EEM Einführung in die Experimental- & Evaluationsmethodik der Wirtschafts- & Sozialpsychologie

Skript zur einstündigen Vorlesung im SS 19

Micha Strack (mstrack@uni-goettingen.de)

Gliederung

Vorbemerkungen: Die Lernziele der Vorlesung sind auf den Vorwissensstand von sich meist im zweiten Semester befindenden Bachelor-Studierenden abgestimmt, das Skript soll jedoch auch später bspw. bei der Planung einer empirischen Qualifikationsarbeit zum Nachschlagen dienen können. Vielleicht leistet die Vermittlung von Grundkenntnissen hier nicht mehr als ein Vorstellen methodischer Fachbegriffe („Fach-Chinesisch“, Fachbegriffe werden durch Kursivschrift gekennzeichnet). Da Verstehen leichter ist als Produzieren und die Vorlesung auf eine SWS beschränkt ist, müssen sich Vorlesung und Skript auf das „Wortverständnis“ konzentrieren. Die Vermittlung von Fertigkeiten in statistischer Auswertungsmethodik wird nicht geleistet (auch Kap. 4 nennt nur Namen). Erreicht werden sollen: Kompetenzen im kritischen Lesen empirischer Untersuchungen aus der Wirtschafts- und Sozialpsychologie, die Entwicklung der Fähigkeit aus *Theorien* empirische *Hypothesen* abzuleiten, die in *Hypothesen* genannten *Variablen* zu identifizieren, sich alternative *Operationalisierungen* für diese *Variablen* zu überlegen und solche sowie ganze Untersuchungen auf ihre *Güte* hin zu bewerten.

Fragen und Verbesserungsvorschläge bitte an: mstrack@uni-goettingen.de

1. Einführendes

1.1 Empirische Forschung - warum?

Die Psychologie ist eine *empirisch arbeitende* Wissenschaft.
Da die meisten Personen an anderen und sich selbst sehr interessiert sind, betätigen sich viele im Alltag als *Laienpsychologe/in*: man findet *Erklärungen* für eigenes Verhalten oder das von anderen relativ leicht. Es gibt also viele Erklärungen, allerdings auch einander widersprechende, zumindest unterschiedlich komplizierte. Welche ist richtig? Welche sind falsch? Forschung in der Psychologie versucht, falsche von eventuell richtigen Erklärungen zu unterscheiden. Empirische Forschung tut dies, indem sie die Erklärungen *mit der Empirie konfrontiert*, in der Empirie überprüft (das klingt wie ´in der Natur´ prüft - oft tut sie es *im Labor*. Warum dies so ist, soll u.a. in dieser Vorlesung klar werden). Oder sollte man besser sagen: *an* der Empirie prüft? Falsche Erklärungen halten dieser Prüfung nämlich nicht stand.
Empirische Wissenschaften (egal ob Naturwissenschaften oder Sozialwissenschaften) versuchen also hauptsächlich, Aussagen als falsch zu kennzeichnen. Die nicht-falschen bleiben als *bisher bewährtes* (aber weiterhin immer vorläufiges) Wissen übrig.

Wissenschaftliche Aussagen lassen sich von unwissenschaftlichen in der Praxis aber nicht leicht unterscheiden. Das Ausmaß der ´Wissenschaftlichkeit´ einer Aussage bemisst sich daran, ob sie *systematisch* mit den aktuell anerkannten Methoden der Fachdisziplin (=*lege artis*) geprüft worden ist.

Während sich die *Wissenschaftlichkeit* also über die Methode definiert, definieren sich die Fächer über ihren Gegenstand (also Inhalt).

Die Definition der *Psychologie* gelingt am günstigen über ihre Aufgabe:

Ψ: Aufgabe der Psychologie ist

das Beschreiben, Erklären und Vorhersagen (sowie das Bewerten von Veränderungstreatments)

des Erlebens & Verhaltens von Individuen.

Für das methodische Vorgehen in Wissenschaften ist die Einhaltung der Reihenfolge von *Beschreibung, Erklärung und Vorhersage* wichtig:
Beschreiben kann schon Beobachtung erfordern, um Antwort auf die Frage zu gewinnen „was ist das eigentlich für ein Phänomen?“ Beschreiben nutz in der wissenschaftlichen Psychologie nur wirklich, wenn es auch gelingt, die Phänomene zu *messen*. Das *Messen* der Ausprägung von *Variablen* wird hier ein großes Thema (s. Kap. 3.2).
Erklärungen haben die Form von *Theorien*. Theorien behaupten eine (*Kausal*-) Beziehung zwischen Konstrukten (und antworten auf diese Weise auf Warum-Fragen). *Gute Theorien sind das Ziel der wissenschaftlichen Bemühung*!
Aus Theorien werden *Vorhersagen* abgeleitet: einerseits (und das sind die wichtigsten Vorhersagen), um Maßnahmen zur Veränderung von Zuständen abzuleiten (sogn. *Treatments,* Interventionen), andererseits (und dies vorher und ´back office´), um die Theorie über ihre Vorhersagen *prüfen* zu können. Die Forschungsmethodik ist somit das Mittel, um zu dem Ziel *geprüfter, guter Theorien* zu gelangen, mit denen man auf wissenschaftlicher Basis etwas verändern kann.
Theorien sind schlecht, wenn sie falsch sind (s. Kap. 5). Um *eine Theorie zu prüfen*, wird also eine Vorhersage abgeleitet und diese sehr präzise als *Hypothese* formuliert (Übung in Kap. 3.1). Die Hypothese wird in einer Untersuchung geprüft. Trifft die Vorhersage ein (= *hypothesenkonformes* Ergebnis), hat sich die Theorie *bewährt*. Wenn sich die aus einer Theorie abgeleiteten Vorhersagen im empirischen Test nicht bewähren (*diskonforme* Ergebnisse), wird entweder die *Güte der Untersuchung* kritisiert (und in Folgeuntersuchungen verbessert) oder die Theorie wird weiterentwickelt oder sogar zugunsten einer besseren verworfen. Wissenschaftlicher Fortschritt bedeutet seit Popper (wikipedia: *Falsifikationismus*), falsche Theorien als falsch zu entlarven.

1.2. Empirische Forschung - wie?

Der forschungslogische Ablauf, also die (ideal-)typische Sequenz von Arbeitsschritten in der Forschung, lässt sich in drei Hauptphasen unterteilen: in den *Entdeckungszusammenhang*, den *Begründungszusammenhang* und den *Verwertungszusammenhang* einer Aussage.

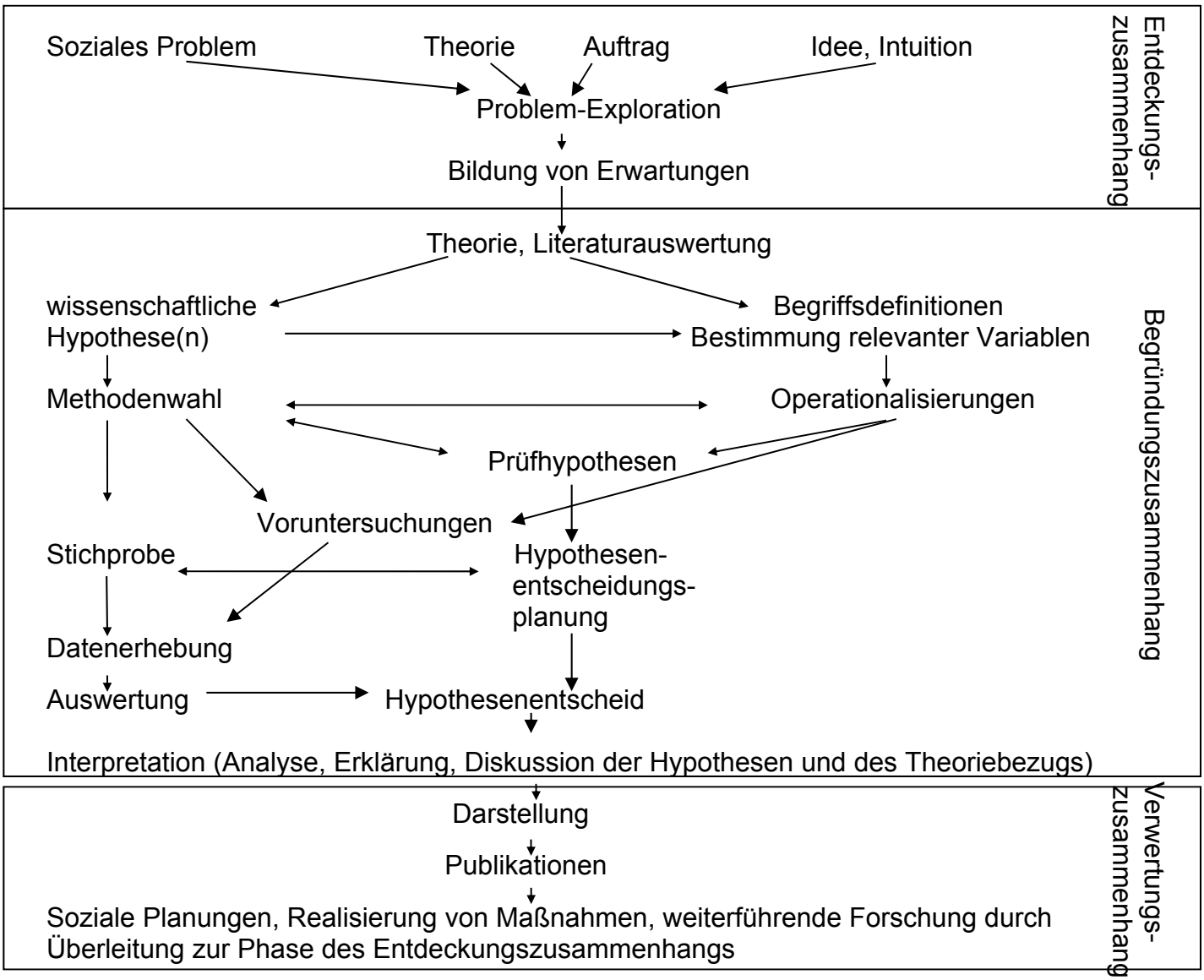


Abb.1_1: Forschungslogischer Ablauf

Der Entdeckungszusammenhang einer Aussage (/einer Theorie), also ob sie wegen eines wichtigen Problems, wegen einer zufälligen Idee oder Intuition, wegen dem Auftrag von anderen zum Gegenstand der Forschung wird (Abb. 1_1 oben), soll bei der Beurteilung der Aussage (oder Theorie) keine Rolle spielen - hier wird die naturwissenschaftliche Ausrichtung der Psychologie deutlich - aber auch schon die Vorsicht, eine Aussage nicht vorschnell nur deshalb als wahr anzusehen, weil sie von einer wichtigen Autorität formuliert wurde oder weil sie bei der aktuellen Problemlage gerade sehr nützlich erscheint. Die *Güte einer theoretischen Aussage* soll vielmehr über ihren *Begründungszusammenhang* bestimmt werden (Abb. 1_1 Mitte): wurde sie mit wissenschaftlichen Methoden *lege artis* überprüft oder nicht? Das Ausmaß der Wissenschaftlichkeit einer Aussage bemisst sich also danach, wie sehr sie *systematisch* und nach dem im Fach jeweils anerkannten Kanon der Methodik *geprüft* wurde (s. Kap. 1.1).

2. Konstrukt & Theorie

Da die Psychologie das Erleben und Verhalten von Personen *erklären* will (s. Kap.1.1), entwickelt sie psychologische *Konstrukte* und bindet sie in *Theorien* ein. *Psychologische Konstrukte* sind dem Wort nach Konstruktionen; die Psychologie behauptet, dass sie existieren und für Erklärungen nützlich sind. Typische Konstrukte (aus den Vorlesungen Sozialpsychologie): Entitäten wie: Selbstkonzept, Selbstwert, Einstellung, Emotion, Motiv, Anspruchsniveau, Beziehung, ... und behauptete Zustände wie: kognitive Balance, Dissonanz, Reaktanz, Motivation, Selbstaufmerksamkeit, und behauptete Prozesse wie Sozialer Vergleich, Impression Management, Kommunikation, Attribution, Dissonanzreduktion, Lernen, Empathie, Compliance, normativer Einfluss, ...

Diese Konstrukte werden behauptet – aber vielleicht gibt es sie gar nicht? Ihre Existenz anzunehmen ist so lange sinnvoll, solange sie in Erklärungen (des Erlebens und Verhaltens von Menschen) eine wichtige Rolle spielen, solange sie als *erklärende Variablen* gebraucht werden.

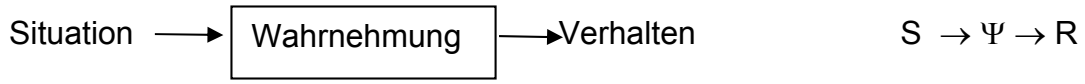


Abb. 2_1: Ein psychologisches Konstrukt als eine konstruierte Ψ Entität.
Eine Theorie *erklärt* ein Phänomen (gibt eine Weil-Antwort auf eine Warum-Frage),
eine Theorie verknüpft Konstrukte durch Kausalaussagen.

In Abb. 2_1 wird dargestellt, wie ein psychologisches Konstrukt (bspw. Wahrnehmung) für die *Erklärung* eines (nachweisbaren) Zusammenhangs zweier objektiv beobachtbarer Variablen (abstrakt: Stimulus → Response) verwendet wird: Jemand stellte fest, dass die Veränderung der Situation das Verhalten der Personen verändert (beobachtbar sind der *Stimulus* S und die *response* R). Die Psychologie behauptet nun, dass Situationen oder Stimuli Verhalten nur deshalb verändern können, *weil* sie von den Personen wahrgenommen werden (Abb. 2_1 links): ohne Wahrnehmung keine Verhaltensveränderung durch Situationsveränderung. In dieser Weise werden psychologische Konstrukte Ψ (erfunden und) in Theorien zur Erklärung von z.B. S → R Phänomenen eingesetzt. So wird z.B. das Konstrukt *Selbstkonzept* eingesetzt, um zu erklären, warum Personen sich nach einem Kompetenztraining, das dem Selbstkonzept Wissen über eigene Fertigkeiten zufügt („ich kann Statistik – ich bin eine Statistikerin!"), mehr zutrauen (-> empirische Abschlussarbeit, Bewerbung in der Marktforschung, etc), während ein anderes Kompetenztraining die Bewerbungsaktivitäten nicht erhöht, weil sich das Selbstkonzept der TeilnehmerInnen nicht verändert hat. Ein zur *Erklärung* verwendetes Konstrukt *´mediiert´* = vermittelt den S-R-Zusammenhang (s. auch Abb. 3_10 und Kap.4 zur *statistischen Mediation*).
Die Theorie behauptet hier also, dass das Training das Selbstkonzept verbessert und das Selbstkonzept die Motivation (z.B. zur Bewerbung) bestimmt.
Theorien verbinden Konstrukte durch Kausalaussagen (= in denen ein „weil" steckt).

Als didaktisches Beispiel einer Theorie wird hier und für die folgenden Kapitel die *Theorie der Sozialen Erleichterung* gewählt (*Theory of Social Facilitation*, Zajonc 1965, z.n. Herkner 1991:474ff o. Aronson et al. 2014, o. https://de.wikipedia.org/wiki/Social_Facilitation). Sie wurde entwickelt, nachdem beobachtet wurde, das Personen *in Anwesenheit anderer* besser arbeiten (zu diesem Thema gab es das erste *Experiment* in der Sozialpsychologie, 1889 von Tripplett, mit der *UV*: allein versus in Anwesenheit anderer). Es folgten Beobachtungen, dass manchmal in Anwesenheit anderer schlechter gearbeitet wurde. Zajonc hat geklärt, dass es auf die Aufgabe ankommt, denn: Anwesenheit anderer erhöht das Erregungsniveau, höheres Erregungsniveau hilft bei leichten, gut geübten, aber hemmt bei neuen, schweren Aufgaben (Abb. 2_2).

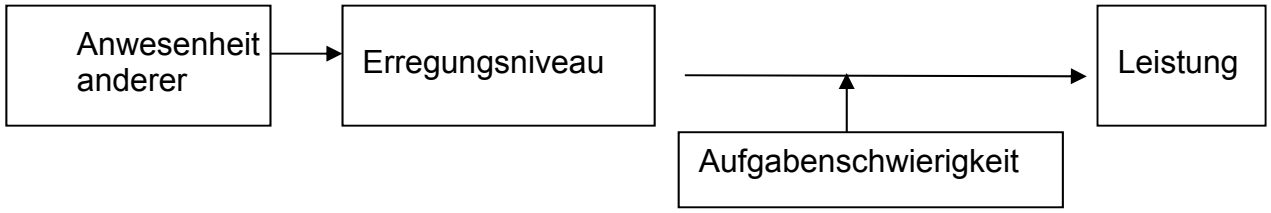


Abb. 2_2: Theorie der Sozialen Erleichterung (als didaktisches Beispiel für das Skript gewählt)

3. Methodische Grundlagen

Theorien werden geprüft, indem aus ihnen Vorhersagen abgeleitet, als *Hypothesen* formuliert und diese Hypothesen empirisch *getestet* werden.

3.1 Hypothesen und Design

3.1.1 Hypothesen

Eine *Hypothese* ist eine *aus einer Theorie abgeleitete* Vorhersage, die sich als falsch herausstellen könnte. Ein trivial-wahrer, tautologischer Satz („Kräht der Hahn auf dem Mist, ändert sich´s Wetter oder es bleibt, wie es ist", Tautologie = immer wahr) ist keine Hypothese. Aus einer Theorie (wie bspw. der in Abb. 2_2) lassen sich verschiedene Hypothesen ableiten. Tabelle 3_1 enthält vier einfache Hypothesen, die sich alle nur auf denselben, einen kleinen Ausschnitt der Theorie der Sozialen Erleichterung (Abb. 2_2) beziehen.

Tab. 3_1: Hypothesen, aus der Theorie Sozialer Erleichterung (Abb. 2_2) abgeleitet.

Hypothesen	ungerichtete, zweiseitige H_1	gerichtete, einseitige H_1
Zusammenhangshypothese	Es zeigt sich ein Zusammenhang zwischen Erregung und Leistung in einer schwierigen Aufgabe	Je höher die Erregung, desto schlechter die Leistung in einer schwierigen Aufgabe
Unterschiedshypothese	Es besteht ein Unterschied in der Leistung in einer schwierigen Aufgabe zwischen der Bedingung hoher und der Bedingungen niedriger Erregung.	Unter der Bedingung hoher Erregung fällt die Leistung in einer schwierigen Aufgabe schlechter aus als unter der Bedingung niedriger Erregung.

Welche dieser Formulierungen sollte man wählen? Hypothesenformulierungen lassen sich einerseits danach unterscheiden, ob sie einen *Zusammenhang* zwischen zwei *Variablen* (=veränderbaren Größen, s.u.) formulieren oder ob *Unterschiede* zwischen Gruppen oder Versuchsbedingungen behauptet werden (*Zusammenhangshypothese* vs. *Unterschieds-hypothese*, s. Tab. 3_1). Ob als endgültige Prüfhypothese eine Zusammenhangs- oder Unterschiedshypothese formuliert wird, ist erst zusammen mit der Planung des Versuchs*designs* (z.B. Abb. 3_3) zu entscheiden, daher enthält der Forschungslogische Ablauf (s. Abb. 1_1) vor der Festlegung der *Prüfhypothese so* viele parallele Überlegungen. Zu Beginn eines Forschungsprojekts sollte man zunächst imstande sein, die eigene Hypothese sowohl als Zusammenhangs- als auch als Unterschiedshypothese zu formulieren, damit man beim Design noch die freie Wahl hat, ob man Gruppen vergleichen oder beide Variablen kontinuierlich variieren lassen und eine Korrelation berechnen wird.
Zweitens, und wichtiger, werden *ungerichtete Hypothesen* von *gerichteten Hypothesen* unterschieden. *Ungerichtete Hypothesen* geben keine Richtung des Zusammenhangs oder Unterschieds an, sagen nicht, ob der Zusammenhang von Erregung und Leistung positiv sein soll (je höher die Erregung, desto besser die Leistung), oder ob der Zusammenhang negativ sein soll (je höher die Erregung, desto schlechter die Leistung). *Gerichtete Hypothesen* geben die Richtung des Zusammenhangs (das Vorzeichen!) an (für schwierige Aufgaben wird ein *negativer Zusammenhang* für Erregung und Leistung vorhergesagt), oder sie sagen, welche Versuchsgruppe die höhere Leistung haben soll (die mit der geringen Erregung, Tab. 3_1). Gerichtete Hypothesen sind daher *spezifischer*, sie behaupten ´mehr Wissen´ (das Wissen zur Richtung des Zusammenhangs/Unterschieds). Daher können gerichtete Hypothesen leichter falsch sein: wenn empirisch für schwierige Aufgaben mit höherer Erregung die Leistung steigt, kann die gerichtete Zusammenhangshypothese aus Tab. 3_1 nicht angenommenen werden, die ungerichtete jedoch schon, da dort ja nur irgendein Zusammenhang behauptet wurde. Das Eingehen des höheren ´Risikos´ durch eine gerichtete Hypothese wird in der Wissenschaft belohnt (s. dazu Kap. 4.2: weniger *N* nötig!), die gerichtete Hypothese hat einen höheren „*empirischen Gehalt*" (Popper). Wissenschaftliche Hypothesen sollten *gerichtet, einseitig* formuliert werden (dadurch wird die Theorie besser, da *strenger* geprüft). Je strenger eine Theorie geprüft wurde, desto besser ihr Begründungszusammenhang (s. Abb. 1_1).

Nullhypothese: Die aus der Theorie abgeleitete Hypothese soll entweder einen von Null verschiedenen Zusammenhang (möglichst bestimmter Richtung, also einseitig = gerichtet, Tab. 3_1) oder einen (möglichst einen bestimmten also gerichteten) Unterschied zwischen Gruppen oder Versuchsbedingungen formulieren. Zur Analyse der genauen Aussage der Hypothese wird im nächsten Schritt die zugehörige Nullhypothese (H0) formuliert. Zu jeder Hypothese gehört eine bestimmte *Nullhypothese* (daher heißt die aus der Theorie abgeleitete Hypothese auch Alternativ-Hypothese: die Alternative zur Nullhypothese, in Absetzung zur H0 auch H1 genannt).

Tab. 3_2: Alternativhypothese mit zugehöriger Nullhypothese

H1 = ´Alternativhypothese´ (s. Tab. 3_1)	zugehörige H0 = Nullhypothese
Es zeigt sich ein Zusammenhang zwischen Erregung und Leistung in einer schwierigen Aufgabe	Es zeigt sich kein Zusammenhang zwischen Erregung und Leistung in einer schwierigen Aufgabe
Je höher die Erregung, desto schlechter die Leistung in einer schwierigen Aufgabe (negativer Zusammenhang zwischen Erregung und Leistung)	Es zeigt sich kein oder ein positiver Zusammenhang zwischen Erregung und Leistung in einer schwierigen Aufgabe, entweder wird die Leistung bei höherer Erregung genauso sein wie bei niedriger, oder mit höherer Erregung wird die Leistung sogar besser.
Es besteht ein Unterschied in der Leistung in einer schwierigen Aufgabe zwischen der Gruppe hoher und der Gruppe niedriger Erregung.	Es besteht kein Unterschied in der Leistung in einer schwierigen Aufgabe zwischen der Gruppe hoher und der Gruppe niedriger Erregung.
In der Gruppe hoher Erregung fällt die Leistung in einer schwierigen Aufgabe schlechter aus als in der Gruppe niedriger Erregung	In der Gruppe hoher Erregung fällt die Leistung in einer schwierigen Aufgabe gleich oder besser aus als in der Gruppe niedriger Erregung

Nullhypothesen tragen ihren Namen ursprünglich, weil sie in ihrer gröbsten Form ´keinen Zusammenhang´ oder ´keinen Unterschied´ behaupten (ein ´Nicht-Wissen´, gegen das die aus der Theorie formulierte Alternative des ´Etwas-Wissens´, die Alternativhypothese, angehen will). Diese ursprüngliche Wortbedeutung gilt sprachlich eigentlich nur, wenn die wissenschaftliche Hypothese (Alternativhypothese, H1) ungerichtet war. Zur Analyse der genauen Aussage der formulierten Hypothese soll nämlich eine Nullhypothese so formuliert werden, dass sie alle Ergebniszustande enthält, die zur wissenschaftlichen Hypothese *diskonform* sind. Tab. 3_2 enthält die jeweilige Nullhypothese zu den Alternativhypothesen aus Tab. 3_1. Die Nullhypothesen zu den *gerichteten* Alternativhypothesen nennen somit zwei Zustände (bspw. ´keinen Zusammenhang oder der gegenteilige Zusammenhang´).
Ziel eines empirischen Forschungsprojekts wird damit, die Nullhypothese als falsch zu entlarven, damit die aus der Theorie abgeleitete Alternativhypothese *angenommen* werden darf (und die Theorie sich damit - zumindest in dieser Untersuchung - *bewährt* hat).

Solange die Nullhypothese nicht verworfen wurde, gilt sie. Zur Ergebnisbeschreibung (s.a. Kap. 4.2) gibt es eine strenge sprachliche Konvention; sie steht in Tab. 3_3.

Tab. 3_3: Sprachliche Konvention zum Hypothesenentscheid

Das Ergebnis	Man rede / schreibe über die H1:	Man rede / schreibe über die Ho:
spricht für die H1	*H1 darf/kann angenommen werden*.	Die *Ho darf/kann verworfen werden*.
passt mit H0	*H1 darf nicht angenommen werden*.	Die *Ho* ist *beizubehalten*.

Da man mit empirischer Forschung zwar falsche Aussagen entlarven kann, zum aktuellen Zeitpunkt angenommenes Wissen aber immer als vorläufig zu gelten hat (s. Kap. 1.1), darf in Forschungsarbeiten nie die Rede vom ´Beweisen´, ´bewahrheitet´, ´wahr´, ´Tatsache´ etc sein. Man schreibe stattdessen immer vorsichtig: ´die H1 kann angenommen werden´, ´die Theorie hat sich bewährt´.

3.1.2 *Variablen* isolieren, Prädiktor- und Kriteriumsvariablen

Nachdem Hypothese und Nullhypothese formuliert sind (Tab. 3_2), werden die in der Hypothese genannten *Variablen isoliert* (=aus dem sprachlichen Hypothesensatz identifiziert und herausgeschrieben, Tab.3_4), um sie anschließend *operationalisieren* zu können.

Tab. 3_4: Beispiele für Hypothesen und aus ihnen isolierten Variablen

H1	Variablen
Der Zusammenhang zwischen der Anwesenheit Anderer und der Erregung ist positiv	Anwesenheit anderer, Erregung
Je höher die Erregung, desto schlechter die Leistung in einer schwierigen Aufgabe.	Erregung, Leistung
In schwierigen Aufgaben nimmt die Leistung mit steigender Erregung ab, in einfachen Aufgaben zu.	Aufgabenschwierigkeit, Erregung, Leistung

Variablen: Variablen heißen so, weil sie in mehreren *Ausprägunge*n *variieren* müssen (= sie mehrere Zustände annehmen müssen), damit die Hypothese prüfbar ist. Die in der mittleren Hypothese in Tab. 3_4 genannte Bedingung der ´schwierigen Aufgabe´ ist in dieser Hypothese keine Variable, weil die Aufgabenschwierigkeit bei Prüfung dieser Hypothese immer gleich bleiben soll (nämlich hoch, also schwierig). Nur zur Prüfung der letzten Hypothese in Tab. 3_4 muss sie auch variieren (oder variiert werden).

Variablenausprägung vs. *Variable*: ´Geschlecht´ ist eine *Variable* mit zwei *Ausprägungen:* ´männlich´ und ´weiblich´. Synonym heissen die Ausprägungen auch *Stufen der Variable*. Sollen zwei bestimmte Gruppen verglichen werden (z.B. die Abteilungen ´Vertrieb ´ versus ´Controlling ´), sollte man einen übergeordneten Namen für diejenige eine *Variable* finden, die zwischen den beiden *variiert* (die Variable könnte ´Kundenorientierungsanforderung´ heißen, falls man aus diesem Grund beide Abteilungen auswählte (hoch für Vertrieb, niedriger fürs Controlling) oder einfach ´Abteilung´, falls man keinen besonderen Grund hatte, genau diese beiden auszuwählen; man sollte sich aber bemühen, einen Namen für die Variable zu finden, der den Grund für die Wahl der Variablenausprägungen, das dahinterliegende psychologische *Konstrukt* bezeichnet (also besser ´Kundenorientierungsanforderung´ als nur ´Abteilung´).

Unterscheidungen von Variablentypen, X & Y: Nachdem die Variablen aus der Hypothese isoliert wurden (Tab. 3_4), folgt die Phase der Versuchsplanung (komplizierte Mitte in Abb. 1_1). Erster Schritt davon ist, die isolierten Variablen in *X- und Y-Variablen* einzuteilen: X sind per Konvention die laut Theorie vermuteten Ursachen, Y die Wirkungen. In der ersten Hypothesen aus Tab. 3_4 ist Erregung die Y-Variable, in den beiden folgenden Hypothesen fungiert Erregung als X-Variable. Synonym werden die X-Variablen *Prädiktoren* (die Vorhersagenden) und die Y-Variablen *Kriterien* (die Resultierenden, meist: *AV*) genannt (Tab. 3_5).

Tabelle 3_5:
Variablen- Benennungen (die unterstrichenen Benennungen machen die wenigsten Voraussetzungen und sind die häufigsten Benennungen)

Ursache	→	Wirkung
X- Variable	→	Y- Variable
<u>Prädiktor</u>- Variable	→	Kriteriums- Variable
UV (unabhängige Variable) (independed var. = IV)	→	AV (<u>abhängige Variable</u>) (depended var. = DV)
Faktor	→	Variate

Die in Tab. 3_5 untereinander stehenden Benennungsvarianten sind nicht perfekt synonym: Kriteriumsvariablen können immer *abhängige Variablen* genannt werden (AV, im statistischen Slang sehr häufig verwendet); Prädiktorvariablen werden jedoch (strenggenommen) nur dann auch *unabhängige Variablen* genannt (UV, eine besondere Markenbezeichnung ☺), wenn es gelungen ist, sie *experimentell zu manipulieren* (s. *Experiment*, Kap. 3.1.3, 3.1.4). *Faktoren* werden die Prädiktor-Variablen immer dann genannt, wenn sie auf *Nominalskalenniveau* operationalisiert wurden (s. *Skalenniveau*, Kap. 3.2), was für experimentelle UV fast immer gilt, aber auch für quasiexperimentelle Prädiktoren (s. *Quasiexperimen*t, Kap. 3.1.4). (Wer über diese Zusatzbedingungen unsicher ist, rede von Prädiktor- und Kriteriumsvariablen, oder von Prädiktor und AV, das geht immer).

Wenn die *Theorie* keine (kausale) Wirkrichtung X→Y zwischen zwei Variablen anzunehmen erlaubt, oder wenn bspw. zwei verschiedene Variablen für ein Konstrukt gemessen werden (z.B. Fluktuationsrate und Krankenstand als Indizes der Arbeits[un]zufriedenheit in Unternehmen) und eine Hypothese einfach deren Zusammenhang behauptet, spricht man vorsichtiger nur von *Covariaten* oder von zwei *AV*.

3.1.3 Spektrum der Untersuchungstypen & Dilemma zwischen interner Validität und gleichzeitig ökologischer Validität

Nachdem Hypothesen formuliert und Variablen isoliert sind, ist die *Prädiktorvariable zu operationalisieren*: gelingt es, sie *experimentell zu manipulieren*? Mit der Operationalisierung der Prädiktorvariablen wird gleichzeitig auch über den *Untersuchungstyp* (*Experiment oder Korrelationsstudie inkl. Quasiexperiment*, s. Kap. 3.1.4) entschieden. Untersuchungstyp und Anzahl von Prädiktor- und Kriteriums-Variablen legen dann das *Design* der Untersuchung fest (z.B. *zweifaktoriell-univariat*, s.u.).

Tab. 3_6: Schritte der Versuchsplanung

Versuchsplanung =Erst Prädiktorvariable(n) *operationalisieren*: gelingt es, die Prädiktorvariablen zu manipulieren? *Untersuchungstyp* festlegen & *Design* aufstellen dabei *Gütekriterien der Untersuchung* (besonders die *interne Validität der Untersuchung, evtl. auch die ökologische Validität*) optimieren. Dann AV operationalisieren, dabei *Gütekriterien der Messung* optimieren.

Entscheidungen während der Versuchsplanung können selten isoliert getroffen werden, oft werden alternative Versuchsdesigns und Operationalisierungen in iterativem Vorgehen gegeneinander abgewogen (s. Abb. 1_1 Mitte), um die Entscheidung über die Hypothesen mit einer möglichst hohen *Güte der Untersuchung* sicherzustellen. Da hierzu vieles gleichzeitig zu bedenken ist, wird im Folgenden zunächst ein sehr grober ‚intuitiver' Überblick über verschiedene Methoden angestellt, den ein Spektrum des sozial- und verhaltenswissenschaftlichen Methodeninventars (Abb. 3_1) anregen soll.

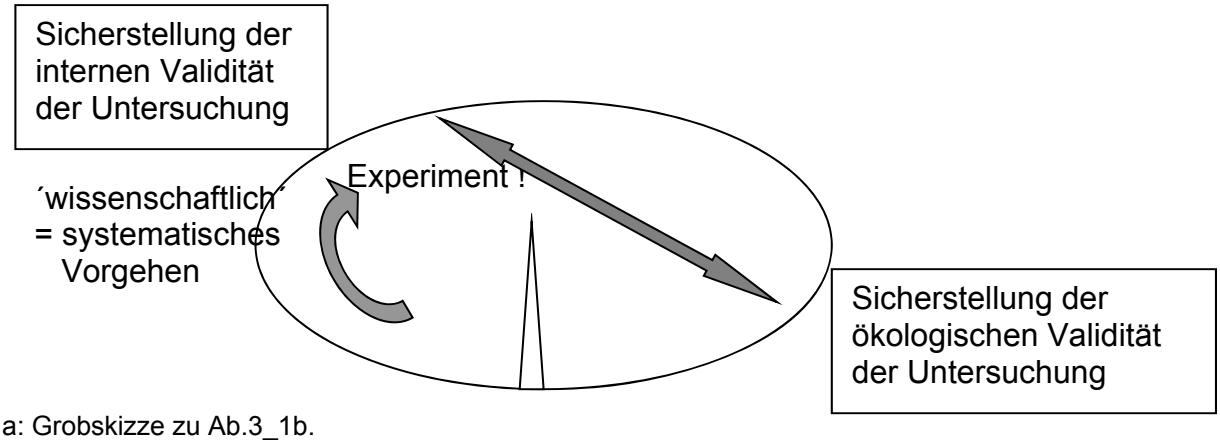


Abb. 3_1a: Grobskizze zu Ab.3_1b.

Abb. 3_1a zeigt die Leserichtung für Abb. 3_1b: Mit der unsystematischen Gelegenheits-beobachtung (Typ A) beginnt die noch *vorwissenschaftliche* Methode, einer Aussage einen *Begründungszusammenhang* zu geben („Erregung verringert wohl die Leistung bei schwierigen Aufgaben, denn bei der letzten Mathe-Klausur war ich wohl schlecht, weil ich zu aufgeregt war"). Über die aufsteigenden Typen B und C der Abb. 3_1b wird der Begründungs-zusammenhang *systematischer* (und daher wissenschaftlich, s. Kap.1.1). Eine systematische Auswertung von zuvor schon vorhandener *Verhaltensspuren*, z.B. Abnutzung des Teppichs in der Nähe der Zimmertür der Mathematikprüfung versus der für ein leichteres Fach, die Auswertung von Internet-Klick-Daten (zur Zeit unter dem Stichwort „Big Data" aktuell, de.wikipedia.org/wiki/Big_Data), von Fluktuations- und Absentismus-Zahlen aus der Betriebsstatistik etc. gehören zu Typ B, weil diese Daten vor der Bildung des Untersuchungsziels schon da waren. Nachteil: man muss mit dem auskommen, was da ist („Reanalyse"). Die verwendeten Methoden in Typ C und höher von Abb. 3_1b geben der Untersuchung eine höhere *interne Validität*, weil die Situationen gezielt zur Entscheidung über Hypothesen *erzeugt* wurden, also geplant genau das erfasst werden kann, was für die Hypothesenprüfung nötig ist und ggf. Störvariablen kontrolliert oder parallelisiert werden können. .

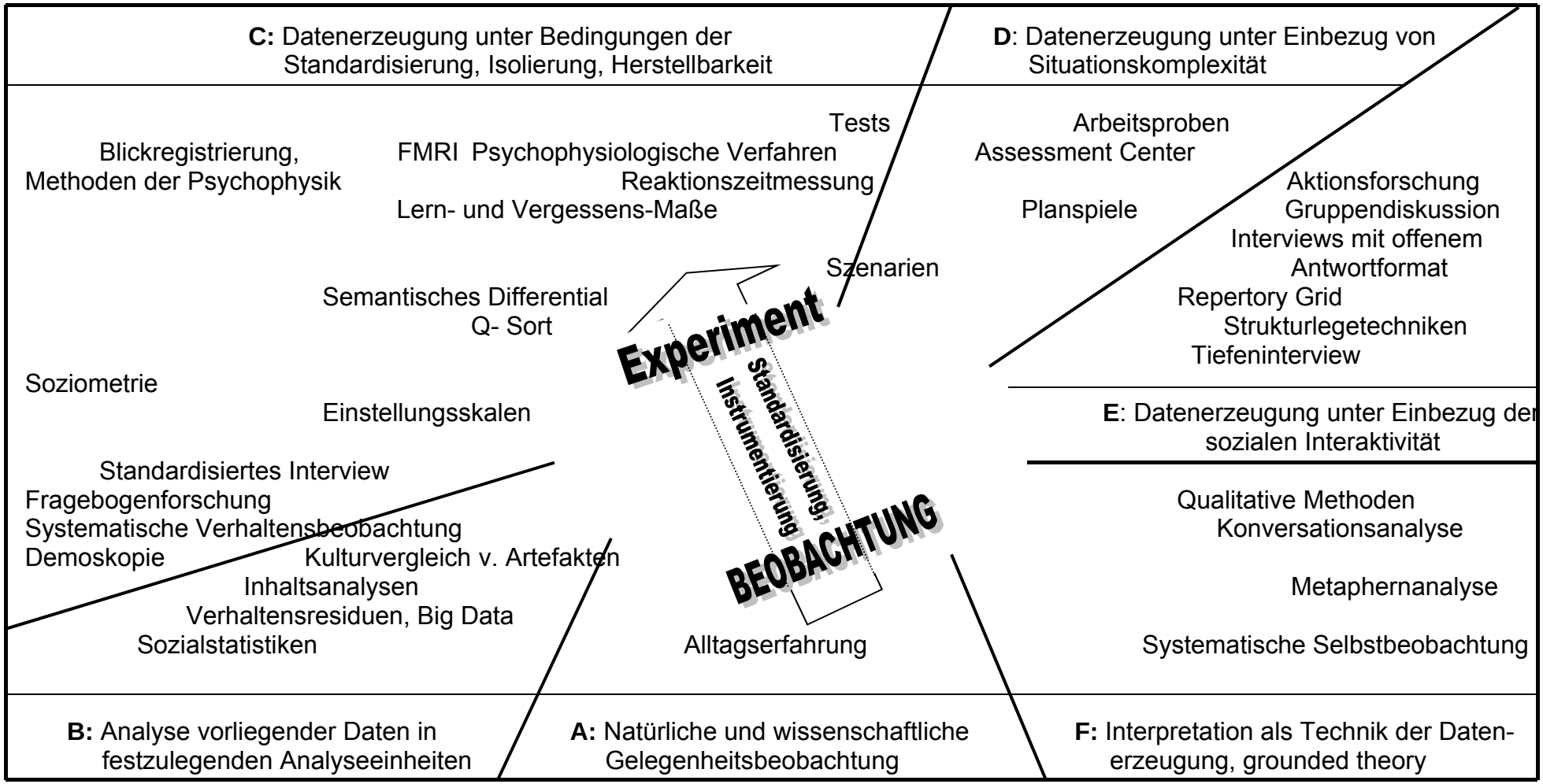


Abb. 3_1a: Spektrum des sozial- und verhaltenswissenschaftlichen Methodeninventars (nach Fassheber et al. 2001, S.7).

Dilemma zwischen interner Validität und gleichzeitig ökologischer Validität

Die ***interne Validität der Untersuchung*** ist hoch, „wenn die Variation der AV eindeutig auf die Variation der UV zurückzuführen ist“, also wenn man sicher sein kann (keine Zweifel aufkommen), dass die Kriteriumsvariable tatsächlich die Wirkung und die Prädiktorvariablen tatsächlich die Ursache im erhaltenen Befund gewesen ist (Abb. 3_2 links). Die *interne Validität einer Untersuchung* ist niedrig (oder *´ist gefährdet´*), wenn der Befund des empirischen Zusammenhangs zweier Variablen, die laut Theorie in einer Kausalitätsrichtung verbunden sein sollen, auch durch eine umgekehrte Kausalität verursacht sein könnte (Abb. 3_2 Mitte; z.B. die Erregung wegen schlechter Leistung stieg), oder wenn die AV gar durch eine Drittvariable verursacht wurde (z.B. Erregung und schlechte Leistung durch geringe Lernzeit; Abb. 3_2 rechts). Auf solche Drittvariablen (*konfundierende Variablen*) geht Kap. 3.1.5 ausführlich ein.

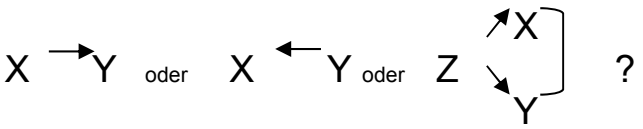


Abb. 3_2: Gibt es Zweifel über die für die Ergebnisse verantwortliche Kausalitätsrichtung, dann war die interne Validität der Untersuchung niedrig.

Das *psychologische Experiment* ist der Untersuchungstyp mit der höchsten *internen Validität* (s. *Experiment*, Kap. 3.1.4). Kausalaussagen sollten daher, wenn möglich, *experimentell* geprüft werden. Da Experimente durch die ***Manipulation der Unabhängigen Variablen*** und die Konstant-haltung von Drittvariablen (bspw. Instruktion, Situation, Kontext; Stichwort: ***ceteris paribus*** = alles andere gleich) häufig ***künstlich*** sind, haben sie - besonders aus Sicht der angewandten Wissenschaften (z.B. Wirtschaftspsychologie) - auch Nachteile.

Die **ökologische Validität der Untersuchung** ist nämlich umso höher, je ähnlicher die Untersuchungssituation der natürlichen Situation ist, auf die generalisiert werden soll (s.a. Stichwort *externe Validität* in Kap. 3.2.3). Wenn eine Laborstudie sehr künstlich ist, sehr stark vereinfachte, von der natürlichen Komplexität abstrahierte Bedingungen vorgibt, ist ihre ökologische Validität gering (geringe *ökologische Validität* = Künstlichkeit).

WissenschaftlerInnengruppen unterscheiden sich darin, wie wichtig ihnen die ökologische Validität im Vergleich zur internen Validität ist. Im Allgemeinen gilt: je angewandter die Wissenschaftsdisziplin, desto wichtiger ist ihr die *ökologische Validität*. Zur Steigerung der ökologischen Validität wurden die Untersuchungstypen D, E, F entwickelt (Abb. 3_1). Eine der komplexen Lebensrealität stärker entsprechende Abbildung des Forschungsgegenstands in Typ D oder gar die Einbeziehung der ProbandInnen in die Beschreibung und Erklärung ihrer Handlungen in Typ E empfiehlt sich einerseits, um bereits laborexperimentell geprüfte Theorien *in der Praxis* zu *bewähren* (´Anwendung´), andererseits wohl auch, um Forschungsgebiete, für die noch keine Theorien entwickelt wurden, zunächst zu explorieren (Theorie-Entwicklung, Entdeckungszusammenhang in Abb. 1_1).

Zwischen den beiden Ansprüchen - hohe *interne Validität* der Untersuchung aber gleichzeitig auch hohe *ökologische Validität* - besteht ein **Dilemma**: man möchte (insbesondere in den angewandten Wissenschaften) beides. Da eine Untersuchung mit geringer interner Validität keinen guten *Begründungszusammenhang* einer Kausalaussage abgeben kann, weil über die drei Kausalitätsrichtungen in Abbildung 3.2 *immer Zweifel bleiben*, sollte man nur nach reiflicher Überlegung auf Mittel zur Erhöhung der internen Validität zu verzichten bereit sein.

Interessant für angewandte Disziplinen sind Studien, die hohe ökologische Validität (Studien ´im Feld´ z.B. am Arbeitsplatz oder am Point of Sale) mit dem Untersuchungstyp des Experiments (höchste interne Validität) verbinden (*Feldexperimente* statt *Laborexperimente*). Abb. 3_15 zeigt ein gelungenes Feldexperiment.

Tab. 3_7: Feldexperiment

	geringe ökologische Validität	hohe ökologische Validität
geringere interne Validität: UV nicht manipuliert	(misslungene Studie)	Feldstudie mit hoher ökologischer Validität
hohe interne Validität: UV manipuliert!	Laborexperiment	**Feldexperiment**

3.1.4 Operationalisierung der Prädiktorvariable als UV: Experiment

Experiment= „Prüfung von Kausalhypothesen durch systematische *Manipulation der UV*".

Das *psychologische Experiment* ist der Untersuchungstyp mit der höchsten *internen Validität*. Grob werden hier nur zwei Untersuchungstypen unterschieden: das *Experiment* und die *Korrelationsstudie*. Im *Experiment* werden die *Prädiktorvariablen* von der Versuchsleitung (absichtlich & systematisch) *manipuliert*, also als *unabhängige Variablen realisiert* (Abb. 3_3). ´*Experiment*´ ist daher in der Psychologie ein geschützter Begriff, der nur für solche Studien verwendet werden darf, in denen mindestens eine *UV manipuliert worden ist*.

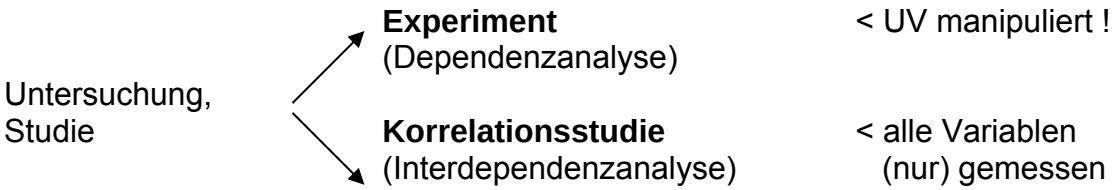


Abb. 3_3: Die wichtigsten beiden Untersuchungstypen.

Prüft man die Hypothese „Bei schwierigen Aufgaben führt Erregung zu schlechteren Leistungen", indem die Erregung vor einer Statistik-Klausur (Puls, Blutdruck, ...) gemessen und die Note der Prüfung erhoben werden, so hat man (trotz eingesetzter Technik bei der Messung des Prädiktors) nur eine Korrelationsstudie durchgeführt (egal ob die Auswertung aus einem Korrelationskoeffizienten, oder bspw. durch einen Mittelwerts-Vergleich der Note von Prüflingen mit hoher versus mit niedriger Erregung durchgeführt wird). Prüft man die Hypothese, indem die Erregung in einer Versuchsgruppe durch zwei Tassen Kaffes absichtlich erhöht und in der *Kontrollgruppe absichtlich* unbeeinflusst gelassen und diese *Bedingungszuordnung* per Münzwurf entschieden wurde, wurde die Erregung (in zwei *Stufen*: erhöht vs. normal) *experimentell manipuliert;* die Hypothesenprüfung erfolgte *experimentell*, weil die *UV manipuliert* worden ist.
Experimente sind am besten geeignet, Kausalhypothesen zu prüfen (X bewirkt Y): Wenn die Variation in der abhängigen Variable (AV) hypothesenkonform von der (experimentell manipulierten) Variation in der UV abhängt (die KandidatInnen nach dem Kaffee schlechter sind als die ohne), ist unwahrscheinlich, dass der Unterschied auf etwas anderes zurückgeht, als auf die Kaffee-Gabe, denn die Zufallszuordnung der Vp zur Kaffee- versus zur Kontrollgruppe (KG) verhindert, dass sich etwas anderes als der Kaffee zwischen beiden unterscheidet (OK, wir geben den Kontrollgruppenmitgliedern Saft, damit sie auch „etwas bekommen" haben). Die zweitbeste Möglichkeit, ohne Experiment für Kausalhypothesen eine umgekehrte Wirkrichtung (s. Abb. 3_2) auszuschließen, erfordert Messungen zu mehreren Zeitpunkten (vgl. Kap. 3.1.7).

Warum heißt die **UV ´unabhängige´ Variable**? Weil ihre Variation von der Versuchsleitung *unabhängig von anderen Versuchsbedingungen oder Versuchspersoneneigenschaften* vorgenommen wird. Unabhängig gelingt diese Manipulation nur, wenn die Versuchsbedingung (z.B. ´mit Kaffee´ versus ´mit Saft´) den Versuchspersonen zufällig (=***randomisiert***) zugewiesen wird (z.B. Münzwurf). Gäbe man einfach denen den Kaffee, die blass aussehen, oder denen, die früh ankommen etc, so ist wahrscheinlich eine wichtige Variable (Erregung? Prüfungsangst?) mit der Kaffee-Gabe *konfundiert*, hier hätte man der Hypothese unfair nachgeholfen (das Schlimmste in der Wissenschaft! Man hätte die Hypothese gegen die Empirie *immunisiert*). Die Kaffee-Gabe muss von Versuchspersoneneigenschaften *unabhängig* erfolgen, echte *Randomisierung* über Münzwurf, Zufallsabfolge aus dem PC etc, ist nötig!

Die Prädiktorvariable Geschlecht lässt sich offensichtlich nicht experimentell variieren (die Variable ´hängt´ an der Versuchsperson, die Versuchsleitung kann sie nicht unabhängig zuweisen). Grob kann man sagen, dass in der Differentiellen Psychologie (= Persönlichkeitspsychologie) Experimente selten (Persönlichkeit ist halt mit der Vp verbunden), in der Allgemeinen Psychologie Experimente häufig sind (Situationen und Stimuli lassen sich herstellen und zuweisen). Da die Sozialpsychologie soziale Bedingungen sowohl in der Person als auch in der Umwelt kennt, gibt es sozialpsychologische Forschungsbereiche, in denen viel, und solche, in denen weniger experimentiert wird.
Wird eine *natürlich vorgefundene* X-Variable wie das Geschlecht (deren Ausprägungen in

bestimmten *Stufen* vorliegt, die also *nominalskaliert* ist, vgl. Kap. 3.2.1), in ein Versuchsdesign aufgenommen, dann wird manchmal von einem **Quasiexperiment** gesprochen, weil der Versuchsplan (= das Versuchsdesign; vgl. z.B. Tab. 3_8, Tab. 3_9) scheinbar (also „quasi“) wie der eines Experiments aussieht. Quasiexperimente sind aber Korrelationsstudien (Abb. 3_3), sie sind, weil die Prädiktorvariable nicht manipuliert wurde, keine Experimente! In der Beschreibung von Studien findet man daher manchmal Sätze der Art: „die Erregung wurde experimentell manipuliert, während die Aufgabenschwierigkeit quasiexperimentell realisiert worden war“ (also gab es vielleicht Kaffee oder Saft per Münzwurf, aber es wurden vorgefundene Fächer verglichen).

Auch wenn *Experimente* das anzustrebende Ideal der Hypothesenprüfung darstellen, ist dieser Untersuchungstyp oft nicht anwendbar, weil die Komplexität des Gegenstands und/oder ethische Grenzen eine gezielte Bedingungsmanipulationen nicht zulassen. Nach den **Ethischen Richtlinien** der Deutschen Gesellschaft für Psychologie (DGPs) und des Berufsverbands deutscher PsychologInnen (BDP) darf die „*Würde und Integrität der teilnehmenden Personen nicht beeinträchtigt werden*“. In einer früheren Fassung hieß es: „*In den Ausnahmefällen, in denen eine vollständige Information vor der Versuchsdurchführung mit dieser nicht vereinbar ist, muss in besonderem Maße sichergestellt sein, dass den Versuchspersonen durch ihre Teilnahme kein Schaden entstehen kann. In diesem Fall sind die Versuchspersonen in allgemeiner Form über die mangelnde Aufklärung zu informieren. Nach Abschluss der Untersuchung sind die Probanden aufzuklären. Versuchspersonen müssen zumindest nachträglich umfassend über Zweck und Vorgehen in der Untersuchung aufgeklärt werden*“- in der aktuellen Fassung ist der Text zum Würdeerhalt bei Experimenten verteilter, lesen Sie bitte Punkt C-III (1)-(9) der Ethischen Richtlinien der DGPs & des BDP: www.bdp-verband.org/bdp/verband/ethik.shtml.

Ist eine experimentelle Bedingungsvariation möglich, dann sind experimentelle Ergebnisse sehr viel wertvoller für den *Begründungszusammenhang* einer Aussage als die von Korrelationsstudien. Daher sind Ideen zur Manipulationen von UV gefragt. Wenn bspw. nicht eigentlich das biologische Geschlecht, sondern die Geschlechtsrollenidentifikation (gender) in einer Kausalhypothese wirkt, ließe sich bei einer Teilgruppe der Vp ihre Geschlechtsrolle vor der AV-Erhebung aktivieren („ich als Frau …“, „ich als Mann, …“); diese beiden Gruppen sollten stärker gender-konform agieren als die beiden Kontrollgruppen ohne Geschlechtsrollenaktivierung.

Warum sichert nur die **Manipulation der UV** die **interne Validität** der Untersuchung?

- werden X und Y nur gemessen, so bleibt die Kausalitätsrichtung zwischen den Variablen (s. Abb. 3_2) letztlich unklar (daher der Name: *Inter*dependenzanalyse in Abb. 3_3).
- in Korrelationsstudien können Befunde evtl. nur *Scheinkorrelationen* sein (s. Störche-Beispiel in Kap. 3.1.5), die Variation der AV kann nämlich durch Variation einer dritten Variable *Z* verursacht sein, die mit *X konfundiert* war (s.u.). Durch die systematische Manipulation der UV im Experiment, also durch die randomisierte Zuweisung der UV-Bedingung zur Vp wird sichergestellt, dass *keine andere Variable mit der UV konfundiert* ist.

Ein Experiment prüft die H1 ´**ceteris paribus**´ = Alles andere gleich. Dazu wird in Experimenten versucht, neben der manipulierten UV andere *Randbedingungen* (Bedingungsvariablen, die nicht in der Hypothese formuliert sind) konstant zu halten (beispielsweise die Aufgabenschwierigkeit bei Prüfung der Hypothesen aus Tab. 3_2). Bedingungsvariablen, die nicht konstant gehalten werden können oder zugunsten der *externen Validität* (z.B. der Situationskomplexität, siehe ´Typ D´ im Methodenspektrum von Abb. 3_1) nicht konstant gehalten werden sollen, vergrößern zwar die *Fehlervarianz* (= der Befund wird weniger deutlich sichtbar), variieren aber in den UV-Bedingungen gleichermaßen (wenn diese den Vp *randomisiert* zugewiesen wurden). Variierende andere Bedingungen, sog. Drittvariablen ´Z´ (Z genannt, da X und Y für hypothesenrelevante Variablen vergeben sind), verringern die interne Validität der Untersuchung nur dann, wenn sie mit der UV korreliert (=*konfundiert*) sein könnten. **Konfundierte Variablen** sind Störvariablen, die mit der Prädiktor-Variable konfundiert = korreliert waren. Um sie zu vermeiden, müssen in Experimenten die Vp zu den Bedingungen *per Zufall* (Münze, Würfel, etc) - **randomisiert** - zugewiesen werden (s.o.).

3.1.5 Konfundierende Variablen, Beispiel für Scheinkorrelation: Bringen Störche die Kinder? Kontrollvariablen, mehrfaktorielle Designs

Geprüft wird die kausal gemeinte Hypothese „Die Störche bringen die Kinder“, statistisch als *gerichteten Zusammenhangshypothese* formuliert: H1: „je mehr Störche, desto mehr Kinder“. Es werden Störcheaufkommen (X) und Geburtenrate (Y) als Variablen isoliert. Da niemandem ein Experiment eingefallen ist, wurde nur eine Korrelationsstudie durchgeführt, in dem beide Variablen in mehreren Ländern gemessen wurden (oder nachgeschlagen: Y gibt’s für 2014 auf https://www.cia.gov/library/publications/the-world-factbook/rankorder/2054rank.html).
Es resultiert eine positive Korrelation, z.B. r = +.62 (MATTHEWS, 2000) in Abb. 3_4 (Das *Korrelationsmaß r* wird öfter erwähnt, erklärt wird es in Kap. 4.2).

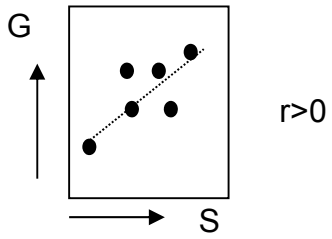


Abb. 3_4: Fiktives Ergebnis der Korrelation von Störcheaufkommen S und Geburtenrate G über mehrere Länder

Auf dem Kongress, auf dem das Ergebnis mit Abb. 3_4 vorgestellt wird, wird eine kritische (sich für Wissenschaftlichkeit einsetzende) Person bemerken, dass der *Begründungszusammenhang* der (Kausal-) Aussage ´Störche bringen Kinder´ nicht stichhaltig, also die *interne Validität der Untersuchung* gering ist, da die Korrelation der AV ´Kinder´ (*operationalisiert* über die Geburtenrate) mit der X-Variable ´Störche´ (gemessenes Störcheaufkommen im Land) statt über Kausalität auch über eine **Drittvariable** verursacht sein kann, z.B. durch die ´Industrialisierung´! Der Industrialisierungsgrad nämlich sei mit dem Störcheaufkommen **konfundiert:** Störcheaufkommen und Industrialisierung sind nämlich korreliert, und zwar negativ korreliert (Ländern mit niedrigem Störcheaufkommen sind oft hoch industrialisierte, Länder mit vielen Störchen noch nicht) und Industrialisierung verursacht auch den Geburtenrückgang (da Frauen in Industrieländern arbeiten gehen). Die ***konfundierende Variable*** (Industrialisierung) war zuvor unentdeckt, ist mit der X-Variable korreliert und wirkt vielleicht anstelle der X auf die Y;
- der Verdacht alleine genügt, um die *interne Validität der Untersuchung* zu ruinieren.

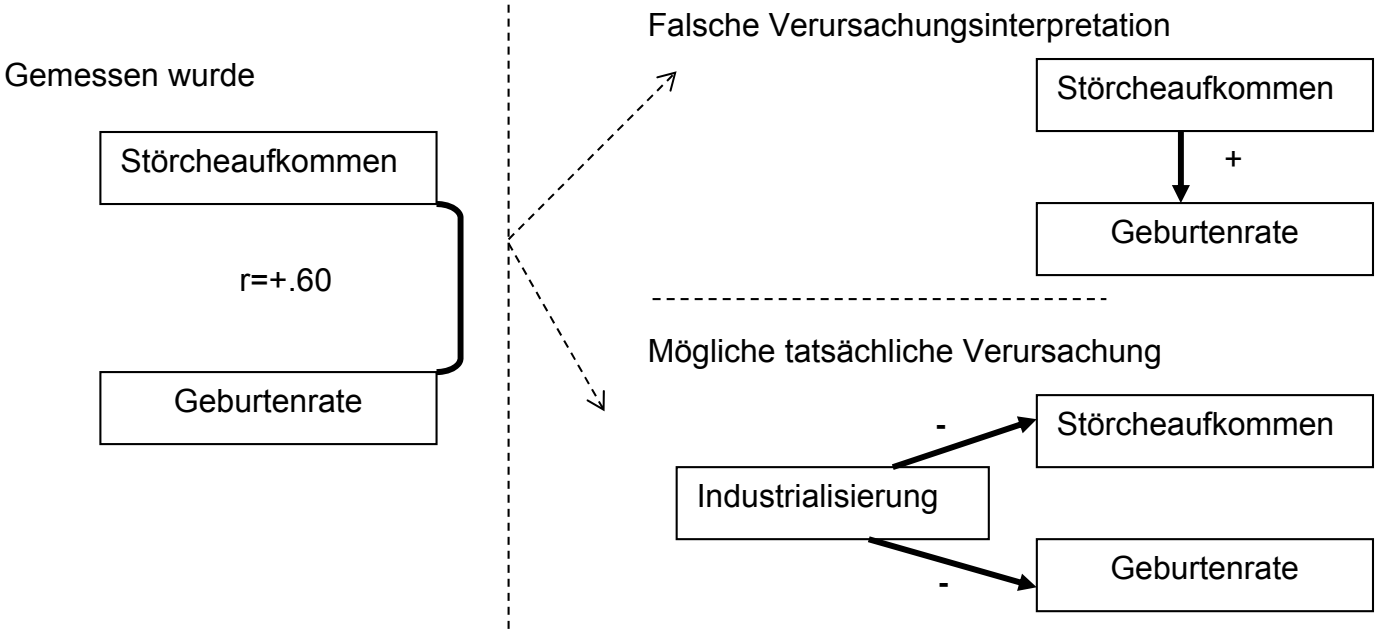


Abb. 3_5a: Wie eine Scheinkorrelation (links) durch die Wirkung einer Drittvariable verursacht wird.

Korrelationen (ein ´Zusammenhangsmaß zwischen zwei Variablen´) sollten daher nicht *kausal interpretiert* werden. Jedenfalls ist die *interne Validität* der Studie gering: die Kausalitätsrichtung könnte auch andersherum gewirkt haben

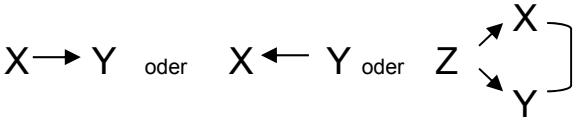


Abb. 3_5b: Eine Korrelation r(x,y) kann drei Ursachen haben.

(Kinder ziehen Störche an) oder von einer Drittvariabel ausgehen (Industrialisierung lässt Störche- und Geburtenrate sinken) (s. Abb. 3_5b, die Abb. 3_2 wiederholt).
Die Zweifel an Kausalitätsrichtung und Drittvariablenbeteiligung (Abb. 3_5b), die jede/r bei Korrelationsstudien immer äußern darf, setzten also die *interne Validität der Studie* zur Prüfung der Kausalaussage X→Y herab. Für einen intern valideren Begründungszusammenhang der (Kausal-)Aussage ´Störche bringen Kinder´ (X bewirkt Y) müsste

(1) das Störcheaufkommen (X) *experimentell manipuliert* werden!
Wenn dies aus z.B. ethischen oder zeit-ökonomischen Gründen nicht machbar ist, sollte versucht werden,
(2) mögliche *konfundierte Variablen ´Z´* zu identifizieren (Nachdenken & Literaturstudium & mit kritischen Personen diskutieren) und diese Z dann
(2.1) *konstant zu halten* (nur Länder gleich hoher Industrialisierung zu untersuchen) oder
(2.2) innerhalb der X zu *parallelisieren* (ebenso viele hochindustrialisierte und niedrig industrialisierte Länder in der Gruppe der Länder mit geringem Störcheaufkommen wie in der Gruppe der Länder mit hohem Störcheaufkommen zu untersuchen) oder/ und
(2.2) Z ebenfalls zu messen und als ***Kontrollvariablen*** in ein dann *zweifaktorielles* Design einzuführen (Tab. 3_9 und Abb. 3_8).

Durch die Aufnahme einer *Drittvariable Z* wird aus der ursprünglich *einfaktoriellen* (nur eine X-Variable: Störcheaufkommen) und *univariaten* (nur eine Y-Variable: Geburtenrate) Untersuchung ein *zweifaktorielles Design (*zwei X-Variablen: Störcheaufkommen und Industrialisierungsgrad). Im Beispiel wird das Ergebnis (s. Abb. 3_8) lauten: **Unter Kontrolle** des Industrialisierungsgrades lässt sich kein Einfluss des Störcheaufkommens auf die Geburtenrate nachweisen (die *Nullhypothese* wird *beibehalten*).

Namen von Versuchsdesigns: Insbesondere in *Experimenten* (aber auch in *Quasiexperimenten*: also *Korrelationsstudien* mit nur *nominal* gestufter X-Variable, z.B. einem Geschlechtervergleich, s. Kap. 3.1.4, oder einem Zeitenvergleich wie in den *Prä-Post-Designs* von Evaluationsstudien, s. Kap. 3.1.7) wird das *Versuchsdesign* durch die Anzahl von *Prädiktorvariablen* (die dann *Faktoren* heißen, s. Tab. 3_5) sowie die Anzahl von *abhängigen Variablen* (= Variaten) bezeichnet (Abb. 3_6).

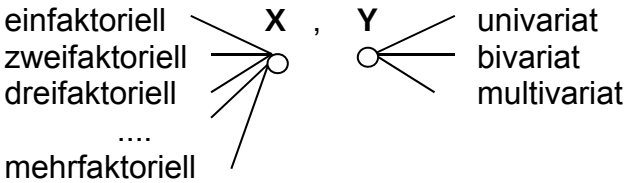


Abb. 3_6:
Benennung von Versuchsdesigns
(dt. Anzahl X-Variablen = Faktoren,
lat. Anzahl Y-Variablen = Variaten, vgl. Tab. 3_5)

Wird die letzte Hypothese aus Tab. 3_4 „H1: In schwierigen Aufgaben nimmt die Leistung mit steigender Erregung ab, in einfachen Aufgaben zu“, durch Manipulation der ´UV1: Erregung´ und der ´UV2: Aufgabenschwierigkeit´ sowie Messung der ´AV: Leistung´ geprüft, so wird ein *zweifaktorielles univariates Design* realisiert.

Tab. 3_8: Zweifaktorielles univariates Versuchsdesign zur Prüfung der letzten Hypothese aus Tab. 3_4.

		UV2:	
		Aufgabe leicht	Aufgabe schwer
UV1:	Erregung hoch		
	Erregung niedrig		

* AV: Leistung

Das *Versuchsdesign*, synonym der *Versuchsplan,* wird im Methodik-Kapitel eines Forschungsberichts (Kap. „3.1 Ziele und *Design“*) oft in Tabellenform aufgeführt: z.B. Tab. 3_8. Im Design in Tab.3_8 wurden die UV1 Erregung und die UV2 Aufgabenschwierigkeit jeweils zweifach *gestuf*t: UV1 *wird in den Stufen* ´hoch / niedrig´ *realisiert*, UV2 in den Stufen ´leicht / schwer´ variiert (z.B. ließen sich die UV1-Stufen durch die Bedingungen „mit Kaffee“ versus „mit Saft“ und die UV2-Stufen durch die (vermutlich quasiexperimentelle) Bedingungen „Kunst- / Matheklausur“ realisieren (=operationalisieren)).
Ein *mehrfaktorielles Design* wird auch gern über die Multiplikation der *Faktorstufen*(anzahl)

benannt. Tab. 3_8 zeigt also ein **2 x 2 - Design**. Wenn jede UV aber bspw. in drei Stufen realisiert wird (z.B. ´niedrig, mittel, hoch´ für UV1 und ´leicht, mittel, schwer´ für UV2), dann handelt es sich um ein **3 x 3 - Design**. Dieses ist weiterhin *zweifaktoriell* und hat nun 3 x 3 = 9 *Zellen*. Soll dann noch bspw. das Geschlecht kontrolliert werden, entsteht ein 3x3x2-Desing mit 18 Zellen.

Design mit Kontrollvariable: Für die Störche-Kinder-Hypothese (s. Abb. 3_4 & 3_5) könnte versucht werden, dem Einwand der Industrialisierungs-*Konfundierung* zu begegnen, indem der Industrialisierungsgrad des Landes berücksichtigt und als *Kontrollvariable* in das dadurch dann *zweifaktorielle Design* von Tab. 3_9 eingeführt wird.

Tab. 3_9: Quasiexperimentelles zweifaktorielles Versuchsdesign zur erneuten Prüfung der Störche-Kinder-Hypothese (H1: positiver Zusammenhang von Störcheaufkommen und Geburtenrate auch unter Kontrolle der Industrialisierungsstufe)

	Störcheaufkommen unterdurchschnittlich	Störcheaufkommen überdurchschnittlich
Industrialisierung überdurchschnittlich	<	
Industrialisierung unterdurchschnittlich	<	

* AV: Geburtenrate

Aus Sicht der Störche-Kinder-Hypothese sollte die Industrialisierung keine Rolle spielen (da die Kinder ja von den Störchen ...), die Industrialisierung ist dann eben (nur) eine *Kontrollvariable*. Das *Versuchsdesign* in Tab. 3_9 ist *zweifaktoriell* und *univariat* (s. Abb. 3_6), es ist ein 2 x 2 Design. Erwartet wird (dies zeigen die ´kleiner als´ - Zeichen in Tab. 3_9), dass die Geburtenrate mit dem Störcheaufkommen (auch innerhalb der Industrialisierungsstufen) variiert. Ergebnis wird allerdings wohl sein, dass sie mit dem Industrialisierungsgrad, aber - unter Kontrolle des Industrialisierungsgrads - nicht mehr mit dem Störcheaufkommen variiert ☺.

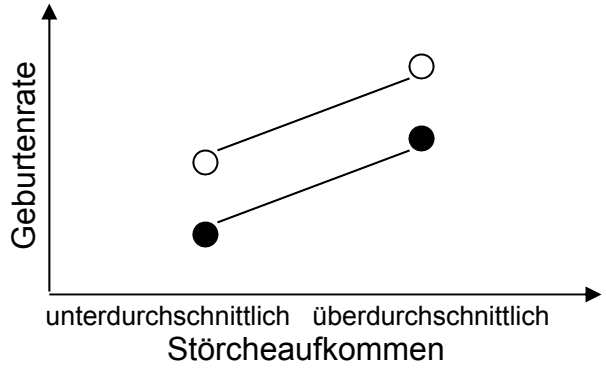


Abb. 3_7: Hypothese: Wirkung der Störche auf die Kinder auch unter Kontrolle des Industrialisierungsgrads

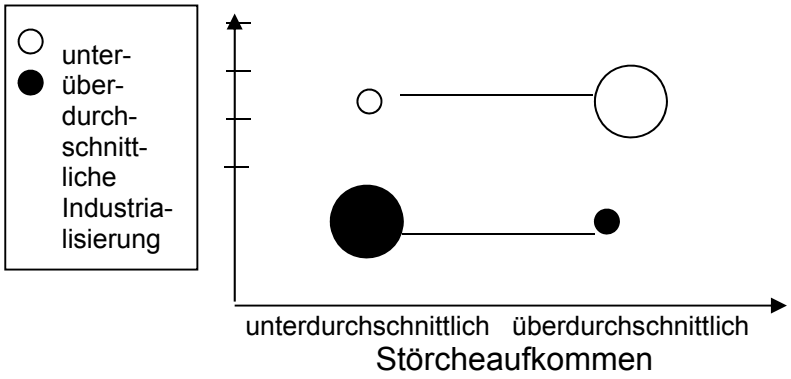


Abb. 3_8: Ergebnis: Unter Kontrolle des Industrialisierungsgrads keine Wirkung der Störche auf die Kinder.

Wie in Tab. 3_9 vorgenommen, empfiehlt es sich, Hypothesen wo möglich in den Versuchsplan einzuzeichnen (größer/kleiner Zeichen für Mittelwertsvergleiche). Um sich über die exakte Prüfhypothese klar zu werden und die Auswertung zu planen, sollte man sogar zusätzlich zum Versuchsplan eine Ergebnisdarstellung entwerfen, in der die erwarteten Ergebnisse hypothesenkonform skizziert werden (Beispiel Abb. 3_8 bzw. 3_7).

Wenn ein hoher Industrialisierungsgrad sowohl das Störcheaufkommen verringert (die Industrialisierung mit der X-Variable *konfundiert* ist), als auch die Geburtenrate senkt (also für die Variation der AV verantwortlich ist), kann die Korrelation von r=+.60 aus Abb. 3_4 als ***Scheinkorrelation*** entlarvt werden: unter Kontrolle des Industrialisierungsgrads (Tab 3_9) tritt sie (die Korrelation) nämlich nicht mehr auf (Abb. 3_8): Die vielen Länder mit hoher Industrialisierung und wenigen Störchen (großer schwarzer Kreis in Abb. 3_8) haben eine ebenso niedrige Geburtenrate wie die wenigen Länder mit hoher Industrialisierung aber vielen Störchen (kleiner schwarzer Kreis). Die wenigen Länder mit geringer Industrialisierung und wenigen Störchen (kleiner weißer Kreis) haben eine ebenso hohe Geburtenrate wie die vielen

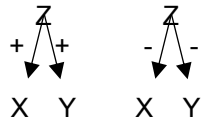
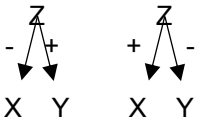
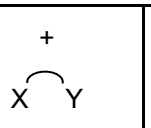
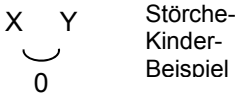
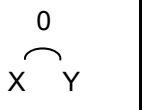
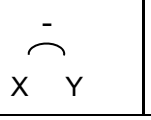
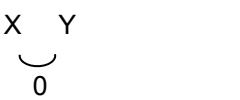
Länder mit geringer Industrialisierung aber vielen Störchen (großer weißer Kreis). => das Störcheaufkommen hat keine Wirkung auf den Geburtenrückgang!

In Übertragung auf eine nichtexperimentelle Prüfung der einseitigen Unterschiedshypothese aus Tab. 3_1 könnte einem *quasiexperimentellen* Design (Prädiktor Erregung, gemessen über Puls vor der Prüfung, anschließend Einteilung in die Gruppen hohe / niedrige Erregung) noch eine Kontrollvariable ´Dauer des Lernens´ zugefügt werden, um die *interne Validität der Untersuchung* zu steigern. Denn das ´Lernen´ könnte mit der X-Variable *konfundiert* sein: vielleicht sind alle, die wenig gelernt haben, deshalb erregt; alle, die viel gelernt haben, ruhig? Und vielleicht bekommen ruhige Personen eine bessere Note in der schwierigen Aufgabe, aber nicht, weil sie ruhig waren, sondern weil sie viel gelernt hatten! Entwerfen Sie als Übung hierzu einen Versuchsplan und ein zur H1 aus Tab. 3_1 konformes Ergebnis in einer Abbildung sowie eine Ergebnis-Abbildung, die die Erregungswirkung als Scheinkorrelation entlarvt.

Und wie sollte Forschung zur Frage, ob Konsum gewalthaltiger Mediendarstellungen zu Gewalt führt, fortgesetzt werden, nachdem sich nicht nur der umgekehrte Kausalzusammenhang ebenfalls bewährt hat (gewalttätige Personen suchen entsprechende Unterhaltung), sondern auch plausible konfundierte Drittvariablen vorgeschlagen wurden (z.B. Herkunfts-Milieu)? Zeichnen Sie ein Variablennetz wie in Abb. 3.5 (und suchen Sie es aus Tab. 3_10 aus), einen Versuchsplan wie in Tab. 3.9 und mögliche Ergebnisse wie in Abb. 3.8- 3.9 zu diesem Problem.

Und üben Sie, konfundierende Variablen in Studien zu finden, die einen Zusammenhangsbefund zwischen X und Y als Scheinkorrelation entlarven können, weil die von Ihnen vorgeschlagene Drittariable Z sowohl auf X als auch Y kausal wirkt, X und Y dann keine Wirkung mehr aufeinander haben brauchen, wie Störche und Kinder eben nicht.

Tab. 3_10: Was mit einer einfachen bivariaten Korrelation passiert, wenn die Drittvariabel Z, die beide Variablen X und Y beeinflusst, auspartialisiert wird.

bivariate Korrelation war	Drittvariable mit **gleicher** Vorhersagerichtung 	Drittvariable Z mit **verschiedener** Vorhersagerichtung 
	 Störche-Kinder-Beispiel	
		
		

Tab. 3_10 zeigt die möglichen Fälle der Drittvariablen-Wirkung abstrakt: hat die Drittvariable auf X und Y gleiche Wirkung (die Industrialisierung erniedrigt Störche- und Geburtenrate), so verschwindet eine ursprünglich positive Korrelation zwischen X und Y, eine ursprünglich vorhandene Nullkorrelation wird zu einem negativen Zusammenhang. Sich das Störche-Kinder-Beispiel zu merken (Tabelle 3_10 links oben) erlaubt, sich Tab. 3_10 selbst zu konstruieren.

Und ein noch komplizierteres Denk-Beispiel (`Simpson-Paradox`): H1: Schulen in Süddeutschland scheinen besser zu fördern, denn sowohl in Gymnasien als auch in Realschulen (Kontrollvariable Schultyp) sind SchülerInnen aus Süd- im Vgl. zu Norddeutschland in bundeseinheitlichen Mathetests im Schnitt besser. H1: je Südlicher, desto bessere Leistung auch unter Kontrolle des Schultyps. Die Aussage kann angezweifelt werden, da der Schultyp den Kindern nach anderen Bedingungen zugewiesen wird: in Süddeutschland ist die Aufnahme aufs Gymnasium strenger. Bspw. kann ein Kind mit mittleren Leistungen in Süddeutschland nur auf die Realschule, in Norddeutschland schon aufs Gymnasium. Damit wird es den Leistungsmittelwert der norddeutschen Gymnasien nach unten drücken. Und da ´die mittlere Kinder´ in den norddeutschen Realschulen fehlen (weil sie dort auf dem Gymnasium sind), verbleiben in norddeutschen Realschulen mehr schlechtere; auch der Leistungsmittelwert der norddeutschen Realschulen liegt damit unter dem der süddeutschen, der durch die dortigen mittleren Kinder gehoben wird.

Durch konfundierte Variablen wird die Interpretation von Korrelationen kompliziert: ein guter Grund, um **über Kausalhypothesen per Experiment zu entscheiden**.

3.1.6 Haupt- und Interaktionseffekte in mehrfaktoriellen Designs

Nach Einführung von Kontrollvariablen (wie in Tab. 3_9) und zur Prüfung von komplizierteren Hypothesen (bspw. der letzten aus Tab. 3_4) werden *mehrfaktorielle Designs* benötigt (als mehrfaktoriell wird, in grobem Umgang mit Abb. 3_6, ein Design auch schon ab zwei Faktoren bezeichnet). In *mehrfaktoriellen Designs* können *Haupteffekte* und *Interaktionseffekte* geprüft werden!

Von einem ***Haupteffekt*** (Haupteffekt einer X-Variable) spricht man, wenn die X-Variable auf die AV (die Y-Variable) generell wirkt, also über verschiedenen anderen Bedingungen (z.B. die Stufen einer Kontrollvariable Z) hinweg. Der Faktor Aufgabenschwierigkeit (UV2 in Tab. 3_8) wird vermutlich einen *Haupteffekt* auf die AV Leistung haben: in schwierigen Aufgaben resultiert generell geringere Leistung als in leichten (so in Abb. 3_9 oben-rechts und unten beide: der Mittelwert aller weißen Punkte liegt höher als der Mittelwert alle schwarzen Punkte).

Ebenso kann man nach einem Haupteffekt der Erregung suchen: nur in Abbildung 3.9 unten rechts liegt ein Haupteffekt der Erregung vor (hohe Erregung veringert die Leistung generell, egal welche Schwierigkeit die Aufgabe hat).

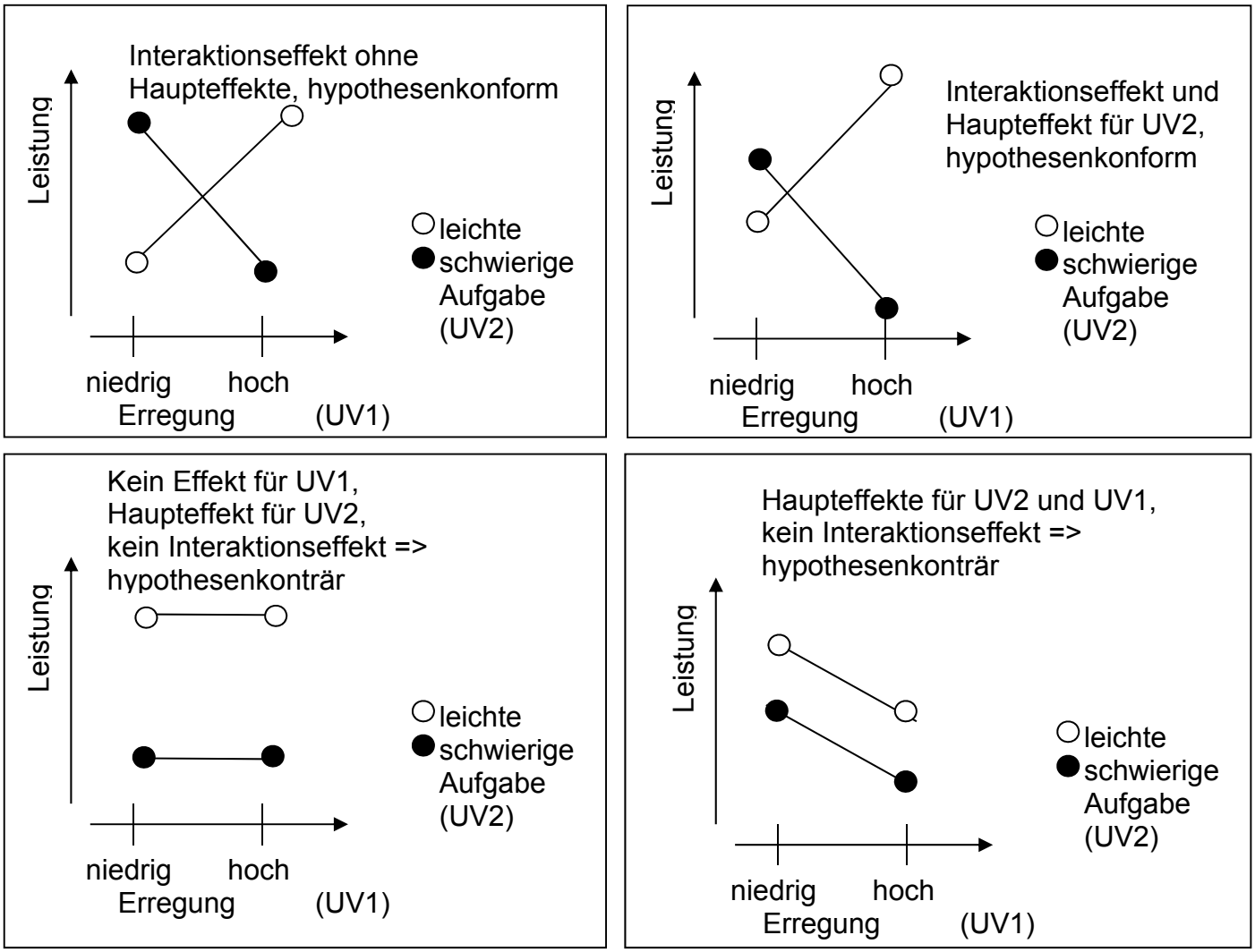


Abb. 3_9: Einige der möglichen Ergebnisse für das Versuchsdesign aus Tab. 3_8 zur Prüfung der untersten Hypothese aus Tab. 3_4 (entwerfen Sie zur Übung andere Mittelwerts-Muster und interpretieren Sie sie).

Drittens frage man nach einem Interaktionseffekt von Aufgabenschwierigkeit und Erregung: Ist der Leistungsrückgang durch schwerere Aufgaben bei hoher und niedriger Erregung gleich? Wenn ja (Abb. 3_9 unten-links), liegt nur ein *Haupteffekt* der Aufgabenschwierigkeit vor und keine Interaktion. Wenn die Linien aber nicht parallel sind, wenn sich also die Wirkung der Aufgabenschwierigkeit in den Bedingungen mit und ohne Kaffee (UV1 Erregung) unterscheidet (Abb. 3_9 oben), dann wird die Wirkung der Aufgabenschwierigkeit auf die Leistung durch die Erregung ***moderiert*** (=verändert), es liegt ein *Interaktionseffekt* von UV1 und UV2 auf die AV

vor. Wenn eine Variable den Effekt einer anderen Variable *moderiert* (in der Theoriedarstellung ein Pfeil auf einen anderen Pfeil zeigt wie in Abb. 2.2), dann gibt es eine **Interaktion** beider Prädiktoren. Zur Prüfung der „H1: In schwierigen Aufgaben nimmt die Leistung mit steigender Erregung ab, in einfachen Aufgaben (aber) zu“ wird ein *Interaktionseffekt* der Variablen Erregung und Aufgabenschwierigkeit (eine Erregungs-*mal*-Aufgabenschwierigkeits-Interaktion) behauptet (egal ob Haupteffekte auch noch vorliegen oder nicht). *Die Prädiktoren interagieren miteinander:* Erregung allein kann die Leistung nicht vorhersagen, denn mal steigt die Leistung mit steigender Erregung, mal fällt sie mit steigender Erregung (Abb. 3_9 oben). Die Aufgabenschwierigkeit allein kann die Leistung auch nicht vorhersagen, die Aufgabenschwierigkeit hat in Abb. 3_9 links-oben sogar keinen Haupteffekt. In Abb. 3_9 rechts-oben hat die Aufgabenschwierigkeit zwar einen *Haupteffekt* (denn der Mittelwert der beiden weißen Punkte dort ist höher als der Mittelwert der beiden schwarzen Punkte), aber es gibt dort zusätzlich auch eine *Interaktion* von Aufgabenschwierigkeit und Erregung (denn die beiden Linien sind nicht parallel). Haupteffekte und Interaktionseffekte können allein oder auch gemeinsam auftreten: AV = Haupteffekt UV1 + Haupteffekt UV2 + Interaktion UV1 x UV2.

Das Gleiche in Schreibweise des *Allgemeinen Linearen Modells* der Statistik (*ALM*, engl. *GLM*):
Y = a + b1 * X1 + b2 * X2 + b3 * X1 * X2.
(mit b1, b2, b3 Gewichte, die zeigen, wie stark Haupteffekte und Interaktion ausfallen).

Nochmal zum Üben: Einen ***Interaktionseffekt*** in Ergebnisbildern wie in Ab. 3.9 erkennt man daran, dass die Linien nicht-parallel sind (obere Bilder in Abb. 3_9). Um über das Vorliegen eines Haupteffekts zu entscheiden (z.B.: liegt ein Haupteffekt der Aufgabenschwierigkeit im Ergebnismuster von Abb. 3_9 rechts-oben vor?), mittelt man über die Stufen der anderen Faktoren (also einen Mittelwert der zwei weißen Punkte rechts oben in Abb. 3_9 und einen weiteren Mittelwert der zwei schwarzen Punkte rechts oben in Abb. 3_9 einzeichnen), dann kann man sehen, ob diese Mittelwerte einen Unterschied aufzeigen (in Abb. 3_9 rechts-oben liegt das Mittel der zwei weißen höher als das Mittel der zwei schwarzen Punkte). Dieses (tatsächliche oder gedankliche) Mittelwerte-Einzeichnen mache man für jede UV gesondert, denn jede UV kann ja ihren eigenen Haupteffekt haben.

Die Begriffe ***moderierende Variable*** **vs.** ***mediierende Variable*** (*moderieren* = verändern vs. *mediieren* = vermitteln) visualisiert Abb. 3_10: Eine ***moderierende Variable*** verändert den Zusammenhang zweier anderer Variablen, sie ist also an einem *Interaktionseffekt* beteiligt. Bspw. moderiert die Variable Aufgabenschwierigkeit die Wirkung der Erregung auf die Leistung (Abb. 2_2; damit moderiert sie auch die Wirkung der Anwesenheit anderer auf die Leistung!). Eine ***mediiererende Variable*** hingegen vermittelt den Zusammenhang zweier anderer Variablen, sie wird also von der einen erhöht und erhöht ihrerseits dann die andere (sie ist ´wie ein Kanal´ zwischen die Wirkung von X auf Y geschaltet, Abb. 3_10 rechts). Die Erregung in Zajoncs Theorie der Sozialen Erleichterung (Abb.2_2) *mediiert die Wirkung* der Anwesenheit anderer *auf* die Leistung („Z mediiert die Wirkung von X auf Y“). Das psychologische Konstrukt (in Abb. 2_1) *mediiert* die Wirkung des Stimulus auf das Verhalten. Die mediierende Variable *erklärt* also, warum Anwesenheit anderer auf die Leistung (oder der Stimulus auf das Verhalten) wirkt: nämlich weil Anwesenheit anderer auf die Erregung wirkt und Erregung auf die Leistung.

Abb. 3_10:
Die statistischen Begriffe *Moderation vs. Mediation*

Interaktionseffekte nachzuweisen erfordert, alle beteiligten *Prädiktor-Variablen* (die *moderierende* Variable gehört dazu) in ein gemeinsames und damit *mehrfaktorielles* Design aufzunehmen (daher das Variablen-Isolieren wie in Tab. 3_4: man darf keinen Prädiktor, keine moderierende Variable übersehen). Die Empfehlung, hypothesenkonforme Ergebnismuster bereits während der Operationalisierung (s. Mitte von Abb. 1_1) als Abbildung darzustellen (Skizze wie in Abb. 3_9 u. 3_12b), wird für mehrfaktorielle Versuchspläne leicht kompliziert, ist

aber zur Auswertungsplanung umso wichtiger. Behauptet die Hypothese einen Interaktionseffekt, liegen aber über Haupteffekte keine Aussagen vor (wie in der letzten Hypothese aus Tab. 3_4), kann man sich mit dem Einzeichnen von ´größer als´ & ´kleiner als´ -Symbolen in den Versuchsplan begnügen (in Tab. 3_8 je ein Zeichen pro Spalte - Versuchen Sie es!). Werden auch Haupteffekte behauptet, so können deren ´>´ oder ´<´ Zeichen an den Spalten- oder Zeilendurchschnitt gezeichnet werden (z.B. Haupteffekt der Aufgabenschwierigkeit ein ´>´ unterhalb der Tab. 3_8; versuchen Sie auch das).

Einige andere Beispiele für Interaktionseffekte (neben dem in Abb. 2_2) enthält Abb. 3_11.

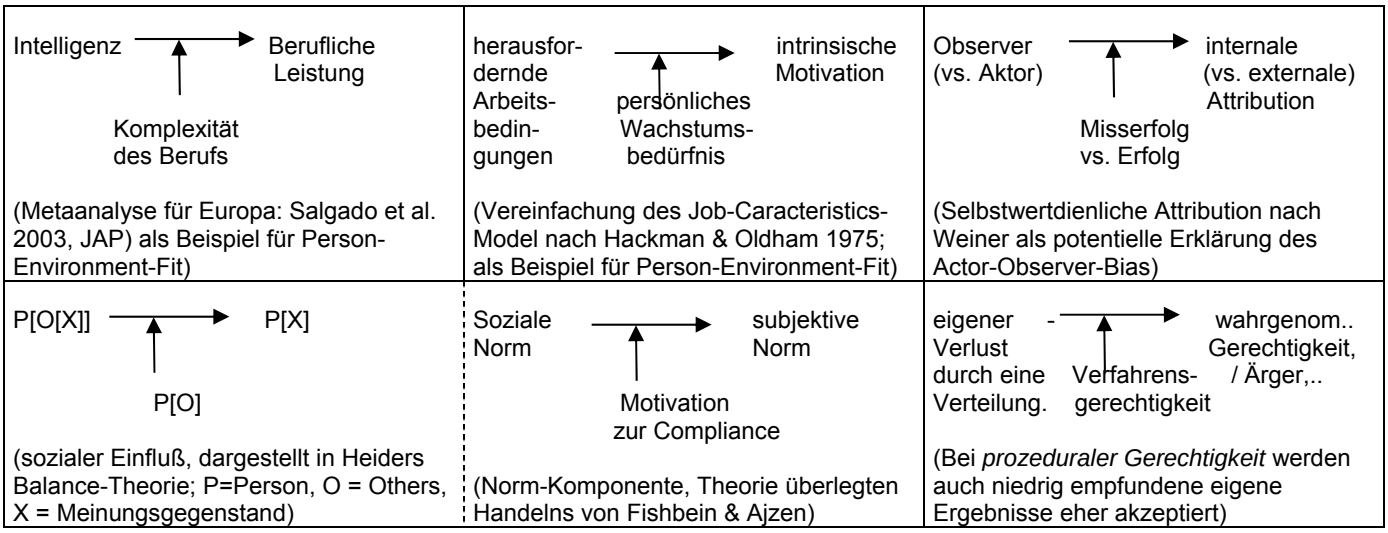


Abb. 3_11: Einige interessante Interaktionseffekte aus sozialpsychologischen Theorien

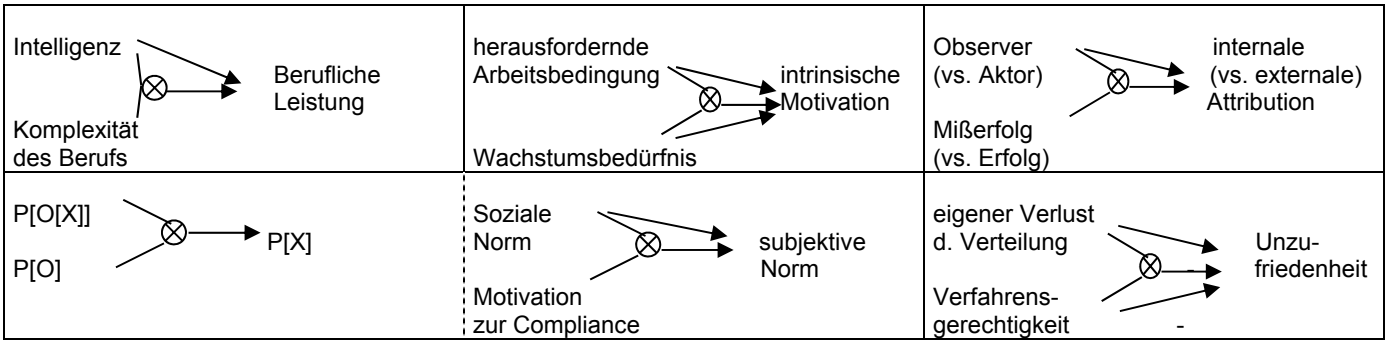


Abb. 3_12a: Andere Darstellung der Interaktionseffekten aus Abb. 3_11 inkl. (von Verf.) erwarteter Haupteffekte.

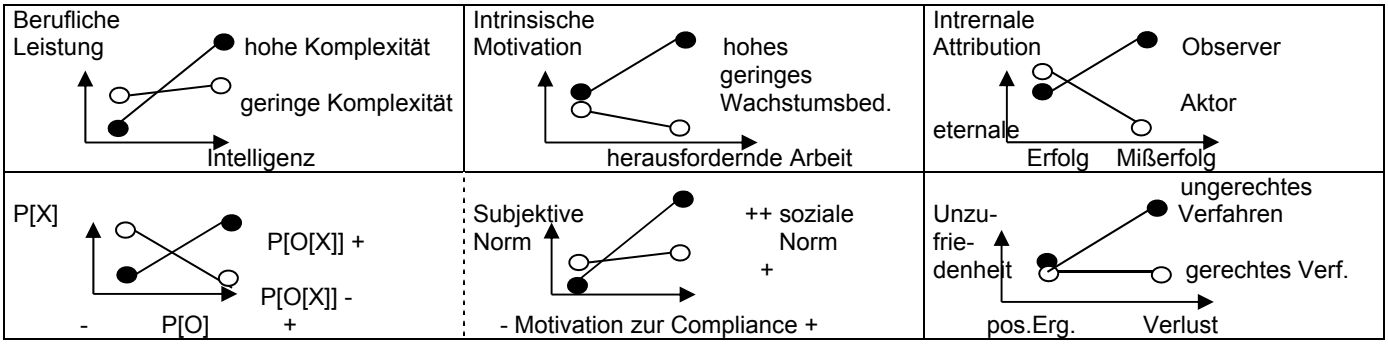


Abb. 3_12b: Erwartete Ergebnismuster zu Abb. 3_11 & 3_12a.

Die Zeichnungsweise „Pfeil auf Pfeil“ in Abb. 3_11 allein lässt nicht mit Sicherheit erschließen, ob der dargestellte Interaktionseffekt ohne Haupteffekt(e) die AV bestimmt (wie in Abb. 3_9 oben links) oder ob es zusätzlich Haupteffekte gibt (Abb. 3_11 oben links: auch über alle Berufe gemittelt erhöht Intelligenz die berufliche Leistung, aber in komplexeren Berufen besonders). Den Theorien in Abb. 3_11 ist der jeweilige Interaktionseffekt das Wichtige. Genauer wird die Vorhersagen-Zeichnung, wenn Haupteffekte als einfache Pfeile und ein Interaktionseffekt als Pfeil aus einem Multiplikationszeichen gemalt werden (Abb. 3_12a). Ob herausfordernde Arbeitsbedingungen generell die intrinsische Motivation erhöhen und dies bei hohem

Wachstumsbedürfnis besonders (so zeichnet es Abb. 3_12a oben-mitte), oder ob solche Arbeit *nur* bei solchen Persönlichkeiten wirkt (dann müsste der Haupteffekt-Pfeil gelöscht werden), ist sich die Theorie (Job-Caracteristics-Model nach Hackman & Oldham) gar nicht so sicher. Den Theorien, die Interaktionseffekte behaupten, ist der Interaktionseffekt meistens das einzig interessante. Daher fällt wenig ins Gewicht, dass der Zeichnung von Zajoncs Theorie sozialer Erleichterung (Abb. 2_2) ein Haupteffekt von Aufgabenschwierigkeit auf Leistungshöhe zugefügt werden müßte (um das wahrscheinliche Ergebnis in Abb. 3_9 oben-rechts abzubilden) – er ist trivial, gehört nicht zur speziellen Theorie, wirkt von ihr unabhängig.

Zusammenfassend werden *mehrfaktorielle Designs* mindestens dann nötig, wenn *Interaktionseffekte* nachzuweisen sind oder wenn *kontrolliert* werden soll, ob eine zuvor konfundierte *Variablen* (wie die Industrialisierung aus Sicht der Störche-Kinder-Hypothese, Kap. 3.1.5) vielleicht an Stelle des hypostasierten *Prädiktors* (Störche) auf die *AV* wirkt.

3.1.7 Designs für Veränderungshypothesen, Solomon-Vier-Gruppen-Design.

Kontrollvariablen zur Sicherung der *internen Validität* der Untersuchung werden besonders in **Versuchsdesigns für Veränderungshypothesen** wichtig, da hier „die Zeit“ (meist in den *Stufen* Vorher vs. Nachher, dazwischen ist dann bspw. ein Treatment) zur Prädiktorvariable wird und mit einer Zeitdauer verschiedenste andere Einflüsse *konfundiert* sein können.
Versuchsdesigns für Veränderungshypothesen sind in der angewandten Forschung häufig, bspw. wenn eine Intervention evaluiert werden soll: also in **Evaluationsstudien**. Die Intervention, die auf ihre Wirksamkeit hin bewertet werden soll, könnte eine organisationale Umstrukturierung, die Einführung von Gruppenarbeit, von Leistungszielvereinbarungen, eine Schulung oder auch eine neue Marketingmaßnahme u.v.a. sein. Treatment ist zu Intervention synonym. Die Hypothese lautet meist: ´Die Intervention wirkt´ (diese Formulierung klingt aber zweiseitig, vgl. Tab. 3_1), oder (einseitig, und damit besser): ´Durch die Intervention wird Y besser´ (mit Y = Zufriedenheit, Leistung, Kaufbereitschaft o.a).
Dass die *Sicherstellung der internen Validität der Untersuchung* zur Prüfung von Veränderungshypothesen schwierig ist, wird schon daraus deutlich, dass Abb. 3_13 und Tab. 3_11 fünf Versuchsdesigns zunehmend höherer interner Validität enthalten.

:

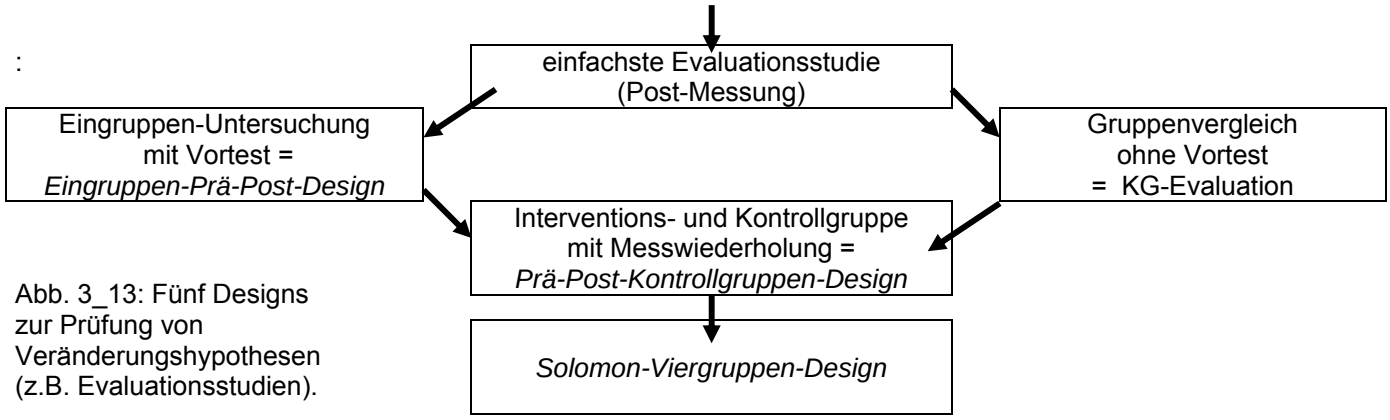


Abb. 3_13: Fünf Designs zur Prüfung von Veränderungshypothesen (z.B. Evaluationsstudien).

Der billigste und daher (leider) sehr verbreitete Untersuchungstyp für Evaluationen ist die **einfache Evaluationsstudie**: die Beteiligten werden nach der Maßnahme (nach dem Treatment, nach der Intervention) zu ihrer Zufriedenheit befragt (vgl. die Lehrevaluation durch Zufriedenheitsurteile am Ende des Semesters) oder es werden nach der Maßnahme Leistungstests (Klausur ☺) angewendet (Post-Messung). Dieser Versuchsplan einer *einfachen Post-Messung* ist aber ungeeignet, eine kausal gemeinte Wirkhypothese zu prüfen; hat also eine *geringe interne Validität*, da nicht ersichtlich ist, ob die AV (Zufriedenheit, Leistung etc) überhaupt auf die ´UV´ (den Einflussfaktor, das Treatment etc.) zurückgeht (Abb. 3_14a), denn *die zu überprüfende Maßnahme hat ja nicht variiert,* sondern war (zu T2 in der *EG*) immer ´da´.
Um eine Zufriedenheit von beispielsweise durchschnittlich 25,8 (oder eine mittlere Klausurnote von bspw. 2,5) auf das Treatment (z.B. die Lehrveranstaltung) zurückführen zu können (und damit das Ergebnis in Abb. 3_14a überhaupt zur Evaluation nutzen zu können: was bedeutet 25,8? ist das viel oder wenig??), ist entweder ein Vergleich mit einer Kontrollgruppe (Abb. 3_14b) oder der Vergleich mit dem Zustand vor dem Treatment (Abb. 14c) nötig.

Tab. 3_11: Fünf Versuchsdesign zur Prüfung von Veränderungshypothesen (z.B. *Evaluation* der Intervention I, eines Treatments, einer Maßnahme).

	Zeit →				
einfache Evaluationsstudie (Post-Messung)			I	→	T_2
Gruppenvergleich ohne Vortest: (Ex Post Facto-Plan, *Kontrollgruppen-Design*)			I	→	T_{2EG}
			-	→	T_{2KG}
Eingruppen-Untersuchung mit Vortest: (Eingruppen-*Prä-Post-Design*) *Messwiederholungsdesign*	T_1	→	I	→	T_2
Interventions- & Kontrollgruppe mit Messwieder-holung = *Prä-Post-Kontrollgruppen-Design*	T_{1EG}	→	I	→	T_{2EG}
	T_{1KG}	→	-	→	T_{2KG}
Solomon-Viergruppendesign:	T_{1EG1}	→	I	→	T_{2EG1}
	T_{1KG1}	→	-	→	T_{2KG1}
	-	→	I	→	T_{2EG2}
	-	→	-	→	T_{2KG2}

I = Intervention, Beeinflussung, Behandlung, Treatment, Maßnahme
EG = Experimental- oder Treatment- oder Interventionsgruppe, KG = Kontrollgruppe.
T = Messung zum Zeitpunkt t (T_1 = Prä-Messung = MZP1, T_2 =Post-Messung = MZP2)

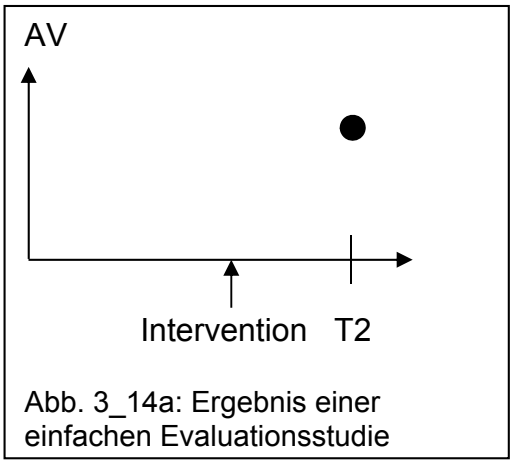


Abb. 3_14a: Ergebnis einer einfachen Evaluationsstudie

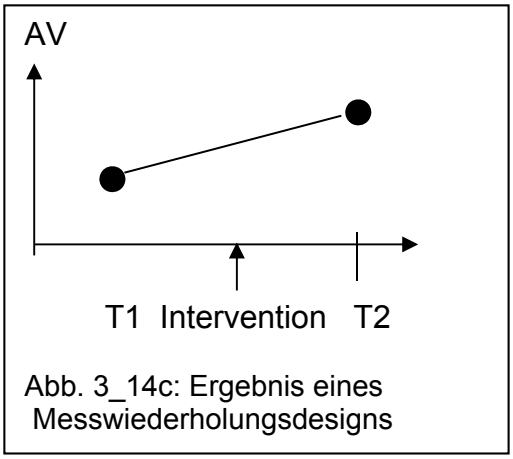


Abb. 3_14c: Ergebnis eines Messwiederholungsdesigns

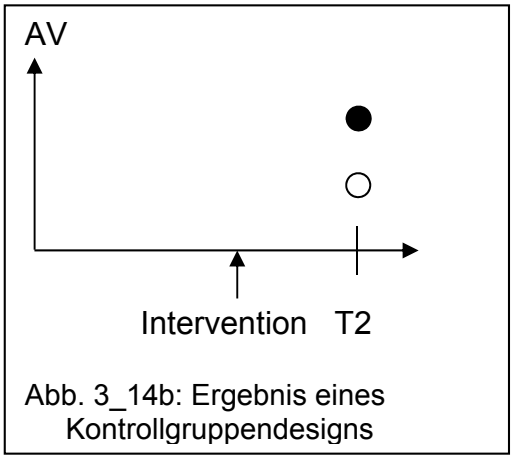


Abb. 3_14b: Ergebnis eines Kontrollgruppendesigns

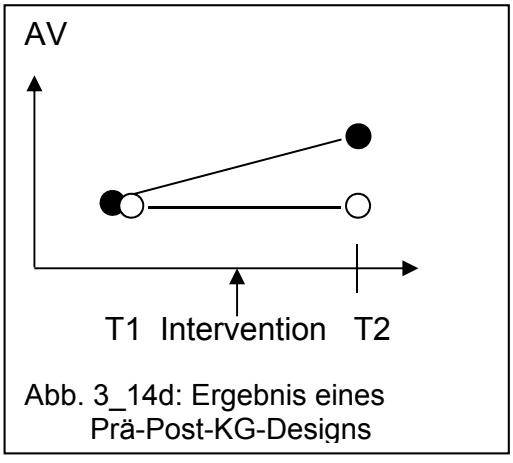


Abb. 3_14d: Ergebnis eines Prä-Post-KG-Designs

Die **Kontrollgruppe** für Abb. 14b kann entweder gesondert erhoben werden (in einer Abteilung, in der die Maßnahme noch nicht durchgeführt wurde, bei KonsumentInnen, die nicht umworben wurden etc.) oder Normwerten aus der Literatur können als Kontrollgruppe dienen, wenn der *Test* mit dem die AV erhoben wird, ein *standardisiertes* Instrument ist (standardisiert = mit bekannten Bevölkerungsnormwerten). Bspw. wird Intelligenz mit Intelligenztests gemessen, die so normiert sind, dass der Bevölkerungsdurchschnitt 100 und die Streuung in der Bevölkerung 10 beträgt (zu *Streuung* s. Kap. 4.1). Dann kann man die in der eigenen Trainingsstudie erhaltenen Intelligenzwerte mit diesem *Erwartungswert* von 100 vergleichen. Wenn Fragebögen zur *AV* Erhebung genutzt werden, für die keine Normwerte vorliegen, müssen Kontrollgruppen-Ergebnisse für Abb. 3_14b selbst erhoben werden.

Praxis-Tipp: Bevor man einen Fragebogen selbst entwickelt, lohnt es sich sehr, einen bereits verwendeten zu recherchieren, für den *Mittelwerte* einer oder mehrerer geeigneter Stichprobe zum Vergleichen vorliegen – man spart ggf. die eigene Erhebung einer Kontrollgruppe!

Eine Gefahr für die interne Validität eines Kontrollgruppendesigns (Abb. 3_14b) besteht dann noch durch alle Merkmale, auf denen sich die beiden Gruppen - ausgenommen die Intervention - noch unterscheiden: ist die *Vergleichbarkeit der Gruppen* gefährdet (sog. *Selektionseffekte;* Tab. 3_13), so sind diese Gruppenunterschiede mit der UV (Intervention ´ja/nein´) *konfundiert*.

Häufige **Selektionseffekte**:

- ein anderer Zeitpunkt der Erhebung der Kontrollgruppe (veraltete Normwerte!);
- eine besondere Auswahl (Selektion) der Treatmentgruppe (z.B.: nur die schlechteren Führungskräfte bekamen das Training etc.)
- **Probandenselbstselektion** (wer bspw. kaum Lernbereitschaft hat, geht in die Kontrollgruppe; wer unzufrieden ist, beteiligt sich nicht am Verbesserungsworkshop), auch selektive *Drop-Outs* (= Unzufriedene machen bei dem Post-MZP nicht mehr mit) nennt man Selbstselektion.

Ebenfalls validitätsgefährdend ist die Konfundierung der Gruppenzugehörigkeit (EG / KG) mit dem Wissen der Tln. um die Gruppenzugehörigkeit: es ändert die jeweils wirksame soziale Erwünschtheit (für Tln. an einer Maßnahme ist ggf. erwünscht, die Bedürftigkeit durch besonders viel Probleme etc. anzuzeigen, während die KG-Mitglieder äußern, auch keine Maßnahme zu brauchen, etc).

Die zweite Möglichkeit, das Ergebnis einer Post-Messung (Abb. 3_14a) auf die Intervention zurückzuführen zu können, besteht in dem **Vergleich mit Prä-Werten** aus derselben Untersuchungsgruppe (Abb. 3_14c). In der Evaluation bspw. einer Organisationsentwicklung muss also darauf geachtet werden, dass in die zur Organisations*diagnose* durchgeführten Befragung, also zu T1 (= MZP1 = Messzeitpunkt 1 = Prä-Messung), bereits solche Items aufgenommen werden, die sich nach der Maßnahme, zu MZP2, erneut beurteilen lassen. Natürlich müssen die selben Fragen und genau das gleiche Antwortformat verwendet werden, damit außer Zeit und Treatment zwischen T1 und T2 in Abb 3_14c nicht auch noch das Fragenformat variiert! Ein hypothesenkonformer Anstieg von Prä nach Post wird dann auf die Intervention zurückführbar.

Veränderungen, die mit der Zeit-Variable *konfundiert* sein und die *interne Validität der Untersuchung* eines **Prä-Post-Designs** (Abb. 3_14c) senken können, sind (vgl. Tab. 3_13):

- Zwischenzeitliche Umwelteinflüsse, die unabhängig vom Treatment auftraten (sog. ***zeitgeschichtliche Effekte***, z.B. eine Gehaltserhöhung, die parallel zur Treatmentphase im Unternehmen stattfand und anstelle des Treatments für die Steigerung der Arbeitszufriedenheit verantwortlich sein könnte; ein unvorhersehbarer Imageverlust des Unternehmens, der die PR-Maßnahme überstrahlt, eine zwischenzeitliche Finanzkrise, etc).
- Eine Entwicklung der Teilnehmenden - unabhängig von dem untersuchten Treatment (*Reifung,* allgemeine Lerneffekte auch ohne Treatment, umgekehrt z.B. Ermüdung, Motivationsverlust).
- Ein *Effekt der Erstmessung* allein - unabhängig vom eigentlichen Treatment - (die Teilnehmenden könnten durch die Einstellungsmessung selbst *sensibilisiert* worden sein, über das Thema nachzudenken und allein dadurch mehr Motivation aufbauen etc).
- *Probandenerwartungseffekte* können dazu führen, dass bspw. vor der Maßnahme besondere ´Bedürftigkeit´ geäußert wird (um die Maßnahme zu bekommen), und/oder nach ihr

besonderer Erfolge (um sie zu rechtfertigen, etc).

- *Drop-Outs* nennt man diejenigen Teilnehmenden, die während des Treatments (nach der Prä-Messung, vor der Post-Messung) aus dem Programm ausgestiegen sind. Die Post-Messung enthält dann evtl. nur noch die Teilstichprobe der Zufriedenen oder Erfolgreichen etc.

Um die *interne Validität* einer Evaluation mit Kontrollgruppe (Abb. 3_14b) oder auch die eines Prä-Post-Designs (Abb. 3_14c) zu erhöhen, sollte ein *zweifaktorieller Versuchsplan*, ein **Prä-Post-Design mit Kontrollgruppe** (Abb. 3_14d) realisiert werden. Die bisher zum Kontrollgruppendesign genannten und die zum Prä-Post-Design genannten Fehlereinflüsse werden damit nicht verhindert, aber die davon wichtigsten können *kontrolliert* (= wenn sie vorhanden sind, aufgedeckt) werden! Der hypothesenkonforme Anstieg nur der Werte der Treatmentgruppe wird im *Prä-Post-Design mit Kontrollgruppe* auf die Intervention zurückgeführt. Sind die Ergebnisse der Prä-Messung für beide Gruppen gleich (wie im idealen Ergebnis von Abb. 3_14d), können die Störvariablen *Probandenselektion, Probandenselbstselektion* und *selektive Probandenerwartung* ausgeschlossen werden (Tab. 3_13). Um dies sicherzustellen, wäre ideal, die Teilnehmenden per Zufall (=*randomisiert*!) auf die beiden Gruppen zuweisen zu dürfen, also ein *Feldexperiment* durchzuführen (Kap. 3.1.5).

In betrieblichen Studien ist das Experimentieren zwar schwierig, aber bei guter Planung dennoch möglich. Vorbildlich ist hier das Feldexperiment von Atwater et al. (2000): Führungskräfte nahmen an einem Upward-Feedback teil. Ihnen wurde mitgeteilt, dass nur die Hälfte von ihnen das ausführliche Feedback der von den Unterstellten erhobenen Bewertungen erhält (die andere Hälfte im nächsten Jahr). Die Zuweisung zu den beiden Gruppen (UV1: mit/ohne Feedback nach t1) wurde - erst nach Durchführung der Prä-Messung - per Los bestimmt (wegen der *randomisierten* Gruppenzuweisung ist die Studie ein *Experiment*!). In der Folgeerhebung nach 10 Monaten zeichneten sich die Führungskräfte mit Feedback durch weniger Selbstüberschätzung und etwas höhere Mitarbeiterurteile aus, die Selbst- und Fremdbilder der Führungskräfte ohne Feedback blieben praktisch unverändert (Abb. 3_15; ob man die Mittelwerte durch Balken wie in Abb. 3_15 oder durch Punkte und Striche wie in Abb. 3_14 visualisiert, ist nicht so wichtig. Wenn auf der X-Achse die Zeit variiert, sind eigentlich Striche angemessener, wenn eine beliebige *Nominalvariable* (s. Kap. 3.2.1) variiert, Balken. Übung: Zeichnen Sie für jede AV aus Abb. 3_15 gesondert ein Ergebnisbild im Stil von Abb. 3_14).

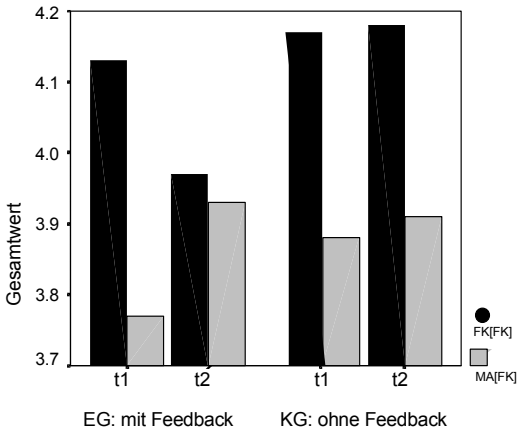


Abb. 3_15: Ein **Feldexperiment** zur Evaluation der Wirkung von Upward-Feedback (Atwater et al. 2000). Das Design ist zweifaktoriell (ein Prä-Post-KG-Design), wobei UV1 (mit/ohne Feedback an Führungskräfte) experimentell realisiert wurde, und bivariat (Selbstbild der FK & Fremdbild von den MitarbeiterInnenn (MA) sind die beiden AV).

Sind die Ergebnisse der Prä-Post-Messung in der Kontrollgruppe konstant (wie im idealen Ergebnis von Abb. 3_14d), können *zeitgeschichtliche Effekte*, *Reifung* und *Erstmessungseinfluß* (als Gegen-Erklärungen für den hypothesenkonformen Anstieg der Werte in der Treatmentgruppe) ausgeschlossen werden (Tab. 3_13, deshalb ist die *interne Validität* des Prä-Post-Kontrollgruppen-Designs höher als die des einfachen Prä-Post-Designs).

Findet sich ein Gruppenunterschied in den Prä-Werten (im Experiment von Atwater et al. 2000 scheinen sich die Führungskräfte der EG zu T1 tendenziell schon etwas weniger überschätzt zu haben als die der KG, s. Abb. 3_15, andererseits aber erhielten sie zu Beginn etwas weniger Wertschätzung von ihren Mitarbeitenden als die Führungskräfte der KG; Atwater et al. 2000 hatten die Gruppenzugehörigkeit aber randomisiert zugewiesen) und / oder fand eine Werteveränderung auch in der Kontrollgruppe statt (vgl. den ganz geringfügigen Anstieg der Mitarbeitendenbewertungen auch in der KG in Abb. 3_15), sind die Befunde nicht mehr so sicher zugunsten der Treatmentwirkung zu interpretieren. Die hypothesenkonforme Änderung in der Treatmentgruppe muss dann wenigsten stärker sein als die in der Kontrollgruppe (ist für die Fremdbilder in Abb. 3_15 gegeben). Das *zweifaktorielle Design* (Abb. 3_14c) wird dann meist

mit einer *Varianzanalyse (s. Kap. 4.2)* ausgewertet, um den ***Interaktionseffekt*** (s. Kap. 3.1.6) zwischen Treatment (mit/ohne) und Messzeitpunkt (T1/T2) abzusichern.

In der angewandten Forschung wird eine Evaluationsstudie in einem Prä-Post-Kontrollgruppen-Design (Abb. 3_14d) häufig durchgeführt, um anschließend das Treatment als „wissenschaftlich geprüft" vermarkten zu können (s. Verwendungszusammenhang, Kap. 1.2) – ohne bei Folgeanwendungen des Treatments die Evaluation dauernd weiterführen zu wollen. Hierfür ist auch ein hypothesenkonformes Ergebnis wie das in Abb. 3_14d keine völlig ausreichende Grundlage; die *interne Validität* des *zweifaktoriellen Designs* für den *Begründungszusammenhang* der Aussage „das Treatment wirkt (auch ohne die Messungen) " ist noch immer nicht optimal, denn der Werteanstieg in der Treatmentgruppe von Abb. 3_14d wurde ja nach einer Realisierung von Treatment und Prä-Messung erhalten. Vielleicht ist der Anstieg nur nach einem gemeinsamen Auftreten von Prä-Messung und Treatment zu erzielen? Vielleicht wirkt das Treatment allein nicht! Vorstellbar ist eine solche *Interaktionen zwischen Messung und Treatment (Tab. 3_13)* gerade bei der Evaluation von Organisationsentwicklungsmaßnahmen: die erste Mitarbeitendenbefragung, die bspw. der Diagnose dient (und als Prä-Messung im Prä-Post-Design auftaucht), kann die Mitarbeitenden sensibilisieren (sie unterhalten sich über die abgefragten Bereiche, denken darüber nach, achten im Alltag mehr auf die im Fragebogen genannte Aspekte). Die Befragung selbst erreicht schon eine Einbindung der MitarbeiterInnen, das *Auftauen* bisherige Gewohnheiten (Borg, 1995; Lewin nannte die drei Phasen der Veränderung: ´unfreezing-change-refreezing´; Sensibilisierung als notwendige Voraussetzung einer Treatmentwirkung: ohne Auftauen kein Change). Das spätere Treatment wirkt dann vielleicht nur bei zuvor Sensibilisierten!

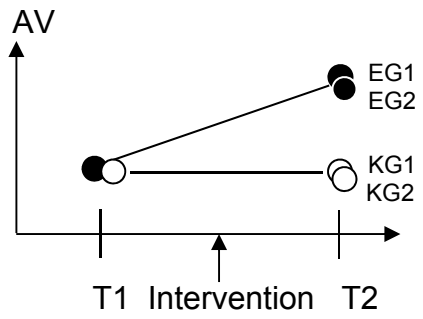


Abb. 3_16: Ideales Ergebnis im Solomon-Viergruppen-Design

Um nachzuweisen, dass das Treatment auch ohne Prä-Messung seine Wirkung entfaltet, ist das **Solomon-Viergruppen-Design** (Abb. 3_16 & Tab. 3_11 unten) notwendig. Das zum Annehmen der Treatmentwirkungs-Hypothese ideale Ergebnis zeigt Abb. 3_16: beide Versuchsgruppen mit Intervention (EG1 & EG2) zeigen hohe Postwerte, beide Gruppen ohne Treatment (KG1 & KG2) zeigen unverändert niedrige Postwerte.

Wenn dann noch durch *randomisierte Zuweisung* der Teilnehmenden zu den vier (damit *experimentellen*) Gruppen sichergestellt werden kann, dass sich die Prä-Werte von EG1 und KG1 nicht unterscheiden (wie in Abb. 3_16), dann genügt es auch, allein die Postmessung in einem *2x2 - Design* auszuwerten (Tab. 3_12), in dem dann ein *Haupteffekt* für die UV ´Treatment-Wirkung´ erwartet wird, aber kein *Haupteffekt* für die UV ´Prämessung (mit/ohne)´ und auch keine *Interaktion* der beiden UV.

Tab. 3_12: Zweifaktorielles Design der Post-Werte zur Prüfung der Treatmentwirkung in einem *experimentellen* Solomon-Viergruppen-Design

AV: Zufriedenheit, Leistung zu **T2**	ohne Treatment		mit Treatment
Mit Prä-Messung	**KG1**	<	**EG1**
ohne Prä-Messung	**KG2**	<	**EG2**

Wenn das Experimentieren (*randomisierte Zuweisung* der Teilnehmenden auf die Bedingungen) nicht geht (ethische Gründe, Betriebsratsvereinbarung u.a.), dann prüfe man zunächst, ob sich die Prä-Messungen in EG1 und KG1 unterscheiden: wenn die EG1 zu T1 nicht besser oder schlechter war als die KG1, dann können *Probandenselektionseffekte* ausgeschlossen und allein die Post-Werte nach Tab. 3_12 mit der Hypothese: „Haupteffekt fürs Treatment, kein Haupteffekt für und keine Interaktion mit der Prä-Messung" ausgewertet werden. Eine Ergebnisabbildung wie in Abb. 3_16 muss es auf jeden Fall geben.

In betrieblichen Evaluationsstudien, die leider oft weniger geplant durchgeführt werden als Forschungsstudien, kann das Kontrollprinzip des Solomon-Vier-Gruppen-Designs dennoch nützlich sein: nämlich wenn bspw. in einem laufenden Prozess (z.B. kontinuierliche Führungs-

kräftebeurteilung, zeitversetzte Personalschulungen, jährliche Zufriedenheitsumfragen, Kampagnentracking etc.) laufend Evaluationsdaten anfallen (die Personalabteilung sammelt Beurteilungsbögen etc.), die Betroffenen aber zu unterschiedlichen Zeiten in das Programm einsteigen. Die Teilnehmenden werden nach erfolgtem Teilnahme-Muster in Untergruppen sortiert (z.B. Personen, die nur in MZP1 teilnahmen, solche, die in MZP1 & MZP2, solche die in MZP1 – MZP3, solche, die nur MZP2, solche, die MZP2 & MZP3 usw.). Für jede dieser Teilstichproben werden Punkte (Mittelwerte) und Linien (wo mehrere Messungen vorliegen) analog zu Abb. 3_16 aufgezeichnet: Fortschritte derjenigen, die vor ihrer jew. Post-Messung ein Treatment erhielten, sowie potentielle drop out – Gründe (z.B. Unzufriedene scheiden aus), lassen sich erkennen.

Tabelle 3_13: Zusammenfassung von Fehlerquellen, die die interne Validität der Untersuchung reduzieren, und Möglichkeiten ihrer Eingrenzung

Name	Vorkommen	Beschreibung	Kontrollmöglichkeiten
1. Probanden-Erwartungs-Effekt	oft möglich (außer bei *non-reaktiven Verfahren*)	Reaktivität der Vp, sich pro oder kontra der von ihr vermuteten sozialen Erwünschtheit zu verhalten (s.a. HAWTHORNE-Effekt).	Nichtreaktive Verfahren, Gestaltung der Instruktion, Begleituntersuchung Bogus-Pipeline – Verfahren
2. VL-Erwartungs-Effekt	bei Interviews & qualitativen Verfahren, offenen Antworten, ..	Versuchsleiter/in behandelt die Versuchspersonen selektiv oder bewertet selektiv zugunsten der H_1-Hypothese (geringe Objektivität) (bspw.. Rosenthal-Effekt, evtl. Self-fulfilling Prophecy).	Bedingungen standardisieren: geschlossene Fragen, Doppelt-Blind-Versuche / Doppelt-Blind-Kodierungen.
3. Probanden-selektionseffekt u. Probanden-selbst-selektions-Effekt	möglich, wenn Vp. nicht zu den Versuchsbedingungen randomisiert zugewiesen wurden, sondern freiwillig teilnehmen und zwischenzeitlich aussteigen können	Der Unterschied zwischen Treatment- und Kontrollgruppe kann auf systematische Unterschiede der Teilnehmenden zurückgeführt werden (bspw. wenn im Betrieb bestimmte Abtl. für die Maßnahme vorausgewählt werden). Selbstselektion führt zu einem systematischen Effekt, wenn die Motivation für das Ergebnis wichtig ist.	*Kontrollgruppendesign.* *Parallelisierung* der Tn.; randomisierte Zuweisung (Los!, dadurch wird der Gruppenvergleich zum *Experiment*!). Muss *Selbstselektion* in Kauf genommen werden, verschieden werben!
u. Selektive Probanden-verringerung= Drop-out-Effekte		Der Ausfall von TIn. in der Zweit- im Vergleich zur Erstmessung kann für bessere Post-Werte verantwortlich sein.	nur Daten solcher Probanden auswerten, die in Prä- und Postmessung verfügbar waren, die Prä-Werte der Drop-Outs mit denen der anderen vergleichen.
4. Regression zur Mitte	Messwiederholungs-Designs, besonders bei Extremgruppen auswahl	Werden aus T1 Extremgruppen ausgewählt, resultiert bereits artifiziell eine Konvergenz beider Gruppen zu T2, da ein Teil der Extremwerte zu T1 nur durch Messfehler (geringe Reliabilität der AV) verursacht war, zu T2 diese Leute ihre mittleren Werte zu recht wieder erhalten. Bekommen bspw. nur schlechte Kräfte die Maßnahme, ist für sie ein Werteanstieg (´zur Mitte´) auch ohne Treatment zu erwarten.	In der Praxis wird die zu erwartende artifizielle Regression zur Mitte aus der Reliabilität des Messinstruments errechnet (spezielle Formel nötig), die gemessene Veränderung muss dann deutlich größer als die artifiziell zu erwartende sein. Veränderungsstudien mit Extremgruppen vermeiden.
5. Zeitgeschichtliche Effekte	Messwiederholungs-Designs	Simultan zur Maßnahme können andere Ereignisse gewirkt haben (Gehaltserhöhung, Krise, Skandal u.v.a).	Vergleich mit einer Kontrollgruppe, die diesen Einflüssen auch unterlag (Abb 3_14c)
u. Reifungsprozesse	"	interne Entwicklungen (Lernen, Ermüdung etc) die unabhängig vom Treatment auftritt.	"
u. Testeinfluß	"	Wenn die Erstmessung allein eine Sensibilisierung für das Thema auslöst, die allein den T2- Effekt verursacht hat.	"
6. Interaktion zwischen Messung und Treatment	Messwiederholungs-Designs	Der erzielte Effekt tritt nur bei Kombination der Treatments mit dem sensibilisierenden Vortest (Prämessung) auf. Dann ist er nicht Wirkung „des Treatments".	Solomon- Viergruppen- Design (in Abb. 3_16 keine Interaktion zw. Messung & Treatment).

3.2 Operationalisierung der (abhängigen) Variablen

Ist das *Versuchsdesign* festgelegt (Kap. 3.1), so wurde über die *Operationalisierung* der *unabhängigen Variablen* weitgehend entschieden (vgl. z.B. Tab. 3_8), nun müssen noch die *abhängigen Variablen* ***operationalisiert***, also messbar gemacht werden. Wurde entschieden, die Hypothese(n) in einer *Korrelationsstudie* zu prüfen (s. Kap. 3.1), so sind alle Variablen noch zu messen. Bei der Überlegung, wie eine bestimmte Variable zu messen ist, sollte das *Skalenniveau der Messung* (kann ein hohes Skalenniveau erreicht werden? Kap. 3.2.1) und sollen die (anderen) *Gütekriterien der Messung* (Kap. 3.2.2) optimiert werden.

3.2.1 Skalenniveau der Messung

Es werden vier Stufen des *Skalenniveaus* unterschieden: *Nominalskala, Ordinalskala, Intervallskala, Verhältnisskala.* Die Nominalskala ist die ´niedrigste´, die Verhältnisskala die ´höchste´ (Tab. 3_14 links). Je *höher* das Skalenniveau, desto höher der Präzisionsgrad der Messung. Bei höherem Skalenniveau sind mehr (und ´bessere´) statistische Verfahren erlaubt, (s. Kap. 4: Auswertungsverfahren, Signifikanztests); in der Praxis optimiert man das Skalenniveau (der AV) auf Intervallskalenniveau, um übliche Statistiken rechnen und mehrfaktorielle Designs auswerten zu dürfen☺.

Tab. 3_14: Skalenniveaus

Bezeichnung	Beispiele	Eigenschaften, Voraussetzungen		erlaubte Berechnungen..
Verhältnisskala (*ratio-scale*)	Zeit, z.B. Alter, Reaktionszeit... Gewicht, Entfernung, Temperatur in Kelvin Prozentzahlen	natürlicher Nullpunkt (0 bedeutet wirklich: ´Nichts´, ´keins´, ´Null´)	x1= b • x2	Bruchrechnung, Bestimmung von Verhältnissen (´z.B. doppelt so groß wie...´) und alles, was bei Intervallskalen erlaubt ist.
Intervallskala	psychometrische Ratingskalen, IQ-Werte etc	Gleichheit von Unterschieden. Bspw. ist der Abstand von 1 nach 2 genauso groß wie der von 4 nach 5.	x1-x2=x3-x4	Berechnung von Streuung, Varianz, normaler Korrelation r, u..a. „parametrische Verfahren“, mehrfaktorielle Designs & alles, was bei Ordinalskalen erlaubt ist.
Ordinalskala (Rangskala)	Rangreihen, Präferenzen, höchster Bildungsabschluss, ungleichabständige Antwortskalen wie: „pro Jahr, pro Monat, pro Woche, …“	Nur „größer als“, „kleiner als“ kann festgestellt werden, aber der Abstand zwischen bspw. einer 1 und einer 2 kann ganz anders als der zwischen 4 und 5 sein	x1>x2>x3	Aussagen wie „größer als“, Berechnung von Mittelwert, Median, Rangkorrelation Rho und anderen sog. „nonparametrischen Verfahren“ & alles, was bei Nominalskalen erlaubt ist.
Nominalskala	Namen, Geschlecht, kategoriale Qualitäten wie Haarfarbe etc..	(nur) Gleich oder Ungleich bestimmbar	x1≠x2≠x3	Häufigkeiten (⇒Prozent, Chi^2-Verfahren), Modalwert (=häufigste Kategorie)

Ein Mittelwert über bspw. kategorisierte Haarfarben zu rechnen, ergibt keinen Sinn, auch wenn sie im Computer mit Zahlen kodiert wurden (z.B. 1=schwarz, 2=blond, 3=braun, 4=rot). Solche Zahlen sind bei einer ***Nominal*skala** eben nur *Namen (Nomen)* für Kategorien (statt 1,2,3,4 hätte man auch a,b,c,d oder gleich s,bl,br,r eingeben können, dann wäre klar, dass damit nicht viel zu ´rechen´ ist: nominal heißt ´nur dem Namen nach´). Um zu wissen, welche Berechnungen mit den im Computer dann aber vorgefundenen Zahlen*kodes* erlaubt sind, muss eben das *Skalenniveau* der Variable bekannt sein. Hat man eine *nur nominal skalierte* AV, so kann man die relative Häufigkeit in jeder Kategorie dieser AV berechnen und eine schöne Balken- oder Tortengraphik zeigen, über die Häufigkeiten lässt sich auch ein Test rechen (Chi-Quadrat-Test, s. Kap. 4.2), aber sonst nichts (für Fortgeschrittene: logistische Regression).
Besser wird es (mit der Freiheit, zwischen verschiedenen statistische Verfahren zu wählen), wenn Nominal-Variablen in Binär-Variablen zerlegt werden. Eine natürliche **Binär-Variable** ist bspw. das biologische (gender-unreflektierte) Geschlecht (w/m, wird oft mit 1/2 oder 0/1 im

Computer kodiert). Für Binär-Variablen, die dann günstig als 0 / 1 kodiert werden sollten, lässt sich der Mittelwert insofern berechnen, als er die relative Häufigkeit eines Attributs verrät: wurden 20 Männer und 30 Frauen untersucht, die Frauen mit 1, die Männer mit 0 kodiert, resultiert ein Mittelwert von 0.60, der angibt, dass 60% Frauen teilgenommen haben (Binärvariablen sind also etwas ganz besonderes, da sie aussehen wie Nominalvariablen, aber mit ihnen – wenn sie 0/1 kodiert sind - Rechnungen erlaubt sind, die tlw. sogar Verhältnisskalenniveau voraussetzen).

Exkurs: In einem typischen *Experiment* (vgl. Kap. 3.1) ist jede ***UV*** auf Nominalniveau operationalisiert (es gibt bspw. eine *Experimentalgruppe* mit Treatment und eine *Kontrollgruppe* ohne Treatment in Abb. 3_14 - 3_15, man kann die UV als ´0´/´1´ kodieren, sogenannte Dummy-Kodierung). Bei *experimentellen UV* gilt ein Nominalskalenniveau nicht als niedrig, denn möglichst hohe Berechnungsverfahren anwenden will man ja über die Daten der *A*V (s. Kap.4.: wichtigstes Merkmal zur Auswahl des statistischen Verfahrens ist das Skalenniveau der AV).

Für die abhängige(n) Variable(n) aber (und gern für Kontrollvariablen, Mediatoren, Moderatoren usw., für Prädiktoren, die nicht experimentell manipuliert werden), sollte ein möglichst hohes Skalenniveau angestrebt werden (möglichst *Intervallniveau*). Die Variablen der Zusammenhangshypothese „Je schwieriger die Aufgabe, desto schlechter die Leistung“ (X= Aufgabenschwierigkeit, Y= Leistung) könnten beide ´**nur ordinal**´ operationalisiert werden: X = von den Prüfern als ´leicht´, ´mittel´ und ´schwer´ bezeichnete Prüfungsfächer, Y = Prädikat / bestanden / durchgefallen. Dann muss man korrekt bleiben und eine **Rangkorrelation** ausrechnen, also eine Formel benutzen, die für *ordinale Daten* zugelassen ist (z.B. Spearmans Rho oder Kendalls Tau, s. Kap. 4. bzw. im Statistik-Buch oder web, oder der Statistik-Software unter Korrelation für *Rangdaten*, *nonparametrische Korrelation* oder *Rangkorrelation* nachschlagen). Mit ordinalen Prädiktoren lassen sich aber keine Interaktionseffekte und keine Mediationsanalysen rechnen!

Bei der Messung von AV wird angestrebt, **Intervallskalenniveau** zu erreichen (Verhältnisskalenniveau ist für die gebräuchlichen sogn. *parametrischen* Auswertungsverfahren nicht notwendig). Mit AV auf Intervallskalenniveau lassen sich die ´üblichen statistischen Verfahren berechnen (z.B. t-Test, Varianzanalyse und die ´normale´ Korrelation r = Pearson-r, die für Regressionen, Faktorenanalysen usw. nötig ist).

Bei gebräuchlichen psychologischen Messverfahren (z.B. Einstellungsmessung und anderen Fragebogen aus Typ C im Spektrum von Abb. 3_1) ist ohne aufwendige Skalierungsuntersuchungen häufig nicht sicher, ob auf Intervallskalenniveau oder doch nur auf Ordinalskalenniveau gemessen wurde. In der Praxis ist es aber üblich, *Ratingskalen* (in Fragebogen verwendetes *geschlossenes* Antwortformat mit mehreren Antwortmöglichkeiten z.B. wachsender Zustimmung, Abb. 3_17) *als intervallskaliert zu behandeln* (wobei diese Formulierung („als ... behandelt“) im Methodikteil des Berichts, Kap. 3.2, durchaus auftauchen sollte, man demonstriere Selbstkritik).

Praxis-Tipp: Damit eine **Ratingskala** möglichst als *Intervallskala* behandelt (also mit den besseren ´parametrischen Verfahren´ ausgewertet) werden darf, sollte *nicht* jeder Skalenpunkt beschriftet sein (z.B. 0= nie, 1=selten, 2=manchmal, 3=häufig, 4= sehr häufig) - denn die Worte könnten von der Vp. so interpretiert werden, als lägen verschiedene Abstände dazwischen. Sondern es sollten nur die (gleichanständigen) Zahlen vorgeben und nur die ´Ränder´ beschriften werden (z.B. „nie 0 1 2 3 4 häufig“ oder „dagegen −2 −1 0 +1 +2 dafür“). Dann nämlich versucht die Vp selbst intuitiv die Bedeutung der Antwortabstufungen so zu empfinden, dass die Abstände zwischen zwei numerisch benachbarten Zahlen gleich groß sind (zwischen 1 und 2 so groß wie zwischen 2 und 3; Voraussetzung der Intervallskala, s. Tab. 3_14).

Abb. 3_17: Intervallskalen dürfen nicht wie Verhältnisskalen interpretiert werden

Dann muss nur noch beachtet werden, Ergebnisse einer *Intervallskala* nicht im Sinne von Verhältnisskala oder Bruchrechnung zu interpretieren: Hat bspw. die EG bei Kodierung einer fünfstufigen *Ratingskala* von 0 bis 4 einen Durchschnittswert von 3,0 erreicht (Mittelwert, per Konvention als ´x-quer´ gekennzeichnet, s. in Abb. 3_17), die KG nur einen von 1,5 (Abb. 3_17 links), so darf man nicht sagen, die EG stimme dem Item „doppelt so stark zu“, denn der Nullpunkt ist nicht absolut, die Ratingskala hätte ebenso gut von 1-5 kodiert werden können (Abb. 3_17 rechts), das gleiche Ergebnis lässt die (eben ja falsche) *Verhältnisaussage* nicht zu.

3.2.2 Gütekriterien der Messung

Neben dem *Skalenniveau* (ist für die AV *Intervallskalenniveau* erreicht?) bestimmen drei *Gütekriterien der Messung* die Güte der AV: die *Objektivität, die Reliabilität und die Validität der Messung* (für *Validität der Messung* unbedingt immer das „der Messung“ dazusagen/schreiben, damit sie nicht mit der „Validität der Untersuchung“ verwechselt wird, s. bspw. Kap. 3.1.3 *interne Validität der Untersuchung*!).

Objektiv wird eine AV gemessen, wenn bei der Auswertung kein Interpretationsspielraum mehr bei der auswertenden Person liegt. Muss aber bspw. Verhalten *beobachtet* werden, müssen PrüferInnen durch Zuhören die Leistung von Prüflingen feststellen und dann mit einer Note die wahrgenommene Leistung ´messen´, oder müssen in einem Fragebogen mit ***offenen Fragen*** (z.B.: „Was mögen Sie an der Statistik? Schreiben Sie bis zu drei Stichwörter“☺) die Antworten einer Inhaltsanalyse unterzogen werden (sogn. ***qualitative Verfahren***, siehe Untersuchungsarten A und wieder F im Spektrum von Abb. 3_1), dann gibt es Interpretationsspielraum! Die Objektivität der Messung ist geringer als wenn das Eintippen der Kreuze als Antworten auf ***geschlossene Fragen*** (z.B. einer Ratingskala, Abb. 3_17) den Versuchspersonen übertragen wird oder gar bei psychophysiologischen Messungen über technische Geräte (Typ D in Abb. 3_1), bei deren Auswertung zunächst keine Interpretation nötig scheint.

Objektivität: Maß der intersubjektiven Übereinstimmung mehrerer Versuchsleitenden (Auswertenden) bei der Datengewinnung, Datenauswertung und Dateninterpretation. Je *standardisierter*, desto *objektiver*.

Objektivitätskontrolle: Urteilsübereinstimmung berechnen.

Objektivitätsoptimierung: Testsituation standardisieren, z.B. nur geschlossene Fragen. Auswertung standardisieren und trainieren. Urteile von mehreren Personen mitteln.

Ist aber Interpretation der Versuchsleitung zur Messung einer Variablen nötig (qualitative Forschung), so sollte die *Objektivität* über ein Maß der **Urteilerübereinstimmung** (= Inter-Rater-Übereinstimmung**)** kontrolliert werden. Dazu müssen mehrere, im Prinzip gleichgute UrteilerInnen die gleichen (mehrere! möglichst verschiedene!) Gegenstände unabhängig voneinander (also jede/r allein) beurteilen, z.B. die offenen Antworten unabhängig kodieren. Kommen die Urteilenden meistens zum übereinstimmenden Ergebnis (ist das berechnete Maß der Urteilerübereinstimmung hoch), dann gelingt die Auswertung *ausreichend objektiv* (=intersubjektiv). Zur Berechnung der Urteilerübereinstimmung gibt es - *je nach Skalenniveau* der Variable (s. Kap. 3.2.1) verschiedene Maße (mit zugehörigen Formeln. Urteilerübereinstimmungsmaß bei Nominaldaten: *kappa* (unter diesem Stichwort in Statistikbuch suchen), bei Ordinaldaten: Rangkorrelation (z.B. Kendalls W), bei Intervalldatendaten: Intraklassenkorrelation u.a.).

Muss eine AV-Messung über Interpretation erfolgen und erweist sich die *Urteilerübereinstimmung* in einer dazu durchgeführten Übereinstimmungsstudie als gering, kann zu dem aufwendigen Verfahren gegriffen werden, immer mehrere KodiererInnen parallel arbeiten zu lassen und deren Kodes zu ´mitteln´ (bzw. bei Nominalskalenniveau den Modalwert verwenden). Im Assessment-Center bspw. ist es daher üblich, immer mehrere BeobachterInnen einzusetzen. Noch besser aber ist, die Urteilerübereinstimmung zu erhöhen, indem die Kodierenden trainiert und die Kodiervorschriften (die erst mal aufzuschreiben sind!) verbessert (vereinfacht und präzisiert) werden. Dann wieder eine Studie zur Kontrolle der Objektivität durchführen, Urteilerübereinstimmung berechnen...- ist sie nun besser geworden?

Es gibt auch Objektivitätsprobleme, die sich durch Urteilerübereinstimmungsstudien nicht

aufdecken lassen würden, nämlich wenn alle Urteilenden dem gleichen ´Erwartungseffekt´ unterliegen: der **Versuchleitungs-Erwartungseffekt** (Tab. 3_13) tritt bspw. auf, wenn offene Antworten kodiert werden müssen und den VL die Hypothese bekannt (und plausibel) ist (bspw. dass nach dem Training die Zufriedenheit steigt): alle erwarten dann bessere Antworten, kodieren die Antworttexte also vielleicht milder gestimmt, achten stärker auf positives etc. Man kann sich schwer zur Objektivität zwingen, Menschen sind unzuverlässig. Hier lässt sich zum Verfahren der **Doppelt-Blind-Studie** greifen: ´einfach blind´ ist eine Untersuchung schon, wenn die Versuchsperson nicht weiß, in welcher Versuchsbedingung sie ist. Das ist günstig, damit sie nicht per Versuchspersonen-Erwartungseffekt, s. Tab. 3_13, nur aus Gründen der sozialen Erwünschtheit oder Reaktanz ihre Ergebnisse lenkt. *Doppelt blind* wird die Studie, wenn auch die versuchsdurchführenden oder/und kodierenden Personen nicht wissen, ob sie gerade einen Text einer Antwort vor dem Training oder nach dem Training kodieren (die Texte wurden bspw. transkribiert, mit künstlichen Codes versehen und gemischt, bevor die Kodierenden sie bekamen). Ist man blind gegenüber den Versuchsbedingungen, kann man die Ergebnisse nicht für oder gegen die Hypothese verzerren - bei Kodierung offener Antworten das Mittel der Wahl!

Die *Objektivität* der Messung ist Voraussetzung für die (ausreichende) *Reliabilität* der Messung: eine Messung, die nicht *objektiv* ist, kann auch nicht *reliabel* sein.

Reliabilität: Maß der *Zuverlässigkeit*, mit der das Erhebungsinstrument misst. Die Reliabilität des Instruments ist hoch, wenn der *Messfehler* klein ist.
Reliabilitätskontrolle: Mehrfachmessung durchführen und Reliabilität berechnen:
z.B. Retest-Reliabilität oder ein Maß der interne Konsistenz (meist Cronbachs Alpha).
Reliabilitätsoptimierung: Mehrere Messungen (Situationen, Items..) mitteln: Fehlerausgleich!

Die Reliabilität eines Messinstruments gibt an, ob das gleiche Merkmal immer wieder das gleiche Messergebnis erhielte (und unterschiedliche Merkmalsausprägungen unterschiedliche Messergebnisse natürlich). Wird bspw. die Breite eines Tisches mit einem Gummiband (auf das cm aufgemalt sind) gemessen, wird man vermutlich trotz gleichen Tisches immer unterschiedliche Ergebnisse erhalten: das Gummi-Band ist **unreliabel**; ein Zentimetermaß aus Metall ist sicherlich **reliabler**.
Um die Reliabilität eines Messwerts berechnen zu können, gibt es verschiedene Methoden. Für die **ReTest-Reliabilität** müssen mehrere Gegenstände, die sich in der Zeit nicht verändern (z.B. Tische unterschiedlicher Größe), oder mehrere Personen auf einem für stabil gehaltenen Personmerkmal (Intelligenz und Extraversion werden für relativ stabil gehalten, Einstellungen aber sind nicht besonders stabil, Stimmungen gar nicht) zweimal gemessen werden: *korrelieren* die Messergebnisse beider Zeitpunkte hoch (z.B. r_{tt} = +.90, der Index *tt* drückt die beiden Meßzeitpunkte aus, r_{tt} steht also immer für eine Meßwiederholungs-Korrelation), scheint das Instrument das Merkmal *reliabel* erfassen zu können. In der wirtschaftspsychologischen Praxis sind Retest-Reliabilitätsuntersuchungen selten (erstens teuer, zweitens taugen sie nur bei stabilen Merkmalen, das Psychische wandelt sich aber oft. Für Fähigkeitstest z.B. Intelligenztests aber sollten und werden Retest-Reliabilitätswerte erhoben).
Die Reliabilität eines an sich unreliablen Messinstruments kann jedoch recht einfach erhöht werden: misst man den Tisch mit demselben unreliablen Gummiband öfter, so *streuen* die Messwerte *unsystematisch* um den *´wahren Wert´* der Tischbreite (Abb. 3_18). Der *wahre Wert* kann (näherungsweise) ermittelt werden, indem der **Durchschnitt aus mehrere unreliablen Messergebnissen** gebildet wird: die Messfehler gleichen sich aus! - Wenn die Messfehler unsystematisch und normalverteilt sind (Abb. 3_18, Normalverteilung = Gauß-Kurve, von Gauß entdeckt wegen Messfehlern beim Sterne-Messen). In der Einstellungsmessung bspw. macht man sich das Prinzip des Messfehlerausgleichs zunutze, indem der Inhalt einer Variable (z.B. die AV Führungserfolg für Abb. 3_15)

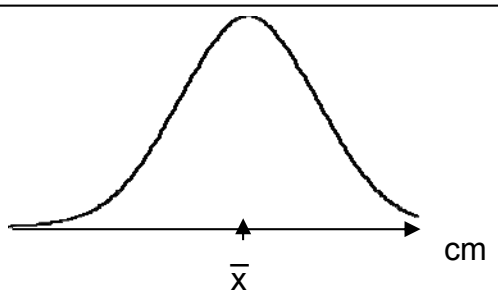


Abb. 3_18: Gummibandbeispiel .
Der Durchschnitt aus vielen unreliablen Messungen ergibt wegen *Messfehlerausgleich* einen reliableren Wert

mehrfach abgefragt wird (mit leichter Umformulierung, ähnlichen Adjektiven etc) und die gleich skalierten Ratings (s. Abb. 3_17) anschließend gemittelt werden.
Der Mittelwert der Antworten zu mehreren ähnlichen Fragen ist *reliabler* als die Antwort auf eine Einzelfrage! Die Reliabilität der Messung eines Konstrukts ist daher per definitionem gering, wenn nur eine einzige Frage zu dem Konstrukt gestellt wurde. Entdeckt man in einem wissenschaftlichen Text, dass eine AV mit nur einem einzigen Item gemessen wurde, kann man das immer kritisieren! Die Reliabilität eines Mittelwerts steigt mit der Anzahl zusammengefasster Fragen. Wie ähnlich die zur Zusammenfassung vorgesehen Fragen von den Versuchspersonen beantwortet wurden - *wie reliabel also der Mittelwert über diese Items ist* -, kann über ein **Maß der internen Konsistenz** berechnet werden (für intervallskalierte Variablen oft ´**Cronbachs Alpha**´), das die Korrelation der Items (mit einer Formel) zusammenfasst. Zu Fragebogen-Skalen ist es also üblich, die *interne Konsistenz* jeder Skala anzugeben (Einstellungsskalen erreichen oft nur Cronbachs Alpha ≈ .50, sollten aber etwa Cronbachs Alpha ≈ .70 erreichen, für Intelligenztests und ähnliche Tests, die zur folgenreichen Beurteilung einzelner Personen verwendet werden, ist man strenger und fordert Cronbachs Alpha ≈ .90). Je mehr positiv korrelierende Items zusammengefasst werden, desto höher die interne Konsistenz der Skala, desto höher die Reliabilität des Skalenwerts.

Exkurs zu dem Begriff Skala: Im letzten Satz war das Wort **Skala** in einer anderen Bedeutung verwendet worden, als bspw. in Abb. 3_17: das Wort ´Skala´ bezeichnet **(1)** die Höhe des *Skalen*niveaus, z.B. Ordinalskala oder Intervallskala, und meint dort, ob die Zahlen der Antwortalternativen einer einzigen Messung als ´richtige Zahlen´ interpretiert und daher mit normaler Mathematik verrechnet werden dürfen, oder nicht. Damit in Zusammenhang spricht man **(2)** davon, für die Antworten eine Rating*skalen* vorgegeben zu haben, z.B. eine fünfstufige Rating*skala* wie in Abb. 3_17. Diese Wortbedeutung (Skalierung der Antwortalternativen) war aus (1) abgeleitet: die Antwortalternativen in Rating*skalen* werden meist als intervall*skaliert* behandelt, eine Noten*skala*, deren *Stufen* einzeln benannt sind (1=sehr gut, 2=gut, .. 5=mangelhaft) hingegen wird meist als nur ordinal*skaliert* aufgefasst (besser also: nur die Ränder benennen wie in Ab.3_17). **(3)** Sollen mehrere ähnliche Items (Fragen, deren Antworten z.B. auf Rating*skalen* gegeben werden) zu einem Mittelwert zusammengefasst werden, so spricht man von der „Liste dieser Items“ als von *einer Skala* für die zu messende AV (meist Likert-Skala, es gibt andere Formen von Zusammenfassungsvorschriften, z.B. Guttman-Skala, Rasch-Skala u.a.m.). Bspw. kann die Y-Achse in Abb. 3_15 über eine *Skala zum Führungserfolg* erfasst worden sein, die aus vielleicht 10 Items bestand, die unterschiedliche Aspekte erfolgreicher Führung formulieren. Die MitarbeiterInnen mussten dann 10 Antworten geben, z.B. jew. auf einer fünfstufigen Rating*skala*. Um nachzuweisen, dass die *Skala* zum Führungserfolg *reliabel* ist (meint: .. der Mittelwert über die 10 Items reliabel ist), wird ein Maß der *internen Konsistenz* berechnet, das prüft, ob die 10 Items miteinander hoch korrelieren, also etwa das Gleiche messen; die *Skala* zum Führungserfolg also *eindimensional* ist. Items, die Verschiedenes messen (miteinander kaum korrelieren), darf man nicht zu einem Mittelwert zusammenfassen, sie bilden keine *gemeinsame Skala* (man kann eine Faktorenanalyse rehnen, bspw. zwei Faktoren identifizieren, dann bilden die Items je eines Faktors eine Skala und die des anderen Faktors eine andere Skala).

War eine Messung *objektiv* und *reliabel*, so kann ihre *Validität* nur noch dadurch beeinträchtigt sein, dass etwas inhaltlich Falsches gemessen wurde.

Validität der Messung: Gültigkeit, mit der die Messung das messen konnte, das mit ihr gemessen werden sollte. Die Validität der Messung ist gering, wenn (Objektivität oder Reliabilität herabgesetzt sind oder/und) etwas inhaltlich anderes als das beabsichtigte gemessen wurde.

Validitätskontrolle: Berechnung von *konvergenter Validität* (evtl. zusätzlich auch noch *diskriminanter Validität*, s.u.) mit anderen Maßen, die dasselbe (bzw. absichtlich diskriminat = etwas anderes) messen sollen.

Validitätsoptimierung: ? -schwierig: Verändern der Messung, bis die Validitätskontrollen gute *Validität der Messung* ergeben.

Wenn die Teilnehmenden in einem Fragebogen nicht ihre eigene Meinung, sondern nur die *sozial erwünschte* Meinung angeben, ist die Einstellungsmessung (oder Zufriedenheitsmessung etc) *wenig valide* (bspw. *Probandenerwartungseffekt*, Tab. 3_13). Die Messung der Sprechgeschwindigkeit in einer Abiturprüfung ist sicherlich keine besonders valide Operationalisierung der Variable ´Leistung´ (als AV zu den Hypothesen aus Tab. 3_4), Sprechgeschwindigkeit könnte ja statt von der Leistung(sfähigkeit) auch oder nur von der Erregung direkt abhängen (aber auch für Erregung gibt es sicherlich validere Maße, z.B. die Herzschlagrate). Eine inhaltlich besser passende (=*inhaltsvalide*) Messung des gewünschten Inhalts zu finden, ist oft wirklich schwierig! Hier lohnt das Literaturstudium (gibt es schon ein *valides* oder zumindest bewährtes Maß für eine Variable?). In der Studie von Abb. 3_15 wurde die Selbstauskunft der Führungskräfte über ihren eigenen Führungserfolg nicht als valides Maß für tatsächlichen Führungserfolg bewertet (sondern hohe Werte wurden als Selbstüberschätzung interpretiert, mit der Intervention sollen die Scores der Selbstaussagen ja sogar sinken!). Ob also ein Fragebogen (hier die *Skala* zum Führungserfolg) als *valides Maß* für Führungserfolg angesehen werden kann, wird über Ableitungen aus *Theorien* mitbestimmt (hat man eine Theorie über Selbstwertschutz, wird man Selbstauskünfte nicht zur Leistungsmessung verwenden). Aus Messproblemen entwickeln sich in der Praxis oft weitere Theorien oder die bisherigen werden präzisiert (s. Kap. 5), in dem sie zu den behaupteten Konstrukten (s. Kap. 2) auch Messvorschriften (also Operationalisierungen) angeben.
Gerade Leistungsmessungen variieren weit über das Spektrum von Abb. 3_1. Antworten in Fragebögen sind *verfälschbar*. Intelligenztests sind manchen Leuten zu künstlich (s. Abb. 3_1 & 3_2: geringe *ökologische Validität* wird auch künstlichen Messungen zugesprochen), insbesondere, wenn die Theorie, dass spezifische Leistung durch abstrakte (allgemeine) Intelligenz ermöglicht werde, nicht akzeptiert ist... (man hält dann die *allgemeine Intelligenz* für wenig valide *zur Prädiktion* einer bestimmten beruflichen Leistungsfähigkeit, hier geht es dann schon um *prädiktive Validität*). Wie valide ist die Beurteilung des Umgangs mit der Postkorb-Aufgabe in einem Assessment-Center für die Berufseignung? Die *inhaltliche Übereinstimmung zwischen dem Gemessenen und dem zu Messenden* ist meist leicht zu kritisieren aber schwierig abzuschätzen - denn dazu muss man das zu Messende (hier z.B. ´Leistung´) noch ein zweites Mal gut (=valide) gemessen haben, um zu vergleichen, ob die neue Messung nahe an die alte herankommt, beide also hoch korrelieren. Z.B. könnte die Note eine *validere Operationalisierung* der Leistung sein als die Sprechgeschwindigkeit. Nachgewiesen werden kann die geringe Validität der Sprechgeschwindigkeit, wenn gezeigt wird, dass die Sprechgeschwindigkeitswerte mehrerer Personen (wenig valides Maß) schlecht (also gegen Null) mit den Noten dieser Personen (valideres Maß) und den späteren Leistungen (die aber auch zu messen sind) korrelieren. Auf diese Weise kann man *wenig valide Maße begründet kritisieren.*
In der Praxis wird die inhaltliche Übereinstimmung zwischen dem Gemessenen und dem zu Messenden häufig einfach erstmal behauptet: man sagt dann, es läge **Augenscheinvalidität** vor (= „das sieht man doch!“). Etwas objektiver (objektiv meint in der Statistik immer: *intersubjektiv*, s.o.) wird diese Behauptung, wenn mehrere ExpertInnen die (Augenschein-)validität des Messinstruments abgesichert haben (ExpertInnen beurteilen, ob die Items ´inhaltlich geeignet´ erscheinen). Toll ist das natürlich immer noch nicht, die ExpertInnen könnten sich alle irren. Sinnvoller ist, sich an *bewährten* Instrumenten zu orientieren, bewährte Fragebögen aus der Literatur zu übernehmen. Das sichert die Validität der Messung nicht unbedingt, aber man bekommt weniger Kritik, da es die anderen auch so machen und so immerhin ***Kommensurabilität*** (= Vergleichbarkeit der Ergebnisse über mehrere Forschungsgruppen) sichergestellt wird.
Soll ein neues Maß für eine Variable eingeführt werden, so sollte eine **Validierungsstudie** durchgeführt werden (extra um die *Validität* des Maßes nachzuweisen): an vielen Personen wird einerseits das neue Maß (z.B. Assessment-Center Scores) erhoben, dann vielleicht zwei der bisherigen (wenn auch vielleicht hoffentlich etwas schlechteren) Masse für die Berufseignung (z.B. Abitursnote und IQ), sowie Maße, die wenig mit der Berufseignung zu tun haben sollen (z.B. Gesichtsattraktivität, Jahreseinkommen der Eltern, Test auf Fähigkeit zum Lügen), denen aber Konfundierungen unterstellt werden könnten. Damit dass Assessment-

Center als gutes Berufseignungsmaß gelten darf (wirklich das *Konstrukt ´*Berufseignung´ gemessen wird und nicht irgendetwas anderes), muss es mit den beiden anderen Berufseignungsmaßen, die dasselbe Merkmal (Konstrukt) messen sollen, hoch korrelieren (**konvergente Validität,** Dreiecksmatrix oben links in Tab 3_15), mit den Maßen inhaltlich anderer Merkmale Attraktivität, Einkommen, Lügekompetenz möglichst niedrig (**diskriminante Validität**, rechteckiger Matrixausschnitt unten links in Tab 3_15).

Tab. 3_15: Korrelationsmatrix zum Nachweis der Konstruktvalidität des Assessment-Center (fiktiv)

AC – Score	1.0					
Abiturnote	**+.6**	1.0				
IQ	**+.7**	**+.4**	1.0			
Attraktivität	+.2	+.1	.0	1.0		
Einkommen	.0	+.1	+.1	+.1	1.0	
Lügentest	+.2	+.0	+.1	.0	-.1	1.0
	AC – Score	Abiturnote	IQ	Attraktivität	Einkommen	Lügentest

Objektivität, Reliabilität und *Validität der Messung* sind die drei Haupt-*Gütekriterien der Messung*. Daneben werden noch **Neben-Gütekriterien** diskutiert, die zwar erst nach ausreichender Sicherstellung der Haupt-Gütekriterien zu beachten sind, aber in der wirtschafts-psychologischen (und auch klinischen, pädagogischen usw.) Praxis eine große Rolle spielen: Fragebögen und Tests sollen *ökonomisch*, *fair/akzeptierbar, nützlich* und *breit anwendbar* sein. Die **Ökonomie** erklärt sich von selbst: wenn drei Items genügen, um ein Konstrukt mit guter *interner Konsistenz*, also *Reliabilität* (s.o.), zu erfassen, warum dann den Fragebogen unnötig um viele weitere Items aufblähen? Die **Fairness** eines Tests erhöht seine **Akzeptanz** bei allen Beteiligten: werden Aufgaben im AC als ungerecht und nicht mit dem späteren Job verbunden empfunden, sinkt dessen Akzeptanz (und auch die der beauftragenden Organisation) bei den BewerberInnen. Scheint ein Intelligenztest weiße vor farbigen Amerikanern oder Männer vor Frauen zu bevorzugen, sinkt seine Akzeptanz bei verantwortlichen AnwenderInnen. In der Praxis ist die Akzeptanz eines Messinstrumentariums bei den Entscheidungsgremien auch von seiner Ökonomie und Nützlichkeit (/Relevanz der Variablen) abhängig. In der wirtschafts-psychologischen Praxis ergibt sich hier ein interessanter Konflikt zwischen den methodisch geschulten AnbieterInnen und den methodisch weniger geschulten NachfragerInnen: die Nachfrageseite wünscht zur Evaluation ihrer Maßnahme oder zur Auswahl unter BewerberInnen bspw. gern ein sehr spezifisches, auf die jeweiligen Verhältnisse zugeschnittenes Instrument (*tailor-made* = maßgeschneidert). Von *tailor-made,* also extra hergestellten Instrumenten weiß man aber über ihre Messgüte noch gar nichts! Auf der anderen Seite können Instrumente mit einer hohen **Anwendungsbreite** zu vielen verschiedenen Zwecken eingesetzt werden, sodass bei späteren Einsätzen die Messgüte (zumindest bspw. die interne Konsistenz von Skalen etc.) schon zuvor bekannt ist. Das speziell konzipierte AC und maßgeschneiderte, selbst-entwickelte Fragebogen scheinen auf den ersten Blick Validität (aber nur *Augenscheinvalidität*) sowie Ökonomie und wegen beidem dann Akzeptanz zu optimieren. Allerdings lassen sich die Ergebnisse solcher maßgeschneiderten Instrumente („es kommt 3,8 raus“, s. den Text zu Abb. 3_14a) mit nichts vergleichen und daher nicht interpretieren („ist 3,8 nun gut oder nicht?“). Instrumente mit einer hohen Anwendungsbreite (das Semantische Differential von Ertel, der SYMLOG-Fragebogen von Bales, die Big Five gemäß Neo-PI, die Werteskala von Schwartz u.a.) haben den unschätzbaren Vorteil, Vergleichsdaten aus anderen Studien zu besitzen und damit vergleichende Aussagen („in anderen Betrieben wurden Mittelwerte von 4,1, 4,2 und 4,3 erreicht, also ist 3,8 eher wenig“) zu erlauben (vgl. die Argumentation zu Abb. 3_14b). Für eine valide Interpretation ist nach der Erfahrung der Verf. die Verwendung eines in der Literatur vorgefundenen Messinstruments mit Vergleichswerten einer Neukonstruktion immer vorzuziehen. Vergleichswerte steigern die *Interpretationsvalidität*.

3.2.3 Zusammenfassung der Gütekriterien der Untersuchung

Mit den *Gütekriterien der Messung* (*Objektivität, Reliabilität, Validität der Messung*) und den in Kap. 3.1 besprochenen, auf das Design (die Realisation der UV) bezogenen *Maßnahmen zur*

Steigerung der internen Validität der Untersuchung (*Konfundierungen* vermeiden, *Kontrollvariablen* einführen, *UV* wo möglich *experimentell* variieren) sind die wichtigsten *Gütekriterien der Untersuchung* genannt.

Güte der Untersuchung:

<= interne Validität der Untersuchung
(Ergebnis ist nur auf geplante Bedingungen zurückzuführen)
keine Konfundierungen mit der UV, ceteris paribus, siehe Güte der UV.

<= Güte der (AV-) Messung:

<= Validität der Messung
(inhaltliche Übereinstimmung von Zumessendem mit Gemessenem)

<= Reliabilität (Zuverlässigkeit) der Messung
(wenig Messfehler oder Messfehlerausgleichsmittelung)

<= Objektivität der Messung
(kein Interpretationsspielraum der VL bei der Messung)

<= Güte der (UV-) Realisation
(Kontrolle der Anfangsbedingungen)

<= Validität der UV-Realisation
(inhaltliche Übereinstimmung von zu Realisierendem und der Realisation,
Kontrolle potentieller systematischer Störfaktoren: konfundierte Variablen)

<= Reliabilität d. UV-Realisation
(Zuverlässigkeit der Realisation, Kontrolle unsystematischer Störfaktoren)

<= Objektivität d. Realisation
(bei anderer Versuchsleitung gleiche Bedingungen)

<= externe Validität der Untersuchung
(Ergebnis ist generalisierbar)

<= Repräsentativität der Probanden
(Generalisierbarkeit auf folgende Population ..)

<= Probandenselektion

<= Probandenselbstselektion

<= ökologische Validität
(Dimension Künstlichkeit - Natürlichkeit der Untersuchung,
Repräsentativität der Stimuli / Situationen, Generalisierbarkeit auf folgende Situationen ..)

<= Ökonomie
(Ergebnis hätte nicht einfacher/klarer erbracht werden können)

<=: hängt ab von

Abb. 3_19: Zusammenfassende Ordnung der Gütekriterien der Untersuchung

Abb. 3_19 fasst die **Gütekriterien der Untersuchung** zusammen, man lese sie ´von rechts nach links´: Die *Objektivität der Messung* ist notwendig (aber nicht hinreichend) für die *Reliabilität der Messung*, die Reliabilität notwendig aber nicht hinreichend für die *Validität der Messung* und die *Validität der Messung* notwendig aber nicht hinreichend für *die Validität der Untersuchung*. Die Kontrolle der Versuchsbedingungen, die ja notwendig für die *interne Validität der Untersuchung* ist (vgl. Kap. 3.1) kann als ´Validität der Realisation der Versuchsbedingungen´ (oder Validität der UV-Realisation) aufgefasst werden (inhaltliche Übereinstimmung von zu Realisierendem und der Realisation). Wird bspw. eine UV über *Konföderierte* der Versuchsleitung realisiert, müssen bezahlte Hilfskräfte bspw. ´freundlich´ oder ´unfreundlich´ zu Versuchspersonen sein, oder müssen sie sich, wie in Experimenten zur Theorie der Sozialen Erleichterung (Abb. 2_2) in einer Versuchsbedingung neben die Versuchsperson stellen (während in der KG die Versuchspersonen alleine bleiben), so bedeutet ein unkontrolliertes Verhalten der Konföderierten (z.B. Kichern) eine ´geringe Objektivität der Versuchsbedingungen´. Um die Objektivität der Versuchsbedingungen zu steigern, werden Instruktionen schriftlich vorgegeben (vielleicht klappt es auch, die Konföderierten *blind* über die

Hypothesen zu lassen? vgl. Kap. 3.2.2). Um die Reliabilität der UV „Anwesenheit anderer“ zu steigern, sollte pro UV-Stufe immer die gleiche Anzahl anderer anwesend sein usw. *Standardisierung der Bedingungen sichert die Reliabilität der UV-Realisation.* In Untersuchungsverfahren, die ökologische Situationskomplexität oder gar soziale Interaktionen in ihre Bedingungen wieder einbeziehen möchten (Segmente D und E in Abb. 3.1) ist die Reliabilität der UV wirklich *das* Problem: wer im Assessment Center das Pech hat, mit einer zufällig sehr durchsetzungsfähigen extravertierten Person zusammen eine simulierte Verhandlung vorführen zu müssen, wird über die unfair unterschiedlichen (also wenig objektiven) Bedingungen klagen.

Wurde im Treatment inhaltlich etwas anderes als beabsichtigt realisiert (bspw. haben in den Hawthorne-Experimenten nicht die in der EG vorgenommenen Beleuchtungsveränderungen gewirkt, sondern einfach die empfundene Wertschätzung durch die Führung), dann war die Validität der UV-Realisation suboptimal. Zur Sicherung der Validität der UV-Realisation diente fast das ganze Kap. 3.1 (vor allem 3.1.4 – 3.1.7): Wenn eine Drittvariable mit der UV konfundiert ist, war die UV nicht valide realisiert worden. Die externe Validität der Untersuchung (Abb. 3_19) wird in Kap. 3.3 noch einmal aufgegriffen, wobei ihr Bestandteil der ökologische Validität in Kap. 3.1.3 schon angesprochen worden war.

3.3 Durchführung und Stichprobe

In einem klassisch gegliederten Versuchsbericht (insbesondere für Laborexperimente) wird über die Teilnehmenden (im Laborexperiment meist: Versuchspersonen, **Vp**, Teilnehmende, **Tn.**, im englischen auch **Ss** = *subjects* oder *participants* ***Pp*** oder *sample*) erst in Kapitel 3.3 berichtet (Überschriften für Kap. 3.1, 3.2, 3.3 etwa so wie hier im Skript, Unterabschnitte aber, wenn, dann inhaltsangepasst). In bspw. betrieblichen Untersuchungen, die nicht zum Test einer Theorie durchgeführt werden, sondern um in diesem Betrieb etwas (z.B. die Arbeitszufriedenheit etc.) zu diagnostizieren oder eine Maßnahme zu evaluieren, wird die geplante Stichprobe (z.B. ´Totalerhebung´ oder: ´freiwillige Teilnahme bei Anschreiben an alle 70 Beschäftigten´, oder ´Teilnehmende an der Pilotphase des Trainings´ etc.) oft schon im Design (Kap. 3.1) genannt.

Was ist an der Teilnehmenden-Auswahl wichtig? Eine *hypothesentestende* Untersuchung kann an Wert verlieren, wenn *zu wenige* Personen teilnahmen (im Slang: „das ***N*** zu klein ist“, um fair zwischen H1 vs. H0 zu entscheiden, s. Kap. 4.2) oder wenn es *die falsche Auswahl* war. Die Anzahl der Versuchspersonen bestimmt ganz praktisch darüber, ob ein bestimmtes hypothesenkonform aussehendes Ergebnis (z.B. die Korrelation von r = +.7 in Tab 3_15) ´bedeutsam´ („signifikant“, s. *Signifikanztest* Kap. 4.2) ist, oder aber unbedeutend („nicht signifikant“), da das Ergebnis bei zu wenigen Vp auch *zufällig* zustande gekommen (oder zufällig ausgeblieben) sein könnte. Daher wird zur Planung von Studien gern vorher berechnet (sogn. Power-Analyse, s. Kap. 4.2), wie viele Vp nötig sind, um ein erwartetes Ergebnis bestimmter Stärke (zu Effektstärken auch Kap. 4.2) signifikant werden zu lassen, nicht dass es an einer zu kleinen Stichprobengröße scheitert.

Gemäß Abb. 3_19 setzt sich die *Güte der Untersuchung* aus ihrer *internen Validität* und *externen Validität* (inkl. Ökonomie) zusammen. Wurden zu wenige und zu *homogene* Personen untersucht, so kann das Ergebnis zwar für die untersuchten Personen perfekte Gültigkeit haben = die *interne Validi*tät der Untersuchung kann hoch sein (= über die intendierte Kausalaussage kann für die Stichprobe entschieden werden). **Externe Validität** erhält eine Untersuchung, wenn sich ihr Ergebnis auf die Gesamtbevölkerung bzw. eine vorher angegeben bestimmte Population generalisieren lässt. Dazu müssen die Teilnehmenden hinreichend **repräsentativ** für diejenige *Population* sein, auf die generalisiert werden soll. Will man bspw. (nur) auf die Population der Studierenden generalisieren, braucht man natürlich nur Studierende zu untersuchen. Nicht die Anzahl (s. Tab. 4_6) sondern die Art der Auswahl der Vp (*Teilnehmendenselektion*, s. Tab. 3_13) bestimmt, auf welche anderen Personen (auf welche **Population**) die Ergebnisse *generalisiert* werden dürfen. Soll auf die bundesdeutsche Bevölkerung generalisiert werden, so muss die Stichprobe der untersuchten Personen möglichst *repräsentativ* für die Bevölkerung sein. Hier ist gar nicht ratsam, einfach besonders

viele Personen zu untersuchen (dies widerspricht auch dem Gütekriterium der Ökonomie, s. Abb. 3_19), sondern Personen mit einer Merkmalsverteilung, wie sie in der bundesdeutschen Bevölkerung vorliegt (und aus Bundesstatistiken etc. vorher eruiert werden muss), bspw. eine Schichtenstichprobe nach Alter, Geschlecht, Berufsgruppen und Region (die für Deutschland repräsentative ´Allgemeinen Bevölkerungsumfragen der Sozialwissenschaften ALLBUS´ oder der seit 2002 durchgeführte ‚European Social Survey ESS' erreichen Repräsentativität für Deutschland mit Stichproben von um zweitausend Personen). Bestimmte ProbandInnen-gruppen sind dabei schwieriger als andere zu rekrutieren (z.B. fehlen immer Männern mit geringerem Einkommen u. geringer Bildung – sie antworten nicht so gerne). In einer Schichten- oder Quotenstichprobe können die genannten soziodemographischen Variablen als *Kontrollvariablen* betrachtet werden (wie der Industrialisierungsgrad der Länder zur Prüfung der Störche-Hypothese, Tab. 3_9 & Abb. 3_8).
Probleme bereiten Stichprobenverteilungen besonders, wenn Personen selber entscheiden, ob sie an der Untersuchung teilnehmen: die Variablen, die die Teilnahme-Motivation bestimmen, können mit untersuchungsrelevanten Variablen *konfundiert* sein! Solche Probleme der **Stichprobenselbstselektion** (s. Tab. 3_13) sind schwer zu lösen. Wenn nur Neugierige bei Produkttests teilnehmen, wird die Akzeptanz des neuen Produkts überschätzt (Neugierige sind nämlich meist neophil (=Neues liebend)). Wenn nur Personen mit hoher Konformität oder nur solche mit Motivation zu freiwilligem Engagement bei politischen Meinungsumfragen teilnehmen, wenn nur Personen mit besonderer Unzufriedenheit oder umgekehrt nur begeisterte Kunden bei Zufriedenheitsbefragungen teilnehmen, sinkt auch die interne Validität des ermittelten Zufriedenheitsmittelwerts. Evaluationsstudien (s. Kap. 3.1.8) kämpfen fast immer mit diesem Problem. Selbst in Kontrollgruppendesigns (Abb. 3_14b, 3_14d) darf die Maßnahme selten *randomisiert* zugeteilt werden (Ausnahme in Abb. 3_15): oft werden Personalmaß-nahmen nur für MA mit z.B. besonderer Weiterbildungsbedarf (besonderem Entwicklungs-potential ☺) bezahlt – die Trainingsgruppe unterscheidet sich dann schon zum Prä-Messzeitpunkt von einer mühsam rekrutierten Kontrollgruppe (s. Fehlerquelle *Selektionseffekte* in Tab. 3_13), die Befunde können dann gar nicht mehr so optimal wie in Abb. 3_ 14d o. 3_16 ausfallen. Die meisten Studien müssen vollständig freiwillige Teilnahme zusichern (s. Ethischen Richtlinien der DGPs & des BDP). Um trotzdem *Stichprobenselbstselektionseffekte* zu verringern, sollte versucht werden, durch verschiedene Ansprache *systematisch* verschiedene Teilnehmendengruppen zu gewinnen (und vielleicht über erhobene Kontrollvariablen zu vergleichen). In einer Mitarbeitendenbefragung kann man sowohl über die Geschäftsführung als auch über die Personalvertretung – verschieden – werben lassen; in Internetstudien kann man auf verschiedenen Portalen, mit verschiedenen Incentives (für eine Gruppe auch ohne) werben. Es empfiehlt sich immer, im Vorfeld über *Stichprobenselbstselektion* nachzudenken und einige der befürchteten *konfundierenden Variablen* mit zu erheben, um hinterher entweder Stichprobenselbstselektionseffekte erleichtert auszuschließen (z.B.: wenn die Prä-Werte der KG, wie in Abb. 3_ 14d, gleich zu denen der Treatmentgruppe ausfallen) oder zumindest konfundierte Variablen aus den Ergebnissen ´herauszurechnen´ (wie in Abb. 3.8).
So wie die *ökologische Validität* die Generalisierbarkeit der Untersuchungs*situation* auf andere (´externe´) Situationen beurteilt (Kap. 3.1.3), bestimmt die *Repräsentativität* der Untersuchungsteilnehmenden die Generalisierbarkeit der Ergebnisse auf andere (´externe´) Personen, daher macht beides (in Abb. 3_19) die *externe Validität* der Untersuchung aus.

4. Ergebnisdarstellung

Die Ergebnisse einer empirischen Studien umfassen die *deskriptive* = beschreibende *Statistik (*die Stärke der gefundenen *Effekte* anzugeben, gehört dazu, Kap. 4_1) und die *inferenzstatistische Statistik*, mit der über die *Hypothese*(n) entschieden wird (s. Tab. 3_3 „muss die H0 beibehalten werden oder darf sie abgelehnt werden?": Kap. 4.2).

4.1 Überblick zur deskriptiven Auswertung, Effektgrößen d und r

Die drei in unserem Fach wohl häufigsten deskriptiven (= beschreibenden) Statistiken sind der **Mittelwert** (Abk.: MW oder M; engl. mean, Symbol: $\overline{x}$ (gesprochen: x-quer)), die **Streuung** (=*Standardabweichung*, Abk: *s*, engl.: standard deviation: SD) und die **Korrelation** (engl.

correlation. Symbol: r, für die normale, also die Pearson Korrelation, auch Produkt-Moment-Korrelation, die *Intervallskalenniveau* voraussetzt (s. Kap. 3.2.1). Bei *Ordinalskalenniveau* wird meist die Spearman-Korrelation (Rang-Korrelation) berechnet und mit Rho abgekürzt.
Die vielleicht häufigste deskriptive Statistik ist die Versuchspersonenanzahl = **N** (werden nicht viele Personen, sondern z.B. viele Situationen oder Targets untersucht, so wird die Anzahl der *Fälle* auch mit *n* abgekürzt. Es existieren viele deskriptive Statistiken mit konventionell festgelegten Abkürzungsbuchstaben, siehe Statistik–Handbücher).
Der Mittelwert ist das wichtigste Beschreibungskriterium einer Häufigkeitsverteilung für eine intervallskalierte AV (s. Abb. 4_1).

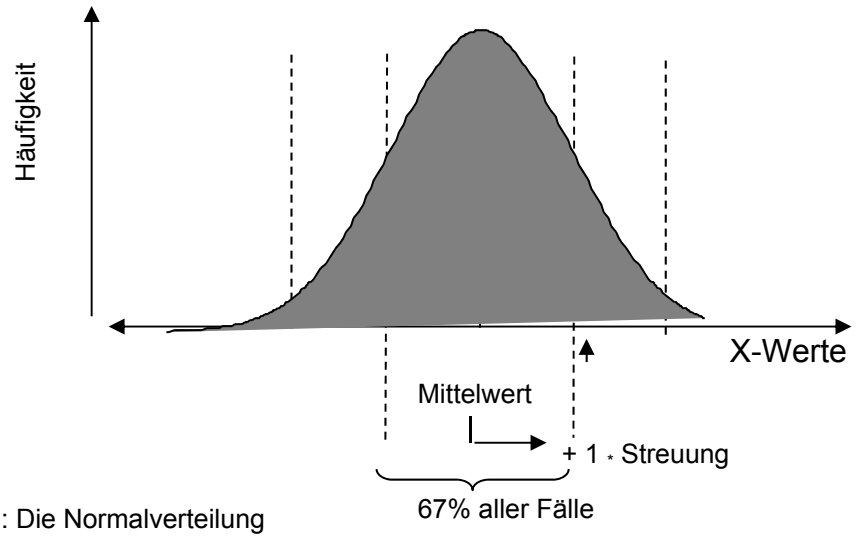


Abb. 4_1: Die Normalverteilung

Die *Streuung* s oder synonym *Standardabweichung* SD (englisch: standard deviation, daher SD) einer idealen Häufigkeitsverteilung, die der vom Göttinger Astronomen Gauss (1777-1855) gefundenen ***Normalverteilung*** entspricht***,*** reicht vom *Mittelwert* bis etwa zum Wendepunkt der sogn. Glockenkurve (Abb. 4_1 und auch Abb. 3_18). Im Bereich von M ± **1*s** liegen in einer *normalverteilten Häufigkeitsverteilung* etwa **67%,** im Bereich von M ± **2*s** etwa **95%** aller untersuchten Fälle (wer 67% und 95% auswendig weiß, kann mit Mittelwert & Streuung in Ergebnisberichten von Publikationen mehr anfangen. Daher ist die *Streuung* das zweitwichtigste Beschreibungskriterium der *Verteilung* einer *intervallskalierten AV*). (Formel für Streuung siehe Statistik-Buch, die quadrierte Streuung s^2 heißt *Varianz*). Die *Normalverteilung* mit einem Mittelwert von 0 und einer Streuung von 1 spielt eine große Rolle bei den Signifikanztests (s. Kap. 4.2), sie wird **Standardnormalverteilung** und auch gerne *z-Verteilung* genannt (wenn der Mittelwert 0 und die Streuung 1 beträgt, kann an der X-Achse von Abb.4_1 also *z-Wert* statt X-Wert stehen)[1].

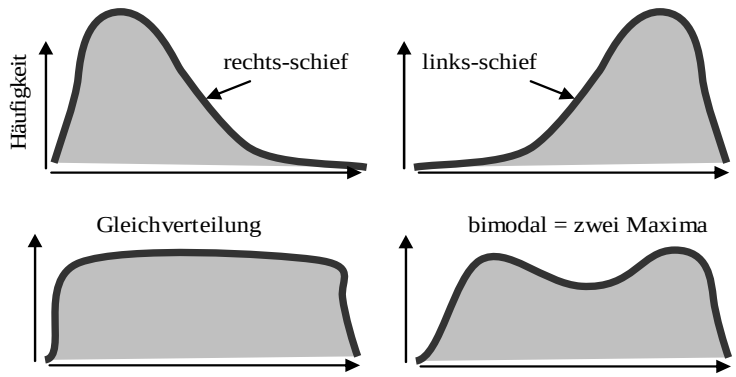


Abb. 4_2: Wenn Häufigkeiten nicht-normal verteilt sind.

Viele statistische Auswertungsverfahren und viele *Signifikanztests* (s. Kap. 4.2) verlangen nicht nur, dass die AV auf *Intervallskalenniveau gemessen* wurde (s. Kap. 3.2.1), sondern auch, dass sie empirisch *normalverteilt* ist. Also darf sie nicht *gleichverteilt*, nicht *bimodal*, nicht *rechtsschief* oder *linksschief* verteilt sein (Abb. 4_2). Beispielsweise sind Zeit-Messungen oft *rechtsschief* (wer länger braucht, kann leicht sehr viel länger brauchen). Arbeits- oder Kunden-zufriedenheit ist oft *linksschief* verteilt

[1] Man kann jede *Normalverteilung* in einer *Standardnormalverteilung* umrechnen: $z_i = (x_i - MW) / s$. Der *z-Wert* +1 bedeutet also: um eine Streuung überdurchschnittlich. Und z= -2: um zwei Streuungen unterdurchschnittlich, für genauer z=-1.98 gilt: nur etwa 2,5% der Fälle haben noch kleinere Werte (ca 95% innerhalb ± 1.06*s, also 5% außerhalb, davon die Hälfte, 2,5%, an jedem Zipfel von Abb. 4_1).

(die meisten Personen geben sich zufrieden; vermutlich, weil sie sich arrangiert haben = sogn. resignative Zufriedenheit).

Hinter *bimodalen* Verteilungen wie in Abb. 4_2 unten rechts (*Modalwert* nennt man den häufigsten X-Wert, bi-modal = zwei Hochpunkte), verbergen sich häufig zwei verschiedene Populationen (z.B. Männer & Frauen bzgl. der Schuhgröße, oder Y vor und nach einer sehr wirksamen Maßnahme etc). Bevor man Signifikanztests rechnet (s. Kap. 4.2), die eine Normalverteilung voraussetzen, sollte man sich die Verteilungsform der eigenen Daten also ansehen (Statistik-Software: Graphiken: Häufigkeitsverteilung = *Histogramm*)[2].

Soll eine **Unterschiedshypothese** (s. Tab. 3_1) geprüft werden, sind nach dem Check der Verteilung (ist die intervallskalierte AV hinreichend *normalverteilt*?) Berechnungen von *Mittelwerten* und *Streuungen* pro Bedingung angesagt (Streuung verlangt schon *Intervallskalenniveau*, Tab. 3_14). Sollen nur zwei Gruppen verglichen werden (z.B. die Schuhgröße von Männern und Frauen), kann allein aus Mittelwerten und Streuung bereits eine *Effektgröße* berechnet werden. Die ***Effektgröße* d** gibt an, ob der gefundene Mittelwertsunterschied (z.B. Geschlecht → Schuhgröße laut wikipedia.org: m=44 und w=40), als ein *schwacher, mittlerer oder starker Effekt* zu bewerten ist, denn dafür gibt es von Cohen schon 1969 vorgeschlagene aber erst seit ca den 1980ern breit akzeptierte Konventionen (Abb. 4_3).

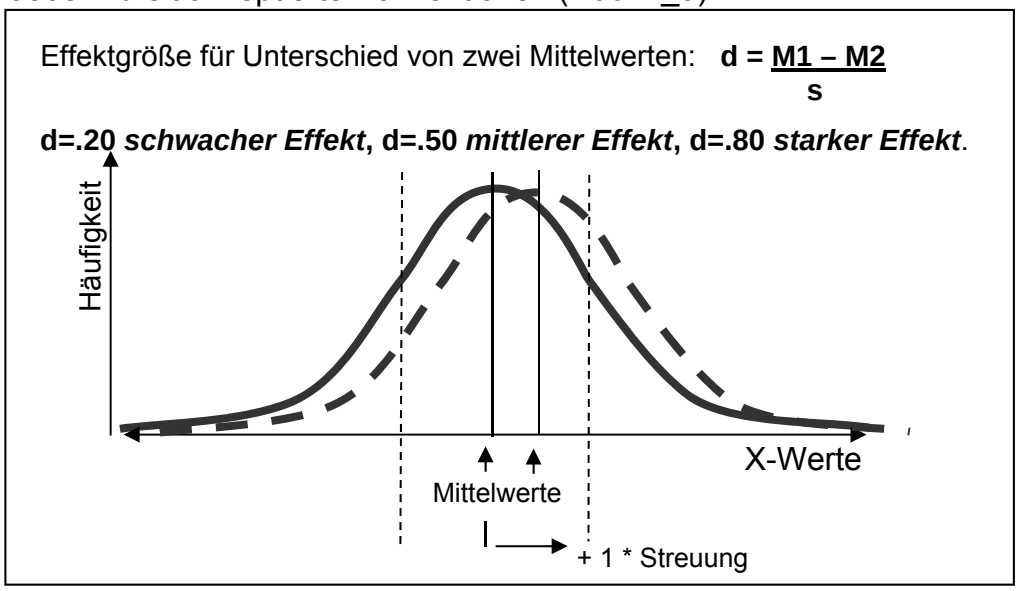


Abb.4_3: Diese Verteilungen unterscheiden sich in ihren Mittelwerten um etwa eine halbe Streuung d= .50.

In Abb. 4_3 sind die Streuungsgrenzen für die durchgezogene Normalverteilung als gestrichelte Senkrechte eingezeichnet. An den durchgezogenen Senkrechten für die beiden Mittelwerte wird erkennbar, dass sie um die Größenordnung einer halben Streuung verschieden sind: also ist die Effektgröße d=.50, also liegt gemäß Cohen-Konvention ein *mittlerer Unterschiedseffekt* vor.
Die Streuung der Schuhgrößenverteilung in Dt. konnte Verf. bisher nicht eruieren. Aber angenommen die Streuung der Schuhgröße bei Erwachsenen sei s= 2 (dann haben 67% der Männer eine Schuhgröße im Bereich 42-46 und 67% der Frauen eine im Bereich von 38-42, s. Abb. 4_1), dann ist der Geschlechtseffekt auf die Schuhgröße: d= (44 – 40) / 2 = 2.0, also als starker Effekt zu bewerten (ab d ≥ .80 wird in der Psychologie das Urteil ´starker Effekt´ gar nicht weiter differenziert). Daraus folgt: in der Psychologie untersuchte Effekte sind üblicherweise sehr viel kleiner als der Geschlechts-Dimorphismus bzgl. Schuhgrößen![3].

[2] es gibt Wege, bspw. schiefe Verteilungen wie in Abb. 4_2 oben künstlich in eine Normalverteilung, s. Abb. 4_1, umzurechnen und dann mit der transformierten Variable den Signifikanztest durchzuführen

[3] Abb. 4_3 enthält ausnahmsweise eine Formel! Viele Lehrbücher und fast alle Forschungsartikel geben Mittelwerte und Streuungen an, berechnen aber oft keine Effektgröße. Referiert man Forschungsarbeiten von anderen, soll man zwar deren numerische Ergebnisse nicht noch einmal abschreiben, man könnte aber vom allerwichtigsten Befund die *Effektgröße* ausrechnen und berichten – vielleicht war es nur ein schwacher Effekt? Sind die beiden *Streuungen* verschieden, darf man sie leider nicht einfach mitteln, aber jeweils quadrieren (das ergibt die *Varianz*, geschrieben s^2), diese dann mitteln, dann Wurzel ziehen – so erhält man die *Streuung s* für

Für den Vergleich von mehr als zwei Gruppen oder für Mittelwertsunterschiede in mehrfaktoriellen Designs gibt es andere Effektgrößen samt Formeln und Konventionen, z.B. *f*, eta^2, etc (-> neues Statistikbuch oder die *G*Power* Software von Erdfelder et al., siehe Netz).

Soll eine **Zusammenhangshypothese** (s. Tab. 3_1) geprüft werden, so ist die Berechnung einer **Korrelation** nötig (bei Intervallskalenniveau beider Variablen die Pearson-Korrelation r, bei *Ordinalskalenniveau* die Spearman-Korrelation Rho oder auch eine andere Rang-Korrelation, z.B. Kendalls Tau → s. Statistikbücher; auch für Binärvariablen gibt's eine ganze Reihe von deskriptiven Zusammenhangsmaßen). Das macht man mit einer Statistiksoftware (z.B. SPSS, statistica, SAS, R, Stata u.v.a.). Am wichtigsten ist, Korrelationen interpretieren zu können: der **Korrelationskoeffizient** variiert zwischen -1 über 0 bis +1. Tab. 3_15 hatte bereits eine Korrelationsmatrix (mehrere Variablen, jede mit jeder korreliert, also eine dreieckige Matrix) gezeigt. Abbildung 4_4 zeigt unterschiedlich starke Zusammenhänge von je zwei Variablen (jeder Punkt im Streudiagramm ist ein *Fall*, z.B. je eine Versuchsperson mit ihren Werten für x_1 und x_2). Wäre der Zusammenhang perfekt positiv (r=1), so lägen alle Punkte auf einer aufsteigenden Diagonalen, wäre der Zusammenhang perfekt negativ (r= -1), so lägen alle Punkte auf einer abfallenden Diagonalen. Ein negativer Zusammenhang bedeutet: „je höher x_1, desto niedriger x_2" (links in Abb. 4_4).

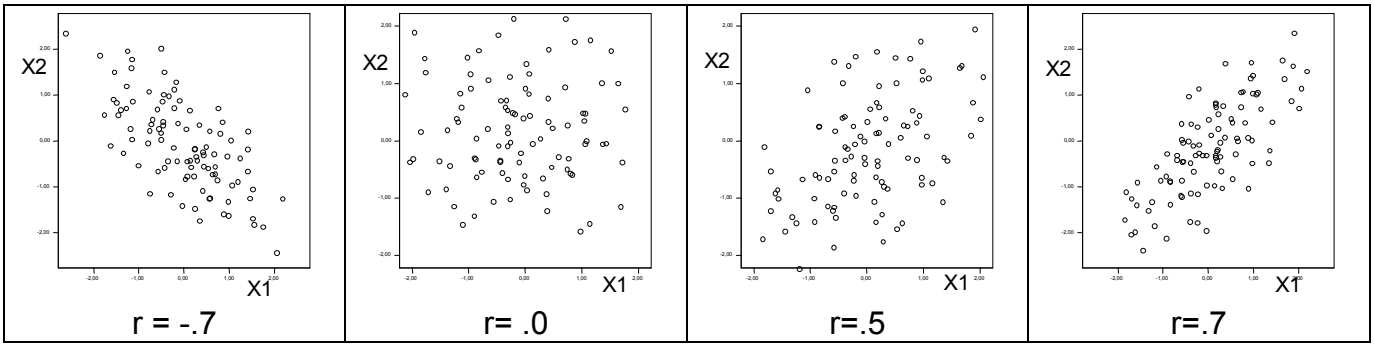


Abb. 4_4: Streudiagramme zur Illustration einiger Korrelationskoeffizienten.

Ergibt sich r=.0, so liegt *kein Zusammenhang* vor. Im zweiten Plot in Abb. 4_4 kann man entsprechend erkennen, dass bei hohen x_1 Werten sowohl hohe als auch niedrige x_2 Werte vorkommen usw. Alle vier Ecken des Plots sind etwa gleich gefüllt, aus x1 lässt sich x2 nicht vorhersagen. Zu *nullkorrelierten Variablen* sagt man auch: x_1 und x_2 sind *orthogonal* (wie in der Geometrie: 90°-Winkel). Bei positiver Korrelation hingegen sind die Ecken links oben und rechts unten im Streudiagramm (Abb. 4_4 rechts) fast leer, bei negativer Korrelation gibt es rechts oben und links unten kaum Fälle (Abb. 4_4 links). *Starke* Korrelationen kann man in Streudiagrammen wie denen in Abb. 4_4 also ´sehen´, *mittlere* und *schwache* aber von *Nullkorrelationen* mit dem Auge allein nicht unterscheiden. Stattdessen bewertet man die Zahl: das Schöne am der Korrelation r ist, dass es für sie als **Effektstärke r** auch Konventionen nach Cohen gibt: ein *schwacher Zusammenhang* besteht ab $|\mathbf{r}| = \mathbf{.10}$ (sprich: „Punkt – Zehn". Das Betragszeichen um r ist notwendig, da r = -.10 als *schwach-negativer Zusammenhang* gilt). Ein *mittlerer Zusammenhang* besteht ab $|\mathbf{r}| = .\mathbf{30}$, ein *starker Zusammenhang* ab $|\mathbf{r}| = \mathbf{.50}$.
Diese Effektstärkenkonventionen für r (.10 / .30 / .50) werden in der psychologischen Forschung dann angewendet, wenn für den Zusammenhang der beiden Variablen die Nullhypothese, r = 0 (es besteht kein Zusammenhang zwischen x_1 und x_2; vgl. die ersten beiden Zeilen in Tab. 3_2), sinnvoll ist. Die genannten Effektstärken-Konventionen für r gelten nicht, wenn die Korrelation als Maß der *Urteilerübereinstimmung* (zur Bewertung der Objektivität der Messung, s. Kap. 3.2.2) oder wenn sie als Maß der *internen Konsistenz* (zur Bewertung der Reliabilität einer Skala, s. Kap. 3.2.2), als *Retestreliabilität* (s. Kap. 3.2.2) oder als Maß der *konvergenten Validität* (s. Kap. 3.2.2 zu Tab. 3_15) berechnet wird; in diesen Fällen ist man anspruchsvoller: Die interne Konsistenz von Fragebogenskalen soll etwa Cronbachs α=.70

---

die d-Formel in Abb. 4_3. Wenn zwei Bedingungen verglichen werden, in denen die gleichen Personen zu beiden Bedingungen AV liefern (z.B. Prä-Post-Design in Abb. 3_14c), der Prädiktor also *within-subject* variierte (=abhängige Daten), berechnet sich d_w (w für within subject, im Gegensatz zum d_b aus Abb. 4_3, b für *between subject = unabhängige Gruppen*) aus der Mittelwertsdifferenz geteilt durch die *Streuung der Differenzen*.

betragen, für Intelligenztests (also wenn Einzelpersonen diagnostiziert werden sollen, denen dadurch relevante Konsequenzen drohen) sogar Cronbachs α=.90.

Auf Korrelationen bauen viele der *höheren statistischen Verfahren* auf (Abb. 4_5a u.4_5b). Es interessieren häufig Korrelationen zwischen sehr vielen Variablen gleichzeitig. Zunächst werden die Korrelationen zwischen je zwei Variablen berechnet und alle in einer dreieckigen **Korrelationsmatrix** zusammengestellt (Tab. 3_15 zeigt die Korrelationsmatrix für sechs Variablen, im Slang: die *Variableninterkorrelationen*).

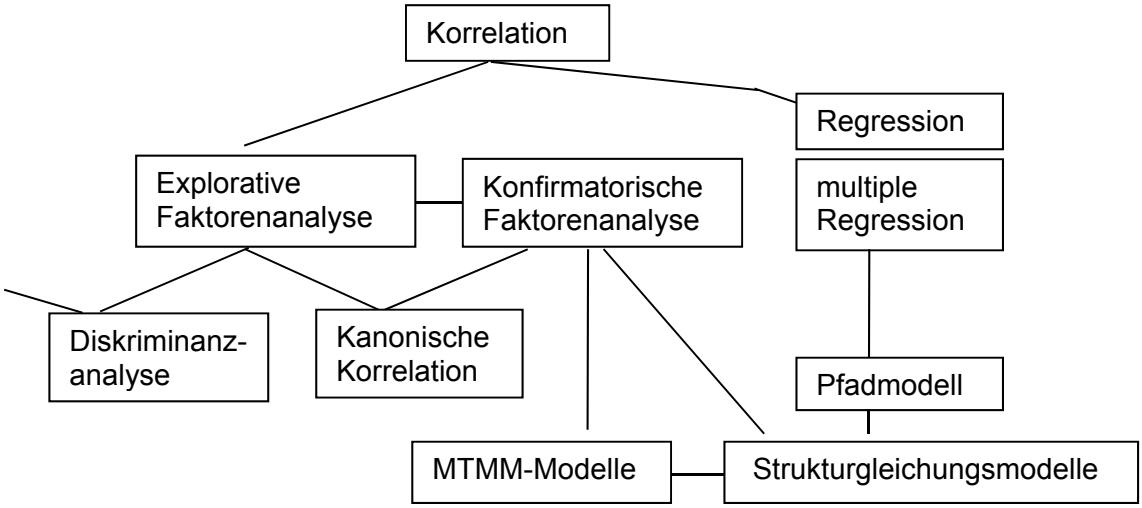


Abb. 4_5a: Korrelationsstatistische Verfahren und ihre Beziehungen zueinander.

Korrelieren bspw. alle betrachteten Variablen miteinander recht *hoch* (=egal ob positiv oder negativ), könnten sie vielleicht alle das Gleiche messen. Sie wären dann *redundant* – oder, ins Positive gewendet: man könnte statt der einzelnen Items ihren Mittelwert weiterverwenden, da sie bei hoher *Interkorrelation* ja gemeinsam eine *reliable Skala* bilden (s. Kap. 3.2.2). Um festzustellen, welche Items man zusammenfassen darf (oder sollte), welche also durch den gleichen *Faktor* bestimmt werden, wird eine *Faktorenanalyse* gerechnet (eine *explorative Faktorenanalyse*, Abb. 4_5a links oben, Abb. 4_5b links oben, Tab. 4_1 oben). *Explorativ* bedeutet, dass die Prozedur selber bestimmen darf, welche Items ihren Interkorrelationen entsprechend auf denselben Faktor gehören, welche aber auf einen zweiten, dritten usw. (explorativ in bayrisch heißt: „schau mer mal“). In *konfirmatorischen* Analysen hingegen muss man vorher festlegen, welche Zusammenhänge bestehen sollen, das *konfrmatorische* Verfahren prüft dann, ob diese ganzen Festlegungen (=Hypothesen) auf die Daten (meist die Korrelationsmatrix) passen (der „Fit“ hoch ist).

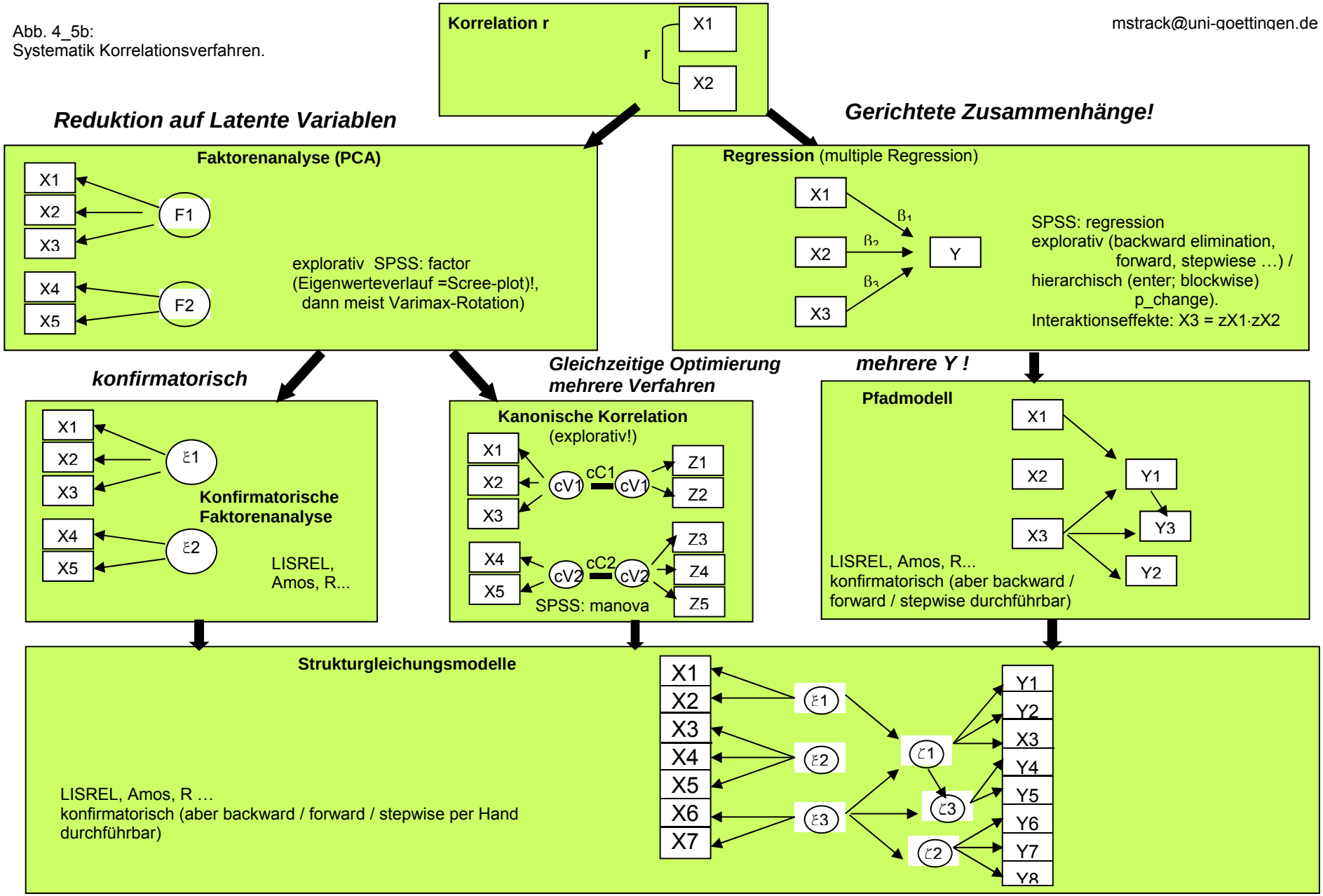
In der Fragebogenforschung mit ihren Rating-Skalen (s. Begriff *Skala* in Kap. 3.2.2)) sind Faktorenanalysen unabdingbar. Daher in Tab. 4_1 eine Kurzbeschreibung innerhalb der Liste der drei gebräuchlichsten *datenreduzierenden Verfahren* (*datenreduzierend* heißt die Faktorenanalyse, da aus vielen Items nur eine oder wenige Skalen gewonnen und dann mit den Skalenwerten anstatt mit den Rohitems weitergerechnet wird).

Da die meisten Hypothesen entweder *Unterschiedshypothesen* oder *Zusammenhangshypothesen* formulieren, wurden in diesem Abschnitt der einfache Mittelwertsvergleich (Relativierung der Mittelwertsdifferenz auf die Streuung in der *Effektgröße d*) sowie die *Korrelation* - unter Hinweis auf einige daraus abgeleitete Verfahren - genannt. Innerhalb dieses Einführungsskripts hat die Nennung der Verfahren (in Abb. 4_5ab und Tab. 4_1) nur den Zweck, das Selbststudium durch geeignete Suchbegriffe zu unterstützen.

Abb. 4_5b:
Systematik Korrelationsverfahren.

mstrack@uni-goettingen.de

a



Tab. 4_1 Datenreduzierende statistische Verfahren, die viele Items (Fragen, Begriffe, Stimuli oder auch Personen) zu wenigen Faktoren, Skalen, Dimensionen, Clustern .. zusammenfassen.

Ziel (typisches)	**Verfahren** (Kurzbeschreibung)
Zusammenfassung vieler Items in wenige Faktoren, Dimensionen, Skalen... Fragebogen-Überprüfung: laden alle diese Items auf einen Skala = ist die Skala eindimensional? Oder müssen doch mehrere unabhängige Faktoren in der Item-Liste angenommen werden? Wie viele Dimensionen (= Faktoren) hat dieser Itemsatz? Und welche Items messen das gleiche Konstrukt?	**Faktorenanalyse** (meist Hauptkomponentenanalyse; engl: factor analysis: FA, oder principal component analysis: PCA). Input: Rohdaten von Ratings etc. (Intervallskalen-Niveau!, links Fälle, oben Items) oder eine fertige Korrelationsmatrix. Software: z.B. SPSS: Dimensionsreduktion / Statistica: mutivar. expl. Techniken. Ergebnis: (1) Eigenwertplot zur Entscheidung über die Anzahl an Faktoren. (2) Tabelle mit (rotierten) Faktorladungen, um die Faktoren zu interpretieren und Items zusammenzufassen. Bei zwei Faktoren auch Ladungsplot möglich. Meist wird der Mittelwerte der Items, die zum selben Faktor gehören, als neue Variable weiterverwendet (oder die neuen Dimensionen werden automatisch gespeichert). (konfirmatorische Versionen können z.Z. noch nicht in SPSS aber in LISREL, AMOS, SEM o.ä. Software gerechnet werden).
Zusammenfassung vieler Personen (seltener: Items, Begriffe) in wenige Gruppen, Cluster, Segmente... (z.B. Marktanalyse: Segmentierung).	**Clusteranalyse** (engl: Cluster) Input: Rohdaten von Ratings etc. (oder eine fertige Ähnlichkeitsmatrize (dreieckig) von z.B. Personen oder Begriffen). Die Ähnlichkeit zwischen dem Antwortmuster je zweier Personen (…) kann zwar über Korrelationen berechnet werden, aber auch über Ähnlichkeitsmaße, die kein Intervallniveau voraussetzen (am häufigsten: *city block* o. euklidische Distan). Software: z.B. SPSS: Cluster / Statistica: mutivariate explorative Techniken. Ergebnisbild: Dendrogramm zur Entscheidung über die Anzahl an Cluster, Zuordnung jeder Vp in ein Cluster.
Darstellung der Bedeutungsähnlichkeit vieler Begriffe (Stimuli, Wörter, Items) in wenigen Dimensionen. (bspw. untersucht die Ethnologie Verwandtschaftsbegriffe, die Sozialpsychologie begründet ein Inhaltsmodell für Werte, die Marktforschung visualisiert Bestandteile einer Marke)	**Multidimensionale Skalierung (MDS)** Input: eine Ähnlichkeitsmatrize (dreieckig, z.B. city block o. euklidische Distanz) der Begriffe (oder Rohdaten von Ratings etc) Software: z.B. SPSS: Skalieren – Alscal / Statistica: mutivariate expl. Techniken. Output: ein zwei oder mehrdimensionales Bild, in dem die euklidische Distanz der Punkte den empirisch gewonnenen Antwortmuster-Ähnlichkeiten der Begriffe entspricht. Die Dimensionen oder Raumregionen können von den Forschenden interpretiert werden, oft werden aber einfach zusammen liegende Begriffe als Cluster umkreist und „Regionen“ im Bild benannt.

Neben solchen (und wie Abb. 4_5ab zeigen sollte, teils recht komplex werdenden) deskriptiven Statistiken wird *zur Entscheidung über eine Hypothese* aber neben der (Effekt-)Stärke des Befunds in der untersuchten (in der Psychologie oft kleinen) Stichprobe eine Abschätzung nötig, ob der Befund in der ganzen *Population*, hätte man sie untersuchen können, etwa ähnlich ausfallen könnte. In solche *Inferenz-Statistik* (Inferenz = Schluß, hier: Schließen von Stichprobenergebnis auf zu erwartendes Populationsergebnis) einzuführen, ist Ziel des nächsten Abschnitts.

4.2 Hypothesenentscheid, Signifikanztest

Wenn die *deskriptiven Statistiken* für das realisierte Versuchsdesign ausgewertet sind (z.B. Mittelwerte, SD und Effektgröße *d* für Unterschiedshypothesen oder Korrelationskoeffizienten für Zusammenhangshypothesen *r*; wobei die ein Konstrukt operationalisierenden Items evtl. zuvor über eine *Faktorenanalyse* zusammengefasst wurden oder zumindest Cronbachs Alpha angegeben wurde, s. Kap. 4.1) und die zugehörigen Graphiken erstellt sind (z.B. Mittelwertsdiagramme wie in Abb. 3_9, 3_12b, 3_14-3_16), und die deskriptiven Daten mit der Hypothese (der H1) *konform* erscheinen (Mittelwerte sich in der vorhergesagten Richtung unterscheiden und $d \geq .20$, oder ein Korrelationskoeffizient das vorhergesagte Vorzeichen hat und $|r| \geq .10$), dann bleibt dennoch eine Unsicherheit darüber bestehen, ob das erzielte Ergebnis ausreichend signifikant ist, ´bedeutsam genug ist´, um die Nullhypothese zugunsten der H1 verwerfen zu dürfen. *Ein hypothesenkonformes* Ergebnis erreicht – nach der Signifikanz-Theorie von Fischer[4] - eine genügende ´Bedeutsamkeit´, also ***Signifikanz***, wenn es

[4] Es gibt eine konkurrierende Signifikanztest-Theorie von Neyman & Pearson (s. z.B. Hager 1987), die besser - aber für AnfängerInnen schwieriger ist. Sie basiert auf Tab. 4_4 und wurde für Tab. 4_6 benutzt.

nicht allein durch Zufall (also nicht ´in einer Welt, in der die (ungerichtete) Nullhypothese wahr wäre´) hätte entstehen können. Genauer: ein Ergebnis ist **signifikant**, wenn es allein durch Zufall nur sehr selten aufgetreten würde (daher darf dann hier, wo es auftrat, die ´H1-Welt´ angenommen werden ☺). Ob ein Ergebnis *auch vom Zufall* erzeugt worden sein könnte, hängt davon ab, wie ´stark´ es ist (*Effektstärke*, s. Kap. 4.1) und wie viele N (Personen oder ´Fälle´) untersucht wurden.

Übliche ***Signifikanztests*** berechnen also die *Wahrscheinlichkeit, mit der ein* (so schönes, wie das vorliegende) *Ergebnis bei gegebenem N auch in einer H0-Welt hätte zufällig auftreten können*. Kurz: ***ein Signifikanztest ermittelt die Zufallswahrscheinlichkeit***. Je kleiner die errechnete *Zufallswahrscheinlichkeit*, desto unwahrscheinlicher war das schöne Ergebnis (z.B. der starke Effekt) nur durch Zufall entstanden. Die mit dem *Signifikanztest* ermittelte ***Zufallswahrscheinlichkeit*** wird in Forschungsberichten durch die Abkürzung **p** (kleines p für Wahrscheinlichkeit) oder α (Alpha, Fehler erster Art, s. Tab. 4_4) bezeichnet. Die ForscherInnengemeinde in der Psychologie hat sich darauf geeinigt, eine Zufallswahrscheinlichkeit von 5%, also **p<.05** (oder α<.05 oder **α<5%**) als so gering anzusehen, dass der Zufall nicht für das Ergebnis verantwortlich sein dürfte. Im professionellen Text ausgedrückt wird die Angabe der Zufallswahrscheinlichkeit bescheiden in Klammern: „die Schuhgröße der N=70 untersuchten Männer ist mit ∅=44 *signifikant* größer als die von ∅=40 der N=70 Frauen *(p<.05)*“. Oder: „die Leistung korreliert mit der Erregung *signifikant* negativ (r = -.30, *p<.05*)“.

Tab. 4_2: Signifikanz erst durch genügend Fälle

Beispiel Münzwurf:

H1: Diese Münze ist gezinkt, sie wird auf eine der beiden Seiten häufiger fallen. (ungerichtete Unterschieds-Hypothese)

H0: Kein Wissen über die Münze, die Seiten werden gleichhäufig auftreten.

Prüfung: zwei Würfe.

Ergebnis: Zahl-Zahl.

Frage: Das Ergebnis sieht zunächst aus, als sei es mit der H1 konform, denn die relative Häufigkeit für Zahl und Bild entspricht nicht der unter H0 erwarteten (50:50). Aber ist das Ergebnis auch bedeutsam (=signifikant)?

Antwort-Weg: Es ist (nur) dann signifikant, wenn es in der H0-Welt (=normale Münze) zu unwahrscheinlich wäre. Bei nur zwei Würfen hat das Ergebnis Zahl-Zahl aber in der H0-Welt eine Wahrscheinlichkeit von 25%! Und die von ´Bild´-´Bild´ (die H1 war ja ungerichtet, ein Bild-Bild-Ergebnis wäre zunächst ja auch zugunsten der H1 interpretiert worden) noch mal von 25%. => Ein Ergebnis „gleiche Seite bei zwei Münzwürfen“ ist nicht signifikant, da es in der H0-Welt auch häufig vorkommt (eben sogar in der Hälfte aller Fälle!). Damit $2_*.50^n$ <.05, wären immerhin n=5 Würfe mit identischem Ausgang nötig. Erst nach fünf Würfen mit identischem Ergebnis darf die H0 (normale Münze) zugunsten der H1 (gezinkte) verworfen werden.

(Also sagt ein Fussballspiel erst beim 5::0 Ergebnis etwas über Leistungsunterschiede der Mannschaften aus ☺) Wer sich interessiert: hier ging es um den Binomialtest (s. Statistik-Buch).

Von der Konvention α<5% wird in manchen Fällen abgewichen (z.B. in angewandten Bereichen, wo selten genügend Personen untersucht werden (können), um bspw. mittelstarke Effekte noch ´signifikant zu bekommen´ - um für solche Forschungsbereiche die Inkaufnahme höherer Zufallswahrscheinlichkeit zu begründen, bspw. schon ab p<.10 zugunsten der H1 zu entscheiden, bedarf es der neuen Signifikanztest-Theorie, s. Fußnote 4).

Werden sehr viele Personen untersucht, etwa tausend bis zweitausend in sozialwissenschaftlichen Repräsentativ-Umfragen für Dt, so ist ein α<.05 unnütz, um über Hypothesen zu entscheiden (weil dann auch mini-Effekte von d<<.20 oder r<<.10 signifikant würden, obwohl sie keine praktische Bedeutsamkeit mehr besitzen). Bei großem N sollte man lieber auf die Effektgröße achten (s. Kap. 4:1, ein kleinerer als ein *schwacher Effekt* ist *kein Effekt!*). Fast alle hypothesen-prüfenden Studien in der Psychologie aber halten sich an die **5%-**Konvention.

Tab. 4_3: Merksatz Signifikanz (zum Sprachstil vgl. Tab. 3_3)

Ein empirisches Ergebnis ist ...	
signifikant, wenn das Ergebnis unter Gültigkeit der H0 zu unwahrscheinlich ist (p ≤ 5%): die Nullhypothese darf verworfen und die H1 darf daher angenommen werden.	**nicht signifikant**, wenn es auch unter Gültigkeit der H0 vorkommen kann (p>5%), also muss die H0 beibehalten werden.

Das Resultat eines *Signifikanztests* ist somit die Wahrscheinlichkeit, mit der ein (vielleicht zu-

nächst schön aussehendes) Ergebnis auch ´unter der H0´ (Slang für: unter Gültigkeit der H0-Hypothese, also „durch Zufall“) auftreten kann.

Da die meisten Signifikanztests nur die sogn. enge Nullhypothese (die Nullhypothese einer ungerichteten H1, z.B. ´es besteht kein Unterschied zwischen den Mittelwerten´, oder ´kein Zusammenhang zwischen den Variablen´) testen (und nicht eine vollständige Nullhypothese für gerichtete H1 behandeln können, s. zwote und vierte Nullhypothese in Tab. 3_2), sind die Ausgaben von Statistikprogrammen meist **zweiseitige Zufallswahrscheinlichkeiten**. Im Münzwurfbeispiel von Tab. 4_2 wird zum Ergebnis ´Zahl – Zahl´ die Zufallswahrscheinlichkeit p=.50 ausgegeben (*Binomialtest*), weil für eine normale (H0-) Münze, die bei vielen Würfen im Schnitt ausgeglichen auf beide Seiten fällt, die Hälfte aller ´nur-zwei-Würfe´-Versuche mit gleichen Seiten ausgeht (Zahl-Zahl oder Bild-Bild wird nicht unterschieden). Um die ungerichtete H1 „Münze gezinkt“ zu testen, ist diese zweiseitige Zufallswahrscheinlichkeit korrekt (die H0 muss beibehalten werden). Hatte man aber die gerichtete H1 „Münze zeigt häufiger Zahl“, und ist das Ergebnis zunächst konform (Zahl-Zahl), so darf man – bei manchen Signifikanztests (beim t-Test, bei r, bei der multiplen Regression, Statistikbuch lesen!) - die ausgegebene *zweiseitige Zufallswahrscheinlichkeit* halbieren zur **einseitigen Zufallswahrscheinlichkeit**. Wer sich die Mühe gemacht hat, eine einseitige Hypothese zu formulieren (s. Tab. 3_1), bekommt dies (bei manchen Signifikanztests, s. Tab. 4_6) dadurch belohnt, dass zum Hypothesenentscheid bei deskriptiv-konformem Ergebnismuster (Mittelwertsunterschied in der richtigen Richtung, Vorzeichen einer Korrelation erwartungskonform) die e*inseitigen Zufallswahrscheinlichkeit* verwendet werden darf, die ein $\alpha_{einseitig}$<5% schon mit weniger N erreicht (englisch $p_{one\text{-}tailed}$ vs. $p_{two\text{-}tailed}$, Abkz. p_{1t} / p_{2t}) Tab. 4_6 zeigt, dass etwa 20% der ProbandInnen gespart werden können, wenn die Hypothese einseitig formuliert worden war. Daher war in Kap. 3.1 verkündet worden: das Formulieren *einseitiger Hypothesen* zahlt sich aus!

Die Wahrscheinlichkeit, mit der ein bestimmtes Ergebnis unter der H0 auftritt, wird **Alpha-Fehler** (oder ´Fehler erster Art´, Alpha-Niveau, α, Zufallswahrscheinlichkeit, Zufallsniveau, Signifikanzniveau, Signifikanz, p-Wert) genannt. Warum ´Fehler´? Im Beispiel der nur zwei Münzwürfe wäre es ein Fehler, die Münze nach den zwei Würfen mit dem Zahl-Zahl – Ergebnis als gezinkt zu bezeichnen (*die H1 anzunehmen*), denn *unter der H0* ist ein solches Ergebnis ja auch sehr wahrscheinlich.

Tab. 4_4: Alpha-Fehler & Beta-Fehler beim Hypothesenentscheid (Synonym: Fehler 1.Art, Fehler 2 Art).

Das empirische Ergebnis spricht ..	..und/aber die H0 das Ergebnis erzeugte,	.. und/aber die H1 das Ergebnis erzeugte,
…für die H1. Wenn man die H1 dann annimmt....	.. hat man einen Alpha-Fehler (Fehler 1. Art) begangen	.. hat man das Ergebnis richtig interpretiert.
…für die H0. Wenn man die H0 also beibehält....	.. hat man das Ergebnis richtig interpretiert.	.. hat man einen Beta-Fehler (Fehler 2. Art) begangen

In Tab. 4_4 wird aufgezeigt, das bei jeder der beiden Entscheidungen (H1 annehmen oder H0 beibehalten; zur Sprachregelung siehe Tab. 3.3) ein Fehler gemacht werden kann: der Alpha-Fehler, wenn die H1 angenommen wird, obwohl eigentlich die H0-Welt gilt (die Münze weggeworfen wird, obwohl sie OK war). Der **Beta-Fehler** tritt auf, wenn die H0 beibehalten wird, obwohl die H1-Welt gilt (auch eine gezinkte Münze fällt bei so wenigen Würfen manchmal ganz ausgeglichen auf beide Seiten!).

Da wir kein Wissen darüber haben, ob tatsächlich die H0- oder die H1-Welt gilt (also welche Spalte von Tab. 4_4 die richtige ist), sondern wir aus der durchgeführten Studie nur ein empirisches Ergebnis (die Zeile in Tab. 4_4) erhalten und die Psychologie durch eine konservative Haltung (*konservativ* bedeutet in der Wissenschaft = im Zweifelsfall die H0 beibehalten! s. Kap. 1.1) versucht, die Inflation zu vieler und falscher Theorien zu verhindern (s. Kap. 1.1), ist der Alpha-Fehler klein zu halten (üblich ist alpha 5%, beta 10%, also beta/alpha=2). Im *Signifikanztest* wird nur die *Zufallswahrscheinlichkeit* für das erhaltene Ergebnis bei gegebener *Stichprobengröße* (Versuchspersonenanzahl) ausgerechnet. Die Zufallswahrscheinlichkeit wird klein (die H0 darf ab p<.05 verworfen werden), wenn viele

Versuchspersonen untersucht wurden und/oder wenn ein deutliches Ergebnis erzielt wurde (z.B. starke Effekte, s. Kap. 4.1). Die Anzahl von Versuchspersonen, die nötig sind, um eine bestimmte Korrelation oder einen bestimmten streuungsrelativierten Mittelwertsunterschied d (Formel für d in Abb. 4_3)´ nachweisen´ zu können (ihn als *überzufällig* auszuweisen), lässt sich ausrechnen (sog. *Power-Analyse*: je mehr Vp, desto geringer beide Fehler aus Tab. 4_4, je mehr Vp, desto größer die *Power des Designs*; die Power-Berechnung ist zwar nicht Bestandteil dieser Einführung – s. Fußnote 4 -, für Tab. 4_5 wurde sie, um grobe Orientierung zu geben, aber benutzt).

Tab. 4_5 zeigt Grössenordnungen der benötigten Versuchspersonenanzahl, um Unterschiede mittels *t-Test* (s. Tab. 4_6; Effektgrößen d=.20, .50, .80, s. Abb. 4_3) oder Zusammenhänge mittels Korrelation (Effektgrößen r=.10, .30, .50, s. Kap. 4_1) unter Einhaltung der für Hypothesenentscheide geforderten kleinen Entscheidungsfehler (Tab. 4_4) prüfen zu können.

Tab. 4_5: Anzahl benötigter Versuchspersonen (N), um einen Effekt bestimmter Stärke (d bzw. r, Cohen-Konventionen, Kap. 4_1) mit Alpha = 5% und Beta = 10% nachweisen zu können (berechnet mit GPOWER, Erdfelder et al. 1996)

Die Hypothese behauptet	Unterschied (per t-Test) zwischen				Zusammenhang (Korrelation r)	
	zwei Gruppen		zwei Messzeitpunkte			
	einseitige	zweiseitig	einseitig	zweiseitig	einseitig	zweiseitig
schwacher Effekt	429:429	527:527	430	528	850	1043
mittlerer Effekt	70:70	86:86	70	87	88	109
starker Effekt	28:28	34:34	29	35	28	34

Erkennbar wird, dass eigentlich nur sozialwissenschaftliche Bevölkerungsumfragen ausreichende Stichprobengrößen (N um die Tausend, obere Zeile von Tab. 4_6) realisieren, die nötig sind, um *schwache Effekte* gegen die Nullhypothesen zu prüfen. Die typische psychologische Untersuchung mit Grössenordnungen von 150–200 Vp. hat sich mit mindestens *mittleren Effekten* zu beschäftigen. Kleine Labor- oder Praxisstudien (N 30-60) sind zur Hypothesentestung nur geeignet, wenn ein *starker Effekt* erwartet wird (bei unvermeidbar kleinen Stichproben kann eine faire Adjustierung der Entscheidungsfehler in Tab. 4_4 dennoch Aussagen erlauben, dies aber gehört zu den *advanced statistics,* Hager 1987, Erdfelder et al. 1996).

Es gibt viele **Signifikanztests** – einfachere und kompliziertere. Je nach Datensituation (vor allem je nach *Skalenniveau der AV*, s. Tab. 3_14) sind bestimmte Signifikanztests angebracht. In Tab. 4_6 werden nur sieben, die wohl in unserem Fach häufigsten der ´einfacheren´ aufgezählt. Um geeignete Signifikanztest auswählen, berechnen und die Ergebnisse angemessen interpretieren zu können, sind mindestens Statistikbücher zu konsultieren (empfohlen wird: Hinzuziehen einer Expertin mit Hauptfach Psychologie ☺). Ähnlich zur Abb. 4_5 und Tab. 4_1 sollen die Namen in Tab. 4_6 eine Vorstellung von dem Raum üblicher Verfahren geben.

Schöner als eine tabellarische Auflistung wäre ein „System aller Inferenzstatistischen Verfahren“. Für einfachere bivariate Fragestellungen liegt ein Versuch von Blankenberger & Vorberg (1998), für übliche SPSS Verfahren einer von Groner (o.J.) sowie von Schwarz & Enzler (2016) vor. Die Skript-Verf. arbeitet ebenfalls an einem System - es ist für dieses Skript noch nicht verfügbar.

Tab. 4_6: Einfachere Signifikanztests

Häufigkeitsverteilung in einer Nominalvariable zufällig? Zusammenhang nominal skalierter Variablen? (z.B. Vier-Felder-Tafel)	Nach Anleitung aus einem Statistikbuch einen „**Chi-Quadrat-Test**“ berechnen. Ergebnis ist ein Chi-Quadrat -Wert $\chi 2$ mit Freiheitsgraden df und eine Signifikanz p dazu (Test ist immer zweiseitig). Zusätzlich: Nach Studium entsprechender Literatur läßt sich auch eine Effektgröße ω (omega) ausrechnen. Einseitige Tests erlaubt bspw. Binomialtest und Konfidenz für Prozente.
Korrelation r (z.B.: Ist die Korrelation von r=.2 zwischen AC-Leistung und Attraktivität aus Tab. 3_15 bedeutsam größer Null?)	Oft gibt das Programm, mit dem die **Korrelation** berechnet wurde, die Zufallswahrscheinlichkeit mit an (bspw. SPSS: Korrelation, bivariat). Falls nicht, kann man ´im Buch nachschlagen´ (Tabellen im Anhang der meisten Statistik-Bücher, wo für bestimmte r und N die p-Werte tabelliert sind). Ist die H1 gerichtet und das empirische Vorzeichen von r hypothesenkonform, darf die *einseitige Zufallswahrscheinlichkeit*, bei ungerichteter Hypothese oder diskonformem Vorzeichen muss die *zweiseitige Zufallswahrscheinlichkeit* verwendet werden (s.a.Tab. 4_5).
Mittelwertsunterschied für zwei (einigermaßen gleichgroße) Vp.Gruppen (bspw. Schuhgrößen-Geschlechtsunterschied Abb. 4_3 oder auch zu Abb. 3_14b)	Mit einem Statistikprogramm einen „**t-Test für unabhängige Daten**“ (Synonym: t-Test für zwei Gruppen, *between subjects, independent samples*) berechnen. Ergebnis ist ein t-Wert mit Freiheitsgraden (df=N-2) und Zufallswahrscheinlichkeit p. Für gerichtete H1 und hypothesenkonformes Ergebnis darf man die ´zweiseitige Zufallswahrscheinlichkeit´ halbieren, um die *einseitige Zufallswahrscheinlichkeit* zum Hypothesenentscheid zu verwenden (s.a. Tab. 4_5). Und: Effektgröße d ausrechnen (Abb. 4_3) (per Hand bzw. Taschenrechner!).
Mittelwertsunterschied für zwei Messungen an den gleichen Vp (z.B. Prä-Post-Design von Abb. 3_14c)	Mit einem Statistikprogramm einen „**t-Test für abhängige Daten**“ (Synonym: t-Test für gepaarten Stichproben, *pared samples*, *within subjects*) berechnen. Ergebnis ist ein t-Wert mit Freiheitsgraden (df=N-1) und Zufallswahrscheinlichkeit p. Für gerichtete H1 und hypothesenkonformes Ergebnis darf man die ´zweiseitige Zufallswahrscheinlichkeit´ halbieren, um die *einseitige* zum Hypothesenentscheid zu verwenden (s.a. Tab. 4_6). Und: Effektgröße d_w ausrechnen (s. Fußnote 3).
Mittelwertsunterschied für $\geq$ 2 Faktorstufen (z.B. Personengruppen) oder mehrfaktorielle univariate Designs	Mit einem Statistikprogramm eine **Varianzanalyse** rechnen (ANOVA = analysis of variance, ´Allgemeines lineares Modell - Univariat´, *general linear model* GLM). Ergebnis ist ein (o. mehrere) F-Wert(e) mit Freiheitsgraden und Signifikanz p (Varianzanalysen bewerten H0 ungerichteter H1, p darf nicht halbiert werden).
Mittelwertsunterschied für mehrere Messungen in min. einer Gruppe (z.B. Prä-Post-Kontrollgruppen-Design wie in Abb. 3_14d).	Mit Statistikprogramm eine **Messwiederholungs-Varianzanalyse** rechnen (´Allgemeines lineares Modell / general linear model GLM - Messwiederholungen´). Ergebnis sind je nach Faktorenanzahl mehrere F-Werte, z.B. für Haupteffekte u. Interaktionen, s. Abb.3_9) mit Freiheitsgraden und Zufallswahrscheinlichkeit p (Varianzanalysen bewerten H0 von ungerichteten H1, p darf nicht halbiert werden).
Mittelwertsunterschied(e) für apriori ungleich großen Gruppen (oder empirisch deutlich verschieden großen) Wahrscheinlichkeitsunterschiede bei nominalen AV, Streuungsunterschieden, Korrelationsunterschieden oder Regressionskoeffizientenunterschieden verschiedener Gruppen, u.v.a.	**Konfidenzintervall** (´Vertrauensintervall´): für Mittelwerte bei intervallskalierten AV (und für viele statistische Parameter, s. links) können Intervalle berechnet werden (von ... bis ...), die angeben, wo der Parameter mit einer bestimmten Wahrscheinlichkeit (=Sicherheit = Konfidenz) wohl liegen wird, wenn statt der kleinen Stichprobe die ganze Population untersucht worden wäre. Das Intervall wird größer (das in der Stichprobe erhaltene Ergebnis unsicherer), wenn wenig N untersucht wurden und wenn man hohe Sicherheit haben will, dass der Populationsparameter doch nicht ausserhalb des Intervalls liegt. Will man zu 90% sicher sein, beträgt der Alpha-Fehler 10%, für ein 95%-Konfidenzintervall beträgt er 5%. Bei einseitiger Hypothese interessiert nur eine Seite des Intervalls (also für einseitiges Alpha 5% das 90%-Konfidenzintervall ☺). Einige Verfahren geben zu einem statistische Parameter dessen Standardfehler (standard error, SE) aus. Nimmt man ihn mal 1.98 erhält man die halbe Konfidenzintervallbreite für 95% (mal 1.68, die für 90% also für p-einseitig-5%)

Zusammenfassend liegt die Leistung des Signifikanz-Testens darin, Ergebnisse von zu kleinen Stichproben nicht vorschnell zur Ablehnung der Nullhypothese und Bewährung einer vielleicht doch falschen Theorie zu missbrauchen. Gerade psychologische Theorien - mit dem ihnen entgegengebrachten grossen Alltagsinteresse (vgl. Kap. 1) - sollten nur dann zum Gütekriterium der *empirischen Bewährung* gelangen (s. Kap. 5), wenn es ihnen gelingt, die nicht-wissende „H0-Welt“ deutlich (eben mit *Signifikanz*) zurückweisen zu können.

5. Was ist eine gute Theorie?

Nicht nur für empirische Untersuchungen (s. Tab. 3_19), sondern auch für *Theorien* gibt es[5] Gütekriterien. Das (in einer empirischen Wissenschaft wie der Psychologie) wichtigste Kriterium fordert, dass sich Vorhersagen der Theorie empirisch *bewähren* (Nullhypothesen verworfen und die aus der Theorie abgeleiteten *H1* angenommen werden durften, siehe Nummer 3 in Tab. 5_1). Mit ihrer *Bewährung* erhält die Theorie ihren *empirischen Begründungszusammenhang* (vgl. Kap. 1.2). Die empirische *Bewährung* einer Theorie kritisch beurteilen und evtl. durch Konzeption neuer Studien mit besserer *interner Validität* zum Ausschluss falscher Theorien beitragen zu können, waren Lernziele dieses Skripts.

Notwendige Voraussetzungen für die empirische Bewährung sind, dass die Theorie widerspruchsfrei formuliert ist (Kriterium 1 in Tab. 5_1, dies gilt auch für geisteswissenschaftliche, z.B. mathematische Theorien) und dass sich aus ihr ableitbare Vorhersagen prinzipiell als falsch erweisen k ö n n t e n (Kriterium 2 in Tab. 5_1). Eine Theorie, die Tautologien enthält (tautologischer Satz: „kräht der Hahn auf dem Mist, ändert sich´s Wetter oder es bleibt wie es ist"), besitzt *keinen empirischen Gehalt* (Tautologien können prinzipiell einfach nie falsch sein. Dass sie immer zutreffen, liefert daher keinerlei Information!). Theorien mit *hohem empirischen Gehalt* erkennt man daran, dass sie (vor dem bisherigen Wissensstand) ´riskante´, evtl. provokative, jedenfalls nicht-triviale Vorhersagen erlauben. Wenn diese *gehaltvollen* Hypothesen angenommen werden können, ist die Theorie wertvoller als eine mit nur trivialen Hypothesen.

Tab. 5_1: Gütekriterien für Theorien

1) begriffliche & logische *Widerspruchsfreiheit*
2) *Überprüfbarkeit, empirischer Gehalt*
3) *Empirische Bewährung*
4) Kriterien für weitere Theorienkonkurrenz:
4.1) *Sparsamkeit*, einfach, *Parsimonie*
4.2) *Präzision*, Genauigkeit (der Vorhersagen)
4.3) *Allgemeinheit* des Geltungsbereichs

Wenn nun mehrere Theorien zum selben Phänomenbereich vorliegen, die alle *widerspruchsfrei* formuliert sind, *empirischen Gehalt* besitzen und sich bisher ähnlich gut *empirisch bewährt* haben, stehen die Theorien in Konkurrenz. Hier versucht die Forschung aus den *konkurrierenden Theorien* konkurrierende (also sich widersprechende) Vorhersagen abzuleiten. Dies ist selten einfach, da begrifflich verschiedene Theorien oft doch nur die gleichen Vorhersagen zulassen (bspw. lässt sich die gerichtete Zusammenhangshypothese „Ähnlichkeit schafft Sympathie" sowohl aus der Austauschtheorie mit ihrem behavioristischen Hintergrund vorhersagen als auch aus der Balancetheorie mit ihrem kognitiv-konsistenztheoretischen Hintergrund). Solange keine einander widersprechenden Vorhersagen aus konkurrierenden Theorien abgeleitet werden und daher kein *Entscheidungsexperiment* stattfinden konnte, gilt diejenige Theorie als besser, die *sparsam(er)* und/oder *präzise(r)* und/oder *allgemein(er)* ist (Nr. 4 in Tab. 5_1).

Sparsam ist eine Theorie, wenn sie nur wenige *Konstrukte* (s. Kap.2; wenige Behauptungen, wenige Ausnahmen etc) enthält. Zajoncs Theorie der sozialen Erleichterung kommt mit zwei Konstrukten (Aufgabenschwierigkeit und Erregung, s. Abb.2_2) aus, neuere Theorien zur Wirkung der Anwesenheit anderer benötigen mehr Konstrukte. Da der Philosoph Okham im 14. Jhd. erfolgreich für die Sparsamkeit von Theorien argumentierte, nennt man das Sparsamkeits-Argument gegen Theorien mit zu vielen Annahmen auch „**Okhams Rasiermesser**".

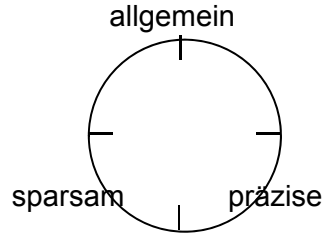


Abb. 5_1: Die Thorngate-Uhr zum Dilemma der Güte von Theorien

Präzision besitzt eine Theorie, deren Vorhersagen sehr genau sind, die bspw. nicht nur aussagt, dass die Leistung bei schwierigen Aufgaben irgendwie mit der

[5] Mit der Bewertung von Theorien beschäftigt sich neben der jeweiligen Fachwissenschaft auch die der Philosophie zugehörige *Wissenschaftstheorie* (dabei beschreibt und erklärt die *deskriptive Wissenschaftstheorie* wie die real existierende Wissenschaft so funktioniert (z.B. dass soziale Moden leider über die Theorieentwicklung mitbestimmen, Kuhn), während die *normative Wissenschaftstheorie* Soll-Aussagen formuliert, also empfiehlt, wie gute Wissenschaft vorgehen soll). Der für die Psychologie einflussreichste Vertreter normativer Wissenschaftstheorie war Karl Popper, ihm verdanken wir das Falsifikations-Ideal.

Erregung fällt (wie Abb. 2_2), sondern die sagen kann, ob sie linear oder quadratisch oder nach welcher Formel eben – oder mit welcher Effektgröße (einem schwachen, mittleren oder starker Effekt?, s. Kap. 4.1) fällt. Physikalische Theorien, die in Form von Formeln vorgestellt werden (z.B. Fallgesetz) sind präziser als (fast) alle psychologischen Theorien (gute Formeln hat die Psychophysik). Präzise Vorhersagen erfordern u.a. die Festlegung des *Skalenniveaus* der Variablen bereits in der Theorie (zwischen einem linearen, beschleunigten oder verlangsamten Abfall der Leistung kann nur entscheiden werden, wenn beide Variablen, Leistung und Erregung, mindestens *intervallskaliert* sind, s. Kap. 3.2.1). Über präzise Vorhersagen kann die Theorie besonders streng geprüft werden (ihr *empirischer Gehalt* wächst). Wenn die Psychologie beginnt, Theorien zu entwickeln, deren Vorhersagen zumindest eine Aussage über die erwarteten Effektgrößen machen, gewinnt sie an Präzision.

Der *Allgemeinheitsgrad* einer Theorie ist hoch, wenn sie in breiten Lebensbereichen Geltung beansprucht und nicht nur in einem sehr engen. Zajoncs Theorie wurde u.a. gewählt, weil Anwesenheit anderer sehr breit gefasst ist (z.B. Großraumbüro vs. virtuelle Teams …) und Vorhersagen sowohl für einfache als auch schwierige Aufgaben intendiert sind.

Wie an der Ausführung zu Sparsamkeit, Präzision und Allgemeinheitsgrad implizit bereits deutlich geworden sein dürfte, stehen auch diese drei Gütekriterien in einer dilemmatischen Beziehung zueinander (Thorngate 1976, z.n. Weick 1995; s. Abb. 5_1): einige Theorien haben zwar gleichzeitig zwei der drei Gütekriterien optimiert (in der vereinfachenden Form von Abb. 2_2 ist Zajoncs Theorie *sparsam* und *allgemein* aber nicht *präzise*, liegt also in Abb. 5_1 oben links); psychologische Theorien können wohl nur selten gleichzeitig sparsam, präzise und allgemein sein!

Nichtsdestotrotz können wir an der Optimierung der Kriterien (aller sechs aus Tab. 5_1) und damit an der Annäherung an ideale Theorien arbeiten. Die Verbreitung von Grundkenntnissen in der Methodik will hierzu bescheiden betragen.

Referenzen

Atwater, L.E., Waldman, D.A., Atwater, D. & Cartier, P. (2000). An Upward-feedback field experiment: supervisors' cynic, reactions, and commitment to subordinates. *Personnel Psychology, 53*, 275-297.

Blankenbeger, S. & Vorberg, D (1999). Die Auswahl statistischer Tests und Maße. Psychologische Rundschau, 50 157-164. http://css-kti.tugraz.at/research/cssarchive/courses/fm2/Entscheidungsbaum.pdf, Web-Umsetzung Lohninger, H. (2012): http://www.statistics4u.info/fundstat_germ/ee_baum_root.html

Borg, I. (1995). *Mitarbeiterbefragungen. Strategisches Auftau- & Einbindungsmanagement*. Göttingen: Verlag für Angewandte Psychologie.

DGPs & BDP (1998). Ethische Richtlinien der Deutschen Gesellschaft für Psychologie und des Bunds Deutscher Psychologen. Erreichbar unter: http://www.dgps.de/dgps/kommissionen/ethik/003.php4

Erdfelder, E.; Faul, F. & Buchner, A. (1996): GPOWER: A general power analysis program. Behavior Research Methods, Instruments, & Computers 28, 1-11. Siehe auch http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3.

Fassheber & Gennerich (2001). Theorien und Konstrukte der Wirtschafts- und Sozialpsychologie. Skript zur Mo-Vorlesung. Erreichbar unter: http://www.psych.uni-goettingen.de/abt/5-alt/lehre/lehrmaterial_n.html

Friedrichs, J. (1973). *Methoden empirischer Sozialforschung*. Hamburg: Rowohlt.

Groner, M. (o.J.). Entscheidungsbaum statistischer Testverfahren. http://etools.fernuni.ch/entscheidungsbaum/entscheidungsbaum.pdf (l.v. IV ,18).

Hager, W. (1987). Grundlagen einer Versuchsplanung zur Prüfung empirischer Hypothesen in der Psychologie. In: Luer, G. (Hrsg). Allgemeine Experimentelle Psychologie (43-264). Stuttgart: Fischer.

Schwarz & Enzler (2016). Datenanalyse mit Entscheidungs-Assistenten http://www.methodenberatung.uzh.ch/de/datenanalyse.html

Weick, K.E. (1995): Der Prozess des Organisierens. Frankfurt: Suhrkamp.

Abbildungsverzeichnis

Tabellenverzeichnis

Kommentare an: mstrack@uni-goettingen.de