Einführung in die Testkonstruktion

Testverfahren, in der Regel Fragebögen, werden als Standarderhebungsinstrumente im klinischen Alltag oder zur Untersuchung neuer Fragestellungen und Forschungsgebiete eingesetzt. Sie werden in Kliniken, bei Beratungsstellen, in der staatlichen Verwaltung, in der Forensik, in Betrieben zur Personalverwaltung, im pädagogischen Bereich (Schulen), beim Militär, in der Marktforschung und Werbepsychologie verwendet (Bühner 2006). Für viele Fragestellungen, bzw. zur Erfassung von psychologisch relevanten Konstrukten, existiert eine große Anzahl von Testverfahren, z. B. Fragebögen zur Erfassung:

- von klinischen Krankheitsbildern,
- des Krankheitsverlaufs bei psychotherapeutischen Behandlungen,
- der Problemlage vor und während einer pädagogischen und psychosozialen Beratung,
- der Schuleignung,
- der Studiumseignung,
- der Potenziale zur Personalentwicklung,
- der Berufseignung durch ein Assessmentcenter (Eignungsdiagnostik) etc.

1 Einsatz von Testverfahren

Amelang (1999) bezeichnet die Testkonstruktion als eine „Schlüsselkompetenz“ für Psychologen. Diese sollten bestehende Testverfahren anhand von Gütekriterien bewerten und neue Testverfahren entwickeln können. Die große Vielfalt von Testverfahren lässt sich den drei Bereichen Leistungstests, psychometrische Persönlichkeitstests und Persönlichkeitsentfaltungsverfahren zuordnen (Brähler et al. 2002).Jeder Bereich kann weiter unterteilt werden:

- Leistungstests:
  - Entwicklungstests,
  - Intelligenztests,
  - allgemeine Leistungstests,
  - Schultests,
  - spezielle Funktionsprüfungs- und Eignungstests.
- Psychometrische Persönlichkeitstests:
  - Persönlichkeitsstrukturtests,
  - Einstellungstests,
  - Interessentests,
  - klinische Tests.
  - Persönlichkeitsentfaltungsverfahren:
  - Formdeutungsverfahren,
  - verbal-thematische Verfahren,
  - zeichnerische und Gestaltungsverfahren.

Unterteilung der Leistungstests: Bei Leistungstests wird die jeweilige Leistung erfasst, wobei es im Gegensatz zur Messung anderer Konstrukte eine Bewertung in die Kategorien „richtige“ oder „falsche“ Antwort gibt. Es werden Schnelligkeitstests (Speedtests) und Niveautests (Powertests) differenziert.

Schnelligkeitstests (Speedtests) bestehen aus leichten oder maximal mittelschweren Aufgaben, die theoretisch von jeder Person in „unendlich“ langer Zeit gelöst werden könnten. Die Bearbeitungsdauer wird jedoch so stark begrenzt, dass eine vollständige Beantwortung praktisch nicht möglich ist. Die Erfassung der Leistung erfolgt über die Auswertung der Anzahl richtig bzw. falsch / nicht bearbeitete Aufgaben, sodass Personen sich in der Anzahl der richtig bearbeiteten Items unterscheiden.

Niveautests (Powertests) beinhalten Aufgaben, deren Schwierigkeitsgrad kontinuierlich zunimmt, sodass nicht alle Items von allen Probanden richtig gelöst werden können. Die Testdurchführung erfolgt zeitunabhängig, da die Fähigkeit des Probanden und nicht die vorgegebene Zeit die Leistung begrenzt. Die Geschwindigkeit ist nicht von Bedeutung, da auch bei unendlicher Zeit nicht alle Aufgaben von einem Probanden gelöst werden könnten.

Eingesetzt werden Testverfahren zur Querschnitts- und / oder zur Längsschnittsdiagnose.

Querschnittsdiagnose: Sie dient der Ermittlung des Zustands zum Zeitpunkt der Erfassung (momentaner Zustand). Dies kann folgende Ziele haben:

- Eine Einordnung der Person in eine Gruppe bezüglich des Merkmals (z. B. durch einen Schuleignungstest).
- Ein Vergleich von mehreren Personen oder Gruppen zur Auswahl der Besten (z. B. durch einen Studieneignungstest).
- Zur Ermittlung von auffälligem Verhalten (z. B. erfüllen die Ausprägungen auf bestimmten Persönlichkeitsvariablen das Kriterium für eine Persönlichkeitsstörung).

Längsschnittsdiagnose: Veränderungen der Person werden über einen größeren Zeitraum mit zwei oder mehreren Messzeitpunkten erfasst. Beispielsweise kann der Schweregrad einer Depression im Verlauf einer Psychotherapie zu Beginn jeder Sitzung erhoben und über die Therapiedauer hinweg verglichen werden.

2 Grundlagen der klassischen Testkonstruktion

Psychologische Theorien sollen primär Konstrukte erklären, nicht nur Zustandsbeschreibungen erstellen. Die Antworten der Personen auf eine Vielzahl von zusammenhängenden Fragen spiegelt die Ausprägung auf der zugehörigen, latenten (= verborgenen) Personenvariablen wider. Diese latenten Variablen sind die zu erklärenden Konstrukte, z. B. „Intelligenz“, „emotionale Labilität“ oder auch „Studiumsmotivation“. Die Antwort auf eine Frage, ein Item eines Fragebogens, wird hingegen als beobachtbare oder manifeste Variable bezeichnet. Wichtig ist bei der Konstruktion eines Fragebogens, dass die Antworten auf verschiedene Items einer Skala systematisch zusammenhängen (hohe Korrelationen zwischen den Items). Die latente Variable soll diese systematischen Zusammenhänge zwischen den Items einer Skala erklären. So kann die Antwort eines Probanden bei jedem einzelnen Item des Fragebogens durch die Ausprägung der Person in der zugrunde gelegten latenten Variablen vorhergesagt werden. Hoch depressive Patienten (hoher Skalenwert) sollten z. B. in allen Items der Skala hohe Werte haben. Diese Voraussetzung wird als lokale stochastische Unabhängigkeit bezeichnet

Lokale stochastische Unabhängigkeit: Wird die Ausprägung der Personen einer Stichprobe in der zugrunde gelegten latenten Variablen konstant gehalten, ergibt sich zwischen den Items der Skala kein Zusammenhang mehr (Nullkorrelation). So sollten z. B. bei der Analyse von Personen mit identischer Ausprägung im Intelligenzquotienten (IQ = 100) die erhobenen Werte nicht miteinander korrelieren (s. Abbildung 1). Dies ist dadurch begründet, dass die gemeinsame Varianz der Items in der latenten Variablen zusammengefasst wird. Lokale stochastische Unabhängigkeit ist Voraussetzung für die Anwendung der klassischen und probabilistischen Testkonstruktion.

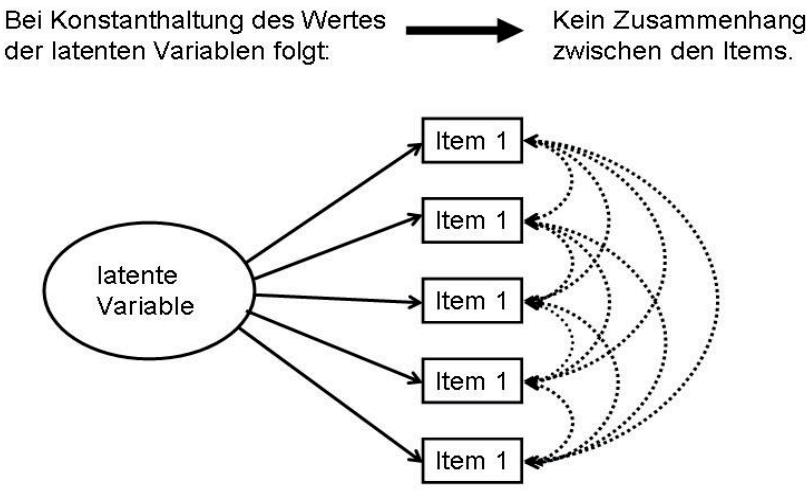

Abb. 1: Lokale stochastische Unabhängigkeit

2.1 Voraussetzungen für die Testkonstruktion

Damit ein Test nach den Regeln der Testtheorie konstruiert werden kann, muss eine Theorie zur Beschreibung von Personen (Messung der Merkmale), bzw. zu den zugrunde gelegten (theoretischen) Konstrukten vorliegen. Das zu messende latente Merkmal sollte so exakt wie möglich definiert werden. Je präziser die Definition, desto besser kann die Messung erfolgen, bzw. der Test konstruiert werden. Die interessierende latente Variable sollte das zu erfassende Merkmal möglichst hoch mit objektiven manifesten und somit leicht zu erhebenden Indikatoren korrelieren.

Die *klassische Testtheorie* (KTT) ist die Grundlage der meisten psychologischen Testverfahren (Basis für ca. 95% der Fragebögen). Sie war die erste Theorie, die zur Konstruktion von Testverfahren entwickelt wurde. Ihr großer Vorteil liegt in der einfachen Anwendbarkeit bei der Testkonstruktion. Grundlegende Kenntnisse in der KTT sind für die Entwicklung, die Bewertung und den Umgang mit Testverfahren notwendig.

Bei der KTT wird berücksichtigt, dass die *Testergebnisse* einzelner Personen bei Messungen mit einem identischen Testverfahren zu verschiedenen Messzeitpunkten variieren können. Würde z. B. einer Person mehrmals der gleiche Intelligenztest vorgelegt, dürften nach der KTT die Ergebnisse innerhalb eines bestimmten Intervalls *variieren*. Ursachen sind z. B.:

- die Verbesserung von Fähigkeiten durch Übung und / oder Transfer(z. B. Rechenleistungen oder Gedächtnisleistungen),
- der Einfluss unsystematischer (nicht erfasster) innerer (z. B. Motivation, Tagesverfassung) oder äußerer Faktoren (z. B. besseres Licht, geeigneter Testraum).

Durch die KTT werden keine Aussagen darüber gemacht, wie einzelne Items eines Fragebogens vom jeweiligen Probanden mit einem bestimmten Testwert beantwortet werden oder wie sich seine Testleistung zusammensetzt. Eine Person mit einem mittleren IQ könnte z. B. diese Leistung entweder durch mittlere Leistung in allen Items oder durch sehr gute Leistung in der ersten Testhälfte und sehr schlechte Leistung in der zweiten Testhälfte erreichen. Die klassische Testtheorie ist somit eine reine Messfehlertheorie.

2.2 Axiome der klassischen Testkonstruktion

Der beobachtete Messwert (X) einer Person setzt sich aus einem konstanten wahren Wert (T) und einem Messfehler (E) zusammen: $X = T + E$

Der Messfehler (E) ist somit die Differenz zwischen beobachtetem Testwert (X) und dem wahren Wert (T) einer Person. Dieser Messfehler repräsentiert alle unkontrollierten, unsystematischen Störeinflüsse und ist somit die Ursache für differierende Ergebnisse bei wiederholter Messung: E = X - T. Dieser Messfehler hat mehrere Eigenschaften:

- Der Mittelwert (M_E) des Messfehlers (E) ist bei unendlich vielen Messungen gleich Null. Dies gilt sowohl für die wiederholte Messung bei einer Person (I) als auch für die wiederholte Erhebung einer beliebigen Population (P):

(1a) $M_{E,I} = 0$ und
(1b) $M_{E,P} = 0$

- Es besteht kein Zusammenhang (keine Korrelation) zwischen dem Messfehler (E) und dem tatsächlichen Wert (T):

(2) $r_{E,T} = 0$

- Die Messfehler verschiedener Testverfahren (A und B) sind voneinander unabhängig:

(3) $r_{E(A), E(B)} = 0$

- Der Messfehler eines Tests A ist unabhängig vom wahren Wert eines weiteren Tests B:

(4) $r_{E(A), T(B)} = 0$

Kritisiert wird an der KTT, dass keine Aussage über das Zustandekommen von Testwerten gemacht wird. Insbesondere werden systematische Einflüsse (Verzerrungen, Bias) auf das Testergebnis nicht berücksichtigt. So kann sich bei einem Intelligenztest eine Verzerrung durch die Anzahl der zuvor schon ausgefüllten, anderen Testverfahren ergeben: Je mehr Intelligenztests ein Proband bearbeitet hat, desto „intelligenter" wird er. Meistens wird die Annahme der Eindimensionalität der Merkmale (lokale stochastische Unabhängigkeit) nicht überprüft. Erhebt z. B. ein Test zur Erfassung der „Studieneignung" die beiden Konstrukte „Motivation" und „Intelligenz", ist das Ergebnis dieses Verfahrens nicht eindimensional. Das wiederum verletzt die Voraussetzung der lokalen stochastischen Unabhängigkeit und bewirkt, dass die Messgenauigkeit des Verfahrens über- oder unterschätzt wird. Darüber hinaus wird der KTT vorgeworfen, dass die Annahme der Nullkorrelation zwischen Fehlerwerten unterschiedlicher Tests nicht zwingend sein muss und dass die Kennwerte der KTT stichprobenabhängig sind, innerhalb verschiedener Teilstichproben also differieren können.

Obwohl die KTT eine Reihe von methodischen Unzulänglichkeiten aufweist, hat sie sich in der Praxis bewährt und wird immer noch häufig angewendet.

3 Grundlagen der probabilistischen Testtheorie

Die KTT wird mehr und mehr von der probabilistischen Testtheorie (PTT) abgelöst. So sind die in der PISA-Studie verwendeten Verfahren nach PTT konstruiert. In der PTT wird ermittelt, wie die Antworten aus den verschiedenen Items begründet werden können.

Grundlegende Annahme: Die Antworten auf einzelne Items sind Indikatoren für latente Fähigkeiten oder Merkmale. Nach der Anwendung eines Testverfahrens können einerseits die Items nach ihrer Schwierigkeit geordnet und andererseits die Testpersonen nach ihrer Leistungsfähigkeit in eine Rangreihe gebracht werden. Streng genommen könnte nun definiert werden, dass ein Proband mit einem bestimmten Leistungsniveau alle Items bis zu einem bestimmten Schwierigkeitsgrad perfekt lösen müsste, während kein schwierigeres Item für ihn lösbar wäre. In der Praxis ist diese Annahme aber nicht realisierbar, da einerseits Personen mit einem niedrigen Leistungsniveau schwierige Aufgaben zufällig lösen können, während andererseits sehr gute Personen leichte Aufgaben zufallsbedingt nicht lösen (z. B. durch nachlässige Bearbeitung). Deshalb wird in der PTT für jede Antwort eine Lösungswahrscheinlichkeit p ermittelt. Diese Lösungswahrscheinlichkeit eines Items hängt von der Fähigkeit der antwortenden Person und der Schwierigkeit des jeweiligen Items ab. Fähige Personen können leichte Aufgaben mit sehr hoher Wahrscheinlichkeit lösen, während weniger fähige Personen sehr schwierige Aufgaben mit hoher Wahrscheinlichkeit nicht richtig bearbeiten. Weil hierbei nur jeweils eine Wahrscheinlichkeit ermittelt wird, wird diese Testtheorie als probabilistische Testtheorie bezeichnet. Die Lösungswahrscheinlichkeit eines Items kann auf einer Item-Characteristic-Curve (ICC) aufgetragen werden.

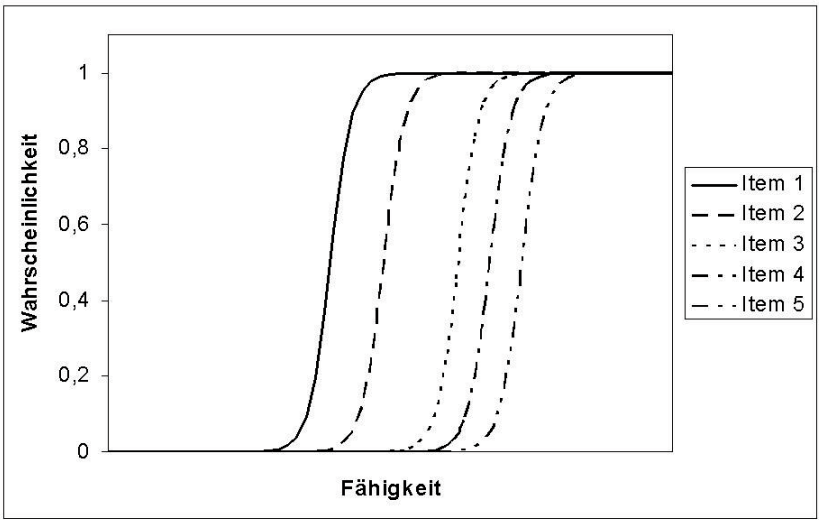

Abb. 2: Item-Characteristic-Curve (ICC)

Abbildung 2 entspricht der Funktion der Lösungswahrscheinlichkeiten eines Items. Die Y-Achse gibt die Wahrscheinlichkeit für die richtige Lösung des Items an, während die X-Achse die Fähigkeit der Personen erfasst. Je höher die Personenfähigkeit, desto größer ist die Lösungswahrscheinlichkeit.

Innerhalb dieses Theoriegebäudes gibt es eine Vielzahl probabilistischer Modelle (z. B. dichotomes Rasch-Modell, Birnbaum-Modell; mehr dazu bei Rost 2004). Die **Wahl des jeweiligen Modells** hängt vom Skalenniveau, der Anzahl der Antwortkategorien und / oder der Einführung von Rateparametern ab. Mit der Gültigkeit eines Rasch-Modells wird die Voraussetzung überprüft, dass mit dem ungewichteten Summenwert der Items die Ausprägung auf der zugrunde gelegten latenten Variablen vorhergesagt werden kann. Bei Gültigkeit des Modells ist der Summenwert einer Person eine erschöpfende Statistik für die Person und liefert Informationen über die Fähigkeitsausprägung.

Zur **Modellprüfung** in der PTT müssen ein Personenparameter (PP) und ein Itemparameter (IP) ermittelt werden, die auf einer eindimensionalen Skala abbildbar sind, wobei der Zusammenhang zwischen den Parametern probabilistisch ist. Diese Schätzung bauen auf die Schwierigkeitsindizes (p) der Items auf (p = N(r)/N). Werden die Items eines Tests auf der Item-Characteristic-Curve (ICC) aufgetragen, so können sich Hinweise für die Rasch-Homogenität der Items ergeben. Die Items sollten sich nur in der Schwierigkeit unterscheiden. Abbildung 3 zeigt eine raschhomogene Skala, während Abb. 4 eine nicht raschhomogene Skala abbildet.

Probabilistische Modelle werden häufig als stichprobenunabhängig bezeichnet, wobei diese Unabhängigkeit nur für interessierende Stichproben gilt. Dies bedeutet, dass jede beliebige Teilstichprobe aus der Gesamtstichprobe zum identischen Ergebnis der Durchführung einer Rasch-Analyse kommt.

Ein *Vergleich von KTT und PTT* ergibt nach MacDonald und Paunonen (2002) keinen entscheidenden Vorteil für eine der beiden Methoden. Bei der PTT erfolgt eine optimale Itemauswahl mit höherer Item- und Personenhomogenität. Die Nachteile der PTT sind die hohe Schwierigkeit bei der Findung modellkonformer Items und die höhere Anforderung an den Stichprobenumfang.

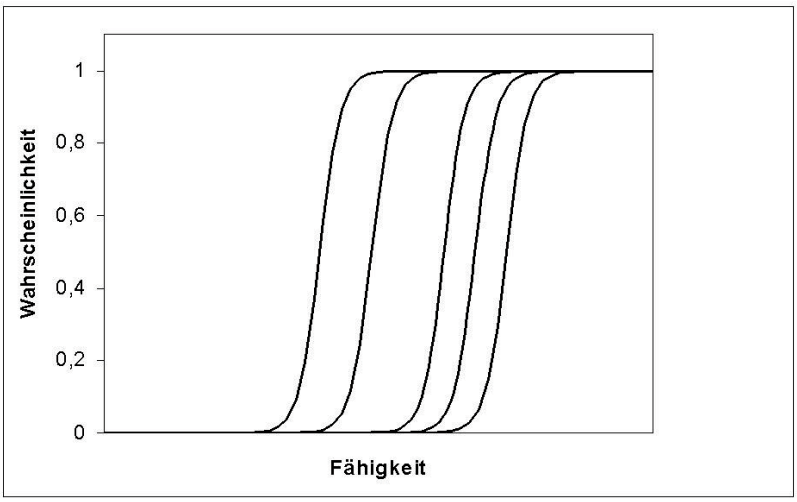

Abb. 3: ICC einer raschhomogenen Skala

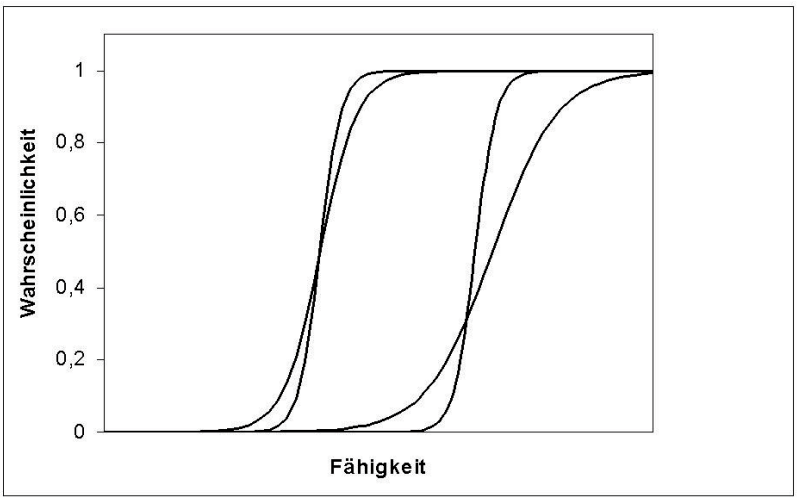

Abb. 4: ICC einer nicht raschhomogenen Skala

4 Stufen der Testentwicklung

Bei einer Testentwicklung werden im Idealfall die folgenden zehn Schritte durchgeführt:

1. Anforderungsanalyse und Problemstellung: Je nach Fragestellung wird durch Beobachtungen, mündliche Befragungen oder Befragungen mittels eines offenen Fragebogens erfasst, welche Problemstellung mit dem neuen, zu entwickelnden Fragebogen erhoben werden soll. Bezogen auf spezifische Tätigkeitsbereiche schlägt Schuler (2001) die folgenden drei Ansätze vor:

- *Erfahrungsgeleitet-intuitive Methode:* Der Entwickler beschäftigt sich mit den Eigentümlichkeiten und Besonderheiten der Tätigkeit und entwickelt ein Testverfahren aufbauend auf seinen Vor-Erfahrungen (z. B. Befragung von Experten).
- *Analytisch-empirische Methode:* Personen werden bei ihren Tätigkeiten in konkreten Situationen beobachtet und anhand dieser Beobachtungen werden Items definiert (z. B. Beobachtung von Angstpatienten im Alltag).
- *Personenbezogen-empirische Methode:* Der Zusammenhang zwischen Merkmalen der tätigen Personen und Kriterien wird erfasst (z. B. Personenmerkmale und berufliche Leistung).

2. Planung und Literatursuche: Im nächsten Schritt wird das Merkmal mit Hilfe von Überblicksartikeln oder Lehrbüchern eingegrenzt. Auch sollte überprüft werden, ob ein den Anforderungen entsprechendes Instrument entwickelt wurde. Falls es keinen theoretischen Hintergrund zur Beschreibung des Merkmals gibt, kann durch Befragung von Laien oder Experten eine anforderungsbezogene Vorform des Tests entwickelt werden. Zwar fehlt dieser Vorform das theoretische Fundament, aber durch eine breite Fassung der Begrifflichkeiten (viele Items) ist bei der Testentwicklung ein empirisches Herangehen an die Fragestellung möglich.

3. Eingrenzung des Merkmals und Arbeitsdefinition: Die in der Literaturrecherche erfassten Definitionen des Merkmals sollten verglichen werden. Falls es mehrere Theoriemodelle für ein Konstrukt gibt, ist das für die jeweilige Fragestellung geeignete auszuwählen. Werden Items nach inhaltlichen Gesichtspunkten aus bestehenden Verfahren ausgewählt oder neu gebildet, wird dies als rationale Fragebogenkonstruktion bezeichnet. Nach der deduktiven Methode werden die Items theoriegeleitet definiert. Die Merkmalsbereiche, die der Fragebogen erheben soll, werden von denen, die er nicht erheben soll, abgegrenzt. Für die verschiedenen Merkmalsbereiche, bei denen Items zu Skalen zusammengefasst werden, sollten immer gleich viele Items verwendet werden (z. B. fünf Items je Skala).

4. Testentwurf: Vor der Erstellung wird definiert:

- für welche Zielgruppe das Verfahren entwickelt werden soll,
- in welcher Form Informationen erhoben werden sollen (subjektive und / oder objektive Informationen) und
- welchen Zweck das Testverfahren hat (Gruppendiskrimination, z. B. Depression oder keine Depression, oder Eigenschaftsbeschreibungen, z. B. Intelligenzquotient).

Damit die Inhaltsvalidität des Verfahrens zufriedenstellend ist, sollte bei der Auswahl der Items auf eine repräsentative und ausreichende Itemmenge geachtet werden. Die Anzahl der Items des Testentwurfs sollte größer sein als die geplante Itemanzahl der Endversion, sodass in den folgenden Entwicklungsschritten ungeeignete Items ggf. ausgeschlossen werden können. Bei der Erstellung eines Testentwurfs können verschiedene Aufgabenformate verwendet werden (Bühner 2006; Lienert/Raatz 1994).

- **Ratingskalen:** Zur Beantwortung einer Frage können verschiedene, geordnete Antwortkategorien herangezogen werden. Im Idealfall sollte eine Ratingskala mindestes 5 Antwortmöglichkeiten haben.
- **Beispiel für eine unipolare Häufigkeitsskala:** Ich gehe ins Kino: nie – selten – gelegentlich – häufig – sehr oft.
- **Beispiel für eine bipolare Ratingskala:** Wie gut fanden Sie den Kinofilm? sehr gut – - gut – - weder gut noch schlecht – schlecht – sehr schlecht.
- **Richtig-Falsch-Aufgaben:** Bei diesem Fragentyp gibt es nur zwei Antwortmöglichkeiten. Bei Leistungstests wird eine Aufgabe richtig oder falsch gelöst; bei Persönlichkeitstests liegt entweder ein Merkmal vor oder nicht. Beispiel: Ich gehe abends gerne aus: stimmt – stimmt nicht.

Das methodische Problem bei diesen Skalen ist die hohe Wahrscheinlichkeit für eine Zufallslösung und das niedrige Skalenniveau des Antwortformats. Daher sollten in Form von Mehrfachwahlaufgaben bei einem Leistungstest mehrere Antwortkategorien dargeboten werden, wobei mehr als eine Lösung richtig sein kann. Alternativ kann eine Auswertung nach probabilistischer Testkonstruktion erfolgen (s. Abschnitt 3).

- **Zuordnungsaufgaben / Umordnungsaufgaben:** Bei diesem Aufgabenformat müssen Zeichen / Inhalte anderen Zeichen oder Inhalten zugeordnet oder Zeichen / Inhalte in die richtige Reihenfolge gebracht werden.
- **Freie Aufgabenformate:** Der Proband kann in Form von Ergänzungsaufgaben oder Kurzaufsätzen frei antworten.

Die Nachteile dieser Antwortformate sind:

- Soziale Erwünschtheit: Wenn sich eine Person beim Beantworten eines Fragebogens selbst beschreiben muss, gibt es eine allgemeine Tendenz, sich selbst positiv darzustellen (z. B. gibt kaum jemand zu, dass er Müll in den Wald wirft).
- Antworttendenzen: Manche Personen wählen generell extreme Kategorien (z. B. bei einer Skala von 1--5 entscheidet sich die Person nur die Randwerte 1 oder 5).
- Motivation: Die Exaktheit der Beantwortung hängt von der Motivation der Probanden ab. Je nach Länge und Komplexität eines Verfahrens nehmen die Motivation ab und die Tendenz zum „blinden Ankreuzen“ zu.
- Reihenfolgeeffekte: Die Antwort auf ein Item hängt von dessen Platz innerhalb des Fragebogens ab (z. B. kann der Befragte durch vorherige Items für bestimmte Themen sensibilisiert worden sein).

5. Überprüfung des Testentwurfs: Der Testentwurf wird einer ausreichend großen Stichprobe (N > 100) vorgelegt, die in ihren Merkmalen der Zielstichprobe entspricht. Zur Testentwicklung zum Thema „Erfassung der Pflegebedürftigkeit“ ist eine Stichprobe aus Psychologiestudierenden nicht geeignet.

6. Verteilungsanalyse: Nach der Datenerhebung sollte über deskriptive Statistiken und Grafiken die Verteilung jedes Items analysiert werden, damit mögliche Decken- und Bodeneffekte erkannt werden können. Durch Ermittlung des Mittelwerts und der Schiefe können die Items entdeckt werden, bei denen bevorzugt nur „Randwerte“ angekreuzt werden. Zwar sollte aus statistischer Sicht immer bei jedem Item eine Normalverteilung vorliegen, doch ist das gerade bei klinischen Fragestellungen oft nicht möglich (z. B. tritt bei der Erfassung von Zwängen das Merkmal „Zählzwänge“ zwar selten auf, dann aber mit meist hoher Ausprägung).

7. Itemanalyse und Itemselektion: Nach der Verteilungsanalyse erfolgt durch Berechnung von Itemschwierigkeit und Itemtrennschärfe die Itemanalyse. Hierdurch sollen Items, die von allen oder keinem Probanden der interessierenden Stichprobe gelöst werden, sowie Items mit geringer Trennschärfe eliminiert werden. Die Trennschärfe ist die korrigierte Korrelation (Part-whole-Korrektur) eines Items mit der Skala, sozusagen die Korrelation des Items mit der Skala ohne das Item selbst:

$$r_{j(t-j)} = \frac{r_{jt} \cdot s_t - s_j}{\sqrt{s_t^2 + s_j^2 - 2 \cdot r_{jt} \cdot s_t \cdot s_j}} \quad \text{mit}$$

$r_{j(t-j)}$: Trennschärfe des Items j bezüglich der Skala t
r_{jt}: Korrelation des Items j mit der Skala t
s_j: Streuung des Items j
s_t: Streuung der Skala t

8. Kriterienkontrolle: Durch die Bestimmung der primären Gütekriterien „Reliabilität” und „Validität” des Tests wird beurteilt, ob das Verfahren messgenau das zu erhebende Konstrukt auch wirklich erfasst. Hierzu wird das Testverfahren wiederholt eingesetzt und die Testergebnisse werden mit objektiven Kriterien verglichen, z. B. im klinischen Kontext die Übereinstimmung des Testwerts mit dem Fremdurteil eines Therapeuten(s. Abschnitt 5).

9. Revision des Tests: Anhand der bisherigen Ergebnisse sollte das Verfahren verbessert werden. Da die Alltagssprache und der theoretische Hintergrund eines Testverfahrens Veränderungen unterliegen können, sollte jeder publizierte Test in gewissen Abständen revidiert werden. Jede neue Testversion sollte, auch als Kennzeichen für eine Intensive „Pflege", wiederum einer psychometrischen Überprüfung unterzogen werden. Die Stufen der Testentwicklung werden immer wieder neu beschritten. Auch muss überprüft werden, für welche Stichproben das Verfahren geeignet ist, da eine Testanwendung in inhaltlich differierenden Stichproben (z. B. kardiologische und orthopädische Patienten) nicht immer vergleichbar ist.

10. Eichung / Cut-Off-Werte: Liegt die Endform des Testverfahrens vor, werden Normstichproben als Vergleichs- und Bewertungsgrundlage erhoben. Hierbei sollte, ggf. mit Hilfe eines professionellen Anbieters (z. B. Markforschungsinstitut), eine repräsentative und möglichst große Stichprobe (N > 1000) erhoben werden. Oft soll auch „nur" ein Cut-Off-Wert für ein Verfahren definiert werden, anhand dessen dann z. B. das Vorliegen einer psychischen Störung / Erkrankung definiert werden kann. Mit Hilfe der Daten der Eichstichprobe kann ermittelt werden, welche Skalenwerte bezogen auf die Population über- oder unterdurchschnittlich sind.

5 Haupt- und Nebengütekriterien

Bei der Bewertung eines Testverfahrens wird zwischen Haupt- und Nebengütekriterien unterschieden:

- ***Hauptgütekriterien*** sind Objektivität, Reliabilität und Validität.
- ***Nebengütekriterien*** sind Normierung, Vergleichbarkeit, Ökonomie und Nützlichkeit.

5.1 Hauptgütekriterien

Objektivität: Ein Test ist objektiv, wenn die Auswertung der Items, die Ermittlung und die Interpretation der Ergebnisse unabhängig von der Person des Durchführenden ist. Je detaillierter die Angaben im Handbuch zum jeweiligen Testverfahren, desto höher ist die Objektivität bei der Anwendung des Tests. Es wird zwischen drei Formen der Objektivität unterschieden:

- Durchführungsobjektivität: Werden z. B. detaillierte schriftliche Aufgabenstellungen auf dem Fragebogen gegeben, sodass jeder Proband die identischen Anweisungen erhält?
- Auswertungsobjektivität: Werden z. B. im Manual Angaben zum Umgang mit fehlenden Werten oder uneindeutigen Antworten gemacht?
- Interpretationsobjektivität: Erfolgt z. B. aufgrund des Testergebnisses im klinischen Alltag eine identische Diagnose bei der Interpretation durch mehrere Therapeuten?

Reliabilität: Die Reliabilität erfasst den Grad der Exaktheit, mit der eine Messung durchgeführt wird. Dabei ist die Reliabilität eines Testverfahrens unabhängig davon, was der Test wirklich erfasst. Eine hohe Reliabilität kann vorliegen, wenn ein Test bei wiederholter Messung zwar zu identischen Ergebnissen kommt, aber das Merkmal nicht wirklich erfasst. Ein Testverfahren wird als messgenau bezeichnet, wenn die Testwerte einer Person bei wiederholter Messung identisch sind. Hierzu wird ein Verfahren mit zeitlichem Abstand erneut erhoben und die Korrelation zwischen beiden Messungen ermittelt. Da die wiederholte Messung nicht immer durchgeführt werden kann, existieren weitere Methoden zur Schätzung der Reliabilität.

Koeffizienten zur Bestimmung der Reliabilität eines Testverfahrens:

- Innere Konsistenz: Der Zusammenhang zwischen den Items einer Skala (eines Konstrukts) wird ermittelt (Cronbach's Alpha). Cronbach´s Alpha ist definiert über:

$$\alpha = \frac{c}{c-1} g\left(1 - \frac{\sum_{i=1}^{j} s_i^2}{s_x^2}\right) \quad \text{mit}$$

s_i^2 = Varianz der Testitems
c = Anzahl der Testitems
s_X^2 = Varianz des Gesamtwertes der Skala

- Testhalbierungsreliabilität: Der Zusammenhang zwischen den Items wird durch Korrelation von zwei gleich langen Testteilen berechnet, wobei die Testlänge in einen Korrekturfaktor eingeht. Die Unterteilungen sind nach der Split-half-Methode (111…222…) oder der Odd-even-Methode (12121212…) möglich.

- Retest-Reliabilität: Die Korrelation zwischen den Werten zweier Messzeitpunkte wird berechnet.

- Paralleltest-Reliabilität: Die Korrelation zwischen den Ergebnissen zweier vergleichbarer Tests (z. B. Parallelform A und B) wird ermittelt, wobei beide Testverfahren dasselbe Merkmal erfassen sollen.

Validität: Validität beschreibt das Ausmaß, in dem ein Testverfahren misst, was es zu messen vorgibt. Die Validität eines Testverfahrens ist schwer erfassbar, da es sich nicht nur um eine statistische Berechnung, sondern auch um inhaltliche Überlegungen handelt. Als Kennwerte zur Bestimmung verschiedener Validitäten werden Korrelationen mit den Ergebnissen von konstruktnahen bzw. mit konstruktfremden Verfahren herangezogen. Die Berechnung der Zusammenhänge (faktorielle Validität) erfolgt über Faktorenanalysen (Leonhart 2004a).

Inhaltsvalidität: Ein Test ist inhaltsvalide, wenn er das zu messende Merkmal vollständig erfasst. Die ausgewählten Items sollen eine repräsentative Auswahl aus dem „Universum" aller Items darstellen, die das jeweilige Merkmal erfassen. Bei diesem „Repräsentationsschluss" kann kein statistischer Kennwert ermittelt werden; er basiert nur auf logischen und fachlichen Überlegungen.

Kriteriumsvalidität: Sie beschreibt die Korrelation der Testleistung in dem zu bewertenden Test mit einem oder mehreren Außenkriterien (andere, validierte Testverfahren oder objektive Daten).

Es wird zwischen verschiedenen Formen dieser Validität unterschieden:

- Vorhersagevalidität (prognostische/prädiktive Validität): Kann mit dem Verfahren zukünftiges Verhalten vorhergesagt werden (gelingt zum Beispiel die Vorhersage des Studienerfolges über den fraglichen Test)?
- Konstruktvalidität: Wie groß sind die Zusammenhänge eines Testverfahrens mit anderen konstruktverwandten (konvergenten oder konkurrente) und konstruktfremden (diskriminanten oder divergente) evaluierten Messinstrumenten? Bei der Ermittlung der konvergenten Validität sollte die Korrelation möglicht hoch (r = 1.0), bei der Ermittlung der divergenten Validität möglichst gering (r=.0) sein.

Eine geringe Objektivität führt immer zu einer schlechten Reliabilität eines Testverfahrens. Eine geringe Reliabilität hat wiederum eine geringe Validität zur Folge. Wird z. B. bei einem Testverfahren zur Erfassung der Mathematikleistung nicht die Zeit für die Bearbeitung der Aufgaben exakt erfasst, reduziert dies z. B. die Retestreliabilität. Da dies keine vergleichbare Leistung mehr erhebt, ist auch die Validität reduziert.

Die Kriteriumsvalidität hängt von der Reliabilität der beiden verwendeten Verfahren ab. Die maximale Korrelation zwischen zwei Merkmalen ist:

- $r_{max} = \sqrt{(r_{tt1} \cdot r_{tt2})}$ mit
- r_{max} = maximale Korrelation zwischen zwei Variablen oder Testverfahren,
- r_{tt1} = Reliabilität der ersten Variablen oder des ersten Testverfahrens,
- r_{tt2} = Reliabilität der zweiten Variablen oder des zweiten Testverfahrens.

Somit gilt, dass eine gute Objektivität Voraussetzung für eine gute Reliabilität, die wiederum Voraussetzung für eine gute Validität ist. Allerdings muss ein Verfahren mit guter Reliabilität nicht valide sein.

5.2 Nebengütekriterien

Normierung: Liegt eine Normierung für das Testverfahren vor, so kann mit Hilfe der Normdaten das Testergebnis einer Person mit der „Normalbevölkerung" verglichen werden. Durch die Ermittlung des Prozentranges wird die Testleistung einer Person als unter- oder überdurchschnittlich klassifiziert. Das Vorhandensein einer adäquaten Normstichprobe ist ein Kennzeichen bei der Beurteilung eines Testverfahrens.

Vergleichbarkeit: Falls ein Testverfahren in mehreren Parallelformen (Testform A und B) entwickelt wurde, ist aufgrund der Vergleichbarkeit dieser inhaltlich ähnlichen Tests eine objektivere Testung möglich. Ei-

nerseits ist bei Messwiederholung die Gefahr reduziert, dass sich Probanden an ihre vorherigen Antworten erinnern, wodurch Veränderungen besser erfasst werden können. Andererseits kann bei Gruppentestungen mit den Versionen A und B eines Testverfahrens das Abschreiben verhindert werden.

Ökonomie: Da der Einsatz eines Fragebogens für den Durchführenden immer mit einem zeitlichen und finanziellen Aufwand verbunden ist und andererseits der Proband keinen unnötig großen Aufwand haben sollte, wird unter dem Stichwort Ökonomie die Dauer der Durchführung, der Materialaufwand, die Komplexität der Handhabung, die Möglichkeiten zur Gruppentestung und die Praktikabilität der Auswertung bewertet. Bei einem Vergleich konkurrierender Verfahren sollte bei ansonsten vergleichbaren Kennwerten immer das ökonomischere Verfahren verwendet werden.

Nützlichkeit: Die Praxisrelevanz eines Verfahrens, bzw. des erhobenen Konstruktes, wird bewertet.

Zusammenfassung: Da in einer Vielzahl von Arbeitsbereichen für Psychologen Testverfahren eingesetzt werden, sind Kenntnisse der Testkonstruktion essenziell. Bei der Bewertung eines Testverfahrens müssen die Hauptgütekriterien (Objektivität, Reliabilität und Validität) sowie die Nebengütekriterien (Normierung, Vergleichbarkeit, Ökonomie und Nützlichkeit) berücksichtigt werden. Zur Entwicklung eines Testverfahrens kann die klassische oder probabilistische Testtheorie eingesetzt werden. Die probabilistische Testtheorie ist methodisch besser fundiert, hat aber höhere Anforderungen, während die klassische Testtheorie weiter verbreitet ist. Die Entwicklung eines Testverfahrens beinhaltet immer einen zeitintensiven Entwicklungsprozess über mehrere Stufen.

Literaturempfehlungen:

Sehr gute, praxisorientierte Einführung:
Bühner, M. (2006). ***Einführung in die Test- und Fragebogenkonstruktion***. München: Pearson.
Theorieorientiert mit einem Schwerpunkt auf der probabilistischen Testkonstruktion:
Rost, J. (2004). ***Lehrbuch Testtheorie*** – Testkonstruktion. Bern: Huber.
Das „klassische" Lehrbuch:
Lienert, G.A. & Raatz, U. (2004). ***Testaufbau und Testanalyse***. Weinheim: Beltz.

Übungsfragen:

1. Für welche großen Bereiche der psychologischen Diagnostik wurden Testverfahren entwickelt?
2. Welche zeitlichen Vorgaben werden bei der Durchführung eines Niveautests gemacht?
3. Zu welchem Zweck kann ein Testverfahren bei einer Querschnittsdiagnose eingesetzt werden?
4. Definieren Sie den Begriff „lokale stochastische Unabhängigkeit". Warum ist diese Unabhängigkeit eine wichtige Voraussetzung für die klassische Testkonstruktion?
5. Welche Annahmen werden in der klassischen Testkonstruktion bezüglich des Messfehlers gemacht?
6. Wann ist nach der probabilistischen Testtheorie eine Skala raschhomogen?
7. Beschreiben Sie am Beispiel der Entwicklung eines Fragebogens zur Erfassung der Studienmotivation die Stufen der Testkonstruktion.
8. Wie kann z. B. überprüft werden, ob eine Person mit ausreichender Motivation und Exaktheit einen Fragebogen beantwortet hat?
9. Definieren Sie den Begriff part-whole-korrigierte Trennschärfe.
10. Worin unterscheidet sich bei der Berechnung der Testhalbierungsreliabilität die Split-half- von der Odd-even-Methode?
11. Warum kann die Reliabilität eines Testverfahrens leichter erfasst werden als die Validität?

Antworten:

1. Es gibt die folgenden Bereiche für Testverfahren: Leistungstests, psychometrische Persönlichkeitstests und Persönlichkeitsentfaltungsverfahren.

2. Bei einem Niveautest werden theoretisch keine zeitlichen Vorgaben gemacht. Der Test kann von einer Person mit einer bestimmten Fähigkeit bis zu einem bestimmten Schwierigkeitsgrad gelöst werden. Schwierigere Aufgaben sind allerdings nicht von dieser Person lösbar.

3. Die folgenden Ziele sind möglich:
   - Einordnung einer Person in eine Gruppe bezüglich des Merkmals.
   - Ein Vergleich von mehreren Personen oder Gruppen zur Auswahl der Besten.
   - Zur Ermittlung auffälligen Verhaltens.

4. Lokale stochastische Unabhängigkeit ist gegeben, wenn die Ausprägung der Personen einer Stichprobe in der zugrunde gelegten latenten Variablen konstant gehalten wird und dann zwischen den Items der Skala kein Zusammenhang mehr vorliegt (Nullkorrelation). Liegt lokale stochastische Unabhängigkeit vor, so kann davon ausgegangen werden, dass die Skala eindimensional ist.

5.
   - Der Mittelwert (M_E) des Messfehlers (E) ist bei unendlich vielen Messungen gleich Null.
     (1a) $M_{E,I} = 0$ und
     (1b) $M_{E,P} = 0$
   - Es besteht kein Zusammenhang (keine Korrelation) zwischen dem Messfehler (E) und dem tatsächlichen Wert (T):
     (2) $r_{E,T} = 0$
   - Die Messfehler verschiedener Testverfahren (A und B) sind voneinander unabhängig:
     (3) $r_{E(A), E(B)} = 0$
   - Der Messfehler eines Tests A ist unabhängig vom wahren Wert eines weiteren Tests B:
     (4) $r_{E(A), T(B)} = 0$

6. Eine Skala ist raschhomogen, wenn für jede Ausprägung des Personenparameters die Items der Skala in eine identische Rangreihe gebracht werden können. Die Items sollten sich nur bezüglich der Schwierigkeit unterscheiden. Bei der Testentwicklung wird die Rasch-Homogenität mittels statistischer Kennwerte geprüft.

7.
- Anforderungsanalyse und Problemstellung: Was soll der Test können? Soll eine Einzel- oder eine Gruppentestung durchgeführt werden? Soll „nur" ein Screening durchgeführt werden?
- Planung und Literatursuche: Wurden schon Verfahren entwickelt, die das Konstrukt Studienmotivation erfassen?
- Eingrenzung des Merkmals und Arbeitsdefinition: Wie wird das Konstrukt definiert?
- Testentwurf: Entwicklung eines Prototypen-Fragebogens.
- Überprüfung des Testentwurfs: explorativer Einsatz des Tests an einer kleinen Stichprobe.
- Verteilungsanalyse: Untersuchung der deskriptiven Kennwerte der einzelnen Items zur Prüfung, ob Decken- oder Bodeneffekte vorliegen und der Verteilungsform (Schiefe).
- Itemanalyse und Itemselektion: Ausschluss von ungeeigneten Items.
- Kriterienkontrolle: Überprüfung, ob der Test wirklich misst, was er zu messen vorgibt.
- Revision des Tests: Überarbeitung des Tests.
- Eichung: Einsatz der Endversion des Tests an einer Normierungsstichprobe (N > 1000).

8. Durch die Messung der Beantwortungszeit und durch die Verwendung sog. „Lügenitems". Bei diesen Items wird die Aufrichtigkeit der Testpersonen erfasst (z.B.: „Ich habe noch nie im absoluten Parkverbot geparkt.").

9. Die part-whole-korrigierte Trennschärfe ist die Korrelation eines Items mit der Skala, wobei der Itemwert selbst nicht in die Berechnung der Skala eingeht.

10. Bei der Split-half-Methode wird der Mittelwert der Items der ersten Testhälfte mit dem Mittelwert der Items der zweiten Testhälfte korreliert. Bei der Odd-even-Methode werden die Items zunächst abwechselnd in zwei Gruppen aufgeteilt (121212…), bevor dann die beiden Mittelwerte korreliert werden.

11. Die Reliabilität kann mittels statistischer Kennwerte erfasst werden. Die Erfassung der Validität ist allerdings schwieriger, da es sich oft um Augenscheinkriterien handelt und es inhaltlicher Bewertung bedarf.