Echterhoff, G. (2013). Quantitative Erhebungsmethoden. In W. Hussy, M. Schreier & G. Echterhoff (Hrsg.), *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor* (S. 71-90). Berlin: Springer.

2.3 Selbstberichtsverfahren: Befragung und Rating

Lernziele

- Die grundlegenden kognitions- und kommunikationspsychologischen Grundlagen von Selbstauskünften in Befragungen kennenlernen.
- Die kognitiven Prozesse würdigen, die Selbstberichten zugrunde liegen.
- Verschiedene Formen der Befragung unterscheiden und die Anwendungsbereiche dieser Varianten identifizieren.
- Einen Überblick über Ratingverfahren und Varianten von Ratingskalen gewinnen.
- Urteilstendenzen berücksichtigen, die bei Antworten auf Ratingskalen auftreten können.
- Den Einsatzbereich von psychologischen Tests und ihre idiografische Zielsetzung kennenlernen.
- Leistungstests und Persönlichkeitstests unterscheiden.
- Wünschenswerte Eigenschaften von Tests und Testitems würdigen. Darunter: Schwierigkeit und Trennschärfe von Testitems sowie deren Zusammenhang kennenlernen.
- Eine Sensibilität für mögliche Verfälschungen von Tests entwickeln und einen Überblick über Gegenmaßnahmen gewinnen.

Die Befragung ermöglicht einen Zugang zu psychischen Prozessen und Strukturen.

Eine Möglichkeit, Zugang zu den nicht direkt beobachtbaren psychischen Prozessen und Strukturen zu erhalten, besteht darin, Menschen zu befragen und über diese psychologisch relevanten Sachverhalte berichten zu lassen. Gegenstand solcher Berichte sind z. B. die eigenen Einstellungen, Gefühle, Denkprozesse, Motive oder Erinnerungen. Dabei ist man in der Forschung darauf angewiesen, dass die untersuchten Personen überhaupt Zugang zu den interessierenden psychischen Prozessen haben und dass sie diese auch kommunizieren können. Zudem müssen die Selbstauskünfte von Untersuchungsteilnehmern und -teilnehmerinnen hinreichend reliabel und valide sein. Dabei sind zwei grundlegende Aspekte – ein kognitionspsychologischer und ein kommunikationspsychologischer – zu berücksichtigen (Schwarz, 1999), die im Folgenden kurz skizziert werden, und zwar:

- Wie gelangen Befragte zu Selbstauskünften?
- Wie werden Selbstberichte kommuniziert?

Wie gelangen Befragte zu Selbstauskünften?

Selbstauskünfte in Befragungen beinhalten drei Prozesse.

Selbstauskünfte in Befragungen beinhalten drei elementare kognitive bzw. mentale Prozesse (Pelham & Blanton, 2007; Strack & Martin, 1987; Sudman, Bradburn & Schwarz, 1996).

1. Interpretation der Frage

Der erste Prozess beinhaltet die **Interpretation der Frage**.

Zunächst müssen die Befragten eine Frage interpretieren, also erwägen, was der Forscher oder die Interviewerin wohl mit einer Frage meint. Wenn Studierende mit einer Frage wie »Was halten Sie von der Lehre von Dozentin X?« konfrontiert werden, so werden sie zunächst etwa überlegen, welche Aspekte der Lehre der Dozentin gemeint sind (z. B. die fachliche Kompetenz, das Engagement, die didaktische Kompetenz, die Struktur der Veranstaltungen etc.) und welche Informationen sich die Auftraggeber der Befragung von der Befragung erhoffen (▶ Beispiel).

Beispiel

Interpretation der Frage

Eine Frage wie »Wie bewerten Sie auf einer Skala von 1 bis 8 die USA?« lässt offen, auf welchen Dimensionen bzw. in welchen Hinsichten die USA beurteilt werden sollen. Befragte müssen zu plausiblen Hypothesen hierzu gelangen, bevor sie überhaupt über Antworten auf die Frage nachdenken können. Ist das aktuelle globale politische Engagement der USA gemeint oder soll es eher um die historische Rolle der USA gehen? Oder sind vielleicht eher US-amerikanische Personen gemeint, also z. B. bekannte Politiker und Politikerinnen oder andere Personen des öffentlichen Interesses?

Die Herausforderung der Interpretation von Fragen stellt sich für die Untersuchungsteilnehmer in mehr oder weniger großem Ausmaß bei allen Befragungen – man denke an Fragen wie »Wie zufrieden sind Sie mit Ihrem derzeitigen Lebensstandard?«, »Wie leicht fällt es Ihnen, sich an Ihre Kindheit zu erinnern?«, »Wie gut gefällt Ihnen das Buch *Einführung in die psychologische Methodenlehre*?«. Stets müssen Befragte als erstes die Frage verstehen und interpretieren.

Der zweite Prozess umfasst den Abruf und die **Konstruktion eines Urteils**, das die Beantwortung der Frage erlaubt.

2. Bildung eines Urteils

Zudem müssen Befragte eine Antwort auf die Frage finden, also ein eigenes Urteil bilden. Dieser Schritt ist oft nicht minder anspruchsvoll und komplex als der erste (▶ Beispiel).

Beispiel

Urteilsbildung

Zunächst wäre in dem vorherigen Beispiel zu entscheiden, ob mit der Frage »Wie bewerten Sie auf einer Skala von 1 bis 8 die USA?« die Gegenwart oder die historische Rolle der USA angesprochen ist. Wenn die Befragten davon ausgehen, dass die USA der Gegenwart gemeint sind, müssen sie ein Urteil generieren. Dies können sie leisten, indem sie ein bereits bestehendes Urteil zu den gegenwärtigen USA abrufen oder zum Befragungszeitpunkt ein Urteil konstruieren. Um ein Urteil zu konstruieren, können Menschen sehr viele unterschiedliche Wege wählen. Sie können sich Bilder von urteilsrelevanten Situationen oder Personen wachrufen, bei der Beantwortung der Beispielfrage etwa von den weltweit Entsetzen auslösenden Bildern von den Misshandlungen irakischer Gefangener durch US-Soldaten und -Soldatinnen in Abu Ghraib. Man kann auch an bekannte US-amerikanische Politiker, Schauspieler oder Sportler denken, sich an persönliche Kontakte mit Amerikanerinnen und Amerikanern erinnern oder an die Kommentare eines guten Freundes über die Politik der USA.

Die Beispiele zeigen die Vielfalt mentaler Operationen, die an der Beantwortung einer Selbstberichtsfrage beteiligt sind.

Wie diese Beispiele zeigen, können Menschen eine Vielfalt mentaler Operationen ausführen, um zu einem Urteil oder einer Vorstellung zu gelangen, auf deren Basis sie in einer Befragung eine Antwort geben können. Durch die Wortwahl einer Frage, die Befragungssituation oder zuvor dargebotene und damit voraktivierte Informationen können bestimmte Operationen und Informationen nahegelegt werden, um zu einem Urteil zu gelangen. Wenn etwa Gedanken an die Opfer des Zweiten Weltkriegs oder des Holocaust bei einer Meinungsumfrage zu den USA wachgerufen werden, werden die Befragten an die positive historische Rolle der USA denken (z. B. die Unterstützung Europas bei dem Sieg über Nazi-Deutschland während des Zweiten Weltkriegs). In diesem Fall wird ein vergleichsweise positives Urteil über die USA konstruiert, als wenn Gedanken an die Gräuel des Vietnam-Kriegs bei den Befragten aktiviert (im anglophonen Fachjargon: »geprimed«) werden.

Der dritte Prozess beinhaltet die **Übersetzung in eine kommunizierte Auskunft**.

3. Übersetzung in eine kommunizierte Auskunft

Doch mit dem Abrufen oder Konstruieren eines Urteils ist es noch nicht getan. Abschließend müssen Befragte ihre mentalen Reaktionen, Urteile oder Vorstellungen auch in eine offen kommunizierte Auskunft übertragen. Die Herausforderung besteht darin, einen internen psychischen Zustand in eine extern kommunizierte Mitteilung zu überführen. Gerade bei vorgegebenen Antwortformaten (wie z. B. einer numerischen Skala) kann dies ein schwieriges Unterfangen sein. Je nach Annahmen über die Bedeutung der Antwortoptionen bzw. je nach Vergleichsstandards werden Befragte ihre mentalen Reaktionen oder Urteile in sehr unterschiedliche Antworten überführen (▶ Beispiel).

Beispiel

Kommunikation einer Antwort

Wie gelangen Befragte im bereits beschriebenen Beispiel von einem konstruierten Urteil über die USA dazu, einen bestimmten Wert auf einer Antwortskala zur Bewertung der USA anzukreuzen? Dies hängt u. a. von den konkreten Erfahrungen und Informationen ab, die Befragte in der Situation aus ihrem Gedächtnis abrufen. Eine Befragte kann durch die Vorstellung problematischer Aktionen des US-amerikanischen Militärs zu einem tendenziell kritischen Urteil gekommen sein, ein anderer Befragter durch die Erinnerung an positive persönliche Kontakte mit Amerikanern bei einer Reise zu einem tendenziell positiven Bild. Wie lassen sich diese Urteile auf einen Zahlenwert von 1 bis 8 bringen? Bei einer kritischen Auffassung kann es naheliegen, einen Wert unterhalb der Skalenmitte (4, 5) zu wählen. Aber wählt man nun die 4 oder 3 oder gar die 2 oder 1? Und verglichen mit anderen Ländern stehen die USA trotz aller Kritik vielleicht positiver da als der Durchschnitt. Aber wählt man in diesem Fall die 5 oder die 6 oder noch größere Werte? Dies kann von verschiedenen Faktoren der Skalenkonstruktion abhängen, z. B. schon allein von der Beschriftung der Skala. Wenn die Endpunkte der Skala mit 1 = »extrem negativ« bis 8 = »extrem positiv« beschriftet sind, ist etwa mit weniger extremen Antworten zu rechnen als bei der Beschriftung 1 = »vorwiegend negativ« bis 8 = »vorwiegend positiv«.

Wie werden Selbstberichte kommuniziert?

Bei einem Selbstbericht handelt es sich um einen Akt der Kommunikation. Eine untersuchte Person teilt sich bewusst mit, sie wendet sich mit einem Bericht an einen oder mehrere Adressaten. Menschliche Kommunikation besteht nicht nur darin, dass ein Sender Informationen vermittelt, sondern impliziert auch eine Absicht des Senders, etwas mitzuteilen (z. B. Luhmann, 1995). Das bedeutet, dass ein Selbstbericht nicht auf die übermittelte Information reduziert werden sollte, sondern stets auch vor dem Hintergrund der (vermutlichen) **Mitteilungsabsicht** des Senders zu interpretieren ist. Bei Erhebungen auf der Basis von Selbstberichten ist also zu bedenken, welche Intention die über sich berichtende Person mit ihrer Mitteilung verfolgt (▶ Beispiel).

Ein Selbstbericht ist als **intentionaler Kommunikationsakt** zu verstehen.

Beispiel

Selbstbericht als Kommunikation mit den Forschenden

Wenn eine studentische Versuchsperson in einem von der Fakultät veranlassten Interview auf die Frage »Was halten Sie von der Lehre von Dozentin X?« antwortet »Ich finde, sie kann ihr Spezialgebiet gut darstellen«, dann spielt nicht nur die wörtlich kommunizierte Information (die Meinung, dass die Professorin ihr Spezialgebiet gut darstellen kann) eine Rolle. Vielmehr ist auch die vermutliche Mitteilungsabsicht in der Befragungssituation zu berücksichtigen, also z. B. das Bemühen, sich nicht zu negativ oder kritisch über Fakultätsangehörige zu äußern oder dem Interviewer zu signalisieren, dass man über negative Aspekte keine direkte Auskunft geben möchte. Bei Antworten in Befragungen zum Lernverhalten von Studierenden ist auch die mögliche Absicht einer positiven Selbstdarstellung zu berücksichtigen; wenn Studierende angeben, intensiv und wiederholt die englischsprachige Prüfungsliteratur zu lesen, so könnte diese Äußerung zum Teil durch diese Selbstdarstellungsabsicht bedingt sein.

Wie das Beispiel illustriert, ist also zu überlegen, weshalb und aus welchen Motiven eine Person eine Mitteilung macht. Die Herausforderung besteht nun darin, bei der Gestaltung und Konstruktion einer Befragung, zumal einer mit standardisierten und strukturierten Antwortoptionen, sowie bei der Auswertung der Ergebnisse diese kognitions- und kommunikationspsychologischen Prozesse hinreichend zu berücksichtigen. Folgende Fragen dienen als Leitfaden bei der Analyse von Daten aus Selbstberichten:

- Interpretieren die Befragten die Frage oder die Themenstellung so, wie es die Forschenden vorgesehen haben?
- Wie wird sichergestellt, dass die Befragten möglichst an diejenige Aspekte oder Themen denken, die die Forschenden bei der Formulierung der Frage im Blick hatten?
- Werden durch die Art der Befragung (z. B. durch die Wortwahl oder die Reihenfolge von Fragen, oder die Befragungssituation) Informationen nahegelegt oder vor-

Bei der Konstruktion und Auswertung einer Befragung ist zu berücksichtigen, aus welchen Motiven eine Person Auskunft über sich gibt.

aktiviert, die die Befragten zur Konstruktion einer Antwort oder eines Urteils heranziehen?
- Wie gut ist gewährleistet, dass die Befragten ihre intern gebildeten Urteile auch in eine Antwort umsetzen können? Können die Antworten adäquat in das vorgegebene Format (z. B. die Skala und deren Beschriftung) übersetzt werden?
- Lassen sich die Absichten oder Motive der Befragten abschätzen, die der Mitteilung von Antworten zugrunde liegen?

Bei der **Befragung** handelt es sich um eine allgemeine Form der Datenerhebung, das **Rating** ist eine spezielle Form der Befragung.

Es gibt eine große Bandbreite von Verfahren, um Selbstauskünfte von Untersuchungsteilnehmern zu erheben. Diese Verfahren lassen sich unterteilen in Befragungs- und Ratingverfahren (»to rate« = beurteilen). Die Befragung stellt eine allgemeine Grundform der Datenerhebung dar; das Ratingverfahren ist eine spezielle Variante der Befragung. Eine Befragung erfordert gegenüber vielen anderen Erhebungsmethoden vergleichsweise weniger finanziellen und zeitlichen Aufwand und ist auch aus diesem Grund das in der Psychologie und den Sozialwissenschaften wohl am meisten verwendete Verfahren. Ratings stellen ein gängiges standardisiertes Format der Befragung dar und werden gesondert behandelt (▶ Abschn. 2.3.2). Aspekte von Befragungsmethoden kommen auch beim Testen (▶ Abschn. 2.4) oder bei der Erhebung verbaler Daten im Bereich qualitativer Methoden zum Zuge (▶ Abschn. 6.1).

2.3.1 Befragung: Varianten und Prinzipien der Konstruktion

Bevor wir das spezielle Befragungsformat des Ratings besprechen, befassen wir uns zunächst mit der Befragung als grundlegender Form der Datenerhebung. Die Fragen in Fragebögen werden auch als **Items** bezeichnet.

▶ **Definition Item**

Definition

Ein **Item** ist eine als Frage oder als Urteil formulierte Aussage, zu der die befragte Person ihre Zustimmung oder Ablehnung – ggf. in unterschiedlicher Intensität – äußern kann.

Unterscheidungskriterien von Befragungen:

Die schriftliche Befragung erfolgt über einen **Fragebogen**, die mündliche in Form eines **Interviews**.

Befragungsverfahren lassen sich nach den folgenden **Kriterien** unterteilen:

Schriftliche vs. mündliche Befragung Die schriftliche Befragung wird mittels **Fragebogen** durchgeführt; sie findet sich häufiger in der quantitativen als in der qualitativen Forschung. Die mündliche Form der Befragung (**Interview**) ist dagegen für die qualitative Forschung typischer.

Im Gegensatz zur **unstandardisierten** werden bei der **standardisierten** Befragung den Befragten verschiedene Antwortalternativen vorgegeben.

Standardisierte vs. nichtstandardisierte Befragung Der Grad der Standardisierung einer Befragung bezieht sich auf die Freiheitsgrade der befragten Person. Bei einer standardisierten Befragung werden (in Form **geschlossener Fragen**) Antwortmöglichkeiten vorgegeben, unter denen die befragte Person auswählt. Bei der unstandardisierten Befragung (in Form **offener Fragen**) kann die befragte Person die Antwort in ihren eigenen Worten formulieren. Quantitative Befragungsverfahren sind standardisierte Verfahren.

Bei der **strukturierten** im Gegensatz zur **unstrukturierten** Befragung muss sich der Forschende bei der Durchführung der Befragung an exakte Vorgaben halten.

Strukturierte vs. unstrukturierte Befragung Der Grad der Strukturiertheit einer Befragung bezieht sich auf die **Freiheitsgrade der Forschenden**. Bei einer strukturierten Befragung sind Wortlaut und Reihenfolge der Fragen bzw. Items genau vorgegeben. Bei einer **halbstrukturierten** Befragung ist als Orientierung ein Leitfaden mit vorformulierten Fragen vorgegeben; die genaue Formulierung und die Reihenfolge der Fragen

kann jedoch variieren. In der **nichtstrukturierten** Befragung schließlich muss der Forscher bzw. die Forscherin hinsichtlich Frageformulierung und Reihenfolge der Items keine Vorgaben einhalten. In der quantitativen Forschung dominiert die strukturierte Befragung.

Anzahl der befragten Personen An einer **Einzelbefragung** nimmt nur eine befragte Person teil. Bei einer **Gruppenbefragung** oder -diskussion werden die Meinungen und Einstellungen einer kleinen Gruppe (zwischen 5 und 15 Personen) erhoben. Bei einer **Umfrage** (»Survey«) schließlich kann die Anzahl der Befragten sehr hoch sein (▶ Abschn. 3.3).

Befragungen unterscheiden sich bezüglich der **Anzahl** der befragten Personen.

Unterschiede zwischen schriftlicher und mündlicher Befragung (Interview)

Schriftliche Befragungen weisen oft einen höheren Grad an Standardisierung auf als mündliche Interviews. Mündliche Interviews haben in vielen Fällen keinen vorgegebenen Ablauf, der für alle Befragten identisch ist. Fragebögen enthalten meist geschlossene Fragen mit vorgegebenen Antworten. Um solche Fragen zu formulieren, ist Vorwissen über den Gegenstand erforderlich. Da schriftliche Befragungen im Vergleich zu einem Interview eine größere Anonymität erlauben, sind sie potenziell in einem geringeren Maß mit dem Problem der Reaktivität behaftet (▶ Abschn. 2.1.2). Daher ist die Wahrscheinlichkeit, dass die untersuchten Personen verfälschte (so etwa sozial erwünschte) Antworten geben, bei (anonymen) schriftlichen Befragungen in der Regel geringer als bei mündlichen Befragungen. Denn gerade bei Face-to-Face-Interviews ist die Möglichkeit sozialer Beeinflussungs- oder Übertragungseffekte aufgrund des Verhaltens der befragenden Person nicht von der Hand zu weisen. Andererseits sind Befragte bisweilen in mündlichen Interviews eher bereit oder in der Lage, ausführlich über sich Auskunft zu geben. Allerdings sind Interviews in der Regel aufwändiger und daher auch kostenintensiver. Die Entscheidung für eine mündliche oder schriftliche Befragung hängt davon ab, welche Rolle die beschriebenen Faktoren für die Erreichung des Forschungsziels spielen und welche (finanziellen, personellen etc.) Ressourcen zur Durchführung vorhanden sind.

Die Konstruktion schriftlicher Befragungen erfordert mehr Vorwissen über den Gegenstand als eine offene mündliche Befragung; zudem ist bei schriftlichen Befragungen das Reaktivitätsproblem geringer, jedoch äußern sich Befragte bisweilen eher und ausführlicher in mündlichen Befragungen; Letztere sind aufwändiger.

Varianten und wesentliche Aspekte der schriftlichen Befragung

Die mündliche Form der Befragung, also das Interview, wird im Kapitel zu qualitativen Methoden (▶ Abschn. 6.1) ausführlich vorgestellt. Im Folgenden wird die schriftliche Befragung erläutert, wobei wir insbesondere auf Ratings eingehen, weil diese ohne größere Datenaufbereitung quantitative Informationen liefern.

Die Validität der Ergebnisse einer schriftlichen Befragung kann leiden, wenn die Versuchsleiter keine Kontrolle über die Befragungssituation haben. Dies ist der Fall, wenn die Befragten darüber entscheiden, wann, wo und unter welchen Bedingungen (alleine, in Anwesenheit anderer, mit oder ohne Zeitdruck) sie die Fragen beantworten – so etwa bei postalischen Befragungen oder Befragungen im Internet (▶ Abschn. 2.6). Daher sollte, wenn möglich, eine schriftliche Befragung unter standardisierten Bedingungen und in Anwesenheit eines Versuchsleiters durchgeführt werden. Diese Bedingungen können gut bei Befragungen in Untersuchungsräumen einer Forschungsinstitution gewährleistet werden. Jedoch erfordern Befragungen in Institutsräumen eine effektive Terminabsprache. Die Terminproblematik kann verringert werden, indem man die Befragung in einem Raum einer Organisation (z. B. Schule, Unternehmen etc.) durchführt, in der die Befragten ohnehin anwesend sind.

Wenn möglich, sollte eine schriftliche Befragung unter standardisierten und kontrollierten Bedingungen erfolgen.

Forschende sollten bei der Konstruktion einer Befragung, insbesondere bei der Formulierung der Fragen, die oben genannten kognitions- und kommunikationspsychologischen Aspekte berücksichtigen. Eine gut konstruierte Befragung erlaubt es der Forscherin bzw. dem Forscher zu ermessen bzw. zu kontrollieren,

Kognitions- und kommunikationspsychologische Aspekte sollten bei Befragungen berücksichtigt werden.

1. wie die Befragten die Fragen interpretieren,
2. wie und worüber sie sich daraufhin ein Urteil gebildet bzw. relevante mentale Zustände generiert haben,
3. ob und wie zutreffend sie das intern generierte Urteil bzw. die mentalen Zustände in eine offen kommunizierte Antwort überführen sowie
4. welche Absichten oder Motive der Mitteilung von Antworten vermutlich zugrunde liegen.

Tipps zur Formulierung von Fragen

Ein guter Fragebogen ist gekennzeichnet durch:

– einfache Formulierung und gute Verständlichkeit,

– keine zu hohen Anforderungen an die mentale oder kognitive Leistungsfähigkeit der Befragten,

– adressatenorientierte Formulierung,

– keine Verneinungen in den Fragen,

– keine überfrachteten Fragen,

– keine »Forced Choice« bei unabhängig beantwortbaren Aspekten,

– keine Fragen, die die Befragten sehr ähnlich beantworten,

– Einsatz mehrerer Items zur Beantwortung einer Frage,

Folgende konkrete Hinweise zur Formulierung von Fragen sollten beherzigt werden:

- Die Frage sollte möglichst einfach formuliert sein und damit gut verständlich. Zu vermeiden sind:
  a. ungebräuchliche Begriffe oder Ausdrücke, Fachbegriffe oder Fremdwörter (lexikalische Ebene),
  b. lange und verschachtelte Sätze, ungewöhnliche Satzkonstruktionen (syntaktische Ebene),
  c. zu abstrakte oder komplizierte Sachverhalte (semantische Ebene).
- Problematisch sind Fragen, deren Beantwortung möglicherweise zu hohe Anforderungen an die mentale oder kognitive Leistungsfähigkeit der Befragten stellt. Hierzu zählen Fragen, die eine übermäßig präzise Gedächtnisleistung voraussetzen wie z. B. »Wie viele Minuten haben Sie in der vergangenen Woche vormittags im Nahverkehr verbracht?« Hilfreich kann es sein, bei solchen Fragen zuerst den breiteren Kontext anzusprechen und dann das interessierende Verhalten im Einzelnen zu erfragen (»Denken Sie an Ihren Weg zum Arbeitsplatz bzw. zur Universität. – Wie lange sind Sie üblicherweise unterwegs? – An wie vielen Vormittagen legen Sie diesen Weg zurück?«)
- Diese ersten beiden Hinweise gelten nicht absolut, sondern sind relativ: Die Formulierung der Fragen sollte adressatenorientiert sein, also an die sprachlichen Gewohnheiten und Kompetenzen der Zielgruppe angepasst sein.
- Eine Frage sollte möglichst keine Verneinungen, zumindest keinen doppelten Verneinungen enthalten.
- Zu vermeiden sind überfrachtete oder vervielfachte Fragen, also Formulierungen, die mehr als einen Aspekt enthalten, auf den sich eine Antwort beziehen kann. Beispiele für solche Fragen sind: »Wie sehr mögen Sie Semesterferien und Ferienjobs?«, »Befürworten Sie das generelle Verbot von Ladenöffnungszeiten an Feiertagen, um die Interessen der Angestellten zu schützen?« Solche Fragen sollten möglichst in ihre Einzelaspekte aufgeschlüsselt werden (z. B.: »Befürworten Sie das generelle Verbot von Ladenöffnungszeiten an Feiertagen?«, »Würde das generelle Verbot von Ladenöffnungszeiten an Feiertagen die Interessen der Angestellten schützen?« und »Wie wichtig finden Sie es, beim Ladenschlussgesetz die Interessen der Angestellten zu schützen?«).
- In der Regel sollten auch Fragen vermieden werden, die die Wahl von sich logisch nicht ausschließenden Antwortoptionen erzwingen (»forced-choice items«), z. B. »Bevorzugen Sie Äpfel oder Orangen?« oder »Lesen Sie lieber die *Odyssee* oder *Einführung in die psychologische Methodenlehre*?« Da die Alternativen bei diesen Frageformulierungen in einer Relation zueinander stehen, lässt sich durch die Antwort nicht gut beurteilen, wie sie für sich genommen beurteilt würden.
- Ein Fragebogen sollte möglichst keine Fragen enthalten, die von praktisch allen Befragten in gleicher Weise beantwortet werden. Solche Fragen liefern keine Streuung innerhalb einer Stichprobe und sind daher nicht geeignet, um zwischen den Untersuchungsteilnehmern/innen zu differenzieren.
- Wenn erforderlich und möglich sollte nicht nur eine einzelne Frage verwendet werden, um einen interessierenden Aspekt oder Sachverhalt zu beleuchten. Zur

Abfrage des Alters reicht zwar eine Frage; aber viele psychologische Konstrukte, z. B. Selbstwert, Erfolgsmotivation oder Ängstlichkeit, sind weniger leicht zu erfassen als gemeinhin klar definierte Konstrukte wie »Alter«. Der Einsatz mehrerer Items zur Beantwortung einer Frage kann oft die Reliabilität (▶ Abschn. 1.4) der erhobenen Daten erhöhen.

- Bei der Anordnung bzw. Reihenfolge der Fragen sollte eine relativ einfache »Warm-up«-Phase zu Beginn vorgesehen werden. Beispielsweise sollte ein Fragebogen nicht mit allzu schwierigen Fragen beginnen; auch Fragen, die sehr persönlich sind, sollten nicht gleich am Anfang des Fragebogens gestellt werden.
- Gerade bei schriftlichen Fragebögen ist eine klare, informative und nicht zu ausschweifende Instruktion erforderlich, mit der den Befragten erläutert wird, welche Arten von Fragen gestellt werden und welche Möglichkeiten der Beantwortung zur Verfügung stehen.

– Beachtung der Ausgewogenheit in der Reihenfolge der Fragen,

– eine klare und informative Instruktion.

Diese grundlegenden Tipps sollten bei der Formulierung von Fragen berücksichtigt werden. Hilfreiche Hinweise zur Generierung und Auswahl von Fragen sowie zur Konstruktion von ganzen Fragebögen finden sich bei Pelham und Blanton (2007, Kap. 4).

2.3.2 Ratings: Beurteilungen auf Skalen

Bei einem **Rating** geben Befragte Urteile auf einer numerisch interpretierbaren Skala ab.

Wohl kaum ein Format wird in schriftlichen Befragungen häufiger verwendet als das Rating. Ratingverfahren verlangen von den Befragten, einen Gegenstand (sich selbst, andere Personen, unbelebte Objekte, Sachverhalte) hinsichtlich eines bestimmten Merkmals auf einer Skala zu beurteilen (z. B. Stimmung, Sympathie etc.). Ratingskalen geben gleich große markierte Abschnitte des Merkmalskontinuums vor. Die Befragten werden gebeten, diejenige Stufe der Ratingskala anzukreuzen, die ihrem persönlichen Empfinden hinsichtlich der Merkmalsausprägung bei dem zu beurteilenden Gegenstand entspricht. Beispielsweise können Befragte auf einer Ratingskala angeben, wie gut ihre aktuelle Stimmung ist (◘ Abb. 2.4).

Wie ist Ihre aktuelle Stimmung? (Zutreffendes bitte ankreuzen)

1	2	3	4	5	6
sehr schlecht					sehr gut

◘ **Abb. 2.4** Beispiel für eine Ratingskala

Ratingdaten wird oft Intervallskalenniveau zugebilligt.

Die große Beliebtheit von Ratingskalen in der Psychologie und den Sozialwissenschaften liegt u. a. darin, dass sie vergleichsweise direkt quantitative Daten (im Beispiel Zahlen zwischen 1 und 6) liefern und durch die Unterteilung der Merkmalsausprägungen in gleich große Abschnitte das Vorliegen eines datenanalytisch günstigen Skalenniveaus nahelegen (des Intervallskalenniveaus). Alleine die Vorgabe einer Ratingskala per se garantiert zwar noch keineswegs Intervallskalenniveau – denn ob der Abstand zwischen 1 und 2 dem zwischen 2 und 3 entspricht, ist nicht aus dem Format der Skala selbst, sondern nur inhaltlich psychologisch und/oder empirisch zu begründen. Jedoch wird in der Scientific Community selten Kritik geübt, wenn Daten aus Ratingskalen auf Intervallskalenniveau behandelt werden.

Besonderes Augenmerk gilt der Validität von Ratingdaten.

Zudem haben Ratingskalen als Instrument der quantitativen Sozialforschung eine hohe Augenscheinvalidität. Dabei ist die Validität von Ratingdaten in größerem Maß als ihr Skalenniveau umstritten. Vor der Verwendung von Ratingverfahren ist sorgfältig zu überlegen, inwiefern diese Selbstauskünfte über die zu untersuchenden mentalen Prozesse oder Strukturen Auskunft geben.

Abb. 2.5 Beispiele für eine unipolare und eine bipolare Ratingskala

Unipolare Ratingskala

Wie ruhig fühlen Sie sich im Moment? (Zutreffendes bitte ankreuzen)

1	2	3	4	5
ruhig				unruhig

Bipolare Ratingskala

Wie ruhig oder angespannt fühlen Sie sich im Moment? (Zutreffendes bitte ankreuzen)

1	2	3	4	5
ruhig				angespannt

Wesentliche Aspekte bei der Konstruktion von Ratingskalen

Bei der Konstruktion von Ratingskalen sind eine Reihe von Entscheidungen zu treffen, die Antworten auf die folgenden Fragen liefern:

Items können die Form einer Frage oder Aussage haben.

1. Wird ein Item als Frage (»Wie angespannt fühlen Sie sich gerade?«) oder als Aussage (»Ich fühle mich gerade angespannt.«) formuliert?
   Fragen sprechen die Befragten in der Regel in der zweiten Person (»du« oder »Sie«) an, während Aussagen häufig in der ersten Person (»ich«) formuliert sind. Je nach gewünschter Art der Ansprache ist die eine oder andere Form zu wählen.

Eine Skala kann im Hinblick auf einen Pol (unipolar) oder auf einen Bereich zwischen Pol und Gegenpol (bipolar) formuliert sein.

2. Soll die Skala hinsichtlich der Endpunkte unipolar oder bipolar sein?
   Unipolare Skalen umfassen geringe bis starke Ausprägungen auf einem einzigen Merkmal (z. B. von »ruhig« bis »unruhig«, »kein Lärm« bis »viel Lärm«). Bipolare Skalen reichen von einem Pol bis zu einem Gegenpol (von »ruhig« bis »angespannt« oder »links« bis »rechts«) (Abb. 2.5). Bipolare Skalen haben den Vorteil, dass die Begriffe an den Endpunkten der Skala einander wechselseitig definieren. Unipolare Skalen sind vor allem dann sinnvoll, wenn der Gegensatz eines Begriffs unklar ist (z. B. bei »schüchtern«); sie eignen sich auch zur Beurteilung von Merkmalen bzw. Variablen mit einem natürlichen Nullpunkt (z. B. Höhe des Lärmpegels).

Die Anzahl der Stufen muss zum Gegenstand passen; häufig werden 4 bis 9 Stufen verwendet.

3. Wie soll die Skala abgestuft werden?
   a. Wie viele Stufen (Zahlen) sollen verwendet werden?
      Ratingskalen können nur wenige (3 oder 4) oder viele (20 oder mehr) Stufen aufweisen. Bei einer zu geringen Anzahl von Stufen sind relevante Unterschiede zwischen verschiedenen Beurteilungen nicht abbildbar; andererseits sollte die Anzahl der Stufen auch die Differenziertheit des Urteils nicht übersteigen. In der Praxis haben sich 4- bis 9-stufige Ratingskalen gut bewährt.

Eine ungerade Anzahl von Stufen suggeriert einen neutralen Mittelpunkt, dessen Interpretation durch das Ambivalenz-Indifferenz-Problem erschwert werden kann.

   b. Wird eine gerade oder ungerade Anzahl von Stufen verwendet?
      Bei einer geraden Anzahl von Stufen (für eine bipolare Skala) wird ein Urteil in Richtung des einen oder anderen Pols der Skala quasi erzwungen. Bei einer ungeraden Anzahl von Stufen steht dagegen auch eine neutrale Urteilsmöglichkeit zur Verfügung. Ein solches neutrales Urteil ist allerdings problematisch, wenn diese Urteilsmöglichkeit zu häufig gewählt wird. Problematisch sind neutrale Urteile auch insofern, als sie nicht eindeutig interpretierbar sind: Eine neutrale Beurteilung kann ebenso Ausdruck einer gleichgültigen wie auch einer ambivalenten bzw. zwiespältigen Einstellung im Hinblick auf den fraglichen Gegenstand sein (Ambivalenz-Indifferenz-Problem).

Die Skalenstufen können numerisch, verbal und grafisch bezeichnet werden.

4. Wie werden die Abstufungen der Skalen gekennzeichnet?
   Die Bedeutung der Stufen kann durch Zahlen (numerische Marken), durch Begriffe oder Wörter (verbale Marken) und durch sprachfreie Zeichen (grafische bzw. ikonische Marken, z. B. Smileys oder Frownies) angegeben werden. Ausschließlich numerische Marken haben den Vorteil, dass sie eindeutig sind und dass die Abstände zwischen den Stufen der Ratingskala gleich sind. Verbale und grafische Marken sind dagegen leichter verständlich, vor allem für Personen, die mit Forschung wenig

	1 2 3 4 5 6 7	
Weich	o x	Hart
Heiter	o x	Traurig
Klar	x o	Verschwommen
Stark	x o	Schwach
Gut	x o	Schlecht
Kühl	x o	Gefühlvoll
Redselig	o x	Verschwiegen
Mächtig	x o	Ohnmächtig
Dominant	xo	Unterwürfig
Aktiv	x o	Passiv
Einfach	x o	Schwierig
	1 2 3 4 5 6 7	
Urteilsobjekt:	x Ingenieur o Heilpraktiker	

Abb. 2.6 Beispiel für ein semantisches Differenzial einschließlich eines fiktiven Polaritätsprofils für die Berufe Ingenieur und Heilpraktiker

vertraut sind; jedoch ergibt sich bei verbalen Marken das Problem, wie der gleiche Abstand zwischen den Stufen der Skala sichergestellt werden kann.

Semantisches Differenzial

Eine spezielle und klassische Form von Ratingverfahren ist das semantische Differenzial, das von Osgood, Suci und Tannenbaum (1957) erstmals vorgestellt wurde. Durch Antworten auf mehreren bipolaren Items entsteht ein Polaritätsprofil, das die mentale Repräsentation des beurteilten Objekts in seinen wesentlichen Merkmalen anschaulich wiedergeben soll. In Abb. 2.6 ist als fiktives Beispiel ein semantisches Differenzial einschließlich eines Polaritätsprofils für die Berufe Ingenieur und Heilpraktiker als Beurteilungsobjekte zu sehen. Diese Art der Darstellung erlaubt einen schnellen Überblick über Unterschiede auf zentralen Dimensionen (z. B. Valenz, Aktivität, Macht) zwischen mehreren Objekten.

Das **semantische Differenzial** liefert **Polaritätsprofile**, die eine schnelle Orientierung über zentrale Merkmale bzw. Unterschiede zwischen Merkmalsträgern erlauben.

Urteilstendenzen bei der Beantwortung von Ratingskalen

Bei der Interpretation von Antworten auf Ratingskalen sind die generellen kognitionspsychologischen Herausforderungen und kommunikationspsychologischen Aspekte zu berücksichtigen, die weiter oben vorgestellt wurden. Darüber hinaus sind noch weitere spezielle *Urteilstendenzen* zu beachten, die mit dem Format der Antwortskala in Zusammenhang stehen.

Antworten auf Ratingskalen können durch spezielle Urteilstendenzen beeinflusst werden.

Tendenz zur Mitte Eine Urteilstendenz, mit der stets zu rechnen ist, ist die Tendenz zur Mitte. Darunter wird die Tendenz verstanden, Extremurteile zu vermeiden. Dieser Fehler tritt besonders dann auf, wenn die Urteilsobjekte wenig bekannt sind oder wenn die Skalen an den Endpunkten nicht verankert sind (wenn die Extreme also unklar bleiben). Der Fehler lässt sich vermindern, indem die Untersuchungsteilnehmer und -teilnehmerinnen hinreichend über die zu beurteilenden Objekte informiert werden.

Gerade bei wenig vertrauten Urteilsobjekten oder Unklarheit über die Endpunkte kann eine **Tendenz zur Mitte** auftreten.

Gedankenlose Reproduktion Das Format der Skala kann Befragte auch dazu verleiten, bei ähnlich erscheinenden Items nach einer ersten Antwort bei den folgenden Fragen denselben Wert zu reproduzieren. Das Risiko einer solchen gedankenlosen Reproduktion kann durch eine geeignete Mischung von Fragen bzw. eine Umpolung der Fragerichtung (die Benennung der Pole wird vertauscht: von »angespannt … ruhig« zu »ruhig … angespannt«) verringert werden.

Eine Folge ähnlicher Items kann zur gedankenlosen Reproduktion einladen.

Primacy-Effekt Eine verwandte weitere Urteilstendenz ist der Primacy-Effekt, der auch mit der Reihenfolge der zu beurteilenden Objekte zusammenhängt. Diese Verzerrung

Wenn anfängliche Urteile folgende, ähnliche Urteile gleichsinnig beeinflussen, spricht man vom **Primacy-Effekt.**

tritt auf, wenn zuerst ein oder mehrere Objekte mit nicht neutraler bzw. extremer Merkmalsausprägung beurteilt werden. Die Beurteilung der folgenden Objekte kann dann von der Beurteilung des ersten Objekts abhängen: Bei wahrgenommener Ähnlichkeit der folgenden Objekte kommt es zu Urteilen in die gleiche Richtung, ggf. auch mit ähnlich extremer Ausprägung; bei wahrgenommener Unähnlichkeit können die Urteile durch eine Kontrastierung in die entgegensetzte Richtung verzerrt sein. Der Einfluss solcher Primacy-Effekte auf der Ebene einer ganzen Stichprobe kann verringert werden, indem die Itemreihenfolge über die Befragten hinweg systematisch variiert (interindividuell ausbalanciert) wird.

Wenn die Beurteilung eines Objekts hinsichtlich verschiedener Merkmale durch das Urteil auf einem zentralen Merkmale beeinflusst wird, spricht man vom Halo-Effekt.

Halo-Effekt Schließlich kann gerade bei Ratingskalen der sog. Halo-Effekt auftreten. Dieser Fehler bezieht sich auf die Tendenz, die Beurteilung eines Objekts hinsichtlich verschiedener Merkmale von der Beurteilung auf einem dieser Merkmale abhängig zu machen. Das Urteil für dieses eine Merkmal strahlt auf die Beurteilung der anderen Merkmale aus, es bildet sozusagen einen Hof (einen »Halo«, ähnlich dem Hof des Mondes) für die Beurteilung auf den anderen Merkmalen. Der Fehler tritt insbesondere bei der Personenbeurteilung auf: Wenn Befragte beispielsweise eine Person für gutmütig halten, können sie infolgedessen diese auch als zuverlässig, freundlich usw. beurteilen. Zu einem Halo-Effekt kommt es vor allem dann, wenn das »Schlüsselmerkmal« ungewöhnlich oder unklar definiert ist. Verringern lässt sich der Effekt, indem klare Informationen über die Unterschiede zwischen den einzelnen Merkmalen sowie über den Fehler selbst gegeben werden.

Kontrollfragen

1. Selbstauskünfte erfordern kognitive Prozesse aufseiten der Befragten. Welche wesentlichen Prozesse werden angenommen?
2. Mit welchen Einflüssen der Kommunikationssituation ist bei Selbstauskünften in Befragungen zu rechnen? Welche kommunikationspsychologischen Prozesse sollten bei der Interpretation von Selbstberichten berücksichtigt werden?
3. Wie unterscheiden sich schriftliche und mündliche Befragungen?
4. Welche Aspekte sollten bei der Formulierung von Fragen in Befragungen beachtet werden?
5. Welche wesentlichen Entscheidungen sind bei der Konstruktion von Ratingskalen zu treffen?
6. Was ist ein Rating? Welche Urteilstendenzen können die Antworten auf Ratingskalen verzerren?
7. Was ist ein semantisches Differenzial?

Weiterführende Literatur

Pelham, B., & Blanton, H. (2007). *Conducting research in psychology: Measuring the weight of smoke* (3rd ed.).Belmont: Thomson Wadworth.

Schwarz, N. (1999). *Self-reports: How the questions shape the answers*. American Psychologist, 54, 93–105.

Sudman, S., Bradburn, M. N., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

2.4 Testen

Lernziele

- Die Ziele und Eigenschaften psychologischer Tests kennenlernen.
- Klassische und probabilistische Testtheorie unterscheiden.
- Die Funktionsweise und Ziele von Leistungstests und Persönlichkeitstests verstehen.
- Einen Überblick über die Gütekriterien von Tests und Testitems gewinnen.
- Mögliche Quellen der Beeinflussung und Verfälschung von Testergebnissen sowie Gegenmaßnahmen kennenlernen.

Die bisher vorgestellten Verfahren der Datenerhebung wie Beobachten, Zählen, Messen und Befragung sind sowohl auf der Ebene von Einzelfällen als auch auf der Ebene von Gruppen (Stichproben) von mehreren Versuchsteilnehmenden einsetzbar. So kann durch eine Befragung die Einschätzung von Berufsaussichten weiblicher vs. männlicher Studierender erfasst und daraus auf die Verhältnisse in einer Grundgesamtheit geschlossen werden. Eine Befragung kann aber auch dazu dienen, die Einschätzungen der Berufsaussichten *einer* Person zu erfassen, um damit beispielsweise Hintergrundinformationen für ein Beratungsgespräch zu gewinnen. Da in der psychologischen Forschung die Suche nach allgemeinen, überindividuell gültigen Aussagen (nomothetischer Ansatz) gegenüber der Suche nach Einzelfallaussagen (idiografischer Ansatz) dominiert, steht in der Forschung die Untersuchung auf der Gruppenebene im Vordergrund. In der klinischen oder diagnostischen Praxis ist das Gewicht hingegen zugunsten der Einzelfallebene verschoben.

Die Forschung strebt vor allem nach Aussagen, die auf überindividueller Ebene (für Grundgesamtheiten bzw. Gruppen) gelten (nomothetischer Ansatz); in der klinischen oder diagnostischen Praxis spielen Aussagen über Einzelfälle (idiografischer Ansatz) eine größere Rolle.

Das Testen nimmt im Spektrum der Datenerhebungsverfahren eine Sonderrolle ein: Testen im wissenschaftlich-psychologischen Sprachgebrauch ist ein Verfahren zur möglichst genauen quantitativen Erfassung von Merkmalsausprägungen bei Individuen, vor allem ihrer überdauernden Persönlichkeitseigenschaften wie etwa Intelligenz, Neurotizismus, Gewissenhaftigkeit oder Extraversion. Das Ziel des Testens ist also in erster Linie eine Diagnose auf individueller Ebene (▶ Tests in den Anfängen der Psychologie). Testergebnisse werden zwar auch auf der Gruppenebene analysiert; jedoch dient dies meist dem gerade genannten Hauptzweck, nämlich der möglichst genauen Einschätzung individueller Merkmale. Eine solche Einschätzung erfolgt durch den Vergleich eines individuellen Testwerts mit den Testwerten geeigneter Referenzpopulationen. Um das Ergebnis eines Intelligenztests bei einer Schülerin oder einem Schüler einzuordnen, wird es beispielsweise mit den Intelligenztestwerten aller gleichaltrigen Schülerinnen und Schüler verglichen.

Ziel des **Testens** ist die präzise Erfassung von Merkmalsausprägungen von Individuen.

Definition

Ein **Test** ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch unterscheidbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst genauen quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung. Ein Test besteht in der Regel aus mehreren Aufgaben oder Fragen (Items), die von verschiedenen Menschen mit unterschiedlichen Fähigkeiten oder Eigenschaften unterschiedlich gelöst bzw. beantwortet werden. In einem abstrakteren methodischen Sinn wird ein Test auch als eine standardisierte Verhaltensstichprobe definiert, die aus Antworten auf eine Mehrzahl von Items besteht. Aus den Antworten wird der Testwert der untersuchten Person aggregiert.

▶ **Definition Test**

Exkurs

Tests in den Anfängen der Psychologie

Die Vorstellung, dass die psychischen Eigenschaften einzelner Personen durch Tests ermittelt werden können, lässt sich bis in die Anfänge der Psychologie zurückverfolgen. So formulierte etwa Galton in den 1880er Jahren eine Reihe von Aufgaben, deren Beantwortung Rückschlüsse auf die intellektuelle Begabung liefern sollte. Und Binet entwickelte im Jahr 1894 den ersten Intelligenztest für Kinder, der eine Dekade später von Binet und Simon durch eine Normierung der Aufgaben weiter verbessert wurde: Aufgaben wurden als altersnormiert angesehen, wenn drei Viertel der Kinder einer Altersklasse und nur ein Viertel der Kinder aus der darunter liegenden, jüngeren Altersklasse die Aufgaben erfolgreich bearbeiten konnten. Bei Zehnjährigen erfüllten z. B. folgende Aufgaben dieses Kriterium: Fünf Gewichte sind in eine Reihenfolge vom leichtesten (6 g) zum schwersten (18 g) zu bringen; aus drei gegebenen Worten (z. B. Wiese, werfen, Ball) sind zwei Sätze zu bilden. Aus der Menge gelöster Aufgaben ergab sich dann das »Intelligenzalter« eines getesteten Kindes, das von dessen biologischem Lebensalter mehr oder weniger stark abweichen konnte.

Man unterscheidet Leistungs- und Persönlichkeitstests.

Tests, die (vor allem kognitive) Merkmale zu einem objektiven Gütestandard in Beziehung setzen, heißen **Leistungstests**.

2.4.1 Arten von Tests

Zwei Hauptarten von Tests werden heutzutage unterschieden: Leistungs- und Persönlichkeitstests.

Leistungstests

Leistungstests erfassen Merkmale im Hinblick auf einen objektiven Maßstab zur Beurteilung der Güte der Antworten. Die Antworten können also »richtig« oder »falsch« sein. Bei der Beantwortung sind kognitive Prozesse im weitesten Sinne und Leistungsmotivation entscheidend. Zu dieser Art des Tests zählen Intelligenztests und Eignungstests. Um zwischen mehr oder weniger guten Leistungen differenzieren zu können, müssen in dem Test verschiedene Schwierigkeitsgrade realisiert sein. Dies kann auf zwei verschiedene Arten erfolgen:

- Speed-Tests und
- Power-Tests

Bei **Speed-Tests** ist die Bearbeitungszeit zu knapp angesetzt, sodass in der Regel nicht alle Aufgaben bearbeitet werden können. Hierzu zählt der »d2-Aufmerksamkeits-Konzentrationstest« von Brickenkamp (2002). Die Testpersonen bearbeiten eine große Zahl von Zeichen, nämlich die Buchstaben d und p, die jeweils mit bis zu vier kleinen Strichen umgeben sind. Anzustreichen sind nur jene d, die zwei Striche aufweisen.

Bei **Power-Tests** wird das Niveau der Aufgaben sukzessive gesteigert, so etwa bei den meisten Skalen des Intelligenztests HAWIE (Tewes, 1991).

Der Intelligenztest HAWIE Die Wechsler-Intelligenz-Tests, benannt nach dem US-amerikanischen Psychologen David Wechsler, gehören zu den bekanntesten Leistungstests. Zur Testung von Erwachsenen existieren für den deutschsprachigen Raum der »Hamburg-Wechsler-Intelligenztest« (HAWIE; Tewes, 1991) bzw. der »Wechsler-Intelligenztest für Erwachsene« (WIE; Aster, Neubauer, & Horn, 2006), die auf der »Wechsler Adult Intelligence Scale« (WAIS) basieren. Der HAWIE umfasst insgesamt 11 Untertests (Subskalen), die dem Bereich der verbalen Intelligenz oder der Handlungsintelligenz zugeordnet sind:

Verbalteil

1. Allgemeines Wissen, z. B. »Was ist der Koran?«, »Wer erfand das Flugzeug?«
2. Zahlennachsprechen: Folgen von 3–9 Ziffern sind vorwärts und rückwärts nachzusprechen, z. B. »5-8-2« oder »4-2-7-3-1-8-2«
3. Wortschatz: »Was ist die Bedeutung von ...?« z. B. »anonym, Prestige, konkor-dant«
4. Rechnerisches Denken, z. B. »Ein Zug fährt 275 km in 5 h. Wie groß ist seine Durchschnittsgeschwindigkeit in km/h?«, »Mit 8 Maschinen kann man eine Arbeit in 6 Tagen erledigen. Wie viele Maschinen sind nötig, um die gleiche Arbeit in einem halben Tag zu erledigen?«
5. Allgemeines Verständnis, z. B. »Was bedeutet das Sprichwort ‚Stille Wasser sind tief'?«, »Wenn Sie sich am Tag im Wald verirrt hätten, was würden Sie tun, um herauszufinden?«
6. Gemeinsamkeiten finden: »Was haben die folgenden Begriffe gemeinsam?« z. B. »Mantel – Anzug«, »Lob – Strafe«, »Zoo – Bücherei«

Handlungsteil

1. Bilder ergänzen: »Welcher wichtige Teil fehlt auf den Bildern«, z. B. Abb. 2.7a
2. Bilder ordnen: »Ordnen Sie die Bilder bitte so, dass sich die sinnvollste Geschichte ergibt!«, z. B. Abb. 2.7b
3. Mosaik-Test: »Legen Sie die Würfel so zusammen, dass sie ein Muster zeigen, wie das auf der Karte«, z. B. Abb. 2.7c,d

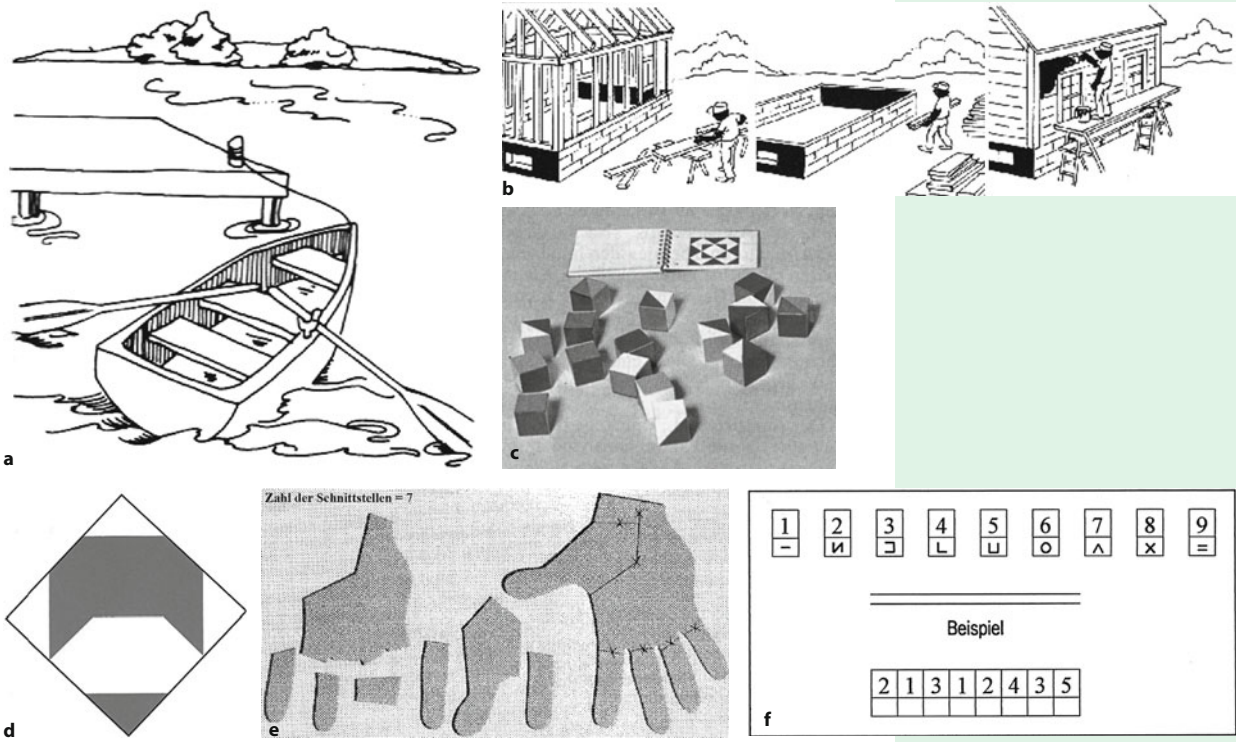

Abb. 2.7 a–f. HAWIE. Bilder ergänzen (**a**); Bilder ordnen (**b**), Mosaik-Test (**c, d**); Figurenlegen (**e**), Zahlen-Symbol-Test (**f**). Aus: *Wechsler Intelligence Scale for Children – Revised (WISC-R)*. Copyright © 1974 NCS Pearson, Inc. Reproduced with permission. All rights reserved. »*Wechsler Intelligence Scale for Children*« and »WISC« are trademarks, in the US and/or other countries, of Pearson Education, Inc. or its affiliates(s)

4. Figurenlegen: »Setzen Sie die Teile so zusammen, dass sie etwas darstellen!«, z. B. Abb. 2.7e
5. Zahlen-Symbol-Test: Symbole wie ∧ (für 7) oder × (für 8) müssen schnell einer Reihe von Zahlen zugeordnet werden; z. B. Abb. 2.7f

Der Gesamt-Intelligenzquotient (IQ) hat einen Populationsmittelwert von 100, bei einer Standardabweichung von 15 (▶ Abschn. 4.2.2). Die Reliabilität des Gesamt-IQ gilt als sehr hoch, die Validität als relativ hoch. Die Werte für die Teilbereiche sind nur bedingt aussagekräftig. Der Test differenziert nicht genügend zwischen Personen, deren IQ im oberen Bereich liegt. Im unteren und mittleren Bereich gilt der Test als wichtiges Instrument, das diagnostisch oder klinisch tätige Psychologen und Psychologinnen kennen sollten.

Persönlichkeitstests

Tests, die die Ausprägung von Eigenschaften wie Extraversion oder Offenheit erfassen, heißen **Persönlichkeitstests**.

Persönlichkeitstests liefern Daten, die nicht im Hinblick auf objektive Leistungsmaßstäbe, sondern auf emotional, motivational und sozial relevante Persönlichkeitseigenschaften hin interpretiert werden. Beispiele sind das »NEO Five Factor Inventory« (NEO-FFI; Costa & Macrae, 1992) zur Erfassung von fünf zentralen Persönlichkeitsdimensionen (Neurotizismus, Extraversion, Offenheit, Gewissenhaftigkeit und Verträglichkeit; ▶ Beispiel) oder das »Freiburger Persönlichkeitsinventar« (FPI-R, Fahrenberg, Hampel & Selg, 1994). Es wird zwischen subjektiven und objektiven Persönlichkeitstests unterschieden. Bei subjektiven Persönlichkeitstests ist der Zweck des Tests für die getesteten Personen leicht durchschaubar. Bei objektiven Persönlichkeitstests wird versucht, den Zweck zu verschleiern, um so die Reaktivität der Datenerhebung zu minimieren und die Validität der Ergebnisse zu erhöhen.

Beispiel

Der NEO-FFI

Das »NEO Five Factor Inventory« (NEO-FFI; Costa & Macrae, 1992) dient der Erfassung von fünf Persönlichkeitseigenschaften. Die Grundannahme besteht darin, dass diese fünf Eigenschaften (Faktoren) nötig und ausreichend sind, um eine Persönlichkeit zu beschreiben. Sie werden als weitgehend voneinander unabhängig angesehen. In der deutschen Fassung (Borkenau & Ostendorf, 1993) werden diese mit Neurotizismus, Extraversion, Offenheit, Gewissenhaftigkeit und Verträglichkeit bezeichnet. Die einzelnen Items werden auf einer 5-stufigen Ratingskala von »stimmt« bis »stimmt nicht« beantwortet.

1. **Neurotizismus**: Die Skala erfasst emotionale Stabilität (Ausgeglichenheit, Sorgenfreiheit) bzw. Labilität (Ängstlichkeit, Betroffenheit). Hohe Werte spiegeln das erschwerte Verarbeiten von negativen Emotionen wider.
2. **Extraversion**: Die Skala erfasst u. a. Geselligkeit, Selbstsicherheit, Aktivität, Gesprächigkeit, Heiterkeit.
3. **Offenheit**: Die Skala erfasst u. a. Wissbegierde und Interesse an neuen Erfahrungen (z. B. im Bereich von Kultur und Literatur). Personen mit hohen Werten sind eher intellektuell, phantasievoll und weniger bereit, Normen zu akzeptieren.
4. **Gewissenhaftigkeit**: Die Skala misst u. a. Zielstrebigkeit, Ehrgeiz, Ausdauer, erwünschte Kontrolle der Handlungsdurchführung und Grad der Planung im Unterschied zu Gleichgültigkeit, Nachlässigkeit und Unbeständigkeit.
5. **Verträglichkeit**: Die Skala erfasst u. a. Altruismus (vs. Egoismus), Kooperativität, Vertrauen (vs. Misstrauen). Verträglichkeit ist in der Regel sozial erwünscht.

2.4.2 Testtheorie

Klassische Testtheorie

Der **klassischen Testtheorie** zufolge gehen in einen gemessenen Testwert der wahre Wert der Person und ein Fehleranteil (Messfehler) ein. Ziel ist die möglichst direkte und präzise Schätzung des wahren Werts. Durch den Einsatz mehrerer Testitems soll der Fehleranteil insgesamt minimiert werden.

Die meisten Skalen und Tests basieren auf der klassischen Testtheorie (z. B. Gulliksen, 1950). Die klassische Testtheorie geht davon aus, dass jeder Wert einer Person auf einem konkreten Item aus zwei Komponenten zusammengesetzt ist: aus dem wahren Testwert der Person und aus einem Fehleranteil. Den wahren Testwert kann man sich als den mittleren Testwert vorstellen, den eine Person in einer unendlichen Serie von Testwiederholungen erzielen würde. Der wahre Wert ist keine praktisch erzielbare, sondern eine theoretische Größe, die durch das konkrete empirische Testergebnis einer Person geschätzt werden kann. Die Abweichung dieses empirischen Schätzwerts vom wahren Wert wird als Messfehler bezeichnet. Die klassische Testtheorie setzt somit voraus, dass wahrer Wert und Fehlerwert getrennt bestimmt werden können (▶ Axiome der klassischen Testtheorie). Der Fehleranteil am beobachteten Wert ist einer der Gründe dafür, weshalb zur Erfassung eines Merkmals mehrere Items formuliert werden: Mehrere Items ermöglichen eine bessere Annäherung an den wahren Testwert einer Person. Außerdem erlauben erst mehrere Items eine differenzierte Erfassung unterschiedlicher Merkmalsausprägungen.

Exkurs

Axiome der klassischen Testtheorie

Die Axiome (theoretisch gesetzte Vorannahmen) der klassischen Testtheorie lauten vereinfacht:

1. Ein Testergebnis besteht aus der Summe von wahrem Wert und Messfehler. Der IQ einer Person in einem Test setzt sich aus ihrem wahren Wert und Fehleranteilen (durch Ablenkung, fehlende Motivation, Störeinflüsse der Umwelt etc.) zusammen.
2. Der mittlere Messfehler ist gleich null. Bei wiederholten Testanwendungen gleichen sich die verschiedenen Messfehler sozusagen aus.
3. Der wahre Wert und der Messfehler sind voneinander unabhängig (nicht miteinander korreliert). Demnach sind z. B. Fehleranteile bei Personen mit hohem und niedrigem IQ gleich groß.

▼

4. Der Messfehler in einem Test ist nicht mit dem wahren Wert in einem anderen Test korreliert. Demnach hängt etwa das Ausmaß der Ablenkung einer Person bei einem Intelligenztest nicht mit den Werten in anderen Tests (wie z. B. Werte für Neurotizismus, Extraversion, Offenheit oder Konzentrationsfähigkeit) zusammen.
5. Die Messfehler aus verschiedenen Tests sind voneinander unabhängig (nicht miteinander korreliert). Demnach sind Personen, die bei einem Test überdurchschnittlich abgelenkt sind, bei einer Testwiederholung nicht ebenfalls überdurchschnittlich abgelenkt.

Probabilistische Testtheorie

Der **probabilistischen Testtheorie** zufolge sind Antworten auf Testitems Indikatoren von latenten Merkmalen; die Wahrscheinlichkeit, dass eine Person ein Item zur Erfassung eines Merkmals X positiv beantwortet, hängt von der (latenten) Ausprägung von X ab.

Während es in der klassischen Testtheorie um die direkte Erfassung des wahren Werts unter Berücksichtigung des Messfehlers geht, steht in der probabilistischen Testtheorie (auch: Item-Response-Theorie; Embretson & Reise, 2000; Rasch, 1980) die **Wahrscheinlichkeit** im Zentrum, mit der eine Person mit einer bestimmten Merkmalsausprägung ein Item positiv beantwortet (z. B. eine Aufgabe löst). Gemäß probabilistischen Ansätzen in der Testtheorie sind die untersuchten Merkmale latent, also nicht direkt beobachtbar; die Antworten auf Testitems stellen Indikatoren dieser latenten Merkmale dar. Die Wahrscheinlichkeit, dass eine Person ein Testitem zur Erfassung des latenten Merkmals X positiv beantwortet, hängt von der Ausprägung von X bei dieser Person ab. Eine Person mit höherer Intelligenz löst demnach ein geeignetes Item in einem Intelligenztest mit höherer Wahrscheinlichkeit als eine Person mit niedrigerer Intelligenz. Zudem löst eine Person mit größerer Wahrscheinlichkeit ein Testitem, das von vielen Personen gelöst wird, als ein Testitem, das nur von wenigen Personen gelöst wird. Um einen Test statistisch zu beschreiben, wird in der probabilistischen Testtheorie die Beziehung zwischen dem Antwortverhalten (der Lösungswahrscheinlichkeit) und dem latenten Merkmal (der Fähigkeit) von Personen für verschiedene Items bestimmt und grafisch in sog. Item-Characteristic-Curves (ICC) dargestellt (Abb. 2.8).

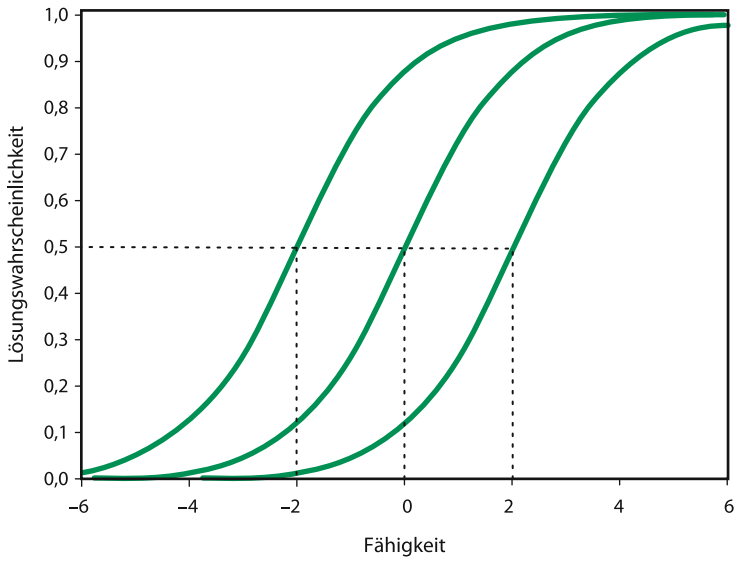

Abb. 2.8 Probabilistische Testtheorie: Item-Characteristic-Curves (ICC), die den Zusammenhang zwischen dem Antwortverhalten (der Lösungswahrscheinlichkeit) auf der y-Achse und der Ausprägung des latenten Merkmals (der Fähigkeit) auf der x-Achse für verschiedene Items darstellen

Die **Güte eines Tests** wird anhand zentraler Kriterien bewertet: Homogenität, Differenzierung, Distinktionsfähigkeit, Objektivität, Reliabilität und Validität.

2.4.3 Wünschenswerte Eigenschaften von Tests und Testitems

Ein Test besteht stets aus einer Mehrzahl von Items. Diejenigen Items, die zur Erfassung eines bestimmten Merkmals dienen, werden als **Itemsatz** bzw. als **Testskala** bezeichnet. Eine Testskala muss den nachfolgend genannten Bedingungen (Kriterien) genügen. Die ersten drei Bedingungen ergeben sich aus der generellen Anforderung der Individualdiagnose. Die letzten drei Bedingungen entsprechen den allgemeinen Gütekriterien wissenschaftlicher Datenerhebung (▶ Abschn. 1.4).

Kriterien zur Güte von Tests als Instrument zur Individualdiagnose

Homogenität Die Items einer Skala müssen homogen sein. Mit Homogenität ist gemeint, dass die Items ein einziges und nicht mehrere verschiedene Merkmale erfassen. Wenn ein Merkmal mehrere Dimensionen beinhaltet (wie vermutlich Intelligenz), dann ist es erforderlich, mehrere Testskalen zu erstellen.

Differenzierung Die Testskala soll eine differenzierte Erfassung des interessierenden Merkmals erlauben, d. h., möglichst viele Ausprägungsgrade des Merkmals erfassen.

Distinktionsfähigkeit Die Testskala muss eine eindeutige Unterscheidung zwischen Personen mit hoher und mit geringer Merkmalsausprägung ermöglichen.

Allgemeine Kriterien zur Güte wissenschaftlicher Datenerhebung

Gemäß dem Kriterium der **Objektivität** müssen Testergebnisse unabhängig von der Person sein, die den Test durchführt, auswertet und interpretiert.

Objektivität Der Test und das ermittelte Ergebnis müssen objektiv sein (▶ Abschn. 1.4). Dabei wird je nach den wichtigsten drei Phasen des Testens zwischen Durchführungsobjektivität, Auswertungsobjektivität und Interpretationsobjektivität unterschieden. Dem Kriterium der Objektivität zufolge müssen die Testergebnisse unabhängig von der Person sein, die den Test durchführt, auswertet und interpretiert.

Ein Test ist reliabel, wenn er bei wiederholter Durchführung bei denselben Personen dieselben Ergebnisse erbringt.

Reliabilität Der ermittelte Testwert muss verlässlich bzw. reliabel sein (▶ Abschn. 1.4). Das Kriterium der Reliabilität besagt, dass der Test die Merkmalsausprägung ohne zu große Schwankungen erfassen soll. Ein uneingeschränkt reliabler Test kommt bei wiederholter Durchführung bei denselben Personen zu exakt denselben Ergebnissen. Die Reliabilität für einen konkreten Test lässt sich auf verschiedene Weisen berechnen: durch die Retest-Reliabilität (Übereinstimmung bzw. Stabilität der Ergebnisse bei Wiederholung des Tests), die Paralleltest-Reliabilität (Übereinstimmung bzw. Äquivalenz von zwei parallelen Testversionen), die Testhalbierungsreliabilität (Übereinstimmung der einen mit der anderen Hälfte eines Tests) oder die interne Konsistenz (Übereinstimmung der Ergebnisse aller einzelnen Testitems untereinander, quantifiziert durch Cronbachs Alpha; Cronbach, 1951). Für weitere Erläuterungen sei auf die einschlägige Literatur verwiesen (z. B. Amelang & Schmidt-Atzert, 2006).

Ein Test ist **valide**, wenn er das erfasst, was er erfassen soll. Unterschieden werden Konstruktvalidität und Kriteriumsvalidität.

Validität Der ermittelte Testwert muss valide sein (▶ Abschn. 1.3.1). Das Kriterium der Validität betrifft die Gültigkeit des Tests: Ein Test ist valide, wenn er auch das erfasst, was er erfassen soll. Eine typische Frage, die sich bei der Überprüfung dieses Kriteriums stellt, lautet z. B., ob ein Intelligenztest tatsächlich mentale Fähigkeiten misst, die als Intelligenz gelten können, oder nur die Leistungsmotivation oder die Fähigkeit, Bildungschancen konsequent zu nutzen. Die Validität eines psychologischen Konstrukts, das einem Test zugrunde liegt (Konstruktvalidität), ergibt sich aus den Relationen zu theoretisch verwandten und entfernten Konstrukten bzw. dem Vergleich mit Tests, die diese anderen Konstrukte erfassen. Dabei bezeichnet die konvergente Validität die Übereinstimmung mit als ähnlich angenommenen Tests; die diskriminante Validität bezeichnet die Abweichung oder die Unabhängigkeit von als unähnlich angenommenen Tests. Zudem können neben anderen Tests auch externe Kriterien herangezogen

werden, um die Validität eines Tests zu bestimmen (Kriteriumsvalidität): Zum Beispiel kann der Berufserfolg als Kriterium für die Validität eines Berufseignungstests verwendet werden. Dabei kann die Übereinstimmung zwischen Test und Kriterium zum selben Zeitpunkt (simultane Kriteriumsvalidität, »concurrent validity«) oder das Zutreffen einer Vorhersage des Kriteriums aus dem Test (Validität der Kriteriumsvorhersage, »predictive validity«) unterschieden werden.

Auf der Ebene einzelner Items interessieren zudem auch die **Schwierigkeit** sowie die **Trennschärfe** eines Items.

► **Definition Schwierigkeit von Testitems**

Definition

Die **Schwierigkeit** ist der Prozentsatz aller untersuchten Personen, die das Item gelöst bzw. positiv beantwortet haben. Obwohl der Begriff einen engen Bezug zu Leistungstests aufweist, kann er doch auch auf Persönlichkeitstests angewendet werden – obgleich er in diesem Fall nicht mit dem Alltagsverständnis von Schwierigkeit übereinstimmt. Bei der Testkonstruktion wird oft eine breitere Streuung der Schwierigkeiten angestrebt, damit der Testwert über das gesamte Spektrum zwischen Personen mit verschiedenen Merkmalsausprägungen differenziert.

► **Definition Trennschärfe von Testitems**

Definition

Der **Trennschärfe** eines Items ist zu entnehmen, wie gut das gesamte Testergebnis aufgrund der Beantwortung dieses einzelnen Items vorhersagbar ist. Damit gibt die Trennschärfe an, wie gut ein einzelnes Item den gesamten Test repräsentiert. Somit erreichen Personen, die einen hohen (niedrigen) Gesamttestwert aufweisen, auf einem trennscharfen Einzelitem ebenfalls einen hohen (niedrigen) Wert.

Je größer die **Schwierigkeit**, desto kleiner die **Trennschärfe**.

Grundsätzlich sind möglichst hohe Trennschärfen wünschenswert. Die Trennschärfe hängt aber auch mit der Schwierigkeit eines Items zusammen: Je größer die Schwierigkeit, desto geringer die Trennschärfe. Wenn man eine breite Streuung der Schwierigkeiten erreichen will, muss man also – bei den besonders leichten und schweren Items – Abstriche bei der Trennschärfe hinnehmen.

Arten von Testitems

Drei verschiedene Arten von Items eines Tests lassen sich unterscheiden:

- Items mit offener Beantwortung,
- Items mit halboffener Beantwortung und
- Items mit Antwortvorgaben.

Innerhalb der **Testitems** werden Items mit offener und halboffener Beantwortung bzw. mit Antwortvorgaben unterschieden. Bei Letzteren liegen Objektivität und Reliabilität im Durchschnitt höher.

Bei **Items mit offener Beantwortung** wird der untersuchten Person keinerlei Antwortoption dargeboten. Beispielsweise wird sie nach ihren freien Assoziationen zu einem dargebotenen Reiz (z. B. einem Tintenklecks im »Rorschach-Test«) gefragt oder sie wird gebeten, die Situation genauer zu schildern, die auf einer Reizgrundlage ihrer Ansicht nach dargestellt wird (z. B. beim »Thematischen Apperzeptionstest«; Murray, 1943). Items mit offener Beantwortung schneiden in Bezug auf die Testgütekriterien der Objektivität und Reliabilität in der Regel schlechter ab als Items mit Antwortvorgaben.

Bei **Items mit halboffener Beantwortung** werden die untersuchten Personen gebeten, eine vorgegebene Formulierung mit eigenen Worten zu vervollständigen bzw. zu ergänzen. Auch bei diesen Items kann es zu Problemen bezüglich der Objektivität und der Reliabilität kommen.

Bei **Items mit Antwortvorgaben** (auch Multiple-Choice-Items genannt) werden verschiedene Antwortmöglichkeiten angeboten, zwischen denen die untersuchte Per-

son auswählen soll. Für solche Items sind Objektivität und Reliabilität vergleichsweise leichter zu gewährleisten. Items mit Antwortvorgaben sind unter den verschiedenen Arten von Testitems am weitesten verbreitet.

2.4.4 Verfälschungen und Gegenmaßnahmen

Testergebnisse können Verfälschungen und Verzerrungen unterliegen. Verfälschungen haben bei Leistungs- und bei Persönlichkeitstests jeweils verschiedene Ursachen.

Leistungstests können durch das Erraten der richtigen Antwort verfälscht werden; Gegenmaßnahmen sind der Einsatz von **Distraktoren** oder die **Ratekorrektur**.

Bei **Leistungstests** kommt es vor allem dann zu Verfälschungen, wenn es den Versuchspersonen gelingt, die richtige Antwort zu erraten. Bei Multiple-Choice-Items kann die Wahrscheinlichkeit solcher Verfälschungen verringert werden, indem neben der korrekten Antwort geeignete Distraktoren eingesetzt werden. Distraktoren sind Antwortmöglichkeiten, die auch durchaus plausibel erscheinen; sie werden nur von solchen Personen als falsch erkannt, die über das relevante Wissen verfügen. Eine weitere Möglichkeit, Verfälschungen durch Erraten richtiger Antworten entgegenzuwirken, besteht darin, bei der Auswertung des Tests eine Ratekorrektur vorzunehmen. Die Korrektur besteht darin, dass von der Punktezahl, die eine Person in dem Test erreicht hat, so viele Punkte abgezogen werden, wie die Person allein durch Erraten hätte erzielen können.

Die Ergebnisse von **Persönlichkeitstests** können verfälscht werden durch …

– das Bemühen um positive Selbstdarstellung,

– die Orientierung an sozialer Erwünschtheit,

– schematische Antworttendenzen der untersuchten Personen.

Da es bei **Persönlichkeitstests** keine objektiv richtigen Antworten gibt, kommt das Erraten solcher Antworten als Fehlerquelle natürlich nicht in Frage. Verfälschungen können bei diesen Tests im Wesentlichen aus drei Quellen resultieren:

a. Die getestete Person versucht, einen bestimmten Eindruck von der eigenen Person zu vermitteln, sich in einer bestimmten Weise darzustellen (**Selbstdarstellung**, »impression management«). Möchte eine Studierende nicht die klassische Frauenrolle erfüllen, so wird sie gezielt Antworten wählen, die sie als »hart«, »stark«, »technisch interessiert« oder »rational« erscheinen lassen.
b. Im Fall der Verfälschung durch **soziale Erwünschtheit** ist die untersuchte Person darum bemüht, sich selbst so darzustellen, wie es den (angenommenen) Erwartungen und Normen der Gruppe entspricht, die von ihrem Testergebnis erfährt. Bei dieser speziellen Form der Selbstdarstellung geht es also darum, die Anerkennung durch andere sicherzustellen. Da Normen je nach sozialer Gruppe und Interaktionskontext variieren, kann eine Selbstdarstellung im Sinne der sozialen Erwünschtheit in verschiedenen Kontexten auch ganz unterschiedlich ausfallen. Wenn sich eine Person bei einem alteingesessenen und konservativen Familienbetrieb bewirbt, wird sie sich anders darzustellen versuchen als bei einer Bewerbung bei einem Start-up-Unternehmen.
c. Zu Testverfälschungen kann es schließlich auch aufgrund von schematischen **Antworttendenzen** kommen. Manche Menschen tendieren z. B. dazu, immer möglichst neutral zu antworten; andere wiederum stimmen Testitems in der Regel zu (Ja-Sage-Tendenz), andere antworten meist ablehnend (Nein-Sage-Tendenz), wiederum andere tendieren dazu, Items zu überspringen. Solche Antworttendenzen müssen den betreffenden Personen nicht immer bewusst sein.

Bei den Arten (a) und (b) handelt es sich um prototypische Fälle von Reaktivität bei psychologischen Untersuchungen (► Abschn. 2.1.2). Die Mehrzahl der Maßnahmen zur Kontrolle der Testverfälschung richtet sich auch gegen diese beiden Arten der Verfälschung.

Eine Gegenmaßnahme zur Verringerung von Testverfälschungen besteht darin, die Antwortvorgaben auszubalancieren.

Im Folgenden werden die **wichtigsten Gegenmaßnahmen** skizziert:

Ausbalancierte Antwortvorgaben Diese Kontrolltechnik besteht darin, dass unterschiedlich gepolte Items formuliert werden, d. h. Items, bei denen eine Zustimmung mit

einer je unterschiedlichen inhaltlichen Antwortrichtung einhergeht. Zudem versucht man, die Formulierungen so zu wählen, dass beide Antwortrichtungen, beide Pole des Items, gängigen gesellschaftlichen Werten und Normen entsprechen. Die weiteren Maßnahmen dienen vorwiegend der Kontrolle von Tendenzen zur Selbstdarstellung (▶ Beispiel).

Beispiel

Ausbalancierte Antwortvorgaben

Man nehme an, es solle die Motivation zum Studium getestet werden. Die folgende Ausbalancierung dient dazu, alle drei oben genannten Fehlerquellen zu eliminieren:

- »Mit meinem Studienfach beschäftige ich mich täglich mehrere Stunden« und
- »Ich finde es wichtig, neben meinem Studium auch andere Interessen nicht zu vernachlässigen«.

Aufforderung zu korrektem Testverhalten Die Instruktion zum Test fordert die Untersuchten klar und unmissverständlich auf, die Items unbedingt ehrlich zu beantworten. Bisweilen wird unterstützend angemerkt, dass die Möglichkeit besteht, Lügen als solche zu erkennen. Jedoch stellt diese Aussage selbst eine Lüge dar; und zudem wird dadurch psychischer Druck auf die Probandinnen und Probanden ausgeübt, sodass ein solcher Zusatz ethisch bedenklich ist. Die bloße Aufforderung zu korrektem Testverhalten ist dagegen in keiner Weise ethisch problematisch.

Eine weitere Gegenmaßnahme zur Verringerung von Testverfälschungen besteht darin, die Untersuchten zu korrektem Antwortverhalten aufzufordern.

Kontrollskalen Kontrollskalen (auch »Lügenskalen« genannt) sollen die Tendenz von Personen erfassen, sozial erwünschte Antworten zu produzieren (z. B. die »Social Desirability Scale« von Crowne & Marlowe 1964). Die Items stellen Verhaltensweisen dar, die gesellschaftlich eindeutig negativ bewertet werden – die zugleich aber so häufig sind, dass es ausgesprochen unwahrscheinlich ist, dass jemand die entsprechenden Verhaltensweisen selbst tatsächlich noch nie gezeigt hat. So haben die meisten Menschen schon einmal zu einer Notlüge gegriffen, obwohl Lügen gesellschaftlich negativ sanktioniert sind. Je häufiger jemand angibt, sich »nie« in der beschriebenen Weise zu verhalten, desto höher ist seine oder ihre Tendenz zur sozialen Erwünschtheit.

Kontrollskalen erfassen anhand von Angaben zu eindeutig antinormativen, aber trotzdem geläufigen Verhaltensweisen die Tendenz von Personen, sozial erwünscht zu antworten.

Randomized-Response-Technik Die Randomized-Response-Technik (Warner, 1965) basiert auf der Hypothese, dass befragte Personen weniger zu Verfälschungen neigen, wenn sie sicher davon ausgehen können, dass ihre konkreten Antworten unbekannt sind. Mit einer Verfälschung ist insbesondere bei Berichten über peinliche oder strafbare Handlungen (Steuerhinterziehung, Alkohol am Steuer) zu rechnen. Die befragten Personen werden bei der Randomized-Response-Technik aufgefordert, den Wahrheitsgehalt ihrer Antwort bei jeder einzelnen Frage von einem nur ihnen bekannten Zufallsereignis (z. B. Augenzahl eines Würfels) abhängig zu machen. Zum Beispiel könnte eine Instruktion lauten, eine Frage nur dann wahrheitsgemäß zu beantworten, wenn sie eine Eins, Zwei, Vier oder Sechs würfeln, und die Frage in jedem Fall zu bejahen, wenn sie eine Drei oder Fünf würfeln. Da die Testleiterin bzw. der Testleiter das Ergebnis des Würfelns nicht kennt, ist bei der Auswertung nicht bekannt, ob eine Frage wahrheitsgemäß beantwortet wurde oder nicht. Da die befragte Person dies weiß, kann sie getrost je nach Zufallsereignis wahrheitsgemäß antworten. Im Unterschied zur individuellen Ebene kann auf der Ebene einer Gruppe bzw. Stichprobe die Wahrscheinlichkeit einer Verfälschung geschätzt werden. Wenn eine Stichprobe, die den Test nach der Randomized-Response-Technik bearbeitet hat (»ehrliche« Stichprobe), andere Ergebnisse liefert als eine Stichprobe, die den Test ohne diese Technik beantwortet hat (potenziell »unehrliche« Stichprobe), dann ist mit einer Verfälschung zu rechnen. Zudem kann der Anteil von Befragten, die durch das Zufallsereignis (Würfelzahl) zu einer »Ja«-Antwort veranlasst wurden, geschätzt werden, da die zugrunde liegende Zufalls-

Durch die Vorgabe einer Regel für zufällige unehrliche Antworten erlaubt die **Randomized-Response-Technik** auf der Gruppenebene die Schätzung von Antwortverfälschungen.

verteilung (z. B. Wahrscheinlichkeit einer Drei oder Fünf beim Würfeln) bekannt ist. Aufgrund der Wahrscheinlichkeitsschätzungen können die Antworten in zukünftigen Testdurchführungen entsprechend korrigiert werden.

Kontrollfragen

1. Was ist ein psychologischer Test und welche Ziele werden mit ihm verfolgt?
2. Worin unterscheiden sich die klassische Testtheorie und die probabilistische Testtheorie?
3. Was ist eine ICC und was stellt sie dar?
4. Worin bestehen die wesentlichen Unterschiede zwischen Leistungstests und Persönlichkeitstests?
5. Welche wünschenswerten Eigenschaften sollten Tests und Testitems aufweisen? Was bezeichnet die Reliabilität eines Tests? Welche Arten der Validität eines Tests können unterschieden werden?
6. Was ist mit der Schwierigkeit und der Trennschärfe von Testitems gemeint? Worin besteht ihr Zusammenhang?
7. Welche möglichen Verfälschungen können bei Tests auftreten und welche Gegenmaßnahmen können eingesetzt werden?
8. Wie funktioniert die Randomized-Response-Technik?

► Weiterführende Literatur

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4. Aufl.). Heidelberg: Springer.

Moosbrugger, H. & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion* (2. Aufl.). Heidelberg: Springer.

Steyer, R. & Eid, M. (2001). *Messen und Testen* (2. Aufl.). Berlin: Springer.

2.5 Biopsychologische und neurowissenschaftliche Messungen

Lernziele

- Die Ergänzung des Methodeninventars der Psychologie durch biopsychologische Methoden verstehen.
- Einen Überblick über Verfahren zur Messung der Aktivität außerhalb und innerhalb des Zentralen Nervensystems gewinnen.
- Die Einsatzgebiete biopsychologischer Messungen kennenlernen.
- Die Vor- und Nachteile der biopsychologischen Verfahren verstehen.
- Biopsychologische Methoden kritisch würdigen.
- Typische Fehlerquellen bei biopsychologischen Messungen kennenlernen.

Biopsychologische Verfahren versprechen eine möglichst **direkte Erfassung** psychischer Prozesse. Die Vorstellung, dem Gehirn beim Denken zuschauen zu können, übt eine große Faszination aus.

Biopsychologische und neurowissenschaftliche Messungen sind ein integraler Bestandteil des Methodeninventars in vielen Bereichen der Psychologie. Sie ergänzen zunehmend die anderen Datenerhebungsverfahren in weiten Bereichen der Psychologie. Beispielsweise spielt in der Erforschung von Stress oder Lernen die Erfassung physiologischer bzw. biologischer Prozesse, z. B. der Ausschüttung von Kortisol bei Stress oder der präsynaptischen Bahnung beim Lernen, eine zentrale Rolle. Die wachsende Bedeutung biopsychologischer Messungen liegt u. a. darin, dass sie direktere Einblicke in die zugrunde liegenden Prozesse verspricht und der psychologischen Forschung den Charakter einer »echten« Naturwissenschaft zu verleihen scheint. So üben etwa die Darstellungen des arbeitenden Gehirns durch bildgebende Verfahren (»brain imaging«) eine große Faszination aus: Noch nie zuvor konnten Menschen dem menschlichen Gehirn beim Denken zuschauen (► Die Debatte zum Verhältnis von Körper und Geist).