2. Testkonstruktion

Dieses Kapitel befaßt sich mit Fragen und Problemen der Entwicklung und Konstruktion von Fragebögen und Testinstrumenten. Die Erörterung von Fragen der Testkonstruktion wendet sich dabei nicht allein an diejenigen LeserInnen, die tatsächlich selbst einen *Test entwickeln* möchten (obwohl das im Rahmen vieler Diplomarbeiten im Fach Psychologie der Fall ist). Die Darstellung soll vielmehr auch dem Verständnis und der kritischen Beurteilung *existierender Verfahren* dienen.

Das Gliederungsprinzip des Kapitels ergibt sich aus den Phasen, die bei einer Testentwicklung zu durchlaufen sind. Kapitel 2.1 befaßt sich mit den Gütekriterien für Tests, d.h. mit der Frage, wodurch sich ein 'guter' Test auszeichnet. Kapitel 2.2 beschreibt die *Schritte*, die idealerweise durchlaufen werden sollten, wenn man ein neues Testinstrument konstruiert. Kapitel 2.3 geht dann konkret auf Fragen der *Itemkonstruktion* ein, d.h. auf die praktische Seite der Testentwicklung. Nach der Darstellung von Problemen der *Datenerhebung* in Kapitel 2.4 beschäftigt sich Kapitel 2.5 abschließend mit der *Kodierung* der Testantworten, d.h. mit der Transformation der Itemantworten in Zahlen, auf die man Testmodelle anwenden kann.

2.1 Gütekriterien für Tests

Wenn man einen Test konstruieren will, muß man eine Vorstellung davon haben, was einen 'guten' Test auszeichnet. Das ist die Frage nach den sogenannten *Gütekriterien* für Tests. Klassischerweise werden hier drei Gütekriterien genannt, nämlich Objektivität, Reliabilität und Validität, die jeder Test zu einem Mindestausmaß zu erfüllen hat.

Mit *Objektivität* ist gemeint, inwieweit das Testergebnis unabhängig ist von jeglichen Einflüssen außerhalb der getesteten Person, also vom Versuchsleiter, der Art der Auswertung, den situationalen Bedingungen, der Zufallsauswahl, von den Testitems usw. Es ist ersichtlich, daß es sehr viele verschiedene Arten von Objektivität bei Tests zu unterscheiden gilt.

Mit *Reliabilität* (Zuverlässigkeit) ist das Ausmaß gemeint, wie genau der Test das mißt, was er mißt (egal, was er mißt). Es ist hier lediglich die *Meßgenauigkeit*, die numerische Präzision der Messung angesprochen, unabhängig davon, was der Test überhaupt mißt. Als Meßgenauigkeit wird dabei nicht die Zahl der Dezimalstellen der Meßwerte bezeichnet, sondern die Zuverlässigkeit, mit der bei einer wiederholten Messung unter gleichen Bedingungen dasselbe Meßergebnis herauskommt.

Mit *Validität* ist gemeint, inwieweit der Test das mißt, was er messen soll. Es geht also um den Grad der *Gültigkeit* der Messung oder der Aussagefähigkeit des Testergebnisses bezüglich der Meßintention.

Diese *klassische Trias* von Testgütekriterien entstammt einer testtheoretischen Tradition, die die Auswertung von Tests noch *nicht* aus dem Blickwinkel der *Anwendung eines Testmodells* sah. Trotzdem lassen sich die drei Konzepte der Objektivität, Reliabilität und Validität weiterhin zur Beschreibung, der Güte eines Tests verwenden.

Alle drei Gütekriterien haben verschiedene Teilaspekte, und es gibt für jedes auch

verschiedene Arten, es zu operationalisieren und in konkrete Zahlen zu fassen. Die konzeptuellen Ausdifferenzierungen werden in den folgenden drei Unterkapiteln beschrieben, die konkreten Berechnungsmöglichkeiten erst in Kapitel 6 (Testoptimierung). Kapitel 2.1.4 geht auf *die logischen Beziehungen* ein, die zwischen diesen Konzepten bestehen. Kapitel 2.1.5 behandelt schließlich ein weiteres Gütekriterium, nämlich die *Normierung*. Geordnet sind die Kapitel nach der Wichtigkeit der Gütekriterien, beginnend mit dem wichtigsten, der Validität.

2.1.1 Validität

Unter der Validität eines Test versteht man das *Ausmaß, in dem der Test das mißt, was er messen soll.*

Wie beurteilt man aber, inwieweit der Test mißt, was er messen soll? Hier gibt es prinzipiell zwei Möglichkeiten. Die eine Möglichkeit setzt voraus, daß eine *andere Messung* dessen, was der Test messen soll, *verfügbar* ist. In diesem Fall braucht man nur an einer Stichprobe von Personen beide Arten der Messung vorzunehmen und zu prüfen ob die Ergebnisse bei allen Personen übereinstimmen.

Beispiel

Man möchte einen besonders ökonomischen (kurzen) Intelligenztest entwickeln und hat zufällig eine Stichprobe von Personen zur Verfügung, die schon vor einiger Zeit hinsichtlich ihrer Intelligenz untersucht worden sind und deren Intelligenzgrad daher bekannt ist. Diesen Personen gibt man dann auch den Kurztest vor. Die *Korrelation* zwischen beiden Meßwertreihen ist dann ein Maß für die Validität des Kurztests.

Nach dieser Möglichkeit entspricht die Validität eines Tests der Korrelation des Testergebnisses mit einer anderen Variable, die eine Messung desselben Merkmals darstellt.

Was ist eine Korrelation?

Als Korrelation bezeichnet man den Zusammenhang zwischen zwei quantitativen Variablen. Es handelt sich also um eine spezielle Art der Kontingenz (s. Kap. 1.2.4). Die Höhe der Korrelation, also die Stärke des Zusammenhangs wird durch den Korrelationskoeffizienten ausgedruckt.

Dieser kann Werte zwischen -1 und +l annehmen, wobei eine Korrelation von 0 bedeutet, daß zwischen den beiden Variablen kein Zusammenhang besteht. Eine negative Korrelation bedeutet, daß hohe Werte auf der einen Variable mit niedrigen Werten auf der anderen Variable einhergehen, während eine positive Korrelation bedeutet, daß hohe Werte auf beiden Variablen bzw. niedrige Werte auf beiden Variablen miteinander gepaart sind.

Der Korrelationskoeffizient zwischen zwei Variablen X und Y wird folgendermaßen berechnet

$$\mathrm{Korr}(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}}.$$

Die Kovarianz im Zähler ist das durchschnittliche Produkt der Abweichungen beider Meßwerte von ihrem jeweiligen Mittelwert,

$$\bar{x}=\frac{1}{N}\sum_{v=1}^{N} x_v\,,\ \text{bzw.}\ \bar{y}=\frac{1}{N}\sum_{v=1}^{N} y_v$$

$$\mathrm{Cov}(X,Y)=\frac{1}{N}\cdot\sum_{v=1}^{N}\left(x_v-\bar{x}\right)\cdot\left(y_v-\bar{y}\right).$$

N bezeichnet die Anzahl und v den Summationsindex der Personen.

Im Nenner des Korrelationskoeffizienten steht die Wurzel aus dem Produkt beider Varianzen, wobei die *Varianz* einer Variable X definiert ist als die durchschnittliche quadrierte Abweichung aller Meßwerte von ihrem Mittelwert:

$$\mathrm{Var}(X) = \frac{1}{N} \sum_{v=1}^{N} \left(x_v - \bar{x}\right)^2 .$$

Diese Art der Validität nennt man *externe Validität,* weil ihre Bestimmung anhand der Testergebnisse eine externe (d.h. außerhalb des Tests liegende) Variable voraussetzt, und zwar genau die Variable, die der Test erfassen will.

Die zweite Möglichkeit zu prüfen, ob der Test das mißt, was er messen soll, benutzt allein die Daten, die aufgrund der Testdurchführung vorliegen. Hier wird geprüft, ob die Personen auf die Items so antworten, wie man es aufgrund der Theorie über die zu messende Personeneigenschaft erwarten würde. Dabei wird natürlich *nicht* vorausgesetzt, daß man die Ausprägungen der Personeneigenschaft bereits kennt.

Beispiel

Ein Steinzeitmensch möchte die Muskelkraft seiner Stammesgenossen testen und sucht sich hierfür eine Reihe unterschiedlich großer Steine und Felsbrocken zusammen. Die Größe der Steine ist unterschiedlich genug, so daß man sie 'mit dem Auge' der Größe nach ordnen und (sofern die Zahlen schon erfunden sind) der Größe nach durchnumerieren kann. Jeder Stammesgenosse muß versuchen, alle Steine anzuheben und die Nummer des größten Steins, den er anheben kann, ist der Meßwert für seine Muskelkraft.

Aus der Steinzeittheorie über die Personeneigenschaft 'Muskelkraft' folgt, daß jede Person alle Steine bis zu einer Größe, die ihrer Kraft entspricht, anheben kann. Beobachtet der Steinzeitmensch nun, daß jede Person alle Steine, die kleiner sind als der größte, den sie heben kann, auch anheben kann, *so* ist der Test *intern valide.*

Dies ist nur *ein* Beispiel für einen möglichen Zusammenhang zwischen der Personeneigenschaft und dem Testverhalten. Es macht deutlich, daß der Begriff der *internen Validität* gleichzusetzen ist mit der Gültigkeit des jeweils zugrunde gelegten Testmodells.

> Ein Test heißt *intern valide,* wenn sich die Annahmen über das Antwortverhalten anhand der Datenmatrix bestätigen lassen.

Je strenger die Annahmen über das Antwortverhalten, desto überzeugender läßt sich die interne Validität eines Tests nachweisen. Während man zum Nachweis der externen Validität ein *Validitätskriterium* braucht (so nennt man die externe Variable, die das repräsentiert, was der Test messen soll), erfordert der Nachweis der internen Validität *präexperimentelle Annahmen* über das Antwortverhalten bei den einzelnen Items.

Beide Aspekte der Validität *bedingen sich nicht* unbedingt gegenseitig. So kann es z.B. sein, daß mit einem intern validen Test, für den das angenommene Testmodell sehr gut paßt, eine Variable gemessen wird, die keinerlei Erklärungswert für das sonstige Verhalten der Personen hat. Genauso kann irgendein Testergebnis einen guten Vorhersagewert für bestimmte an-

dere Variablen besitzen, ohne daß man eine Theorie über die Itemantworten hat.

Es wird deutlich, daß interne und externe Validität zwei sehr unterschiedliche Seiten desselben Gütekriteriums sind. Während die Prüfung der internen Validität ein zentrales Thema der Testtheorie darstellt, überschreitet die Frage der *externen Validität* den Bereich der Testtheorie. Ob ein Test extern valide ist, kann mit den üblichen Methoden statistischer Datenanalyse untersucht werden. Kapitel 6.4 geht auf die einfachsten Arten der Validitätsberechnung ein und beschreibt den Einfluß der Meßgenauigkeit eines Tests auf die Höhe der errechneten externen Validität.

Ganz anders verhält es sich mit der *internen Validität* eines Tests. Sie ist abzulesen an der Geltung des jeweiligen Testmodells für den Datensatz der Testentwicklung. Hier gibt es jedoch andere Probleme. Ob ein Testmodell gilt oder nicht gilt, ist oft keine Ja-Nein-Entscheidung, sondern kann durchaus ein graduelles Urteil sein, d.h. ein Testmodell kann mehr oder weniger gut passen. Oft ist die Entscheidung, ob ein Testmodell paßt oder nicht, auch *nur relativ zu anderen Testmodellen* zu beantworten, d.h. es läßt sich lediglich sagen, ob ein bestimmtes Testmodell besser paßt als ein bestimmtes Vergleichsmodell. In diesen Fällen gibt es nicht mal mehr eine quantitative Aussage, wie gut ein Modell paßt, sondern lediglich eine relative Aussage, die davon abhängt, welche Vergleichsmodelle man überhaupt geprüft hat. Hierauf wird in Kapitel 5 im Detail eingegangen.

2.1.2 Reliabilität und Meßgenauigkeit

Die Reliabilität oder zu deutsch die *Zuverlässigkeit* eines Tests bezeichnet die Präzision oder Genauigkeit, mit der ein Test eine Personeneigenschaft mißt. *Reliabilität im engeren Sinne* meint jedoch eine bestimmte Definition von *Meßgenauigkeit*, die nicht die einzig mögliche ist und auch nicht bei jedem Testmodell Sinn macht. Um diese Definition verständlich zu machen, wird zunächst dargestellt, was ein *Meßfehler* ist.

Angenommen, man hat bei einer Anzahl von N Personen intervallskalierte Meßwerte erhoben. Der Meßwert einer Person v wird mit x_v bezeichnet und stellt den *sog. beobachteten Wert* dar.

Von diesem Meßwert nimmt man an, daß er die 'wirkliche' Eigenschaftsausprägung ziemlich genau, aber nie 'ganz genau' widerspiegelt. Die hypothetische wirkliche Eigenschaftsausprägung einer Person wird mit t_v bezeichnet (t wie true = wahr) und stellt den sog. *wahren Wert* dar.

Den kleinen Betrag, um den der beobachtete Wert von dem wahren Wert abweicht, nennt man *Meßfehler* und bezeichnet ihn mit e_v (e wie error = Fehler). Aus diesen Überlegungen ergibt sich die *Grundgleichung der Meßfehlertheorie:*

$$x_v = t_v + e_v$$

Da die Grundgleichung der Meßfehlertheorie für alle Personen v gelten soll, läßt sie sich auch als Beziehung zwischen den Variablen schreiben:

$$X = T + E$$

X bezeichnet die Meßwertvariable, T die Variable der wahren Werte und E die Fehlervariable.

Anmerkung
Variablen werden hier und im folgenden mit großen lateinischen Buchstaben bezeichnet, ihre Ausprägungen mit den zugehörigen Kleinbuchstaben.

Diese Gleichungen zerlegen das beobachtete Testergebnis x_v in zwei Komponenten, t_v und e_v, die man beide nicht kennt. Über den wahren Wert kann man nicht viel sagen, aber einen Meßfehler zeichnen zwei Eigenschaften aus:

Erstens zeichnet sich ein Meßfehler dadurch aus, daß er dazu beiträgt, den wahren Wert manchmal zu überschätzen, und manchmal zu unterschätzen. Genau das unterscheidet einen Meßfehler von einem systematischen Fehler, auch *'Bias'* genannt: er ist im Mittel 'neutral'. Mit anderen Worten, der Mittelwert oder *Erwartungswert der Meßfehlervariable E* über eine große Anzahl von Personen *ist 0:*

$$\text{Erw}(E) = 0. \qquad \text{(I)}$$

Was ist ein Erwartungswert?

Der Erwartungswert ist eine Kenngröße einer numerischen Variable. Treten die Werte x einer Variable X mit der Wahrscheinlichkeit p(x) auf, so ist der Erwartungswert definiert durch

$$\text{Erw}(X) = \sum_x x\, p(x).$$

Summiert wird hier über alle möglichen Werte der Variable X. In dem Beispiel

x	1	2	3
p(x)	0.5	0.3	0.2

beträgt der Erwartungswert $1 \cdot 0.5 + 2 \cdot 0.3 + 3 \cdot 0.2 = 1.7$. Kennt man nicht die Wahrscheinlichkeitsverteilung einer Variable, also die Werte p(x), sondern hat man N Ausprägungen der Variable x beobachtet, so entspricht der Mittelwert dieser Werte

$$\overline{X} = \frac{1}{N} \sum_{v=1}^{N} x_v$$

näherungsweise dem Erwartungswert von X. Hat man in dem o.g. Beispiel von 10 Beobachtungen 5-mal die 1, 3-mal die 2 und 2-mal die 3 beobachtet, so beträgt der Mittelwert von X ebenfalls $x = 1.7$.

Zweitens gehört zum Konzept eines Meßfehlers, daß er *nicht mit dem wahren Wert korreliert* ist, d.h. es darf nicht sein, daß z.B. hohe wahre Werte überschätzt werden und niedrige wahre Werte unterschätzt werden. In einem solchen Fall würde man ebenfalls von einem systematischen Fehler oder Bias sprechen. Meßfehler zeichnen sich daher auch dadurch aus, daß:

$$\text{Korr}(E,T) = 0. \qquad \text{(II)}$$

Überhaupt gehört zum Konzept eines Meßfehlers dazu, daß er *mit keiner anderen Variable* korreliert, also nicht mit den wahren Werten einer Variable Y:

$$\text{Korr}(E_x,T_y) = 0, \qquad \text{(III)}$$

und auch nicht mit deren Meßfehler E,:

$$\text{Korr}(E_x,E_y) = 0. \qquad \text{(IV)}$$

Diese vier Gleichungen (1) bis (IV) nennt man auch die *Axiome der klassischen Testtheorie.* Sie wurden von Gulliksen (1950) formuliert und beschreiben nichts anderes als die *Eigenschaften eines Meß-*

fehlers. Aus ihnen ist keine Testtheorie im Sinne von Kapitel 1 ableitbar, sondern nur eine Theorie über das Verhalten des Meßfehlers. Sie wird daher im folgenden als *Meßfehlertheorie* bezeichnet.

Im Unterschied zu einer Testtheorie, die sich auf nominale oder ordinale Itemantworten bezieht, geht die Meßfehlertheorie von fertigen *Meßwerten* X und Y aus. Die vier Gleichungen (1) bis (IV) stellen genauso wie Testmodelle ein *formales Modell* dar, nur bezieht sich dieses formale Modell auf eine *andere Datenstruktur.*

Aus dem formalen Modell wird (wie immer) eine Theorie, wenn man es auf einen konkreten Inhalt (konkrete Meßwerte X und Y) anwendet und die Parameter des Modells schätzt (S.O. Kap. 1). In diesem Fall macht die Meßfehlertheorie die Aussage, daß ein Meßwert *X nur eine latente Variable* T_x repräsentiert und die Abweichungen der Meßwerte von den wahren Variablenausprägungen den *Erwartungswert 0 haben und mit nichts anderem korrelieren.*

Diese Theorie ist z.B. dann *falsifiziert,* wenn sich herausstellt, daß der (vermeintliche) Meßfehler die soziale Erwünschtheit beinhaltet (S.O. Kap. 1). Dann gibt es nämlich eine Variable, die mit dem Meßfehler korreliert, nämlich die Tendenz der Personen, sozial erwünscht zu antworten.

Wie im Fall von Testmodellen, so gibt es auch bei Meßfehlermodellen nicht nur *ein* Modell. Vielmehr entstehen durch verschiedene Zusatzannahmen und Erweiterungen viele *Meßfehlermodelle,* die hinsichtlich ihrer Gültigkeit miteinander verglichen werden können. Auf diese unterschiedlichen Meßfehlermodelle wird in Kapitel 3.1.1.2.1 kurz eingegangen.

Testmodelle und Meßfehlermodelle schließen sich also nicht gegenseitig aus, sondern sie ergänzen einander:

Testmodelle wendet man auf Itemantworten an, um daraus Meßwerte zu machen, Meßfehlermodelle wendet man auf die erhaltenen Meßwerte an, um deren Fehleranteil zu bestimmen.

Die gerade dargestellte Meßfehlertheorie bezieht sich ausschließlich auf *quantitative, mindestens intervall-skalierte Personenvariablen.* Das gleiche gilt für die folgende Definition der Reliabilität eines Tests. Nach dieser Definition ist *Reliabilität* ein Varianzanteil, nämlich das Verhältnis von wahrer Varianz zu beobachteter Varianz.

$$\text{Reliabilität} = \frac{\text{wahre Varianz} \quad \text{Var}(T_x)}{\text{beobachtete Varianz} = \text{Var}(X)}$$

Mit *wahrer Varianz* bezeichnet man die Varianz der nicht beobachtbaren, imaginären wahren Testergebnisse und mit *beobachteter Varianz* die Varianz der tatsächlich in einem Test erhaltenen Ergebnisse.

Aus den Annahmen der Meßfehlertheorie folgt, daß die wahre Varianz stets kleiner ist als die beobachtete Varianz.

Beweis

Die Varianz der Summe zweier Zufallsvariablen X und Y läßt sich nach der Formel berechnen:

$\text{Var}(X+Y)=\text{Var}(X)+\text{Var}(Y)+2\text{Cov}(X,Y)$.

Mit Cov(X,Y) wird die Kovarianz von X und Y bezeichnet (s. Kap. 2.1.1). Diese ist

definitionsgemäß gleich 0, wenn die Korrelation von X und Y gleich 0 ist.

Die Grundgleichung der Meßfehlertheorie zerlegt den Meßwert in die Summe zweier *unkorrelierter* Variablen (laut Axiom II):

$$X = T_x + E_x,$$

deren Varianzen sich folglich auch addieren:

$$Var(X)=Var(T_x)+Var(E_x),$$

Da Varianzen stets positiv sind, ist auch die wahre Varianz stets kleiner als die beobachtete Varianz:

$$Var(T_x) < Var(X).$$

Dies mag auf den ersten Blick nicht einleuchten, kann man sich doch z.B. folgenden Fall vorstellen:

x	5	10	15	20	25
T_x	2	8	15	22	28

Hier wäre die wahre Varianz größer als die beobachtete. Bei näherer Betrachtung sieht man jedoch, daß das Axiom II verletzt ist, denn die Fehlervariable, im Beispiel:

Ex	3	2	0	-2	-3

ist natürlich hoch (negativ) mit dem wahren Wert korreliert.

Bei Geltung der Voraussetzungen ist tatsächlich die wahre Varianz stets kleiner als die beobachtete Varianz, so daß nach der obigen Definition die Reliabilität eines Tests stets zwischen Null und Eins liegt:

$$0 < Rel. < 1.$$

Dieses Maß für die Meßgenauigkeit eines Tests gibt an, welcher *Anteil an der Varianz der Meßwerte* wirklich auf Personenunterschiede zurückgeht und ist als Varianzanteil daher ähnlich interpretierbar wie das Quadrat eines Korrelationskoeffizienten (= Anteil gemeinsamer Varianz) oder ein Erblichkeitsindex (= Anteil der erblich bedingten Varianz).

Wie man die Reliabilität eines Tests konkret berechnet, wird in Kapitel 6 beschrieben.

Soviel zu dem klassischen Reliabilitätsbegriff, der nur *eine* Art der Definition der Meßgenauigkeit von Tests darstellt. Für Tests mit einer *kategorialen Personenvariable* gibt es keine vergleichbare einheitliche Definition der Meßgenauigkeit. Hier kann sich eine hohe Meßgenauigkeit z.B. darin ausdrucken, daß die Anzahl der 'Fehlklassifikationen' der Personen zu den Valenzen der kategorialen Personenvariable sehr gering ist (s. Kap. 6).

2.1.3 Objektivität

Wenn ein Testergebnis nicht unabhängig vom Testleiter, von Situationsmerkmalen, von störenden Randbedingungen, vom Testauswerter oder sonstigen Personen ist, so wird der Test auch keine interne Validität und keine besonders hohe Meßgenauigkeit erlangen können. Insofern ist *Objektivität* der Testdurchführung eine *logische Voraussetzung für Reliabilität und Validität.* Eine hohe Objektivität bei der Testentwicklung zu erreichen, ist somit kein Selbstzweck im Sinne eines positivistischen Wissenschaftsbegriffes, sondern lediglich Mittel, um Genauigkeit und Validität zu erreichen.

Im einzelnen ist bei der Testentwicklung anzustreben, daß das Testergebnis unabhängig davon ist,

- wer den Test vorgibt *(Durchführungsobjektivität),*

- wer den Test auswertet *(Auswertungsobjektivität)* und
- wer den Test interpretiert *(Interpretationsobjektivität).*

Zusätzlich gibt es verschiedene Unterformen dieser Objektivitätsaspekte wie z.B. die *Signierobjektivität*, die sich auf die Objektivität bei der Kodierung freier Antworten bezieht. Sie ist ein Teilaspekt der Auswertungsobjektivität.

Neben diesen Objektivitätsaspekten, die sich auf die Unabhängigkeit von anderen *Personen* beziehen, gibt es auch die Objektivität im Sinne einer Unabhängigkeit von anderen *Dingen.* So sollte das Testergebnis z.B. weitgehend davon unabhängig sein, in welcher *Situation* der Test durchgeführt wurde. Damit kann natürlich nur eine Unabhängigkeit innerhalb eines Spektrums 'normaler' Situationen gemeint sein.

Vermutet man eine starke Situationsabhängigkeit des Testergebnisses und hält deshalb die Testsituation konstant, indem man den Test quasi unter Laborbedingungen durchführt, so hat das unterschiedliche Auswirkungen auf die *interne und externe Validität* des Testergebnisses. Während die interne Validität sogar steigen kann, je stärker man die situationalen Bedingungen konstant hält (da ein bestimmtes Testmodell unter Idealbedingungen vielleicht besser paßt), dürfte die externe Validität im allgemeinen sinken: Wenn ein Testergebnis nicht mal auf andere Testsituationen *generalisierbar* ist, so wird auch seine Korrelation mit externen Variablen nicht hoch sein.

Hier zeigt sich die semantische Ähnlichkeit des Begriffspaares 'interne und externe Validität' mit den gleichlautenden Begriffen aus der *Versuchsplanung* besonders deutlich: bei Experimenten bezeichnet man als externe Validität ebenfalls die Aussagekraft und die Generalisierbarkeit des Ergebnisses über die Experimentalsituation hinaus.

Ebenfalls eine Objektivität im Sinne einer Unabhängigkeit von anderen Dingen ist mit dem Begriff der *spezifischen Objektivität* gemeint. Spezifische Objektivität bezeichnet die Unabhängigkeit eines Testergebnisses von der Itemauswahl aus einem hypothetischen Item-Universum. Dahinter steht die Überlegung, daß jeder Test nur eine sehr begrenzte Anzahl von Items umfassen kann, das Testergebnis aber nicht nur etwas über die Fähigkeit zur Beantwortung *dieser* Items aussagen soll, sondern über die Fähigkeit zur Beantwortung *dieses Typs von Items.*

Eine Eigenschaftsmessung bezieht sich also immer auf ein ganzes Itemuniversum, das unendlich viele Items umfaßt. Ein wichtiger Objektivitätsaspekt ist daher mit der Frage angesprochen, ob bei jeder beliebigen Itemauswahl stets dasselbe Testergebnis (abgesehen vom Meßfehler) herauskommt.

Diese sogenannte spezifische Objektivität ist nicht nur eine Eigenschaft eines Tests, sondern auch des jeweiligen *Testmodells:* Bei den meisten der in Kapitel 3 behandelten quantitativen Testmodelle sind die Meßwerte spezifisch objektiv, sofern das Modell für die Daten gültig ist.

2.1.4 Logische Beziehungen zwischen den drei Gütekriterien

Zwischen den drei Gütekriterien eines Tests bestehen verschiedene *logische Beziehungen,* die sich unter bestimmten mathematischen Annahmen sogar in Formeln beschreiben lassen. Und zwar ist die Objektivität eine logische Voraussetzung für die Reliabilität und diese wiederum ist logische Voraussetzung für die externe Validität:

Validität (ext.)
↑
Reliabilität
↑
Objektivität

Ein Test, der bei einem anderen Testleiter oder in einem anderen Raum bei denselben Personen gänzlich andere Resultate erbringt, also nicht objektiv ist, kann auch keine hohe Meßgenauigkeit haben, d.h. nicht reliabel sein.

Ebenso kann ein Test mit einer sehr geringen Meßgenauigkeit (Reliabilität) keine besonders hohe *externe Validität* erreichen. Soll z.B. ein Test entwickelt werden, der die Schulleistung vorherzusagen gestattet, so kann diese Vorhersage nicht besonders gut ausfallen, wenn der Test nur sehr ungenau mißt.

Eine solche Voraussetzungsbeziehung besteht nicht zwischen der Meßgenauigkeit und der *internen Validität.* Auch ein ungenau messender Test kann intern valide sein.

Andererseits besteht zwischen Meßgenauigkeit, interner und externer Validität auch ein *kontradiktorisches Verhältnis:*

Das Streben nach einer möglichst hohen Meßgenauigkeit bei der Testentwicklung kann in einem Widerspruch stehen zum Ziel einer möglichst hohen Validität. Dieser Widerspruch ergibt sich daraus, daß sich die Meßgenauigkeit im allgemeinen dadurch steigern läßt, daß man den *Test verlängert,* d.h. zusätzliche Items aufnimmt (s. Kap. 6.1.2).

Durch eine *Testverlängerung,* die den Test rein theoretisch beliebig genau machen könnte, können Items hineinkommen, die einen etwas anderen Aspekt der latenten Variable ansprechen, es können Bearbeitungseffekte wie Ermüdung, Konzentrationsmangel, Wechsel der Antwortstrategie, Erinnerungseffekte, Lerneffekte und ähnliches eintreten. Diese Effekte können sowohl die präexperimentelle Theorie über das Antwortverhalten, d.h. das Testmodell, in seiner Gültigkeit einschränken, als auch die Korrelation mit einem Validitätskriterium, also die externe Validität, beeinträchtigen.

Auch die Ziele einer möglichst hohen internen und externen Validität können bei der Testentwicklung miteinander in einem Konflikt stehen. So läßt sich die interne Validität im allgemeinen dadurch steigern, daß man den Test *homogener* macht, d.h. möglichst ähnliche Aufgaben auswählt. Damit erfaßt man aber eine sehr eng gefaßte, spezielle Personeneigenschaft, die nur noch geringe Korrelationen mit einem Validitätskriterium aufweist.

Die immanenten Widersprüche zwischen Reliabilität und Validität werden auch als *Reliabilitäts-Validitäts-Dilemma* der Testtheorie bezeichnet (s. Kap. 6.4.3). Dieses Dilemma ist letztlich Ursache für den weitverbreiteten Argwohn, daß Tests entweder mit einer hohen Präzision etwas

völlig Irrelevantes messen, oder eine Personeneigenschaft in ihrer ganzen Breite, aber völlig unzuverlässig erfassen.

2.1.5 Normierung

Neben den drei klassischen Gütekriterien gibt es das mehr pragmatische Kriterium der Normierung. Dieses betrifft die Frage, inwieweit es für die Ergebnisse eines Tests *Vergleichsdaten* gibt, anhand derer sich Einzelergebnisse interpretieren lassen. Solche Vergleichsdaten, die an repräsentativen Stichproben verschiedener Teilpopulationen der Bevölkerung erhoben worden sind, bilden dann die *Normen,* anhand derer das Ergebnis einer einzelnen Person beurteilt und interpretiert werden kann.

Interpretiert man ein Ergebnis indem man es mit der Norm einer Referenzpopulation vergleicht, so spricht man auch von *normorientiertem Testen.* Das Gegenstück hierzu ist das sogenannte *kriteriumsorientierte Testen.* Hier wird das einzelne Testergebnis nicht über den Vergleich mit den Werten einer Referenzpopulation interpretiert, sondern anhand eines inhaltlichen, vorher vom Testkonstrukteur gesetzten Kriteriums.

Ein prominentes Beispiel für diesen Unterschied ist die *Zensurenvergabe* in der Schule. Ein normorientiertes Vorgehen besteht darin, daß zunächst für jeden Schüler Punkte vergeben werden, um dann anhand der Punkteverteilung eine Notenzuordnung durchzuführen. Diese soll sicherstellen, daß auf jeden Fall ein paar Einsen und ein paar Fünfen dabei sind.

Beispiel:

Punkte in Klausur

2 2 3	5 7 7	9 11 12 14 16	20 20	25
5	4	3	2	1

Note

Eine solche Zensur ist für den Schüler normorientiert, d.h. sie informiert ihn nur über seine *relative Stellung* in der Klasse, seiner Referenzpopulation, aber nicht relativ zum Leistungsziel.

Bei einer *kriteriumsorientierten* Zensurenvergabe würde der Lehrer *vorher* festlegen, bei welcher Punktzahl es eine Eins, eine Zwei usw. gibt. Die resultierenden Zensuren sagen etwas darüber aus, wie der einzelne Schüler zum gesteckten Leistungsziel, dem Kriterium, steht, aber nicht unbedingt, wie er im Vergleich zu den anderen Schülern dasteht.

Wie bei diesem Beispiel der Schulleistungsbewertung gibt es generell bei psychologischen Tests die Alternative zwischen einer Normorientierung und einer Kriteriumsorientierung.

Soll ein Test später für individual diagnostische Zwecke eingesetzt werden, so sind im allgemeinen Normtabellen sehr hilfreich, da sie über die Verteilung der Testergebnisse in verschiedenen Referenzpopulationen Aufschluß geben, z.B. in Altersgruppen, Geschlechtsgruppen oder nach der Schulbildung definierten Gruppen. Von daher wird die *Normierung* eines Tests im allgemeinen *als ein Gütekriterium* angesehen, da diese die Interpretation erleichtert und in einem gewissen Sinne auch objektiver macht (objektiv in dem Sinne der Unabhängigkeit von der subjektiven Setzung eines inhaltlichen Kriteriums).

Dennoch ist die Normierung eines Tests kein für alle Zwecke einer Testentwicklung sinnvolles Gütekriterium. Auch bei individualdiagnostischen Fragestellungen kann eine rein kriteriumsorientierte Interpretation des Testergebnisses wesentlich

sinnvoller sein, z.B. bei der Frage, ob ein bestimmtes Krankheitsbild vorliegt oder nicht. In diesem Falle ist vom Testkonstrukteur nicht eine Normierung des Tests vorzunehmen, es muß vielmehr ein inhaltliches Kriterium oder mehrere solcher *Kriterien für die Interpretation* der Testresultate bereitgestellt werden.

Für viele Zwecke der Testentwicklung stellt eine Normierung keine Notwendigkeit und auch kein Gütekriterium dar. Geht es z.B. darum, im Rahmen von *Forschungsarbeiten* eine Personenvariable mit einem Test zu messen, um sie mit anderen Variablen in Beziehung zu setzen, so ist eine Normierung der Testresultate überflüssig: Sollen etwa zwei verschiedene Personengruppen hinsichtlich ihres Ergebnisses in einem Test miteinander verglichen werden oder soll eine quantitative Personenvariable mit einer anderen Persönlichkeitsvariable wie Extraversion oder Intelligenz korreliert werden, so ist es völlig unerheblich, ob ein Test normiert ist oder nicht. Eine Normierung schlägt sich weder in Mittelwertsdifferenzen noch in Korrelationen nieder.

Das Gütekriterium der Normierung wird oft *überbewertet,* d.h. man begeht leicht den Fehlschluß anzunehmen, daß ein normierter Test auch etwas Sinnvolles mißt. Das kann, muß aber nicht der Fall sein: Das Gütekriterium der Normierung steht in keinerlei logischer Beziehung zu den anderen drei Gütekriterien der Objektivität, Meßgenauigkeit und Validität. Auch ein wenig objektiver, wenig reliabler und wenig valider Test läßt sich einer repräsentativen Bevölkerungsstichprobe vorgeben und an ihr normieren (s. Kap. 6.5).

Literatur

Die Gütekriterien für Tests werden von Lienert und Raatz (1994) aber auch in den meisten Diagnostik-Lehrbüchern (z.B. Guthke, Böttcher & Sprung 1990) behandelt. Fischer (1974) und Steyer & Eid (1993) führen aus, daß die Axiome der Meßfehlertheorie keine Axiomatik im mathematischen Sinne darstellen. Das Konzept der spezifischen Objektivität wird von Rasch (1977) und Fischer (1987) diskutiert. Die Abhängigkeit der Testergebnisse von Situationen wird von Eid (1995) systematisch in die Formalisierung von Testmodellen einbezogen.

Übungsaufgaben

1. Sie haben die Meßwerte einer Variablen X, die wahren Werte derselben Variable, T_x, und die Meßwerte eines Validitätskriteriums Y von 5 Personen:

	Personen				
	1	2	3	4	5
T_x:	1	1	5	9	9
x :	2	0	5	10	8
Y:	3	4	4	4	5

   Prüfen Sie, ob für den Meßfehler der Meßwerte X die ersten beiden Axiome der Meßfehlertheorie gelten. Berechnen Sie die Reliabilität und die externe Validität von X.

2. Nennen Sie 5 möglichst unterschiedliche Faktoren, die die Objektivität eines Tests beeinträchtigen können.

3. Ein Schüler hat in einer Klausur 84% aller gestellten Aufgaben richtig gelöst Ermöglicht dieses Ergebnis bereits eine normorientierte oder kriteriumsorientierte Interpretation? Welche Zusatzinformation benötigt man, um das Ergebnis normorientiert oder kriteriumsorientiert interpretieren zu können?

2.2 Schritte der Testentwicklung

Jede Testentwicklung nimmt ihren Ausgangspunkt in einer Theorie über die Personeneigenschaft, die der Test erfassen soll. Eine solche Theorie ist oft sehr wenig präzise und muß hinsichtlich verschiedener Aspekte konkretisiert werden, um Grundlage einer Testentwicklung sein zu können.

Idealerweise findet diese Präzisierung in fünf Schritten statt:

- *Erstens* muß man sich darüber klar werden, welcher Art die *Personenvariable* überhaupt ist.
- *Zweitens* kann man sich darüber Gedanken machen, über welche Art von Testverhalten man diese Personeneigenschaft am besten erfassen könnte.
- *Drittens* sollte man den Typ von Items, die den gewünschten Schluß vom Testverhalten auf die Personeneigenschaft zulassen, als Itemuniversum formulieren.
- Den *vierten* Schritt stellt die Auswahl einer geeigneten Itemstichprobe aus diesem Universum dar.
- Schließlich sollte man sich *fünftens* auch schon vor der Testkonstruktion Gedanken über das Testmodell machen, das auf diese Daten passen soll.

Diese Punkte werden in den folgenden Unterkapiteln abgehandelt.

2.2.1 Arten von latenten Variablen

Aus der Theorie sollte ableitbar sein, ob die zu testende Personeneigenschaft *quantitativer Natur oder qualitativer Natur* ist.

Mit quantitativer Natur ist gemeint, daß sich die Personen hinsichtlich eines 'mehr oder weniger' voneinander unterscheiden, das zu testende Personenmerkmal also graduelle Abstufungen annimmt.

Mit qualitativer Natur ist gemeint, daß Personenunterschiede getestet werden sollen, die sich darin ausdrucken, daß sich *Gruppen von Personen* qualitativ voneinander unterscheiden. Das zu messende Personenmerkmal ist dann lediglich *nominal skaliert.*

Weiterhin sollte aus der Theorie ableitbar sein, ob es sich um ein *univariates oder ein multivariates* Persönlichkeitsmerkmal handelt. Univariat bedeutet, daß nur *eine* Variable variiert, multivariat heißt ein Merkmal, das sich nur mit Hilfe von *mehreren* Variablen beschreiben läßt. Im Fall von mehreren quantitativen Personeneigenschaften spricht man auch von einer *mehrdimensionalen* Personenvariable.

Ein Beispiel ist das Konstrukt *Ängstlichkeit*, das sich als eine mehrdimensionale Variable definieren läßt. Die einzelnen Dimensionen ergeben sich aus den Gegenstandsbereichen, in denen sich Ängstlichkeit manifestiert, also z.B. Angst vor physischer Verletzung, Angst vor sozialer Ablehnung, Angst vor medizinischer Behandlung etc.

Auch bei *kategorialen* oder qualitativen Eigenschaften gibt es *multivariate Konzeptionen.* Ein Beispiel hierfür ist die Messung des Attributionsstils, welcher als eine bivariate Personeneigenschaft aufgefaßt werden kann:

Die erste kategoriale Personenvariable unterscheidet, ob die Person primär intern oder primär extern attribuiert ('es liegt

alles an mir' oder 'es lag an den äußeren Umständen'). Die zweite kategoriale Personenvariable unterscheidet stabile versus labile Attributionen ('das ist immer so' oder 'in diesem einzelnen Fall war das so').

Es gibt also sowohl bei kategorialen als auch bei quantitativen Personenvariablen univariate und multivariate Konzeptionen von Personeneigenschaften. Sind die Personenvariablen *kategorial, so* läßt sich aus ihnen eine einzelne Variable konstruieren, die als Kategorien die möglichen Kombinationen der Kategorien der Ausgangsvariablen hat. Im obigen Beispiel würde man also eine latente Variable mit vier Ausprägungen bilden:

intern - labil
intern - stabil
extern - labil
extern - stabil

Sofern die Variablen *quantitativ* sind, es sich also um eine *mehrdimensionale* Variable handelt, sind die möglichen Implikationen für die Testentwicklung vielfältig.

Der einfachste Fall besteht darin, daß man versucht, die verschiedenen Dimensionen *mit unterschiedlichen Items* zu erfassen. Im oben genannten Beispiel eines Angstfragebogens konstruiert man also Fragen zur Angst vor physischer Verletzung, zur Angst vor sozialer Ablehnung etc. In diesem Fall kann man jede Teilmenge von Items, jeden sogenannten Subtest, als eigenständigen Test konstruieren und auswerten. Anschließend können die Zusammenhänge der Meßwerte auf den verschiedenen Dimensionen analysiert werden.

Komplizierter ist der Fall, daß *dieselben Items* mehrere Dimensionen ansprechen. Z.B. wird die Beantwortung der Frage:

Wie unangenehm ist es Ihnen, sich nachts in einem Gasthaus in einer fremden Gegend nach dem Weg erkundigen zu müssen?

sowohl von der Angst vor physischer Verletzung, als auch von der Angst vor sozialer Ablehnung beeinflußt sein. Generell ist von der Konstruktion derartiger mehrdimensionaler Tests abzuraten, obwohl es Testmodelle gibt, mit denen man auch solche Tests auswerten kann (s. z.B. Kap. 3.4.2).

Der dritte Fall mehrdimensionaler Tests besteht darin, daß man nicht die Items sondern die *Antwortkategorien* danach unterscheidet, welche Dimension sie ansprechen. Ein Beispiel ist die Frage:

Was würden Sie heute abend um liebsten unternehmen?

- ins Theater gehen
- Freunde besuchen
- gutes Essen zubereiten

Die Auswahl der Antwort wird in diesem Beispiel von drei Dimensionen des Freizeitinteresses bestimmt: das Interesse an kulturellen Aktivitäten, an sozialen Aktivitäten und an gestaltenden Beschäftigungen. Diese Art der Erfassung mehrdimensionaler Eigenschaft birgt gewisse Schwierigkeiten, auf die im Kapitel 3.2.2 eingegangen wird, und kommt in der Praxis selten vor. Trotzdem gibt es auch für diesen Fall geeignete Testmodelle (Kap. 3.2.2).

Schließlich gibt es einen speziellen Fall der *Kombination* einer kategorialen und einer quantitativen Personenvariable. Dieser ist dann gegeben, wenn eine *quantitative Personenvariable* gemessen werden soll, aber damit zu rechnen ist, daß *verschiedene Personengruppen* diesen

Test auf unterschiedliche Art und Weise bearbeiten.

Beispiel

Die Messung des *räumlichen Vorstellungsvermögens* ist zweifellos ein Beispiel für die Messung einer quantitativen Personenvariable. Es folgt allerdings aus der Theorie, daß es zwei unterschiedliche Arten von Lösungsstrategien für die Testitems gibt, nämlich eine analytische und eine holistische Strategie. Weiterhin wird angenommen, daß jede Person eine dieser beiden Strategien bevorzugt und daher auch einen Raumvorstellungstest primär mit der von ihr bevorzugten Strategie löst. In diesem Falle gibt es eine kategoriale Personenvariable (holistische versus analytische Strategiepräferenz) und eine quantitative Personenvariable, nämlich die Fähigkeit, mit der jeweiligen Strategie Raumvorstellungsaufgaben zu lösen.

Auch für diesen Spezialfall einer Kombination von kategorialer und quantitativer Personenvariable gibt es spezielle Testmodelle, die in Kapitel 3.1.3 und 3.3.5 behandelt werden.

Die Klärung der Frage, welcher Art die latente Variable ist, die der Test erfassen soll (kategorial oder quantitativ, ein- oder mehrdimensional) stellt deswegen den ersten Planungsschritt bei der Testentwicklung dar, weil die Beantwortung dieser Frage weitgehend von der psychologischen *Theorie* über die betreffende Persönlichkeitseigenschaft bestimmt sein sollte. Die konkreten Implikationen für die Testkonstruktion ergeben sich aber erst aus der Kenntnis der Testmodelle, die man für den jeweiligen Zweck heranziehen kann.

2.2.2 Arten von Tests

Hat man sich Klarheit darüber verschafft, welcher Art die zu messende Personeneigenschaft ist, so stellt sich als nächstes die Frage, welcher Art das zu beobachtende *Testverhalten* ist, und wie es mit der Personeneigenschaft zusammenhängt. Je nach der Art des im Test erfaßten *Verhaltens* lassen sich folgende Arten von Tests unterscheiden:

Leistungstests
Persönlichkeitsfragebögen
objektive Persönlichkeitstests
Projektive Tests
Situationsfragebögen
Symptomlisten
Einstellungstests
Motivations- und Interessensfragebögen
Verhaltensfragebögen

Im folgenden soll das Charakteristische der Beziehung zwischen Personeneigenschaft und Testverhalten bei diesen Testarten dargestellt werden.

2.2.2.1 Leistungstests

Leistungstests zeichnen sich dadurch aus, daß von den Personen die Lösung von Aufgaben oder Problemen verlangt wird, die Reproduktion von Wissen, das Unterbeweisstellen von Können, Ausdauer oder Konzentrationsfähigkeit. So heterogen diese Aufzählung klingen mag, Leistungstests haben die wichtige Eigenschaft gemeinsam, daß die getesteten Personen das Ergebnis willentlich *nur in einer Richtung verfälschen* können, nämlich 'nach unten'. Man kann sich 'dümmer' stellen als man ist, man kann sich keine Mühe geben bei der Testbearbeitung oder die Antworten einfach zu erraten versuchen. Man kann

aber nicht in Verfälschungsabsicht eine höhere Leistung erbringen als die, zu der man imstande ist.

Leistungstests sind daher schon von vornherein als 'halb-objektiv' zu bezeichnen, obwohl die Verfälschungsmöglichkeit 'nach unten' aufgrund mangelnder Testmotivation, z.B. bei Felduntersuchungen, sehr gravierende Einschränkungen der Interpretierbarkeit der Ergebnisse mit sich bringen kann. Das Phänomen des *Erratens* der richtigen Lösung kann mit Mitteln der Itemkonstruktion eingeschränkt und mit geeigneten Testmodellen kontrolliert werden.

Innerhalb der Kategorie der Leistungstests gibt es eine weitere Unterteilung in sogenannte *speed- und power-Tests.* Bei speed-Tests wird durch eine begrenzte Zeitvorgabe neben der Qualität der Leistung auch die Geschwindigkeit erhoben, mit der eine Leistung erbracht wird. Bei power-Tests zählt dagegen nur, ob die Aufgaben richtig oder falsch gelöst wurden, und nicht wieviel Zeit die Person dafür benötigt.

Reine *power-Tests* sind schon aus technischen Gründen kaum durchführbar, da jede Testvorgabe eine zeitliche Begrenzung haben muß. Diese Grenze sollte aber so bemessen sein, daß in der Regel alle Personen bis zur letzten Testaufgabe vordringen. Nur in diesem Fall lassen sich die meisten Testmodelle auf die resultierenden Daten anwenden: die unterschiedliche Anzahl von nicht bearbeiteten Aufgaben wirft rechnerische Probleme, vor allem aber auch Interpretationsprobleme auf.

Relativiert man die erbrachte Leistung an der Zahl der *bearbeiteten Aufgaben, so* bewertet man die langsamen Personen zu gut, da sie sich für jede Aufgabe mehr Zeit als die schnellen Personen genommen haben. Relativiert man die Leistung an der *Gesamtzahl der angebotenen Aufgaben, so* bewertet man die Qualität der Leistung von langsamen Personen zu schlecht, da man nicht berücksichtigt, wieviele der nicht bearbeiteten Aufgaben sie noch hätten lösen können.

Auf jeden Fall stellt die Verquickung von Qualität und Geschwindigkeit bei 'gespeedeten' Leistungstests ein Problem dar, für dessen Lösung es zwar einige Ansätze in der Testtheorie gibt, von denen aber keiner ganz befriedigend ist. Günstiger ist *es,* die *Bearbeitungszeit pro Aufgabe zu* begrenzen. Hier hat jede Person dieselben Bedingungen für jede Aufgabe und es lassen sich die meisten Testmodelle problemlos anwenden.

Für einige Varianten von *Speed-Tests* benötigt man allerdings auch keine Testmodelle. Mißt man etwa die Zeit, die eine Person für eine vorgegebene Menge von Aufgaben benötigt, so hat man mit der gemessenen Zeitdauer bereits eine metrische Personenvariable. Um den Qualitätsaspekt aus dieser Zeitmessung ganz zu eliminieren, kann man falsch gelöste Aufgaben wiederholt vorlegen (z.B. beim computerunterstützten Testen), so daß die Zeit für die *richtige Lösung aller Aufgaben* gemessen wird.

Mit einer gewissen Berechtigung läßt sich bei Vorgabe eines *festen Zeitintervalls* auch die Anzahl der richtig gelösten Aufgaben als eine Häufigkeit, und somit als eine metrische Variable auffassen und man kann ebenfalls von der Anwendung eines Testmodells absehen.

Spielt also die Geschwindigkeit eines Verhaltens die zentrale Rolle bei der Messung einer Personeneigenschaft, so läßt sich die physikalische Größe 'Zeit' auch zur (metrischen) Operationalisierung dieser Personeneigenschaft nutzen.

2.2.2.2 Persönlichkeitsfragehögen

Persönlichkeitsfragebögen sind dadurch charakterisiert, daß von der befragten Personen eine *Selbstauskunft* (self report) verlangt wird. Fragen wie

> *Sorgen Sie sich um schreckliche Dinge, die vielleicht geschehen könnten ?*
> *ja - nein*
> *(Item aus dem EPI, Eggert 1974)*

stellen verschiedene Anforderungen an den Beantwortungsprozeß, wenn der Schluß von der Itemantwort auf die Personeneigenschaft (hier: Neurotizismus) gerechtfertigt sein soll.

Zunächst einmal muß die erfragte Selbstkenntnis vorhanden sein, d.h. die Person muß *wissen*, ob sie sich um schreckliche Dinge sorgt. Dies ist ein Aspekt der *Metakognition* (das ist die Einsicht in eigene kognitive Prozesse) der befragten Person, die durchaus nicht immer vorhanden sein muß. Ist diese Metakognition nicht vorhanden, oder entspricht sie ganz und gar nicht der Realität, so kann man von der Itemantwort bestenfalls auf das *Selbstbild* der Person, aber nicht auf ihre Persönlichkeit schließen. Beispiel: eine Person meint, sie mache sich Sorgen über schreckliche Dinge, sorgt sich aber tatsächlich nur darum, daß das Geld nicht bis zum Monatsende reichen könnte.

Sodann muß eine *Offenbarungsbereitschaft* vorhanden sein, d.h. die Person muß bereit sein, gemäß ihrer Metakognition zu antworten. Es kann z.B. sein, daß eine Person zwar die Bereitschaft hat, den Test auszufüllen, aber ihr *Ideal-Selbstbild* anstelle des Real-Selbstbildes wiedergibt. Mit Ideal-Selbstbild ist hier dasjenige Selbstbild gemeint, das die Person gerne gegenüber demjenigen, der den Test vorgibt, zeichnen möchte.

Man hat den Tendenzen, sich in einem Persönlichkeitsfragebogen anders darzustellen als man wirklich ist, verschiedene Namen gegeben: Beantwortet eine Person die Fragen so, daß ein positives, in unserer Gesellschaft allgemein akzeptiertes Bild entsteht, so beeinflußt die Variable der *sozialen Erwünschtheit* ihr Antwortverhalten. Eine zweite Variable, die die Ehrlichkeit der Antworten in einem Persönlichkeitsfragebogen beeinflußt, ist die Tendenz zur *Selbstpräsentation (self monitoring)*. Personen, die sich stets der jeweiligen Situation angepaßt darstellen, also die Eigenschaft eines Chamäleons haben, tun dies eventuell auch bei der Beantwortung eines Persönlichkeitsfragebogens.

Ist die Metakognition und die Offenbarungsbereitschaft gegeben, so ist als drittes eine geeigneter *Beurteilungsmaßstab* vorauszusetzen, der Daten *aus sozialen Vergleichsprozessen* erfordert. So beinhaltet z.B. die Frage, ob man sich Sorgen um schreckliche Dinge macht, auch den Aspekt, ob die befragte Person das häufiger oder intensiver tut als andere Personen. Dies setzt bei der Beantwortung der Frage voraus, daß die Person es einschätzen kann, inwieweit sich *andere* Menschen Sorgen um schreckliche Dinge machen.

Wird eine Frage *ohne* einen solchen Beurteilungsmaßstab beantwortet, so sagt die Antwort zwar auch etwas über die Person aus (nämlich, daß sie *meint*, daß sie sich mehr als andere Personen Sorgen macht). Sie sagt dann aber weniger über den 'tatsächlichen' Neurotizismusgrad der Person aus, sondern vielleicht etwas über ihren Leidensdruck oder ihren Glauben, daß es anderen Leuten besser geht als ihr.

Neben diesen drei Voraussetzungen für eine brauchbare Selbstauskunft ist ein weiteres Charakteristikum von Persönlichkeitsfragebögen ihre *Durchschaubarkeit*. Jugendliche und Erwachsene mit einem gewissen psychologischen Reflexionsniveau werden bei vielen Fragen aus Persönlichkeitsfragebögen durchaus richtig raten, auf welche Personeneigenschaften aus der Antwort geschlossen werden soll.

Diese Durchschaubarkeit beinhaltet eine leichte *Verfälschbarkeit* im Sinne einer gezielten Beeinflussung des gesamten Testresultates. Im Gegensatz zu Leistungstests, kann diese Beeinflussung in beide Richtungen gehen, z.B. kann man sich aufgrund der Durchschaubarkeit bewußt neurotischer oder weniger neurotisch darstellen.

2.2.2.3 Objektive Persönlichkeitstests

Dieser Begriff geht auf den Persönlichkeitsforscher R.B. Cattell zurück. Dieser forderte als Ergänzung und Kontrolle von Persönlichkeitsfragebögen noch eine zweite Art von Tests, die er objektive Persönlichkeitstests nannte. Sie sind in dem Sinne *objektiv*, als eine Verfälschung wegen Undurchschaubarkeit ausgeschlossen sein soll.

Schmidt (1975, S. 19) definiert objektive Tests folgendermaßen:

> *'Objektive Tests zur Messung der Persönlichkeit und Motivation sind Verfahren, die unmittelbar das Verhalten eines Individuums in einer standardisierten Situation erfassen, ohne daß diese sich in der Regel selbst beurteilen muß. Die Verfahren sollen für den Probanden keine mit der Meßintention übereinstimmende Augenscheinvalidität haben.'*

Unter *Augenscheinvalidität* versteht man die Eigenschaft von Tests, daß man ihnen 'ansieht' was sie messen sollen und welches Verhalten man damit vorhersagen möchte.

Die Idee objektiver Persönlichkeitstests besteht also darin, aus Itemantworten auf Personeneigenschaften zu schließen, die gar nicht Gegenstand der Fragen waren. Z.B. soll die befragte Person in einem Untertest der Cattell'schen Testbatterie (Schmidt et al. 1994) beurteilen, ob jede Feststellung einer vorgegebenen Liste 'sinnvoll ist und einen guten Eindruck hinterläßt' oder aber 'sinnlos ist und ein schlechtes Licht auf den wirft, der sie benutzt'. Es folgt dann eine Reihe klischeehafter Feststellungen, wie

Frauen können sich nie entscheiden
Jeder Mensch braucht Freunde
Geld ist der Grund vieler Bosheit
Wer Geld hat, hat auch Freunde...u.s.w.

Während der Befragte gemäß der Testinstruktion nach Sinn und Unsinn jeder einzelnen Feststellung ringt, wird am Ende nur ausgezählt, *wieviele* Feststellungen

man für sinnvoll hält - als Maß für die 'Hausbackenheit' des Befragten.

Auch viele der wöchentlich neu konstruierten 'Psyche-Tests' in Illustrierten sind von diesem Typ, etwa wenn aus der Beantwortung der Frage

Sind Sie eigentlich 'wetterfühlig' ?

darauf geschlossen wird, wie 'rücksichtsvoll' die befragte Person ist. Diese Tests sind ein gutes Beispiel dafür, daß ein Test zwar das Gütekriterium der Objektivität (oder einen Aspekt davon) erfüllen kann, aber trotzdem keine große Validität besitzt.

Das Konzept objektiver Tests setzt eine *nicht-triviale Theorie* über den Zusammenhang von unverfänglich erfragbaren Verhaltensaspekten und relevanten Persönlichkeitseigenschaften voraus. Vielleicht liegt es daran, daß die Erfassung von Persönlichkeitseigenschaften mit solchen Tests im akademischen Bereich derzeit kaum eine Rolle spielt. Es ist offenbar sehr schwer, von Verhaltensaspekten zuverlässig auf Persönlichkeitseigenschaften zu schließen, bezüglich derer die Fragen *keine* Augenscheinvalidität besitzen.

Trotzdem stellen diese Tests wohl am ehesten das dar, was der Laie von psychologischen Tests erwartet: auf geheimnisvolle Weise aus ein paar banalen Antworten auf tiefliegende Strukturen der Persönlichkeit schließen zu können.

2.2.2.4 Projektive Tests

Ein ganz anderer Weg, zu objektiven Testresultaten zu gelangen, wird mit den sogenannten projektiven Verfahren begangen. Der Name leitet sich aus der psychoanalytischen Theorie ab, in der mit *Projektion* ein Abwehrmechanismus bezeichnet wird, mit Hilfe dessen sich das Ich gegen angstauslösende oder verbotene Triebregungen wehrt. Die Abwehr besteht darin, daß diese inneren Regungen und Impulse nach außen, meistens auf andere Personen projiziert werden und dadurch nicht mehr mit den Normen des eigenen Über-ich in Konflikt geraten können.

Bei *projektiven Tests* wird angenommen, daß dieser Vorgang auch in der Situation einer Testvorgabe stattfinden kann und man somit über die Itemantworten zu Erkenntnissen über Persönlichkeitseigenschaften gelangt, die der Person selbst gar nicht bewußt sind oder in einem direkten Fragebogen nicht geäußert ('zugegeben') würden.

Um den Vorgang der Projektion zu ermöglichen, stellen die Items *Stimuli* dar (das sind auslösende Reize), welche möglichst *unstrukturiert* sein müssen. Sie müssen einerseits innere Vorgänge stimulieren, die dann zum Inhalt einer Projektion werden können. Andererseits muß das Item so vage (unstrukturiert) sein, daß man in diesen Stimulus auch Eigenes 'hineinlesen' oder projizieren kann. Bezüglich dieser Eigenschaften unterscheiden sich projektive Verfahren graduell.

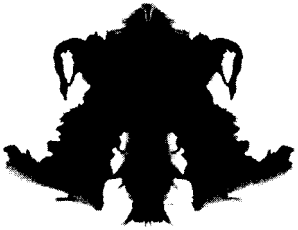


Abbildung 7: Ein Item aus dem Rorschach Test (Rorschach, 1954)

Während die Items des Rorschach Tests nur aus einem (spiegelsymmetrischen) Tintenklecks bestehen, sind die Bilder des thematischen Apperzeptionstests (TAT) photographisch genau, jedoch bezüglich der Interpretation ihres Inhaltes offen und unstrukturiert.

Abbildung 8: Ein Item aus dem thematischen Apperzeptionstests (TAT; Revers & Taeber 1968)

Unstrukturiert sind die Items des Rosenzweig Picture-Frustration Tests dadurch, daß es sich um sehr sparsame Strichzeichnungen handelt, die keine Ähnlichkeit mit existierenden Personen haben. Dadurch wird es der befragten Person erleichtert, sich selbst mit der antwortenden Person in der Zeichnung zu identifizieren.

Abbildung 9: Ein Item aus dem Picture-Frustration Test (Hörmann & Moog 1957)

Das Konzept von projektiven Tests ist auch über den engen psychoanalytischen Begriff der Projektion hinaus sinnvoll. So ist es eine unbestreitbare Tatsache, daß man leicht 'von sich auf andere schließt' oder Dinge assoziiert, die dem eigenen Erleben und Denken entspringen.

Projektive Tests sind immer dann in Betracht zu ziehen, wenn Persönlichkeitseigenschaften gemessen werden sollen, die mit einer *starken positiven oder negativen Wertung* verknüpft sind, sei diese gesellschaftlicher oder individueller Natur. Beispiele sind etwa die Messung der Ag-

gressivität, die man ungern zugibt oder auch nur wahrhaben will, oder die Messung des Leistungsmotivs, über dessen Stärke man sich oft nicht im Klaren ist, und dessen hohe Ausprägung in unserer Gesellschaft eine positive Norm darstellt.

Die Stärke der Tendenz, mit der sich eine Person an der gesellschaftlichen oder sozialen Norm orientiert, nennt man die Variable der *sozialen Erwünschtheit* (social desirability). Die soziale Erwünschtheit beeinflußt potentiell jedes Ergebnis einer direkten Befragung. Projektive Verfahren können als Versuch aufgefaßt werden, den Einfluß der sozialen Erwünschtheit auf das Testergebnis dadurch möglichst gering zu halten, daß der Befragte in der Itemantwort nicht über sich selbst sprechen muß (und sich somit sozial erwünscht darstellt), sondern über einen abstrakten Stimulus oder eine fremde Person.

2.2.2.5 Situationsfragebögen

Eine andere Art von Projektion stellt das *'Sich-hinein-versetzen'* in eine beschriebene Situation dar. Hier werden nicht innere Triebregungen nach außen projiziert, sondern die eigene Person versetzt sich in der Vorstellung in eine hypothetische Situation. Sodann wird das Erleben und Verhalten in dieser Situation erfragt. Derartige Tests heißen *Situationsfragebögen.*

Ein Beispiel ist etwa das Angstbewältigungsinventar (ABI, Krohne et al. 1989), in dem die Person aufgefordert wird, sich folgende Situation vorzustellen:

> *Stellen Sie sich vor, Sie fahren als Beifahrer mit einem offensichtlich ungeübten Autofahrer. Es herrschen durch Schnee und Glatteis ungünstige Straßenverhältnisse.*

Die Person hat dann für 18 Verhaltensbeschreibungen anzugeben, ob diese für sie in der Situation zutreffend sind oder nicht, z.B.:

- *denke ich: 'Mir bleibt auch nichts erspart.'*
- *sage ich mir: 'Es wird schon nichts Schlimmes passieren.'*
- *schaue ich einfach nicht mehr auf die Fahrbahn, sondern denke an etwas anderes oder betrachte die Gegend.*

Verlangt wird von der befragten Person - wie bei Persönlichkeitsfragebögen - eine Selbstauskunft, jedoch ohne die Voraussetzungen der Metakognition und des sozialen Maßstabs (S.O. Kap. 2.2.2.2). Voraussetzung ist im allgemeinen nur die Erinnerung an ähnliche Situationen und die Fähigkeit, das Wissen aus dieser Erinnerung heraus auf die vorgegebene, hypothetische Situation zu übertragen. Natürlich ist das auch ein Stück 'Selbstkenntnis', jedoch wird von der befragten Person *keine Einschätzung* der eigenen Person verlangt, sondern *Auskunft über potentiell beobachtbares Verhalten* und potentielles Erleben.

Die Voraussetzung der Offenbarungsbereitschaft und die Möglichkeit einer Beeinflussung durch die soziale Erwünschtheit ist bei Situationsfragebögen genauso gegeben, wie bei Persönlichkeitsfragebögen.

2.2.2.6 Einstellungstests

Die Messung von Einstellungen (attitudes) ist ein sehr altes Kapitel in der Geschichte der Messung psychischer Merkmale. Im Unterschied zu generellen Persönlichkeitseigenschaften sind Einstellungen auf

ein bestimmtes *Objekt gerichtet,* das Einstellungsobjekt. Das Einstellungsobjekt muß nicht eine Person oder Sache sein, sondern kann auch ein abstraktes Prinzip, ein Paragraph oder Ähnliches sein. Beispiele sind etwa:

Einstellung gegenüber Kernkraftwerken
Einstellung zur Abtreibung
Einstellung gegenüber Ausländern
Einstellung zum Recht auf Freie Meinungsäußerung

Üblicherweise wird bei Einstellungen eine Pro-Contra-Dimension oder eine *Zustimmungs-Ablehnungs-Dimension* gemessen. Dies geschieht in der Regel dadurch, daß verschiedene Statements über das Einstellungsobjekt vorgegeben werden und die befragten Personen angeben sollen, inwieweit sie der jeweiligen Aussage zustimmen oder sie ablehnen. Diese Aussagen stellen die Items des Tests dar und jede Aussage (jedes Item) druckt eine bestimmte Position auf der zu messenden Pro-Contra-Dimension aus.

In einem Fragebogen zur Einstellung gegenüber der Nutzung von Kernenergie markieren z.B. die drei folgenden Items unterschiedliche Positionen auf der zu messenden Einstellungsdimension:

- *Kernkraftwerke sichern langfristig unsere Energieversorgung.*
- *Die derzeit in Betrieb befindlichen Kernkraftwerke sollten innerhalb der nächsten 10 Jahre abgeschaltet werden.*
- *Kernkraftwerke stellen eine Technologie dar, die gegenüber den nachfolgenden Generationen unVerantwortbar ist.*

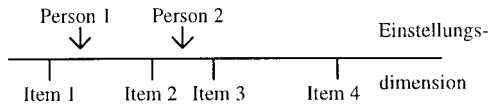
Während das erste Item am positiven Ende der Einstellungsdimension liegt, hat das zweite Item eine mittlere Position und das dritte Item liegt am negativen Ende. Die Zustimmung zu der jeweiligen Aussage kann im einfachsten Fall mit einer ja-nein Antwort erfaßt werden, wird aber in der Regel mit einer mehrstufigen Ratingskala erfaßt (vgl. Kap. 2.3.1.3), z.B.:

- *stimme völlig zu*
- *stimme eher zu*
- *lehne eher ab*
- *lehne völlig ab*

Über den Zusammenhang von Antwortverhalten und latenter Variable gibt es zwei unterschiedliche Annahmen, die jeweils auch unterschiedliche Testmodelle für die Testauswertung erforderlich machen. Sie werden nach den beiden 'Pionieren' der Einstellungsmessung, L.L. Thurstone und R. Likert benannt.

Die Annahme der Thurstone-Skalierung

Thurstone (Thurstone & Chave 1929) hat eine Methode zur Einstellungsmessung angewendet, deren zentrale Annahme darin besteht, daß die Personen denjenigen Items zustimmen, die ihrer eigenen Position auf der Einstellungsdimension *am nächsten* liegen. Items, die von der eigenen Position weiter entfernt liegen, werden dagegen abgelehnt. In der folgenden Graphik würde also Person 1 den Items 1 und 2 zustimmen und die Items 3 und 4 ablehnen.



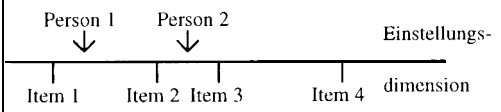
Person 2 würde den Items 2 und 3 zustimmen und die Items 1 und 4 ablehnen.

Diese Annahme ist zwar sehr plausibel, hat aber für die Testauswertung die schwerwiegende Konsequenz, daß man die Positionen der Items genau kennen muß, um zu Meßwerten für die Personen zu gelangen.

Die **Annahme der Likert-Skalierung**

Likert (1932) hat eine andere Methode der Einstellungsmessung verwendet, bei der man *nicht* die Position jedes einzelnen Items kennen muß. Sie basiert vielmehr auf der Annahme, daß jedes Item entweder eine *positive* oder eine *negative Haltung* gegenüber dem Einstellungsobjekt ausdrückt. Eine Zustimmung zu einem positiven Item kann dann genauso gewertet werden wie eine Ablehnung eines negativen Items.

Die grundlegende Annahme über das Antwortverhalten besagt, daß eine Person allen positiv formulierten Items umso mehr zustimmt, und alle negativ formulierte Items umso mehr ablehnt, je positiver ihre Einstellung zu dem betreffenden Objekt ist. Handelt es sich in dem folgenden Beispiel um vier positiv formulierte Items, so würde nach dieser Annahme die Person 1 dem ersten Item zustimmen, die anderen drei eher ablehnen.



Die Person 2 stimmt den Items 1 und 2 zu, wobei die Zustimmung zu Item 1 deutlicher ausfallt ('stimme völlig zu'), weil die Einstellung der Person noch positiver ist. als es das erste Item ausdrückt. Die Zustimmung zu einem Item *sinkt* also nicht mit zunehmender Distanz zu der Position des Items (wie bei der Thurstone-Skalierung), sondern sie *steigt* mit zunehmender Distanz *in positiver Richtung.*

Auch diese Annahme ist für viele Einstellungstests sehr plausibel. Sie hat den Vorteil, daß die Testauswertung vergleichsweise unkompliziert ist, sofern alle Items eindeutig positiv oder negativ formuliert sind.

Welche der beiden Annahmen über das Antwortverhalten zutreffend ist, hängt weitgehend auch von der Formulierung der Items ab. In einem Test zur Messung der Einstellung zum Umweltschutz löst das folgende Item sicherlich umso mehr Zustimmung aus, je positiver die Einstellung ist:

> *Jeder Bürger sollte seinen privaten Energieverbrauch so weit wie möglich reduzieren.*

Dagegen wird die folgende Formulierung vermutlich sowohl von Personen abgelehnt, die eine geringe Ausprägung der Einstellung haben, als auch von Personen, die weitaus drastischere Maßnahmen zur Erhaltung der Umwelt für notwendig halten:

> *In der Reduktion des privaten Energieverbrauchs liegt der Schlüssel zum Schutz der Umwelt.*

Bei der Konstruktion eines Einstellungstests muß man sich frühzeitig für eine der beiden Annahmen entscheiden und die Items entsprechend formulieren. Die meisten der in Kapitel 3 behandelten *quantitativen* Testmodelle eignen sich nur zur Analyse von Items, für die die Annahme der Likert-Skalierung gilt. Modelle für Einstellungstests, die nach der Thurstone-Methode konstruiert sind, werden in Kapitel 3.1.1.3 behandelt. Allerdings eignen

sich auch Testmodelle mit einer *kategorialen Personenvariable* für die Auswertung von Einstellungstests. Für die Anwendung dieser Testmodelle spielt es *keine* Rolle, ob die Itemantworten nach Thurstone oder nach Likert zustandegekommen sind.

Darüber hinaus ist es bei der Anwendung klassifizierender Testmodelle auch nicht erforderlich, von einer *Einstellungsdimension* auszugehen. Individuelle Unterschiede in der Einstellung gegenüber einem Einstellungsobjekt können sich auch in Form von qualitativen Unterschieden äußern. Man spricht dann auch von der *Einstellungsstruktur*. Damit ist gemeint, daß sich Personen darin unterscheiden, bei *welchen* Items sie im Sinne einer positiven Einstellung antworten und bei welchen im Sinne einer negativen Einstellung. In diesem Fall sind weder die *Personen* noch die *Items* auf einer Dimension anordenbar, wie das zuvor stets vorausgesetzt wurde.

2.2.2.7 Motivations- und Interessensfragebögen

Interessen sind wie Motivationen Eigenschaften, die eine Antriebsqualität für das Handeln von Personen haben, also gleichsam Motor des Verhaltens sind. Fragebögen, die diese Eigenschaften direkt erfassen sollen, bestehen oft aus Fragen der folgenden Art:

Was machst Du am liebsten.....
Was würdest Du gerne tun....
Wozu hast Du Lust.....

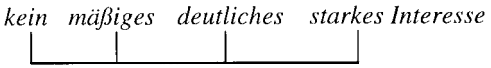
Dieser Typ von Tests hat vieles gemeinsam mit den bisher beschriebenen Testarten. So wird eine *Selbstauskunft über innere Zustände oder Vorgänge* erfragt, die vergleichbare Voraussetzungen erfordert und Verfälschungsgefahren birgt, wie Persönlichkeitsfragebögen. Interessen sind wie Einstellungen *objektbezogen,* d.h. man hat ein Interesse *an* etwas oder *für* etwas, und man druckt sein Interesse wie seine Meinung gerne *graduell abgestuft* aus (z.B. 'mein Interesse daran ist eher gering').

Ein besonderer Aspekt von Interessensfragebögen besteht jedoch darin, daß Interessen *zukunftsorientiert* sind und in der Regel auch zukunftsbezogen erfaßt werden. Damit ist gemeint, daß sich die Tätigkeit, auf die sich das Interesse bezieht, erst in der Zukunft ausgeführt wird: '..was möchtest Du (...gleich... später... morgen...) tun?'.

Besonders deutlich wird die Zukunftsorientierung, wenn man etwa Schülerinteressen erhebt, um den späteren Unterricht für diese Schüler zu planen, oder Berufsinteressen, um die Probanden bezüglich ihrer Berufswahl zu beraten. Hier bezieht sich die Itemantwort auf eine innere Vorliebe für etwas, was der Befragte *noch gar nicht kennt* und kennen kann, weil es ihm noch bevorsteht.

Dem trägt man bei der Testkonstruktion dadurch Rechnung, daß entweder die einzelnen Items aus einem längeren Text bestehen oder mehreren Items ein gemeinsamer *Text* vorangeht, in dem das Interessensobjekt beschrieben ist. Diese Beschreibung dient dann als *Stimulus,* der das Interesse 'wecken' oder zumindest bewußt machen soll.

Die Itemantwort kann dann - wie bei einem Einstellungstest - in einem Urteil auf einer mehrstufigen Ratingskala bestehen, z.B.



Das Problem dabei ist, daß die befragten Personen einen vergleichbaren *Beurteilungsmaßstab* haben sollten, nämlich was 'mäßiges', 'deutliches' und 'starkes' Interesse ist. Man verläßt sich hier auf die intersubjektive Gültigkeit der Sprache, die bei solchen differenzierten Urteilen oft fraglich ist.

Die Stärke des Interesses muß nicht unbedingt auf einer Ratingskala eingestuft werden. Es kann z.B. auch eine *Präferenzwahl* (Präferenz = Bevorzugung) aus vorgegebenen Alternativen erfolgen, die als Ausdruck des *relativen* Interesses gewertet wird, z.B.:

Was würden Sie jetzt am liebsten lesen:

- *das nächste Kapitel über Verhaltensfragebögen*
- *etwas über Ergebnisse der Interessensforschung*
- *etwas darüber, wie man Interessenstests mit Präferenzwahlen auswertet*
- *eine Übersicht, welche Interessenstests schon entwickelt und erprobt worden sind*

Wählen Sie eine Alternative aus!

Der Nachteil von solchen Präferenzwahlen besteht darin, daß die Itemantwort nur den Schluß zuläßt, daß die gewählte Alternative *relativ zu den anderen Alternativen* als interessant gilt. Es können somit nur relative Interessensausprägungen gemessen werden und das Testergebnis hängt völlig von den angebotenen Vergleichsalternativen ab. Die generelle Problematik von Antwortformaten mit nominal-skalierten Antwortvariablen wird in Kapitel 3.2 behandelt.

2.2.2.8 Verhaltensfragebögen

Aufgrund der Probleme, die mit der Einschätzung und Beurteilung eigener innerer Zustände und Vorgänge verbunden sind, bietet sich als Alternative an, statt innerer Zustände das *tatsächliche Verhalten* der Personen mit Fragebögen zu erfassen. Es hat einigen Reiz, sich nicht mit Introspektion, sozialen Vergleichen, Beurteilungsmaßstäben und Präferenzurteilen auseinandersetzen zu müssen, sondern den Probanden schlicht zu fragen:

Was hast Du getan ?

Prominentes Beispiel ist etwa die Erfassung des Umweltbewußtseins, wo man einsehen mußte, daß Selbsturteile über umweltrelevante Einstellungen, Verantwortungszuschreibungen und sogar Handlungsabsichten nicht das tatsächliche Verhalten im Umweltbereich vorherzusagen gestatten.

Mit Verhaltensfragebögen möchte man erfassen, was die befragten Personen tatsächlich in der *Vergangenheit* getan haben. Im Gegensatz zu Situationsfragebögen ist selbst die Übertragung auf hypothetische Situationen ausgeschlossen.

Voraussetzungen für die Interpretierbarkeit der Itemantworten sind ein hinreichend zuverlässiges *Gedächtnis* der Probanden für das eigene Verhalten und die Bereitschaft, *ehrlich* Auskunft zu geben. Die soziale Erwünschtheit kann natürlich auch die Ergebnisse eines Verhaltensfragebogens beeinflussen, jedoch ließe sich diese Beeinflussung nur über eine bewußte Lüge realisieren. Hier ist die Hemmschwelle sicherlich höher, als bei der Selbsteinschätzung einer Persönlichkeits-

eigenschaft, welche sich leichter 'verzerren' läßt.

Diese Ehrlichkeit vorausgesetzt, kann das Antwortverhalten im Test gleichgesetzt werden mit dem tatsächlichen Verhalten. Damit ist man aber nur scheinbar 'dichter' an der zu messenden Persönlichkeitseigenschaft dran. Das Problem beim *Schluß* vom tatsächlichen Verhalten auf *eine Persönlichkeitseigenschaft* besteht darin, daß das gezeigte Verhalten außer von der vermuteten Persönlichkeitseigenschaft von einer Vielzahl von *situationalen Bedingungen* abhängt.

So kann man z.B. keine Gelegenheit gehabt haben, ein Verhalten zu zeigen, daran gehindert worden sein, von anderen veranlaßt worden sein, es zu zeigen, oder es aus ganz anderen Gründen 'zufällig' gezeigt haben. Kurzum, der Schluß von Verhalten unter Realbedingungen auf Personeneigenschaften ist extrem *fehlerbehuftet.* Um diesen Fehler klein zu halten, sollte man nach solchen Verhaltensweisen fragen, bei denen die situationalen Bedingungen für alle Befragten möglichst gleich sind. Dies kann wiederum die Aussagekraft bezüglich der zu messenden Eigenschaft einschränken.

Sieht man von der Beeinträchtigung durch situationale Faktoren einmal ab, setzt der Schluß vom Testverhalten (= erfragtes Verhalten) auf Personeneigenschaften Annahmen darüber voraus, unter welchen Eigenschaftsausprägungen welches Verhalten zu erwarten ist. Im Falle einer *quantitativen Eigenschaft* kann dies - wie bei Einstellungstests - darüber geschehen, daß die erfragten Verhaltensweisen unterschiedliche Punkte auf der Eigenschaftsdimension markieren. Auch hier gibt es wieder die beiden Alternativen, die den Annahmen der Thurstone- und der Likert-Skalierung analog sind (vgl. Kap. 2.2.2.6):

Erstens, das Verhalten tritt nur dann auf, wenn die Eigenschaftsausprägung der Person *in* der *Nähe* der Position der Verhaltensweise ist.

Zweitens, das Verhalten wird von einer bestimmten Eigenschaftsausprägung an *aufwärts* gezeigt.

Beispiel

In einem Fragebogen zum Umwelthandeln wird gefragt:

- *haben Sie in letzter Zeit Geld für eine Umweltschutzorganisation gespendet?*
- *sind Sie Mitglied in einer Umweltschutzorganisation?*
- *arbeiten Sie in einer Umweltschutzorganisation mit?*

Die erste Verhaltensweise zeigt sich im Sinne der ersten, oben genannten Alternative vermutlich nur bei einer mittleren Handlungsbereitschaft, aber nicht bei einer sehr schwachen oder sehr starken Handlungsbereitschaft. Bei einer sehr starken Handlungsbereitschaft spendet man nicht mehr, sondern arbeitet selbst mit.

Das zweite Item wird vermutlich im Sinne der zweiten Alternative beantwortet, da man auch bei einer *aktiven* Mitarbeit in einer Umweltorganisation Mitglied in dieser Organisation ist.

In Verhaltensfragebögen, die sehr viele Verhaltensweisen abfragen, ist die erste Annahme über das Antwortverhalten sehr viel realistischer, da auch von Personen mit hoher Eigenschaftsausprägung (Handlungsbereitschaft) nicht erwartet werden kann, daß sie *alle* Verhaltensweisen zeigen. Dafür reicht oft die zur Verfügung

stehende Zeit nicht aus und auch bei einer starken Handlungsbereitschaft werden Akzente auf *bestimmte* Aktivitäten gesetzt. Entsprechendes gilt z.B. auch für sog. *Symptomlisten,* bei denen ebenfalls nicht erwartet werden kann, daß Patienten mit einer starken Ausprägung der Störung *alle* Symptome eines Krankheitsbildes zeigen.

Für die Testauswertung bedeutet dies, daß *quantitative* Testmodelle mit monoton steigenden Itemfunktionen (s. Kap. 3) nicht geeignet sind, die Handlungsbereitschaft zu messen. Testmodelle mit *kategorialer* Personenvariable sind hier sehr viel unproblematischer, da sich bei diesen Modellen die Personen hinsichtlich ihres *Musters* an Verhaltensweisen unterscheiden und nicht nur hinsichtlich der *Anzahl* an Verhaltensweisen.

2.2.3 Definition des Itemuniversums

Aus der inhaltlichen Theorie über die zu messende Personeneigenschaft sollte auch ableitbar sein, in welchen *Situationen* sich ein Verhalten äußert, das Rückschlüsse über die Ausprägung der Personeneigenschaft zuläßt. Diese Beschreibung einer *Klasse von Situationen,* in denen sich ein bestimmtes Verhalten zeigen kann, und einer *Klasse von Verhaltensweisen,* die Rückschlüsse auf die Personeneigenschaft zulassen, muß dann transformiert werden in eine Beschreibung des *Itemuniversums.*

Beispiel

Bei der Messung der Fähigkeit zum analogen Schließen ist die Menge der Situationen durch alle Problemstellungen definiert, die die formale Struktur

A : B = C : ?

(A verhält sich zu B wie C zu ?)

haben. Die Klasse der Verhaltensweisen unterscheidet lediglich zwei Arten von Verhalten, nämlich sinnvolle und sinnlose Ergänzungen der Analogie. Sinnvolle sind dadurch definiert, daß das für das Fragezeichen gefundene Element in derselben Relation zu C steht wie das Element B zu A.

Diese Situations- und Verhaltensbeschreibungen für die Fähigkeit des analogen Schließens sind natürlich noch keine Definition eines Itemuniversums. Hier müssen im Sinne einer *operationalen Definition* (S.O.) pragmatische und formale Festlegungen getroffen werden, die allerdings die Gültigkeit des Tests für die in der Theorie behandelte Persönlichkeitseigenschaft einschränken.

So ließe sich im vorliegenden Beispiel das Itemuniversum als die Menge aller deutschsprachigen Drei-Wort-Analogien definieren, bei denen es ein viertes Wort geben muß, das zu C in derselben Relation steht wie B zu A. Damit sind alle nichtsprachlichen und fremdsprachlichen Analogien ausgeschlossen, sowie solche Analogien, die mehrere Worte pro Element des Analogieschlusses benötigen.

Bei der Definition des Itemuniversums hat man sich davon leiten zu lassen, welche Art von Items homogen genug zu sein scheint, um die Messung der gewünschten Persönlichkeitseigenschaft zu ermöglichen. Eine solche *Homogenitätsvermutung* ist natürlich eine sehr subjektive Angelegenheit und resultiert gewöhnlich aus einer Mischung von Erfahrung mit Testkonstruktionen und einer weiteren Elaboration der Theorie über das zu messende Persönlichkeitsmerkmal.

Die Definition eines Itemuniversums ist deswegen von Bedeutung, weil ein Testergebnis nicht nur etwas über die Beantwortung der im Test enthaltenen Items aussagen will, sondern eine generalisierende Aussage über das Antwortverhalten bezüglich einer ganzen Klasse von Situationen (Items) ermöglichen soll. Das Item-Universum definiert den *Geltungsbereich* des Testergebnisses.

2.2.4 Ziehung einer Itemstichprobe

Wenn es um die Ziehung von Stichproben geht, denkt man zunächst an eine Zufallsstichprobe, da deren Ergebnisse am ehesten generalisiert werden dürfen. Die Ziehung einer Zufallsstichprobe aus der Menge aller möglichen Items (Itemuniversum) ist in der Regel weder möglich noch sinnvoll.

Möglich ist eine Zufallsziehung oft deswegen nicht, weil das Itemuniversum zwar theoretisch definiert werden kann, jedoch nicht in einem physischen Sinne existiert wie etwa die Population eines Landes. Wo keine Grundmenge existiert, ist es technisch zumindest schwierig, eine Stichprobe zu ziehen.

Auch sinnvoll wäre eine Zufallsstichprobe nicht, da man einen Test im allgemeinen für eine bestimmte Adressatengruppe konstruiert und man eine Itemauswahl treffen sollte, die speziell zu dieser Adressatengruppe 'paßt'. Das *Prinzip der Passung* von Personenstichprobe und Itemstichprobe zielt in erster Linie auf die *Maximierung der Varianz der Antwortvariablen* ab. Das bedeutet, daß solche Items ausgewählt werden sollten, von denen erwartet wird, daß es eine starke Streuung der Itemantworten in der betreffenden Personenstichprobe gibt.

Items, auf die sämtliche befragten Personen einer Stichprobe dieselbe Antwort geben, bei denen also die Varianz der Itemantwort 0 beträgt, sind schlicht wertlos. Es läßt sich im Rahmen von vielen Testmodellen zeigen, daß tatsächlich diejenigen Items die *meiste Information* zur Messung eines Personenmerkmals beitragen, bei denen die Variation der Itemantworten am größten ist (s. Kap. 6.1).

Im Falle von Leistungstestitems, bei denen nur zwischen einer korrekten und einer falschen Antwort unterschieden wird, ist die Varianz der Itemantworten dann maximal, wenn das Item in der betreffenden Stichprobe eine *relative Lösungshäufigkeit von* 50 % hat. Dies läßt sich direkt aus der Formel für die Varianz einer 0- 1 -Variable ablesen. Diese lautet nämlich

$$\mathrm{Var}(X) = p\,(1-p),$$

wenn X nur die Werte 0 oder 1 annimmt und p die Wahrscheinlichkeit bezeichnet, daß X den Wert 1 annimmt, also $p(X = 1)$.

Die folgende Tabelle zeigt, daß diese Varianz mit einem Wert von 0.25 bei $p = 0.5$ maximal ist.

p(X=1)	.1	.2	.3	.4	.5	.6	.7	.8	.9
Var (X)	.09	.16	.21	.24	.25	.24	.21	.16	.09

Dieses Prinzip der Passung von Item- und Personenstichprobe gilt jedoch *nicht nur für Leistungstests* und auch nicht nur für die Messung von quantitativen Personenvariablen. Will man etwa mit Hilfe eines Verhaltensfragebogens umweltpolitisch aktive Personen von umweltpolitisch nicht aktiven Personen unterscheiden (eine zweikategorielle Personenvariable), so

wäre es im wahrsten Sinne des Wortes 'unpassend', relativ mittellose Gymnasiasten zu fragen, ob sie schon einmal einer Umweltschutzorganisation einen größeren Geldbetrag gespendet haben.

Neben dem Prinzip der Passung muß auch davon ausgegangen werden, daß es einfach *bessere und schlechtere* Vertreter des Itemuniversums gibt. D.h. keine noch so sorgfältige Definition eines Itemuniversums wird ausschließen können, daß es Items gibt, bei denen der Schluß vom Antwortverhalten auf die Personeneigenschaft zwingend und eindeutig ist, und solche, bei denen andere Faktoren als die zu messende Personeneigenschaft das Antwortverhalten beeinflussen können. Diese Frage geht jedoch in den Bereich der Itemkonstruktion hinein, der in Kapitel 2.3 behandelt wird.

Auch über die *Größe der Itemstichprobe* läßt sich wenig Allgemeingültiges aussagen. Generell gilt, daß eine höhere Meßgenauigkeit durch eine größere Itemanzahl erreicht werden kann. Andererseits hat eine größere Itemanzahl auch negative Auswirkungen wie Ermüdung, Redundanz, Konzentrationseinbußen, Minderung der Antwortbereitschaft, Lern- und Übungseffekte, und vieles andere mehr.

Zusammenfassend sei festgehalten, daß die Ziehung einer Itemstichprobe anderen Prinzipien folgt und generell sehr viel schwieriger ist als etwa die Ziehung einer Personenstichprobe aus einer definierten Personenpopulation. Dennoch ist es sinnvoll, die Menge der in einem Test enthaltenen Items *als Stichprobe* aus einer hypothetischen Grundgesamtheit zu betrachten und, soweit es geht, auch so zu behandeln, da sonst die Frage der Generalisierbarkeit des Testergebnisses schwer zu beantworten ist.

2.2.5 Auswahl eines geeigneten Testmodells

Auch die Auswahl eines geeigneten Testmodells gehört in die Planungsphase, d.h. in die Phase der Konstruktion eines Testinstrumentes. Idealerweise sollte auch hier die Theorie über das jeweilige Personenmerkmal so präzise sein, daß die Annahmen über den Zusammenhang von Antwortverhalten im Test und latenter Personenvariable direkt ableitbar sind.

Dies ist in der Praxis nicht immer der Fall, *so* daß *das umgekehrte Vorgehen* gewählt wird: Man überlegt sich, welche Testmodelle man kennt und welches am ehesten zu der Theorie über die Persönlichkeitseigenschaft paßt. Dieses setzt natürlich einen Überblick über ein möglichst breites Spektrum bestehender Testmodelle voraus.

Sich in der Phase der Testkonstruktion auf ein bestimmtes Testmodell festzulegen, ist deswegen von Bedeutung, weil bestimmte formale Annahmen des jeweiligen Modells auch spezielle Anforderungen an die Itemformulierung und Testkonstruktion stellen. Z.B. macht es einen Unterschied, ob man einen *deterministischen* Zusammenhang zwischen Antwortverhalten und latenter Variable annimmt oder einen *probabilistischen* Zusammenhang.

Ein Item wie

Ich könnte mir vorstellen, einmal gegen die Errichtung eines großtechnologischen Projektes Einspruch zu erheben
(Antwort: ja - nein)

steht wohl kaum in einem deterministischen Zusammenhang mit einer politischen Einstellungsdimension wie *Protestbereitschuf?*. Dies könnte bei einem Item wie

Ich habe schon einmal an einer Demonstration gegen ein Kernkraftwerk teilgenommen (Antwort: ja - nein)

dagegen eher möglich sein.

Die Auswahl eines passenden Testmodells in der Planungsphase kann auch damit enden, daß man zwei oder drei alternative *Testmodelle zur Auswahl* hat und man empirisch darüber entscheiden will, welches Modell am besten paßt. Dies ist im Sinne eines Entscheidungsexperimentes nicht nur ein legitimes Vorgehen, sondern kann ausgesprochen interessante Fragestellungen einer empirischen Klärung zuführen.

Dies reicht hin bis zu der Grundfragestellung, ob eine angenommene Persönlichkeitseigenschaft dimensionaler oder typologischer Natur ist (ob die Personenvariable quantitativ oder kategorial ist). Für die weitere Konstruktion des Testinstrumentes hat dies jedoch die Konsequenz, daß das Testinstrument mit den Annahmen der gewählten Testmodelle kompatibel sein muß. Was das im einzelnen bedeuten kann, wird im Laufe des Kapitels 3 deutlich.

Literatur

Nährer, W. (1986) stellt Konzeptionen von Leistungstests mit Zeitbegrenzung dar (Speed-Tests). Auf die Messung von Persönlichkeitseigenschaften mit Fragebögen gehen Angleitner & Wiggings (1986) ein, das Konzept objektiver Persönlichkeitstests diskutieren Schmidt (1975) und Schmidt & Schwenkmezger (1994). Die Problematik projektiver Verfahren wird von Allesch (1991), Asendorpf (1994) und Tent (1991) erörtert, Westmeyer (1994) stellt das Selbstverständnis der Verhaltensdiagnostik dar. Dawes (1977) behandelt die Grundlagen der Einstellungsmessung und Edwards (1957) den Einfluß der 'sozialen Erwünschtheit' in Persönlichkeitsfragebögen. Eine Beschreibung der Likert- und der Thurstone-Skalierung findet sich z.B. bei Roskam (1983) und Schnell et al. (1989).

Übungsaufgaben

Sie sollen drei Testinstrumente neu entwickeln, und zwar zu den drei Personeneigenschaften:
- Freundlichkeit im zwischenmenschlichen Umgang
- die Eigenschaft, in Persönlichkeitsfragebögen sozial erwünscht zu antworten
- Prüfungsangst.

Wählen Sie für jedes Instrument eine andere Testart aus (begründen Sie die Wahl), beschreiben Sie die Art der Per sonenvariable und konstruieren Sie je zwei Beispielitems.

2. Welche Varianten von Speed-Test! gibt es? (Vor- und Nachteile)

3. Welche Voraussetzungen müssen bei der Beantwortung von Persönlichkeitsfragebögen seitens der befragten Per son gegeben sein?

4. Worin unterscheiden sich die Annahmen einer Thurstone-Skalierung und einer Likert-Skalierung? Bei welchen Testarten kann diese Unterscheidung eine Rolle spielen?

2.3 Itemkonstruktion

Nach den Planungsüberlegungen, die die Anlage und Konstruktion des gesamten Testinstrumentes betreffen, stellt die Formulierung und Konstruktion der einzelnen Items die 'eigentliche' Arbeit der Testkonstruktion dar. Es ist nicht leicht, etwas über die Konstruktion von Items zu sagen, ohne sich zumindest auf einen bestimmten Typ von Tests zu beziehen oder sogar auf eine bestimmte zu messende Personeneigenschaft. Trotzdem gibt es einige übergreifende *Konstruktionsprinzipien,* die bei sehr vielen Testarten zu berücksichtigen sind.

Hierzu soll zunächst dargestellt werden, was ein Item überhaupt ist und welche *Bestundteile* es hat. Danach wird in getrennten Unterkapiteln auf verschiedene Arten von Antwortformaten, auf die sprachliche Formulierung der Items und auf die Zusammenstellung des Tests eingegangen.

Das Item ist die *kleinste Beobachtungseinheit* in einem Test, sozusagen der elementare Baustein, aus dem ein Test aufgebaut ist. An einem Item lassen sich zwei Komponenten unterscheiden, nämlich der sogenannte *Itemstamm* und das *Antwortformat.*

Der Itemstamm kann aus einer Frage, einer Aussage, einem Bild, einer Geschichte, einer Zeichnung oder einer Rechenaufgabe bestehen und stellt ganz allgemein die Situation dar, in der die Person ihr Testverhalten zeigt.

Demgegenüber dient das Antwortformat der Registrierung eben dieses Testverhaltens. Es kann aus anzukreuzenden Alternativen bestehen, aus einer leeren Zeile, in die man etwas eintragen muß, aus einer mehrstufigen Antwortskala, auf der man eine Stufe ankreuzen muß, oder einem weißen Blatt Papier, auf das man etwas zeichnen soll.

Diese beiden Bestandteile gehören aus logischen Gründen zu einem Item, denn man möchte in einem Test das Verhalten unter standardisierten Situationen erfassen (durch den Itemstamm vorgegeben), und man möchte das Verhalten der Personen in diesen Situationen in einem vergleichbaren Format registrieren, dem Antwortformat. Dennoch kann einer der beiden Bestandteile eines Items bei einzelnen Tests bis zur Unkenntlichkeit *degeneriert* sein.

So bestehen z.B. die Items bei dem bekannten Tintenkleckstest von Rorschach (vgl. Kap. 2.2.2.4) allein aus den Tafeln, die in diesem Sinne den Itemstamm darstellen. Das Antwortformat ist schlicht das offene Ohr des Testleiters, meist eines Therapeuten, für die gesprochenen Ausführungen des Probanden zu dieser Tafel. Im anderen Extrem kann ein Item *nur* aus den Alternativen bestehen, zwischen denen man auswählen soll, also dem Antwortformat, eventuell mit dem Hinweis als 'Itemstamm', daß man die geeignete Alternative anzukreuzen habe.

Der *Normalfall* besteht jedoch tatsächlich darin, daß im Itemstamm eine Aufgabe gestellt wird, eine Frage gestellt wird oder eine Situation dargestellt ist, und mit einem geeigneten Antwortformat das Verhalten in dieser Situation, zu dieser Frage oder zu dieser Aufgabenstellung registriert wird.

Mit der Definition eines Items als kleinste Beobachtungseinheit ist auch gemeint, daß ein Item tatsächlich eine Einheit im Sinne von *'Einheitlichkeit'* darstellen muß. Ein

Item, das nach *zwei* Dingen gleichzeitig fragt, zwei unterschiedliche Aufgaben in einem stellt oder gleichzeitig zwei sehr unterschiedliche Stimuli beinhaltet, ist in der Regel ein unbrauchbares, zumindest problematisches Item: Das im Antwortformat registrierte Verhalten muß *eindeutig* auf die im Itemstamm vorgegebene Situation (Frage) zurückzuführen sein, wenn das Testverhalten Rückschlüsse auf die Personeneigenschaft erlauben soll.

2.3.1 Arten von Antwortformaten

Die wichtigste Unterscheidung bei Antwortformaten ist die Trennung nach freien (oder offenen) und gebundenen Antwortformaten.

In einem *freien Antwortformat* wird die Itemantwort von der getesteten Person selbst in einem allgemein verständlichen Zeichensystem formuliert wie z.B. in der Sprache, in Form von Zahlen, in Bildern, in Gesten oder in Lauten. Es bleibt dann dem Testleiter vorbehalten, diese wie auch immer registrierte Itemantwort zu verschlüsseln, d.h. in ein vorgefertigtes Kategoriensystem einzuordnen. Diesen Vorgang nennt man *Signierung* (s. Kap. 2.5.1). Der typische Fall von freien Antworten besteht in einer kurzen schriftlichen Antwort auf dem Testformular.

Auch freie Antworten erfordern ein *Format,* denn es wird ja vorgegeben, welche Art von Verhalten die Person produzieren soll, etwa ein Bild malen, einen Satz ergänzen, ein Muster fortsetzen, eine Zahlenreihe ergänzen oder eine Geschichte erzählen.

Ein *gebundenes Antwortformat* bietet demgegenüber eine Auswahl von Verhaltensalternativen an. Die Person braucht die Itemantwort nicht zu formulieren, sondern hat einen eingeschränkten Verhaltensbereich zur Verfügung, aus welchem eine Auswahl zu treffen ist. Der Vorteil dieser Antwortformate liegt darin, daß der Prozeß der Signierung, also der Einordnung der Itemantwort in Verhaltenskategorien entfällt.

2.3.1.1 Freie Antwortformate

Ein *freies Antwortformat* ist vorzuziehen, wenn es um die Erfassung *spontaner Reaktionen* geht, denn das Durchlesen von Verhaltensalternativen kann die Spontanität einschränken. Es ist auch bei der Erfassung *kreativer Leistungen* (was sich von selbst versteht) oder bei *Assoziationstests* sinnvoll, wo es darum geht, welche Assoziationen man zu einem vorgegebenen Stimulus hat. Auch in *projektiven* Testverfahren sind im allgemeinen freie Antwortformate angebracht, da das Durchlesen vorgegebener Antwortalternativen den Prozeß der Projektion stören kann.

Bei *Leistungstests* sind freie Antworten ein Mittel, um die Wahrscheinlichkeit einzuschränken, daß die richtige Antwort erraten wird. Generell ist auch bei solchen Befragungsinhalten ein freies Antwortformat vorzuziehen, bei denen sich die *Wichtigkeit* des Erfragten darin manifestieren kann, daß es der befragten Person *zuerst* einfällt.

Ein Beispiel hierfür ist die Erhebung von *Wertvorstellungen:* gibt man diese in einem gebundenen Format vor, so werden in der Regel *alle* als wichtig eingestuft.

Läßt man dagegen in einem freien Antwortformat diejenigen Werte nennen, die für die befragte Person wichtig sind, so fallen der Person unter Umständen wirklich nur diejenigen Werte ein, von denen sie sich leiten läßt.

Ein ganz anderes Kriterium für freie Antworten ist das *Alter* der befragten Personen. So kann es für Kinder durchaus schwierig sein, vor die Entscheidungssituation eines gebundenen Antwortformates gestellt zu werden, jedoch sehr viel einfacher, eine freie Antwort zu produzieren.

Innerhalb der freien Antwortformate lassen sich drei *Arten* von Antwortformaten unterscheiden.

Eine Art ist dadurch gekennzeichnet, daß - außer der Angabe des Mediums - so gut wie *keine weiteren Vorgaben* gemacht werden. D.h. die Person bekommt ein weißes Blatt Papier hingelegt mit dem Auftrag, z.B. die Mitglieder ihrer Familie als Tiere zu zeichnen (Familie-in-Tieren Test).

Ein zweiter Typ freier Antwortformate macht eine *formale Vorgabe* für die Produktion des Verhaltens, wie z.B. ein Wort aufzuschreiben, genau einen Satz zu formulieren, genau drei Dinge zu nennen, so viele Antworten wie möglich zu produzieren und diese so schnell wie möglich aufzuschreiben usw. Mit diesen formalen Vorgaben für die freie Produktion der Antwort kann eine gewisse *Standardisierung* des Tests erreicht werden und es können Fehlerquellen wie die Eloquenz (Redegewandtheit) der befragten Person kontrolliert werden.

Ein dritter Typ freier Antwortformate macht eine sogenannte *Lückenvorgabe*, d.h. die erfragte Itemantwort soll eine Leerstelle im vorgegeben Itemstamm ausfüllen. Dies ist z.B. der Fall, wenn die Aufgabe darin besteht, ein unvollständiges Bild oder einen Satz zu ergänzen oder Geschichten oder vorgegebene Muster fortzusetzen.

Der *Vorteil* von einschränkenden Vorgaben bei freien Formaten liegt zum einen in einer größeren Sicherheit für die getestete Person hinsichtlich dessen, was von ihr verlangt wird. Zum anderen lassen sich die Antworten leichter signieren, da sie, zumindest äußerlich, homogener sind.

Der *Nachteil* einschränkender Vorgaben ist darin zu sehen, daß die freie Produktion der Antworten behindert werden kann.

Bei der Auswahl eines freien Antwortformates ist unbedingt schon in der Planungsphase genau festzulegen wie die freien Antworten zu signieren sind.

Wird z.B. bei einem Kreativitätstest mit freiem Antwortformat lediglich ausgezählt, *wieviele* Ideen zu einem Stimulus produziert werden, unabhängig davon, wie ähnlich sich die Ideen, wie neu oder wie nützlich sie sind, so sollte das Antwortformat eine Zeitbegrenzung enthalten. Bei unbegrenzter Beantwortungszeit dürfte sich die Anzahl der Produktionen einander angleichen. Soll hingegen auch die Qualität der Produktion (Neuartigkeit, Brauchbarkeit) signiert werden, so ist ein Antwortformat ohne Zeitbegrenzung sinnvoller.

2.3.1.2 Gebundene Antwortformate

Gebundene Antwortformate haben zunächst den *Anschein einer höheren Objektivität* und sind tatsächlich oft auch objektiver, da die Auswertungsobjektivität sehr hoch ist. Die durch die vorgegebenen Antwortalternativen erzwungene Objektivität kann jedoch auch leicht zu *Lasten der Validität* des Tests gehen: Die vorgegebenen Alternativen schöpfen vielleicht nicht alle Reaktionsmöglichkeiten aus, das Durchlesen der Alternativen erzeugt bzw. beeinflußt die Antwort oder die vorgegebenen Antworten entsprechen in Formulierung und Stil nicht der natürlichen Reaktion der befragten Person.

Der Hauptvorteil gebundener Formate besteht in der *Auswertungsökonomie* des Tests, d.h. solche Tests sind schnell, von ungeschulten Auswertem und mit Schablonen auswertbar und somit bei Massenuntersuchungen einsetzbar. Tests mit freien Antworten können prinzipiell den gleichen Grad an Objektivität (und wissenschaftlicher Dignität) erreichen, aber verbunden mit einem höheren Aufwand.

Ein gebundenes Antwortformat besteht aus einem vorgefertigten System von Antwortmöglichkeiten. Die befragte Person ist an diese Antwortkategorien *gebunden*, also nicht frei in ihren Reaktionen.

Wie bei jedem *Kategoriensystem*, so stellt sich auch bei den vorgegebenen Antwortkategorien eines gebundenen Antwortformates die Frage, ob die Kategorien *disjunkt* sind, d.h. einander ausschließen, und ob die Menge der vorgegebenen Kategorien *exhaustiv* ist, d.h. den Bereich aller Verhaltensmöglichkeiten ausschöpft. Im Prinzip kann es bei Testitems alle vier Kombinationsmöglichkeiten von disjunkten und nicht-disjunkten und exhaustiven und nicht-exhaustiven Antwortkategorien geben. Beispiel:

Wie groß ist die Wurzel aus 2?

mit den Antwortkategorien:

	disjunkt	nicht disjunkt
exhaustiv	kleiner als 1.3. 1.3 bis 1.5 größer als 1.5	kleiner als 1.3. 1.2 bis 1.8 größer als 1.6
nicht exhaustiv	1.2 1.69 1.41	1.41 oder 1.73 1.21 oder 1.73 1.21 oder 1.41

Während bei Leistungstests nicht-exhaustive Formate sehr gebräuchlich sind, können sie bei anderen Testarten problematisch sein, da die befragte Person in die Situation kommen kann, eine Itemantwort geben zu wollen, die in den Antwortkategorien gar nicht vorgesehen ist.

Manchmal möchte man bewußt *keine Exhaustivität*, wenn man nämlich die befragte Person dazu zwingen will, eine Auswahl aus den vorgegebenen Alternativen zu treffen. Solche Antwortformate nennt man *'forced choice'* Formate (deutsch: erzwungene Wahl). Beispiel:

Was machen Sie, wenn ein guter Freund ein lang geplantes Treffen absagt ?

- *Ich verabrede mich mit jemand anderem.*
- *Ich gehe allein spazieren.*
- *Ich verrichte eine seit langem notwendige Arbeit.*

Die Funktion von solchen forced choice Formaten besteht darin, nur solche Reaktionen zuzulassen, die man nach der vorliegenden Theorie über die zu messende Personeneigenschaft auch *eindeutig interpretieren* kann.

Ihr Nachteil liegt selbstverständlich darin, daß sich die *Validität* des Tests verschlechtert, wenn die Personen in Ermangelung einer passenden Kategorie eine beliebige der vorgegebenen Kategorien ankreuzen.

Die Exhaustivität der Antwortkategorien ist jedoch nicht nur eine Frage des Antwortformates, sondern *auch des Itemstamms*. So kann ein exhaustiv formuliertes Antwortformat wie

- ja
- nein
- ich weiß nicht

für die befragte Person zu einem Problem werden, wenn sie am liebsten '*sowohl als auch*' oder '*weder noch*' antworten würde. Beispiel:

Sind Sie immer noch so glücklich wie früher?
Ja - Nein - Ich weiß nicht

Hier werden alle befragten Personen vor ein Problem gestellt, die früher gar nicht glücklich waren.

Bei *Leistungstests* sind die vorgegebenen Antwortkategorien im allgemeinen nicht exhaustiv und können es meistens auch gar nicht sein. Beispiel:

Welche Zahl setzt die folgende Zahlenreihe am besten fort?
2 3 4 9 8 27 16?
Antwortalternativen: 32, 18, 54 oder 81.

(Die richtige Zahl ist 81, da sie die Reihe $3 = 3^1$, $9 = 3^2$, $27 = 3^3$ fortsetzt.)

Solche Aufgaben haben eine *unendlich große Anzahl* möglicher Itemantworten, aus der nur eine kleine Anzahl zur Auswahl angeboten werden kann. Die richtige Itemantwort sollte natürlich darunter sein. Die aus der großen Anzahl *möglicher falscher Antworten* ausgewählten Antwortalternativen nennt man *Distraktoren.*

Wie wichtig die Auswahl geeigneter Distraktoren für die Itemkonstruktion ist, wird sofort einsichtig, wenn man sich vorstellt, die Antwortalternativen zum vorangehenden Beispiel lauteten:

1, 2, 3, 4 und 81.

Distraktoren haben die Funktion, die Identifikation der richtigen Antwort zu *erschweren.* Dies ist deswegen notwendig, weil der Lösungsprozeß bei gebundenen Antwortformaten grundsätzlich ein anderer ist als bei freien Antworten. Bei vorgegebenen Antwortalternativen werden in der Regel *alle* vorgegebenen Alternativen daraufhin geprüft, ob sie die angemessene Itemantwort darstellen. Je 'ähnlicher' die Antwortkategorien sind, desto schwieriger ist dieser Auswahlprozeß für die befragte Person.

Bei Leistungstestitems wird der Auswahlprozeß nicht nur durch die Ähnlichkeit der Antwortalternativen erschwert, sondern auch durch die *Plausibilität* der Distraktoren auf den ersten Blick. So kann der zeitliche Aufwand zur Lösungsfindung nahezu beliebig durch das Angebot sehr schwieriger Distraktoren gesteigert werden. Ein Beispiel hierfür ist das folgende Item aus dem Test für medizinische Studiengänge (TMS):

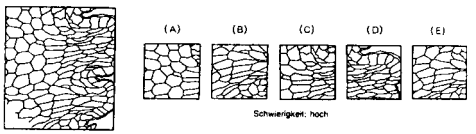


Abbildung 10: Ein Item aus dem TMS, (Inst. f. Test- und Begabungsforschung 1989)

Mit der Auswahl von Distraktoren kann jedoch nicht nur die Schwierigkeit eines Items variiert werden, sondern es können auch gezielt *halbrichtige* Lösungen oder bestimmte *Denkfehler* der befragten Personen erfaßt werden.

Ein Beispiel hierfür stellt der Würfeltest aus dem Intelligenzstrukturtest (IST) dar, bei dem es neben der richtigen Lösung auch immer einen Distraktor gibt, in dem der Würfel zwar die richtigen Flächen, aber in einer falschen Anordnung hat:

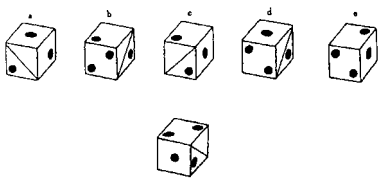


Abbildung 11: Ein Item aus dem IST (Amthauer 1970)

Während die Antwortkategorien bei Leistungstests im allgemeinen *nicht exhaustiv* sind, sollten sie jedoch stets *disjunkt* sein, wenn man nur *eine* Antwortalternative auswählen darf. Dies ist notwendig, damit die befragte Person ihre Itemantwort eindeutig in genau einer der vorgegebenen Antwortkategorien wiederfindet.

Nun gibt es aber Antwortformate, wo bewußt *mehrere* Antwortkategorien anzukreuzen sind oder sogar eine *beliebige* Anzahl, einschließlich der Möglichkeit gar keine anzukreuzen,

Bei Leistungstests bedient man sich oft dieses Tricks, um die *Ratewahrscheinlichkeit* zu senken. Bei Auswahl von nur einer Kategorie aus k vorgegebenen Kategorien beträgt die Ratewahrscheinlichkeit nämlich 1/k, also bei 5 Antwortalternativen 20%.

Soll man aus fünf Antwortalternativen zwei auswählen, sinkt die Ratewahrscheinlichkeit bereits auf 10%, da die Anzahl der möglichen Zweierkombinationen aus fünf Elementen (5.4)/2 = 10 beträgt.

Soll man eine *beliebige* Anzahl aus 5 Antwortkategorien auswählen, beträgt die Ratewahrscheinlichkeit nur noch 1/32, da es jeweils

5 Einerauswahlen,
10 Zweierauswahlen,
10 Dreierauswahlen und
5 Viererauswahlen

gibt, wo noch die beiden Möglichkeiten hinzukommen, daß gar keine Alternative oder alle Alternativen richtig sind.

Die Anzahl möglicher Kombinationen aus n Antwortalternativen läßt sich mit Hilfe des Binomialkoeffizienten

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{1 \cdot 2 \cdot 3 \ldots \cdot k}$$

berechnen (sprich ‘n über k’), der die Anzahl der Kombinationen von k Elementen aus einer Menge von n Elementen definiert. Die Anzahl *aller* möglichen Kombinationen ist dann über folgende Summe zu berechnen:

$$\sum_{k=0}^{n} \binom{n}{k}.$$

Die Ratewahrscheinlichkeit wird *minimal*, wenn man die Anzahl richtiger Antworten nicht vorgibt, sondern es dem Befragten überläßt, wieviele Alternativen er für richtig hält.

Solche Antwortformate werden nicht nur zur Senkung der Ratewahrscheinlichkeit in Leistungstests eingesetzt. Sie können auch im Rahmen von *Einstellungsmessungen* verwandt werden, z.B. wenn man aus einer Liste von Politikern die fünf erfolgreichsten oder eine beliebige Anzahl von vertrauenswürdigen Politikern auszuwählen hat.

Eine Auswahlanweisung, bei der die Anzahl auszuwählender Alternativen nicht vorgegeben ist, bezeichnet man auch als 'Pick any out of n'-Format.

Die in Kapitel 3 behandelten Testmodelle können mit solchen Mehrfachantworten nicht direkt umgehen, da sie *genau eine* Reaktion pro Person-Item-Kontakt voraussetzen. Diese Voraussetzung läßt sich auf zweierlei Weise nachträglich herstellen.

Erstens kann bei der *Kodierung* der Daten die Mehrfachantwort in eine Antwortvariable (mit disjunkten Kategorien) transformiert werden (s. Kap. 2.5). Bei Leistungstests wird dies in der Regel auch getan, indem nämlich nur die richtige Kategorienkombination als Itemlösung kodiert wird und alle anderen Kombinationen als Nicht-Lösung. Es sind aber auch Transformationen in eine ordinale Antwortvariable möglich, indem z.B. die *Anzahl* der angekreuzten richtigen Alternativen als Antwortvariable fungiert.

Der zweite Weg ist nur bei 'Pick *any* out of n' möglich und besteht darin, die Antwortalternativen selbst *als Items* mit einem dichotomen Antwortformat (gewählt oder nicht gewählt) aufzufassen. Im Falle einer vorgegebenen Anzahl von Auswahlen (Pick k out of n) ist dieser Weg nicht gangbar, da die experimentelle Unabhängigkeit zwischen den Items verletzt ist (s. Kap. 2.3.3).

Beispiel: Wenn man nur drei Politiker von 20 vorgegebenen auswählen kann, so haben nach drei erfolgten Wahlen die restlichen Politiker keine Chance mehr gewählt zu werden. Die 'Items' würden also keine unabhängigen Beobachtungseinheiten des Tests mehr darstellen.

2.3.1.3 Ratingformate

Unter den gebundenen Antwortformaten bilden die sogenannten Ratingformate eine häufig benutzte Untergruppe. Ein Ratingformat zeichnet sich durch zwei Eigenschaften aus. Erstens handelt es sich um mehrere, d.h. *mehr als zwei abgestufte Antwortkategorien*, von denen angenommen wird, daß sie für die befragte Person eine Rangordnung darstellen. Zweitens sind diese Antwortkategorien *item-unspezifisch* formuliert, d.h. dieselbe Benennung der Antwortkategorien gilt für mehrere oder *alle* Items eines Fragebogens. Diese itemunspezifischen, ordinalen Antwortkategorien nennt man *Ratingskala.*

Beispiel: 2 Items aus dem State-Trait-Anxiety-Inventory (STAI, Laux et al. 1981)

	fast nie	manch-mal	oft	fast immer
Item 34: Ich mache mir Sorgen über mögliches Mißgeschick	1	2	3	4
Item 38: Enttäuschungen nehme ich so schwer, daß ich sie nicht vergessen kann	1	2	3	4

Ratingformate haben gegenüber dichotomen Antwortformaten, bei denen nur zwischen Ja/Nein oder Zustimmung/Ablehnung unterschieden wird, den Vorteil, daß sie *informationsreicher* sind. Die befragte Person hat die Möglichkeit, sich gegenüber dem Iteminhalt differenzierter zu äußern und verschiedene Abstufungen ihrer Zustimmung oder Ablehnung auszudrücken.

Trotz der relativ klaren Definition einer Ratingskala und ihrer Vorteile, gibt es eine Vielzahl von *Varianten von Ratingskalen* und ebenso viele Punkte, die es bei der Testkonstruktion zu bedenken gilt. Die meisten dieser Überlegungen hängen damit zusammen, daß die Ratingskala eine *Ordinalskala* sein soll, d.h. von der befragten Person als solche benutzt und bei der Datenauswertung entsprechend verrechnet werden soll.

Oft besteht sogar der weitergehende Anspruch, daß die Ratingskala *Intervallskalenqualität* besitzt. Wenn man dies (ungeprüft) annehmen will, kann man auf die Itemantworten normale statistische Verfahren anwenden, die Intervallskalen voraussetzen. Bei den in Kapitel 3 behandelten Testmodellen wird *keine* Intervall-Skalenqualität von Ratingskalen vorausgesetzt, sondern lediglich Ordinalskalenqualität. Mit den geschätzten Modellparametern erhält man Information über die Kategorienabstände (also auch, ob sie äquidistant sind und somit eine Intervallskala bilden) *und* darüber, ob die Annahme des Ordinalniveaus gerechtfertigt ist.

Folgende Aspekte gilt es bei der Konstruktion einer Ratingskala zu beachten:

Erstens, soll die Skala *unipolar oder bipolar* aufgebaut sein?

Eine unipolare Skala geht von einem Nullpunkt lediglich in eine Richtung, d.h. zum Beispiel in Richtung auf eine starke Zustimmung. Die Ratingskala im o.g. Beispiel aus dem STAI ist unipolar in Richtung auf zunehmende Häufigkeit.

Bei bipolaren Ratingskalen gehen die Kategorien von einem negativen Pol (z.B. sehr starke Ablehnung) über einen fiktiven oder als Mittelkategorie vorgegebenen Nullpunkt bis hin zu einem positiven Pol (z.B. sehr starke Zustimmung). Eine bipolare Ratingskala ist im allgemeinen *symmetrisch,* d.h. sie hat gleich viele Kategorien auf jeder Seite. Sie muß es aber nicht sein, wie das folgende Beispiel aus dem Interaktions-Angst-Fragebogen (IAF, Becker 1982) zeigt:

> *Item 9: Sie denken daran, daß Sie von Ihrem Vorgesetzten abends eingeladen sind.*
> *Item 11: Es soll Ihnen vom Arzt mit einer dicken Nadel Blut entnommen werden.*

Die Ratingskala für alle Items lautet:

angenehm			*unangenehm*			
ziem-lich	ein wenig	weder noch	ein wenig	ziem-lich	sehr	äußerst

Dieser Fragebogen ist nicht nur ein Beispiel für eine bipolare *asymmetrische* Ratingskala, er zeigt auch, daß es manchmal problematisch sein kann, ein item-unspezifisches Antwortformat für alle Items zu verwenden. Es hängt sehr stark vom jeweiligen *Iteminhalt* ab, ob eine unipolare oder eine bipolare Ratingskala angemessen ist.

Darüber hinaus hängt die Entscheidung 'unipolar oder bipolar?' auch von der zu messenden *Personeneigenschaft* ab, die ihrerseits unipolar oder bipolar definiert sein kann (z.B. 'Extraversion-Introversion' als bipolares, 'Ängstlichkeit' als unipolares Konstrukt). Eine Korrespondenz der Art 'unipolares Konstrukt - unipolare Ratingskala' ist zwar nicht zwingend, aber es kann Schwierigkeiten bereiten, wenn man für eine unipolare Eigenschaft einen Gegenpol auf der Ratingskala konstruieren will.

So kann im obigen Beispiel der Gegenpol 'angenehm' zum Ängstlichkeit anzeigenden Pol 'unangenehm' auch zu konzeptuellen Problemen führen: Müssen Personen, die *extrem wenig* ängstlich sind, die genannten Situationen wirklich als 'ziemlich angenehm' einstufen? Wenn die Ratingskala Ordinalniveau haben soll, müßte man das erwarten.

Zweitens lassen sich Ratingskalen danach unterscheiden, wie differenziert sie das abgestufte Urteil erfassen, d.h. also *wie viele Stufen* die Ratingskala aufweist. Die Anzahl der Stufen sollte sich daran orientieren, welchen Differenziertheitsgrad im Urteil man den zu befragenden Personen 'zutrauen' kann. Dabei kommt so ziemlich jede Anzahl zwischen 3 und 10 in Frage.

Neben dem vermuteten Grad der kognitiven Differenziertheit der zu befragenden Personen spielt bei der Entscheidung über die Kategorienanzahl einer Ratingskala auch die Vermeidung sogenannter *Antworttendenzen* oder *response sets* eine große Rolle.

Response sets

Unter einem **response set** versteht man die von der zu messenden Personeneigenschaft unabhängige Neigung einer Person. die Ratingskala in einer bestimmten Art und Weise zu gebrauchen.

Es lassen sich folgende response sets unterscheiden:

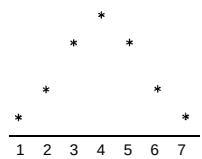
Tendenz zum mittleren Urteil
Tendenz zum extremen Urteil
Ja-sage-Tendenz (Aquieszenz)

oder auch deren jeweiliges Gegenteil, d.h.

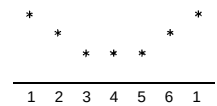
Vermeidung des mittleren Urteils,
Vermeidung eines extremen Urteils
und
Nein-sage-Tendenz.

Graphisch lassen sich response sets durch die jeweils entstehende Häufigkeitsverteilung der Antwortkategorien darstellen:

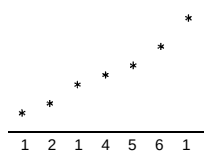
Tendenz zur Mitte



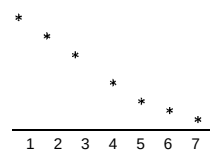
Tendenz zum Extrem

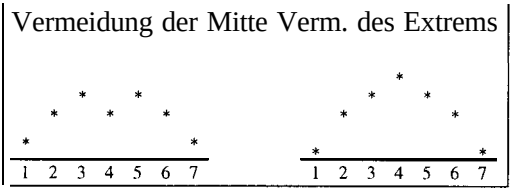


Ja-sage Tendenz



Nein-sage Tendenz





Auch *Kombinationen* aus diesen response sets oder *weitere Formen* können die Benutzung einer Ratingskala systematisch prägen.

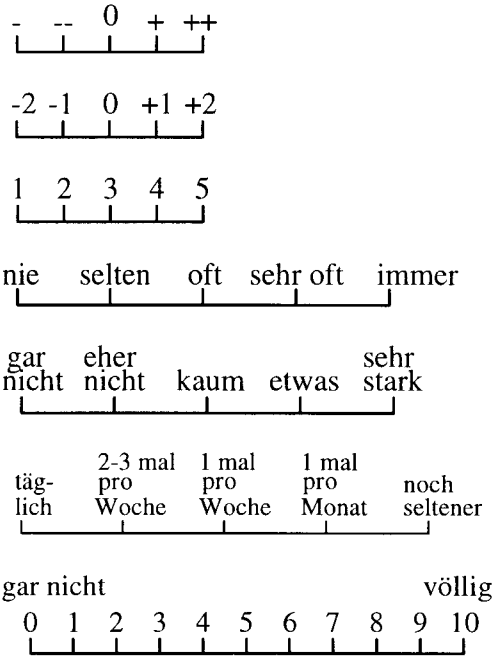
Bei der Konstruktion und Auswahl einer Ratingskala ist der Einfluß von response sets deswegen *möglichst gering* zu halten, weil er den Schluß von den Itemantworten auf die zu messende Personeneigenschaft beeinträchtigt, d.h. unsicherer macht. Die Beeinflussungsmöglichkeiten hängen insofern mit der Anzahl der Stufen der Ratingskala zusammen, als sich z.B. bei nur drei oder vier Antwortstufen eine Tendenz zum extremen Urteil weniger gravierend bemerkbar macht als etwa bei 7 Stufen.

Drittens unterscheiden sich Ratingskalen dahingehend, ob sie eine ungerade Anzahl von Kategorien - und damit eine neutrale oder *mittlere Kategorie* - haben oder eine gerade Anzahl.

In vielen Untersuchungen hat sich die Verwendung einer mittleren, neutralen Kategorie als ungünstig erwiesen. Diese Kategorie wird von den Personen oft nicht oder nicht nur als Ausdruck einer mittleren Position zwischen zwei Polen benutzt, sondern sie drückt aus, daß die Person das Item für unpassend hält oder die Antwort verweigert. Insofern ist die Persönlichkeitseigenschaft, von der die Benutzung dieser Kategorie abhängt, oft eine andere als die, die gemessen werden soll. Der Test ist in diesem Sinne dann *zweidimensional.*

Von Personen, die motiviert sind den Test zu bearbeiten, wird die mittlere Kategorie oft *gemieden,* d.h. sie tritt seltener auf als es aufgrund der Verteilung der zu messenden Eigenschaft zu erwarten ist. Dies führt dazu, daß die Parameter entsprechender Testmodelle anzeigen, daß die mittlere Kategorie mit den anderen Kategorien der Ratingskala keine Ordinalskala bildet. Die Qualität der Messung wird dann durch diese Kategorie eher beeinträchtigt als erhöht.

Viertens und letztens unterscheiden sich Ratingskalen nach der Benennung ihrer Kategorien. Im folgenden einige Beispiele:



Eine Benennung mit *Zahlen* wird oft verwendet, um zu bewirken, daß die Ratingskala wie eine *Intervallskala* benutzt wird.

Dies ist jedoch nicht automatisch garantiert, da in der subjektiven Wahrnehmung

von Personen auch aufeinanderfolgende ganze Zahlen nicht unbedingt gleichen Abstand haben.

Verbale Etikettierungen haben demgegenüber den Vorteil, daß die Bedeutung der Antwortstufen durch eine sprachliche Umschreibung intersubjektiv vereinheitlicht wird, was bei einer Kennzeichnung durch Zahlen nicht gegeben ist. Die Schwierigkeit bei der sprachlichen Benennung liegt jedoch darin, solche Beschreibungen zu finden, die eindeutig eine Rangordnung der vorgegebenen Kategorien ausdrucken.

Von einer *Kombination* aus beidem, d.h. numerische Bezeichnung der Stufen und verbale Beschreibung der Pole (siehe die obigen Beispiele), erhofft man sich die Vorteile von beiden Varianten.

Eine Belegung mit *Symbolen* wie Plus- und Minuszeichen soll wiederum die subjektiven Schwankungen in der Bedeutung sprachlicher Benennungen ausschließen und - gegenüber einer numerischen Etikettierung - den Eindruck übertriebener mathematischer Exaktheit vermeiden.

Häufigkeitsangaben als Etikettierungen der Ratingkategorien haben den Vorteil, daß sie einen verbindlichen, intersubjektiv definierten Maßstab als Beurteilungsskala anbieten und somit Urteilsfehler und den Einfluß von response sets auf ein Minimum reduzieren.

Ratingformate haben den Vorteil, daß sie nicht für jedes einzelne Item konstruiert werden müssen, sondern für alle Items eines Tests gelten. Dies ist auch für die befragte Person ein Vorteil, denn sie kann sich auf einen Antwortmodus einstellen und *gleichartige Maßstäbe für alle Items* bei ihrer Antwort benutzen.

Dennoch kann es sinnvoll sein, ordinal abgestufte Itemantworten *spezifisch für jedes Item* zu formulieren. Solche itemspezifischen ordinalen Antwortalternativen würde man nicht mehr als Ratingskala (im engeren Sinne) bezeichnen. Für die Auswertung solcher Daten müssen Testmodelle herangezogen werden, die unterschiedliche Kategorienabstände für jedes Item vorsehen.

Das Problem von response sets ist bei itemspezifischen Formaten differenzierter. Einerseits treten response sets seltener auf, da sich *Antwortgewohnheiten* schlechter herausbilden und manifestieren, wenn die Antworten bei jedem Item anders lauten. Andererseits können sie - wenn sie auftreten - schwerer identifiziert werden.

Der Vorteil itemspezifischer Antwortkategorien liegt jedoch darin, daß man sie auf den jeweiligen *Iteminhalt* beziehen kann. Ein Beispiel sind die beiden folgenden Items eines Interessentests:

Wenn Sie sich Ihre Freizeit allein nach Ihren Interessen gestalten könnten, wie häufig würden Sie...

ein Buch lesen		
mindestens 1 Std. tägl.	*etwa 5-8 Std. in der Woche*	*etwa 1-2 Std. in der Woche*

mit Freunden ausgehen			
mindestens 3mal pro Woche	*einmal pro Woche*	*1-2mal pro Monat*	*seltener*

2.3.2 Die sprachliche Formulierung der Items

Auch die einfachste Frage bleibt stets mehrdeutig und läßt dem Befragten einen Interpretationspielraum. Daher muß man eine gewisse Bereitwilligkeit voraussetzen, daß der Befragte die Frage auch so versteht wie sie gemeint ist. Bereits eine einfache Frage wie

Warum haben Sie dieses Buch gekauft?

hat je nach Betonung mindestens vier Interpretationen:

... *Was war die Motivation?*
... *Warum Sie und kein anderer?*
... *Warum gerade dieses Buch?*
... *Warum gekauft und nicht geklaut oder geborgt?*

Für derartige Probleme der sprachlichen Formulierung von Items kann es keine allgemeingültigen Anweisungen geben außer der, daß jede Frage oder *jedes Item nur einen einzelnen Aspekt* ansprechen sollte und nicht zwei oder drei gleichzeitig.

Im folgenden sollen einige Dichotomien dargestellt werden, nach denen sich Items einteilen lassen und die auch bei der Auswahl und Formulierung der Iteminhalte dienlich sein können.

Der Unterschied zwischen *direkten und indirekten* Fragen besteht darin, daß man in einem Item die zu messende Personeneigenschaft selbst ansprechen kann, z.B.

Halten Sie sich für rücksichtsvoll?

oder man Indikatoren erfragt, über die man indirekt auf die zu messende Eigenschaft schließt:

Halten Sie mit dem Auto an, wenn am Straßenrand eine Person steht, die offensichtlich die Straße überqueren möchte ?

Das Item kann sich auf einen *hypothetischen* oder *tatsächlichen* Sachverhalt beziehen, also z.B.

Was würden Sie tun, wenn...
oder
Haben Sie schon einmal . . . getan ?

Hypothetische Inhalte sind anfälliger gegenüber *Fehleinschätzungen* der eigenen Person, sozialer Erwünschtheit und anderen Fehlerquellen. Erfragt man tatsächliche Sachverhalte, so erhält man zwar 'harte Fakten' und ist von subjektiven Einschätzungen unabhängiger, jedoch ist die Itemantwort außer von der Personeneigenschaft noch von *situationalen Bedingungen* der befragten Person abhängig: Eine Person kann z.B. keine Gelegenheit gehabt haben, die erfragte Tätigkeit zu zeigen.

Das Item kann sich auf einen eher *konkreten* oder *abstrakten* Sachverhalt beziehen. Beispiel:

Sammeln Sie Briefmarken?
oder
Sammeln Sie gerne irgendwelche Sachen?

Auch hier stellen die konkreten Inhalte eher harte Fakten dar, die situationsabhängiger sind. Die allgemeinen Inhalte sind eher 'Einschätzungssache' und somit anfälliger für Urteilsfehler.

Die Frage kann *personalisiert* oder *depersonalisiert* gestellt werden. Beispiel:

Würden Sie gegen ein geplantes Kernkraftwerk demonstrieren?
oder

Sollten möglichst viele Menschen gegen geplante Kernkraftwerke demonstrieren?

Personalisierte Fragen lassen einen besseren Rückschluß auf die zu messende Eigenschaft zu, wenn sie ehrlich beantwortet werden. Sie können aber von der befragten Person als ein zu starker Eingriff in die *Privatsphäre* betrachtet werden und Widerstand gegen den Test bewirken.

Depersonalisierte Fragen wahren die Distanz, bergen aber die Gefahr, daß die Antworten nur allgemeine *Unverbindlichkeiten* ausdrucken ('Ja, ja, man sollte das tun . . . aber ich doch nicht').

Items können versuchen einen inneren Zustand neutral abzufragen, d.h. möglichst keine *Stimulusqualität* haben, oder sie können bewußt einen solchen Stimulus setzen, um die Reaktion darauf zu erfragen. Beispiel:

Sind Sie manchmal wütend über die lasche Haltung der Polizei gegenüber dem Rechtsradikalismus?
oder
Was empfinden Sie, wenn Sie hören, daß Jugendliche den Hitler-Gruß zeigen, ohne von der danebenstehenden Polizei behelligt zu werden?

Items mit Stimulusqualität (die zweite Formulierung) haben sicherlich den Vorteil, daß auch Personen ohne eine entsprechende Metakognition beim Durchlesen des Itemtextes *ihre eigene Reaktion* auf diesen Stimulus beobachten können und die Antwort daher gültiger ist.

Andererseits ist die Itemantwort bei solchen Items sehr stark *vom jeweiligen Stimulus abhängig,* was die Zuverlässigkeit des Testergebnisses schmälern kann. Bei dem obigen Beispiel hält vielleicht eine Person den Hitler-Gruß für eine vom Grundgesetz erlaubte freie Meinungsäußerung, während sie ansonsten für eine stärkere Bekämpfung des Rechtsradikalismus ist.

Nicht zuletzt kann man mit der sprachlichen Einkleidung des Iteminhaltes die *Schwierigkeit des Items* beeinflussen und gezielt steuern. Mit Schwierigkeit ist dabei gemeint, wie schwer es einer durchschnittlichen Person fällt, dem Iteminhalt zuzustimmen oder die Frage zu bejahen. Beispiel:

Die Polizei sollte Bundesbürger, die den Hitlergruß zeigen, sofort festnehmen und strafrechtlich verfolgen
oder
Das Zeigen des Hitlergrußes sollte vom Staat mit den zur Verfügung stehenden Rechtsmitteln geahndet werden.

Eine Manipulation der Itemschwierigkeit durch die sprachliche Einkleidung (die erste Formulierung dürfte 'schwieriger' sein) ist nichts Ungewöhnliches. Man *muß* als Testkonstrukteur sogar die Schwierigkeit in einem gewissen Rahmen beeinflussen, wenn man einen zuverlässigen Test entwickeln will: Bei einigen Testmodellen in Richtung auf eine *mittlere Schwierigkeit,* bei anderen Testmodellen in Richtung auf eine gleichmäßige Streuung oder *Staffelung* der Schwierigkeiten aller Items.

Weil die Itemschwierigkeit von der sprachlichen Formulierung abhängt, sind *deskriptive Ergebnisse von einzelnen Items,* z.B. '80% der Bevölkerung tolerieren den Hitler-Gruß', relativ wertlos, wenn nicht der vollständige Wortlaut und das Antwortformat der Frage mit genannt werden.

2.3.3 Die Zusammenstellung des Tests

Wie fügt man Items zu einem Test zusammen? Damit ist im wesentlichen die Frage gemeint, welche *Abhängigkeiten zwischen den Items* erlaubt sind und welche nicht.

Betrachtet man die Durchführung eines Tests als ein *Experiment* (S.O. Kapitel l), so stellt die Beobachtung des Verhaltens mehrerer Personen bei verschiedenen Items - in der Terminologie der Versuchsplanung - eine *Meßwiederholung* dar. Da alle Itemantworten von denselben Personen stammen, und *durch die zu messende Personeneigenschaft bedingt sind,* werden *keine unabhängigen* Beobachtungen realisiert.

Hält man die zu messende Personeneigenschaft jedoch konstant, z.B. indem man nur *eine* Person betrachtet oder nur Personen mit *derselben Ausprägung* der latenten Variable, so müssen die Items experimentell *unabhängig* bearbeitet werden.

Diese spezielle Art von Unabhängigkeit nennt man lokale stochastische Unabhängigkeit (stochastisch = wahrscheinlichkeitsmäßig). 'Lokal' bedeutet, daß die stochastische Unabhängigkeit nur für einen festen 'Ort' (locus = Ort) oder Wert der Personenvariable gilt.

Betrachtet man nur Personen mit demselben Wert der latenten Variable, so versteht man unter *stochastischer Unabhängigkeit* von zwei Items A und B, daß die Wahrscheinlichkeit einer bestimmten Antwort A_l bei Item A und Antwort B_l bei Item B gleich dem Produkt der beiden Einzelwahrscheinlichkeiten ist:

$$p(A_l \text{ und } B_1) = p(A_1) \cdot p(B_1).$$

Was diese Definition bedeutet, wird klar, wenn man sich anschaut, wie sich die Wahrscheinlichkeit der Antwortkombination *ohne* die Annahme der stochastischen Unabhängigkeit berechnen würde. Dann müßten die Wahrscheinlichkeiten der Antwortkombinationen auf *bedingte Wahrscheinlichkeiten* zurückgeführt werden:

$$p(A_1 \text{ und } B_1) = p(A_l) \cdot p(B_l \mid A_l)$$

oder

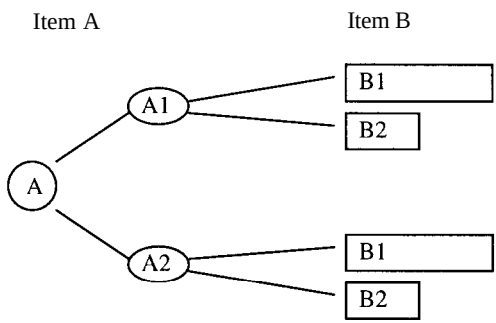
$$p(A_l \text{ und } B_l) = p(B_1) \cdot p(A_l \mid B_1).$$

$p(B_l \mid A_l)$ bezeichnet die Wahrscheinlichkeit von B_1 unter der Bedingung von At.

Vergleicht man diese beiden Gleichungen mit der obigen Definition, so zeigt sich, daß stochastische Unabhängigkeit nichts anderes bedeutet, als daß die *bedingten* Antwortwahrscheinlichkeiten gleich den *unbedingten* sind. Noch anders ausgedruckt:

Die Wahrscheinlichkeit einer Antwort auf Item B darf nicht davon abhängen, was die Person auf Item A (tatsächlich) geantwortet hat.

Folgende Graphik soll das demonstrieren:



Die Graphik stellt die Antwortwahrscheinlichkeiten des Items B mit den beiden Alternativen B_1 und B_2 in Abhängigkeit von den *tatsächlichen* Antworten auf Item A, A_1 und A_2, dar. Die Länge der Kästchen symbolisiert die Größe der Wahrscheinlichkeiten. Das Wahrscheinlichkeitsverhältnis von B_1 zu B_2 muß gleich bleiben, egal was bei Item A geantwortet wurde.

Es wird jedoch *keine* Aussage darüber gemacht wie das Wahrscheinlichkeitsverhältnis von A_1 zu A_2 ist. Ein häufiges *Mißverständnis* besteht darin, daß man meint, die *Wahrscheinlichkeiten* von A und B dürften nicht zusammenhängen. Über die Personen hinweg betrachtet ist das natürlich der Fall: wer Item A eher löst, wird auch Item B eher lösen, wenn es sich um einen homogenen Leistungstest handelt. Nur: durch die Lösung von Item A darf sich die Wahrscheinlichkeit für B nicht *verändern!*

Wann ist die Annahme der lokalen stochastischen Unabhängigkeit verletzt? Auf jeden Fall bei *logischen Abhängigkeiten* zwischen den Items. *Logische Unabhängigkeit* zwischen den Items bedeutet, daß die Beantwortung eines Items nicht *eine bestimmte Antwort auf ein anderes Item* voraussetzen darf. Ein Beispiel für logisch abhängige Items ist:

Item 1: Haben Sie schon einmal das Gefühl gehabt, daß Sie keinem Menschen trauen können ?

Item 2: Haben Sie daraufhin mit jemandem darüber gesprochen?

Es ist klar, daß die Beantwortung von Item 2 nur Sinn macht, wenn Item 1 bejaht wurde. Solche logischen Abhängigkeiten bilden das Prinzip von sog. *verzweigten Fragebögen*, bei denen jeweils vorgeschaltete 'Filterfragen' abklären sollen, ob die befragte Person überhaupt den folgenden Fragenkomplex zu bearbeiten hat (z.B. '...wenn nein, gehen Sie weiter zu Frage XY'). Zur testtheoretisch fundierten Messung *einer* Personeneigenschaft eignen sich solche abhängigen Items nicht, da die Konstruktion geeigneter Testmodelle sehr kompliziert ist.

Wenn sichergestellt ist, daß zwischen den Items keine logischen Abhängigkeiten bestehen, stellt sich als nächstes die Frage, wie die Items zu einem Test zusammengesetzt werden können, ohne die lokale stochastische Unabhängigkeit der Itemantworten zu gefährden. Hier gilt es, *Positionseffekte* und *Reihenfolgeeffekte zu* berücksichtigen.

Positions- und Reihenfolgeeffekte

Unter Positionseffekten versteht man die Veränderung der Schwierigkeit oder anderer Merkmale eines Items infolge seiner Plazierung im Test. Mit solchen Positionseffekten ist besonders bei den Items *am Testanfang* (mangelndes Instruktionsverständnis oder 'warming-up' Prozesse) oder *am Testende* (Ermüdung, Zeitmangel, schwindende Testmotivation und Abbruch) zu rechnen.

Unter Reihenfolgeeffekten versteht man die Beeinflussung der Itemantwort dadurch, *welche* anderen Items zuvor bearbeitet wurden. So sind bei vielen Leistungs- und Intelligenztests die Aufgaben *nach aufsteigender Schwierigkeit* geordnet, was sicherlich dazu beiträgt, daß die schwierigen Aufgaben infolge der Übung an leichteren Aufgaben ebenfalls etwas leichter zu lösen sind.

Solche Effekte können, müssen aber nicht die lokale stochastische Unabhängigkeit verletzen. Diese besagt lediglich, daß die Wahrscheinlichkeit einer Itemantwort nicht davon abhängen darf, *was* bei den vorangehenden Items *geantwortet* wurde. Sehr wohl darf sie davon beeinflußt sein, *welche* Items vorher *bearbeitet* wurden.

Solange sich ein Positions- oder Reihenfolgeeffekt darin ausdrückt, daß ein Item durch seine Position im Test oder durch vorangehende Items *für alle Personen* gleichermaßen leichter oder schwerer wird, ist die stochastische Unabhängigkeit *nicht* verletzt.

Bewirken diese Effekte dagegen *reaktionskontingente Veränderungen* der zu messenden Personeneigenschaft (reaktionskontingent = mit der Reaktion zusammenhängend), so ist die lokale stochastische Unabhängigkeit verletzt.

Hier ist insbesondere an *reaktionskontingentes Lernen* zu denken, also Lernvorgänge, die bei einer richtigen Itemlösung anders ablaufen als bei einer falschen Lösung. Leider sind die meisten oder zumindest die interessanteren Lernvorgänge reaktionskontingent. *So* etwa *Lernen durch Einsicht*, das sich einstellt, wenn man eine Aufgabe 'zufällig' gelöst hat (Aha-Erlebnis), oder *Verstärkungslernen*, wenn die richtige Lösung (sofern man auch merkt, daß sie richtig ist) als Verstärker für den richtigen kognitiven Prozeß fungiert. Auch wenn man bei einer *erfolglosen Bearbeitung* von Aufgaben mehr lernt als bei einer richtigen Lösung, liegt reaktionskontingentes Lernen vor. Sollten solche Lernprozesse massiv auftreten, wäre die stochastische Unabhängigkeit der Items nicht gegeben.

Findet dagegen lediglich Lernen im Sinne von *Üben* statt, was im wesentlichen von der Anzahl und Qualität der Aufgaben aber nicht von den eigenen Reaktionen abhängt, so ist das eine Form von Lernen, die mit der Annahme der stochastischen Unabhängigkeit *vereinbar* ist.

Auch bei anderen Tests als Leistungsstests kann es zu reaktionskontingenten Veränderungen der zu messenden Personeneigenschaft kommen. Ein Beispiel wäre ein Aggressionstest, bei dem aggressive Reaktionen auf frühere Items einen *kathartischen Effekt* haben (Katharsis = Läuterung) und somit die Wahrscheinlichkeit aggressiver Reaktionen auf spätere Items senken.

Was folgt aus diesen Überlegungen für die Zusammenstellung von Items zu einem Test?

Erstens dürfen Items, die dieselbe Personeneigenschaft messen, *nicht logisch voneinander abhängig* sein.

Zweitens sollte man möglichst eine *Zufallsabfolge* wählen. Möchte man durch eine *gezielte Anordnung* bestimmte Reihenfolge- oder Positionseffekte ausnutzen, so muß sichergestellt sein, daß diese Effekte auf alle Personen gleichermaßen wirken und *nicht* davon beeinflußt sind, *wie* eine Person bestimmte Items beantwortet.

Zur Vermeidung unerwünschter Abhängigkeiten zwischen den Items gibt es einige *Tricks*. So kann man z.B. *Scheinitems* in den Test einstreuen,

- die eine befürchtete Kontingenz zwischen aufeinanderfolgenden Items durchbrechen sollen *(Puffer-Items)*,
- die die zu messende Persönlichkeitseigenschaft *verschleiern* sollen,

- oder die 'ganz nebenbei' vermutete Störvariablen *(Tendenz zur sozialen Erwünschtheit, Ja-sage-Tendenz)* erfassen sollen.

Weiterhin kann man *sensible* Items, deren Antwort durch vorangehende Items beeinflußt werden könnte, *an den Anfang* stellen. Und man kann *reaktive* Items, d.h. solche deren Beantwortung Effekte auf spätere Itemantworten ausüben können, *an den Schluß* stellen.

Soll ein Testinstrument zur Messung mehrerer Personeneigenschaften zusammengestellt werden, ergeben sich weitere Möglichkeiten, wie z.B. die Items verschiedener Untertests *zu mischen.*

Literatur

Eine detaillierte Diskussion der Konstruktion von Fragebögen und Tests findet sich bei Lienert (1969), Kline (1986), Mummendey (1987), Roid & Haladyna (1982) und Tränkle (1983). Hornke und Rettig (1992) diskutieren am Beispiel von Analogieitems Ansätze einer theoriegeleiteten Itemkonstruktion, Esser (1977) geht auf die Problematik von response sets ein, Couch & Keniston (1960) gehen speziell auf die Ja-sage Tendenz ein. Dubois & Bums (1975) analysieren die Rolle einer 'Ich-weiß-nicht' Kategorie.

Übungsaufgaben

1. Man möchte in einer schriftlichen Befragung von *Ihnen* wissen, welche übergeordneten Werte für Sie in Ihrem Leben und für Ihr Handeln wichtig sind. Wie sollten die Items aussehen, auf die Sie am ehesten und am ehrlichsten antworten würden? Formulieren sie 3 Beispielitems mit unterschiedlichem Antwortformat und diskutieren Sie Vor- und Nachteile.
2. Formulieren Sie für folgende Items Distraktoren, die es Ihnen ermöglichen, auch Denkfehler zu erfassen:
   Wieviel ist 4^3 (4 hoch 3)?
   Wer war zur Zeit der großen Koalition Bundespräsident?
   Wieviel kostet es, eine 60 Watt Lampe 5 Stunden brennen zu lassen, wenn die Kilowattstunde 20 Pfennige kostet?
3. Wie groß ist die Ratewahrscheinlichkeit bei einem Item mit 6 Antwortkategorien, wenn genau 3 richtige Antworten dabei sind (und das Item nur als gelöst gilt, wenn alle richtigen angekreuzt werden)? Ist die Ratewahrscheinlichkeit kleiner, gleich oder größer, wenn es genau 4 richtige Antworten gibt?
4. Sie möchten als Indikator für Ausländerfeindlichkeit die Bereitschaft erfragen, direkt neben einem Asylbewerberheim zu wohnen. Formulieren sie 3 möglichst unterschiedliche Items für diesen Indikator und diskutieren Sie die Vor- und Nachteile.
5. In einer Stichprobe von Personen mit derselben Ausprägung der zu messenden Fähigkeit erhalten Sie die folgenden Lösungshäufigkeiten von 2 Items A und B:

   A und B gelöst: 35%
   A gelöst, B nicht: 5%
   B gelöst, A nicht: 25%
   Weder A noch B gelöst : 35%

   Zeigen Sie, daß hier die Annahme der lokalen stochastischen Unabhängigkeil nicht gilt. Wie müßten die 4 o.g. prozentualen Häufigkeiten aussehen, wem die Annahme gilt?

2.4 Datenerhebung

Mit Datenerhebung ist in diesem Kapitel die Sammlung von Testdaten zum Zwecke einer *Testentwicklung* oder im Rahmen einer *Forschungsarbeit* gemeint. Fragen der Datenerhebung im Sinne einer *Testanwendung* für diagnostische Zwecke werden hier nicht behandelt.

2.4.1 Stichprobenprobleme

Jede Stichprobenziehung fängt mit der Definition der *Population* an, über die die Stichprobe etwas aussagen soll. Wie bei jeder empirischen Untersuchung ist auch bei einer Testentwicklung eine *repräsentative Stichprobe* optimal. Repräsentativität bedeutet, daß *alle denkbaren Variablen* in der Stichprobe genauso verteilt sind wie in der Population. Eine repräsentative Stichprobe ist somit nur durch eine völlig *zufällige Auswahl* der Individuen aus der Population herzustellen.

Eine solche Zufallsauswahl ist in der Praxis so gut wie nie erreichbar und es stellt sich daher die Frage, *welche Eigenschaften* einer repräsentativen Stichprobe für eine Testentwicklung wirklich wichtig sind und mit welcher Art der Stichprobenziehung diese Eigenschaften gewonnen werden können.

Hier muß wieder nach den *Zielen* der Testentwicklung unterschieden werden:

Soll der Test *normiert* werden (s. Kap. 2.1.5), *so* muß die *Verteilung der zu messenden Personenvariable* in der Stichprobe völlig identisch sein mit der Verteilung in der Population.

In Abbildung 12 symbolisiert die durchgezogene Linie die Häufigkeitsverteilung der Meßwerte X in der Population und die gestrichelte Linie die Stichprobenverteilung.

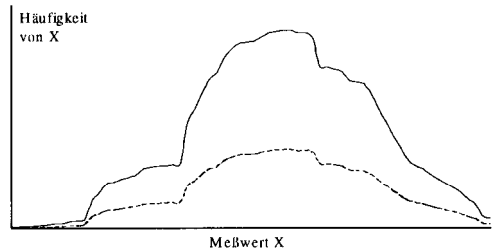


Abbildung 12: Populations- und Stichprobenverteilung

So ein nahezu identisches Abbild der Populationsverteilung des zu messenden Merkmals ist tatsächlich nur notwendig, wenn Normen für die Testinterpretation entwickelt werden sollen (s. Kap. 2.1.5 und 6.5). Dann sind allerdings sogar oft *mehrere* repräsentative *Stichproben* für verschiedene Teilpopulationen erforderlich, je nachdem für welche Referenzpopulationen getrennte Normen gewünscht werden.

Die Art der Stichprobenziehung soll in diesem Fall lediglich sicherstellen, daß die *zur Selektion benutzten Variablen* nicht mit der zu messenden Variable zusammenhängen. So darf z.B. die Tatsache, daß jemand ein Telefon besitzt, nicht mit der zu messenden Eigenschaft zusammenhängen, wenn die Stichprobe durch telefonische Anfrage rekrutiert werden soll (aus Telefonbüchern lassen sich leicht Zufallsstichproben ziehen).

Sollen demgegenüber keine Normen entwickelt werden, sondern soll ‘lediglich’ ein *meßgenauer und valider Test* entwickelt werden, so schwächen sich die Erfor-

dernisse an die Verteilung der Eigenschaft in der Stichprobe deutlich ab. Es sind im wesentlichen zwei Dinge zu gewährleisten:

Erstens, sollte die *Variation* der zu messenden Eigenschaft in der Stichprobe gegenüber der Populationsvariation *nicht eingeschränkt* sein. Dieser Punkt ist besonders wichtig, wenn eine *externe Validität* des Tests berechnet wird. Jede Einschränkung der Varianz der Meßwerte bewirkt nämlich eine *Unterschätzung der Validität.*

Dies wird in der folgenden Abbildung veranschaulicht.

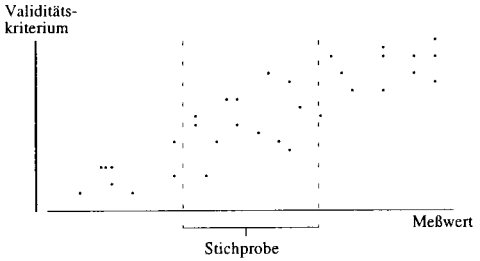


Abbildung 13: Korrelation zwischen Meßwert und Validitätskriterium

Die Graphik zeigt die Korrelation zwischen Meßwerten und externem Validitätskriterium in der Population. Innerhalb des eingeschränkten Variationsbereiches der Stichprobe fällt die Punktewolke wesentlich 'runder' und damit die Korrelation (sprich: externe Validität) niedriger aus.

Obwohl die Konsequenzen *gegen* die Intentionen des Testkonstrukteurs gerichtet sind, ist die Varianzeinschränkung wohl *einer der häufigsten Fehler,* der bei der Stichprobenziehung begangen wird. Man denke nur an die vielen Testentwicklungen, die ausschließlich an studentischen Stichproben vorgenommen werden.

Eine eingeschränkte Varianz der zu messenden Eigenschaft in der untersuchten Stichprobe wirkt sich auch auf andere Berechnungen im Rahmen einer Testentwicklung nachteilig aus. So kann die Qualität des Items nicht so gut beurteilt werden, wenn die Varianz der latenten Variable eingeschränkt ist (vgl. Kap. 6.2.1).

Für die Stichprobenziehung kann man daraus die Konsequenz ableiten, *mehrere* möglichst unterschiedliche *Teilstichproben* zu untersuchen, um so die Variation zu erhöhen.

Der zweite Punkt, der auch bei einer nichtrepräsentativen Stichprobe gewährleistet sein sollte, besteht darin, daß die Art der Abhängigkeit von Testverhalten und Personeneigenschaft in der Stichprobe *nicht untypisch* für die Art der Abhängigkeit in der Gesamtpopulation ist. Entwickelt man etwa einen Angstfragebogen ausschließlich an einer Stichprobe von Personen mit akademischer Bildung, so ist damit vielleicht nicht die Variation der Eigenschaft 'Ängstlichkeit' eingeschränkt. Es kann aber sein, daß der rationale Umgang mit dem Phänomen 'Angst' und somit die Beziehung von Ängstlichkeit und Testverhalten in dieser Stichprobe anders aussieht als in anderen Teilpopulationen.

Der letztgenannte Punkt betrifft primär die Sicherstellung der *internen Validität* des Tests. Diese ist aber Voraussetzung für jegliche sinnvolle Verwendung des Tests.

Abschließend noch ein paar Antworten auf die zentrale Frage: *Wie groß soll die Stichprobe sein?*

Diese Frage läßt sich unter drei Gesichtspunkten beantworten, je nachdem welches

Ziel oder Gütekriterium eines Tests man vor Augen hat:

- die Prüfung der Modellgeltung (die *interne Validität* des Tests)
- die *Genauigkeit* der Parameterschätzungen
- die Entwicklung von *Normen.*

Strebt man eine möglichst *exakte Prüfung der Modellgeltung* an, so kann das leicht zu astronomischen Stichprobengrößen führen. So lautet die (Maximal-) Antwort auf die o.g. Frage, daß man ein *Mehrfaches* (z.B. 5-faches) der *Anzahl möglicher Antwortmuster* in einem Test braucht.

Diese Antwort hat folgenden Hintergrund: Die theoretisch befriedigendste Methode, ein Testmodell vollständig auf Gültigkeit zu prüfen, verlangt, daß man die *beobachteten Häufigkeiten unterschiedlicher* Antwortmuster mit den vom Modell vorhergesagten *Häufigkeiten aller möglichen* Antwortmuster vergleicht.

Besteht ein Test z.B. aus zehn Items mit je zwei Antwortmöglichkeiten, so gibt es 2^{10} = 1024 unterschiedliche Antwortmuster:

Item:	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	0	1	1
5	0	0	0	0	0	0	0	1	0	0
6	0	0	0	0	0	0	0	1	0	1
7	0	0	0	0	0	0	0	1	1	0
1020	1	1	1	1	1	1	1	0	1	1
1021	1	1	1	1	1	1	1	1	0	0
1022	1	1	1	1	1	1	1	1	0	1
1023	1	1	1	1	1	1	1	1	1	0
1024	1	1	1	1	1	1	1	1	1	1

Ein Vergleich von theoretisch erwarteten und beobachteten Häufigkeiten der Antwortmuster würde voraussetzen, daß jedes dieser 1024 Antwortmuster eine reelle Chance hätte beobachtet zu werden. Dies ist wohl erst bei *ein paar tausend getesteten* Personen der Fall.

Natürlich gibt es auch 'sparsamere' Formen, die Geltung eines Testmodells zu testen, aber dann wird die Antwort auf die Frage 'Wieviel Personen?' zur Ermessenssache.

Eine sparsamere Form der Geltungsprüfung besteht darin, die *Stichprobe* bei der Testauswertung *in zwei Hälften zu teilen* und die Parameterschätzungen in beiden Teilstichproben miteinander zu vergleichen. Diese Methode setzt voraus, daß die halbe Stichprobengröße ausreicht, die Modellparameter zu schätzen. Da dies bei vielen Modellen schon mit etwa 50 Personen möglich ist, kommt man zu einem minimalen *Stichprobenumfang von ca. 100 Personen.*

Hat man sehr *starke a priori Hypothesen* (a priori (lat.) = im vorhinein), z.B. über die Rangordnung der Schwierigkeiten der Testitems, so reichen auch *40-50 Personen* aus. Die Prüfung der internen Validität des Modells kann dann über den Vergleich der empirisch geschätzten Modellparameter mit den hypothetischen erfolgen.

Innerhalb dieses Spielraumes von 50 bis 5000 Personen kann man nur differenziertere Aussagen machen, wenn man sich auf ein spezielles Testmodell bezieht. So reichen für Modelle mit einer quantitativen Personeneigenschaft im allgemeinen etwas *kleinere Stichprobenumfänge*

aus als für Modelle mit kategorialer Personeneigenschaft.

Geht man von der *Genauigkeit der Parameterschätzungen* aus, so lassen sich Empfehlungen für Stichprobengrößen ebenfalls nur modellspezifisch ableiten. Anzumerken ist hier, daß für die Genauigkeit der Meßwerte der Items ausschließlich die *Anzahl der Personen* maßgeblich ist. Umgekehrt wird die Genauigkeit der Personenmeßwerte ausschließlich von der *Anzahl der Items* beeinflußt.

Insofern ist die Erreichung einer hohen Meßgenauigkeit (Reliabilität) des Tests keine Frage der Größe der Personenstichprobe. Man kann jedoch Ansprüche an die *Genauigkeit der Itemmeßwerte* stellen und daran die Stichprobengröße orientieren. In welcher Weise die Stichprobengröße mit der Meßgenauigkeit der Items zusammenhängt, wird in Kapitel 6.1 behandelt.

Nimmt man die *Ableitung von Normen* als Kriterium für die Bestimmung der Stichprobengröße, so stellt sich zunächst die Frage, *wie differenziert* man denn die Normen haben möchte. Im einen Extrem kann man allein daran interessiert sein, wie groß der *Mittelwert* einer quantitativen Personeneigenschaft in einer Referenzpopulation ist. Hier können schon 20 bis 30 Personen ausreichen, um den Populationsmittelwert einigermaßen genau zu bestimmen.

Im anderen Extrem kann man z.B. alle 100 Prozentmarken der Verteilung der Meßwerte in einer Population bestimmen wollen. Hierfür sind dann schon ca. 2000 Personen erforderlich; das ist eine Stichprobengröße, die sich auch für Meinungsumfragen und Wahlprognosen als hinreichend erwiesen hat. Nicht zuletzt muß berücksichtigt werden, für wieviele Teilpopulationen, die z.B. nach Geschlecht, Alter oder Berufsgruppe aufgeschlüsselt sind, Normtabellen entwickelt werden sollen. Hier lassen sich keine allgemeingültigen Empfehlungen geben.

2.4.2 Durchführungsprobleme

Bei der Durchführung der Datenerhebung sind einige Probleme zu bedenken, die es bei einem Einsatz des Tests zu individualdiagnostischen Zwecken so nicht gibt.

Hierzu gehört zunächst die *Aufklärung über den Gegenstand der Befragung*. In vielen Fällen ist es sinnvoll, wenn die befragten Personen möglichst wenig über den Gegenstand der Befragung wissen, damit die *Itemantworten unbeeinflußt* bleiben von Vorkenntnissen. Das bewußte Verschweigen des eigentlichen Gegenstands einer Befragung oder gar die Vorspiegelung einer falschen Testabsicht wirft jedoch *ethische Probleme* auf.

Wie bei vielen Experimenten, bei denen man vor demselben Problem steht, wird es im Allgemeinen für ethisch vertretbar gehalten, wenn man die Befragten vorher informiert, *daß* man den Gegenstand der Befragung vor der Testbearbeitung nicht offenbaren kann, aber ankündigt, daß man dies *im Anschluß* nachholt. Eine falsche Cover-story erfordert in jedem Fall eine nachträgliche Richtigstellung.

Neben den ethischen Problemen hat ein Verschweigen oder eine Falschinformation auch den Nachteil, daß *fälschliche Vermutungen* über den Befragungsgegenstand die Itemantworten ebenso nachteilig oder

noch ungünstiger beeinflussen können, wie die richtige Information.

Ein *Beispiel* wäre, wenn man einen Fragebogen zu moralischen Wertvorstellungen damit zu kaschieren versucht, daß man vorgibt, es handele sich um einen Fragebogen zum politischen Konservativismus. Die Testergebnisse über die Moralvorstellungen hängen dann auch davon ab, wie konservativ sich die Befragten darstellen möchten.

Eine Information über den Sinn der Befragung ist unter anderem auch deshalb notwendig, um eine Bereitschaft zur Testbearbeitung zu schaffen, die sogenannte *Testmotivation.* Jeder Befragte braucht irgendeinen Grund, eine Motivation, den Test möglichst sorgfältig und ehrlich zu beantworten.

Ein solches Motiv ist bei der späteren individualdiagnostischen Verwendung eines Tests in der Regel automatisch gegeben. Bei der Testentwicklung mittels zufälliger Personenstichproben ist diese Testmotivation erst herzustellen.

Da die *ethischen Richtlinien* für die Durchführung von Humanexperimenten (und um solche handelt es sich bei Tests) verlangen, daß die Teilnahme *freiwillig* ist, sollte man sich der *Bereitschaft* der zu befragenden Personen vorher vergewissern und gegebenenfalls *Anreize* zur Bearbeitung des Tests schaffen. Inwieweit die dabei induzierte Testmotivation die *Beantwortung der Items beeinflussen* kann, ist im Einzelfall abzuwägen.

Ein weiterer Punkt, in dem sich die Entwicklungsphase eines Tests von seiner individualdiagnostischen Verwendung unterscheidet, liegt in der Zusicherung der *Anonymität*. Dabei ist die Zusicherung leichter gegeben als eingehalten, denn es müssen *organisatorische Maßnahmen* getroffen werden, um zu verhindern, daß der Testleiter im nachhinein die Identität der Befragten rekonstruieren kann (nicht zu viele demographische Variablen, wie Alter, Geschlecht, Beruf etc. erfragen).

Schließlich müssen die *Bearbeitungshinweise* für den Test so einfach und so genau wie möglich formuliert werden. Hierzu gehören im allgemeinen

- ein oder zwei *Itembeispiele* mit möglicher Antwort
- eine Angabe, wieviel *Zeit* die Bearbeitung insgesamt in Anspruch nimmt,
- Hinweise, was man tun soll, wenn man ein Item *nicht beantworten* will, und
- bei Leistungstests, ob man bei zu schweren Items die Antwort *raten* oder das Item lieber überspringen soll.

Spezielle Arten der Datenerhebung bringen auch spezifische Durchführungsprobleme mit. So stellt sich bei einer *postalischen Befragung* das Problem, eine hohe *Rücklaufquote* zu erreichen. Darunter versteht man den prozentualen Anteil zurückgesandter Fragebögen an der Gesamtzahl versandter Fragebögen. Je nach Umfang und Inhalt der Fragebögen muß man manchmal schon mit einer Rücklaufquote von 50% zufrieden sein.

Das Problem einer geringen Rücklaufquote ist nicht die Verkleinerung des Stichprobenumfangs. Diese kann dadurch ausgeglichen werden, daß man von vornherein mehr Personen anschreibt als benötigt werden. Das Problem stellt die sogenannte *Eigenselektion* dar. Damit ist gemeint, daß die befragten Personen selbst

entscheiden, ob sie den Fragebogen beantworten und zurücksenden. Die Kriterien, nach denen diese Auswahl (Selektion) erfolgt, hängen in der Regel mit dem Gegenstand der Befragung zusammen, so daß die zurückerhaltenen Fragebögen eine verzerrte Stichprobe des *Antwortverhaltens* darstellen.

Hat man den befragten Personen Anonymität zugesichert, läßt sich diese Verzerrung auch nicht dadurch beeinflussen, daß man säumige Personen anmahnt oder Ersatzpersonen sucht, die in demographischen Merkmalen vergleichbar sind.

Telefonische Befragungen eignen sich naturgemäß nur für Erhebungen von geringem zeitlichen Umfang. Sie werden insbesondere im Bereich soziologischer Untersuchungen eingesetzt.

Spezielle Möglichkeiten und Probleme ergeben sich auch durch den Einsatz des Computers bei der Testvorgabe. Das *computerunterstützte Testen* stellt eine Form der Datenerhebung dar, die es erlaubt, die Auswahl der Testitems individuell auf jede Person abzustimmen. Die höchste Stufe dieses maßgeschneiderten Testens (taylored testing) besteht darin, jede Itemantwort sofort zu verarbeiten und für die Auswahl des nächsten Items zu nutzen.

Das *Prinzip der Passung* von Itemschwierigkeit und Personenfähigkeit (s. Kap. 2.2.4) kann dadurch optimal realisiert werden, daß schon nach wenigen bearbeiteten Items eine erste Schätzung der Fähigkeit der Person vorgenommen wird. Die folgenden Items werden dann so ausgewählt, daß die betreffende Person in etwa eine 50%-ige Lösungswahrscheinlichkeit hat. Auf diese Weise kann eine relativ hohe Meßgenauigkeit realisiert werden und die getesteten Personen müssen sich nicht mit zu leichten oder zu schweren Items beschäftigen.

Außerdem dient das computerunterstützte Testen der Standardisierung der Testdurchführung und damit der Objektivität der Ergebnisse. Auf die vielen technischen Aspekte der Computernutzung beim Testen kann hier jedoch nicht eingegangen werden. Mehrere Beiträge zum computerunterstützten Testen finden sich in dem Sammelband von Kubinger (1988).

Literatur

Allgemeine Fragen der Stichprobenziehung werden in Lehrbüchern der empirischen Forschung behandelt, s. z.B. Bortz (1984), Schnell et al. (1989). Auf einige Aspekte der Testdarbietung geht Lienert (1969) ein.

Übungsaufgaben

1. Wie wirkt sich eine eingeschränkte Varianz des zu messenden Merkmals in der Stichprobe auf die Validität und die Reliabilität des Tests aus? Ziehen Sie zur Beantwortung der Frage die Definition von Reliabilität in Kapitel 2.1.2 heran.
2. Sie haben einen Test mit 3 dichotomen, 3 dreikategoriellen und 3 vierkategoriellen Items. Wieviele Personen müßte Ihre Stichprobe umfassen, damit alle möglichen Antwortmuster mindestens einmal beobachtet werden können?

2.5 Kodierung der Antworten

Den Vorgang, die Itemantworten der befragten Personen aus dem Testheft oder dem Antwortblatt derart in Zahlen zu verschlüsseln, daß diese Daten dann mit einem entsprechenden Testmodell analysiert werden können, nennt man *Kodierung* der Itemantworten. Die Kodierung der Antworten ist bereits ein Vorgang, bei dem berücksichtigt werden muß, *wie* die Daten ausgewertet werden sollen. Im Zweifelsfalle empfiehlt es sich, möglichst die ganze, in den Antworten vorhandene Information differenziert zu kodieren, denn eine Rekodierung durch *Zusammenlegen* von Kategorien ist jederzeit möglich. Der umgekehrte Weg, d.h. eine nachträgliche Ausdifferenzierung von zu groben Kategorien ist dagegen nur unter erheblichem Aufwand möglich.

Für Items mit freien Antwortformaten läßt sich der Prozeß der Kodierung in *zwei Phasen* unterteilen, nämlich die Zuordnung der freien Antworten zu bestimmten Kategorien und die Zuordnung von Zahlencodes zu diesen Kategorien. Man bezeichnet den ersten Schritt als *Kategorisierung* oder mit einem älteren Begriff als *Signierung* der freien Antworten. Der Begriff Signierung stammt aus der Auswertung projektiver Tests, bei denen die Kategorisierung der Itemantworten ein hohes Maß an psychologischer Schulung erfordert. Die beiden Phasen der Transformation einer Itemantwort in die Antwortvariable zeigt Abbildung 14. Die dritte Phase ist die Transformation der Antwortvariablen in einen Meßwert mittels eines Testmodells (s. Kap. 3) und der Schätzung seiner Parameter (Kap. 4).

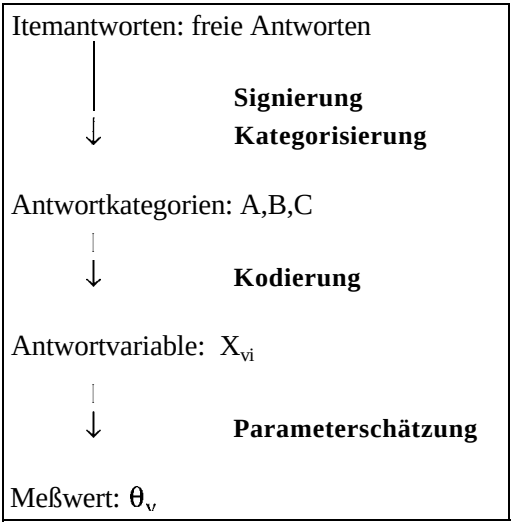
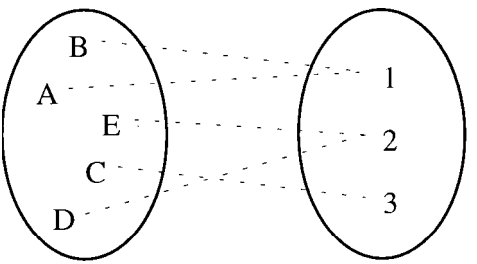


Abbildung 14: Phasen der Transformation einer freien Antwort in einen Meßwert

Das Ziel der beiden ersten Phasen besteht darin, für jedes Item i eine *Antwortvariable* X_{vi} zu erhalten.

Was ist eine Variable?

In der Sprache der Mengenlehre versteht man unter einer *Variable* eine eindeutige Zuordnung (Abbildung) einer Menge von Objekten zu einer Menge von Zahlen. Das bedeutet, daß dieselbe Zahl zwar mehreren Personen (Objekten) aber nicht dieselbe Person mehreren Zahlen zugeordnet werden kann:



Das Wesen einer *Variable* besteht darin, jedem Objekt, in diesem Fall: jeder Per-

son, *genau einen* Wert aus einer Menge von Zahlen zuzuordnen.

> *Eine Antwortvariable ordnet jeder Person hinsichtlich jeder Itemantwort genau einen Wert zu.*

Das hat zum Beispiel zur Konsequenz, daß auch Mehrfachantworten auf ein Item nur durch *eine* Kodezahl verschlüsselt werden dürfen, es sei denn man unterscheidet mehrere Signierungsaspekte (s. Kap. 2.5.1)

Die beiden folgenden Unterkapitel gehen getrennt auf den Prozeß der Signierung und der Kodierung ein.

2.5.1 Die Signierung freier Antworten

Freie Antworten können aus Bildern, Worten, Satzergänzungen, verbalen Bildinterpretationen oder ähnlichem bestehen. Eine erste Frage betrifft die Anzahl der *Signierungsaspekte,* hinsichtlich derer jede Antwort signiert oder kategorisiert werden soll. Im einfachsten Fall handelt es sich nur um einen einzelnen Signierungsaspekt, also z.B. welche Art von Aggressivität in der Itemantwort zum Ausdruck kommt. Ein Beispiel für mehrere Signierungsaspekte ist die Auswertung freier Textproduktionen nach Textlänge, Merkmalen des Satzbaus und nach Inhalten des Textes. Jeder Signierungsaspekt ergibt in der Regel *eine* Antwortvariable.

Für die weitere Auswertung ist es sinnvoll, daß die Signierungsaspekte *logisch unabhängig* voneinander sind, d.h. daß die Zuordnung einer Itemantwort zu einer Kategorie des einen Aspektes nicht zur Folge haben darf, daß bestimmte Kategorien eines anderen Signierungsaspektes auftreten müssen oder nicht auftreten können. Derartige logisch voneinander abhängige Signierungsaspekte sind schwierig auszuwerten, da die *logischen* Abhängigkeiten *zu statistischen* Abhängigkeiten führen, welche keine empirischen Gegebenheiten widerspiegeln, sondern nur die Definition der Signierungsaspekte.

Innerhalb jedes Signierungsaspektes gilt es, einen Satz von mindestens zwei Kategorien derartig klar und eindeutig zu definieren, daß jede Itemantwort *in genau eine* dieser Kategorien entfällt bzw. ihr zuordenbar ist. Bisweilen werden auch *Mehrfachsignierungen* vorgenommen, d.h. Zuordnungen der Itemantwort zu mehr als einer Kategorie desselben Signierungsaspektes. Solche mehrfach signierten Itemantworten müssen aber im nächsten Schritt der Kodierung derart verschlüsselt werden, daß tatsächlich eine Antwortvariable entsteht (s. o.).

Das *Kategorienschema,* welches man für einen Signierungsaspekt entwickelt, kann sehr unterschiedlich aussehen. Es reicht von lediglich *dichotomen* Antwortkategorien (ein bestimmtes Merkmal ist in der Itemantwort enthalten oder nicht), über *qualitativ* unterschiedliche Kategorien (Merkmal A, B oder C ist in der Antwort enthalten) bis hin zu mehrfach *gestuften* Ratingskalen, anhand derer die Itemantworten beurteilt werden. Generelle Empfehlungen, welche Art von Kategorienschema für welche Signierungsaspekte am sinnvollsten sind, lassen sich schwer geben. Ein wichtiges formales Kriterium besteht darin, daß das Kategorienschema *einfach genug* sein muß, damit eine hinreichende Signierobjektivität erreicht werden kann.

Unter *Signierobjektivität* versteht man das Ausmaß, in dem zwei voneinander unabhängig arbeitende Signierer die Itemantworten denselben Antwortkategorien zuordnen. Die Signierobjektivität muß bei jeder Testentwicklung kontrolliert, d.h. berechnet werden und gilt als Gütekriterium des Tests (vgl. Kap 2.1.3). Die Berechnung der Signierobjektivität geschieht mittels eines geeigneten Übereinstimmungskoeffizienten.

Ausgangspunkt für die Berechnung eines Übereinstimmungskoeffizienten ist eine *sog. Übereinstimmungsmatrix,* in der die Häufigkeiten stehen, mit denen zwei Signierer die Antwortkategorien zugeordnet haben.

Beispiel

Die *Übereinstimmungsmatrix*

		Signierer 2					
		A	B	C	D	E	
Signierer 1	A	10	1	2	0	1	14
	B	0	15	1	0	0	16
	C	1	1	20	2	2	26
	D	3	0	0	8	0	11
	E	0	2	0	2	13	17
		14	19	23	12	16	84

gibt an, daß von den 84 zu signierenden Itemantworten 10 übereinstimmend von beiden Signierern der Kategorie A, 15 Antworten der Kategorie B etc. zugeordnet wurden. Die Übereinstimmung ist perfekt, wenn nur die Felder der *Hauptdiagonale* in dieser Matrix besetzt sind. Im vorliegenden Fall hat z.B. Signierer 2 vier Antworten anderen Kategorien zugewiesen, die Signierer 1 der Kategorie A zugeordnet hat (nämlich eine in B, zwei in C und eine in E). Aus den Randsummen der Matrix ist ersichtlich, daß Signierer 2 die Kategorie B häufiger und Kategorie C seltener verwendet als Signierer 1.

Eine solche Übereinstimmungsmatrix kann man *itemspezifisch* aufstellen (in diesem Fall wären die 84 Kodierungen im obigen Beispiel auf 84 Personen und nur ein Item bezogen) oder für mehrere bzw. alle Items (z.B. könnte es sich um die Antworten von 21 Personen auf 4 Items handeln). Ob die Signierobjektivität itemspezifisch oder für alle Items gemeinsam berechnet werden sollte, hängt davon ab, ob man besondere Schwierigkeiten der Signierung bei einzelnen Items erwartet. In diesem Fall sollte die Objektivitätskontrolle itemspezifisch erfolgen, so daß man Items mit einer zu geringen Signierobjektivität bei einer Testrevision *modifizieren* oder *eliminieren* kann, bzw. deren Antworten bei der Testauswertung *unberücksichtigt* läßt.

Es gibt mehrere *Übereinstimmungskoeffizienten,* die man anhand einer solchen Matrix berechnen kann, von denen hier nur einer dargestellt werden soll. Es handelt sich um Cohen's κ (Kappa), der folgendermaßen definiert ist

$$(1) \qquad \kappa = \frac{p - p_e}{1 - p_e} \quad ,$$

wobei p die relativen Häufigkeiten der übereinstimmenden Kategorisierungen bezeichnet:

$$p = \frac{\sum_x f_{xx}}{N} \quad .$$

Die Häufigkeiten in den Diagonalfeldern werden mit f_{xx} bezeichnet und die Anzahl der kodierten Itemantworten mit N.

Mit p_e werden die zu erwartenden Häufigkeiten von Übereinstimmungen bezeichnet, die allein per Zufall auftreten, d.h. wenn beide Signierer würfeln würden. Diese erwarteten Häufigkeiten lassen sich

anhand der Randsummen f_{1x} und f_{2x} der Matrix berechnen

$$p_e = \frac{\sum_x f_{1x} f_{2x}}{N^2} \quad .$$

Zur Berechnung von κ benötigt man also nur die Häufigkeiten aus der Hauptdiagonale und die Randsummen der Übereinstimmungsmatrix.

Beispiel

Für die oben aufgeführte Übereinstimmungsmatrix ergibt sich für p der folgende Wert:

p = (10+15+20+8+13) / 84 = 0.7857

Diese Zahl besagt, daß die beiden Signierer 78,5 % aller Itemantworten übereinstimmend signiert haben. Unter Zufallsbedingungen würde bei den gegebenen Randverteilungen die folgende Übereinstimmung erreicht:

p_e = (14·14+16·19+26·23+11·12+17·16) / 84^2 = 0.213

Der Koeffizient κ korrigiert die beobachtete Übereinstimmung um diesen Zufallseffekt:

$$\kappa = \frac{0.785-0.213}{1-0.213} = 0.727 \quad .$$

Dieser Koeffizient κ berücksichtigt nicht, *welche* andere Kategorie ein Signierer wählt, wenn er nicht mit einem anderen Signierer übereinstimmt: Wie die Häufigkeiten in den Feldern außerhalb der Diagonale verteilt sind, geht in die Berechnung nicht ein. Dies ist dann problematisch wenn die Kategorien eine *Rangordnung* darstellen, also eine Vertauschung von B und D gravierender ist als eine Vertauschung von B und C.

Das ist relativ häufig gegeben, nämlich immer dann, wenn mittels abgestufter Kategorien das *Ausmaß* signiert wird, in dem eine freie Antwort z.B. Aggression oder Angst ausdrückt. Für diese Fälle ge*ordneter Signierungskategorien* kann ein gewichteter K-Koeffizient berechnet werden, der eine unterschiedliche Signierung in benachbarten Kategorien weniger stark gewichtet als eine Signierung in weiter auseinander liegenden Kategorien.

Um diese Gewichte in dem Übereinstimmungsmaß κ berücksichtigen zu können, wird κ zunächst so transformiert, daß es anhand der Häufigkeiten außerhalb der Hauptdiagonalen berechnet wird und nicht anhand der Diagonalfelder selbst:

$$\kappa = 1 - \frac{1-p}{1-p_e}$$

$$(2) \qquad = 1 - \frac{\frac{1}{N}\sum_{x \neq y} f_{xy}}{\frac{1}{N^2}\sum_{x \neq y} f_{1x} f_{2y}} \quad .$$

Im Zähler des zweiten Summanden steht die relative Häufigkeit der *nicht* übereinstimmenden Kategorisierungen, also die Summe aller Häufigkeiten f_{xy} aus Zeile x und Spalte y der Übereinstimmungsmatrix, wobei x und y nicht identisch sein darf, also $x \neq y$. Im Nenner stehen die unter Zufallsbedingungen zu erwartenden Nicht-Übereinstimmungen, wobei f_{1x} die Randhäufigkeit der Zeile x (also von Signierer 1) und f_{2y} die Randhäufigkeit der Spalte y (also von Signierer 2) bezeichnet.

In dieser Schreibweise von κ lassen sich jetzt leicht Gewichte einführen, um den 'Schweregrad' einer Abweichung der beiden Signierer einzubeziehen:

$$(3) \qquad \kappa_w = 1 - \frac{N \cdot \sum_{x \neq y} w_{xy} f_{xy}}{\sum_{x \neq y} w_{xy} f_{1x} f_{2y}}$$

Der Index w von κ steht für 'weighted' also gewichtetes Kappa und die Gewichte w_{xy} sind so zu wählen, daß ein größeres Gewicht eine gravierendere Abweichung bezeichnet. Bilden die Signierungskategorien eine Rangordnung und kann man weiterhin davon ausgehen, daß die Abstände zwischen den Kategorien gleich groß sind, so verwendet man als Gewicht die *quadrierten Abweichungen.* Hierfür werden die Kategorien mit aufsteigenden ganzzahligen Werten kodiert, also z.B. mit den Werten 1 bis 5. Die Gewichte lauten dann:

$$(4) \qquad w_{xy} = (x-y)^2.$$

Für das Datenbeispiel mit fünf Signierungskategorien sieht die Matrix der Gewichte wie folgt aus:

	A	B	C	D	E
A	0	1	4	9	16
B	1	0	1	4	9
C	4	1	0	1	4
D	9	4	1	0	1
E	16	9	4	1	0

Prinzipiell lassen sich auch andere Gewichte wählen, jedoch bedarf es dafür recht präziser Vorstellungen, wie ähnlich sich die Kategorien sind und wie 'zulässig' daher Verwechslungen sind. Wählt man die quadrierten Abweichungen (4) als Gewichte, so ist κ_w bei großem N identisch mit der *Intraklassen-Korrelation,* einem Übereinstimmungsmaß, das man für intervallskalierte Signierungskategorien verwendet (s. Fleiss und Cohen (1973)).

Datenbeispiel

Für die oben genannte Übereinstimmungsmatrix soll das gewichtete Kappa mit den quadrierten Abweichungen als Gewichte berechnet werden. Hierfür werden zunächst die Zellen der Übereinstimmungsmatrix mit den Zellen der Gewichtematrix multipliziert:

	A	B	C	D	E
A	0	1	8	0	16
B	0	0	1	0	0
C	4	1	0	2	8
D	27	0	0	0	0
E	0	18	0	2	0

Da die Hauptdiagonalelemente dieser Matrix durch das Gewicht 0 aus der Gewichtematrix ohnedies gleich 0 sind, ergibt die Summe aller Matrixelemente, 88, multipliziert mit N = 84 den Zählerausdruck von κ_w, $88 \cdot 84 = 7392$. Den Nenner ergibt die Summe aller Elemente einer Matrix, in deren Zellen die *erwarteten* Häufigkeiten, f_{1x} f_{2x}, multipliziert mit den Gewichten stehen:

	A	B	C	D	E
A	0	266	4322	9.168	16.224
B	224	0	368	4.192	9.256
C	4.364	494	0	312	4·416
D	9·154	4·209	2 5 3	0	176
E	16.238	9.323	4.391	204	0

Diese Summe lautet 25374, so daß sich für κ_w der folgende Wert ergibt:

$$\kappa_w = 1 - \frac{7392}{25374} = 0.709.$$

Die Übereinstimmung zwischen den beiden Signierern ist demnach unter der

Annahme geordneter Kategorien mit gleichen Abständen etwas niedriger als unter der Annahme nominalskalierter Kategorien (0.727). Dies liegt daran, daß den 7 Verwechslungen zwischen benachbarten Kategorien (s. die Übereinstimmungsmatrix) immerhin 10 Verwechslungen zwischen weiter auseinanderliegenden Kategorien gegenüberstehen.

Den Berechnungen der Signierobjektivität unter der Annahme geordneter Signierungskategorien liegt bereits eine bestimmte Zuordnung von Zahlencodes zu den Antwortkategorien zugrunde. Dieser Auswertungsschritt der 'Kodierung' von Antwortkategorien wird im folgenden Kapitel ausführlicher behandelt.

2.5.2 Die Kodierung von Antwortkategorien

Den Kategorien der Itemantworten, seien sie durch das Antwortformat vorgegeben oder seien sie das Resultat der Signierung freier Antworten, müssen Zahlen zugeordnet werden, um sie weiter verarbeiten zu können. Dieser Vorgang wird als *Kodierung* bezeichnet und hat zum Ziel, die *Antwortvariablen* herzustellen (s.o.). Die Arten der Kodierung von Antwortkategorien lassen sich daher anhand der Arten der durch sie erzeugten Antwortvariablen unterscheiden.

Die wichtigsten Unterscheidungsmerkmale sind hierbei, ob die Antwortvariable

- dichotom (zweigeteilt) oder polytom (mehrgeteilt) ist, und
- ob sie ungeordnete oder geordnete Kategorien hat.

Dichotome Antwortvariablen sind weitaus die häufigsten. Sie nehmen nur 2 Werte (Valenzen) an, nämlich 0 und 1. Diese beiden Codes haben sich durchgesetzt, weil sie rechnerisch am leichtesten handhabbar sind (anders als etwa die Codes 1 und 2).

Unterscheidet das Antwortformat (oder die Signierung) von vornherein nur 2 Kategorien, z.B. richtig - falsch, ja - nein, stimme zu - stimme nicht zu etc., so stellt sich bei der Kodierung in eine dichotome Antwortvariable nur ein Problem, nämlich das der *Polung*. Für die meisten Arten der Testauswertung, insbesondere für die Messung quantitativer Personenmerkmale, ist es nämlich wichtig, daß die Antwortvariablen für alle Items *gleichsinnig gepolt* sind, d.h. der Code '1' immer auf denselben Pol des zu messenden Merkmals hinweist (z.B. Extraversion). Das bedeutet, daß eine ja-Antwort durchaus nicht immer mit einer '1' zu kodieren ist, nämlich dann nicht, wenn sie auf den entgegengesetzten Pol der zu messenden Eigenschaft hinweist (z.B. Introversion).

Ob eine derartige Umpolung negativ formulierter Items bei der Kodierung erfolgen sollte, hängt von dem anzuwendenden Testmodell ab. So sollten die Antworten bei einem quantitativen Testmodell mit nicht-monotonen Itemfunktionen (s. Kap. 3.1.1.3) *nicht* umgepolt werden, bei klassifizierenden Testmodellen dient eine Umpolung lediglich der Übersichtlichkeit der Ergebnisse.

Gibt es mehr als zwei Antwortkategorien, so kann eine *Dichotomisierung*, also eine Kodierung in eine dichotome Antwortvariable sinnvoll sein, was aber stets mit einem *Informationsverlust* verbunden ist. Werden in einem Leistungstest etwa 5

Alternativen vorgegeben, so verzichtet man mit der Kodierung der richtigen Alternative mit ‘1’ und aller anderen mit ‘0' auf die Information, *welcher Distraktor* gewählt wurde. Die Wahl eines schwierigen Distraktors ist zwar auch eine ‘falsche’ Itemantwort, sie weist aber darauf hin, daß sich die Person bei der Beantwortung ‘etwas gedacht’ und nicht nur geraten hat. Die Alternative zu einer Dichotomisierung wäre in diesem Fall die Kodierung mit

0 = Wahl eines unplausiblen Distraktors
1 = Wahl eines plausiblen Distraktors
2 = Wahl der richtigen Alternative,

also die Herstellung einer polytomen Antwortvariable.

In Fragebögen mit geordneten Antwortkategorien, z.B. Ratingformaten, empfiehlt sich generell *keine* Dichotomisierung. Fast alle in Kapitel 3 behandelten Testmodelle gibt es auch in einer Version für polytome Antwortvariablen mit geordneten Kategorien. Solche Testmodelle für ordinale Daten bieten nicht nur genauere Meßwerte für die Personeneigenschaft, sondern auch bessere Möglichkeiten der Prüfung, ob ein Testmodell auf die Daten paßt.

Entscheidet man sich für die Dichotomisierung mehrerer Antwortkategorien, stellt sich die Frage *wie* man dichotomisiert.

Bei einem Leistungstest mit *mehreren* richtigen Antwortkategorien kann man unterschiedlich streng dichotomisieren, indem man entweder nur die Kombination der richtigen Alternativen mit ‘1’ kodiert oder auch Kombinationen, in denen die richtigen Alternativen und ein Distraktor enthalten sind. Hierzu kann es keine generellen Empfehlungen geben, außer der, daß es aus statistischen Gründen vorteilhaft ist, wenn beide Codes etwa gleich häufig auftreten.

Bei Ratingformaten kann sich die Notwendigkeit einer Zusammenlegung von Kategorien, und im Extremfall einer Dichotomisierung stellen, wenn einige Antwortkategorien *zu selten* gewählt wurden. Bei sehr vielen Testmodellen gibt es nämlich Probleme mit der Parameterschätzung, wenn einzelne Antwortkategorien bei einem Item gar nicht oder nur 2- oder 3-mal auftreten.

Bei *polytomen Antwortvariablen* sind diejenigen mit *geordneten Kategorien* weitaus häufiger als solche mit ungeordneten. Werden Ratingformate verwendet, so nimmt man im allgemeinen an, daß die Kategorien der Ratingskala geordnet sind. Ihre Kodierung erfolgt mit aufsteigenden ganzzahligen Werten, wobei ebenfalls mit 0 begonnen wird. Die Antwortvariablen X_{vi} nehmen also Werte von 0 bis m an,

$$x_{vi} \in \{0,1,2,\ldots,m\} \quad ,$$

wenn es m+l Ratingkategorien gibt.

Die Zuordnung *aufeinanderfolgender ganzzahliger Werte* zu den Stufen einer Ratingskala (sog. *integer scoring)* wird oft als willkürlich empfunden und es wird argumentiert, man könnte den Stufen mit gleichem Recht auch die Werte 1, 3, 9, 10 und 27 zuordnen. Mit dieser Kritik ist gemeint, daß die Zuordnung gleichabständiger *(äquidistanter) Codes* ein Skalenniveau für die Itemantworten voraussetzt, (nämlich des Niveau einer Intervallskala), das den Daten gar nicht zukommt. Würde diese Kritik zutreffen, so wäre das tatsächlich ein gravierender Nachteil polytomer Antwortvariablen, denn eine *Äquidistanzannahme*, die bereits in die Kodie-

rung der Itemantworten eingeht, kann nicht nachträglich über die Gültigkeit eines Testmodells geprüft werden. Der Testauswertung würde in diesem Fall ein willkürliches Element anhaften, das ihre 'Wissenschaftlichkeit' in Frage stellt.

Diese Kritik trifft jedoch nur dann zu, wenn man die Antwortvariablen selbst zu Meßwerten erklärt. Das ist etwa dann der Fall, wenn man die Summe der Itemantworten (bzw. deren Codes) als Meßwert für die Personeneigenschaft nimmt. Berechnet man jedoch die Meßwerte mit Hilfe eines Testmodells für polytome Daten (s. Kap. 3.3), so stellen die Codes der Antwortkategorien *keine* Werte auf einer Intervallskala dar, sondern sie bezeichnen die Anzahl der *Schwellenüberschreitungen,* die einer Itemantwort zugrundeliegen.

Damit ist gemeint, daß zwischen den Kategorien einer (m+l)-stufigen Ratingskala genau m Übergänge, sog. *Schwellen* liegen. Ein Kreuz auf einer m-stufigen Ratingskala zu machen, setzt bei der befragten Person m-mal die Entscheidung voraus, eine Schwelle zu überschreiten oder nicht. Ein Kreuz in Kategorie x gibt an, daß die Person x-mal eine Schwelle überschritten hat. Der Code x ist also eine *Häufigkeitsangabe* und kein intervallskalierter Meßwert.

Diese Art der Kodierung mit Werten von 0 bis m setzt lediglich voraus, daß die Antwortkategorien tatsächlich *geordnet* sind, so daß die Überschreitung einer höheren Schwelle nur möglich ist, wenn alle niedrigeren Schwellen überschritten wurden. Andernfalls würde der Code x für die (x+l)-te Stufe der Ratingskala nicht mehr die Anzahl der Schwellenüberschreitungen kennzeichnen. Ob diese *Schwellen* dann für ein bestimmtes Item oder ein bestimmtes Antwortformat *äquidistant* sind, ist eine ganz andere Frage, die mittels geeigneter Testmodelle beantwortet werden kann (s. Kap. 3.3.2 und 3.3.4).

Auch für polytome Antwortvariablen mit geordneten Kategorien stellt sich die Frage einer gleichsinnigen *Polung* aller Items, die dasselbe Merkmal messen. Für die wichtigsten Testmodelle ist eine solche gleichsinnige Polung Voraussetzung. Ausnahmen bilden Testmodelle mit nicht-monotonen Itemfunktionen (s. Kap. 3.1.1.3), die sich auch für polytome Daten verallgemeinem lassen und klassifizierende Testmodelle mit itemspezifischen Schwellendistanzen (s. Kap. 3.3.4).

Das Problem bei einer *Umpolung* negativ formulierter Items mit geordneten Antwortkategorien besteht darin, daß mit der Umpolung auch die Reihenfolge der Schwellen umgekehrt wird.

Beispiel

Ein Angst-Fragebogen enthält die beiden folgenden Items:

- *Vor einer Prüfung kann ich meistens nichts essen.*
- *Wenn ich zum Zahnarzt gehe, lese ich im Wartezimmer in aller Ruhe die Illustrierten.*

Das Antwortformat lautet:

- *trifft nicht zu*
- *trifft selten zu*
- *trifft oft zu*
- *trifft immer zu.*

Da die beiden Items offensichtlich in unterschiedlicher Richtung formuliert sind erfordert eine gleichsinnige Polung der

Antwortvariablen die Kodierung von 0 bis 3 beim ersten, und von 3 bis 0 beim zweiten Item. Ein Wert von x=l bedeutet daher beim ersten Item, daß die Schwelle von 'trifft nicht zu' nach 'trifft selten zu' überschritten wurde. Beim zweiten Item bedeutet derselbe Wert, daß die Schwelle von 'trifft immer zu' nach 'trifft oft zu' überschritten wurde.

Diese Umkehrung der Schwellenreihenfolge infolge der Umpolung negativ formulierter Items ist für solche Testmodelle problematisch, bei denen die Schwellenabstände als *für alle Items* konstant angenommen werden (s. Kap. 3.3.4). Diese Modelle können in solchen Fällen nicht angewendet werden.

Die Kodierung *ungeordneter Kategorien* in polytome Antwortvariablen wirft ganz andere Fragen auf. Soll auf die Daten ein *klassifizierendes Testmodell* angewendet werden, d.h. soll eine qualitative Personenvariable erfaßt werden, so ist die Kodierung ungeordneter Kategorien völlig unproblematisch: den Kategorien jedes Items werden in beliebiger Reihenfolge die Werte 0 bis m zugeordnet, wobei es nicht nur egal ist, *wie* die Antwortkategorien von jedem Item *definiert* sind, sondern sogar *wieviele* Antwortkategorien bei jedem Item unterschieden werden.

Beispiel

Mit einem Fragebogen sollen die leistungsbezogenen Kognitionen von Schülern erfaßt werden, wobei von einer Typologie von Schülern ausgegangen wird, die 3 Muster von leistungsbezogenen Kognitionen unterscheidet. In dem Fragebogen kommen ganz unterschiedliche Items vor:

Erstens, Items die die *Attribution von Mißerfolg* (schlechte Noten) erfassen sollen und vier Antwortkategorien unterscheiden: intern-labile, intern-stabile, extern-labile und extern-stabile Attribution.

Zweitens, Items die die *Leistungsmotivation* erfassen sollen und zwei Antwortkategorien unterscheiden: Hoffnung auf Erfolg und Furcht vor Mißerfolg.

Drittens, Items die die *subjektiven Kontrollüberzeugungen* der Schüler erfassen sollen und drei Antwortkategorien unterscheiden: Kontrolle liegt beim Schüler, Kontrolle liegt bei anderen Personen, Kontrolle liegt beim Zufall.

Die drei erwarteten Typen von Schülern zeichnen sich durch folgende Antwortmuster aus:

Typ 1: intern-labile Attribution von Mißerfolg, Hoffnung auf Erfolg, Kontrolle beim Schüler

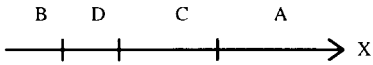
Typ 2: extern-labile Attribution von Mißerfolg, Furcht vor Mißerfolg, Kontrolle beim Zufall

Typ 3: extern-stabile Attribution von Mißerfolg, Furcht vor Mißerfolg, Kontrolle bei anderen Personen

Die Items können mit den Werten 0-1-2-3, 0-1 und 0-1-2 kodiert werden und mit einem klassifizierenden Testmodell ausgewertet werden, um die Schülertypen zu erfassen.

Ganz anders sind die Erfordernisse an die Kodierung, wenn mit ungeordneten Kategorien *quantitative Personenmerkmale* erfaßt werden sollen. Hierzu muß man sich zunächst klarmachen, daß man *nicht* mit mehreren ungeordneten Antwortkategorien nur *eine* quantitative Eigenschaft

erfassen kann. Hat ein Item die Kategorien A, B, C und D und soll mit allen 4 Kategorien nur die eine Eigenschaft X erfaßt werden, so geht das nur, wenn die Antwortwahrscheinlichkeiten der 4 Kategorien auch tatsächlich von der Eigenschaft X abhängen. Das bedeutet, daß jede der Kategorien A, B, C und D einen Abschnitt auf der zu messenden Dimension haben muß, in dem diese Kategorie auch mit relativ hoher Wahrscheinlichkeit gewählt wird



In diesem Fall handelt es sich aber bereits um *geordnete* Kategorien und diese müssen gemäß ihrer Ordnung kodiert werden, d.h. B=0, D=1, C=2 und A=3.

Ungeordnet sind Kategorien nur dann, wenn jede Kategorie eine *andere* Eigenschaft anspricht, z.B.

A → die Tendenz intern-labil zu attribuieren
B → die Tendenz intern-stabil zu attribuieren
C → die Tendenz extern-labil zu attribuieren
D → die Tendenz extern-stabil zu attribuieren

Aus diesem Sachverhalt leiten sich auch die Konsequenzen für die Kodierung ungeordneter Antwortkategorien ab: die Kategorien werden wiederum mit den Werten von 0 bis m kodiert, jedoch muß jeder Code bei allen Items *dieselbe Antwortkategorie* bezeichnen, also z.B. die Codezahl '2' bezeichnet diejenige Antwort, die auf eine extern-labile Attribution hinweist.

Zusammenfassend läßt sich sagen, daß die Kodierung bei der Messung qualitativer Personeneigenschaften *itemspezifisch* erfolgen darf, während sie bei der Messung quantitativer Eigenschaften *itemübergreifend* erfolgen muß.

Eine letzte Anmerkung noch zur Kodierung *'fehlender Itemantworten'*, also zu dem Fall, daß Personen einzelne Items ausgelassen oder übersprungen haben. Es hat sich eingebürgert, diese *sog. missing data* mit der Codezahl '9' zu kodieren, bzw. mit '99' wenn mehr als 9 Antwortalternativen zu kodieren sind. Man sollte eine solche getrennte Kodierung fehlender Antworten in jedem Fall vornehmen, auch wenn es bei der Anwendung eines Testmodells oft erforderlich ist, diesen Wert zu *recodieren,* d.h. mit einer zulässigen Itemantwort zusammenzulegen (z.B. 9→0). Nicht nur die verfügbare *Software* unterscheidet sich hinsichtlich ihrer missing-data Optionen, also dem Angebot mit fehlenden Werten umzugehen. Auch hängt es von den jeweiligen *Testmodellen* ab, wie sinnvoll überhaupt mit fehlenden Werten umgegangen werden kann.

In den folgenden Kapiteln wird diese Problematik nicht weiter erörtert. Es wird vielmehr davon ausgegangen, daß die Anzahl fehlender Werte im allgemeinen so gering ist, daß eine Zusammenlegung mit einer zulässigen Kategorie (z.B. '0' bei Leistungsitems oder einer mittleren Kategorie bei Ratingskalen) zu keiner gravierenden Veränderung der Ergebnisse führt.

Literatur

Das Übereinstimmungsmaß Kappa wurde von Cohen (1960) für Nominaldaten und Cohen (1968) für Ordinal- oder Intervalldaten (weighted Kappa) entwickelt. Fleiss (1971) und Light (1971) diskutieren die Verallgemeinerung dieses Maßes für mehr

als 2 Signierer und Fleiss et. al (1969) geben an, wie man den Standardfehler von Kappa berechnen kann. Neuere Methoden zur Berechnung von Signier- oder Rater-Übereinstimmung bedienen sich der latent class Analyse (Dillon & Mulani 1984). Asendorpf & Wallbott (1979) sowie Zegers (1991) geben einen Überblick über verschiedene Koeffizienten. Matschinger und Angermeyer (1992) diskutieren Effekte der Itempolung auf das Antwortverhalten.

Übungsaufgaben

1. In einem Satzergängzungstest wurden zu dem Satzanfang : 'Wenn mich auf dem Gehweg jemand anrempelt und sich nicht mal entschuldigt, dann ...' folgende Ergänzungen produziert:

   ... ist das unverschämt
   ... rufe ich ihm/ihr etwas hinterher
   ... gehe ich einfach weiter
   . . kann das mal passieren
   . . ärgere ich mich
   ... sage ich 'hoppla'
   . . . bleibe ich stehen und wundere mich.

   Schlagen Sie mehrere Signierungsaspekte vor und signieren Sie die Antworten danach.

2. Zwei Signierer erhalten bei der Signierung nach 4 Kategorien die folgende Übereinstimmungsmatrix:

	A	B	C	D
A	5	1	0	1
B	2	7	1	0
C	3	0	4	0
D	0	4	1	8

   Berechnen Sie den Übereinstimmungkoeffizienten κ.

3. In einem Leistungstest zum Physikwissen ist als ein Item eine Batterie, zwei Lämpchen und ein Ein-/Aus-Schalter abgebildet. Die Aufgabe besteht darin, eine Kabelverbindung zwischen diesen 4 Teilen so einzuzeichnen, daß bei einer Realisierung dieser Schaltung beide Lämpchen möglichst hell leuchten. Schlagen Sie eine Signieranleitung und eine Kodierung vor, mit der eine polytome Antwortvariable entsteht.