ps psychologie

Markus Bühner

Einführung in die Test- und Fragebogenkonstruktion

3., aktualisierte Auflage

PEARSON Studium

Erstellung eines Testentwurfs

3

Wie gehe ich bei der Erstellung des Testentwurfs vor?

Zur Testkonstruktion sind eine Reihe sorgfältig geplanter Schritte nötig. Wie man dabei vorgehen kann, wird im folgenden Kapitel beschrieben. Auf einer übergeordneten Ebene lässt sich der Prozess der Testkonstruktion in drei grobe Teilabschnitte untergliedern: (1) **Erstellung des Testentwurfs**, (2) **empirische Überprüfung des Testentwurfs** (siehe *Kapitel 5.1* und *5.2*), (3) **Normierung/Cut-Off-Ermittlung** (siehe *Kapitel 5.4*) der endgültigen Testversion. Der erste Schritt ist dabei besonders wichtig, denn Fehler, die dem Testkonstrukteur in dieser Phase unterlaufen, lassen sich später bei der empirischen Überprüfung nicht mehr korrigieren. In den folgenden Abschnitten dieses Kapitels wird nun dieser **erste Schritt „Erstellung eines Testentwurfs"** näher erläutert, der sich in weitere Einzelstufen, wie in *Abbildung 3.1* dargestellt, aufgliedern lässt.

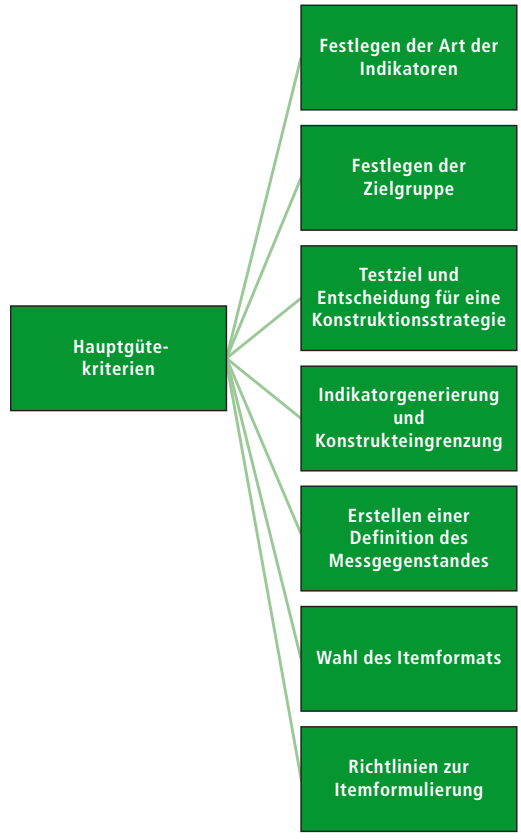


Abbildung 3.1: Schritte der Testkonstruktion.

Zunächst wird vom Testautor festgelegt, ob der Test **objektive** oder **subjektive Indikatoren** einer Eigenschaft oder Fähigkeit erfassen soll. Im Anschluss daran wird entschieden, für welche **Zielgruppe** der Test entwickelt werden soll. Von den ersten beiden Teilschritten hängen wesentliche Entscheidungen für die Itemgenerierung ab. Danach entscheidet sich der Testkonstrukteur für eine Konstruktionsstrategie. Hier stellen sich grob drei Fragen, (1) ob der Test **Gruppen trennen**, (2) zur **Erfassung der**

Ausprägung einer Eigenschaft oder Fähigkeit dienen oder (3) **Wissen erfassen** soll. Dann werden gemäß der Konstruktionsstrategie **Indikatoren** des Konstrukts **generiert**. Indikatoren können beispielsweise Verhaltensweisen, Einstellungen oder Eigenschaften sein. Sind die Indikatoren des Konstrukts definiert, kann eine **Definition des Messgegenstandes** erfolgen. Aus den Indikatoren des Konstrukts werden dann Fragen formuliert und zu einem Testentwurf mit Instruktion und Auswertungsregeln zusammengefasst. Dieser Testentwurf wird dann einer **empirischen Prüfung** unterzogen. Das heißt eine relevante Analysestichprobe bearbeitet den Test und die Analyse der Itemantworten wird zu einer Optimierung des Testentwurfs genutzt. Der Test wird dann schließlich **normiert** und/oder ein **Cut-Off-Wert** ermittelt, ab dem eine Person eine bestimmte Diagnose erhält oder der Person eine bestimmte Leistung bescheinigt wird. Die Testkonstruktion als erste Stufe wird in diesem Kapitel dargestellt und die letzten beiden Stufen in *Kapitel 5*.

Einstiegshilfe Für das Lesen dieses Kapitels ist ein Basisverständnis von Skalenniveaus nützlich (vgl. Bühner & Ziegler, 2009, Kapitel 2.1). Als weiterführende Lektüre zur Erstellung und Formulierung von Fragen eignet sich das Buch „Was ist eine gute Frage?“ von Faulbaum, Prüfer und Rexroth (2009).

3.1 Festlegung der Art der Indikatoren

Zunächst muss die Frage beantwortet werden, ob der Test anhand **objektiver** oder **subjektiver Indikatoren** ein Konstrukt erfassen soll. Dies wird im Folgenden näher erläutert.

Cattell (1965) differenzierte Daten anhand der Quellen, die ein diagnostisches Verfahren zur Informationssammlung nutzt. So bezeichnete er Daten aus **Tests**, bei denen die Antwort eindeutig als richtig oder falsch klassifiziert werden kann, als **T-Daten** (Test data). Daten aus **Fragebögen**, die Selbst- oder Fremdratings erfassen, bezeichnete er als **Q-Daten** (Questionnaire data). Daten aus **Verhaltensbeobachtungen** nannte er **L-Daten** (Life data bzw. L-Daten). In dieser Einteilung ist auch eine Differenzierung zwischen objektiven und subjektiven Daten oder besser objektiven oder subjektiven **Indikatoren** enthalten. Ein Indikator für ein Konstrukt ist ein Item, das beispielsweise Verhalten, Einstellungen oder Eigenschaften messen kann. Meist kann vereinfachend angenommen werden, dass Leistungstests (T-Daten) aus objektiven und Persönlichkeitsfragebögen (Q-Daten) aus subjektiven Indikatoren bestehen. Es gibt jedoch auch Versuche, Persönlichkeitseigenschaften mithilfe von **objektiven Persönlichkeitstests** zu erfassen. Bei diesen objektiven Tests können wiederum verschiedene Herangehensweisen unterschieden werden. Diese sollen am Beispiel der Leistungsmotivation kurz im folgenden Exkurs beschrieben werden.

Exkurs 3.1 Testarten am Beispiel Leistungsmotivation

Das historisch gesehen älteste Verfahren ist hier sicher der Thematische Apperzeptionstest (TAT). Mittlerweile existieren jedoch auch Verfahren, die andere Techniken nutzen. So werden auch implizite Assoziationstests (IATs) oder einfache motorische Aufgaben (OLMT) verwendet. Allerdings muss hierbei gesagt werden, dass empirische Untersuchungen zeigen, dass die verschiedenen Herangehensweisen nicht zwingend dasselbe Konstrukt erfassen (Ziegler, Schmukle, Egloff & Bühner, 2010). Am häufigsten werden sicherlich subjektive Fragebögen zur Persönlichkeitsmessung eingesetzt. In solchen Fragebögen wird der Proband gebeten, Stellung zu verschiedenen Aussagen zu nehmen, um sich selbst einzuschätzen. Im Leistungsmotivationsinventar von Schuler und Prochaska (2001) wird vom Probanden beispielsweise verlangt, auf einer siebenstufigen Skala, die von „trifft gar nicht zu“ bis zu „trifft vollständig zu“ reicht, anzugeben, inwieweit sie oder er bestimmte Aussagen als für sich zutreffend ansieht. Solche Aussagen sind z.B.: „Mein Ehrgeiz ist leicht herauszufordern.“ Bevor nun der Test entwickelt wird, muss sich der Testkonstrukteur also entscheiden, ob er einen objektiven Test oder einen subjektiven Fragebogen entwickeln möchte. Die Tragweite dieser Entscheidung wird dadurch deutlich, dass der Zusammenhang zwischen objektiven und subjektiven Tests, die eigentlich dasselbe Konstrukt erfassen sollen, häufig gering ausfällt. Mögliche Erklärungen hierfür können in der Literatur nachgelesen werden (Gschwendner, Hofmann & Schmitt, 2006; Thrash & Elliot, 2002). Die Entscheidung für eine der beiden Testarten kann auch von der Art des vorherzusagenden Kriteriums abhängen. So zeigt eine Studie von Brunstein und Schmitt (2004) beispielsweise, dass ein objektiver Leistungsmotivationstest (IAT) andere Kriterien vorhersagt als ein subjektiver Fragebogen.

Beispiel 3.1 Konstruktion eines Konzentrationstests

Im Rahmen des Kapitels soll exemplarisch die Entwicklung eines Konzentrationstests beschrieben werden. Im ersten Schritt entscheiden wir uns dafür, einen Test zu konstruieren, der mithilfe objektiver Indikatoren Konzentration erfassen soll (T-data). Das heißt es müssen Aufgaben gesucht werden, für die es eine eindeutig richtige oder falsche Lösung gibt.

3.2 Festlegen der Zielgruppe

Es ist im nächsten Teilschritt wichtig festzulegen, welche Zielgruppe den zu entwickelnden Test später einmal bearbeitet sowie deren Merkmale zu beschreiben, die für die Testbearbeitung relevant sein könnten. Hierbei sind verschiedene Merkmale der Zielgruppe zu berücksichtigen, die sich auf die Testerstellung auswirken können:

- Alter und Bildung
- Erlebens- und Verhaltensspektrum
- Sprachbeherrschung der Zielgruppe
- Item-/Testfairness

Die hier genannten Einflussgrößen sind nicht isoliert zu sehen, sondern können sich gegenseitig bedingen oder überschneiden. Neben den hier genannten Einflussgrößen sind bei der Testbearbeitung und -entwicklung von speziellen Gruppen weitere Aspekte zu berücksichtigen. Betrachten wir beispielsweise neurologische Patienten, müssen Tests den neurologischen Status der Patienten berücksichtigen, beispielsweise Gesichtsfeldeinschränkungen, Halbseitenlähmungen oder Ähnliches. Im Weiteren soll zunächst nur auf die oben aufgeführten Einflussgrößen eingegangen werden.

Alter und Bildung Alter und Bildung der Zielgruppe haben Auswirkungen auf formale Aufgabenmerkmale, die Itemschwierigkeit sowie die Itemformulierung. Dies wird im Folgenden näher läutert.

Definition: Itemschwierigkeit
Unter hoher Itemschwierigkeit versteht man, wenn dem Item von vielen Personen zugestimmt wird (bei Fragebögen) bzw. es viele Personen richtig lösen (bei Leistungstests). Eine hohe Itemschwierigkeit drückt sich also in einem hohen Prozentsatz an Richtiglösungen oder einem hohen Itemmittelwert aus. Diese Definition von Itemschwierigkeit entspricht nicht dem, was wir im Alltagsverständnis unter hoher Schwierigkeit verstehen (wenige Personen lösen ein Item). Es wird deshalb in Abgrenzung dazu auch von **psychometrischer Itemschwierigkeit** gesprochen (vgl. *Kapitel 5.1*).

Die Bedeutung des Alters für die Zielgruppe kann am Beispiel eines Konzentrationstests, dem Test d2 (Brickenkamp, 2002), veranschaulicht werden. In diesem Test muss eine Person unter Zeitdruck in 14 Zeilen den Buchstaben d, wenn er mit zwei Strichen versehen ist, durchstreichen. Allerdings befinden sich im Test auch die Buchstaben d und p mit anderen Strichkombinationen, die nicht durchgestrichen werden sollen. Da die Zeichen sehr klein sind, sind diese gerade für ältere Personen in der Regel nicht gut zu erkennen. Wird der Test in seiner ursprünglichen Form Personen ab 60 Jahren vorgegeben, kann sich die Anzahl von ungültigen Testbearbeitungen häufen. Daher wurde eine Version mit einer größeren Schrift entwickelt, wenn die Zielgruppe des Tests in einem höheren Altersbereich liegt. Bühner und Schmidt-Atzert (2004) konnten zeigen, dass die Anzahl der Fehler in einer Stichprobe älterer Probanden bei der Bearbeitung des Tests d2 abhängig von der Größe des Testmaterials ist. Hier ist also auf das **Format** der Aufgaben zu achten, das heißt auf Aspekte wie Größe. Durch die Vergrößerung des Testblatts von DIN A4 auf DIN A3 sank die Anzahl der Fehler deutlich.

Sowohl Alter als auch Bildung wirken sich auf die Konstruktion der Items insofern aus, als dass Items mit unterschiedlicher **Itemschwierigkeit** angemessen sein können. Betrachten wird dazu Intelligenztestitems. Für Kinder im Alter von sechs Jahren müssen Intelligenztestaufgaben mit einem geringeren Schwierigkeitsgrad konstruiert werden als für 16-jährige Gymnasiasten. Für 16-jährige Hauptschüler wiederum sind möglicherweise andere Aufgaben informativ als für 16-jährige Gymnasiasten. In manchen Fällen versucht der Testautor, das Problem der zielgruppenspezifischen Itemkonstruktion zu vermeiden, indem Tests konstruiert werden, die ein sehr breites Spektrum an unterschiedlich schweren Items beinhalten. Dies ist eine Möglichkeit, mit diesem Problem umzugehen. Dabei können jedoch unerwünschte Nebeneffekte auftreten, beispielsweise werden schwächere Personen frustriert, weil sie viele Items des Tests nicht lösen können, und leistungsstarke Personen werden durch die hohe Anzahl einfacher Aufgaben gelangweilt. Im klinischen Bereich ist darauf zu achten, dass die Personen nicht unnötig durch die Bearbeitung von nicht informativen Items zusätzlich belastet werden. Das heißt in diesem Schritt werden erste Weichen für die Nebengütekriterien Ökonomie und Zumutbarkeit gestellt.

Auch die Instruktion (sprachliche Gestaltung, Fachausdrücke, Länge, Anzahl von Erklärungen, Beispiele) des Tests sollte so weit wie möglich alters- und bildungsgerecht sein. Offensichtlich ist das Alter auch für die **Itemformulierung** relevant. Je nach Altersgruppe, in der ein Test oder Fragebogen eingesetzt wird, sind bestimmte Worte mehr oder weniger geläufig. So verstehen Jugendliche unter dem Begriff *fett* unter Umständen etwas anderes als Senioren.

Erlebens- und Verhaltensspektrum Weiterhin muss die Itemkonstruktion an das Erlebens- und Verhaltensspektrum der Zielgruppe angepasst werden. Soll beispielsweise ein Schulleistungstest entwickelt werden, der deutschlandweit und schulübergreifend einsetzbar ist, müssen die unterschiedlichen Lehrpläne der Länder und Schulen berücksichtigt werden. So kann es sein, dass das Thema Stochastik in manchen Ländern in anderen Schuljahren behandelt wird und in manchen Schularten vielleicht sogar überhaupt nicht. Damit haben die Schüler in diesen Ländern oder Schulen keine Möglichkeit, Stochastik zu erlernen bzw. zu erleben. Somit ist es von vornherein sehr unwahrscheinlich, dass diese Schüler Aufgaben zur Stochastik lösen können: Das von ihnen im weitesten Sinne erwartete Verhalten, das Wissen über Stochastik, ist in diesen Teilpopulationen nicht präsent. Offensichtlich müssen die Inhalte also dem jeweiligen Lehrplan angepasst werden. Es können nur zwischen den Ländern überschneidende Inhalte abgeprüft werden. Würde hierauf nicht geachtet, würden sich in Ländern, in denen Stochastik gelehrt wird, die **Itemschwierigkeiten** gegenüber Ländern, in den Stochastik im Unterricht nicht behandelt wird, wahrscheinlich unterscheiden. Dies liegt daran, dass die Gruppe der Schüler ohne Stochastikunterricht im Durchschnitt weniger Aufgaben richtig lösen würde als Schüler mit Stochastikunterricht.

Nehmen wir als weiteres Beispiel an, wir sollen einen Fragenbogen entwickeln, der Extraversion misst. Die Zielgruppe sind Erwachsene, die sich auf eine Stelle als Animateur in einer großen Hotelkette bewerben. In der Gruppe der Bewerber wird sicherlich im Vergleich zur Gesamtbevölkerung eher eine überdurchschnittliche Merkmalsausprägung für Extraversion vorliegen. Verwenden wir nun für den Fragebogen ein Item wie „Ich mag es, in der Disco auf der Box zu tanzen“ und eine fünfstufige Antwortskala (*starke Ablehnung* bis *starke Zustimmung*), so werden Personen aus der Zielgruppe mit ihren Antworten diesem Item wohl eher zustimmen. Überlegen Sie

sich, wie leicht es Ihnen fallen würde, dieser Aussage zuzustimmen. Würde man dieser Personengruppe das Item „Ich fühle mich unter Leuten wohl“ vorgeben, würde diesem Item wahrscheinlich nahezu jeder Bewerber zustimmen, da er sich sonst von vornherein nicht auf eine solche Stelle bewerben würde. Das heißt werden hier Items falsch formuliert oder gewählt, unterscheiden sich Personen der angestrebten Zielgruppe im Antwortverhalten möglicherweise nicht oder nur geringfügig, und das Item enthält keine Information über das Verhalten der Zielgruppe. Somit kann sich das Erlebens- und Verhaltensspektrum der Zielgruppe in Abhängigkeit der **Itemformulierung und -konstruktion** unmittelbar auf die **Itemschwierigkeit** auswirken. Dies hat wiederum Auswirkungen auf die psychometrischen Eigenschaften der Items (siehe *Kapitel 5.2*).

Sprachbeherrschung der Zielgruppe Auch die sprachliche Fähigkeit der Zielgruppe ist bei der Testkonstruktion zu bedenken. Diese wirkt sich natürlich vor allem auf die Formulierung der Items, der Antwortmöglichkeiten und der Instruktion aus. Wird eine Frage nicht verstanden, kann dies zu einem zufälligen Antwortverhalten oder einem Antwortverhalten führen, dass sogar in die entgegengesetzte Richtung als die eigentlich beabsichtigte Antwort führt. Solche sprachlichen Effekte sind nicht immer durch eine Analyse der Aufgabenschwierigkeit zu finden, hier eignen sich vor allem die **kognitiven Interviewtechniken** zum Auffinden ungeeigneter Items.

Exkurs 3.2 Kognitive Interviewtechniken

Die gründliche Itemauswahl und präzise Formulierung erlauben noch keinen Rückschluss auf die Praxistauglichkeit eines Fragebogens. Aus diesem Grunde ist es sinnvoll, Vortests durchzuführen. Es ist unbedingt darauf zu achten, dass der Vortest mit **Personen aus der Zielgruppe** durchgeführt wird, für die der Test konstruiert wurde. Dieser Vortest beinhaltet die Vorgabe eines Fragebogens unter möglichst realistischen Bedingungen (z.B. Schüler, Patienten, Manager), um technische Probleme (z.B. unausgefüllte Seiten bei zweiseitigem Druck), Verständlichkeit (z.B. Fremdwörter, widersprüchliche Angaben, ungünstiger Satzbau) und Akzeptanz (z.B. Rücklaufquoten) oder Antworttendenzen (z.B. Neigung zu Extremen oder zur Mitte) bzw. die Eignung verschiedener Antwortformate (z.B. dichotom, mehrstufig) zu sondieren. Es wird häufig vernachlässigt, welche Rolle kognitive Prozesse bei der Beantwortung von Fragebogenitems spielen. Die Methode des lauten Denkens und des Nachfragens als wichtige Interviewtechniken sollen über diese für uns sonst nicht sichtbaren Prozesse näheren Aufschluss geben (vgl. Fowler, 1995). Zwei dieser Techniken werden im Folgenden kurz beschrieben. Eine sehr gute und sehr praxisnahe Zusammenfassung verschiedener kognitiver Interviewtechniken geben Prüfer und Rexroth (2005) in einem frei zugänglichen Dokument. Es findet sich auf der folgenden Webseite:
http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/How_to15PP_MR.pdf

Zunächst wird die „Think-aloud-Technik“ (Technik des lauten Denkens) kurz vorgestellt, bei der der Proband die Fragen des Fragebogens laut liest, seine Überlegungen äußert und schließlich die Antworten laut formuliert, um dann den Fragebogen entsprechend zu markieren (Concurrent-think-aloud). Dieser Prozess wird durch den Interviewer auf Tonträger aufgezeichnet; visuelle Eindrücke werden zusätzlich schriftlich dokumentiert.

Darüber hinaus können **Nachfragetechniken** (Probing) eingesetzt werden, um das Verständnis der Frage, Teile der Frage oder einzelner Wörter zu prüfen. Dabei kann auch nachgefragt werden, warum eine Person eine bestimmte Antwortkategorie gewählt hat.

Mit diesen Techniken wird getestet, ob der Proband die Fragen richtig versteht und mit den Antwortkategorien zurechtkommt oder Probleme im Umgang mit ihnen hat. Es kann beispielsweise sein, dass das Antwortformat zu differenziert ist und der Betroffene eigentlich nur Ja-Nein-Antworten geben kann anstatt vier Abstufungen. Sie bewähren sich insbesondere, wenn man einen Fragebogen für Patienten entwickeln möchte. Darüber hinaus treten weitere wichtige Erkenntnisse zutage, z.B. die Tendenz des Patienten, eigene Defizite zu entschuldigen, oder auch persönliche Einstellungen, durch die das Antwortverhalten beeinflusst wird.

Unabhängig davon, nach welcher Methode man Items konstruiert, ist ein Vortest nach einer kognitiven Interviewtechnik sehr sinnvoll. Eine schlechte Itemkonstruktion oder ein falsch gewähltes Antwortformat ist nicht durch komplexe statistische Auswertungen auszugleichen. Es ist sogar zu empfehlen, die **kognitiven Interviews zweimal durchzuführen**, einmal mit dem Test in der ersten Rohfassung und danach mit dem verbesserten Testverfahren.

Item- und Testfairness. Inwieweit einzelne Gruppen, die sich in den genannten oder auch weiteren Merkmalen unterscheiden, durch einen Test bevorzugt oder benachteiligt werden, wird unter dem Stichwort **Item-** oder **Testfairness** diskutiert. Es handelt sich hier um ein sehr komplexes Thema, wozu sich ein eigenes Kapitel schreiben ließe. Falls Testfairnessüberlegungen relevant sind, ist vor der Testkonstruktion festzulegen, anhand welcher Merkmale die Items hinsichtlich der Testfairness überprüft werden sollen. Diese Merkmale müssen im Rahmen der Datenerhebung miterfasst werden. Die Item- oder Testfairness kann mithilfe des Rasch-Modells überprüft werden. Das Rasch-Modell bietet Möglichkeiten, die Itemschwierigkeiten von unterschiedlichen Gruppen zu vergleichen (siehe grafischer Modelltest in *Kapitel 8.3.7*). Liegen im Rahmen einer solchen Analyse **Schwierigkeitsunterschiede** einzelner Items vor, ist dies ein Hinweis auf **mangelnde Testfairness**. In einem solchen Fall ist es günstig, die Items so umzuformulieren, dass solche Effekte nicht mehr auftreten. Wenn es nicht gelingt, solche Effekte zu eliminieren, das entsprechende Item aber für die Sicherung der Inhaltsvalidität unerlässlich ist, muss das Item beibehalten werden. Betrachten wir zur Verdeutlichung den Test zur Führerscheinprüfung. Die Fragen des Tests enthalten ausschließlich Items, die für die Teilnahme am Straßenverkehr notwendige Voraussetzungen schaffen. Items können aufgrund der inhaltlich klar definierten Itemmenge nicht aus dem Test entfernt werden. Es kann nun jedoch sein, dass manche Fragen sich für Gruppen von Personen mit unterschiedlichem Geschlecht,

unterschiedlicher Bildung oder unterschiedlichem Alter in ihrer Itemschwierigkeit unterscheiden. Beispielsweise sind manche Verkehrsschilder im Test so klein gedruckt, dass sie ältere Menschen benachteiligen und ihnen damit die richtige Lösung schwerer fällt. Hier sind zur Herstellung von Testfairness **sprachliche** oder aber auch **formale Modifikationen** möglich.

Darüber hinaus ist nachvollziehbar, dass eine bestimmte **Vertrautheit mit Tests** jeglicher Art auch einen Vorteil bei der Bearbeitung darstellt, z.B. durch wiederholte Testbearbeitung oder dadurch, dass bestimmte Gruppen mit bestimmten Inhalten eher vertraut sind (z.B. Rechenkonzentrationstest bei Servicekräften in der Gastronomie).

Ein weiterer übergeordneter Aspekt der Testfairness ist die **kulturelle Fairness** eines Tests. Es wird hier berücksichtigt, dass sich Kulturen in der Vertrautheit mit bestimmten Testmaterialien oder auch in der Wahrnehmung und Beurteilung bestimmter Fragen unterscheiden können. Um zu vermeiden, dass ein Test neben dem eigentlich zu messenden Konstrukt auch Fach- bzw. Sprachwissen oder Zugehörigkeit zu einem anderen Kulturkreis erfasst, sind besondere Anstrengungen nötig.

ZUSAMMENFASSUNG

Für die Testkonstruktion werden schon durch die Definition der Zielgruppe die Weichen für das Format des Tests, die Itemschwierigkeit und die Itemformulierung gelegt. Relevant sind dabei die folgenden Einflussgrößen (1) Alter und Bildung, (2) Erlebens- und Verhaltensspektrum, (3) Sprachbeherrschung der Zielgruppe sowie (4) Aspekte der Item-/Testfairness. In *Abbildung 3.2* ist dargestellt, welche der einzelnen Einflussgrößen sich auf das Format des Tests, die Itemschwierigkeit und die Itemformulierung auswirken.

	Format	Formulierung	Schwierigkeit
Alter und Bildung	•	•	•
Erlebens- und Verhaltensspektrum		•	•
Sprachliches Niveau		•	
Fairness			•

Abbildung 3.2: Einflussgrößen auf die Itemgenerierung.

ZUSAMMENFASSUNG

Beispiel 3.2 **Konstruktion eines Konzentrationstests**

In unserem Beispiel wollen wir ein Messinstrument zur Erfassung von Konzentration von jungen Erwachsenen konstruieren. Die Zielgruppe soll aus 15- bis 17-jährigen Berufsschülern bestehen. Daher ist bezüglich Alter und Bildung von einer vergleichsweise homogenen Gruppe auszugehen. Besonderer Wert muss auf die Konstruktion der Instruktion gelegt werden, um ungültige Testbearbeitungen zu minimieren. Dabei müssen Formulierungen an das Alter und die Bildung der Zielgruppe angepasst werden. Hier könnte die Think-aloud-Methode angewandt werden sowie Probetestungen mit einem anschließenden Interview. Da Konzentrationstestaufgaben in der Regel Aufgaben enthalten, die von allen Personen ohne Zeitbegrenzung richtig lösbar sind, ist die Aufgabenschwierigkeit hier von untergeordneter Bedeutung. Aber auch hier sind Probetestungen mit besonders leistungsfähigen Personen notwendig, um die Zeitbegrenzung für den Test optimal festzulegen. Es handelt sich ja um einen so genannten Speed-Test (siehe *Kapitel 1.3*). Würden bei der Testbearbeitung viele Personen alle Aufgaben bearbeiten, könnten wir zwischen diesen Personen keine Unterschiede mehr feststellen und damit nicht mehr differenzieren.

3.3 Testziel und Entscheidung für eine Konstruktionsstrategie

Im Prinzip unterscheidet man drei Ziele, die man mit einer Testkonstruktion erreichen will. Ein Ziel ist es, die Eigenschafts- oder Fähigkeitsausprägung einer Person festzustellen, ein weiteres ist die Gruppentrennung oder Klassifikation von Personen und das dritte die Erfassung von Wissen.

Bestimmung der Eigenschafts- oder Fähigkeitsausprägung Im ersten Fall gilt es vor allem, inhaltsvalide Items des Konstrukts zu finden. Dabei ist darauf zu achten, dass die Items nur ein Konstrukt erfassen und nicht mehrere. Darüber hinaus handelt es sich um reflektive Indikatoren. Das heißt alle Indikatoren des Konstrukts müssen miteinander korrelieren, und eine latente Variable erklärt die Zusammenhänge der Items vollständig.

Gruppentrennung Ist eine reine Gruppentrennung geplant, ist es vor allem wichtig, Items oder Aufgaben zu generieren, in denen sich die beiden Populationen auch tatsächlich unterscheiden. Das bedeutet es muss analysiert werden, worin die **markanten Unterschiede** zwischen den Populationen bestehen und wie dies in den Items oder Aufgaben berücksichtigt werden kann. Sollen beispielsweise Hypochonder von Gesunden mittels eines Fragebogens getrennt werden, ist anzunehmen, dass die Items auf Verhaltensweisen eines Hypochonders abzielen. Dabei spielt es keine Rolle, ob diese Verhaltensweisen tatsächlich alle einem Konstrukt zugeordnet werden können oder nicht. Vielmehr sollten sich die Antworten der Hypochonder und der Gesunden möglichst stark unterscheiden. Es handelt sich also um formative Indikatoren eines Konstrukts. Die Indikatoren können, müssen jedoch nicht untereinander korrelieren. Im Vordergrund steht die Inhaltsvalidität.

Wissenstests Bei der Konstruktion eines Wissenstests können sowohl reflektive als auch formative Indikatoren generiert werden. Auch hier steht zunächst die **Inhaltsvalidität** bei der Konstruktion **im Vordergrund**. Bei Wissenstests mit formativen Indikatoren können nachträglich mithilfe von exploratorischen Faktorenanalysen homogene Teilbereiche zusammengefasst werden (vgl. *Kapitel 6*). Es kann dann mithilfe von Strukturgleichungsmodellen oder dem Rasch-Modell geprüft werden, ob es sich innerhalb der homogenen Teilbereiche um reflektive Indikatoren handelt (vgl. *Kapitel 7* und *Kapitel 8*). Man könnte auch gleich eine Wissenstheorie zugrunde legen, die von reflektiven Indikatoren ausgeht. In diesem Fall würde es sich a priori um reflektive Indikatoren handeln.

In den nächsten Abschnitten werden nun die verschiedenen Methoden der Testkonstruktion näher erläutert und miteinander verglichen.

3.3.1 Rationale Testkonstruktion

Die rationale Methode der Testkonstruktion, oder auch deduktive Methode genannt, eignet sich vor allem dann zur Testkonstruktion, wenn eine gut ausgearbeitete Theorie für das zu untersuchende Konstrukt vorliegt. Ein sehr anschauliches Beispiel für die Vor- und Nachteile der rationalen Methode liefert der I-S-T 2000 R (Liepmann & Beauducel, Brocke, Amthauer 2007). Diesem Intelligenztest wurde ein Intelligenzmodell zugrunde gelegt, in dem unter anderem die beiden Intelligenzfaktoren fluide und kristalline Intelligenz unterschieden werden. Innerhalb dieser beiden breiten Faktoren werden zudem Facetten in Abhängigkeit der in den Aufgaben verwendeten Materialarten unterschieden. So gibt es jeweils verbale, numerische und figurale Facetten. Alle erfassten Intelligenzkomponenten werden im Handbuch definiert. Ein Blick auf die Definitionen verdeutlicht unmittelbar, dass die Güte der Definition darüber entscheiden kann, wie leicht sich Items generieren lassen. Wird Intelligenz nur als „gut denken“ definiert, haben wir sicher Probleme, uns Aufgaben zu überlegen. Ist die Definition jedoch spezifischer, wird die Aufgabenerstellung erleichtert. Im I-S-T 2000 R (Amthauer, Brocke, Liepmann & Beauducel, 2007, S. 92) wird Numerische Intelligenz z.B. folgendermaßen definiert: „... erfasst die Rechenfertigkeit und die Fähigkeit, logische Beziehungen zwischen Zahlen herzustellen.“ Sicher fällt es anhand dieser Definition deutlich leichter, Aufgaben abzuleiten. Das Beispiel macht deutlich, wie wichtig eine spezifische Definition bei der rationalen Fragebogenkonstruktion sein kann. Die Güte dieser Definition hängt sicher unmittelbar mit dem Expertenwissen des Testkonstrukteurs zusammen. Ist dieses begrenzt oder das Konstrukt generell zu wenig erforscht, empfiehlt es sich, durch Methoden wie die **Critical Incident Technique** (CIT, siehe *Kapitel 3.4.3*) eine möglichst verhaltensbasierte Definition zu nutzen.

3.3.2 Externale Testkonstruktion

Die externale Methode der Testkonstruktion wird häufig auch als kriteriumsorientierte Testkonstruktion bezeichnet. Diese Methode ist direkt mit dem Testzweck verbunden, zwischen verschiedenen Gruppen trennen zu können, zunächst unabhängig von der Frage der beteiligten Konstrukte. Bei der externalen Konstruktion suchen wir demnach Items, anhand derer wir Gruppen möglichst gut trennen können. Üblicherweise wird bei der Testkonstruktion zunächst eine große Anzahl an Items gesammelt, die potenziell zwischen den Gruppen unterscheiden könnten. In die finale Testversion werden dann

die Items übernommen, die aufgrund empirischer Untersuchungen tatsächlich zwischen den Gruppen trennen können. Dies kann sich beispielsweise in großen Mittelwertsunterschieden der Items zwischen den Gruppen äußern. Die vor allem in den USA sehr beliebten Integritytests (Verfahren zur Erfassung kontraproduktiven Verhaltens) wurden ursprünglich mit dieser Methode entwickelt. Ziel war es, zwischen Personen, die sich integer verhalten, und solchen, die zu kontraproduktivem Verhalten neigen, zu unterscheiden. In Deutschland existiert mit dem Inventar berufsbezogener Einstellungen und Selbsteinschätzungen (IBES, Marcus, 2006), dem bisher einzigen deutschen Integritätstest, ein solches Verfahren. Betrachtet man die Items dieses Fragebogens, wird deutlich, dass sie aus sehr verschiedenen Bereichen des menschlichen Erlebens stammen. So finden sich z.B. folgende Items: „Wenn sich zufällig die Gelegenheit ergibt, würde ich schon einmal Haschisch oder Marihuana probieren“ oder „Am Ende tun die Leute meistens, was ich von ihnen will“. Aufgrund der inhaltlichen Heterogenität kann es sehr schwierig sein, das Ergebnis eines solchen Tests (den Summenwert einer Skala) wie gewohnt zu interpretieren. Dazu beinhaltet der Fragebogen einfach zu viele unterschiedliche Aspekte. Somit ist oft lediglich eine Aussage möglich, die angibt, wie wahrscheinlich es ist, dass der Proband einer bestimmten Gruppe angehört, unabhängig davon, aufgrund welcher Verhaltensweisen diese Klassifikation vorgenommen wird. Für die Interpretation, warum er dieser Gruppe angehört, müsste auf die Interpretation der einzelnen Items zurückgegriffen werden. Offensichtlich kann das Ergebnis dieser Methode dazu führen, dass Items zwar gut trennen, aber sehr unterschiedlichen Konstrukten zugeordnet werden können und es sich damit um einen mehrdimensionalen Test handelt. Auch hier kann mithilfe der Faktorenanalyse versucht werden, die Fragen in homogenere Teilbereiche zusammenzufassen. Gelingt dies, ist eine Interpretation der Testskalen wiederum eindeutiger.

3.3.3 Induktive Testkonstruktion

Die induktive Methode der Testkonstruktion hat in der Persönlichkeitsforschung der letzten 40 Jahre deutliche Spuren hinterlassen. Ausgangspunkt dieser Methode ist wiederum eine große Itemmenge. Diese wird einer Stichprobe vorgegeben und die Daten werden dann mittels exploratorischer Faktorenanalysen ausgewertet. Ziel der Analyse ist es, Dimensionen zu finden, die den Itemantworten zugrunde liegen. Zu Beginn liegt also außer einer Idee, dass bestimmte Items dieselben Konstrukte erfassen könnten, keine theoretische Ausarbeitung vor. Vielmehr wird nach der Datenanalyse aufgrund der gefundenen Dimensionen ein theoretisches Modell entwickelt. Dies ist oft kritisiert worden, findet aber in der neueren Wissenschaftstheorie zunehmend Anklang. Haig (2005) schlug vor, psychologische Phänomene zunächst durch empirische Untersuchungen zu beschreiben und zu ergründen und aus diesen Erkenntnissen Theorien zu formulieren. Ein prominentes Beispiel für eine induktive Testkonstruktion sind sicher die Big 5. Die Ursprungsidee hierfür lag in der Sedimentationshypothese Klages, 1926: „Diejenigen Persönlichkeitseigenschaften, die besonders wichtig für den Alltag sind, finden Eingang in die naive Persönlichkeitstheorie. Je wichtiger sie sind, desto eher werden sie in einem einzigen Wort – einem Adjektiv oder Substantiv, seltener ein Verb – abgebildet.“ Verschiedene Forscher nutzten diese Idee, stellten Adjektiv- oder Itemlisten auf und gaben sie Probanden vor. Basierend auf den statistischen Analysen wurde schließlich das Modell der Fünf Faktoren entwickelt. Eine sehr gute Beschreibung findet sich bei John, Angleitner und Ostendorf (John, Angleitner & Ostendorf, 1988).

3.3.4 Prototypenansatz

Wenn wir an eine bestimmte Verhaltensweise denken, z.B. Extraversion, dann hat jeder Mensch eine prototypische Vorstellung von extravertiertem Verhalten. Empirische Untersuchungen zeigen, dass sich diese Vorstellungen meist ähneln (Cantor & Mischel, 1977; Rosch, 1975). Diesen Sachverhalt nutzt der Prototypenansatz. Dieser Konstruktionsansatz nutzt also die Idee, dass Menschen für jede Eigenschaft eine prototypische Vorstellung haben. Um nun Items zu generieren, müssen die prototypischen Vorstellungen verschiedener Personen gesammelt werden. Aus diesen lassen sich dann wiederum Items formulieren. Offensichtlich ist dieses Vorgehen vor allem bei Persönlichkeitsfragebögen und weniger bei Leistungstests geeignet.

Im Zusammenhang mit dem Prototypenansatz wird auch oft der so genannte Act-Frequency-Approach (Buss & Craik, 1983) eingesetzt. Diese Methode ähnelt ein wenig der **C**ritical **I**ncident **T**echnique (**CIT**, siehe *Kapitel 3.4.3*). Zunächst wird eine Gruppe von Probanden aus der späteren Zielpopulation gezogen. Den Probanden wird eine bestimmte Eigenschaft vorgegeben und sie werden gebeten, an eine Person zu denken, die sie gut kennen und die ihrer Meinung nach diese Eigenschaft am prototypischsten verkörpert. Die Probanden sollen dann die Verhaltensweisen, die die Person besonders prototypisch für die Eigenschaft machen, beschreiben. Aus diesen Informationen lassen sich dann Items generieren. Es empfiehlt sich, diese Items einer weiteren Stichprobe vorzugeben und bezüglich der Prototypizität raten zu lassen. Entsprechend der Ergebnisse können die Items ausgewählt werden. Um das gesamte Merkmalsspektrum abzudecken, kann dasselbe Vorgehen auch durchgeführt werden mit der Aufforderung, an eine Person zu denken, die möglichst wenig prototypisch für die zu messende Eigenschaft ist.

3.3.5 Vergleich der Methoden

Ein Vergleich dieser Methoden kann unmöglich zum Ziel haben, die beste Methode zu finden. Burisch (1984) konnte zeigen, dass keine der Methoden einen Vorteil in Bezug auf Validität erzielen kann. Lediglich bezüglich der Ökonomie sieht er einen Vorteil für rational konstruierte Instrumente. Allerdings ist diese Studie bereits mehr als 25 Jahre alt. Daher sollten die Ergebnisse nicht überinterpretiert werden. Jede Methode hat ihre Vor- und Nachteile. Durch eine Kombination der Methoden können hier sicher optimale Ergebnisse erzielt werden. Schließt sich an eine rationale Methode eine induktive an, lassen sich sicher schlechte Items besser finden als lediglich aufgrund theoretischer Überlegungen. Auch die Kombination mit dem Prototypenansatz ist sowohl für den rationalen als auch für den induktiven Ansatz zu empfehlen. Sicher nimmt die externale Methode einen gewissen Sonderstatus ein, da hier die Gruppentrennung im Vordergrund steht und nicht die inhaltsvalide Erfassung eines bestimmten Konstrukts. Dennoch ist eine Kombination beispielsweise mit dem Prototypenansatz denkbar.

Z U S A M M E N F A S S U N G

In den vorangegangenen Unterkapiteln wurde dargestellt, dass zunächst entschieden werden muss, ob objektive oder subjektive Indikatoren erhoben werden sollen. Im Anschluss wurde die Zielgruppe genauer definiert und Einflussgrößen benannt, die sich auf die weitere Testkonstruktion auswirken könnten. In diesem Unterkapitel wurde dargestellt, dass im nächsten Schritt das Testziel definiert wird. Dabei werden grundsätzlich drei Ziele unterschieden: (1) **Eigenschafts- oder Fähigkeitsausprägung einer Person feststellen**; (2) **Gruppen trennen**; (3) **Wissen erfassen**. Um diese Ziele zu erreichen, werden verschiedene Konstruktionsmethoden beschrieben: (1) Rationale oder deduktive, externale und induktive Methode sowie der Prototypenansatz. Die **rationale oder deduktive Methode** ist die Ableitung eines Tests aus einer Theorie heraus. Bei der **externalen Konstruktionsmethode** geht es darum, Items zu finden, die Gruppen bestmöglich trennen. Die **induktive Methode** basiert auf empirischen Untersuchungen von größeren Itemmengen und der Reduktion oder Ordnung dieser so analysierten Items. Als Methode der induktiven Testkonstruktion wird häufig die Faktorenanalyse verwendet. Der **Prototypenansatz** geht davon aus, dass jeder Mensch eine prototypische Vorstellung von einem Konstrukt hat. Um im Rahmen dieses Ansatzes Items zu generieren, müssen die prototypischen Vorstellungen verschiedener Personen gesammelt werden. Es konnte gezeigt werden, dass keine der genannten Methoden einen Vorteil bezüglich der Validität des späteren Tests bietet. In manchen Fällen kann es sinnvoll sein, die Methoden zu kombinieren.

Z U S A M M E N F A S S U N G

Beispiel 3.3 **Konstruktion eines Konzentrationstests**

Das Ziel der Testkonstruktion ist es, die Ausprägung von Konzentration bei jungen Erwachsenen festzustellen. Als Konstruktionsmethode entscheiden wir uns für die rationale Konstruktionsmethode und versuchen, ausgehend von einem Modell und Modelldefinitionen einen Konzentrationstest zu entwickeln. Wir verwenden die folgende Definition als Basis: „Konzentration ist die Fähigkeit, unter Bedingungen, die das Erbringen einer kognitiven Leistung normalerweise erschweren, schnell und genau zu arbeiten“ (Schmidt-Atzert, Büttner & Bühner, 2004; Schmidt-Atzert, Krumm & Bühner, 2008). Ein konkretes Modell (siehe *Abbildung 3.3*), worauf sich Konzentration in Abgrenzung zu Aufmerksamkeit bezieht, findet sich ebenfalls bei Schmidt-Atzert, Krumm und Bühner (2008, S. 11). Es stellt nach Angaben der Autoren einen Minimalkonsens aus rationalen und (induktiven) empirischen Ansätzen zur Modellbildung dar.

Zunächst wird ein Reiz präsentiert, der eine Reaktion erfordert. Dazwischen laufen die Wahrnehmung und die Weiterverarbeitung des Reizes ab. Im Rahmen der Wahrnehmung und Weiterverarbeitung kann nun Konzentration erfasst werden. Wird Konzentration im Rahmen der Wahrnehmung erfasst, wird dies als „Konzentrierte Aufmerksamkeit“ bezeichnet. Wird hingegen Konzentration im Rahmen der Weiterverarbeitung gemessen, wird dies als „Konzentrierte Weiterverarbeitung“ bezeichnet. Eine „Konzentrierte Weiterverarbeitung“ kann beispielsweise die Anwendung mathematischer Operationen darstellen. Damit ist

„Konzentrierte Weiterverarbeitung“ eher wissensbasiert als wahrnehmungsbasiert. Tests zur „Konzentrierten Aufmerksamkeit“ erfordern neben einer Aufmerksamkeitsanforderung auch eine willentliche Anstrengung, die beispielsweise durch die Bedingungen, unter denen die Aufgaben vorgegeben werden, sichergestellt werden kann. Hier wird willentliche Anstrengung als Bestandteil einer konzentrativen Leistung nicht subjektiv als „erlebte Anstrengung“ definiert, sondern objektiv über Aufgabenmerkmale. Dabei enthält die „Konzentrierte Aufmerksamkeit“ auch Anforderungen an die Alertness, Selektion oder Daueraufmerksamkeit, wenn diese unter willentlicher Anstrengung („cognitive effort“) zu erbringen sind.

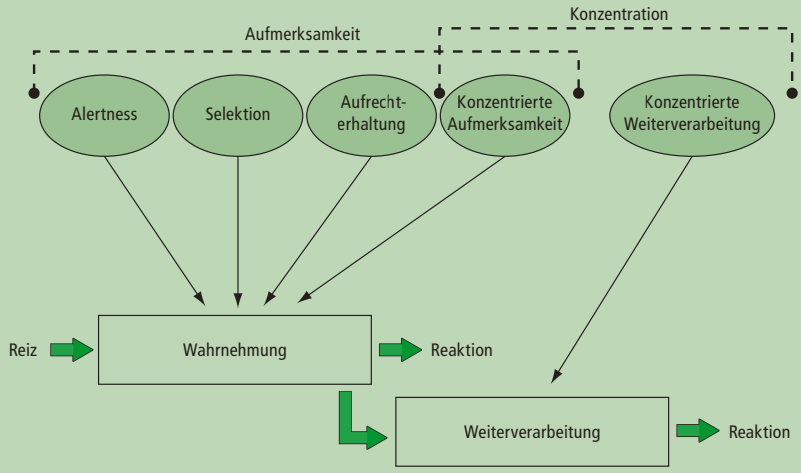


Abbildung 3.3: Konzentrationsmodell von Schmidt-Atzert, Krumm und Bühner (2008).

3.4 Generieren von Indikatoren und Eingrenzen des Konstrukts

Nehmen wir zunächst an, die Aufgabe bei einer Testkonstruktion bestünde darin, ein diagnostisches Verfahren für die Auswahl von Universitätsprofessoren zu konstruieren. Es wird schnell klar, dass dieser Beruf vielfältige und komplexe Anforderungen mit sich bringt. Um diese genauer zu ergründen, wird im Rahmen der Eignungsdiagnostik eine Anforderungsanalyse durchgeführt. Die Erfassung von Depressivität würde uns wahrscheinlich vor weitaus weniger Probleme stellen, da hier durch entsprechende Klassifikationssysteme die Anforderungen zur Erfassung von Depressivität schon definiert wurden. Aus diesen Beispielen wird deutlich, dass eine Anforderungsanalyse ein wichtiger erster Schritt auf dem Weg der Testkonstruktion darstellt. Sie dient damit zur Bestimmung des Konstruktraums und zur Eingrenzung des Konstrukts. Mit Bestimmung des Konstruktraums und Eingrenzung des Konstrukts ist gemeint, aus welchen Indikatoren (z.B. Verhalten, Einstellungen, Eigenschaften) das Konstrukt besteht, wo es Anknüpfungspunkte oder gar Überlappungsbereiche zu anderen Konstrukten gibt und welche Indikatoren dem Konstrukt zugrunde liegen sollen und welche nicht. Man spricht auch davon, das zu messende Konstrukt in ein

nomologisches Netzwerk einzuordnen. Das Ergebnis der Eingrenzung des Konstrukts bzw. der Indikatorensuche bildet dann die Basis zur Erstellung einer Arbeitsdefinition für das Messziel. Es gibt verschiedene Techniken, um die Basis für eine Konstruktdefinition zu erhalten, die im Folgenden näher erläutert werden. In der Regel werden die hier dargestellten Techniken kombiniert.

Top-Down- und Bottom-Up-Technik. Zwei Techniken haben sich besonders bewährt, um das zu erfassende Merkmal einzugrenzen: Top-Down-Technik und Bottom-Up-Technik. Im Rahmen beider Techniken ist das Messziel bereits definiert, es unterscheidet sich jedoch der Weg, wie die Indikatoren für das zu messende Konstrukt ermittelt werden. Im Rahmen der **Top-Down-Technik** werden Indikatoren des Konstrukts mithilfe von Expertenurteilen oder der Sammlung von Definitionsmerkmalen aus der bestehenden Literatur zusammengestellt. Das heißt es wird von oben nach unten vorgegangen, das Konstrukt (oben) definiert die Indikatoren (unten). Im Rahmen der **Bottom-Up-Technik** werden zunächst Verhaltensweisen oder Indikatoren mithilfe der analytisch-empirischen und/oder der personenbezogen-empirischen Ansätze ermittelt. Aus diesen Indikatoren (unten) wird das Konstrukt (oben) dann entwickelt:

1. Top-Down-Techniken:
   - Erfahrungsgeleitet-intuitiver Ansatz
   - Sammlung und Analyse von Definitionen/Literaturrecherche
2. Bottom-Up-Techniken:
   - Analytisch-empirischer Ansatz
   - Personenbezogen-empirischer Ansatz

Drei der oben genannten Ansätze stammen aus der **Eignungsdiagnostik**. Es sind dies der **erfahrungsgeleitet-intuitive**, der **analytisch-empirische** und der **personenbezogen-empirische Ansatz**. Diese Ansätze sind wiederum integrale Bestandteile von **Anforderungsanalysen** (Schuler, 2006). Im Rahmen der Eignungsdiagnostik muss genau definiert werden, welche Anforderungen ein potenzieller Bewerber erfüllen muss, um für einen bestimmten Beruf geeignet zu sein. Auf Basis der Anforderungsanalyse werden dann entsprechende eignungsdiagnostische Methoden ausgewählt und eingesetzt.

Die **Idee der Anforderungsanalyse** lässt sich auch auf die Testkonstruktion übertragen. Eine Anforderungsanalyse kann auch dazu dienen, einzelne Elemente (Indikatoren) des zu messenden Konstrukts zu bestimmen. Das Wort Anforderung bezieht sich nun nicht mehr darauf, welche Anforderungen ein möglicher Bewerber im Job später erfüllen muss. Vielmehr ist damit nun gemeint, welche Anforderungen ein Test oder Fragebogen erfüllen muss, um das angestrebte Konstrukt möglichst repräsentativ und trennscharf zu erfassen. Repräsentativ meint im günstigsten Fall, dass eine Auswahl aller relevanten Indikatoren des Konstrukts im Test enthalten ist und diese Indikatoren möglichst trennscharf sind. Trennscharf meint, dass sie keine anderen konstruktdivergenten Konstrukte gleichzeitig miterfassen. Strenger definiert meint repräsentativ, dass alle Indikatoren des Konstrukts im Vorhinein bekannt sind und eine repräsentative Auswahl aus all diesen Verhaltensweisen quasi zufällig aus einer Urne gezogen wird. Dies ist im Rahmen der Testkonstruktion jedoch nicht immer möglich, da selten der gesamte Konstruktraum bekannt ist. Wie jedoch bereits in *Kapitel 2.4.1* dargestellt, existieren Ansätze, das Itemuniversum darzustellen, wie die automatische Itemgenerierung (vgl. Arendasy & Sommer, 2010). In den folgenden Abschnitten wer-

den nun diese Ansätze zur Eingrenzung des Messziels näher erläutert. Um die einzelnen Methoden besser besprechen zu können und den gesamten Konstruktionsprozess anschaulicher zu machen, soll hier durchgehend ein Beispiel genutzt werden.

3.4.1 Erfahrungsgeleitet-intuitiver Ansatz

Die Idee hinter diesem Ansatz ist es, dass Experten, die eine Eingrenzung des Messziels vornehmen, sehr genaues und detailliertes Wissen über das zu erfassende Konstrukt besitzen. Diese werden in die Phase der Eingrenzung und Definition des Konstrukts einbezogen und benennen spezifische Indikatoren oder Elemente des Konstrukts.

Definition: Erfahrungsgeleitet-intuitiver Ansatz
Beim erfahrungsgeleitet-intuitiven Ansatz wird auf Basis von **Expertenwissen** definiert, welche Elemente oder Indikatoren eines diagnostischen Verfahrens das angestrebte Konstrukt messen.

Das Ergebnis dieses Ansatzes hängt direkt von der Qualität des Expertenwissens ab. Daher ist es günstig, mehrere Experten zu einem gemeinsamen Workshop zusammenzubringen oder zumindest zu befragen. Es handelt sich bei diesem Ansatz also um eine Top-Down-Technik. Damit ist gemeint, dass die Indikatoren für das zu erfassende Konstrukt nicht unmittelbar aus Beobachtungen stammen, sondern vielmehr aufgrund des Wissens einiger Personen über das Konstrukt ermittelt werden.

Exkurs 3.3 **Laien als Experten**

Auch die Meinung von Laien kann zur Eingrenzung des Konstrukts informativ sein. Zum einen hat jede Person eine intuitive oder prototypische Vorstellung (siehe *Kapitel 3.3.4*) von dem, was beispielsweise unter Depression zu verstehen ist. Zum anderen wird der Test oder Fragebogen vor allem von Laien bearbeitet werden. Im Rahmen der Personalauswahl kann es für die Akzeptanz des Tests abträglich sein, wenn die Aufgaben oder Fragen eines Tests keinen offensichtlichen Anforderungsbezug besitzen. Man spricht auch von mangelnder Augenscheinvalidität eines Tests. In diesen Fällen können die getesteten Personen nicht einschätzen, was eigentlich gemessen wird, oder sie ziehen falsche Schlüsse. Da aber wichtige Entscheidungen für die Person vom Test abhängen, kann eine solche Fehleinschätzung zu Reaktanz oder sogar dazu führen, dass das Auswahlverfahren als unfair empfunden wird. Liegt es also im Interesse des Testanwenders, dass die Zielgruppe, in der der Test oder Fragebogen später eingesetzt werden soll, erkennt, was erfasst wird, dann sollten auch Laien befragt werden. Auch, wenn es die Absicht ist, die Zielgruppe im Dunkeln über das Testziel zu lassen, kann eine solche Befragung sinnvoll erscheinen. Sie sichert in diesem Fall ab, dass Laien tatsächlich nicht das Testziel erahnen können. Vor allem bei der Konstruktion von objektiven oder projektiven Tests kann dies sinnvoll sein.

3.4.2 Sammlung und Analyse von Definitionen/Literaturrecherche

Ziel der Sammlung und Analyse von Definitionen/Literaturrecherche ist es, Indikatoren bzw. Definitionsmerkmale eines Konstrukts zu finden. Dazu können Literaturdatenbanken genutzt werden oder bereits publizierte Testmanuale.

Literaturdatenbanken Eine genaue Beschäftigung mit der existierenden Literatur ist im Rahmen der Testkonstruktion unerlässlich, wenn es um die Eingrenzung des zu messenden Merkmals geht. Dazu stehen verschiedene Quellen zur Verfügung. Um eine gute Einführung in eine bestimmte Thematik zu erhalten, können beispielsweise einschlägige Lehrbücher genutzt werden. Außerdem finden sich oft auch Übersichtsartikel, so genannte Reviews, in verschiedenen Fachzeitschriften, die den Stand der Literatur aktuell und detailliert umreißen. Dort findet man auch Hinweise auf speziellere Literatur. Sie lässt sich ebenso mithilfe von Literatursuchmaschinen finden. Zu diesen Suchmaschinen gehören beispielsweise PsycInfo, Psyndex, GoogleScholar oder Web of Science. Das Prinzip dieser Suchmaschinen besteht darin, anhand der vom Benutzer vorgegebenen Suchwörter Artikel aus Fachzeitschriften zu suchen. Um nicht direkt im Angebot der verschiedenen Suchmaschinen zu versinken, empfiehlt es sich, zunächst auf vorhandene Literatur, z.B. in Lehrbüchern, zurückzugreifen, um dann das Erstellen von Suchbegriffen zu erleichtern.

Bestehende Testverfahren Eine weitere Informationsquelle sind bereits existierende Testverfahren, die ein ähnliches Konstrukt erfassen sollen. Wie haben die Testautoren hier die Eingrenzung vorgenommen? Welche Definitionen geben sie an und welche Konstrukte werden abgegrenzt? Das Sammeln und Analysieren solcher Definitionen ermöglicht es, wiederkehrende Definitionsmerkmale zu identifizieren. Solche Merkmale stellen offensichtlich Kernbereiche des zu messenden Konstrukts dar, da sie in mehreren Definitionen Verwendung finden.

Beispiel 3.4 **Konstruktion eines Konzentrationstests**

Für unser Beispiel entscheiden wir uns, aus der Literatur Definitionsmerkmale herauszusuchen und zu systematisieren. Dies ist bei Bühner (2001) bereits geschehen. Man kann folgende Definitionsmerkmale der Konzentration unterscheiden und zusammenfassen.

Der Beginn einer konkreten Konzentrationsleistung besteht darin, ein Handlungsziel, beispielsweise eine spezifische Reizkonfiguration, zu entdecken und auf diese in einer bestimmten vorgegebenen Weise zu reagieren. Konzentration beginnt demnach mit der zweckgebundenen intentionalen Zuwendung zu Reizen (Beckmann, 1991; Beckmann & Strang, 1993; Berg, 1991; Berg & Imhof, 2001; Brickenkamp, 2002; Brickenkamp & Karl, 1986; Hoffmann, 1993; Neumann & Simon, 1994; Posner & Rafal, 1987; Westhoff, 1991, 1995). Ein Teil der Wahrnehmung wird dabei bewusst auf einen begrenzten Teil der Umgebung ausgerichtet (Selektion), wobei störende (für die Aufgabenbearbeitung irrelevante) Reize bewusst ausgeblendet bzw. abgeschirmt werden (Beckmann, 1991; Beckmann & Strang, 1993;

Berg & Imhof, 2001; Brickenkamp, 2002; Brickenkamp & Karl, 1986; Hoffmann, 1993; Neumann & Simon, 1994). Die wahrgenommenen Reizkonfigurationen werden fortlaufend und zielorientiert beurteilt (Berg, 1991; Berg & Imhof, 2001; Brickemkamp, 2002; Brickenkamp & Karl, 1986; Westhoff, 1991) und entsprechende (instruktionsgemäße) Handlungsreaktionen werden mental koordiniert sowie meist motorisch ausgeführt (Berg, 1991; Berg & Imhof, 2001; Neumann & Simon, 1994; Posner & Rafal, 1987; Westhoff, 1991). Konzentriertes Arbeiten wird meist als anstrengend erlebt (Berg & Imhof, 2001; Posner & Rafal, 1987; Westhoff 1991, 1995) und von internen Zustandsvariablen (z.B. Ermüdung) oder externen Zustandsvariablen (z.B. Lärm) sowie von Persönlichkeitsvariablen (z.B. Leistungsmotivation oder Intelligenz) beeinflusst (Westhoff, 1995).

Nun müssen wir versuchen, bei der Aufgabenkonstruktion diese Definitionsmerkmale so weit wie möglich umzusetzen. Es müssen dabei nicht zwingend alle Definitionselemente umgesetzt werden. Dies sollte bei der Definition des Messgegenstandes des Tests mitberücksichtigt werden und wird im nächsten Unterkapitel näher erläutert.

3.4.3 Analytisch-empirischer Ansatz

Dieser Ansatz der Merkmalseingrenzung ist wahrscheinlich vor allem im Bereich der Arbeits-, Betriebs- und Organisationspsychologie vertreten. Er wird eigentlich als Arbeitsplatz-analytisch-empirischer Ansatz bezeichnet. Laut Schuler (2006, S. 48) untersucht diese Methode „... die beruflichen Tätigkeiten und Situationen mittels formalisierter Vorgehensweisen (Fragebogen) an konkreten Arbeitsplätzen". Das heißt zur Bestimmung des Messgegenstands sollen standardisierte Beobachtungs- und Klassifikationsverfahren eingesetzt werden. Ein Beispiel hierfür wäre der *Fragebogen zur Arbeitsanalyse* (Frieling & Hoyos, 1978). Im Rahmen der Test- und Fragebogenentwicklung ist es sicher angemessener, allgemeiner von einer analytisch-empirischen Analyse zu sprechen.

Definition: Analytisch-empirischer Ansatz
Die Grundidee des analytisch-empirischen Ansatzes auf die Testkonstruktion verallgemeinert lautet, dass die Elemente des Messgegenstands durch die Verwendung standardisierter Beobachtungs- oder Befragungsinstrumente identifiziert werden.

Allerdings wird die Verwendung von standardisierten Fragebögen im Rahmen einer Testkonstruktion außerhalb der Arbeits-, Betriebs- und Organisationspsychologie in den meisten Fällen zu aufwendig sein. Daher wird im Folgenden eine Methode vorgestellt, die auch zu diesem Ansatz gezählt werden kann, aber mehr Gestaltungsfreiraum bietet.

Critical Incident Technique Dabei handelt es sich um die so genannte Critical Incident Technique, CIT (Flanagan, 1954). Die Ausgangsidee der CIT ist, dass sich das Verhalten von Personen in verschiedenen Situationen unterscheidet. Solche Situationen

können nun bezogen auf das zu messende Konstrukt mehr oder weniger relevant sein. So sind Situationen in einem sozialen Kontext sicher geeigneter, extravertiertes Verhalten bei einer Person zu beobachten, als Situationen, in denen eine Person alleine ist. Situationen können also mehr oder weniger kritisch im Sinne des zu erfassenden Konstrukts sein. In kritischen Situationen zeigen Menschen mehr Verhalten, das dem zu messenden Konstrukt zugeordnet werden kann. Die genaue Analyse der unterschiedlichen Verhaltensweisen sollte es ermöglichen, Rückschlüsse darüber zu ziehen, welche Verhaltensweisen dem Konstrukt zugrunde liegen. Das heißt zu Beginn ist es notwendig, kritische Situationen und vor allem die entsprechenden Verhaltensweisen zu sammeln. Dies erfolgt meist retrospektiv. Experten werden also gebeten, sich an solche Situationen zu erinnern. Dabei bezieht sich der Begriff Experte nicht zwingend auf ein überlegenes Fachwissen. In einem eignungsdiagnostischen Kontext wären z.B. auch Personen als Experten anzusehen, die auf der zu besetzenden Stelle selbst gearbeitet haben. Auch deren Vorgesetzte und Mitarbeiter verfügen sicher über einen reichen Erfahrungsschatz in Bezug auf kritische Verhaltensweisen, die in diesem Beruf auftreten.

Beispiel 3.5 **Critical Incident Technique (CIT)**

In unserem Beispiel, der Konstruktion eines Instruments zur Erfassung von Konzentration, könnte man also einschlägige Forscher, Therapeuten und eventuell einschlägige Lehrbuchautoren bitten, eine Reihe von Fragen zu beantworten. Es wäre hier auch möglich, direkt Personen mit Konzentrationsproblemen zu befragen, wie sich diese bemerkbar machen. Je nach Zielgruppe kann eine direkte Befragung sehr hilfreich sein. Da neben den Situationseigenheiten vor allem auch das Verhalten der Handelnden erfasst werden soll, zielen die Fragen meist auch darauf ab. Die folgenden Fragen stellen Beispielfragen dar, wie sie im Rahmen einer solchen **Critical Incident Technique** eingesetzt werden können:

Frage 1: Erinnern Sie sich bitte an eine Situation, in der Sie unkonzentriert waren!

- Schildern Sie bitte die genauen Umstände!
- Woran haben Sie gemerkt, dass Sie unkonzentriert waren?
- Was waren die Konsequenzen?

Frage 2: Erinnern Sie sich bitte an eine Situation, in der eine andere Person unkonzentriert war!

- Schildern Sie bitte die genauen Umstände!
- Wie hat sich die Person verhalten?
- Was waren die Konsequenzen?

Das Ergebnis solcher Befragungen, die als **strukturiertes Interview** oder auch mithilfe eines **Fragebogens mit offenen Fragen** durchgeführt werden können, sind genaue Situations- und Verhaltensbeschreibungen. Um aus diesen Beschreibungen Indikatoren für das zu messende Konstrukt abzuleiten, muss auf Basis der Information auf

zugrunde liegende Personeneigenschaften geschlossen werden. Flanagan (1954, S. 335) schreibt hierzu: „The incidents must be studied in the light of relevant established principles of human behavior and of the known facts regarding background factors and conditions operating in the specific situation." Mit anderen Worten, an dieser Stelle ist es wichtig, dass nicht nur Expertenwissen bezüglich des zu erfassenden Konstrukts, sondern auch bezüglich anderer psychologischer Konstrukte vorhanden ist. Andernfalls besteht die Gefahr, dass man wichtige Konstruktüberlappungen übersieht. Das Ergebnis sind dann Testverfahren, die Konstrukte erfassen, die wenig trennscharf zu anderen konstruktdivergenten Konstrukten sind und verschiedene Konstrukte gleichzeitig messen. Im Lauf des Kapitels wird immer wieder auf die CIT Bezug genommen und aufgezeigt, wie die Ergebnisse dieser zunächst aufwendig erscheinenden Technik sehr hilfreich bei der Testkonstruktion sein können. Im Gegensatz zu den vorher beschriebenen Top-Down-Techniken werden bei der CIT also zunächst vom Verhalten ausgehend Verhaltensindikatoren abgeleitet und diese dann Konstrukten zugeordnet. Es wird vom Verhalten auf die Konstrukte geschlossen und nicht umgekehrt von den Konstrukten auf das Verhalten, daher bezeichnen wir dieses Vorgehen auch als Bottom-Up-Technik.

3.4.4 Personenbezogen-empirische Methode

Bei der personenbezogen-empirischen Methode wird auf bereits existierende empirische Befunde zurückgegriffen. Das heißt aus den empirisch gefundenen Zusammenhängen zwischen Personenmerkmalen und bestimmten Kriterien wird auf die zugrunde liegenden Verhaltensindikatoren geschlossen. Es gibt im Rahmen der Testkonstruktion zwei Hauptanwendungsbereiche für dieses Vorgehen. Zum einen kann es zur Verortung des Konstrukts in ein **nomologisches Netzwerk** genutzt werden, zum anderen kann es helfen, Konstrukte zur **Formung eines Indexes** zu finden, um die **Kriteriumsvalidität** eines Indexes zu **maximieren**.

Definition: Personenbezogen-empirische Methode
Bei der personenbezogen-empirischen Methode wird aufgrund von empirisch in der Literatur gefundenen Zusammenhängen zwischen Personenmerkmalen und bestimmten Kriterien auf zugrunde liegende Indikatoren geschlossen.

Nomologisches Netz Es ist also durchaus ratsam, bereits bei der Feststellung der Verhaltensindikatoren eine tiefer gehende Literaturrecherche zu machen. Informationen über Zusammenhänge zwischen verschiedenen Konstrukten helfen, mögliche Konstruktüberlappungen zu finden. Dies wiederum ist bei der genauen Definition des zu erfassenden Konstrukts von großer Bedeutung und auch bei der Wahl divergenter Testverfahren zur Konstruktvalidierung. Ein Nachteil dieser Methode ist jedoch, dass sie nur sinnvoll eingesetzt werden kann, wenn es bereits verlässliche Verfahren zur Erfassung des zu messenden Konstrukts gibt. Andernfalls sind die empirischen Befunde wenig aussagekräftig. Existieren solche Verfahren, muss die Entwicklung eines neuen Verfahrens gut begründet werden. Der neue Test sollte schließlich etwas leisten, was bisherige Verfahren noch nicht können.

Maximierung der Kriteriumsvalidität Man kann diese Methode insbesondere nutzen, wenn man eine besonders hohe Kriteriumsvalidität mit dem Test erzielen möchte und dabei die Itemhomogenität im Sinne von Eindimensionalität eines Tests von untergeordneter Bedeutung ist. Dies klingt zunächst abstrakt, kann aber an einem Beispiel verdeutlicht werden. Nehmen wir an, es soll ein Screeninginstrument (Index) zur Fahrtauglichkeit entworfen werden. Dabei ist angestrebt, Items zu entwickeln, die mit Fahrtauglichkeit als Kriterium hoch korrelieren, aber nicht unbedingt ein eindimensionales Konstrukt darstellen, sondern aus Indikatoren unterschiedlicher Konstrukte bestehen oder sogar Personenmerkmale darstellen. Hat man viele solcher kriteriumsvaliden Items zu einem Index formiert, wird folglich auch die Kriteriumsvalidität des Gesamttests maximiert. Die Indikatoren selbst müssen dabei nicht miteinander korrelieren, so wie man es für reflektive Indikatoren eines möglichst eindimensionalen Konstrukts erwarten würde, sondern können Indikatoren verschiedener Konstrukte sein. Man könnte als Kriterium in diesem Fall beispielsweise eine Reihe kritischer Situationen in einem Fahrsimulator zusammenstellen und die Fahrfehler als Kriterium heranziehen. In Metaanalysen werden häufig Informationen über solche Zusammenhänge zwischen Prädiktoren und Kriterien dargestellt. Diese Metaanalysen liefern also verschiedene Konstrukte und Einflussgrößen als Anhaltspunkte für eine Testkonstruktion, die die Kriteriumsvalidität maximiert. Es ist also durchaus ratsam, bereits bei der Feststellung der Indikatoren eine tiefer gehende Literaturrecherche durchzuführen. Ein Nachteil dieser Methode ist jedoch, dass bei einem extremen Indexwert nicht klar ist, welcher Indikator oder welche Indikatoren dafür verantwortlich sein können. Hierzu muss jedes Element des Indexes betrachtet und interpretiert werden.

Z U S A M M E N F A S S U N G

In den letzten Abschnitten wurden vier Methoden vorgestellt, die dabei helfen können, Indikatoren eines zu erfassenden Konstrukts zu finden bzw. zu definieren. Je nach Fragestellung ist es nicht in jedem Fall möglich, direkt eine Definition des zu messenden Konstrukts zu erstellen. Häufig ist es notwendig, durch weitere Schritte zunächst zu (be-)schließen, was eigentlich genau gemessen werden soll. Die Techniken, die hier vorgestellt wurden, sind zum einen die **Top-Down-Techniken** mit dem **erfahrungsgeleitet-intuitiven Ansatz** sowie der **Literatursuche** und die **Bottom-Up-Techniken** mit den **analytisch-empirischen** und **personenbezogen-empirischen** Ansätzen. Dabei wurde aufgezeigt, dass die Top-Down-Techniken vorhandenes Expertenwissen nutzen, um Bestandteile des zu erfassenden Konstrukts zu identifizieren. Die Arbeitsplatz-analytisch-empirische Methode nutzt standardisierte Instrumente, um solche Anforderungen zu identifizieren. Zu diesem Ansatz wird auch die **Critical Incident Technique (CIT)** gezählt. Sie hat sich besonders bewährt, um das Messziel trennscharf zu benennen. Schließlich wurde dargestellt, dass der **personenbezogen-empirische** Ansatz Erkenntnisse aus bereits vorhandenen empirischen Untersuchungen nutzt, um das Konstrukt in ein nomologisches Netzwerk einzuordnen oder Konstrukte und/oder Personenmerkmale zu finden, die einen möglichst kriteriumsvaliden Index bilden.

Das **Ziel** der genannten Ansätze besteht darin, eine Reihe von **Indikatoren** für das zu erfassende Konstrukt **zu erhalten**. Existiert eine erschöpfende Liste solcher Indikatoren bzw. ist eine Eingrenzung des Merkmals gelungen, muss im Anschluss daran das Konstrukt genau definiert werden.

Z U S A M M E N F A S S U N G

3.5 Erstellen einer Definition des Messgegenstandes

Im letzten Abschnitt wurde dargestellt, welche Methoden es gibt, um Indikatoren für das zu erfassende Konstrukt bzw. das nomologische Netzwerk des Konstrukts zu erhalten. Sobald dies geklärt ist, sollte ganz genau festgelegt werden, was unter dem zu messenden Konstrukt zu verstehen ist. Warum ist das so wichtig? Ein kurzer Blick in die Geschichte der Intelligenzmessung verdeutlicht die Bedeutsamkeit einer genauen Definition des Messgegenstandes. Für das Konstrukt Intelligenz existieren so viele Definitionen und Beschreibungen, dass der Begriff eigentlich kaum noch eindeutig verwendbar ist. Je nachdem, welcher Definition bzw. welchem theoretischen Modell man sich als Testkonstrukteur verschreibt, wird der zu konstruierende Test sehr verschieden ausfallen. Als extreme Beispiele können sicherlich der I-S-T 2000 R (Amthauer, Brocke, Liepmann & Beauducel, 2001) und der Zahlen-Verbindungs-Test (Z-V-T, Oswald & Roth, 1987) miteinander verglichen werden. Während Ersterer vorwiegend schlussfolgerndes Denken erfasst, erfasst Letzterer eher Konzentration (vgl. Schmidt-Atzert, Bühner & Enders, 2006). Dieses einfache Beispiel zeigt, dass es enorm wichtig ist, das zu messende Konstrukt genau zu definieren.

Neben der Definition sind auch Vorüberlegungen günstig, welche anderen Konstrukte in Zusammenhang mit dem zu messenden Konstrukt stehen (**personenbezogen-empirische Methode**). Die Konstrukte können zum Teil überlappen, wie es bei Konzentration, Intelligenz und Koordination sowie Arbeitsgedächtniskapazität der Fall ist (Bühner, Krumm, Ziegler & Plücken, 2006). Es gibt aber auch nicht überlappende Konstrukte: So wird beispielsweise angenommen, dass die Persönlichkeitseigenschaften Neurotizismus und Extraversion nicht korrelieren (Eysenck & Eysenck, 1964). Zur Feststellung von Konstruktüberschneidungen können, wie oben beschrieben, auch Informationen gesammelt werden. Idealerweise liegen nach dem Durchführen verschiedener Top-Down- und Bottom-Up-Ansätze viele Informationen vor, die sich zum einen direkt auf Indikatoren eines Konstrukts beziehen und zum anderen eher auf die Lage des Konstrukts in einem nomologischen Netz. Nun ist es wichtig sicherzugehen, dass die verschiedenen Indikatoren des zu erfassenden Konstrukts tatsächlich dem Konstrukt zuzuordnen sind bzw. inwieweit sie mit anderen Konstrukten zusammenhängen. Dies ist ein wichtiger Schritt im Definitionsprozess, und an dieser Stelle sollte ruhig etwas mehr Aufwand betrieben werden, da durch eine genaue Definition und eine umfangreiche Abgrenzung zu anderen Konstrukten unter anderem die Wahrscheinlichkeit für eine hohe Inhaltsvalidität des Tests steigt.

Eine große Hilfe hierbei ist es, die Ergebnisse solcher Recherchen grafisch festzuhalten (siehe *Abbildung 3.4*). Dabei ist zu beachten, dass die Größe der Zusammenhänge zwischen den angrenzenden Konstrukten in der Darstellung aus Gründen der Übersichtlichkeit ebenso nicht berücksichtigt wurde wie die Überlappungen zwischen den einzelnen Konstrukten. Wird eine solche Grafik auf eine Tafel oder eine Metaplanwand übertragen, lässt sich hierin später auch die Definition des Messgegenstandes hineinschreiben und so auch eine Abgrenzung zu konstruktdivergenten Konzepten vornehmen.

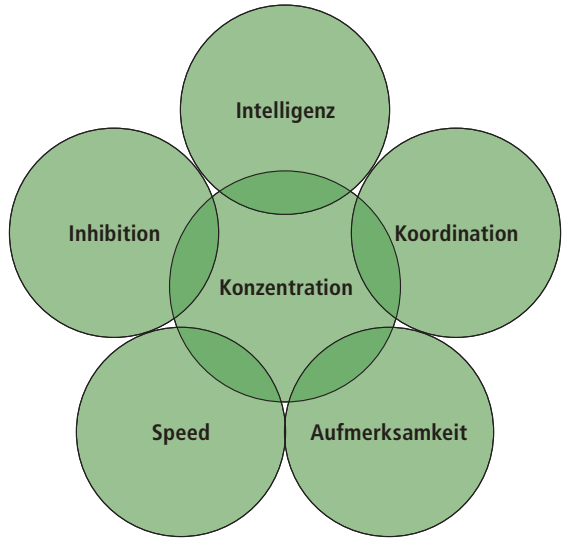


Abbildung 3.4: Nomologisches Netzwerk der Konzentration.

Nun kann anhand dieses Schaubilds für jedes Item oder jede Testart geprüft werden, ob er bzw. sie im Konstruktbereich von Konzentration oder von einem anderen Konstrukt liegt. Würde man eine solche Grafik für ein Konstrukt mit subjektiven Indikatoren, beispielsweise Depressivität, erstellen, könnte man nun die konstruierten Items den Kreisen oder Überlappungsbereichen zuordnen, um einen Überblick über die Items und deren Einordnung zu erhalten. Wenn möglich, sollten Items in den Test aufgenommen werden, die im Kernbereich des Konstrukts liegen und nicht in den Überlappungsbereichen. Werden Items aus den Überlappungsbereichen und den Kernbereichen vermischt, führt dies in der Regel zu mehrdimensionalen Konstrukten. Manchmal liegen jedoch Items gänzlich in den Überlappungsbereichen: Ein Item, dass Konzentration misst, misst auch Aufmerksamkeit, da die Aufmerksamkeitsleistung der konzentrativen Leistung vorgeschaltet ist. Trifft dies für alle Items in gleichem Maße zu, ist dies weniger ein Problem (siehe *Kapitel 2.1*). In diesem Fall ist die Konstruktüberlappung der Items mit der Korrelation der Konstrukte zu erklären. Insgesamt wird hier deutlich, dass extensives Wissen über psychologische Konstrukte vorhanden sein und genutzt werden muss, um dieses nomologische Netz zu konzipieren. Gelingt eine Abgrenzung zu anderen Konstrukten nicht, besteht die Gefahr, dass Items, die eigentlich unterschiedliche Konstrukte erfassen, für ein und dasselbe Konstrukt genutzt werden oder dass Tests mit anderen Leistungen über die Korrelation zwischen den Konstrukten hinaus verschmutzt sind.

Häufig ist es im Rahmen der Testkonstruktion hilfreich, eine so genannte Testdefinition zu erstellen. Damit ist gemeint, dass mit der Definition keine Allgemeingültigkeit angestrebt, sondern vielmehr eine Arbeitsgrundlage geschaffen wird, die eine Basis für die Testkonstruktion und einen wichtigen Bestandteil des Testhandbuchs darstellt. Im nächsten Abschnitt wird kurz erläutert, wie eine Definition des Messgegenstandes erstellt werden kann. Basis zur Erstellung der Arbeitsdefinition sind entweder die Ergebnisse der **Literatursuche** oder der Befragung von **Experten** sowie der **Critical Incident Technique**.

Top-Down-orientierte Testdefinition Im Rahmen der Literatursuche haben wir bereits verschiedene Definitionen gefunden und haben unter den unterschiedlichen Definitionen nach wiederkehrenden Kernstücken gesucht, die dann genutzt werden, um eine eigene Definition aufzustellen. Diese Definition vereint oder abstrahiert die übereinstimmenden Definitionsbestandteile aus den Forschungsarbeiten. Für die Erstellung der Definition werden dann entweder alle Definitionsbestandteile oder eine gut begründete, reduzierte Anzahl der Definitionsbestandteile genutzt. Eine Arbeitsdefinition kann aber auch mithilfe von Experten bestimmt werden. Diese erarbeiten beispielsweise in einem Workshop eine Testdefinition. Beide Strategien entsprechen Top-Down-Strategien.

Bottom-Up-orientierte Definition Eine weitere Möglichkeit, eine Definition zu erstellen, besteht in der Anwendung einer Bottom-Up-Technik, der Critical Incident Technique. Wie bereits beschrieben, werden mit dieser Technik **Verhaltensindikatoren** aus unterschiedlichen Situationen gesammelt und in Gruppen zusammengefasst. Aus diesen Gruppen von Verhaltensweisen lassen sich dann nicht nur Rückschlüsse über die zugrunde liegenden Konstrukte ziehen. Vielmehr lassen sich diese Konstrukte dann auch anhand der Verhaltensweisen definieren. Diese verhaltensbasierte Definition hat sicher den Vorteil, dass sie sehr verständlich ist und vor allem trennscharf. Ein Verhaltensanker, der einem Konstrukt zugeordnet wird, wird keinem weiteren Konstrukt zugeordnet. Ein Nachteil dieser Methode besteht darin, dass die Definition eines Konstrukts sehr breit ausfallen kann, wenn viele Verhaltensindikatoren gesammelt wurden und einem Konstrukt zugeordnet werden. Es kann aber auch sein, dass Konstrukte nur durch wenige Verhaltensindikatoren definiert sind. In einem solchen Fall wird nur ein kleiner Verhaltensausschnitt gemessen.

Beispiel 3.6 **Konstruktion eines Konzentrationstests**

Für unser Beispiel wählen wir folgende Definition des Messgegenstandes. Eine globale Definition wurde bereits vorgestellt: „Konzentration ist die Fähigkeit, unter Bedingungen, die das Erbringen einer kognitiven Leistung normalerweise erschweren, schnell und genau zu arbeiten." Man könnte nun ausgehend von dieser Definition und den immer wiederkehrenden Definitionselementen folgende Definition des Messgegenstandes des zu konstruierenden Tests entwickeln:

Der Test erfasst die Fähigkeit, sich zu konzentrieren. Darunter wird das Ausführen einer zielgerichteten fortlaufenden Selektion von Reizkonfigurationen unter Bedingungen verstanden, die normalerweise das Erbringen dieser Leistung erschweren. Erschweren meint, dass bei der Aufgabenbearbeitung irrelevante, störende Reize ausgeblendet werden müssen.

Auf die Definitionselemente, dass eine bewusste Ausblendung von irrelevanten Reizen erfolgen muss, und auch, dass die Aufgabe subjektiv als anstrengend erlebt werden sollte, wird verzichtet, da diese Informationen nur durch die Befragung der Personen gewonnen werden könnten und nur schwerlich objektiv zu erfassen sind.

Z U S A M M E N F A S S U N G

Um eine Arbeitsdefinition zu erstellen, können verschiedene Schritte durchgeführt werden. Basis sind entweder **Top-Down-** oder **Bottom-Up-Ansätze**. Bei den Top-Down-Ansätzen werden anhand einer umfassenden **Literatursuche** immer wiederkehrende Definitionselemente eines Konstrukts gesammelt und systematisiert. Danach wird der Test gemäß dieser oder einer reduzierten Anzahl von Definitionselementen konstruiert. Unter diesen Ansatz fällt auch das Erarbeiten einer Definition des Messgegenstandes im Rahmen eines **Expertenworkshops**. Bei den **Bottom-Up-Ansätzen** werden zunächst mithilfe der **Critical Incident Technique** (CIT) Verhaltensanker für ein Konstrukt ermittelt und dann verschiedenen Konstrukten zugeordnet. Dies führt zu besonders trennscharfen Definitionen.

Z U S A M M E N F A S S U N G

3.6 Wahl des Itemformats

Welches Itemformat soll ich auswählen?

Es werden **gebundene** und **freie Antwortformate** unterschieden. Bei gebundenen Antwortformaten sind konkrete Lösungsmöglichkeiten oder Antwortalternativen vorgegeben. Die freie Aufgabenbeantwortung ist nicht oder nur wenig durch Antwortvorgaben eingeschränkt. Auf beide Aufgabentypen wird im Folgenden näher eingegangen. Zuvor wird jedoch im Exkurs geschildert, wie man sich den Antwortprozess auf eine Frage vorstellen kann.

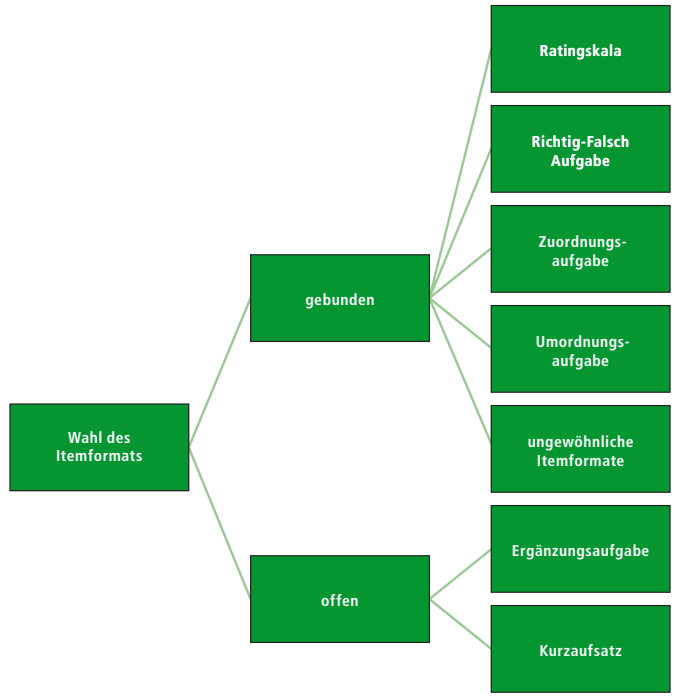


Abbildung 3.5: Darstellung verschiedener Itemformate.

Exkurs 3.4 Wie beantworten wir eigentlich die Items in einem Fragebogen?

Routinemäßig werden Fragebögen zur Erfassung von Persönlichkeitseigenschaften eingesetzt. Die Idee ist, dass aufgrund der Itemantworten Rückschlüsse auf die zu messende Persönlichkeitseigenschaft möglich sein sollen. Mit anderen Worten, die Unterschiede zwischen den Personen in dieser Eigenschaft verursachen auch die Unterschiede im Ankreuzverhalten bzw. dem Lösen oder Nichtlösen einer Aufgabe. Diese Annahme liegt zunächst dem in *Kapitel 2* beschriebenen Ansatz einer latenten Variablen zugrunde.

Aber wie genau soll es funktionieren, dass die Persönlichkeitseigenschaft sich in den Antworten niederschlägt? Was genau passiert, wenn wir Fragebogenitems beantworten? Die Antwort auf diese Frage findet sich in einigen Modellen (siehe *Abbildung 3.6*), deren empirische Grundlage mehr oder weniger gesichert ist. Beispielhaft sei hier ein Modell von Krosnick (1999) aufgeführt, der die Vorstellungen verschiedener anderer Forscher zusammenfasste. In diesem Modell setzt sich der Antwortprozess aus vier Stufen zusammen. In der ersten Stufe, **Verstehen**, geht es darum, das Item zu lesen, den Inhalt zu verstehen und eine mentale Repräsentation des Items bzw. seiner Bedeutung zu generieren. Im nächsten Schritt werden nun Informationen aus dem Gedächtnis abgerufen, die in Zusammenhang mit dem Iteminhalt stehen. Diese Phase nennt Krosnick **Retrieval**. Allerdings bleibt es unklar, worin genau diese Informationen bestehen. Denkbar ist, dass Menschen sich an vergangenes Verhalten erinnern oder vielleicht auch daran, was andere von ihnen halten. Aus dem Abgleich mit dem Iteminhalt und der Information bildet sich dann ein **Urteil** (Stufe drei). Im letzten Schritt erfolgt dann laut Krosnick die **Wahl der Antwortkategorie**, die nach Meinung der Person am besten passt. Offensichtlich ist dieser Prozess weitaus komplexer, als man zunächst meinen kann. Das Ergebnis ist eine optimale Antwort, und Krosnick spricht daher auch von **Optimizing**. Ob es immer zum Ablauf dieses Prozesses kommt, ist fraglich und unter anderem abhängig von der kognitiven Beanspruchung, dem Wunsch, sich selbst auszudrücken, interpersonalen Reaktionen, der intellektuellen Herausforderung, dem Self-understanding, Altruismus, emotionaler Katharsis und weiteren Punkten (Warwick & Lininger, 1975). Es existieren jedenfalls eine Menge Gründe anzunehmen, dass Personen nicht immer Optimizing betreiben. Hier geht Krosnick davon aus, dass im schlimmsten Fall die beiden mittleren Stufen ausgelassen werden und die Person direkt nach dem Lesen eine Antwort wählt. Dies nennt er starkes **Satisfizing**. Weniger starke Formen des Satisfizing (weak Satisfizing) zeichnen sich dadurch aus, dass zwar eine Retrieval- und eine Urteilsphase stattfinden, diese aber stark verkürzt sind. Wie bereits angedeutet, sind diese Antwortmodelle meist empirisch wenig untersucht. Ein Grund hierfür ist sicher, dass ein quantitativer Zugang hier eher unbrauchbar erscheint. Dennoch liefern diese Modelle eine gute Heuristik des Antwortprozesses auf ein Item.

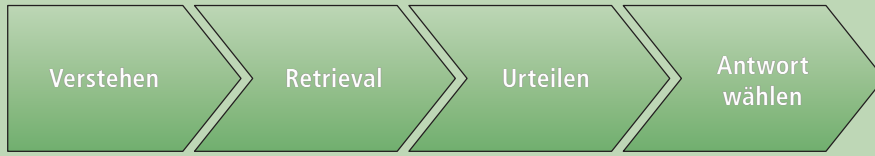


Abbildung 3.6: Ablaufschema einer Itembeantwortung.

3.6.1 Gebundene Aufgabenbeantwortung

Eignet sich ein gebundenes Antwortformat für meinen Test? Wenn ja, welches?

Bei der gebundenen Beantwortung werden festgelegte Antwortkategorien vorgegeben. Es gibt keinen Freiraum für eigene Antworten. Im Anschluss werden folgende Arten gebundener Aufgabenformate beschrieben sowie deren Vor- und Nachteile erörtert:

- Ratingskala
- Richtig-Falsch-Aufgabe
- Zuordnungsaufgabe
- Umordnungsaufgabe
- Ungewöhnliche Itemformate

Ratingskala

Ratingskalen weisen verschiedene Benennungen ihrer Kategorien bzw. Abstufungen auf, beispielsweise von *trifft zu* bis *trifft nicht zu* oder von *sehr gut* bis *sehr schlecht*. Sie ermöglichen eine quantitative Beurteilung der Eigenschaftsausprägung einer Person. Bei der Konstruktion einer Ratingskala sind folgende Punkte zu beachten:

- Differenzierungsgrad des Items
- Polarität des Items (unipolar vs. bipolar)
- Benennung der Antwortalternativen
- Adjustierung der Itemschwierigkeit

Diese Punkte sollen im Folgenden näher erläutert werden.

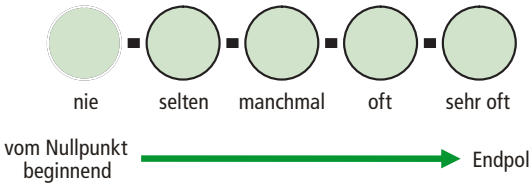
Differenzierungsgrad Vor der Konstruktion von Ratingskalen muss festgelegt werden, wie differenziert die Antwortkategorien abgestuft werden sollen, beispielsweise dreistufig, vierstufig, fünfstufig, sechsstufig oder siebenstufig. Dies hängt unter anderem davon ab, wie genau bzw. differenziert die Probanden die entsprechende Frage beantworten können und wie genau der Testkonstrukteur zwischen den Probanden unterscheiden will. Um dies zu verdeutlichen, nehmen wir an, der Testkonstrukteur stellt zehn Fragen mit einem dichotomen Antwortformat, bei dem die Personen entweder mit *Ja* oder *Nein* antworten. Die Antwort *Ja* wird mit eins codiert und die Antwort *Nein* mit null. Es ergibt sich ein Wertebereich für den Summenwert von 0 bis 10. Werden diese zehn Fragen mit einem fünfstufigen Antwortformat vorgegeben, wobei *starke Ablehnung* mit 0, *Ablehnung* mit 1, *neutral* mit 2, *Zustimmung* mit 3 und

starke Zustimmung mit 4 codiert wird, ergibt sich ein Wertebereich von 0 bis 40 Punkten. Das heißt mit einem differenzierten Antwortformat ergeben sich mehr Möglichkeiten, zwischen Personen Unterscheidungen zu treffen.

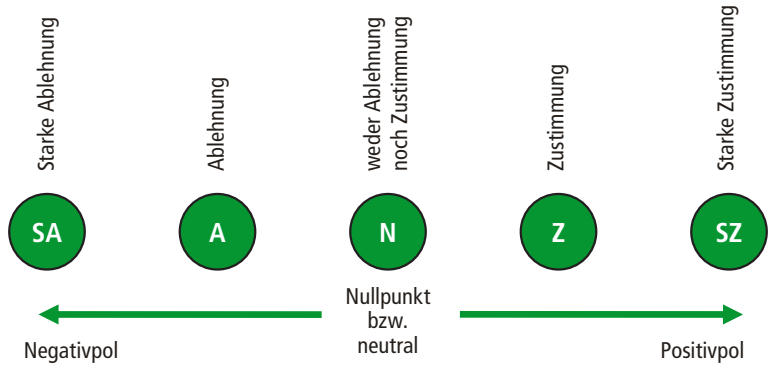
Validität und **Reliabilität steigen**, wenn **mehr Antwortkategorien benutzt werden** (Bortz & Döring, 2002, S. 179; Matell & Jacoby, 1971). Preston und Colman (2000) berichten, dass eine siebenstufige Antwortskala sowohl unter Reliabilitäts- als auch Validitätsgesichtspunkten besonders vorteilhaft sei. **Reliabilitäts- und Validitätssteigerungen sind ab sieben Antwortkategorien gering** (Faulbaum, Prüfer & Rexroth, 2009). Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol und Coffman (2009) konnten zeigen, dass mit zunehmender Anzahl von Antwortkategorien die Reliabilität zwar zunimmt, jedoch gleichzeitig die Modellpassung (beispielsweise im Rahmen konfirmatorischer Faktorenanalysen oder von Rasch-Modellen), insbesondere bei längeren Tests, abnimmt. Zu viele Antwortkategorien können sich auch negativ auf Messeigenschaften eines Items auswirken. Damit ist insbesondere dann zu rechnen, wenn die Probanden mit dem Differenzierungsgrad des Antwortformats überfordert sind. Liegen die verbalen Abstufungen der Skala zu nah beieinander bzw. sind die verbalen Abstufungen nicht eindeutig, kann dies zu Problemen führen: beispielsweise *gelegentlich* gefolgt von *manchmal*. Für das eben genannte Beispiel ist nicht klar, welcher Begriff die höhere Intensität bedeutet.

Ein anderes Problem bei der Wahl von mehr als zwei Antwortkategorien stellen Antwortstile dar. Rost, Carstensen und Davier (1999) berichten von **Mittel- und Extremkreuzern** beim Ausfüllen eines Persönlichkeitsfragebogens. Dieser Befund ist von besonderer Bedeutung und wiegt besonders schwer bei der Wahl des Antwortformats. Kemper (2009) hat herausgefunden, dass dieser Antwortstil bei vielen Fragebögen auftritt und relativ stabil ist. Das bedeutet auch, dass ein Punktwert von beispielsweise 20 Punkten im NEO-FFI für einen Mittel- und einen Extremkreuzer eine völlig unterschiedliche Eigenschaftsausprägung bedeuten kann. Der einzige Weg, damit umzugehen, besteht im Moment darin, ein dichotomes Antwortformat (z.B. *Ja*/*Nein*) zu verwenden.

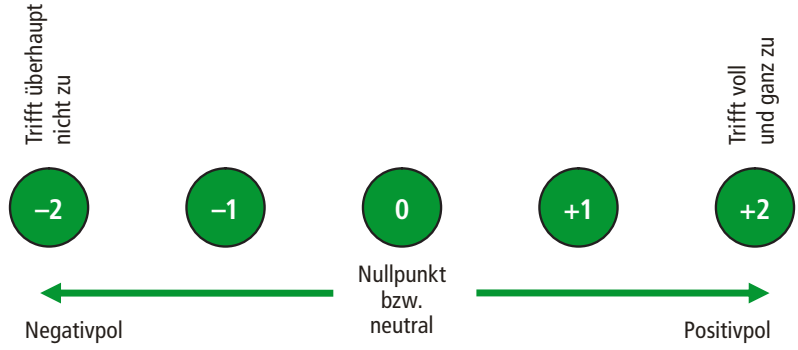
Polarität Zum anderen muss geklärt werden, ob die Items unipolar oder bipolar (Extremausprägungen sind durch gegensätzliche Begriffe gekennzeichnet) vorgegeben werden sollen. Eine **uniploare Skala** geht nach Rost (2004, S. 65) von einem Nullpunkt aus in nur eine Richtung. Betrachten wir dazu die folgende Frage „Wie oft benutzen Sie das Flugzeug für Geschäftsreisen?“ und das dazugehörige Antwortformat.



Davon abzugrenzen ist eine **bipolare verbale Antwortskala** aus dem Persönlichkeitstest NEO-FFI (Borkenau & Ostendorf, 1993). Betrachten wir aus diesem Persönlichkeitstest als Beispiel die Feststellung: „Ich fühle mich oft angespannt und nervös.“ Für alle Fragen dieses Tests geht die Antwortskala von einem Nullpunkt in zwei entgegengesetzte Richtungen zu zwei entgegengesetzten Polen.



Eine solche Skala kann auch mit Zahlen beschriftet sein:



In manchen Fällen sind dabei auch nur die Endpole mit verbalen Ankern benannt.

Benennung der Antwortalternativen Es muss auch entschieden werden, ob und welche Antwortkategorien verbal umschrieben werden bzw. ob darüber hinaus auch noch Zahlen oder weitere visuelle Hilfsmittel verwendet werden. Krosnick (1999) berichtet, dass eine Benennung jeder Stufe bei einer Ratingskala zu Verbesserungen der Messgenauigkeit und der Validität führt. Diese Technik machen sich auch verhaltensverankerte Skalen zu eigen. Solche Skalen enthalten konkrete Verhaltensbeispiele für jede oder die meisten Antwortkategorien (vgl. Marcus & Schuler, 2001, S. 410 f.; Bortz & Döring, 2002, S. 177). Die Auswahl an solchen Ratingskalen ist vielfältig, so können numerische und verbale Abstufungen gemeinsam verwendet bzw. auch mit Prozentzahlen ergänzt werden. In manchen Fällen werden auch nur die Pole der numerischen Abstufungen verbal benannt.

Es gibt auch **bipolare Antwortskalen**, die **Symbole** verwenden. Man findet dies häufig bei Beurteilungen zur Qualitätssicherung, beispielsweise auf die Frage: „Wie gut fanden Sie den Vortrag?“

Auch rein numerische Notenskalen können zur Beurteilung der Qualität herangezogen werden. Es besteht die Hoffnung, dass Noten allen Beurteilern noch bekannt sind. Bei-

spielsweise könnte folgende Frage mithilfe einer Notenskala bewertet werden: „Beurteilen Sie bitte das Essen im Restaurant in Schulnoten.“

Im Rahmen von Internetbefragungen werden auch visuelle Analogskalen vorgegeben. Hierbei kann die Person einen Schieberegler auf einer Linie bewegen. Eine Frage könnte z.B. lauten: „Wie stark stimmen Sie der folgenden Aussage zu? Politiker kümmern sich um die Bedürfnisse von Studenten.“

Rohrmann (1978) untersuchte, bei welchen Antwortformaten der Bedeutungsgehalt der verwendeten sprachlichen Skalenbeschriftung möglichst optimal mit den verwendeten Zahlencodierungen korrespondiert. Er untersuchte dabei vier Urteilsdimensionen: Häufigkeit, Intensität, Wahrscheinlichkeit und Bewertung. Er schlägt für diese Bereiche die im Kasten aufgeführten Skalen vor.

Beispiel 3.7

Vorschläge für Ratingskalen nach Rohrmann (1978)

Häufigkeitsskalen:

❍ - - - - - - ❍ - - - - - - ❍ - - - - - - ❍ - - - - - - ❍

nie selten gelegentlich oft immer

❍ - - - - - - ❍ - - - - - - ❍ - - - - - - ❍ - - - - - - ❍

sehr selten selten gelegentlich oft sehr oft

Intensität:

❍ - - - - - - ❍ - - - - - - ❍ - - - - - - ❍ - - - - - - ❍

gar nicht wenig mittelmäßig überwiegend völlig

Wahrscheinlichkeit:

❍ - - - - - - - - ❍ - - - - - - - - ❍ - - - - - - - - ❍ - - - - - - - - ❍

keinesfalls wahrscheinlich nicht vielleicht ziemlich wahrscheinlich ganz sicher

Bewertung:

❍ - - - - - - - - ❍ - - - - - - - - ❍ - - - - - - - - ❍ - - - - - - - - ❍

trifft gar nicht zu trifft wenig zu trifft teils-teils zu trifft ziemlich zu trifft völlig zu

Zu seiner Untersuchung merkt Rohrmann (1978) selbstkritisch an, dass sie auf einer kleinen und nicht repräsentativen Stichprobe beruht. Darüber hinaus ist die Studie schon vor 28 Jahren publiziert worden. Aus diesem Grund können die vorgeschlagenen Skalen nur Anhaltspunkte für die Benennung der Antwortstufen bieten.

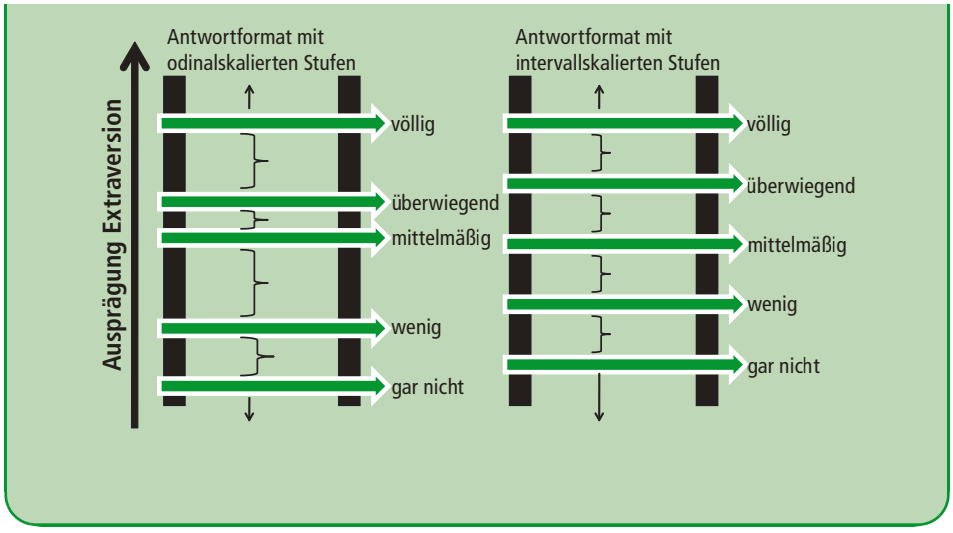
Exkurs 3.5 Skalenniveau von Ratingskalen

Man kann jedoch in den meisten Fällen davon ausgehen, dass das Antwortformat **Ratingskalen nicht intervallskaliert**, sondern ordinalskaliert sind. Betrachten wir dies an einem konkreten Beispiel:
Ich bin ein gut gelaunter Mensch.

Nehmen wir an, das Item soll Extraversion messen. Nun kann man dieses Item im Konstrukt Extraversion verorten. Zur Veranschaulichung nehmen wir eine Hochsprunganlage als Vorbild. Die Grundannahme ist, dass mit steigender Extraversion dem Item mehr zugestimmt wird, das heißt die Intensität der Extraversion zunimmt. Betrachtet man Extraversion als Eigenschaft, kann man sich fragen, ab welchem Grad an Extraversion man in der nächsthöheren Antwortkategorie antwortet. Das heißt jede Person besitzt eine gewisse Ausprägung auf der Eigenschaft Extraversion und kreuzt entsprechend eine Antwortkategorie an. Welche Antwortkategorie eine Person wählt, hängt also von der Ausprägung auf dem Konstrukt Extraversion ab. Dies ist in *Abbildung 3.7* dargestellt. Intervallskaliert sind die Antwortkategorien der Items dann, wenn die Eigenschafts- oder Fähigkeitsbereiche für alle Antwortkategorien, die gewählt werden können, gleich breit ausfallen (siehe rechte Abbildung). Nehmen wir an, wir wollen Extraversion messen und die Eigenschaftsausprägung von Extraversion läge auf einer zunächst fiktiven E-Skala zwischen minus und plus unendlich. In diesem Fall müsste der Bereich der Eigenschaftsausprägung, um *wenig* anzukreuzen, zwischen der „Latte“ *wenig* zu *gar nicht* derselbe sein wie zwischen *mittelmäßig* und *wenig*, *überwiegend* und *mittelmäßig* sowie *völlig* und *überwiegend* (vgl. *Abbildung 3.7*), beispielswiese immer jeweils zwei Einheiten:

gar nicht und *wenig*“	$-\infty$ bis -2.0
mittelmäßig und *wenig*	ab -2.0 bis 0.0
überwiegend und *mäßig*	ab 0.0 bis 2.0
völlig und *überwiegend*	ab 2.0 bis $+\infty$

Wie man sieht, ist dies in der linken Abbildung nicht der Fall, daher liegt auch kein Intervallskalenniveau der Antwortkategorien vor.



Adjustierung der Itemschwierigkeit Man kann die Schwierigkeit eines Items als Grad der Zustimmung zur Itemaussage in Schlüsselrichtung des zu erfassenden Merkmals definieren. Betrachten wir das Item „Wie häufig hat Sie in den letzten vier Wochen das Gefühl der Wertlosigkeit verfolgt?“. In einer gesunden Population werden diesem Item wenige Personen zustimmen.

Man kann nun auch versuchen, die Zustimmung zu verringern, indem man das Item wie folgt umformuliert: „Wie häufig hatten Sie in den letzten vier Wochen das Gefühl, extrem wertlos zu sein?“ Man kann die Zustimmung aber auch erhöhen, indem man fragt: „Wie häufig hat Sie in den letzten vier Wochen das Gefühl geplagt, dass Sie weniger wert sind als andere?“

Vorteile von Ratingskalen Mit Ratingskalen erhält man sehr differenzierte Informationen über die Ausprägung eines Merkmals. Die Durchführung und die Auswertung sind ökonomisch. Man kann die Differenziertheit der Fragen dem Untersuchungszweck und der Differenzierungsfähigkeit der Probanden angleichen.

Nachteile von Ratingskalen Eventuell werden die **Abstufungen subjektiv unterschiedlich aufgefasst**, z.B. was versteht man unter *sehr häufig* oder unter der Aussage *trifft vollkommen zu*? Zu dieser Problemstellung kann ein einfaches Experiment durchgeführt werden. Dazu geben Sie vier oder fünf Personen folgende Frage vor: Wie oft gehen Sie im Monat aus? Als Antwortformat geben Sie *nie – selten – manchmal – oft – sehr oft* vor. Jede Person soll nun eine Kategorie ankreuzen und unter die Kategorie die konkrete Zahl schreiben, wie häufig sie im Monat ausgeht. Personen, die eine Kategorie wählen, z.B. *oft*, werden wahrscheinlich sehr unterschiedliche Häufigkeitsangaben machen. So kann für eine Person zehnmal im Monat auszugehen *oft* bedeuten,

für eine andere Person ist 20-mal im Monat auszugehen *oft*. Problematisch wird es, wenn für andere Personen zehnmal im Monat auszugehen „selten" bedeutet. Anhand dieses Beispiels wird schnell klar, dass die Wahl einer höheren Antwortkategorie nicht zwangsweise mit einer höheren Verhaltensausprägung einhergeht. Damit ist der ganze Messvorgang infrage gestellt.

Darüber hinaus können **Antworttendenzen** auftreten, z.B. Neigung zu extremen Antworten oder die Tendenz zu mittleren Urteilen. Wie Antworttendenzen erkannt werden können, wird in *Abschnitt 8.3.4* dargestellt. Rost (2004, S. 67) weist darauf hin, dass sich in vielen, aber nicht allen Untersuchungen die neutrale Antwortkategorie als empirisch problematisch erwiesen hat (vgl. Carstensen & Rost & von Davier, 1999, sowie Bortz & Döring, 2006, S.180). So kann die Benutzung der mittleren Antwortkategorie neben einer „mittleren" Ausprägung darauf hindeuten, dass ein Item von der Person für unpassend gehalten wird oder die Person die Beantwortung verweigert. Andererseits kann es problematisch sein, den Probanden eine mittlere oder neutrale Kategorie vorzuenthalten, da sie so zu einer Entscheidung gezwungen werden. Eine weitere alternative Möglichkeit besteht darin, dass eine „Weiß nicht"- oder „Kann ich nicht beantworten"-Kategorie als zusätzliche Antwortmöglichkeit vorgegeben wird. Antworten in diesen Kategorien müssen jedoch als fehlende Werte gewertet werden. Diese können bei der nachfolgenden statistischen Auswertung zu Problemen führen. Aus diesem Grund ist die Anzahl der Antwortkategorien genau abzuwägen. Entscheidende Fragen sind dabei:

- **Soll die Reliabilität einer Skala maximiert werden?** → Fünf- bis siebenstufiges Antwortformat nutzen. Nachteil: unterschiedliche Verwendung des Antwortformats, unterschiedliche Nutzung der Mittelkategorie, höhere Wahrscheinlichkeit einer Modellablehnung im Rahmen von konfirmatorischen Faktorenanalysen oder im Rahmen von Rasch-Modellen.
- **Soll verhindert werden, dass die Mittelkategorie unterschiedlich aufgefasst wird?** → Mittelkategorie streichen und eine sechsstufige Skala ohne Mittelkategorie verwenden. Nachteil: Personen mit wirklich neutraler Eigenschaftsausprägung können nicht korrekt antworten, was jedoch nur einen Teil der Personen in der Stichprobe betreffen wird.
- **Soll ein unterschiedlicher Gebrauch der Antwortskala durch die Personen (beispielsweise Mittel- und Extremkreuzer) verhindert werden?** → Dichotomes Antwortformat verwenden. Nachteil: geringerer Differenzierungsgrad der Items.
- **Wie wird in der Regel das Bearbeiten einer Ratingskala erleichtert?** → Eine sprachliche Verankerung der Antwortstufen sowie Zahlen zur Visualisierung verwenden. Die verbalen Verankerungen sollten dabei möglichst gleichabständig und eindeutig gewählt werden (vgl. Rohrmann, 1978).

Richtig-Falsch-Aufgabe

Richtig-Falsch-Aufgaben bestehen aus nur zwei Antwortmöglichkeiten. Sie kommen als Leistungstestaufgaben (Richtig-Falsch-Aufgaben) oder auch Ja-Nein-Fragen (z.B. *Trifft zu/Trifft nicht zu*) in Persönlichkeitstests vor. Der Antwortmodus kann sehr unterschiedlich sein. Die Bandbreite reicht von Ankreuzen über Durchstreichen bis dahin, ein Item mit einem Haken zu versehen.

Beispiel 3.8 **Richtig-Falsch-Aufgaben**

Persönlichkeitsfragebogen. Beispiel eines Ja-Nein-Fragebogenitems aus dem FPI-R (Freiburger Persönlichkeitsinventar, revidierte Form, Fahrenberg, Hampel & Selg, 2001), einem Persönlichkeitstest (Zutreffendes wird angekreuzt):

Ich gehe abends gerne aus. ❑ stimmt ❑ stimmt nicht

Leistungstest. Beispiel eines Items aus dem Revisions-Test zur Erfassung der Konzentrationsfähigkeit von Marschner (1972) (Additionsaufgaben: Falsche Ergebnisse werden durchgestrichen und richtige Ergebnisse mit einem Haken markiert):

2 5
7 2
~~6~~ 7 ✓

Vorteile von Richtig-Falsch-Aufgaben Die Bearbeitungs-, Auswertungs- und Lösungszeiten sind meist kurz. Für die Probanden ist die Testinstruktion in der Regel leicht zu verstehen und die Items können von den Probanden schnell und auch relativ leicht beantwortet werden.

Nachteile von Richtig-Falsch-Aufgaben Ja-Nein-Items müssen so formuliert werden, dass sie eindeutig beantwortet werden können. Im Gegensatz zum Ratingformat ist ein hoher Prozentsatz an Zufallslösungen möglich (50 Prozent), was insbesondere bei Leistungstests ein Problem darstellt (hier lässt sich aber durch das Vergeben von Minuspunkten bei falschen Antworten entgegenwirken). Man erhält zudem wenig differenzierte Informationen. Allerdings kann dieser Nachteil dadurch ausgeglichen werden, dass mehr Fragen gestellt werden, deren Schwierigkeiten sich unterscheiden. Manche Autoren raten von einem Ja-Nein-Antwortformat ab. Diese Ablehnung ist zum Teil begründet durch Schwierigkeiten bei der statistischen Analyse (siehe *Kapitel 5.1* „Schwierigkeitsanalyse“, *Kapitel 5.2* „Trennschärfenanalyse“ bzw. *Kapitel 6* „Faktorenanalyse“). Dichotome Items sind meist nur mit Rasch-Modellen sinnvoll auf ihre Messeigenschaften zu prüfen. Darüber hinaus gibt es Hinweise dafür, dass bei Ja-Nein-Items eine erhöhte Ja-sage-Tendenz zu beobachten ist (Krosnick, 1999, S. 552). Es gibt jedoch auch Beispiele für eine gelungene Fragebogenkonstruktion mit Ja-Nein-Items, wie das FPI-R (Fahrenberg, Hampel & Selg, 2001) beweist.

Einfach- und Mehrfach-Wahlaufgabe

Mehrfach-Wahlaufgaben haben mehr als zwei Antwortalternativen, wovon entweder nur eine Antwort richtig ist (Single-Choice-Aufgabe) oder mehr als eine bis zu allen Antworten einer Aufgabe richtig sein können (Multiple-Choice-Aufgabe). Auch hier existieren verschiedene Formen.

Beispiel 3.9 Mehrfachwahlaufgabe

Leistungstest. Single-Choice-**Beispielitem aus dem IST-2000 R** (Intelligenz-Struktur-Test, revidierte Form A, Liepmann & Beauducel, Brocke & Amthauer 2007), einem Intelligenztest:

Beispiel:

Wald : Bäume = Wiese : ?
a) Gräser b) Heu c) Futter d) Grün e) Weide

„Gräser" ist offensichtlich richtig. Deshalb ist auf Ihrem Antwortbogen unter Aufgabengruppe 02 in der Beispiel-Zeile das a) markiert.

Ein weiteres Beispiel:

dunkel : hell = nass : ?
a) Regen b) Tag c) feucht d) Wind e) trocken

Da dunkel das Gegenteil von „hell" ist, muss zu „nass" auch das Gegenteil gefunden werden. Also ist e) trocken die richtige Lösung.

Beispiel einer **Leistungstestaufgabe aus dem LPS** (Leistungsprüfsystem, Horn, 1983), einem Intelligenztest (bei dieser Aufgabe muss der Proband Fehler in der Rechtschreibung korrigieren):

Afriga → richtige Antwort: Afrika

Persönlichkeitstests → Klinischer Fragebogen. Single-Choice-**Beispielitem aus dem BDI**, einem Depressionsfragebogen (Beck-Depressions-Inventar, Hautzinger, Bailer, Worall & Keller, 1994) (zutreffende Aussage markieren):

Aufgabengruppe A

- ❒ Ich bin nicht traurig.
- ❒ Ich bin traurig.
- ❒ Ich bin die ganze Zeit traurig und komme nicht davon los.
- ❒ Ich bin so traurig oder unglücklich, dass ich es kaum noch ertrage.

Wissenstest. Multiple-Choice-**Beispielitem aus einer Klausur zur Testkonstruktion**. In dieser Klausur musste Zutreffendes angekreuzt werden:

In einem Test kann man verschiedene Aufgabenarten verwenden. Welche der folgenden Aufgabenarten haben ein gebundenes Antwortformat?

- ❒ Multiple-Choice-Items
- ❒ Ergänzungsaufgabe
- ❒ Ratingskala
- ❒ Zuordnungsaufgabe
- ❒ Kurzaufsatzaufgabe

Vorteile von Mehrfach-Wahlaufgaben Durchführung und Auswertung sind ökonomisch. Eine zufällige Beantwortung der Items durch den Probanden ist umso weniger problematisch, je mehr Antwortalternativen zur Verfügung stehen (siehe Exkurs), und wenn darüber hinaus Kombinationen aus mehreren Antwortalternativen die Richtigantwort bilden. Dadurch wirkt sich Raten weniger auf das Testergebnis aus.

Nachteile von Mehrfach-Wahlaufgaben Antwortalternativen zu finden ist eventuell schwierig (alle „falschen" Antwortalternativen sollten gleich wahrscheinlich gewählt werden). Das Antwortformat sollte ausbalanciert sein: Richtige Antworten sollen auf die Itempositionen gleich verteilt sein, also sollte das richtige Item einmal an Position „eins", „zwei" und an jeder anderen Position stehen. Problematisch an diesem Aufgabentyp kann sein, dass nur ein Wiedererkennen von Material oder Wissen verlangt wird, keine Reproduktion. Dies ist nicht für alle Konstrukte sinnvoll (z.B. „Kreativität"). Außerdem können die Antworten schon Hinweise auf die richtige Lösung enthalten. Zudem existieren nach Kubinger (1996) bei Multiple-Choice-Aufgaben qualitativ unterschiedliche Lösungsstrategien. Dies wirkt sich auf die Dimensionalität der Skala oder des Tests aus. Das heißt der Test ist nicht mehr eindimensional und misst somit mehr als nur eine Eigenschaft oder Fähigkeit. Zu beachten ist, dass die Ratewahrscheinlichkeit bei Mehrfachwahlaufgaben nicht nur von der Anzahl der Antwortmöglichkeiten abhängt, sondern auch von der Qualität der Distraktoren.

Definition: Distraktoren
Unter **Distraktoren** versteht man die falschen Antwortmöglichkeiten, die eine Person von der richtigen Lösung der Aufgabe gleichsam ablenken sollen.

Exkurs 3.6 Multiple-Choice-Tests zur Erfassung von Wissen

Mit Wissenstests können unterschiedliche Ziele verfolgt werden. In manchen Fällen muss jede Frage eines Wissenstests richtig beantwortet werden, um den Test zu bestehen. Es wird mit einem solchen Test **keine Differenzierung in Leistungsbereiche** angestrebt. Werden beispielsweise Schaltelemente im Cockpit eines Airbus A 380 auf einem Blatt dargestellt, muss der Pilot alle richtig benennen und deren Funktionen kennen. Es reicht für einen Piloten wohl nicht aus, nur 70 Prozent der Schaltelemente zu kennen. Auf der anderen Seite kann mit einem Wissenstest auch eine **Differenzierung in Leistungsbereiche gewünscht** werden. Das heißt es wird eine Bestehensgrenze definiert und innerhalb des Bereichs, in dem der Test als bestanden gilt, wird eine Differenzierung in Leistungsbereiche gewünscht.

Homogene und heterogene Wissenstests Weiterhin unterscheidet man heterogene und homogene Wissenstests. Heterogene Wissenstests enthalten Items aus verschiedenen Wissensbereichen, die nicht zwangsweise korreliert sein müssen (formative Indikatoren). Homogene Wissenstests erfassen homogene und im besten Fall eindimensionale Wissensbereiche. Das heißt es wird versucht, die Korrelationen zwischen den Items mithilfe von latenten Variablen zu erklären (reflektive Indikatoren).

Anzahl der Antwortalternativen Die Anzahl der Antwortalternativen bestimmt mit, wie hoch die Ratewahrscheinlichkeit bei einer Aufgabe ist. Raten wird mit zunehmender Anzahl der Antwortalternativen unwahrscheinlicher sowie wenn mehrere Antwortalternativen richtig sind. Man sollte dabei die Anzahl der richtigen Lösungen pro Aufgabe variieren lassen von keiner richtigen Lösung bis dahin, dass alle Antworten richtig sind. Setzt man Personen bei jeder Aufgabe darüber in Kenntnis, wie viele richtige Antworten unter den Antwortmöglichkeiten enthalten sind, kann es dazu kommen, dass Personen ein Ausschlussverfahren nutzen, um die richtige Lösung zu erlangen, und nicht aus ihrem Wissen heraus die richtige Lösung finden. Im Folgenden werden statistische Ratewahrscheinlichkeiten für verschiedene Multiple-Choice-Fragen dargestellt.

Ratewahrscheinlichkeit bei einer richtigen Antwort:
1 aus 3 = 33 %
1 aus 4 = 25 %
1 aus 5 = 20 %

Ratewahrscheinlichkeit bei x richtigen Antworten:
2 aus 4 = 16.6 %
2 aus 5 = 10 %

Betrachten wir konkret, wie hoch die Ratewahrscheinlichkeit ist, wenn sich unter fünf Antworten (n) zwei richtige Antworten (x) befinden.

$$\frac{1}{\frac{n!}{x! \cdot (n-x)!}} = \frac{1}{\binom{5}{2}} = \frac{1}{\frac{5!}{2! \cdot 3!}} = \frac{1}{\frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 3 \cdot 2 \cdot 1}} = \frac{1}{\frac{20}{2}} = \frac{1}{\frac{10}{1}} = \frac{1}{10} = .10$$

Vergleicht man die Ratewahrscheinlichkeiten des Formats fünf Antworten, von denen eine richtig ist, gegenüber fünf Antworten, von denen zwei richtig sind, erkennt man, dass eine Erhöhung der Anzahl der richtigen Lösungen die Ratewahrscheinlichkeit von 20 auf 10 Prozent senkt!

Formulierung von Multiple-Choice-Items und -Aufgaben Es ist wichtig, auf eine angemessene Auswahl und Formulierung der Distraktoren und der Richtiglösungen zu achten. Sind die Distraktoren zu leicht, steigt die Ratewahrscheinlichkeit. Sind die Distraktoren zu schwer, kann sich die Ratewahrscheinlichkeit wieder erhöhen, wenn selbst Personen mit hohem Wissensstand die richtigen Antworten von den falschen nicht mehr unterscheiden können. Dies ist beispielsweise dann der Fall, wenn Distraktoren mehrdeutig interpretierbar sind. Leistungsstarke Personen wählen dann eventuell den Distraktor statt der richtigen Antwort. Aus diesem Grund ist es günstig, die Aufgaben mithilfe einer Think-aloud-Technik vor dem Einsatz an der Zielgruppe zu überprüfen. Distraktoren und Richtigantworten müssen ständig optimiert werden.

Messung nicht intendierter Fähigkeiten oder Eigenschaften Beim Antwortformat ist eine Beanspruchung des Arbeitsgedächtnisses oder der Intelligenz zu vermeiden. Häufig werden Aufgaben dadurch erschwert, dass einfache oder doppelte Verneinungen verwendet werden, z.B.: „Welche Kriterien sind **keine** Hauptgütekriterien von psychologischen Tests?“ Dies kann dazu führen, dass mit dem Test nicht das eigentliche Messziel erfasst wird, sondern kognitive Flexibilität oder gar Intelligenz. Häufig findet man auch das folgende Antwortformat:

Welche Kriterien sind Hauptgütekriterien von psychologischen Tests?

(1) Validität

(2) Reliabilität

(3) Normierung

(4) Objektivität

(5) Ökonometrie

❒ 1, 2, 3 sind richtig

❒ 2, 3, 5 sind richtig

❒ 1, 2, 4 sind richtig

❒ keines ist richtig

❒ alle sind richtig

Bei der Anwendung dieses Antwortformats wird wahrscheinlich neben Wissen auch noch die Merkfähigkeit der Personen gemessen. Es bedarf einer gewissen Arbeitsgedächtnisleistung, um die verschiedenen Wörter zu den Richtlösungen zu kombinieren. An dieser Stelle sind auch nur Wörter genannt. Man stelle sich vor, dass dieser Aufgabentyp mit längeren Aussagesätzen verwendet wird (wie es bei medizinischen Tests oft vorkommt). Es kann nicht das Ziel des Testkonstrukteurs sein, andere Fähigkeiten oder Eigenschaften als das angestrebte Wissen noch mitzuerfassen.

Kriterienorientierte Bewertung Ein normorientiertes Vorgehen bei der Auswertung von Wissenstests ist meist unangemessen. Man stelle sich vor, dies würde in der Schule so durchgeführt. Die besten 10 Prozent bekommen die Note eins und man geht dann in 10 Prozentstufen jeweils eine Note nach unten. Nehmen wir nun an, es gibt zwei Jahrgänge. Der eine besitzt deutlich mehr Wissen als der andere Jahrgang. In beiden aufeinanderfolgenden Jahrgängen gibt es aber jeweils immer 10 Prozent der Schüler, die eine sehr gute Note bekommen haben. Ein „sehr gut“ im ersten Jahrgang kann aber eine deutlich schlechtere Leistung als ein „sehr gut“ im zweiten Jahrgang bedeuten.

Möglichkeiten der Bewertung von Multiple-Choice-Aufgaben Mit den zu erwartenden Rateeffekten bei Multiple-Choice-Aufgaben kann auf zwei Arten umgegangen werden. Dazu sollen im Folgenden zwei Bewertungssysteme vorgestellt werden, die Rateeffekte mitberücksichtigen. Sie sollen anhand der folgenden Aufgabe veranschaulicht werden. Die Personen sollen bei dieser Aufgabe ankreuzen, ob die jeweilige Antwort richtig oder falsch ist. In diesem Beispiel ist die Ratewahrscheinlichkeit für jede Antwortalternative 50 Prozent.

Welche Kriterien sind Hauptgütekriterien von psychologischen Tests?		
R	F	Validität
R	**F**	Reliabilität
R	F	Normierung
R	F	Objektivität
R	F	Ökonometrie

System 1: Ratekorrektur durch Minuspunkte Im ersten System wird durch Minuspunkte für Raten korrigiert. Es hat den Nachteil, dass eine negative Gesamtpunktzahl resultieren kann, was auf die Getesteten einen demotivierenden Einfluss haben oder sogar Angst erzeugen kann.

System 2: Ratekorrektur durch Erhöhung der Bestehensgrenze Es ist auch möglich, für jede richtige Lösung Pluspunkte zu vergeben und dabei keine Minuspunkte abzuziehen. Für jede falsche Lösung werden dabei null Punkte vergeben. Man kann dann einfach, um eine indirekte Ratekorrektur vorzunehmen, die Bestehensgrenze von beispielsweise 50 auf 70 Prozent erhöhen. Dieses System hat den Vorteil, dass keine Angst erzeugt wird.

Betrachten wir die **Auswirkungen beider Systeme auf die Punktzahl** für unser Beispiel. Nehmen wir an, eine Person hätte die fett markierten Antwortmöglichkeiten gewählt:

Welche Kriterien sind Hauptgütekriterien von psychologischen Tests?				
R	F	Validität	System 1: +1 Punkt	System 2: +1 Punkt
R	**F**	Reliabilität	System 1: –1 Punkt	System 2: 0 Punkte
R	F	Normierung	System 1: –1 Punkt	System 2: 0 Punkte
R	F	Objektivität	System 1: +1 Punkt	System 2: +1 Punkt
R	F	Ökonometrie	System 1: +1 Punkt	System 2: +1 Punkt

Wir sehen, dass bei Anwendung von **System 1** insgesamt nur **ein Punkt** auf die Aufgabe vergeben wird, während bei **System 2 drei Punkte** vergeben werden. Diese Diskrepanz zu System 1 kann durch eine höhere Bestehensgrenze aufgefangen werden. In einem sehr lesenswerten Artikel beschreiben Michelsen und Cordes (2005), wie hoch die Bestehensgrenze sein muss, um eine bestimmte Wissensquote zu garantieren. Was würde passieren, wenn wir die Bestehensgrenze bei System 2 nicht auf über 50 Prozent erhöhen würden? Nehmen wir an, wir haben eine Klausur mit 100 Punkten. 50 Prozent der Antworten sind richtig. Eine Person hätte das Lernziel erreicht, wenn sie mindestens 50 Punkte erzielt. Wenn die Person nun einfach entweder alle F-Antworten oder alle R-Antworten angekreuzt, hat sie 50 Prozent der Punkte erzielt. Daher muss bei System 2 die Bestehensgrenze auf über 50 Prozent erhöht werden. Alternativ könnte man nur dann einen Punkt auf die Gesamtaufgabe geben (nicht jede Antwort einzeln bewerten), wenn sie komplett richtig gekreuzt wurde. Nachteilig daran ist, dass richtige Teillösungen dabei nicht gewertet werden.

Zuordnungsaufgaben

Bei Zuordnungsaufgaben werden bestimmte Zeichen oder Inhalte anderen Zeichen oder Inhalten zugeordnet.

Beispiel 3.10 **CFT-1 und Wisconsin-Card-Sorting-Test**

Beispiel aus dem Wisconsin-Card-Sorting-Test (Grant & Berg, 1993, siehe *Abbildung 3.8*). Die Aufgabe im Wisconsin-Card-Sorting-Test besteht darin, dass einer Person vier Karten vorgegeben werden. Dabei soll die Person eine fünfte Karte einer der vier vorgegebenen Karten zuordnen. Als Zuordnungskriterium dient Farbe, Form oder Zahl. Dieses Kriterium kennt die Person jedoch nicht. Wenn die Person eine richtige Zuordnung vornimmt, wird dies vom Testleiter an die Person zurückgemeldet und bleibt für die nächsten zehn Durchgänge gleich. Bei der elften Testung wird vom Testleiter ein anderes Zuordnungskriterium gewählt. Die Person muss demnach ihr vorher erworbenes Konzept ändern.

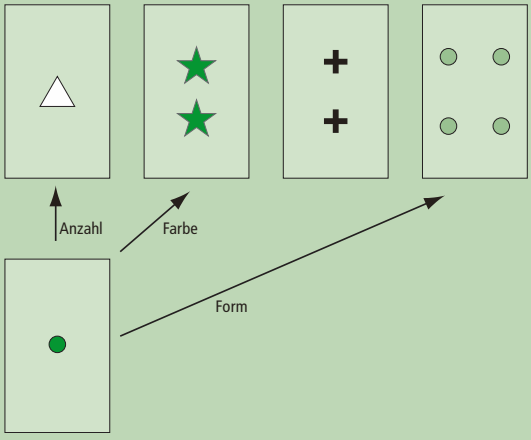


Abbildung 3.8: Karten aus dem Wisconsin-Card-Sorting-Test nach Grant und Berg (1993).

Item aus dem CFT-1-Intelligenztest (Culture-Fair-Intelligence-Test, Weiss & Osterland, 1997): Bei diesem Test müssen den Bildsymbolen (z.B. Bleistift, Uhr) die entsprechenden Zeichen (z.B. I oder O) zugeordnet werden. Der Test besteht aus Zeilen mit Bildsymbolen (siehe *Abbildung 3.9*), unter die dann das entsprechende Zeichen eingetragen werden muss.

Abbildung 3.9: Item aus dem CFT-1-Intelligenztest (Culture-Fair-Intelligence-Test, Weiss & Osterland, 1997).

Auch für Klausurarbeiten ist dieses Antwortformat anwendbar: Eine Aufgabe kann z.B. darin bestehen, verschiedene Formeln jeweils einem Begriff zuzuordnen.

$$(1)\ r_{tt} = 2 \cdot \left(1 - \frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{\hat{\sigma}_x^2}\right)$$

$$(2)\ \alpha = \frac{c}{c-1} \cdot \left(1 - \frac{\sum_{i=1}^{c} \hat{\sigma}_i^2}{\hat{\sigma}_x^2}\right)$$

$$(3)\ r_{tt} = \frac{4 \cdot \hat{\sigma}_1 \cdot \hat{\sigma}_2 \cdot r_{12}}{\hat{\sigma}_x^2 - \left(\frac{\hat{\sigma}_1^2 - \hat{\sigma}_2^2}{\hat{\sigma}_x}\right)^2}$$

Formel von Feldt: _____
Formel Cronbach-α: _____
Formel von Guttman: _____

Vorteile von Zuordnungsaufgaben Durchführung und Auswertung sind ökonomisch. Die zufällige Beantwortung ist bei diesem Aufgabentyp unproblematisch. Dieser Aufgabentyp eignet sich auch zur Überprüfung von Wissen.

Nachteile von Zuordnungsaufgaben Antwortalternativen zu finden ist eventuell schwierig (alle „falschen" Antwortalternativen sollten gleich wahrscheinlich gewählt werden). Statt Reproduktion wird nur Wiedererkennen von Material verlangt, was nicht für alle Konstrukte sinnvoll ist.

Umordnungsaufgabe

Bei Umordnungsaufgaben oder Sortieraufgaben müssen vorgegebene Fragmente nach einer bestimmten Reihenfolge sortiert bzw. umgeordnet werden.

Beispiel 3.11

Der HAWIK-III von Tewes, Rossmann und Schallberger ist im Huber Verlag 2002 erschienen. Bei dem Untertest Bilderordnen sollen die Kinder die dargestellten Kärtchen in eine logische Reihenfolge bringen, so dass die dargestellte Geschichte einen Sinn ergibt.

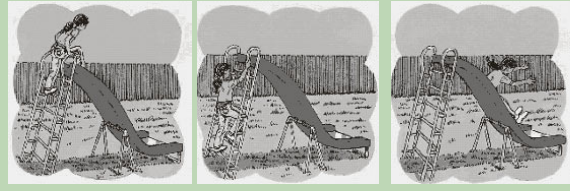


Abbildung 3.10: Untertest Bilderordnen aus dem HAWIK-III nach Tewes, Rossman und Schallberger (2002).

Vorteile von Umordnungsaufgaben Dieser Aufgabentyp kann bei Personen eingesetzt werden, die nicht lesen können (z.B. Kinder unter sechs Jahren).

Nachteile von Umordnungsaufgaben Speziell bei Gruppentestungen muss das Material, falls es sich um Kärtchen oder Ähnliches handelt, in großen Stückzahlen verfügbar sein. Es muss für jeden Probanden vollständig vorhanden sein. Darüber hinaus ist das Itemformat nur für wenige spezifische Fragestellungen anwendbar.

3.6.2 Allgemeine Probleme gebundener Itemformate

Der Nachteil gebundener Itemformate besteht darin, dass sie für manche Konstrukte schwierig anzuwenden sind, z.B. zur Erfassung von Kreativität oder sprachlichen Fertigkeiten. Außerdem können sich absichtliches Verfälschen, Raten und so genannte Antworttendenzen des Probanden auf die Itembeantwortung auswirken. Im Folgenden sollen weitere Probleme gebundener Itemformate näher besprochen werden:

- Verfälschbarkeit
- Antworttendenzen
- Motivation
- Reihenfolgeeffekte und logisch abhängige Items
- Negativ gepolte Items

Verfälschbarkeit

Man unterscheidet zwei grundsätzliche Arten von **Verfälschung**: **Simulation** (fake good) und **Dissimulation** (fake bad). Mit „Simulation“ ist das Vortäuschen von Verhalten oder Symptomen gemeint, welche der Proband normalerweise nicht zeigt. Unter „Dissimulation“ versteht man das Verschleiern oder Verbergen von Symptomen oder Verhalten, welche der Proband normalerweise zeigt. Bortz und Döring (2006, S. 231) verstehen unter Simulation auch den Versuch, besonders **hohe Testwerte** zu erzielen, und unter Dissimulation, durch **„Dummstellen“** besonders niedrige Testwerte zu erzielen. Simulation und Dissimulation betreffen objektive Leistungstests und Fragebogenverfahren.

Dabei ist anzumerken, dass objektive Leistungstests wahrscheinlich nur schwer oder gar nicht nach oben verfälschbar sind, wohl aber nach unten (vgl. Ziegler, Schmidt-Atzert, Bühner & Krumm, 2007). In der Regel können die Personen jedoch nicht einschätzen, wie „dumm“ man sich verhalten soll, um ein schlechtes Ergebnis bei einem Leistungstest zu erzielen. Ihnen unterlaufen dann beispielsweise Fehler oder so geringe Leistungen, wie sie selbst hirngeschädigten Patienten nicht unterlaufen (vgl. Schmidt-Atzert, Bühner, Rischen & Warkentin, 2004).

Fragebögen werden in diesem Zusammenhang oft abgelehnt, weil sie natürlich relativ leicht verfälschbar sind. Daher wird deren Einsatz gerade in der Eignungsdiagnostik im Rahmen von Auswahlentscheidungen kritisch gesehen.

Dass Personen in der Lage sind, Fragebögen zu verfälschen (Ziegler, Schmidt-Atzert, Bühner & Krumm, 2007), und dies auch tun, wenn sie daraus einen Vorteil ziehen können (Griffith, Chmielowski & Yoshita, 2007), ist empirisch mehrfach belegt. Dies alleine verbietet jedoch nicht unmittelbar die Verwendung von Fragebögen. Um hier eine Entscheidung zu treffen ist es wichtig, sich die Auswirkungen solcher Verfälschungen auf den Fragebogen näher zu betrachten. Nehmen wir beispielsweise an, die Stelle einer wissenschaftlichen Hilfskraft am Lehrstuhl für Psychologische Diagnostik in Graz soll besetzt werden. Die Anzahl der Bewerber ist so groß, dass nicht mit allen

Bewerbern ein Interview durchgeführt werden kann. Daher sollen in einem ersten Schritt Personen aufgrund geringer Ausprägungen auf der Persönlichkeitseigenschaft Gewissenhaftigkeit ausgeschlossen werden. Zur Auswahl soll das NEO-FFI (Borkenau & Ostendorf, 2. Auflage, 2008), verwendet werden und die Normstichprobe aus dem Handbuch zur Auswertung. Wie wirkt sich nun das Problem der Verfälschung auf diese konkrete Fragestellung aus?

Zunächst ist anzunehmen, dass Personen ihre Antworten nach oben verfälschen, um sich positiver darzustellen. Nach Fowler (1995, S. 28 f.) ist dies insbesondere dann der Fall, wenn man sich besser darstellen möchte als man ist (z.B. Bewerbungssituation), wenn ein Testergebnis eine negative Konsequenz für die Person nach sich ziehen könnte (z.B. Verlust von sozialem Ansehen, Strafverfolgung, Verlust des Arbeitsplatzes) oder eine Person ein positives Bild von sich aufrechterhalten will (z.B. aus Gründen des Selbstschutzes oder aufgrund von „Wunschdenken"). Dies hat zur Folge, dass der Mittelwert der Bewerbergruppe gegenüber einer Gruppe, in der Personen nicht verfälschen würden, erhöht ist. Metaanalytisch liegt der Effekt zwischen einer halben und einer ganzen Standardabweichung (Birkeland, Manson, Kisamore, Brannick & Smith, 2006; Viswesvaran & Ones, 1999). Demzufolge muss bei der Wahl des Cut-Offs die Situation, in der ein Fragebogen eingesetzt wird, berücksichtigt werden. Würden jedoch alle Personen im gleichen Maße ihre Antworten verzerren, wäre diese Mittelwertsverschiebung unproblematisch. Ist dies jedoch nicht der Fall und unterscheiden sich die Bewerber in ihrem Fakingverhalten, dann kann sich dies auf die psychometrischen Eigenschaften des Fragebogens auswirken.

Um dies genauer darzustellen, ist es notwendig, Faking zu definieren. Ziegler und Bühner (2009) haben Faking als Spurious Measurement Error, welcher einen systematischen Messfehler darstellt, definiert. Er entsteht durch eine Interaktion zwischen Person und Situation. In diesem Sinne verfälschen Personen nur in bestimmten Situationen, z.B. in einer Auswahlsituation und auch nur bei als anforderungsrelevant betrachteten Items. Dabei unterscheiden sie sich jedoch in Abhängigkeit ihrer Persönlichkeit im Ausmaß, in dem sie verfälschen. Beispielsweise tendieren Personen mit hoher Selbstwirksamkeitserwartung eher zum Verfälschen (Ziegler, 2007). Die Definition impliziert auch, dass Verfälschung systematisch das Antwortverhalten beeinflusst. Somit bestimmt nicht nur die zu messen beabsichtigte Persönlichkeitseigenschaft das Antwortverhalten, sondern auch die Fähigkeit oder Eigenschaft zu Verfälschen. Dies wirkt sich auch auf die psychometrischen Eigenschaften des Fragebogens aus. Dies wird im Folgenden näher beschrieben.

Auswirkung von Verfälschung Ergebnisse zur Auswirkung von Faking auf die Reliabilität sind rar. Heggestad, Morrison, Reeve und McCloy (2006) fanden, dass Faking die **interne Konsistenz** eines Fragebogens leicht erhöht. Die Ergebnisse bezüglich der **Konstruktvalidität** sind gemischt. Während bei Laboruntersuchungen aufgrund von Faking die Korrelationen zwischen verfälschten Skalen in einem Fragebogen ansteigen (Pauls & Crost, 2004; Ziegler & Bühner, 2009), ist dies bei Stichproben in tatsächlichen Bewerbungssituationen nicht zu beobachten (Smith & Ellingson, 2002). Es bleibt also vorerst fraglich, welchen Einfluss Faking auf die Konstruktvalidität hat. Die Ergebnisse bezüglich der **Kriteriumsvalidität** legen nahe, dass sich Faking kaum auswirkt. So kann weder eine steigernde noch eine senkende Wirkung nachgewiesen werden (Ones & Viswesvaran, 1998). Zudem lässt sich zeigen, dass die Unterschiede im Antwortverhalten, die aufgrund von Faking auftreten, bei der Vorhersage akademischer Leistungen nicht valide sind (Ziegler & Bühner, 2009). Diese Befunde beziehen sich jedoch meist

ausschließlich auf Fragebogenergebnisse und Persönlichkeitsfaktoren mit einer hohen Abstraktionsebene (Domainelevel). Eine Studie, die unter anderem den Effekt von Faking auf die Validität von Persönlichkeitsfacetten untersuchte, zeigte weniger beruhigende Ergebnisse (Ziegler, Danay, Schölmerich & Bühner, 2010). Zwar blieb die Kriteriumskorrelation für einige Facetten unberührt, für andere sank oder stieg sie jedoch oder drehte sich gar vom Vorzeichen her um. Auf diesem Gebiet ist daher noch viel Forschung nötig. Persönlichkeitsfragebögen werden vor allem in Auswahlsituationen verfälscht. Hier zählt jedoch nicht der Gruppenmittelwert oder die durchschnittliche Vorhersagegüte, sondern vielmehr die Beurteilung einzelner Individuen. Hier zeigt sich, dass Faking die **Rangreihe** der Personen in der Ausprägung des Traits im Vergleich zu ehrlichen Antworten verändert (Holden, 2008). Je weniger Personen selegiert werden, desto größer ist der Anteil von Fakern unter den Ausgewählten.

Insgesamt ist noch viel Forschung nötig, um das Phänomen des Verfälschens von Fragebögen abschließend zu klären. Dennoch wäre es vorschnell, Fragebögen zu verdammen. Der Einsatz von Fragebögen kann vor allem bei hohen Bewerberzahlen in einem frühen Stadium zur Negativauslese herangezogen werden. Hier kann ein vergleichsweise niedriger Cut-Off gewählt werden, so dass die Rangplatzvertauschungen weniger stark ins Gewicht fallen.

Strategien gegen Verfälschung Bortz und Döring (2002, S. 235) berichten, dass die **Aufforderung zu korrekter Testbearbeitung** in der Instruktion zu keinen besseren Ergebnissen führt als die Zusicherung von absoluter Anonymität, in der es für die Untersuchungsteilnehmer keine Veranlassung gibt, den Test zu verfälschen. Auch Andeutungen oder Warnungen, dass nicht korrektes Testverhalten erkannt wird, sind nach Bortz und Döring (2006, S. 235) wenig effektiv. Im Gegensatz dazu berichten Hossiep, Paschen und Mühlhaus (2000, S. 61) von Untersuchungen, in denen eine Warnung vor nicht korrekter Bearbeitung einen reduzierenden Effekt auf die Verfälschung hatte. Auch wenn solche Strategien nicht oder nicht immer helfen, soziale Erwünschtheit oder Antworttendenzen zu eliminieren, heißt dies nicht, dass es nicht notwendig ist, Vertraulichkeit und Datenschutz – wenn möglich – zuzusichern oder auf die Wichtigkeit einer korrekten (ehrlichen) Testbearbeitung hinzuweisen (Fowler, 1995).

Ein Weg, um den Einfluss der Verfälschung auf die Itembeantwortung zu minimieren, besteht in der Verwendung von so genannten **Forced-Choice-Items**. Bei dieser Itemart werden dem Probanden verschiedene Aussagen vorgegeben, und er muss sich für mindestens eine Aussage entscheiden. Dabei ist zu beachten, dass dieses Antwortformat lediglich Aussagen über die relative Ausprägung einer Eigenschaft im Vergleich zu anderen Eigenschaften einer Person zulässt. Nehmen wir als Beispiel an, Matthias füllt einen Berufsinteressenstest aus. Das Ergebnis zeigt, dass Matthias bei neun von zehn Items einen Gärtnerberuf einem Büroberuf vorgezogen hat. Das heißt Matthias präferiert den Gärtnerberuf gegenüber dem Büroberuf. Auch Julia bearbeitet den Fragebogen und wählt jeden Beruf genau fünfmal, bevorzugt also beide Berufe gleichermaßen. Es ist hier allerdings nicht möglich, die absolute Präferenzhöhe zu ermitteln. Somit lässt sich nicht sagen, ob der Wunsch, Gärtner zu werden, bei Matthias größer ist als bei Julia. Vergleiche zwischen Personen sind somit nicht möglich. Es kann z.B. sein, dass der Gärtnerberuf für Matthias zwar interessanter ist als der Büroberuf, aber ihn beide Berufe eigentlich nur wenig interessieren. Für Julia sind beide Berufe gleich attraktiv, jedoch insgesamt vielleicht auch unattraktiv. Dieses Antwortformat ist nicht unumstritten. Beispielsweise werden die vorgefertigten globalen Aussagen auf einzelne Probanden immer mehr oder weniger zutreffen, was bei der Itembeantwortung nicht miterfasst

wird. Die am ehesten zutreffenden Aussagen können bezüglich des Genauigkeitsgrades der Beschreibung einer Person stark variieren und so zu relativ ungenauen Beschreibungen führen. Neue Forschungsergebnisse zeigen allerdings, dass auch Forced-Choice-Items durch bewusste Verfälschung beeinflusst werden können. So kommen Heggestad, Morrison, Reeve und McCloy (2006) nach zwei Studien zum Schluss, dass Forced-Choice-Items zur Kontrolle von Simulation in der Personalauswahl ungeeignet sind.

Beispiel 3.12 **Forced-Choice-Item**

In einem Beispiel werden Ihnen vier Aussagen vorgegeben: A, B, C und D. Ihre Aufgabe besteht darin, jeweils die Aussage auszuwählen, die für Sie am meisten und am wenigsten zutrifft. Markieren Sie die am meisten zutreffende Aussage mit einem „M“ und die am wenigsten zutreffende Aussage mit einem „W“.
Ich bin ein Mensch, der ...
A) gerne Freunde um sich hat. ____
B) gerne Partys organisiert. ____
C) mit Stress gut zurechtkommt. ____
D) Abwechslung sucht. ____

Zur Überprüfung von Effekten der sozialen Erwünschtheit werden häufig spezielle **Fragebögen zur sozialen Erwünschtheit** herangezogen. Pauls und Crost (2004) konnten jedoch zeigen, dass auch diese Fragebögen gegenüber der Verfälschung, die sie messen sollen, selbst anfällig sind. Probanden merken, was gemessen werden soll, und verfälschen den Fragebogen, z.B. in Richtung geringer sozialer Erwünschtheit. Gravierender noch sind die Befunde, die zeigen, dass Fragebögen zur sozialen Erwünschtheit zu einem substanziellen Anteil systematische Unterschiede in den Persönlichkeitsdimensionen Neurotizismus, Verträglichkeit und Gewissenhaftigkeit abbilden (Paulhus, 2002). Ein hoher Wert bedeutet also nicht zwangsweise, dass eine Person sozial erwünscht geantwortet hat. Es könnte auch sein, dass ihre Persönlichkeitskonstellation abgebildet wurde. Das heißt unter anderem, dass die bloße Korrelation zwischen Fragebögen zur sozialen Erwünschtheit und Persönlichkeitsfragebögen nichts über die Verfälschbarkeit des Persönlichkeitsfragebogens aussagt.

Um Testergebnisse besser interpretieren zu können, wurden verschiedene so genannte **Validitätsskalen** entwickelt. Der MMPI-II (Minnesota Multiphasic Personality Inventory, deutsche Bearbeitung von Engel, 2000), ein klinischer Persönlichkeitstest, besitzt sehr viele, aber vor allem drei sehr bekannte Validitätsskalen. Die erste Skala (L-Skala) erfasst sozial erwünschte Antworten. Dazu werden Aussagen vorgegeben, die sozial erwünscht sind, aber selten angekreuzt werden (z.B. „Ich bin immer objektiv“), und Aussagen, die häufig vorkommen, aber sozial unerwünscht sind (z.B. „Manchmal bin ich auch ungerecht“). Wird erstere mit Ja und letztere mit Nein beantwortet, so wird daraus geschlossen, dass der Proband im ganzen Test sozial erwünscht geantwortet hat. Die zweite Skala wird als Validitätsskala (F-Skala) bezeichnet, die sowohl sozial unerwünschte als auch sehr selten vorkommende Aussagen enthält (z.B. „Manchmal rede ich mit Außerirdischen“). Die dritte Skala, die so genannte K-Skala, enthält Items, die häufig vorkommen, aber sozial unerwünscht sind (z.B.

„Zuweilen möchte ich am liebsten etwas kaputt schlagen"). Die Anzahl der Nein-Antworten wird als Indikator für eine Abwehrhaltung gegenüber dem Test aufgefasst. Hohe Ausprägungen auf diesen Validitätsskalen sollen dem Anwender Hinweise darauf geben, ab wann ein Testergebnis mit Vorsicht zu interpretieren ist. Sie liefern jedoch nur Indizien für eine mögliche Verzerrung der Antworten, aber keine Fakten. Sie dürfen keinesfalls inhaltlich interpretiert werden.

Antworttendenzen

Auch Antworttendenzen können Itemkennwerte, wie die psychometrischen Itemsschwierigkeiten und Trennschärfen (Korrelation eines Items mit dem Testwert, vgl. *Kapitel 5.2*), verzerren. Zu den häufigsten Antworttendenzen gehören die Ja-sage- (Zustimmungs-) oder die **Nein-sage-Tendenz** (Ankreuzen von überwiegend mittleren oder extremen Antworten). Die **Ja-sage-Tendenz** wird auch als Akquieszenz bezeichnet. Zur Kontrolle dieser Antworttendenzen wird eine Reihe von Maßnahmen vorgeschlagen, z.B. das Auszählen von Ja-Antworten oder extremen Antwortkategorien. Meist kann damit das jeweilige Antwortverhalten bestimmt werden, jedoch ist die Interpretation kritisch: Das Vorliegen eines bestimmten Antwortverhaltens ist nur ein Indikator für eine Antworttendenz, sichert diese aber nicht ab. Murphy und Davidshofer (2001, S. 224) weisen darauf hin, dass zum Erkennen von Antworttendenzen oftmals die Betrachtung mehrerer Validitätsskalen notwendig ist.

Motivation

Die Genauigkeit der Itembeantwortung hängt auch von der **Motivation** des Probanden ab, den Test zu bearbeiten. Sie kann sich während der Testbearbeitung ändern. Die Motivation, ein einzelnes Item korrekt zu beantworten, kann wiederum auch von den Testeigenschaften abhängen. Je komplexer die Items formuliert sind, je schwerer die Items zu beantworten sind und je länger der Test dauert, desto wahrscheinlicher ist es, dass die Motivation des Probanden abnimmt. Es empfiehlt sich also, die Länge eines Fragebogens auf möglichst **wenige Items** zu begrenzen. Allerdings sollten die späteren Skalen oder Untertests eine ausreichende Messgenauigkeit und Inhaltsvalidität besitzen. Das heißt sie sollten so kurz wie möglich und so lange wie nötig sein.

Reihenfolgeeffekte und logisch abhängige Items

Abhängig von seiner Position kann die Antwort auf ein Testitem zwischen Versuchspersonen variieren. Bei Speed-Tests sind die Items meist gleich schwer und die Reihenfolge der Items hat nur einen geringen Einfluss auf die Itembeantwortung. Bei Intelligenztests hingegen werden die Items oft nach ihrer Schwierigkeit geordnet dargeboten. Diese wird empirisch anhand einer Stichprobe bestimmt. Problematisch ist es, wenn die Itemrangreihe in anderen Stichproben eine andere Itemreihung ergibt. In diesem Fall sind einige Items gemessen an ihrer Position im Test zu schwer. Probanden verlieren hier viel Zeit bei der Itembearbeitung, die ihnen dann vielleicht für die Bearbeitung vergleichsweise leichterer Items, die später im Test folgen, fehlt. Die Fähigkeit dieser Probanden wird dann mit dem Summenwert des Tests unterschätzt. Es sollte zudem bedacht werden, dass eine vorausgegangene Aufgabe keine Hinweise zur Lösung einer nachfolgenden Aufgabe geben sollte. Betrachten wir z.B. folgende Items eines Wissenstests: „Wie viele Zentimeter hat ein Dezimeter?" und „Wie viele Dezimeter ergeben einen Meter?". Wer nicht weiß, was ein Dezimeter ist, wird beide Fragen falsch beantworten. In man-

chen Fällen (z.B. Items eines Persönlichkeitstests), bei denen die Zusammenhänge zwischen den Items nicht so offensichtlich sind (z.B. durch Bewusstmachen eines Problems), können so genannte Pufferitems (Items mit neutralem Inhalt) Reihenfolgeeffekte vermindern (Rost, 1996, S. 75). Bei Fragebögen empfiehlt es sich, eine zufällige Itemreihenfolge zu wählen.

Negativ gepolte Items

Negativ gepolte Items werden herangezogen, um eine Zustimmungs- oder Ja-sage-Tendenz der befragten Personen zu verhindern. Beispielswiese wird die Frage „Ich bin glücklich“ in „Ich bin unglücklich“ umformuliert. Nehmen wir an, mit dieser Frage soll die Neigung, positive Gefühle zu erleben, erfasst werden. Nehmen wir weiter an, es wird ein fünfstufiges Antwortformat verwendet: *starke Ablehnung*, *Ablehnung*, *neutral*, *Zustimmung*, *starke Zustimmung*. Während bei der ersten Frage die Antwort *Zustimmung* die höchste Ausprägung bedeutet, ist dies durch die invertierte Formulierung bei der zweiten Frage *starke Ablehnung*. Viele Lehrbücherautoren (z.B. Rubin & Babbie, 2009) sprechen sich gegen negativ gepolte Items aus (siehe auch *Kapitel 3.7*). Es kann beispielsweise sein, dass Personen die Invertierung des Items überlesen. Sie antworten dann genau entgegengesetzt der von ihnen beabsichtigten Antwort. Es kann besondere Personengruppen geben, die sich mit dem Verstehen eines negativ gepolten Items schwertun. Dies ist insbesondere dann der Fall, wenn weitere Schwierigkeiten in der Itemformulierung vorliegen, beispielsweise „Ich bin nicht oft unglücklich“ und als Antwort „Trifft nicht zu“ bis „Trifft vollkommen zu“. Daher bleibt hier nur das Fazit, dass auf negativ gepolte Items fast ausschließlich verzichtet werden kann.

3.6.3 Die freie Aufgabenbeantwortung

Eignet sich ein freies Antwortformat für meinen Test? Wenn ja, welches?

Für die Aufgabenbeantwortung werden keine festen Kategorien vorgegeben, sondern sie ist frei oder teilstrukturiert. Teilstrukturiert bedeutet, dass Teile der Lösung vorgegeben sind. (Hinweis: Die im Folgenden ohne Quellenangabe dargestellten Itembeispiele wurden nicht aus bestehenden Testverfahren entnommen. Sie sind rein fiktiv.) Im Anschluss werden folgende freie Antwortformate näher erläutert:

- Ergänzungsaufgabe
- Kurzaufsatz

Ergänzungsaufgabe

Im Folgenden werden verschiedene Formen von Ergänzungsaufgaben anhand von Beispielen vorgestellt.

Offene Fragen:

In welchem Land liegt die Stadt Lima? _____
Der absolute Nullpunkt liegt bei? _____
Apfel verhält sich zu Obst wie Weizen zu? _____

Lösungen: Peru, –273 °C, Getreide

Ergänzen Sie bei dem folgenden Text die fehlenden Wörter: Die Reliabilität eines Tests bezeichnet inhaltlich die _____ eines Tests. Dabei wird die Höhe der Reliabilität bestimmt durch die _____ in Kombination mit der _____.

Lösungen: Messgenauigkeit, Anzahl der Items, mittleren Interitemkorrelation
Objektiv sind diese Aufgaben dann, wenn Antwortkategorien vorgegeben sind. Diese sind standardisiert auszuwerten, beispielsweise in dieser Form (vgl. Beispiel Lückentext):

1 **Messgenauigkeit**

2 **Itemanzahl**

3 **mittleren Interitemkorrelation**

Ergänzen Sie bei dem folgenden Text die fehlenden Worthälften:
Die Korrelation einer Var_____ mit einem Krit_____
hängt unter anderem davon ab, ob Aus_____ die Korrelation verfä_____.

Lösungen: iablen, erium, reißer, lschen

Vorteile von Ergänzungsaufgaben Zufallslösungen sind bei dieser Aufgabenform kaum möglich, eventuell ist aber der Lösungsweg erkennbar. Es kann daher eine qualitative Auswertung dieser Aufgaben vorgenommen werden. Inhaltlich besteht die Möglichkeit der Konstruktion komplexer Aufgaben.

Nachteile von Ergänzungsaufgaben Es wird nur eine Reproduktion von Wissen abgefragt, was nicht für alle Konstrukte sinnvoll ist. Eventuell ergibt sich eine Suggestivwirkung der Fragestellung. Ketteneffekte können auftreten. Wird beispielsweise ein Wort nicht erkannt oder gewusst, kann das nächste Wort wahrscheinlich auch nicht gefunden werden. Der Zeitaufwand bei der Bearbeitung ist größer als bei anderen Aufgabentypen, und es kann sich eine eingeschränkte Auswertungsobjektivität ergeben, wenn mehrere Begriffe passen. In einem solchen Fall sollten a priori alle möglichen Begriffe, die als richtig gezählt werden, in einer Musterlösung aufgeführt sein.

Kurzaufsatz

Bei Kurzaufsätzen müssen auf Fragen kurze, freie Antworten niedergeschrieben werden. *Wie kommen Sommer und Winter zustande?* Antworten:

- Die Sonne steht im Winter am tiefsten und im Sommer am höchsten.
- Die Sonne scheint im Sommer lang, im Winter hingegen nur kurz.
- Die nördliche Halbkugel der Erde, auf der wir leben, ist der Sonne im Sommer zugewandt, im Winter abgewandt.
- Die Wetterlage und die klimatischen Bedingungen ändern sich rhythmisch.

Vorteile von Kurzaufsätzen Die freie Reproduktion von Wissen ist möglich. Bei bestimmten Fragestellungen ist diese Methode unerlässlich, z.B. bei der Erfassung von stilistischer Begabung oder der Reproduktion von Wissen beispielsweise durch einen Gedächtnistest. Zufallslösungen sind nicht möglich.

Nachteile von Kurzaufsätzen Eventuell besteht eine eingeschränkte Auswertungsobjektivität, da es sehr schwierig ist, eindeutige Auswertungskriterien (Klassifikation als eindeutig richtig oder eindeutig falsch) festzulegen. Es ist daher sehr aufwendig, Inhaltsanalysen durchzuführen.

3.6.4 Atypische Aufgabenbeantwortung

Welche sonstigen Aufgabenformate gibt es noch?

Bei der atypischen Aufgabenbeantwortung handelt es sich um eine Restkategorie. In dieser sind die Antwortformate aufgeführt, die sich den oben erwähnten Kategorien nicht zuordnen lassen.

Beispiel 3.13 **Ungewöhnliche Antwortformate**

Beispiel aus dem Zahlenverbindungstest, ZVT (Oswald & Roth, 1987), einem speziellen Intelligenztest (Verbinden Sie die Zahlen in aufsteigender Reihenfolge!):

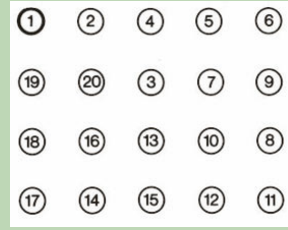


Abbildung 3.11: Beispielaufgaben aus dem Zahlenverbindungstest nach Oswald und Roth (1987).

Beispielitem aus dem I-S-T 2000 R (*Abbildung 3.12*). Dieses Item stammt ursprünglich aus dem Berliner Intelligenz-Struktur-Test (B-I-S, Jäger, Süß & Beauducel, 1997) und sollte Bearbeitungsgeschwindigkeit erfassen. Es war aber empirisch immer stark mit schlussfolgerndem Denken korreliert. Nachdem eine Umgestaltung der Aufgabe, so dass kein schlussfolgerndes Denken mehr miterfasst wird, nicht möglich war, wurde sie im I-S-T 2000 R als Aufgabe zum schlussfolgernden Denken genutzt (persönliche Mitteilung André Beauducel).

Abbildung 3.12: Beispielitem aus dem I-S-T 2000 R Rechenzeichen (Liepmann & Beauducel, Brocke, Amthauer 2007).

Beispiel 3.14 **Konstruktion eines Konzentrationstests**

Man kann verschiedene Arten von Konzentrationstests (vgl. Bühner, 2001) unterscheiden:

- Rechentests (z.B. Revisions-Test von Marschner, 1972, siehe auch 8. Beispiel)
- Durchstreichtests (z.B. Test d2 von Brickenkamp)
- Kartensortierverfahren (z.B. Konzentrationsverlaufstest, K-V-T, von Abels (1965), siehe auch Beispiel 3.10 Wisconsin-Card-Sorting-Test)

An dieser Stelle wird also zunächst entschieden, welche Testart man verwenden möchte. Es wäre auch möglich, eine völlig neue Idee zu entwickeln, um Konzentration zu erfassen. Wir entscheiden uns für die Entwicklung eines Durchstreichtests. Wie wir wissen, ist ein Kartensortierverfahren aufwendig im Materialverbrauch und kann in der Regel nur als Einzeltest durchgeführt werden. Ein Rechentest hat den Nachteil, dass das Einstreuen von Störreizen, um das Definitionsmerkmal der Ablenkung umzusetzen, schwierig ist.

Daher entscheiden wir uns bei der Testart für einen Durchstreichtest. Wir lehnen uns an den Test d2 an und verwenden die Buchstaben n und u und versehen diese mit Punkten. Damit verwenden wir Reizkonfigurationen, wie in der Definition gefordert. Zu suchen ist der Buchstabe u mit zwei Punkten, wobei entweder unter und über dem Buchstaben u jeweils ein Punkt steht oder oben oder unten jeweils zwei Punkte. Mit der Verwendung ähnlicher Distraktoren (Buchstabe n) versuchen wir, eine Situation herzustellen, die das normale Entdecken der Reizkonfigurationen erschwert und Ablenkung erzeugt. So könnten Beispielitems aussehen:

Um eine fortlaufende Reizselektion zu gewährleisten, werden 50 Zeichen in zwölf Zeilen angeordnet, und die Aufgabe der Personen besteht darin, die Zielreizkonfiguration (u mit zwei Punkten) so schnell und genau wie möglich zu entdecken.

3.7 Richtlinien zur Itemformulierung

Worauf muss ich bei der Itemformulierung achten?

Items sollten immer möglichst präzise und leicht verständlich formuliert werden. In diesem Abschnitt soll erläutert werden, auf welche Besonderheiten bei der Formulierung von Items geachtet werden sollte. Dabei unterscheiden wir zwischen dem eigentlichen Itemstamm (Frage) und dem Antwortformat.

- Merkmale der Zielgruppe
- Itempolung

- doppelte Verneinungen
- Registrierung von Verhaltenshäufigkeiten
- weitere Hilfen zur Aufgabenkonstruktion

Merkmale der Zielgruppe Im *Kapitel 3.2* wurde bereits dargestellt, welchen Einfluss bestimmte Eigenschaften der Zielgruppe bei der Itemformulierung haben können. Daher wird an dieser Stelle nicht noch einmal darauf eingegangen.

Itempolung Bei Fragebögen wird üblicherweise immer dasselbe Antwortformat für alle Items genutzt. Oft ist es so, dass höhere Antwortkategorien eher mit Zustimmung zum Iteminhalt einhergehen. Ein Problem, dass bei solchen Items auftreten kann, ist das der Akquieszenz. Hiermit ist gemeint, dass Personen eine Tendenz haben, Items eher zuzustimmen. Um nun zu vermeiden, dass Personen einfach einen zustimmenden Antwortstil nutzen, wurde bereits früh darauf zurückgegriffen, die Items positiv und negativ zu polen. Bei positiv gepolten Items entspricht eine hohe Antwortkategorie einer höheren Merkmalsausprägung. Messen wir beispielsweise Angstsensitivität mit dem Item „Wenn mein Herz schnell schlägt, habe ich Angst, einen Herzinfarkt zu kriegen" und nutzen eine fünfstufige Antwortskala (1 = starke Ablehnung, 5 = starke Zustimmung), dann bedeutet eine höhere Antwortkategorie auch eine höhere Merkmalsausprägung. Um nun Akquieszenz zu vermeiden, haben verschiedene Autoren vorgeschlagen, auch negativ gepolte Items in den Fragebogen einzubauen. Bei diesen Items entspricht eine höhere Antwortkategorie dann einer niedrigeren Merkmalsausprägung. Unser Beispielitem könnte lauten: „Wenn mein Herz schnell schlägt, mache ich mir darüber keine weiteren Gedanken."

So überzeugend dieses Vorgehen erscheinen mag, birgt es ernst zu nehmende Gefahren (Barnette, 2000; Locker, Jokovic & Allison, 2007). Die prominenteste Gefahr besteht darin, dass die Verwendung negativ gepolter Items die Faktorstruktur des Instruments beeinflusst. So kann es sein, dass jeweils die positiven und die negativen Items einen gemeinsamen Faktor bilden. Dies kann daran liegen, dass die negativen Items schwerer zu beantworten sind und somit auch verbale Intelligenz bei der Itembeantwortung erfasst wird (Marsh, 1996). Dies ist natürlich nicht wünschenswert. Ein weiteres Problem liegt darin, dass die unterschiedlich gepolten Items oft unterschiedliche Mittelwerte aufweisen. Das würde bedeuten, wenn die Anzahl der unterschiedlich gepolten Items in einem Fragebogen nicht identisch ist, wird der Summenwert in eine bestimmte Richtung verzerrt. Abschließend sei angemerkt, dass die Korrelationen zwischen Summenwerten aus positiv bzw. negativ gepolten Items und Fremdbeurteilungen verschieden waren. In der Summe kann also nicht zwingend davon ausgegangen werden, dass die beiden Itemarten dasselbe Konstrukt erfassen. Daher ist es fraglich, ob die zahlreichen Nachteile es rechtfertigen, unterschiedliche Polungen zu nutzen. Sollten diese dennoch enthalten sein, empfiehlt es sich, getrennte Analysen durchzuführen, um die Vergleichbarkeit zu gewährleisten.

Doppelte Verneinungen Insbesondere in Übersetzungen von englischsprachigen Fragebögen werden häufig doppelte Verneinungen verwendet. Diese sind unter Umständen schwer zu verstehen und können zu einer nicht gewollten falschen Antwort des Probanden führen (Beispiel: „Ich würde mich nicht als jemanden bezeichnen, der nicht traurig ist"). Von wesentlicher Bedeutung im Zusammenhang mit doppelten Verneinungen ist die verwendete Stichprobe. Während für Studenten z.B. komplexere Fragen oder doppelte Verneinungen ein weniger großes Problem darstel-

len, kann die Beantwortung für bestimmte Personengruppen (z.B. Patienten mit niedrigem Bildungsgrad) problematisch werden. Solche Verständnisschwierigkeiten und infolgedessen eine geringere Motivation, den Fragebogen auszufüllen, können ein zusätzlicher, unerwünschter Effekt doppelter Verneinungen sein. Daher ist von doppelten Verneinungen in der Regel dringend abzuraten.

Registrierung von Verhaltenshäufigkeit Durch Häufigkeitsfragen erfasst man, wie oft ein Ereignis innerhalb eines bestimmten Referenzzeitraums aufgetreten ist. Bei der Gestaltung der Fragen ist es wichtig, dass der Testkonstrukteur das ihn interessierende Ereignis genau definiert, einen geeigneten Referenzzeitraum wählt und ein dazu passendes Antwortformat verwendet.

- *Definition Ereignis*. Eine genaue Beschreibung des interessierenden Ereignisses ist wichtig, damit eine Person die Frage richtig interpretiert. Wird im Rahmen einer Häufigkeitsfrage nur vage formuliert, um welche Ereignisse es geht, nutzen Personen eher Kontextinformation (z.B. Merkmale des Fragebogens wie Instruktion, Antwortvorgaben oder Layout), um die Frage zu interpretieren (Schwarz, 2007).
- *Wahl des Referenzzeitraums*. Der Referenzzeitraum sollte passend zu dem erfragten Ereignis gewählt werden, denn auch er kann beeinflussen, wie die Person die Frage interpretiert (Winkielman, Knäuper & Schwarz, 1998): Werden beispielsweise sehr große Referenzzeiträume verwendet (z.B. „Wie häufig haben Sie sich in den letzten zwölf Monaten geärgert?“), gehen Personen eher davon aus, dass seltenere oder wichtigere Ereignisse abgefragt werden, als bei der Verwendung von kleineren Referenzzeiträumen (z.B. „Wie häufig haben Sie sich innerhalb der letzten Woche geärgert?“).
- *Wahl der Antwortkategorie*. Als Antwortformat können Ratingskalen, die Vorgabe von Zeiträumen oder offene Antwortformate gewählt werden. Ratingskalen sind zwar einfach zu konstruieren und für Personen leicht und intuitiv zu verstehen, sie weisen jedoch auch einige Nachteile auf (Wänke, 2002). Die Person setzt die Häufigkeit des Ereignisses ins Verhältnis zu subjektiven Standards: Welche absolute Häufigkeit eine Person beispielsweise als „sehr häufig“ betrachtet, kann unter anderem davon abhängen, welches Ereignis abgefragt wird und wie häufig dieses Ereignis normalerweise bei der Person beziehungsweise ihrer Referenzgruppe (Normalbevölkerung, Freunde usw.) auftritt. Nach Schwarz (1999) sind Ratingskalen deswegen ungeeignet, um objektive Häufigkeiten zu erfassen.

Bei der Verwendung von Antwortkategorien (z.B. „keinmal“, „ein- bis zweimal“, „drei- bis viermal“, „fünf- bis sechsmal“, „über sechsmal“) ist zu berücksichtigen, dass Personen davon ausgehen, dass die Mittelkategorie einer durchschnittlichen, normalen Häufigkeit entspricht. Dies kann dazu führen, dass sie ihre Häufigkeitsangabe an diese Referenz anpassen (Schwarz, 1999). Bei der Konstruktion ist darauf zu achten, dass die Häufigkeitsvorgaben repräsentativ für das tatsächliche Auftreten des Ereignisses in der Stichprobe sind.

Aufgrund der Nachteile von geschlossenen Antwortformaten empfiehlt Schwarz (1999) die Verwendung von offenen Antwortformaten. Dabei ist es wichtig, dass die Maßeinheit angegeben wird, beispielsweise:

- Wie viele Stunden hören Sie in der Woche Musik? ____ Stunden pro Woche

Dies verhindert, dass ungenaue Antworten wie „ein paar Stunden“ angegeben werden. Dabei ist es wichtig, eine für das Ereignis passende Maßeinheit zu wählen.

Weitere Hilfen zur Aufgabenkonstruktion Angleitner, John und Löhr (1986, S. 69) haben ein sehr nützliches Kategoriensystem entwickelt, das helfen kann, Fragebogenitems zu systematisieren:

- Beschreibung eigener Reaktionen:
  1. prinzipiell **beobachtbar** („Ich gehe oft auf Partys.“)
  2. **internal** und nicht prinzipiell beobachtbar („Ich grüble viel.“)
  3. **Symptome** („Ich schwitze viel.“)
- Eigenschaftszuschreibungen („Ich bin ein geselliger Mensch.“)
- Wünsche und Interessen („Ich wäre gern Blumenhändler.“)
- biografische Fakten („In meiner Jugend bin ich schon mal mit dem Gesetz in Schwierigkeiten gekommen.“)
- Einstellungen und Überzeugungen („Ich glaube an die Wiederkunft Christi.“)
- Reaktionen anderer gegenüber der Person („Meine Familie ist mit dem Beruf, den ich gewählt habe, nicht einverstanden.“)
- bizarre Items („Man wollte mich schon einmal vergiften.“)
- ergänzt von Tränkle (1983, S. 243): Frage nach Motiven („Warum sind Sie dieser Meinung?“)

Beispiel 3.15 **Itemformulierungen**

Im Folgenden werden einige Aspekte der Aufgabenkonstruktion, die bereits behandelt worden sind, anhand von Beispielen erläutert und durch weitere ergänzt.
Begriffe mit **mehreren Bedeutungen** sollten vermieden werden.

Ich bin in Gesprächen *angriffslustig*.

Trifft nicht zu ①–②–③–④–⑤ Trifft zu

– In diesem Beispiel ist unklar, ob „angriffslustig“ positiv im Sinne von „Ich vertrete meine Meinung offensiv“ oder negativ in Form von „Ich mache andere nieder“ gemeint ist.

Begriffe und **Formulierungen vermeiden**, die nur einem **Teil der in Aussicht genommenen Zielgruppe** (im Beispiel: Fragebogen für Kinder) **geläufig** sind.

Ich fühle mich *depressiv*.

Trifft nicht zu ①–②–③–④–⑤ Trifft zu

Jedem **Item** nur **einen sachlichen Inhalt**/Gedanken zugrunde legen.

Ich fahre sehr gerne und sehr schnell Auto.

Trifft nicht zu ①–②–③–④–⑤ Trifft zu

– In diesem Item sind zwei Aussagen vermischt, die voneinander unabhängig sein können. Man kann gerne Auto fahren, aber nicht unbedingt „schnell“ fahren. Besser ist es, eine solche Frage in zwei Teilfragen zu zerlegen:

Ich fahre sehr gerne Auto.

Trifft nicht zu ①–②–③–④–⑤ Trifft zu

Ich fahre sehr schnell Auto.

Trifft nicht zu ①–②–③–④–⑤ Trifft zu

Keine doppelten Verneinungen verwenden, da diese die Verständlichkeit verringern und zu einer längeren Aufgabenbearbeitung führen können.
Ich bin *nicht* oft *un*glücklich.
Trifft nicht zu ①–②–③–④–⑤ Trifft zu

Verallgemeinerungen vermeiden.

Alle Kinder machen Lärm.

Trifft nicht zu ①–②–③–④–⑤ Trifft zu

– Formulierungen wie „immer", „alle", „keiner", „niemals" sollten vermieden werden. Es kann sein, dass Befragte solche pauschalen Aussagen ablehnen. In spezifischen Kontexten, z.B. zur Erfassung irrationaler Einstellungen, können sie jedoch sinnvoll sein.

Umständliche Längen und telegrafische Kürzen vermeiden.

U. U. ist es *m. E.* legitim, gegen Friedensbewegungsbefürworter mit Polizeigewalt vorzugehen.

Trifft nicht zu ①–②–③–④–⑤ Trifft zu

Wichtiges durch **Fettdruck**, **Unterstreichen** oder Ähnliches **hervorheben**. Allerdings sollte mit Hervorhebungen sparsam umgegangen werden, da sie sonst unübersichtlich sind und verwirren.

Für **mich** ist es wichtig, die Kontrolle zu behalten.

Trifft nicht zu ①–②–③–④–⑤ Trifft zu

Bei **positiv und negativ gepolten Items** sollte man sich bei späteren Analysen daran erinnern, dass beide Items bis auf eine unterschiedliche Polung den gleichen Inhalt erfassen sollen, aber bei Analysen zu Artefakten führen können (z.B. Faktorenanalyse: zwei Faktoren, Faktor „positive Items" und Faktor „negative Items").

Beispiel von zwei Items, die etwas Ähnliches wie Extraversion erfassen könnten:

Item 1: Ich gehe gerne aus.

Trifft nicht zu ①–②–③–④–⑤ Trifft zu

Item 2: Ich gehe nicht gerne aus.

Trifft nicht zu ①–②–③–④–⑤ Trifft zu

Der **Zeitpunkt bzw. die Zeitspanne**, auf die Bezug genommen wird, sollte **eindeutig definiert** sein.

In den letzten Wochen war ich häufig niedergeschlagen.

Trifft nicht zu ①–②–③–④–⑤ Trifft zu

– In diesem Beispiel ist der Zeitrahmen nicht klar. Ein Proband könnte als „letzte Wochen" die letzten zwei Wochen als Basis nehmen, ein anderer drei oder vier Wochen. Dementsprechend ändern sich auch die Häufigkeitsangaben. Gleiches gilt natürlich für den Begriff „häufig"; wann etwas „häufig" auftritt, kann von Probanden subjektiv unterschiedlich interpretiert werden. Ein weiteres Problem ist die Verwendung von Häufigkeitsangaben im Antwortstamm. Bortz und Döring (2006, S. 255) geben ein schönes Beispiel für die Unsinnigkeit von quantifizierenden Beschreibungen wie „fast", „kaum" oder „selten": „Ich gehe selten ins Kino." Der Antwortmodus ist dabei „nie-selten-gelegentlich-oft-immer". Besser ist, in einer solchen Frage den Zeitrahmen und die Häufigkeitsangaben genauer zu spezifizieren:

In den letzten zwei Wochen hatte ich jeden Tag mindestens einmal das Gefühl, niedergeschlagen zu sein.

Trifft nicht zu ①–②–③–④–⑤ Trifft zu

Fowler (1995, S. 78 ff.), Osterlind (1983) und Bortz und Döring (2002, S. 255) geben darüber hinaus folgende **weitere Richtlinien** zur Gestaltung von Items:

- Es sollte ein hohes Maß an Kongruenz zwischen jedem Item und dem zu messenden Merkmal bestehen. Erst dann können valide Interpretationen eines Testwerts vorgenommen werden.
- Items, bei denen von vornherein klar ist, dass die Probanden ihnen immer zustimmen bzw. sie immer ablehnen, sollten vermieden werden; sie enthalten keine Information über das Verhalten der untersuchten Personen.
- Das Itemformat sollte an das Ziel des Tests angepasst sein. So erfordern einfache Ziele auch einfache Items, z.B. sollten die Items in einem Speedtest möglichst einfach gestaltet sein, wohingegen die Aufgaben eines Intelligenztests ein breites Spektrum an Schwierigkeiten abdecken sollten, um das komplexe Konstrukt der Intelligenz adäquat in allen Fähigkeitsbereichen zu erfassen.
- Falls es notwendig ist, Definitionen zu geben, sollten diese genannt werden, bevor die eigentliche Frage gestellt wird.
- Die Frage sollte in sich geschlossen sein. Wenn es Antwortalternativen gibt, sollten diese im Anschluss an die Frage dargeboten werden.
- Gleichlautende Items in einer Skala führen zu Verzerrungen bei den späteren statistischen Auswertungen. Die Trennschärfen (Korrelation des Items mit dem Test, vgl. *Kapitel 5.2*) solcher Items werden überschätzt.
- Es sollten Items in einem breiten Fähigkeits- oder Eigenschaftsspektrum konstruiert werden. Eine Einschränkung kann je nach Testziel getroffen werden (beispielsweise nur schwere Aufgaben, weil eine Differenzierung im Rahmen einer leistungsfähigen Stichprobe vorgenommen werden soll, beispielsweise Piloten).

ZUSAMMENFASSUNG

Janke (1973, S. 47) fasst die Problematik bei der Konstruktion von Fragebögen in sehr anschaulicher und systematischer Weise zusammen:

Die Itemsätze unserer Fragebögen stellen aber ein merkwürdiges Gemisch aus Fragen völlig unterschiedlicher logischer und empirischer Bezüge dar. Fragen nach konkreten biographischen Daten, nach Verhalten und Erleben in spezifischen Situationen stehen neben solchen, bei denen der Proband über Zeit und Intensität Verhalten und Erleben in spezifischen oder völlig vagen (etwa „aufregenden") Situationen zu integrieren hat oder zu entscheiden hat, ob ein Verhalten bei ihm häufiger oder intensiver als bei einem nicht-spezifizierten anderen ist. Daneben werden Interessen, Präferenzen, Einstellungen, Bewertungen eigener oder fremder Verhaltensweisen erfragt. Zusätzlich wird er als „Konstrukt-Konstrukteur" oder „Konstruktvalidierer" eingesetzt in direkten Fragen wie „Ich bin ein ängstlicher, ein neurotischer Mensch". Items im Sinne von „habits", „traits" und „types" stehen nebeneinander in einem einzigen Fragebogen; Verhaltensweisen und Teilaspekte eben dieser Verhaltensweisen werden bedenkenlos summiert. Einige Items erfragen Häufigkeiten von Verhalten, andere – im gleichen Fragebogen aufgeführte – erfragen Reaktionsintensitäten. Andere Items wiederum stellen Gemische dar aus Häufigkeit und Intensitäten (Beispiel: „Ich reagiere manchmal in bestimmten Situationen stark").

Zur sprachlichen Formulierung der Items können folgende Empfehlungen gegeben werden:

- Begriffe mit mehreren Bedeutungen vermeiden.
- Keine doppelten Verneinungen verwenden.
- Keine negativ gepolten Items verwenden.
- Verallgemeinerungen vermeiden.
- Wichtiges sparsam hervorheben.
- Keine Abkürzungen verwenden.
- Keine Fremdwörter verwenden und wenn, dann kurz erklären.
- Zeitspannen genau definieren.
- Keine Items verwenden, die zwischen Personen nicht differenzieren.
- Für jedes Item prüfen, ob es das Konstrukt abbildet.
- Für jedes Item prüfen, ob es für die Zielgruppe angemessen formuliert ist.
- Keine gleichlautenden Items in einer Skala verwenden.
- Items mit unterschiedlicher Itemschwierigkeit verwenden.
- Falls nur Items einer Kategorie vewendet werden sollen, dann für jedes Item prüfen, ob es tatsächlich diese Kategorie (z.B. Einstellungen) misst und keine andere Kategorie (z.B. Verhalten).
- Sollen Häufigkeiten möglichst exakt erfasst werden, dann diese konkret erfragen. Bei Verwendung von Kategorien bedenken, dass Personen die Mittelkategorie als durchschnittliche, normale Häufigkeit auffassen. Immer die Einheit nennen (z.B. Stunden pro Woche).

ZUSAMMENFASSUNG

Beispiel 3.16 **Konstruktion eines Konzentrationstests**

Wir haben uns also abschließend für einen Top-Down-Ansatz bei der Konstruktion eines Tests mit objektiven Aufgaben für die Zielgruppe von 15- bis 17-jährigen jungen Erwachsenen entschieden. Der Messgegenstand soll Konzentrationsfähigkeit sein. Es wird eine rationale Konstruktion angestrebt, die auf einem konkreten Modell basiert, das wiederum auf Basis empirischer Befunde und existierender theoretischer Modelle verschiedener Autoren beruht. Um Anhaltspunkte für die Aufgabenkonstruktion zu finden und als Vorstufe zur Erstellung einer Arbeitsdefinition wurde eine Literatursuche mit dem Ziel durchgeführt, immer wiederkehrende Definitionselemente für das Konstrukt Konzentration zu sammeln und zu systematisieren. Daraus ergab sich folgende Definition:

Der Test erfasst die Fähigkeit, sich zu konzentrieren. Darunter wird das Ausführen einer zielgerichteten fortlaufenden Selektion von Reizkonfigurationen unter Bedingungen verstanden, die normalerweise das Erbringen dieser Leistung erschweren. Erschweren meint, dass bei der Aufgabenbearbeitung irrelevante, störende Reize ausgeblendet werden müssen.

Im Anschluss mussten wir uns für die Wahl einer bestimmten Testart zur Erfassung von Konzentration entscheiden. Drei gängige Arten haben wir gefunden: Rechentests, Durchstreichtests und Kartensortierverfahren. Da Kartensortierverfahren sehr materialaufwendig sind und wir mit Rechentests nicht unsere Definition umsetzen können, da wir hier kaum Ablenkungsreize einfließen lassen können, haben wir uns für die Konstruktion eines Durchstreichtests entschieden. Um die Definitionsmerkmale umzusetzen, ordnen wir den Test in Zeilen an und lassen nach Reizkonfigurationen suchen. Durch die Wahl geeigneter Distraktoren versuchen wir die Arbeit zu erschweren:

Um die fortlaufende Reizselektion zu gewährleisten, ordnen wir 50 Aufgaben in zwölf Zeilen.