Tests berücksichtigen nicht die Korrelationen zwischen den Variablen (multivariate Fragestellung).
(3) Untersuchungsgruppen mögen sich für keinen der univariaten Vergleiche unterscheiden, wohl aber in der Menge der Variablen insgesamt; d.h. kleine Unterschiede in den einzelnen Variablen mögen sich zu einer signifikanten „Overall“-Differenz kombinieren.

Man sollte allerdings nicht jede beliebige Variable mit dieselben MANOVA analysieren. Variablen, für die man keine guten Gründe hat, anzunehmen, daß sie Bedingungen oder Gruppen unterscheiden, sollten getrennt von solchen Variablen behandelt werden, für die diese Gründe existieren. Andernfalls könnten (Gruppen-) Unterschiede in letztgenannter Variablenmenge „verschleiert“ werden.

(B) Die multivariate Test-Statistik ist eine Generalisierung des univariaten t-Tests.

Univariater t-Test:

$$H_0: \quad \mu_1 = \mu_2$$

Multivariater Test:

$$H_0: \quad \begin{vmatrix} \mu_{11} \\ \mu_{21} \\ . \\ . \\ . \\ \mu_{p1} \end{vmatrix} = \begin{vmatrix} \mu_{11} \\ \mu_{21} \\ . \\ . \\ . \\ \mu_{p1} \end{vmatrix}$$

1. Index: Variable
2. Index: Gruppe

Annahmen des univariaten t-Tests:

(1) Unabhängigkeit der Beobachtungen
(2) Normalverteiltheit der abhängigen Variablen
(3) Gleichheit der Populationsvarianzen (Varianzhomogenität)

Annahmen des multivariaten Tests:

(1) Unabhängigkeit der Beobachtungen
(2) Multivariate Normalverteiltheit der abhängigen Variablen
(3) Gleichheit der Populationskovarianzen

Die multivariate Teststatistik ergibt sich aus der univariaten, indem Zahlen durch Vektoren und Matrizen ersetzt werden.

Univariater t-Test: $$t^2 = \frac{({}^{M}y_1 - {}^{M}y_2)^2}{s^2 (1/n_1 + 1/n_2)}$$

Hotelling´s T^2: $$T^2 = \frac{n_1 n_2}{n_1 + n_2} ({}^{M}y_1 - {}^{M}y_2)' \, S^{-1} \, ({}^{M}y_1 - {}^{M}y_2)$$

${}^{M}y_1 - {}^{M}y_2$: Differenz der Mittelwerts-*Vektoren*,
S^{-1}: invertierte geschätzte Populations-Kovarianzmaxtrix; die Matrix S entspricht dem Fehlerterm in Hotelling´s T^2-Test; sie wird analog zur Innergruppen-Populationsvarianz beim t-Test ermittelt.

Schätzung des Fehlerterms für t-Test und Hotelling´s T^2:

	t-Test (univariat)	T^2 (multivariat)
Annahme	Die Innergruppen-Populationsvarianzen sind gleich: $\sigma_1^2=\sigma_2^2=\sigma$	Die Innergruppen-Populations-Kovarianzmatrizen sind gleich: $\Sigma_1=\Sigma_2=\Sigma$
Zur Berechnung der angenommenen gemeinsamen Populationsparameter werden die nachfolgenden 3 Schritte vorgenommen:		
Berechnen der Maße der Innergruppen-Variabilität	ss_{g1}, ss_{g2}	W_1, W_2
Zusammenfügen dieser beiden Maße	$ss_{g1} + ss_{g2}$	$W_1 + W_2$ $W_1 = \begin{vmatrix} ss_1 & ss_{12} \\ ss_{21} & ss_2 \end{vmatrix}$ $W_2 = \begin{vmatrix} ss_1 & ss_{12} \\ ss_{21} & ss_2 \end{vmatrix}$
Division durch die Anzahl Freiheitsgrade	$\hat{\sigma}^2 = \frac{ss_{g1} + ss_{g2}}{n_1+n_2-2}$	$\hat{\Sigma} = S = \frac{W_1 + W_2}{n_1+n_2-2}$

^=geschätzt

Hotelling (1931) konnte zeigen, daß die nachfolgende Transformation von T^2 eine exakte F-Verteilung ergibt: $F = [(n_1+n_2-p-1)/(n_1+n_2-2)p]\ T^2$ (p=Anzahl abhängige Variablen)

Eine andere Schreibweise für T^2 ist: $T^2 = kd'S^{-1}d$
(d=Vektor der Mittelwertsdifferenzen; S=Kovarianzmatrix der p abhängigen Variablen)

(C) Ein Rechenbeispiel verdeutlicht das Vorgehen bei der MANOVA.

1. Vektor der Mittelwertsdifferenzen:
$({}^{M}y_1 - {}^{M}y_2)' = (2 - 5, 4 - 8)$

2. Varianz-Kovarianz-Matrizen für die Gruppen 1 und 2:
$W_1 = \begin{vmatrix} 2 & 4 \\ 4 & 14 \end{vmatrix}$ $W_2 = \begin{vmatrix} 4 & 4 \\ 4 & 16 \end{vmatrix}$

3. Bestimmung des multivariaten Fehlerterms S^{-1} (gepoolte Innergruppen-Varianz-Kovarianz-Matrix)

$$S = \frac{\begin{vmatrix} \mathbf{2} & \mathbf{4} \\ 4 & 14 \end{vmatrix} + \begin{vmatrix} \mathbf{4} & \mathbf{4} \\ 4 & 16 \end{vmatrix}}{n_1 + n_2 - 2} = \begin{vmatrix} 6/7 & 8/7 \\ 8/7 & 30/7 \end{vmatrix}$$

4. Bestimmung der Inversen von S
$S^{-1} = \begin{vmatrix} 1.811 & -.483 \\ -.483 & .362 \end{vmatrix}$ (Rechenweg ausgelassen)

5. Berechnen von Hotelling´s T^2:

$$T^2 = \frac{n_1 n_2}{n_1+n_2} ({}^{M}y_1 - {}^{M}y_2)' S^{-1} ({}^{M}y_1 - {}^{M}y_2)$$

$$= \frac{3 * 6}{3 + 6} (2 - 5, 4 - 8) \begin{vmatrix} 1.811 & -.483 \\ -.483 & .362 \end{vmatrix} (2 - 5) \\ (\quad) \\ (4 - 8) = 21$$

6. F-transformation von T^2:

$$F = \frac{n_1+n_2 - p - 1}{(n_1+n_2 - 2)\ p} T^2 = 9$$

7. Post hoc-Tests:
Erhält man einen signifikanten T^2-Wert, so besteht ein post hoc-Test in der Durchführung univariater t-Tests, wobei allerdings jeder t-Test auf dem Signifikanzniveau α/p (p=Anzahl abhängige Variablen; α=insgesamte Fehlerwahrscheinlichkeit) durchgeführt wird (Bonferroni-Korrektur).

Bonferroni-Korrektur: jeder der p post hoc-Tests wird mit einem α-Niveau von α/p durchgeführt. Daher liegt die Typ-1-Fehlerwahrscheinlichkeit der post hoc-Tests *insgesamt* bei α.

(D) Stichprobengröße, Anzahl AVs und die Interkorreliertheit der AVs beeinflussen den Typ-I-Fehler bei ANOVA und MANOVA unterschiedlich.

Fehlerwahrscheinlichkeiten:

Hummel & Sligo gingen der Frage nach, wie die Fehlerwahrscheinlichkeit (a) bei univariaten versus (b) multivariaten mit nachfolgenden univariaten Test abhängt von

- der Stichprobengröße,
- der Anzahl Variablen und
- der Interkorreliertheit der Variablen.

Ergebnisse (siehe nachfolgende Tabelle)

- Bei den univariaten Tests liegt das faktische α-Niveau deutlich über dem nominalen, vor allem wenn „mehr" Variablen berücksichtigt werden.
- Bei den univariaten Tests ist die Erhöhung des faktischen α-Niveaus umso größer, je geringer die Interkorrelation der abhängigen Variablen ist.
- Folgen einem multivariaten Test univariate Tests (im Falle der Signifikanz), so liegt das faktische α-Niveau nahe beim nominalen und ist eher „konservativ"

Konsequenz: Werden univariate Tests signifikant, obgleich ein korrespondierender multivariater „Overall"-Test nicht-signifikant ist, so kann dies die Konsequenz des erhöhten faktischen gegenüber dem nominalen α-Niveau sein!

Fehlerraten bei der Analyse multivariater Daten mit univariaten Tests
oder multivariaten Tests gefolgt von univariaten Tests

Stichprobengröße	Anzahl Variablen	Anteil gemeinsamer Varianz	
		.10	.70
nur univariate Tests			
10	3	.145	.077
10	9	.348	.129
50	3	.138	.083
50	9	.324	.146
Erst multivariater Test, dann univariate Tests			
10	3	.044	.022
10	9	.050	.018
50	3	.038	.028
50	9	.036	.020

Die Ergebnisse für (a) N=30, (b) Anteil gemeinsamer Varianz=.30, .50 und (c) p=6 abhängige Variablen wurden der Übersichtlichkeit halber weggelassen. Nominales α=.05

(E) Multivariate Signifikanzen können ohne univariate Signifikanzen auftreten.

Problem: Es kann der Fall eintreten, daß der multivariate Test signifikant wird ohne daß einer der nachgeschalteten univariaten Tests signifikant wird.

Grund: Die multivariaten und univariaten Tests nutzen unterschiedliche Informationen in den Daten:
- multivariat *vs* univariat: Interkorrelationen der AV werden berücksichtigt
- multivariat *vs* univariat: alle AV werden simultan analysiert

Multivariate Tests sind den univariaten insbesondere dann „überlegen" hinsichtlich der Teststärke, wenn eine (oder beide) der nachfolgenden Bedingungen gegeben ist (sind):

(1) Das „Treatment" (Gruppenvergleich) beeinflusste die AV in unterschiedlicher Weise:
- Die „across-group"-Verbindung zwischen den AV ist schwach, und jede AV liefert einen „einzigartigen" Beitrag zur Trennung der Gruppen
- Dies entspricht einer fehlenden (geringen) Multikollinearität in der multiplen Regression (Kriterium=Gruppe)

Beispiel		Gruppe 1	Gruppe 2	Gruppe 3
	AV1	4.6	5.0	6.2
	AV2	8.2	6.6	6.3

(2) Die Innergruppen-Korrelation der AV ist sehr stark (z.B. r=.88):
- Damit wird der MANOVA-Fehlerterm (korrespondierend mit MS_{within} in der ANOVA) kleiner
  - Die Matrix $|W|$ entspricht MS_{within} in der ANOVA bzw. ist die multivariate Veallgemeinerung der SS_{within}.
  - $|W|$ (Determinante der Matrix Kovarianzmatrix W) ist ein Maß dafür, wie sehr die Variablenausprägungen in der Gruppen variieren, sie entspricht der generalisierten Varianz einer Menge an Variablen.
  - *Beispiel*:
    - $W_1 = \begin{vmatrix} 12.0 & 13.2 \\ 13.2 & 18.8 \end{vmatrix} \rightarrow |W_1| = 51.36$
    - $W_2 = \begin{vmatrix} 12.0 & 5.0 \\ 5.0 & 18.8 \end{vmatrix} \rightarrow |W_2| = 200.6$
- *Fazit*: Bei *geringer* Interkorreliertheit der Variablen werden die Fehler in der einen Variablen nicht durch die der anderen erklärt und addieren sich zur Gesamtfehlervarianz auf; bei *hoher* Interkorreliertheit der Fehler fügen weitere Variablen der Fehlervarianz wenig hinzu.

„Indikation" zur multivariaten Analyse: die Kombination aus einer geringen Zwischengruppen-Assoziation der abhängigen Variablen bei gleichzeitiger hoher Innergruppen-Korrelation.

(F) Die MANOVA für k Gruppen ist eine Verallgemeinerung der ANOVA für k Gruppen.

Problemstellung: Es werden k Gruppen hinsichtlich p abhängigen Variablen miteinander verglichen.

Bei mehrfaktoriellen univariaten ANOVA lautet die Nullhypothese:
H_0: $\mu_1=\mu_2=....=\mu_k$ (Populationsmittelwerte sind gleich)

Bei mehrfaktoriellen MANOVA lautet die Nullhypothese:
H_0: $\mu_1=\mu_2=....=\mu_k$ (Populations-Mittelwertsvektoren sind gleich)

Die univariate Test-Statistik F wird bestimmt über:

$$F = \frac{MS_{between}}{MS_{within}}$$

Die multivariate Test-Statistik Wilk´s Λ (lambda) wird bestimmt über:

$$\Lambda = \frac{|W|}{|T|} = \frac{|W|}{|B+W|}$$

W=Maß der Innergruppenvariabilität
B=Maß der Zwischengruppenvariabilität
T=Maß der Gesamtvariabilität

Anmerkung: Wilk´s Λ ist ein *inverses* Kriterium, d.h. je kleiner Λ, desto größer ist die Treatment-Varianz: 1. Fall: kein Treatment-Effekt⇒ B=0 ⇒ Λ=(|W| / |0+W|)=1.
2. Fall: „großer" Treatment-Effekt⇒ B>>W ⇒Λ→0.

Univariate Quadratsummenzerlegung: $SS_{total}=SS_{between}+SS_{within}$.

Multivariate Quadratsummenzerlegung: T = B + W.
W=SSCP-Matrix within; B=SSCP-Matrix between; T=SSCP-Matrix total
SSCP = Sum of Squares Cross-Product Matrix

Berechnung der Matrizen T, B und W: **W** wird im Falle von k Gruppen genauso berechnet wie im Falle von 2 Gruppen.

Beispiel: 3 Gruppen, 2 Variablen: $W = W_1 + W_2 + W_3$.

Berechnen der Innergruppenvariabilität und Kovarianz:

$$W_1 = \begin{vmatrix} ss_1 & ss_{12} \\ ss_{21} & ss_2 \end{vmatrix} \quad W_2 = \begin{vmatrix} ss_1 & ss_{12} \\ ss_{21} & ss_2 \end{vmatrix} \quad W_3 = \begin{vmatrix} ss_1 & ss_{12} \\ ss_{21} & ss_2 \end{vmatrix}$$

ss_1=Varianz der ersten AV; ss_2=Varianz der zweiten AV; $ss_{12}=ss_{21}$=Kovarianz der ersten und zweiten AV ... in W_1, W_2, W_3 separat berechnet für die 3 Gruppen.

Berechnen der Gesamtvariabilität:

Wegen T=B+W wird zunächst B berechnet.

1. Diagonalelemente von B:

$$B = \begin{vmatrix} \underline{b} & b \\ b & \underline{b} \end{vmatrix}$$

$b_{ii} = {}_j\Sigma^k\, n_j\, ({}^{M}y_{ij} - {}^{GM}y_i)^2$.
n_j=Anzahl Probanden in Gruppe j
${}^{M}y_{ij}$= Mittelwert der Variable i in Gruppe j
${}^{GM}y_i$=Grand mean der Variable i (über alle Gruppen)
b_{ii} = für eine bestimmte Variable i entspricht b_{ii} der $SS_{between}$ für i

2. Nichtdiagonalelemente von B:

$$B = \begin{vmatrix} b & \underline{b} \\ \underline{b} & b \end{vmatrix}$$

$b_{mi} = b_{im} = {}_j\Sigma^k\, n_j\, ({}^{M}y_{ij} - {}^{GM}y_i)\, ({}^{M}y_{mj} - {}^{GM}y_m)$.

3. Nun wird T berechnet: T=B+W

4. Wilk´s Λ:

$$\Lambda = \frac{|W|}{|T|}$$

Post hoc-Verfahren: Sofern die MANOVA für k>2 Gruppen ein signifikantes Ergebnis bringt, ist man i.a. daran interessiert, zu erfahren, auf welche(n) Gruppenvergleich(e) das „overall" signifikante Ergebnis zurückgeht.

Weiteres Vorgehen:

- Hotelling´s T^2 wird für alle (oder alle geplanten) paarweisen Gruppenvergleiche durchgeführt (mit α-Adjustierung)
- Für die signifikanten T^2s wird mit univariaten t-Tests nach den differenzierenden Variablen gesucht (mit α-Adjustierung)
-

(G) Die Anzahl abhängiger Variablen in einer MANOVA sollte so gering wie möglich sein.

Gründe:

1. viele, theoretisch/empirisch wenig gut begründete AV mögen mit kleinen Gruppenunterschieden verbunden sein, die die großen Gruppenunterschiede der wenigen gut begründeten AV „verdecken".
2. Die Teststärke multivariater Tests nimmt mit zunehmender Anzahl AV generell ab.
3. Eine Kombination (z.B. Summe) mehrere AV mag reliabler sein als die Einzel-AV (Aggregationsprinzip)
4. Die Ergebnisse sind i.a. leichter zu interpretieren.
5. Sind die AV redundant, so gibt es vermutlich ein zugrundeliegendes „Konstrukt", über das eine vorgeschaltete PCA Aufschlüsse geben kann.
6.

Zusammenfassung MANOVA

(A) Es gibt Gründe für die Durchführung einer MANOVA anstelle mehrerer ANOVA.
(B) Die multivariate Test-Statistik ist eine Generalisierung des univariaten t-Tests.
(C) Ein Rechenbeispiel verdeutlicht das Vorgehen bei der MANOVA.
(D) Stichprobengröße, Anzahl AVs und die Interkorreliertheit der AVs beeinflussen den Typ-I-Fehler bei ANOVA und MANOVA unterschiedlich.
(E) Multivariate Signifikanzen können ohne univariate Signifikanzen auftreten.
(F) Die MANOVA für k Gruppen ist eine Verallgemeinerung der ANOVA für k Gruppen.
(G) Die Anzahl abhängiger Variablen in einer MANOVA sollte so gering wie möglich sein.

ENDE
Statistik-Exkurs: Multivariate Varianzanalyse (MANOVA)

6.5.3 Korrelative Versuchspläne

Zielsetzung korrelativer Versuchspläne ist die Prüfung von Zusammenhangshypothesen zwischen zwei oder mehreren Variablen, und zwar ohne die Manipulation von experimentellen Variablen und damit ohne den Anspruch auf kausaltheoretische Aussagen.

(1) Prüfung von Merkmalszusammenhängen

Folgende Pläne und Datenanalysestrategien zur Prüfung von Hypothesen über Merkmalszusammenhänge werden häufig verwendet:

- *Einfache (bivariate) Korrelation* (Merkmalszusammenhang) und *Regression* (Merkmalsvorhersage)
- *Partielle Korrelation und Regression*
- *Multiple Korrelation und Regression*
- *Multiple und partielle Regression* als sog. Pfadanalyse, um hypothetische Bedingungsketten bzw. Modellvorstellungen an einem Datensatz zu testen.
- *Kanonische Korrelation* (Kanonische Analyse): zur Prüfung des Zusammenhangs zweier Sätze von Variablen
- *Faktorenanalyse:* Bei der multiplen Korrelation/Regression, Kanonischen Korrelation und Faktorenanalyse wird die simultane Beziehung mehrerer Variablen analysiert; es handelt sich um eine Zusammenfassung mehrerer linear transformierter Variablen (sog. Linearkombinationen).
- *Nicht-lineare Korrelation, polynomiale Korrelation:* nicht-linearer Merkmalsbeziehungen.

(2) Multiple Regressions-Korrelations-Analyse (MRC)
Dieses generelle Datenanalyse-Konzept umfaßt Regressionen und Korrelationen; außerdem Varianzanalyse (ANOVA) und Kovarianzanalyse (ANCOVA), indem die Bedingungen als Variablen kodiert werden; Diskriminanzanalyse, indem die Gruppenzugehörigkeit kodiert wird; außerdem Pläne mit Meßwiederholung. Schließlich sind Generalisierungen auf die Kanonische Analyse und mehrere AV in der multivariaten Varianzanalyse (MANOVA) und Kovarianzanalyse (MANCOVA) möglich.
(3) Diskrimination und Klassifikation
Während es bei der deskriptiven numerischen Taxometrie um die Suche nach Klassen (Clustern) geht, kommt es bei der Diskriminanzanalyse auf die Trennung bereits bekannter Gruppen an. Können diese Gruppen multivariat tatsächlich getrennt werden? Mit welcher Zuordnungs- (Klassifikations-) Wahrscheinlichkeit ist jedes einzelne Individuum einer der Gruppen zuzuordnen?

6.6 Gruppenuntersuchungen mit Meßwiederholung

Psychologische Veränderungsmessung ist überall dort notwendig, wo geprüft werden soll, ob psychologische Interventionen intraindividuell „Wirkung“ zeigen. B*eispiele* sind die Überprüfung der Wirksamkeit einer Psychotherapie, eines bestimmten Trainings etc.
Bei Versuchspläne mit Wiederholungsmessung ("within-subject design" oder "repeated-measurements designs") werden eine oder mehrere Probandengruppen zu verschiedenen Zeitpunkten des Gesamtversuchs unter allen Bedingungen getestet. Zu den wichtigsten Vorteilen gehört bei varianzanalytischer Auswertung die Reduktion der Fehlervarianz; gewichtige Nachteile hinsichtlich der internen Validität können aus etwaigen „carry over“-Effekten entstehen, also der Beeinflussung einer Messung durch eine vorangegangene Messung.

6.6.1 Experimentelle Fehlerquellen bei der Veränderungsmessung

(A) *„Carry-Over“-Effekte*: Diese gehören zu den wichtigsten potentiellen Gefährdungen der internen Validität bei Meßwiederholungs-Designs. Zur Gruppe der „carry over“-Effekte gehören: gehören: Positionseffekte, Sensibilisierung, Übungs-, Ermüdungs, Erinnerungseffekte,

externes zwischenzeitliches Geschehen, Interaktionen zwischen Behandlungsbedingungen und Übertragungseffekten sowie allgemeine Testerfahrung (testing sophistication: die Vp lernt unspezifische, in einer Testsituation benötigte Fähigkeiten durch die Testung (z.B. sich eine Zeit lang zu konzentrieren; die wesentlichen Punkte einer Instruktion erkennen; etc)), zwischenzeitliches Training (die Vp trainiert zwischen den Testungen „unaufgefordert“ und unkontrolliert die getesteten Fähigkeiten, Sättigungseffekte (die Vp verliert Interesse und Motivation an der Untersuchung teilzunehmen). Einige dieser Effekte können auf folgende Weise kontrolliert werden:

(1) *Vollständiges Ausbalancieren*: Alle möglichen Abfolgen werden realisiert (bei k Faktorstufen ergeben sich k! Abfolgen).
(2) *Unvollständiges Ausbalancieren*: (a) Aus der Population aller möglichen Sequenzen werden eine oder mehrere *zufällig* ausgewählte Sequenzen realisiert (Randomisierungsprinzip); (b) Es werden bei k Faktorstufen k verschiedene Reihenfolgen so gewählt, daß jede Bedingung gleich häufig an jeder Stelle erscheint (Anordnung im Sinne des Lateinischen Quadrats). Quadratische VPL finden sowohl bei der intra- als auch bei der interindividuellen Bedingungsvariation Verwendung. Bei der intraindividuellen Bedingungsvariation wird jede Stichprobe einer anderen Sequenz zugewiesen. Bei der interindividuellen Bedingungsvariation werden qudratische Pläne meist aus ökonomischen Gründen realisiert, wenn Interaktionseffekte unwahrscheinlich sind. Hierbei ist jede Zelle des VPL mit unterschiedlichen Stichproben besetzt (s.u.). Ein wesentlicher Nachteil dieses VPL besteht darin, daß Interaktionseffekte (statistisch) nicht überprüft werden können und eine angemessene Interpretation der Ergebnisse nur möglich ist, wenn diese, etwa aufgrund anderer Experimente oder plausibler Hypothesen, ausgeschlossen werden können.

Obwohl die erste Variante die ideale Lösung darstellt, ist sie in der Regel aus ökonomischen Gründen nur bis maximal drei Faktorstufen anwendbar (es ergeben sich 6 mögliche Sequenzen). Bei mehr als drei Faktorstufen stellt die Variante Nr. 2a die Methode der Wahl dar, wobei mehrere zufällig ausgewählte Abfolgen realisiert werden sollten. Einschränkend muß angemerkt werden, daß die Sequenzeffekte durch die genannten Maßnahmen nur ausgeglichen (kontrolliert) werden können, wenn sie symmetrisch sind, d.h. wenn "das Ausmaß der von jeder Behandlungsbedingung ausgehenden Sequenzeffekte gleich dem Ausmaß der von anderen auf sie entfallenden Effekte ist" (Hager & Westermann, 1983, S. 58). Liegen solche asymmetrischen Sequenzeffekte vor, so ergibt sich eine signifikante Interaktion Sequenz*Treatment. Solche Sequenzeffekte können jedoch mit Hilfe der genannten Designs überprüft werden.

(B) *Selektive Ausgangsstichprobe*: Messungen zu mehreren Zeitpunkten (z.B. bei Längsschnitt-Untersuchungen) sind sehr aufwendig hinsichtlich Organisation und zeitlichem Aufwand. Kleinere Ns sind zumeist die Folge und damit eine erhöhte Schwierigkeit, *Repräsentativität* zu erzielen.

(C) *Selektive Stichprobenveränderung*: Wiederholte Untersuchungen setzen hoch-motivierte Probanden voraus, die zudem über einen längeren Zeitraum erreichbar sind. Gründe für das Ausscheiden („drop out“) aus einer Studie können Desinteresse, Wohnortwechsel etc sein.
Sind die Probanden-“Ausfälle“ nicht zufällig, sondern stehen in einem systematischen Zusammenhang mit der AV, so spricht man von einer

selektiven Drop-Out-Rate. Die Kontrolle selektiver „Drop-Outs“ ist kaum möglich. Daher bleibt i.a. nur die Möglichkeit, nachträglich die bereits erhobenen Merkmale für die Ausgeschiedenen und die in der Studie Verbliebenen miteinander zu vergleichen.

(D) *Zeit als Störfaktor*: Zu messende Veränderungen benötigen Zeit. Die „richtige“ Wahl des Zeitintervalls zur sinnvollen Abbildung eines psychologischen Prozesses kann eine erhebliche Schwierigkeit darstellen.

1. *Fall*: Ein zu kurzes Zeitintervall—Es wird keine Veränderung gemessen, da der zu messende Prozeß noch nicht abgeschlossen ist.

2. Fall: Ein zu langes Zeitintervall—Es wird keine Veränderung gemessen, da der zu messende Prozeß zwar abgeschlossen, doch bereits wieder von einem angelaufenen neuen Prozeß überlagert wird, der nichts mit der Fragestellung der Untersuchung zu tun hat.

„Ausweg“: Die Wahl des/der Zeitintervalle kann nur auf der Grundlage theoretischer Erwägungen und bereits gewonnener empirischer Erkenntnisse getroffen werden; liegen diese Grundlagen nicht vor, so ist die Wahl der Zeitintervalle ein „Glücksspiel“.

6.6.3 Differentielle Veränderung

Im Hinblick auf den praktischen Einsatz der Veränderungsmessung kann es oft wichtig sein, die individuelle Veränderung und damit interindividuelle Unterschiede in der Veränderung festzustellen. Verschiedene Ansätze scheinen hier naheliegend zu sein; sie können jedoch mit speziellen Problemen und Gefahren für Trugschlüsse verbunden sein.

(1) Differenzwert-Bildung

Ein naheliegendes Vorgehen bei der Messung von Veränderung ist die Bildung von Differenzen, z.B. „post minus prä“. Dieses scheinbar plausible Vorgehen ist jedoch mit einigen Problemen verbunden.

(A) Skalenabhängigkeit von Differenzen

(B)

Beispiel: Es soll untersucht werden, ob Verkehrserziehung die Unfallgefährdung reduziert und ob dieser Effekt für Männer und Frauen unterschiedlich ist.

Rohwerte				
	Vorher		Nachher	
	Anzahl Unfälle	Anzahl km	Anzahl Unfälle	Anzahl km
Männer	4.0	30.000	5	40.000
Frauen	0.5	10.000	0.9	20.000

Differenzen						
	Anzahl km /Unfall			Anzahl Unfälle /10000km		
	vorher	Nachher	D	vorher	nachher	D
Männer	7.500	8.000	500	1.33	1.25	0.08
Frauen	20.000	22.222	2.222	0.50	0.45	0.05
Fazit			F>M			M>F

Erläuterung: Der Übergang von der Skala „Anzahl km/Unfall“ zur Skala „Kilometer/Unfall“ ist eine zulässige, nicht-lineare Transformation nach der Formel Y=1/X. Die Aussage, die sich aus der Interpretation von Differenzwerten ergibt, ist in diesem Beispiel abhängig von der gewählten Skala und kann durch Wechseln der Skala ins Gegenteil verkehrt werden. Sofern eine Interaktion nicht invariant ist gegenüber monotonen Transformationen, sind folgende Fragen zu prüfen:

1. Welche Maße (AV) außer dem zunächst gewählten sind noch verwendbar? Welche inhaltlich-theoretisch plausiblen (*nicht*: „exotischen“) Skalentransformationen kommen in Betracht?
2. Wie ändern sich ggf. die Ergebnisse, wenn andere AV gewählt bzw. Skalentransformationen durchgeführt werden?
3. Gibt es einen inhaltlichen Grund, weshalb die gewählte Skalierung gegenüber anderen zu bevorzugen ist?

(B) Reliabilität von Differenzen
Beispiel: In einer Studie wird untersucht, ob Lernfähigkeit —definiert als Leistungszuwachs nach einer Lernphase— stärker mit der Schulnote korreliert als eine punktuelle Fähigkeitsmessung. Nach einer ersten Fähigkeits-Messung wurde eine Lernphase durchgeführt, der eine zweite Fähigkeits-Messung folgte. Die Differenz zwischen 2. und 1. Messung korrelierte jedoch erheblich schwächer mit der Schulnote als jede der einzelnen Fähigkeitsmessungen für sich genommen.—Wie ist das zu erklären?

→In die Differenz zweier Messungen gehen die Meßfehler (Unreliabilitäten) beider Messungen ein. Deswegen sind Differenzwerte i.a. relativ unreliabel; und zwar umso unreliabler, je stärker die Meßwerte korreliert sind, für welche die Differenzen gebildet werden.

Testtheorie-Exkurs: Reliabilität von Differenzen
Anfang

In Abschnitt* 4.3 wurde die KTT behandelt. Gemäß der KTT wird die Gesamtvarianz der Testwerte zerlegt in eine Varianz der wahren Werte (Var(T)) und die Fehlervarianz (Var(F)).

Die Varianz der Differenzwerte läßt sich in die Varianz der wahren Differenzen und die Varianz der Fehlerdifferenzen zerlegen.
$$Var(X-Y) = Var(T_x-T_y) + Var(F_x-F_y)$$

Reliabilität ist definiert als Anteil der Varianz der wahren Werte an der Gesamtvarianz.
Für die Varianz der wahren Werte gilt nun
Die Varianz der Differenz zweier wahrer Werte entspricht der Summe der Einzelvarianzen der wahren Werte abzüglich ihrer Kovarianz.
$$Var(T_x-T_y) = Var(T_x) + Var(T_y) - \underline{2Cov(T_x,T_y)}$$
Die Varianz der Differenz zweier Meßfehler entspricht der Summe ihrer Einzelvarianzen.
$$Var(F_x-F_y) = Var(F_x) + Var(F_y) - 0.$$

Das bedeutet insgesamt: Je höher die Korrelation der wahren Werte ist, desto geringer ist die wahre Varianz in den Differenzen. Das heißt: Je höher die Korrelation zwischen erster und zweiter Messung, desto niedriger ist die Reliabilität der Differenzen.
Insgesamt läßt sich zeigen, daß sich die Reliabilität von Differenzen folgendermaßen ergibt:

$$Rel(X-Y)=\frac{\sigma^2(X)Rel(X) + \sigma^2(Y)Rel(Y) - 2\rho(XY)\sigma(X)\sigma(Y)}{\sigma^2 + \sigma^2 - 2\rho(XY)\sigma(X)\sigma(Y)}$$

Die Unreliabilität von Differenzen ist daher ein erhebliches Problem bei der individuellen Vorhersage z.B. von Schulleistungen.

(C) Korrelation von Differenz und Ausgangswert

Eine häufige Frage in klinischen Studien ist, die, ob die Veränderung eines Merkmals im Verlauf der Therapie oder des Trainings mit dem Ausgangswert dieses Merkmals vor Trainings- oder Therapiebeginn korreliert. Wird das Merkmal perfekt reliabel gemessen, so entspricht diese inhaltliche Fragestellung der Korrelation: $\rho[(T_2\text{-}T_1),T_1]$.

Problem: Angenommen, in Wirklichkeit bestünde kein Zusammenhang zwischen dem Ausgangswert und dem Veränderungswert. Wie würde sich dann nicht perfekt reliable Messungen vorgenommen werden?

Ausgangspunkt:

$$\rho[(T_2\text{-}T_1),T_1] = 0$$

Es läßt sich folgendes zeigen:

$$\begin{aligned}\text{Cov}[(X_2\text{-}X_1), X_1] &= \text{Cov}\,[(T_2\text{-}T_1 + F_2\text{-}F_1), (T_1+F_1)]^a \\ &= \text{Cov}\,[(T_2\text{-}T_1), T_1] + \text{Cov}[(F_2\text{-}F_1), F_1]^b \\ &= 0^{\,c} - \text{Var}(F_1) \\ &= -\text{Var}\,(F_1).\end{aligned}$$

Erläuterungen:
a: der beobachtete Wert (X) setzt sich zusammen aus wahrem Wert (T) und Fehler (F); b: Die Kovarianzen von wahrem Wert und Fehler sind additiv; c: „Cov $[(T_2\text{-}T_1), T_1]$“ ist in diesem Beispiel per definitionem gleich 0.

Fazit: Bei nicht perfekt-reliabler Messung geht der Meßfehler bei der ersten Messung (X_1) in Ausgangswert und Differenz mit entgegengesetzten Vorzeichen eingeht und liefert daher einen negativen Beitrag zu Kovarianz. Sind Ausgangslage und Zuwachs nicht korreliert, wird die beobachtete Korrelation negativ; sind sie positiv korreliert, so wird die beobachtete Korrelation reduziert.

Sofern die Reliabilität der Messung bekannt ist, läßt sich eine Korrekturformel einsetzen:
$\rho[(X_2\text{-}X_1),X_1´] = \rho[(X_2\text{-}X_1), X_1] + (\text{Var}(F_1)/\text{sqrt}(\text{Var}(X_2\text{-}X_1)\text{Var}(X_1)))$

$\rho[(X_2\text{-}X_1),X_1´]$: Korrelation zwischen beobachtetem Ausgangswert und beobachtetem Zuwachs bei unabhängigen Meßfehlern. Die Korrelation zwischen den wahren Werten $\rho[(T_2\text{-}T_1),T_1]$ ergibt sich erst nach Anwendung der Minderungskorrektur:
$\rho[(T_2\text{-}T_1),T_1] = \rho[(X_2\text{-}X_1),X_1´] * \text{sqrt}(\text{Rel}(X_2\text{-}X_1)* \text{Rel}(X_1))$

Ende
Testtheorie-Exkurs: Reliabilität von Differenzen

(D) Regressionseffekte
Statistische Regression wurde in Abschnitt* 5.3 als eine Gefährdung der internen Validität bezeichnet und kann nur im Kontext von Meßwiederholungen auftreten. Immer dann, wenn es eine Regression zur Mitte gibt, existiert eine negative Korrelation zwischen Ausgangswert und Zuwachs; beide Probleme sind letztlich nur verschiedene Darstellungen desselben Sachverhalts.

Klassisches Beispiel: Nach Galton haben große Väter durchschnittlich kleinere Söhne, kleine Väter hinngegen durchschnittlich größere Söhne. Der Fehlschluß aus diesem Regressionseffekte ist, daß sich über die Generationen hinweg die Körpergröße der Männer angleicht.

Welche Korrelation besteht nun für den „Zuwachs“ (negativ oder positiv, über 2 Generationen) an Körpergröße und dem „Ausgangswert“ der Körpergröße?

X_1 = Körpergröße der Väter (erste Messung)
X_2 = Körpergröße der Söhne (zweite Messung)
X_2-X_1 = „Zuwachs“ an Körpergröße der Söhne gegenüber den Vätern
$\rho(X_1,X_2)$ = Korrelation der Körpergrößen von Vater und Sohne, z.B. r=.50
$\sigma(X_1) = \sigma(X_2)$ = Streuung der Körpergröße, für Väter und Söhne gleich

Die Korrelation einer Differenz mit irgendeiner anderen Variable errechnet sich aus der Formel:

$$\rho[(X_2-X_1), X_1] = \frac{\rho(X_1X_2)\sigma(X_2)-\sigma(X_1)}{\text{sqrt}(\sigma^2(X_1)+ \sigma^2(X_1) - 2\rho(X_1X_2)\,\sigma(X_1)+ \sigma(X_1)}$$

In diesem Beispiel läßt sich aufgrund $\sigma(X_1) = \sigma(X_2)$ diese Formel nach einiger Umformung vereinfachen zu: $\rho[(X_2-X_1), X_1] = -\text{sqrt}(0.5(1-\rho(X_1X_2)))$

Im Beispiel: $\rho[(X_2-X_1), X_1] = -\text{sqrt}(0.5(1-.50)$
$= -\text{sqrt}(.25) = -.50$

Fazit: Die Korrelation zwischen der Körpergröße der Väter und dem „Zuwachs“ bzw. der Veränderung der Körpergröße der Söhne gegenüber den Vätern beträgt -.50. Das bedeutet, daß bei gleicher Varianz von erster und zweiter Messung die Korrelation von Ausgangswert und Zuwachs alleine von der Korrelation zwischen erster und zweiter Messung abhängt und keinerlei zusätzliche Information enthält.

Gleiche Varianzen bei erster und zweiter Messung sind bzgl. Der Körpergröße von Vätern und Söhnen ein biologisches Faktum, keine mathematische Notwendigkeit. Anders ist es aber bei denjenigen psychologischen Anwendungen, bei denen aufgrund der Standardisierung (z.B. z-Transformation) die Streuungen von erster und zweiter Messung gleichgesetzt werden (Beispiel Intelligenz-Diagnostik: hier wird per Standardisierung der Mittelwert auf 100, die Streuung auf 15 gesetzt). Eine Regression zur Mitte, d.h. eine negative Korrelation zwischen Ausgangswert und Zuwachs, ergibt sich dann zwangsläufig.

Regressionseffekte können auch den Korrelationen von Zuwachs und anderen Variablen überlagert sein. Die Korrelation zwischen Zuwachs und irgendeiner anderen Variablen W wird über folgende Formel berechnet:

$$\rho[(X_2-X_1),W] = \frac{\rho(X_2W)\sigma(X_2) - \rho(X_1W)\sigma(X_1)}{\text{sqrt}(\sigma^2(X_1) + \sigma^2(X_2) - 2\rho(X_1X_2)\sigma(X_1)\sigma(X_2))}$$

Haben X_1 und X_2 gleich Streuungen, so gilt:

$$\rho[(X_2-X_1),W] = \frac{\rho(X_2W) - \rho(X_1W)}{\text{sqrt}(2(1-\rho(X_1X_2)))}$$

Aus dieser Formel folgt, daß bei gleicher Streuung von erster und zweiter Messung jede Variable W, die mit der ersten Messung höher korreliert als mit der zweiten, mit dem Zuwachs negativ korrelieren muß.

Beispiel: Es wird angenommen, daß Menschen im Laufe der Jahre in unterschiedlichem Maße Sport treiben und ein unterschiedliches Gewicht haben; außerdem, daß zwischen Ausmaß an

sportlicher Betätigung und dem Körpergewicht eine negative Korrelation besteht (d.h. je mehr Sport, desto geringer das Gewicht). Nun wird eine Anzahl an Personen im Abstand von 5 Jahren nach dem Ausmaß ihrer sportlichen Betätigung gefragt und gewogen.

Wie sieht die Korrelation zwischen dem Ausmaß der sportlichen Betätigung zu Beginn der Untersuchung und der Veränderung des Gewichts 5 Jahre später aus?

Zwei Annahmen erscheinen plausibel:
(1) das Gewicht hat zu beiden Meßzeitpunken dieselbe Varianz ($\sigma(X_1)=\sigma(X_2)$)
(2) die sportliche Betätigung zum ersten Meßzeitpunkt korreliert stärker mit dem Gewicht zum ersten als dem Gewicht zum zweiten Meßzeitpunkt.

Also: $\rho(X_1W) = -0.3$ $\rho(X_2W) = -0.1$ $\rho(X_1X_2) = 0.7$

W : Ausmaß der sportlichen Betätigung zum ersten Meßzeitpunkt; X_1, X_2 : Gewicht zum ersten bzw. zweiten Meßzeitpunkt

Gemäß der Formel:

$$\rho[(X_2-X_1),W] = \frac{\rho(X_2W) - \rho(X_1W)}{\text{sqrt}(2(1-\rho(X_1X_2)))}$$

ergibt sich nach Einsetzen und Ausrechnen:

$$= \frac{(-0.1) - (-0.3)}{\text{sqrt}(2(1-0.7))}$$

$$= 0.26$$

Das heißt: Wer viel Sport treibt, tendiert zu einer höheren Gewichtszunahme nach 5 Jahren!
Das bedeutet inhaltlich: Es gibt eine Regression zur Mitte. Unter den eher Unsportlichen werden einige sportlicher geworden sein, während manche Sportlichen weniger Sport betreiben werden. Aufgrund des Zusammenhangs von Sport und Körpergewicht, nehmen die Sportlichen daher mehr zu als die Unsportlichen.

Allgemeines Fazit: Die Interpretation von Korrelationen beliebiger Variablen mit der als Differenz gemessenen Veränderung führt immer dann zu Fehlschlüssen, wenn existierende Regressionseffekte nicht erkannt und interpretiert werden.

Testtheorie-Exkurs: Reliabilität von Residualscores
Anfang

(E) Residualscores
Fragen nach interindividuellen Unterschieden in der Veränderung, also z.B. individuelle Lernfortschritte, haben oftmals die Form: „Hat jemand mehr oder weniger dazugelernt als der Durchschnitt vergleichbarer Personen? Und wenn ja: weshalb? Das heißt: gibt es Variablen, mit denen ein über- oder unterdurchschnittlicher Zuwachs zusammenhängt?“

Solche Fragen lassen sich beantworten, wenn man Residualscores bildet. Das heißt, daß aufgrund der ersten Messung eine Regressionsschätzung der zweiten Messung ($\hat{X}_2$) vorgenommen und geprüft wird, inwieweit der Proband von dieser Vorhersage abweicht:
Residualscore = $X_2 - \hat{X}_2$.

Sofern der Regressionsverlauf linear ist (bivariate Normalverteiltheit gegeben?), als Durchschnittswert interpretiert werden, den Personen mit gleichem Ausgangswert wie der interessierende Proband erreichen. Zweck der Berechnung von Residualscores ist die Identifikation von Variablen, die individuelle Unterschiede im Zuwachs unabhängig von der Ausgangslage erklären.

Die Korrelation einer Variablen W mit dem Residualscore ($X_2-\hat{X}_2$) läßt sich mittels der Formel für die Semipartialkorrelation bestimmen:

$$\rho[(X_2-\hat{X}_2),W] = \frac{\rho(X_2W) - \rho(X_1W)\rho(X_1X_2)}{\sqrt{1 - \rho^2(X_1X_2)}}$$

Für die Berechnung der Semipartialkorrelation benötigt man also nur die einfachen Korrelationen der 3 Variablen (X_1,X_2,W) untereinander.

Wie sieht die Reliabilität der Residualscores aus?

$$Rel(X_2-\hat{X}_2) = \frac{Rel(X_2) - \rho(X_1X_2)(2-Rel(X_1))}{1 - \rho^2(X_1X_2)}$$

Analog zur Reliabilität der Differenzen ist auch die Reliabilität der Residualscores niedrig, wenn erste und zweite Messung hoch korrelieren.

	Rel(Residualscore) $X_2-\hat{X}_2$	Rel(Differenzwert) X_2-X_1
$\rho(X_1X_2)$	$Rel(X_1) = Rel(X_2) = 0.90$	
0.50	0.83	0.80
0.60	0.79	0.75
0.70	0.71	0.67
0.80	0.54	0.50
0.90	0.05	0.00

Wie ist die Reliabilität der Residualscores in dem Beispiel zu Sport und Körpergewicht, an dem wir zuvor die Reliabilität der Differenzwerte erläutert hatten?

Zu Erinnerung: $\rho(X_1W) = -0.3$ $\rho(X_2W) = -0.1$ $\rho(X_1X_2) = 0.7$

W : Ausmaß der sportlichen Betätigung zum ersten Meßzeitpunkt; X_1, X_2 : Gewicht zum ersten bzw. zweiten Meßzeitpunkt

Gemäß der Formel für die Semipartialkorrelation:

$$\rho[(X_2-\hat{X}_2),W] = \frac{\rho(X_2W) - \rho(X_1W)\rho(X_1X_2)}{\sqrt{1 - \rho^2(X_1X_2)}}$$

errechnen wir:

$$= \frac{-0.1 - (-0.3*0.7)}{\dots} = \frac{0.11}{\dots} = 0.15$$ (gegenüber 0.26 für den Differenzwert)

$\text{sqrt}(1 - 0.70^2)$ 0.71

Kommentar: Obgleich auf niedrigerem Niveau legt auch dieser Wert denselben Trugschluß nahe. Die Fehlinterpretationen hätten vermieden werden können, wenn man auch zum zweiten Meßzeitpunkt das Ausmaß der sportlichen Betätigung erfaßt hätte. Dann hätte man die sportlich gebliebenen mit den unsportlich gewordenen Probanden vergleichen können.

Ende
Testtheorie-Exkurs: Reliabilität von Residualscores

6.6.4 Allgemeine Veränderung: experimentelle Versuchspläne

Problemstellung: Sehr häufig möchte man in Gruppenuntersuchungen der Frage nachgehen, ob bestimmte Interventionen (Treatments) zu bestimmten Veränderungen führen.

Beispiel: Mittels eines Förderprogramms sollen pädagogische Psychologen die Intelligenzentwicklung von Vorschulkindern aus sozial-benachteiligten Stadtteilen positiv beeinflussen. Der Effekt des Förderprogramms soll jedoch zuvor überprüft werden.
Je nachdem, wie „frei" die Untersuchung planbar ist, ergeben keine oder sehr große methodische Probleme bei deren Auswertung. Klar ist, daß man eine Experimental- und eine Kontrollgruppe benötigt; entscheidend ist, ob die Probanden der Experimental- und Kontrollbedingung (*A*) zufällig zugeteilt werden können oder (*B*) nicht.

(A) Zufallsaufteilung zu den Grupen: Aufgrund der Randomisierung ist die Ausgangslage bei hinreichend großem N in den beiden Gruppen gleich. Es reicht, am Ende des Förderprogramms Leistungsunterschiede zwischen den Gruppen auf statistische Signifikanz und praktische Bedeutsamkeit zu prüfen (vgl. VPL#1, oben). Auch ließe sich ein Vortest durchführen, dessen Ergebnisse kovarianzanalytisch nutzbar wären.
Ist eine randomisierte Zuteilung zu den Gruppen nicht möglich, sondern „teilen" sich die Probanden der Experimental- und Kontrollbedingung „selber zu", so ist von unterschiedlichen Ausgangslagen zu Beginn des Förderprogramms auszugehen.

(B) Nicht-zufällige Zuteilung zu den Gruppen:
Selektion nach dem Vortest: Die Experimental- und Kontrollgruppe werden aufgrund der Leistungen der Probanden in einem Vortest gebildet.
Stichproben aus unterschiedlichen Populationen: Experimental- und Kontrollgruppe stammen gar aus unterschiedlichen Populationen (z.B. Stadtteilen).

In beiden Fällen besteht das Problem darin, daß eine nicht bekannte Vielzahl an Variablen existieren mag, die die Unterschiede zwischen den Gruppen bisher bedingt hat und im Verlauf der Untersuchung fortwirkt.

(VPL#9) Prätest-Posttest Kontrollgruppenplan
Fragestellung & Vorgehen: Zusätzlich zum zuvor dargestellten Versuchsplan wird eine Vorher-Messung durchgeführt. Die Pbn werden *per Zufall* der Experimental- bzw. der Kontrollgruppe zugeteilt. Zunächst werden eine oder mehrere AV(s) werden erhoben (Yb), danach *nur* die Experimentalgruppe dem Treatment „ausgesetzt", und schließlich die AV(s) erneut in beiden Gruppen erhoben (Ya). Unterschiede in Ya zwischen EG und KG werden auf das Treatment zurückgeführt.

Struktur

	Yb	X	Ya	EG
R				
	Yb	~X	Ya	KG

SPSS- Auswertung: Varianzanalyse mit Meßwiederholung, MANOVA

Sonstige Anmerkungen: Wechselwirkung Testen (Yb) *x* X: solche „Sensibilisierungseffekte“ sind insbesondere in motivations- und sozialpsychologischen Untersuchungen zu erwarten. Zur Kontrolle potentiell validitätsmindernder Prätests läßt sich der sog. SOLOMON-4 Gruppenplan einsetzen (siehe unten). Der Prätest-Posttest-Kontrollgruppen-Plan unterscheidet sich vom Prätest-Posttest-Plan *vor allem* in der zufälligen Zuteilung der Pbn zu den Untersuchungsgruppen. Dadurch wird gewährleistet, daß sich die Gruppen *vor* Applikation des Treatments nicht unterscheiden (hinreichend großes N vorausgesetzt).
Aufgrund der Randomisierung sind Unterschiede in Ya zwischen EG und KG eindeutig auf X zurückführbar. Jede Gefährdung der internen Validität („zwischenzeitliches Geschehen, Reifung, usf.), die auftreten könnte, mag zwar Ya beeinflussen, doch dann aufgrund der Randomisierung für beide Gruppen gleichermaßen – und kann daher den Unterschied zwischen den Gruppen in Ya nicht erklären. Daher sind Unterschiede in Ya zwischen den Gruppen auch nicht als Folge etwaiger Gefährdungen der internen Validität interpretierbar.
Es ist denkbar, daß die anfängliche Erhebung der AV(s) (Yb) die Wirkung des Treatments X beeinflußt. Diese denkbare Wechselwirkung tangiert zwar nicht den Gruppenunterschied in Ya, da ja auch die KG Yb erhielt, wohl aber die Generalisierbarkeit der Wirkung von X: sie könnte nur für solche Pbn gelten, die vor dem Treatment Yb „erhielten“.

Statistik-Exkurs: Varianzanalytische Verfahren zur Veränderungsmessung

(A) Einfaktorielle Varianzanalyse mit Meßwiederholung

Vorbemerkungen: Ein Versuchsplan mit Meßwiederholung zeichnet sich dadurch aus, daß die Ausprägungen einer oder mehrerer abhängiger Variablen unter mindestens 2 (experimentellen) Bedingungen (oder Zeitpunkten) gemessen werden.

Ein einfaches statistisches Verfahren zur Auswertung von Versuchsplänen mit Meßwiederholung ist der t-Test für abhängige Stichproben. Eine Erweiterung dieses t-Tests ist die Varianzanalyse mit Meßwiederholung.

Zur Wiederholung aus Statistik II:

Varianzanalytische Modelle	
Modell	Effekte-Konzept
Model I: fixed effects	Die Stufen des Faktors sind festgelegt und bilden die Grundgesamtheit möglicher Faktorstufen.
Model II: random effects	Die Stufen des Faktors werden als Zufallsstichprobe aus einer Grundgesamtheit möglicher Faktorstufen verstanden.
Model III: mixed effects	Feste und Zufallseffekte werden im selben Modell umgesetzt.

Einordnung der Varianzanalyse mit Meßwiederholung: Die Varianzanalyse mit Meßwiederholung ist ein Spezialfall der Situation im Model III: der mixed models mit festen und Zufalls-Faktoren:

Fester Faktor: die *J* experimentellen Treatments
Zufallsfaktor: die *n* (zufällig ausgewählten) Probanden

Wert des Probanden *i* unter dem Treatment *j*: Y_{ij}.

Lineares Modell: $Y_{ij}=\mu+\alpha_j+a_i+e_{ij}$.
Mit: $\alpha_j=\mu_j-\mu$ $\quad {}_j\Sigma\alpha_j=0$
a_i: Zufallseffekt des Probanden *i*,
e_{ij}: Zufallsfehler bei Proband *i* unter Treatment *j*.

Anmerkung: In diesem Modell werden nur der Probanden- und der Treatment-Effekt, nicht aber deren Interaktion behandelt.

Weitere Modell-Annahmen:

Modell-Annahmen	Erläuterung
$E(e_{ij})=0$	Über die gesamten Population der Probanden hinweg ist der Mittelwert des Fehlerterms unter Treatment *j* gleich 0.
$E(a_i)=0$	Der Mittelwert der individuellen Effekte a_i ist in der Population der Individuen i gleich 0.
$Cov(a_i, e_{ij})=0$	Die individuellen Effekte a_i und die Fehlereffekte e_{ij} sind unabhängig voneinander.
$Cov(e_{ij}, e_{ik})=0$	Die Fehler in Treatment *j* sind unabhängig von den Fehlern in Treatment *k*, und zwar für alle Individuen *i* und alle Treatment-Paare *j* und *k*.
Die Verteilung der Effekte a_i ist normal.	
Die Verteilung der Fehler e_{ij} ist für jedes Treatment *j* normal.	
Die Fehlerterme e_{ij} haben dieselbe Varianz (σ^2_e) für alle Treatments *j*.	

Nicht angenommen wird, daß die Beobachtungen bei einem Probanden unter 2 verschiedenen Bedingungen voneinander unabhängig sind, d.h. $Cov(Y_{ij},Y_{ik})$ muß *nicht* gleich null sein. Das bedeutet, daß in diesem Modell ein gewisses Maß an Abhängigkeit zwischen den Beobachtungen aufgrund desselben Probanden zulässig ist.

Varianzanalyse ohne Meßwiederholung

p Treatment-Stufen bzw. Untersuchungsgruppen mit n Individuen

a_1	a_2	...	a_p
y_{11}	y_{12}	...	Y_{1p}
y_{21}	Y_{22}	...	Y_{2p}
y_{31}	Y_{32}	...	y_{3p}
.	.	...	.
.	.	...	.
.	.	...	.
y_{n1}	Y_{n2}	...	y_{rp}
Varianz innerhalb der Gruppen			
← Varianz zwischen den Treatmentstufen→			

Kommentar: Unter jeder der p Treatment-Stufen werden n Versuchspersonen untersucht.
$Y_{ij}=\mu+\alpha_j+e_{ij}$. α_j=fester Effekt, Treatment j

Quadratsummen-Zerlegung & F-Test		
$SS_{total} = {}_i\Sigma\,{}_j\Sigma\,(y_{ij}-{}^{GM}y)^2$	$SS_{total} = SS_{between} + SS_{within}$	$F=\frac{SS_{between}/p-1}{SS_{error}/N-p}$
$SS_{between} = {}_j\Sigma\, n_j({}^{M}y_j-{}^{GM}y)^2$	$SS_{within} = SS_{total} - SS_{between}$	
$SS_{within} = {}_j\Sigma\,{}_i\Sigma\,(y_{ij}-{}^{M}y_j)^2$	"SS_{within}" = "SS_{error}"	
		^{M}=Mean of; ^{GM}=Grand Mean of.

Varianzanalyse mit Meßwiederholung

p Treatment-Stufen ↔ p Meßwerte von n Individuen

a_1	a_2	...	a_p	
y_{11}	y_{12}	...	y_{1p}	↑ Varianz zwischen den Vps ↓
y_{21}	y_{22}	...	y_{2p}	
y_{31}	y_{32}	...	y_{3p}	
.	.	...	.	
.	.	...	.	
.	.	...	.	
y_{n1}	y_{n2}	...	y_{rp}	

← Varianz zwischen den Treatmentstufen→

Kommentar: Für jede Versuchsperson wird unter allen p Treatment-Stufen 1 Meßwert erhoben.

$Y_{ij} = \mu+\alpha_j+a_i+e_{ij}$. α_j=fester Effekt, Treatment j

a_i=Zufallseffekt, Vp i

Quadratsummenzerlegung & F-Test		
$SS_{total} = {}_i\Sigma\,{}_j\Sigma\,(y_{ij}-{}^{GM}y)^2$	$SS_{total} = SS_{between.subjects} + SS_{within.subjects}$	$F=\frac{SS_{between.treatment}/p-1}{SS_{residual}/(N-1)(p-1)}$
$SS_{between.subjects} = {}_j\Sigma\,{}_i\Sigma\,(y_{ij}-{}^{M}y_j)^2$	$SS_{total} = SS_{between.subjects} + SS_{between.treatments} + SS_{residual}$	
$SS_{within.subjects} = {}_j\Sigma\,{}_i\Sigma\,(y_{ij}-{}^{M}y_i)^2$ $SS_{between.treatments} = {}_j\Sigma\, n_j({}^{M}y_j-{}^{GM}y)^2$ $SS_{residual} = SS_{total} - SS_{between.treatments} - SS_{between.subjects}$	"$SS_{residual}$" = "SS_{error}" = "$SS_{betweenSubxA}$" = "*interaction variance*"	
		^{M}=Mean of; ^{GM}=Grand Mean of.

Vergleich der Varianzanalysen mit versus ohne Meßwiederholung: Bei der repeated measures ANOVA wird im Vergleich zur ANOVA ohne Meßwiederholung immer hervorgehoben, daß sie besonders geeignet sei, den Meßfehler bzw. die Störvarianz zu reduzieren (vgl. auch Abschnitt* 6.3).

Hierzu einige Zitate:

Winer (1971, p. 261): „*Because of large differences in experience and background, the responses of people to the same experimental treatments may show relatively large variability. (...) If this latter source of variability can be separated from treatment effects and experimental error, then the sensitivity of the experiment can be increased. If this source* of *variability cannot be estimated, it remaines part of the uncontrolled sources of variability and is thus automatically part of the experimental error.*"

Hays (1973, p. 568): "*The practical limit to the strategy of matching subjects in an experiment is reached when each subject is matched with himself. (...) Conceptually, each group of J observations of a single subject is like a matched group of observations: (...)*"

Stevens (1986, p. 402): „*In repeated measures designs, blocking is carried out to it´s extreme. That is, we are blocking on each subject. Thus, variability among the subjects due to individual differences is completely removed from the error term. This makes these designs much more powerful than completely randomized designs, where different subjects are randomly assigned to the different treatments.*“

Warum die repeated measures ANOVA die Fehlervarianz reduziert.

Ein Rechenbeispiel:

	Treatments				
Vps	1	2	3	4	means
1	30	28	16	34	27
2	14	18	10	22	16
3	24	20	18	30	23
4	38	34	20	44	34
5	26	28	14	30	24.5
Means	26.4	25.6	15.6	32	24.9 (GM)

(nach Winer, 1973 & Stevens, 1986)

Anmerkung: Es ist in der äußerst rechten Spalte leicht zu erkennen, daß sich die Probanden deutlich unterscheiden. Dieser Datensatz sei nun auf zweifache Weise ausgewertet: (1) als befänden sich unter den 4 Treatment-Stufen verschiedene Probanden (keine Meßwiederholung, sondern Gruppenvergleich); (2) als würden 5 Probanden jeweils mit 4 Treatments behandelt (mit Meßwiederholung).

Zu Erinnerung:

Varianzanalyse ohne Meßwiederholung	*Varianzanalyse mit Meßwiederholung*
$SS_{total} = SS_{between} + SS_{within}$ $SS_{within} = SS_{total} - SS_{between}$	$SS_{total} = SS_{between.subjects} + SS_{within.subjects}$ $SS_{total} = SS_{between.subjects} + SS_{between.treatments} + SS_{residual}$ $SS_{residual} = SS_{total} - SS_{between.treatments} - SS_{between.subjects}$
$F = \frac{SS_{between}/p-1}{SS_{within}/N-p}$	$F = \frac{SS_{between.treatment}/p-1}{SS_{residual}/(N-1)(p-1)}$

Rechenbeispiel: Varianzanalyse ohne Meßwiederholung

$SS_{between} = 5[(26.4-24.9)^2 + (25.6-24.9)^2 + ... + (32-24.9)^2] = 698.2$
$SS_{within} = (30-26.4)^2 + (14-26.4)^2 + ...+ (30-32)^2 = 793.6$

$SS_{between}/p-1$ 698.2 / 3 232.73

$F = \frac{SS_{between}/p-1}{SS_{within}/N-p} = \frac{698.2 / 3}{793.6 / 16} = \frac{232.73}{49.6} = 4.7$ (←Fehlerterm)

$SS_{residual}/(N-1)(p-1)$

Rechenbeispiel: Varianzanalyse mit Meßwiederholung

$SS_{between} = 5[(26.4-24.9)^2 + (25.6-24.9)^2 + ... + (32-24.9)^2] = 698.2$
$SS_{within.subjects} = (30-26.4)^2 + (14-26.4)^2 + ...+ (30-32)^2 = 793.6$
$SS_{between.subjects} = 4\,[(27-24.9)^2 + .. + (16-24.9)2 + (24.5-24.9)^2] = 680.8$
$SS_{residual} = 793.6 - 680.8 = 112.8$

$$F = \frac{SS_{between.treatments}/p\text{-}1}{SS_{residual}/(n\text{-}1)(p\text{-}1)} = \frac{698.2 / 3}{112.8 / (4*3)} = \frac{232.73}{9.4} = 24.76$$

(←Fehlerterm)

Die Berücksichtigung der Kovarianzen.

Es läßt sich zeigen, daß die Kovarianz der Beobachtungen unter 2 beliebigen Bedingungen der Varianz der individuellen Effekte über alle Individuen hinweg ist.

$$Cov(Y_{ij},Y_{ik}) = E(a_i^2)$$

Die Varianz der Werte Y_{ij} unter einem gegebenen Treatment j entspricht:

$$\sigma_{Yj}^2 = E(a_i^2) + E(e_{ij}^2) \quad \text{bzw.} \quad \sigma_Y^2 = \sigma_a^2 + \sigma_e^2$$

(keiner dieser Ausdrücke hängt von j ab)

Das bedeutet dann: $\sigma_e^2 = \sigma_Y^2 - \sigma_a^2$
bzw., weil $\sigma_a^2 = Cov(Y_{ij},Y_{ik})$
$\sigma_e^2 = \sigma_Y^2 - Cov(Y_{ij},Y_{ik})$

Die Kovarianz $Cov(Y_{ij},Y_{ik})$ muß für alle Treatmentpaare j und k gleich sein (statistische Annahme). Daher reicht es, von der mittleren Kovarianz ${}^{M}Cov$ zu sprechen: $\sigma_e^2 = \sigma_Y^2 - {}^{M}Cov$.

Das heißt: Die Fehlervarianz entspricht der Varianz der Werte innerhalb eines Treatments abzüglich der mittleren Kovarianz über die Treatments hinweg. Bei der Varianzanalyse ohne Meßwiederholung entspricht die Fehlervarianz (MS_{within}) der Varianz der Werte innerhalb der Treatments.

Eine andere Problemstellung.

Es soll überprüft werden, ob die Gabe eines stimulierenden Medikamentes die Konzentrationsfähigkeit bei Kindern mit Hyperaktivität erhöht. Die Teststärke ist die Wahrscheinlichkeit, mit der bei Zutreffen der Alternativhypothese ein signifikantes Ergebnis eintritt und entspricht (1-β). Die Teststärke ist unter sonst gleichen Umständen um so größer, je größer der Effekt ist. Die Effektgröße entspricht der (ewarteten) Mittelwertsdifferenz relativ zur Standardabweichung innerhalb der Gruppen. Je größer die Mittelwertsdifferenz zwischen den Gruppen im Verhältnis zur Standardabweichung innerhalb der Gruppen, desto größer ist die Teststärke.

Frage: Welcher der beiden nachfolgenden Versuchspläne hat die größere Effektstärke?

Versuchsplan I: Posttest-Kontrollgruppen-Plan	Versuchsplan II: Prätest-Posttest-Kontrollgruppen-Plan
KG: Placebo — K-Test R: EG: Stimul. — K-Test	KG: K-Test^1 — Placebo — K-Test^2 R: EG: K-Test^1 — Stimul. — K-Test^2
AV: Ergebnis im K-Test	AV: *Differenzwert* zwischen K-Test^2 und K-Test^1.

Mittelwertsdifferenz: in beiden Plänen gleich der Medikamentenwirkung

$$Var(X) > Var(X-Y)$$

wenn gilt: $r_{tt} \geq .50$

Konsequenz: Je stärker die erste und zweite Messung korrelieren und je niedriger damit die Reliabilität des Differenzmaßes ist, desto günstiger ist Versuchsplan II gegenüber Versuchsplan I.

Erklärung: Bei der Erfassung von "Treatment"-Effekten (Medikamentenwirkung) ist *jede* interindividuelle Varianz nachteilig. Die Varianz der wahren Werte wird bei der Differenzbildung nun um so kleiner, je stärker erste und zweite Messung korreliert sind.

Fazit: Je höher die erwartete Korrelation zwischen erster und zweiter Messung voraussichtlich sein wird (ist r_{tt} vorab bekannt?), desto größer ist die relative Teststärke von Versuchsplan II gegenüber Versuchsplan I. Alternativ sollte geprüft werden, ob die Voraussetzungen für eine Kovarianzanalyse erfüllt sind. Wenn ja, ist diese vermutlich der Auswertung von Differenzen überlegen.

Nachteile der Varianzanalyse mit Meßwiederholung.

Ein Problem der Varianzanalyse mit Meßwiederholung *kann* in „carry-over"-Effekten von einer zu anderen Treatment-Stufe bestehen (siehe auch Abschnitt* 6.3). *Nicht gemeint ist*: die intraindividuelle Konsistenz im Verhalten.

Gemeint ist: der Einfluß eines Treatments auf die Leistung in einem der nachfolgenden Treatments.

Beispiele sind: Lern-, Übungs-, Ermüdungseffekte; (verzögerte) Wirkungen psychoaktiver Substanzen, generell: „Reifungseffekte", die durch ein Treatment angestoßen werden.

Wichtigste Schutzmaßnahme: „Counterbalancing"

„*Counterbalancing*":= Es werden soviele Testreihenfolgen realisiert, wie erforderlich ist, damit jede Treatmentstufe in jedem Stadium der Untersuchung gleichhäufig auftritt.

	Treatments			
Sequenzen	a_1	a_2	a_3	a_4
1	1	2	3	4
2	2	4	1	3
3	3	1	4	2
4	4	3	2	1

Annahme: Es wird angenommen, daß eine Interaktion „Testreihenfolge *x* Treatment" nicht existiert, d.h. daß die Carry-Over-Effekte für jede Reihung der Treatments dieselbe ist. Eine solche Interaktion läßt sich ggf nur nachträglich feststellen. In diesem Fall sind nur die initialen Aufgaben frei von Carry-Over-Effekten.

Annahmen der Varianzanalyse mit Meßwiederholung.

1. Die Beobachtungen sind unabhängig voneinander:
   - diese Annahme betrifft den „between subjects"-Vergleich
   - innerhalb der Probanden ist eine Abhängigkeit der Beobachtungen „zulässig" (s.o.)
   - die Varianzanalyse mit Meßwiederholung ist wie jedes varianzanalytische Verfahren sehr unrobust gegenüber der Verletzung dieser Annahme
2. Die Beobachtungen (bzw. Effekte a_i) sind normalverteilt.
3. Sphärizität bzw. Zirkularität bzw. „compound symmetry" trifft zu.
   Erläuterung der Sphärizitätsannahme: Ebenso wie der t-Test für abhängige Stichproben, „rechnet" die Varianzanalyse mit Meßwieder-holung mit Differenzwerten, d.h. aus k Originalvariablen enstehen k-1 Differenzwerte.

   Die Sphärizitätsannahme besagt, daß die Varianz-Kovarianz-Matrix der Differenzwerte eine Diagonalmatrix ist mit gleichen Varianzen auf der Diagonalen, d.h.

	Transformierte Variablen (Differenzen)			
	1	2	.	k-1
1	σ^2	0	.	0
2	0	σ^2	.	0
.	.	.	.	.
k-1	0	0	.	σ^2

Varianzen: gleich; Kovarianzen: sämtlich 0.

Vielfach ist diese restriktive Annahme nicht erfüllt, was zu einer zu großen „Liberalität" des F-Tests führt. In dem Maße, in dem die Kovarianzmatrix von der Sphärizität abweicht, kann eine Korrektur der Freiheitsgrade des F-Tests vorgenommen werden. Das Greenhouse-Geisser ε $(1 \geq \varepsilon \geq (1/(k-1))$ stellt jedoch eine extrem konservative Freiheitsgrad-Korrektur dar. Die meisten Statistikprogramme geben auch die —weniger konservativ— nach Huynh-Feldt korrigierten F-Werte aus.

Post hoc-Verfahren für die Varianzanalyse mit Meßwiederholung.

Nach Auffinden eines signifikanten „overall" Meßwiederholungs-effektes bieten sich paarweise Vergleiche zwischen den Treatment-Stufen an. Ein gebräuchliches Verfahren ist das von Tukey.
Wenn gilt, daß ...

$$|{}^{M}y_i - {}^{M}y_j| > q_{.05;\, k;\, (n-1)(k-1)} \, \mathrm{sqrt}(MS_{res}/n)$$

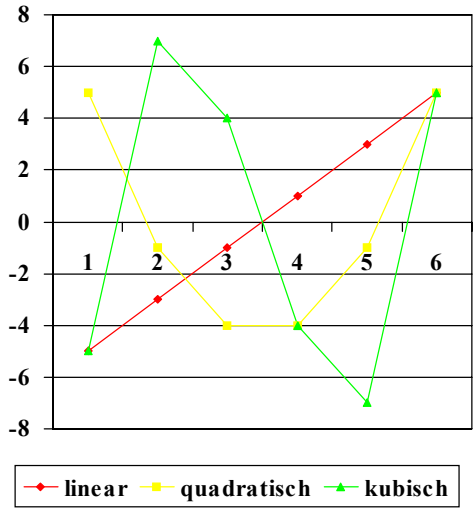
... ist der Bedingungsvergleich signifikant.

k=Anzahl Treatment-Stufen; n=Anzahl Probanden; (n-1)(k-1)= Fehlerfreiheits-grade; .05=5% Signifikanz-niveau; $q_{.05;\, k;\, (n-1)(k-1)}$: Studentized Range Statistic (tabelliert).

Trendanalysen.

Manchmal kann es sinnvoll sein, Veränderungsverläufe nachzuvollziehen und für sie eine statistisches Modell zu entwickeln. Die Varianzanalyse mit Meßwiederholung erlaubt die Entwicklung eines solchen Modells in Form einer Trendanalyse.

Beispiele für Trends:



Erläuterung: Die Abbildung zeigt einen linearen, quadratischen und kubischen Trend.

Trendanalysen innerhalb der Varianzanalyse mit Meßwiederholung bedeuten, daß die Treatment-Quadratsummen ($SS_{between.treatment}$) zerlegt werden in Komponenten, die mit einem bestimmten Trend verbunden sind.

Orthogonale Polynom-Koeffizienten:

	a_1	a_2	a_3	a_4	a_5	a_6	$\Sigma(c_i)^2$
linear	-5	-3	-1	1	3	5	70
quadratisch	5	-1	-4	-4	-1	5	84
kubisch	-5	7	4	-4	-7	5	180
weitere	...	...	...	...	...	...	...

Meßwerte der einzelnen Probanden

	a_1	a_2	a_3	a_4	a_5	a_6	
Vp1	7	3	2	2	1	1	
Vp2	4	8	3	8	1	2	
Vp3	*7*	*6*	*3*	*1*	*5*	*4*	
...							

Mit der Trendkomponente gewichtete Meßwerte

Beispiele:

Vp 1, linearer Trend: -5*7 + -3*3 + -1*2 + 1*2 + 3*1 + 5*1 = -36
Vp 2, linearer Trend: -5*4 + -3*8 + -1*3 + 1*8 + 3*1 + 5*2 = -26
Vp 3, linearer Trend: -5*7 + -3*6 + -1*3 + 1*1 + 3*5 + 5*4 = -20
usw

	Linear	quadratisch	kubisch	...
Vp1	-36	...	...	...
Vp2	-26	...	...	...
Vp3	-20	...	...	...
...	...	...	...	...
	$\Sigma[(c_i)(a_i)]^2$			

Nun kann für jede Trendkomponente die Quadratsumme berechnet werden:

$$SS_{trend} = \frac{\Sigma[(c_i)(a_i)]^2}{n[\Sigma(c_i)^2]}$$

Der Fehlerterm für eine Trendkomponente wird ermittelt über:

$$SS_{Trend*S} = \frac{{}_j\Sigma^s [\,{}_i\Sigma^a\ (c_i)(y_{ij})]^2}{\Sigma\,(c_i)^2} - \frac{[\,{}_i\Sigma^a\ (c_i)(A_i)]^2}{n[\Sigma(c_i)^2]}$$

Schließlich kann ein F-Test gerechnet werden, wenn die Quadratsumme „zu Lasten" der jeweiligen Trendkomponente und der zugehörige Fehlerterm durch ihre Freiheitsgrade geteilt werden.

$$F = \frac{SS_{trend} / 1}{SS_{Trend*S} / (n-1).}$$

(B) MANOVA zur Auswertung von Versuchsplänen mit Meßwiederholung

Die multivariate Varianzanalyse kann als Verallgemeinerung des univariaten t-Tests für abhängige Stichproben aufgefaßt werden.

Vorgehen beim t-Test für abhängige Stichproben.

Vps	Prätest	Posttest	Differenz
1	7	10	3
2	5	4	-1
3	6	8	2
.	.	.	.
n	3	7	4

H_0: $\mu_1=\mu_2$ bzw. $\mu_1-\mu_2=0$

$$t=\frac{{}^Md}{s_d/\mathrm{sqrt}(n)}$$

${}^Md$: mittlerer Differenzwert

s_d : Streuung der Differenzwerte

Verallgemeinerung bei MANOVA-Analysen von Meßwiederholungsdaten.

Die Teststatistik für k Meßwiederholungen wird aus den (k-1) Differenzwerten, ihren Varianzen und Kovarianzen gebildet.

Unabhängige Stichproben	Abhängige Stichproben
$t^2 = \frac{({}^My_1-{}^My_2)^2}{s^2(1/n_1 + 1/n_2)}$	$t^2 = \frac{({}^Md)^2}{s_d{}^2/n}$
$t^2=\frac{n_1n_2}{n_1+n_2} ({}^My_1-{}^My_2)(s^2)^{-1}({}^My_1-{}^My_2)$	$t^2= n\,({}^Md)(s_d{}^2)^{-1}({}^Md)$
Mittelwerte werden durch Mittelwertsvektoren, die gepoolte Within-Varianz (s^2) durch die gepoolte Within-Kovarianzmatrix ersetzt:	Die mittlere Differenz wird durch einen Vektor von Mittelwertsdifferenzen, die Varianz der Differenzwerte durch die Varianz-Kovarianz-

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (y_1 - y_2)' (S)^{-1} (y_1 - y_2)$$

Matrix der (k-1) Differenzvariablen ersetzt.

↓

$$T^2 = n\, y_d'\, S_d^{-1}\, y_d$$

y_d' : Zeilenvektor der (k-1) Mittelwertsdifferenzen; S_d: Varianz-Kovarianz-Matrix der (k-1) Differenzvariablen

Das Rechenbeispiel von vorhin.

	Treatments				
Vps	1	2	3	4	means
1	30	28	16	34	27
2	14	18	10	22	16
3	24	20	18	30	23
4	38	34	20	44	34
5	26	28	14	30	24.5
Means	26.4	25.6	15.6	32	24.9 (GM)

	Differenzen		
Vps	y_1-y_2	y_2-y_3	y_3-y_4
1	2	12	-18
2	-4	8	-12
3	4	2	-12
4	4	14	-24
5	-2	14	-16
Means	0.8	10	-16.4
Varianzen	13.2	26	24.8

$y_d' = (0.8, 10, -16.4)$
Kovarianz der ersten beiden Differenzvariablen:
Beispielrechnung:
$s_{y1-y2,\ y2-y3} = ((2-0.8)(12-10) + (-4-0.8)(8-10) + ... + (-2-0.8)(14-10))/4$
$= -3$

$$S_d = \begin{array}{c|ccc|} & y_1-y_2 & y_2-y_3 & y_3-y_4 \\ & 13.2 & -3 & -8.95^* \\ & -3 & 26 & -19^{**} \\ & -8.95 & -19 & 24.8 \end{array}$$

*=Kovarianz (y_1-y_2) & (y_3-y_4);
**= Kovarianz (y_2-y_3) & (y_3-y_4); usw

$$T^2 = y_d' \quad S_d^{-1} \quad y_d$$

$$T^2 = 5 \begin{pmatrix} .8, & 10, & -16.4 \end{pmatrix} \begin{vmatrix} .455 & .381 & .450 \\ .381 & .407 & .444 \\ .450 & .444 & .536 \end{vmatrix} \begin{pmatrix} .8 \\ 10 \\ -16.4 \end{pmatrix}$$

$$= 169.48$$

$$F = \frac{n-k+1}{(n-1)(k-1)} T^2$$

$$= \frac{5-4+1}{4(3)} (169.48)$$

$F_{(df1=3,\ df2=2)} = 28.25$

Ende
Statistik-Exkurs: Varianzanalytische Verfahren zur Veränderungsmessung

(VPL#9) Solomon Viergruppenplan

Fragestellung & Vorgehen: Per Zufall werden die Pbn den 4 Untersuchungsgruppen zugeteilt. Dies gewährleistet die „Gleichheit" der Gruppen vor Beginn der Untersuchung. Sodann erhält die Experimentalgruppe (EG) einen Vortest (Yb), das Treatment (X) und einen Nachtest (Ya). Die 1. Kontrollgruppe (KG1) erhält Vortest (Yb) und Nachtest (Ya), doch kein Treatment (X); die 2. Experimentalgruppe (EG2) erhält Treatment (X) und Nachtest (Ya), doch keinen Vortest (Yb); die 2. Kontrollgruppe (KG2) erhält nur den Nachtest (Ya).

Struktur

	Yb	X	Ya	EG
	Yb	~X	Ya	KG1
R				
		X	Y	KG2
		~X	Y	KG3

SPSS- Auswertung: Es gibt kein statistisches Verfahren, mit dem alle 4 Gruppen zugleich verglichen werden könnten. (1) alle Ya*s* per 1-faktorieller ANOVA mit 4 Stufen; (2) vollständiger Vergleich für EG und KG1; (3) die Treatmentwirkung läßt sich gleich auf 4 unterschiedliche Weisen testen: (1) EG1/Ya *vs* EG1/Yb; (2) EG1/Ya *vs* KG1/Ya; (3) EG2/Ya *vs* KG2/Ya; (4) EG2/Ya *vs* KG1/Yb. Sofern alle diese Vergleiche zu gleichsinnigen Ergebnissen führen, ist die Wirkung des Treatments sehr „umfassend" gezeigt worden.

Sonstige Anmerkungen: Mit der EG und KG1 ist der Pretest-Posttest-Kontrollgruppen-Plan realisiert—und damit die Kontrolle der Gefährdungen der internen Validität. Der wesentliche Unterschied zu diesem Versuchsplan besteht in der *zusätzlichen* Untersuchung einer Experimental- und einer Kontrollgruppe (EG2 und KG2) ohne Vortest Yb.

6.6.5 Allgemeine Veränderung: quasi-experimentelle Versuchspläne

VPL—#10 Prätest-Posttest Eingruppenplan

Fragestellung & Vorgehen: Bei einer Untersuchungsgruppe wird/werden die AV(s) vor (Yb) und nach (Ya) einem Treatment (X) erhoben. Änderungen in der/den AV(s) werden auf das Treatment zurückgeführt.

Struktur Yb X Ya

Statistische Auswertung:

(a) univariat: t-Test für abhängige Stichproben
(b) multivariat: Hotelling´s T^2.

Sonstige Anmerkungen: Die Änderungen in der/den AV(s) können nicht zweifelsfrei auf das Treatment zurückgeführt werden, da auch „zwischenzeitliches Geschehen", „Reifung" usf. zu Veränderungen in der/den AV(s) geführt haben könnten. Der Vortest (Yb) könnte die Reaktion auf das Treatment (X) beeinflussen; damit wäre die Verallgemeinerbarkeit der Treatment-Wirkung auf diejenigen Treatment-Applikationen eingeschränkt, denen eine Vortestung Yb vorausging. Ebenso könnten Spezifika der Stichprobenauswahl für die gefundenen Treatmenteffekete mit-„verantwortlich" sein, was die Generalisierbarkeit auf andere Stichproben/ Populationen einschränkt.

VPL#11—Prätest-Posttest-Treatmentumkehr Eingruppenplan

Fragestellung & Vorgehen:

Struktur Yb X Ya Yd X Yc

SPSS- Auswertung:
(a) univariat: Varianzanalyse mit Meßwiederholung (4 Stufen)
(b) multivariat: 1-faktorielle MANOVA

Auswertung mit SPSS:

Sonstige Anmerkungen**:**

(VP#11) Prätest-Posttest Kontrollgruppenplan
Fragestellung & Vorgehen:

Struktur

Yb	X	Ya
Yb	~X	Ya

SPSS- Auswertung:
(a) univariat: Split plot-Varianzanalyse (1 within-, 1 between-Faktor)
(b) multivariat: 2-faktorielle MANOVA

6.6.6 Korrelative Versuchspläne

Allgemeine Struktur: Aufgrund der fehlenden Unterscheidung zwischen UV und AV haben korrelative Versuchspläne mit Meßwiederholung eine andere Struktur als (quasi-) experimentelle. Generell gilt, daß mindestens 2 Variablen zu mindestens 2 Meßzeitpunkten erhoben werden (schematische Beispiele in Tabelle).

Struktur ←*m* Meßzeitpunkte→

↑	$y_{1.1}$	$y_{1.2}$ (1)	.	$y_{1.m-1}$	$y_{1.m}$ (3)
	$y_{2.1}$	$y_{2.2}$	.	$y_{2.m-1}$	$y_{2.m}$
k Variablen	.	. (2)	.	.	.
↓	.	.	.	.	.
	.	.	.	.	.
	$y_{k.1}$	$y_{k.2}$	.	$y_{k.m-1}$	$y_{k.m}$

(1)=2 Variablen, 2 Meßzeitpunkte; (2)=3 Variablen, 4 Meßzeitpunkte.

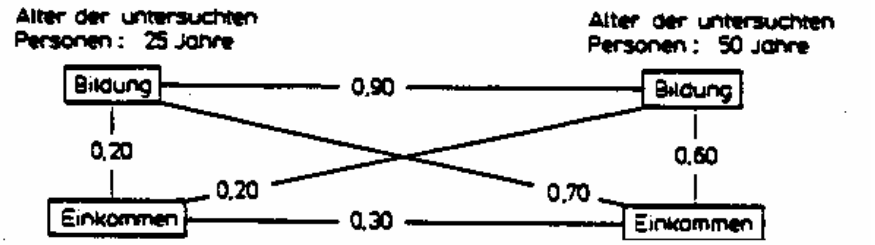
Verschiedene Anordnungen sind möglich.

(1) *Bivariate Korrelations- und Regressionspläne* mögen die Korrelation einer Variablen mit sich selbst zu 2 Meßzeitpunkten oder die Regression einer Variablen zu einem Zeitpunkt auf einen vorherigen Zeitpunkt untersuchen (Feld (1) in der Tabelle).

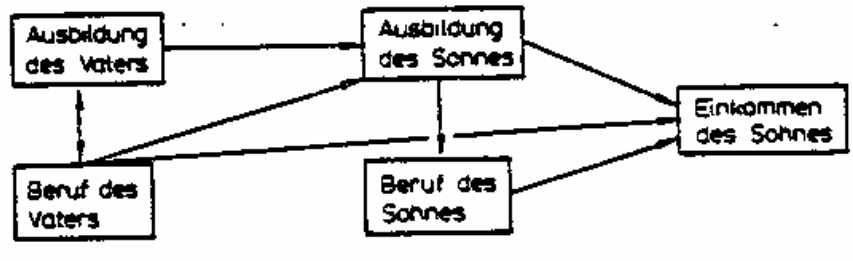
(3) In *„Cross-lagged panel"-Plänen* werden mindestens 2 Variablen zu mindestens 2 Meßzeitpunkten untersucht. Anhand der Korrelationsstruktur sind theoretisch erarbeitete Kausalmodelle empirisch prüfbar (Feld (2) in der Tabelle).

(4)

Ein Beispiel für die Korrelationsstruktur für 2 Variablen zu 2 Meßzeitpunkten:
Die relative Größe der Korrelationskoeffizienten macht die Hypothese plausibel, daß die Bildung das Einkommen beeinflußt und nicht das Einkommen mit 25 Jahren das mit 50 Jahren.



Beispiel eines pfadanalytisch prüfbaren Kausalmodells:



Erläuterung: Die (gerichteten) Pfeile markieren die vermutete Kausalstruktur. Die quantitative Ausprägung des kausalen Einflusses wird durch die sog. *Pfadkoeffizienten* ausgedrückt. Diese Koeffizienten lassen sich ähnlich wie die Partialkorrelationen berechnen. Aufgrund der berechneten Pfadkoeffizienten können die Korrelationskoeffizienten geschätzt werden. Dies ist an sich trivial, da die Pfadkoeffizienten aufgrund der Korrelationskoeffizienten berechnet werden. Allerdings werden pfadanalytische Modelle in der Regel so formuliert, daß zur Berechnung der Pfadkoeffizienten nicht alle Korrelationen benötigt werden. Mit Hilfe der Pfadkoeffizienten können diese (nicht einbezogenen) Korrelationen dann geschätzt und mit den empirisch ermittelten verglichen werden. Ergeben sich signifikante Differenzen, so muß das Kausalmodell als falsifiziert betrachtet werden. (weitere Beispiele: Bortz, 1984, S. 398/399).

(3) *Faktoren- oder Hauptkomponentenanalysen*: Bei Hauptkomponenten- oder Faktorenanalysen mit Meßwiederholungsdaten wird eine Menge an Variablen zu mindestens 2 Meßzeitpunkten erhoben (siehe Cattell´schen Datenbox in Abschnitt* 2.2):

- viele Personen, 2 Meßzeitpunkte, mehr als 2 Variablen: R-Technik & Trait-Bestimmung
- viele Personen, mehr als 2 Meßzeitpunkte, Differenzwerte von mehr als 2 Variablen: differentielle R-Technik (dR-Technik) & Bestimmung von Veränderungs-Dimensionen.

6.7 Weitere Versuchspläne für Gruppenuntersuchungen

6.7.1 Versuchspläne mit Blockbildung

*Zielsetzung:*Es handelt sich hierbei um Versuchspläne mit parallelisierten ("matched") Gruppen. Diese Pläne werden häufig auch unter Zufalls- oder Meßwiederholungsversuchsplänen subsummiert. Ziel ist die Reduktion der Fehlervarianz durch Bilden homogener Blöcke von Versuchspersonen zwecks Konstanthaltung bestimmter Ausgangsbedingungen.
Vor der eigentlichen Untersuchung ist die Erhebung von zusätzlichen Daten erforderlich, die später in die Versuchsplanung eingebracht werden. Meist handelt es sich hierbei um mit der AV hoch korrelierte Organismusvariablen (Alter, Geschlecht, Intelligenz, Konzentrationsleistung etc.). Die Probanden werden anhand dieser Daten in statistisch homogene Blöcke eingeteilt. Die Ausgangsunterschiede zwischen verschiedenen Versuchsgruppen sollen hierdurch minimiert werden.
Ein Sonderfall der Blockversuchsplanung ist die sog. Jochkontrolle ("yoked control"), die eine rigorose Konstanthaltung der Ausgangsbedingungen beabsichtigt. Es wird ein "matching" der unmittelbaren experimentellen Umgebung vorgenommen, d.h. es werden z.B. Zwillingspaare festgelegt, die identische Information (Stimuli, Umgebungsmerkmale etc.) erhalten.
Vorteile: gute Kontrolle bzgl. einer mit der AV korrelierenden Vortestvariable; *Nachteile*: Falls keine geeigneten Vortestvariablen gefunden werden, kann die intendierte Reduktion der Fehlervarianz mißlingen; *Statistische Auswertung*: wie Wiederholungsmessungen, d.h. t-Test für abh. Stichproben oder VA für Meßwiederholung.

6.7.2 Mischversuchsplanung

Es handelt sich um zwei- oder mehrfaktorielle Designs, wobei die einzelnen Faktoren den Haupttypen der zuvor behandelten Designs (Zufallsgruppenfaktoren, Faktor mit wiederholten Messungen, Blockfaktor) entsprechen. Nach Neale und Liebert (1973) umfassen "mixed designs" Kombinationen eines experimentellen und eines korrelativen Faktors (Organismusvariable). Das Hauptziel einer gelungenen Versuchsplanung - die Umsetzung des Max Min Kon-Prinzips nach Kerlinger - wird durch Mischdesigns wesentlich erleichert.
Struktur: Eine Reihe von Kombinationen der Zufallsgruppen (R), Meßwiederholungs- (W) und Blockbildungsdesigns (B) sind denkbar.
Vorteile: Eine gute Sekundär- und Fehlervarianzkontrolle ist möglich; eine Verbindung von experimentellen und korrelativen Forschungsansätzen wird hierdurch erleichtert. *Nachteile*: geringere Sensitivität beim Nachweis von Wechselwirkungen von Zufallsgruppen und Organismusvariable; Probleme bei der Festlegung der Stufen des korrelativen Faktors (zu wenige oder nicht-repräsentative Stufen sind ungünstig). *Statistische Auswertung:* Überwiegend varianzanalytische Verfahren für multifaktorielle Designs.

6.8 Einzelfalluntersuchungen

Zwar ist bei Einzelfall-Untersuchungen die Analyseeinheit ein einzelner Proband, doch können zwecks Prüfung der Verallgemeinerbarkeit der Ergebnisse deren mehrere an Einzelfall-Untersuchungen teilnehmen.

Zahlreiche Fragestellungen können eine Einzelfall-Betrachtung anzeigen:

Mögliche Indikationen zur Einzelfallbetrachtung
Identifikation individueller Parameter in der Diagnostik
Individuelle Therapiekontrolle
Hypothesengenerierende Explorationsstudien
Untersuchung selten auftretender Phänomene
Vorhandensein heterogener Stichproben
Abhängigkeit der Einzelmessungen der Pbn
Unmöglichkeit der Bildung einer Zufallsstichprobe
Durchführung von Langzeitstudien

6.8.1 Planung von Einzelfallstudien

Aufgrund der zahlreichen Meßwiederholungen ist bei Einzelfallstudien besondere Sorgfalt bei der Wahl des Meßverfahrens geboten. Wenn eine Vp mehrfach untersucht wird, so sollte das Meßverfahren *möglichst wenig reaktiv* sein, d.h. eine nur geringe *seriale Abhängigkeit* der Messungen mit sich bringen (positive Beispiele: psychophysiologische Maße, Beobachtungsdaten, (Interviewdaten); negative Beispiele: Leistungstests, (Selbsteinschätzungen)).

6.8.2 Zeitreihen-Versuchspläne

(A) Versuchspläne.

Bei mehr als zwei Beobachtungen kann von einer Zeitreihe, d.h. einer zeitlich geordneten (meist äquidistanten) Menge von Beobachtungen gesprochen werden. Zur Auswertung solcher Zeitreihen abhängiger Meßwerte existieren mehrere Verfahren. Es wurden auch verschiedene Versuchspläne mit Bedingungsvariation entwickelt und speziell in der Einzelfall-Forschung eingesetzt. Typische Zeitreihen-(Einzelfall-) Versuchspläne zur Prüfung von Interventionen sind:

Einzelfall-Design	Behandlungsplan						
1) Case study	X	YYY					
2) AB-Design		YYY	X	YYY			
3) ABA-Design		YYY	X	YYY	X	YYY	
4) Multiple intervention Design		YYY	X_1	YYY	X_2	YYY	
5) ABAB (Operant)-Design		YYY	X	YYY	X	YYY	X YYY
6) Reversal Design		YYY	X_1	YYY	X_2	YYY	
		YYY	X_2	YYY	X_1	YYY	
7) Interaction Design		YYY	X_1	YYY	X_2	YYY	X_1X_2 YYY
8) Multiple baseline Design		YYY	X	YYY	X	YYY	~X YYY
		YYY	~X	YYY	X	YYY	X YYY
		YYY	~X	YYY	~X	YYY	X YYY

(B) Auswertung.
Die Auswertung von Zeitreihenversuchsplänen kann erfolgen durch nicht-parametrische Trendanalysen oder Übergangsmatrizen, d.h. zufallskritische Prüfung der Abfolge von Meßwerten bzw. Kategorien:

- Autokorrelation: Eine Zeitreihe wird in immer größer werdenden Abständen (lags) gegen sich selbst verschoben korreliert; das erhaltene Autokorrelogramm läßt die sequentielle Abhängigkeit der Meßwerte erkennen, außerdem Trends und Periodizitäten (z.B. 24 Stunden, 7 Tage).
- Kreuzkorrelation: Direkte (zeitsynchrone) oder verschobene Korrelation mit einer zweiten Zeitreihe.
- Autoregressive, integrierte moving-average (Gleitmittelwert) - Modelle ARIMA: Zeitreihenanalysen fassen die Merkmalsausprägungen zu verschiedenen Beobachtungszeitpunkten als Ausdruck eines prozessualen Verlaufs auf, der in unterschiedliche Komponenten zerlegbar ist. Holtzman (1977) unterscheidet 3 Komponenten einer Meßsequenz:
    1. Trend / Langzeitbewegung (Richtung entlang der Zeitachse)
    2. Oszillation um den Trend
    3. Fehler- oder Restkomponente (Schwankungen um eine individuum-spezifische Merkmalsausprägung)
- Verschiedene Zeitreihenmodelle werden unterschieden
    - Es können deterministische und stochastische Zeitreihenmodelle unterschieden werden; nur letztere dürften für psychologische Fragestellungen in Betracht kommen.
    - In Zeitreihen abgebildete Prozesse können (a) stationär oder (b) nicht-stationär sein.

      Stationarität: Es treten höchstens Oszillationen um den Mittelwert auf, nicht aber Schwankungen des Mittelwerts oder des Trends.

      Nicht-Stationarität: Es treten Schwankungen des Mittelwerts oder des Trends auf.

(C) Methodische Probleme der Einzelfallbetrachtung

Methodische Probleme
1. Die von Box & Jenkins (1970) geschätzte Mindestzahl von 50 serialen Meßwerten ist vielfach nicht zu erreichen.
2. Die Anordnung der Meßzeitpunkte muß eine sinnvolle Abbildung des psychologischen Prozesses gestatten.
3. Die Verallgemeinerbarkeit der Ergebnisse ist unklar; n-fache Replikationen der Zeitreihen scheinen erforderlich.
4. Die interne Validität muß gewährleistet sein, ist aber oft schwer zu erreichen (z.B. Kontrolle zwischenzeitlichen Geschehens).
5. Gefundene Effekte sollten zufallskritisch abgesichert werden.

6.8.3 Faktoren- und Hauptkomponentenanalyse

Ein Speziallfall von Zeitreihen-Versuchsplänen sind Faktoren- und Hauptkomponentenanalysen, die für die Ausprägungen in einer Zahl an Merkmalen einer Person zu mehreren Meßzeitpunkten gerechnet werden. Die sogenannten O- und P-Techniken der Faktorenanalyse nach Cattell (1946) können zur Analyse von Veränderungen bei Einzelpersonen herangezogen werden.

- *O-Technik*: Über die n Zeitpunkte werden Korrelationen zwischen den Variablenpaaren gerechnet und faktorisiert. Das Ergebnis ist die allen Zeitpunkten gemeinsame Merkmalsfluktuation, die z.B. im Sinne wiederkehrender Entwicklungsphasen interpretiert werden kann. Während die O-Technik bislang kaum angewandt wurde, wird die P-Technik in der Persönlichkeitsforschung zur Analyse dynamischer Prozesse eingesetzt.
- *P-Technik*: Über die m Variablen werden Korrelationen zwischen den n Zeitpunkten gerechnet und faktorisiert. Das Ergebnis ist die allen Variablen gemeinsame zeitliche Fluktuation, die z.B. im Sinne eines durch soziale Umstände bedingten allgemeinen Trends interpretiert werden kann. Die P-Technik vermag nicht die Richtung der Wirkung in einer Meßsequenz offenzulegen (die Faktorenstruktur ist unabhängig von der Reihenfolge, in der die Meßwerte einer Zeitreihe in die Faktorenanalyse eingehen).

Übersicht über Versuchspläne und Strategien der Datenanalyse

Designtyp	Design	Hypothese	Typische stat. Verfahren	Typische Anwendungsfelder
Experiment Sozial-	Zufallspläne Blockpläne Meßwiederholungpläne (balanciert und unbalanciert) Hierarchische Pläne Quadratische Pläne Kombinationen der genannten Pläne (alle Pläne können sowohl ein- als auch multifaktoriell bzw. uni- und multivariat realisiert werden; mit und ohne Kovariate. Dies gilt teilweise auch für die quasi-experimentellen Pläne)	Unterschieds- und/oder Veränderungshypothesen	AN(K)OVA MAN(K)OVA (non-parametrische Verfahren)	Allgemeine und psychologie t-test
Quasi-Experiment außerdem und	Differentielle- und äquivalenten Kontroll-Gruppen (meist mit Meßwiederholung) Evaluationsforschung Zeitreihenversuchspläne Einzelfallversuchspläne (allg. Klein-N-Forschung) (entwicklungspsychologische Sequenzmodelle)	Pläne mit nicht oder Veränderungshypothesen	Unterschieds- und/ GLM ARIMA ABO-Fragestellungen	wie oben, Entwicklungspsychologie Psychotherapie- Pädagogische Psychologie
Korrelationsstudie Korrelationsanalysen	Design wicklungspsychologie nicht weiter festgelegt	Cross-lagged panel Differentielle und Enthypothesen	Zusammenhangs- Kanonische Korrelation Mutiple Regression Cluster- und Faktorenanalysen Pfadanalysen, LISREL	Epidemiologie Klinisch-psychologische Fragestellungen