Anzahl richtig gelöster Aufgaben die durch die Anzahl der Distraktoren (nicht die Anzahl der Antwortalternativen) dividierte Fehleranzahl abzieht:

$$x_{corr} = N_R - \frac{N_F}{k-1},$$

mit k: Anzahl der Antwortalternativen.

Beispiel: Bei 100 Items mit jeweils k=4 Antwortvorgaben wird ein ratender Proband ca. N_R=25 Items zufällig richtig und N_F=75 Items zufällig falsch beantworten. Er hätte damit eine korrigierte Punktzahl von Null:

$$x_{corr} = 25 - 75/(4-1) = 0.$$

Wenn pro Item mehrere Antwortvorgaben richtig sind, können Rateeffekte neutralisiert werden, wenn man für jedes richtige Ankreuzen einen Pluspunkt, für jedes falsche Ankreuzen einen Minuspunkt und für jede nicht angekreuzte Antwortvorgabe keinen Punkt vergibt. Auch hier sollten jedoch die Untersuchungsteilnehmer zuvor darüber informiert werden, in welcher Weise in der Auswertung Falschantworten berücksichtigt werden.

Weitere Ratekorrekturen bei Aufgaben mit vorgegebenen Antwortmöglichkeiten diskutiert Barth (1973). Schaefer (1976) weist auf Möglichkeiten einer probabilistischen Auswertung von Mehrfachantworten hin. Diese läuft auf eine Anwendung des sog. dreiparametrigen logistischen Modells hinaus (kurz: Birnbaum-Modell, vgl. hierzu Rost, 2004, S. 133, oder Baker & Kim, 2004, S. 18 ff.). Eheim (1977) geht der Frage nach, ob die Wahrscheinlichkeit einer richtigen Antwort bei Mehrfachwahlaufgaben von der Position der richtigen Alternative innerhalb der vorgegebenen Alternativen abhängt. Die Frage kann verneint werden. Weniger eindeutig sind die Ergebnisse einer Studie von Buse (1977), der die Abhängigkeit der Testreliabilität von Rateeinflüssen überprüfte. Die Bedeutsamkeit des Ratens für die Reliabilität hängt demnach von der Testlänge, der Trefferwahrscheinlichkeit und der Personenquote, die zum Raten aufgefordert wurde, ab. Jaradat und Tollefson (1988) zeigen, dass die Testreliabilität von der Art der Ratekorrektur unabhängig ist.

Häufig verwendet man auch in Persönlichkeits- und Einstellungstests Items mit mehreren Antwortalternativen. Diese stellen jedoch keine richtigen oder falschen Antwortmöglichkeiten dar, sondern Antwortalternativen, die es dem Untersuchungsteilnehmer erleichtern, bei Meinungsfragen oder subjektiven Einschätzungen seine Positon zum Ausdruck zu bringen. Hierbei erübrigen sich natürlich Ratekorrekturen.

Schwierigkeiten bereiten jedoch Items, die neben der Antwortalternative ja – nein (stimmt – stimmt nicht etc.) eine dritte **neutrale Kategorie** »weiß nicht« (»unentschieden«) vorgeben. Derartige Tests sind schwer auswertbar, wenn viele Untersuchungsteilnehmer – u. U. auch noch aus verschiedenen Motiven – die neutrale Kategorie wählen. Wenn möglich, sollte man derartige Itemkonstruktionen vermeiden oder zumindest durch eine entsprechende Instruktion in ihrer Bedeutung präzisieren, indem man deutlich macht, ob die Mittelposition ausdrückt, dass (a) der Proband etwas nicht weiß, (b) er sich unsicher ist, (c) er die Frage nicht beantworten möchte, oder (d) er zwischen mehreren Antworten schwankt. Erscheint die Verwendung von Mittelkategorien unumgänglich, empfiehlt sich eine Analyse bzw. Revision des Testinstruments nach einem von Heller und Krüger (1976) vorgeschlagenen Verfahren (vgl. hierzu auch die Ausführungen zum Ambivalenz-Indifferenz-Problem auf ▶ S. 180).

Itemanalyse

Die Qualität eines Tests oder Fragebogens ist abhängig von der Art und der Zusammensetzung der Items, aus denen er besteht. Die Itemanalyse (Aufgabenanalyse) ist deswegen ein zentrales Instrument der Testkonstruktion und Testbewertung, in deren Verlauf die psychometrischen Itemeigenschaften als Kennwerte bestimmt und anhand vorgegebener Qualitätsstandards beurteilt werden. Grundlage der Itemanalyse sollte nach Möglichkeit eine sog. **Eichstichprobe** sein, d. h. ein Miniaturabbild genau jener Population, für die der Test konzipiert ist. So führt man die Itemanalyse für einen Test zur Gedächtnisleistung im Alter am besten an einer Stichprobe älterer Probanden durch und nicht etwa an Studenten.

Der Begriff »Itemanalyse« ist in der Literatur nicht eindeutig festgelegt. Meistens werden – bei »klassischen« Testkonstruktionen – die Analyse der Rohwerteverteilung, die Berechnung von Itemschwierigkeit, Trennschärfe und Homogenität sowie die Dimensionalitätsüberprüfung zur Itemanalyse gezählt (zur Durchführung einer Itemanalyse mit SPSS vgl. Bühner, 2004, Kap. 3.4 und 3.5). Für Tests, die nach einem probabilistischen

Testmodell wie z. B. dem dichotomen logistischen Modell von Rasch (1960) konstruiert wurden, erübrigt sich eine Itemselektion auf der Basis der Itemanalyse. Die Selektion erfolgt über Modelltests, die die Verträglichkeit der Items mit den Modellannahmen überprüfen.

Rohwerteverteilung. Die Häufigkeitsverteilung der Testwerte (grafisch darstellbar als Histogramm) vermittelt einen ersten Überblick über das Antwortverhalten der untersuchten Probanden. Am Histogramm ist z. B. abzulesen, wie stark die Testergebnisse streuen, d. h., ob sie den gesamten Wertebereich ausfüllen oder sich um bestimmte Werte konzentrieren. Häufig interessiert man sich dafür, ob die Rohwerteverteilung einer Normalverteilung entspricht. Normalverteilte Testwerte sind erstrebenswert, weil viele inferenzstatistische Verfahren normalverteilte Werte voraussetzen. Ob die empirisch gefundene Verteilung überzufällig von einer Normalverteilung abweicht oder nicht, kann mit dem sog. Goodness-of-Fit-Chiquadrattest (vgl. Bortz, 2005, S. 164 ff.) oder mit dem Kolmogoroff-Smirnov-Test (vgl. Bortz et al., 2000, S. 319 ff., oder Bortz & Lienert, 2003, S. 226 ff.) überprüft werden (zur Problematik dieses Anpassungstests ▶ S. 650 ff. zum Stichwort »Nullhypothesen als Wunschhypothesen«).

Intelligenztests beispielsweise sind extra so angelegt, dass sie normal verteilte Ergebnisse produzieren, was im Einklang steht mit der inhaltlichen Vorstellung, dass die meisten Menschen mittlere Intelligenz aufweisen, während extrem hohe oder extrem niedrige Intelligenz nur selten auftritt. Nicht bei allen Konstrukten ist in dieser Weise »von Natur aus« mit normal verteilten Merkmalsausprägungen zu rechnen. Bei der Erfassung von Lebenszufriedenheit ist z. B. davon auszugehen, dass die meisten Menschen nicht etwa mittelmäßig, sondern eher zufrieden sind.

Stellt sich heraus, dass die Rohwerteverteilung von einer Normalverteilung abweicht, sind folgende Konsequenzen in Erwägung zu ziehen:

- Sofern aus theoretischer Sicht normalverteilte Merkmalsausprägungen zu erwarten sind, modifiziert man die Itemzusammensetzung des Tests in der Weise, dass die revidierte Version normal verteilte Ergebnisse produziert.
- Ist die Nichtnormalverteilung der Testwerte theoriekonform, kann der Test unverändert bleiben. Allerdings muss die statistische Auswertung (z. B. Gruppenvergleiche) auf die Verletzung der Normalverteilungsvoraussetzung abgestimmt werden. Zwei Strategien sind möglich: Entweder man operiert mit größeren Stichproben (ab ca. 30 Untersuchungsobjekten), wodurch sich die Forderung nach normalverteilten Messwerten in der Regel erübrigt (vgl. Bortz, 2005, S. 93 f.), oder man verwendet (vor allem bei kleinen Stichproben) statt der »normalen« (verteilungsgebundenen) statistischen Verfahren die sog. verteilungsfreien Analysetechniken (vgl. Bortz & Lienert, 2003).

Über mögliche Ursachen nicht normal verteilter Testwerte und nachträgliche Normalisierungsverfahren berichten z. B. Lienert und Raatz (1994, Kap. 8 und 12).

Itemschwierigkeit. Items besitzen unterschiedliche Lösungs- bzw. Zustimmungsraten, die als Itemschwierigkeiten (Itemschwierigkeitsindizes) quantifizierbar sind. Schwierige Items werden nur von wenigen Probanden bejaht bzw. richtig gelöst. Bei leichten Items kommt dagegen fast jeder zum richtigen Ergebnis. Die Itemschwierigkeiten beeinflussen also ganz wesentlich die Verteilung der Testwerte. Der Schwierigkeitsindex wird für jedes Item eines Tests bzw. eines Itempools einzeln berechnet, wobei zwischen zweistufigen (dichotomen) und mehrstufigen (polytomen) Antwortalternativen zu unterscheiden ist.

Die Itemschwierigkeit wird durch einen Index gekennzeichnet, der dem Anteil derjenigen Personen entspricht, die das Item richtig lösen oder bejahen.

Bei dichotomen Antwortalternativen erhält man die Schwierigkeit (p_i) von Item i, indem die Anzahl der richtigen Lösungen bzw. Zustimmungen (R) durch die Gesamtzahl der Antworten (N) dividiert wird; die Schwierigkeit entspricht damit dem Anteil der »Richtiglöser« oder »Zustimmer« für das betrachtete Item:

$$p_i = \frac{R_i}{N_i}.$$

Ein Schwierigkeitsindex von p_i=0,5 besagt, dass das Item von 50% der Untersuchungsteilnehmer richtig gelöst (bzw. bejaht) und von 50% falsch beantwortet (bzw.

verneint) wurde (Fisseni, 1990, S. 30 ff.; Lienert & Raatz, 1994).

Für mehrstufige Items lässt sich eine Formel anwenden, nach der die Summe der erreichten Punkte (x_i) auf Item i durch die maximal erreichbare Punktsumme dieses Items zu dividieren ist (Dahl, 1971, S. 140 f.). Die maximal mögliche Punktsumme ergibt sich als Produkt der maximalen Punktzahl (k), die eine Person auf Item i erreichen kann, und der Anzahl der antwortenden Personen (n).

$$p_i = \frac{\sum_{m=1}^{n} x_{im}}{k_i \cdot n}$$

Aus dieser Definition der Itemschwierigkeit folgt ein Wertebereich von 0 (schwerstes! Item) bis 1 (leichtestes! Item). Bei dem leichtesten Item erreichen alle Probanden theoretisch die maximale Punktzahl, während beim schwersten Item niemand einen Punkt erhält. Bei Ratingskalen (z. B. nie–selten–gelegentlich–oft–immer; ▶ S. 177) ist darauf zu achten, dass die unterste Kategorie nicht mit Eins, sondern mit Null kodiert wird. Die übrigen vier Kategorien erhalten dementsprechend die Werte 1 bis 4.

Beispiel: Angenommen, die Schwierigkeit von Item 10 eines Persönlichkeitsfragebogens (»Ich halte mich gerne im Freien auf«) soll ermittelt werden. Das Item ist auf einer Ratingskala von 1 (stimmt gar nicht) bis 5 (stimmt völlig) zu beantworten. Diese Ratingskala ist zunächst umzukodieren mit 0 für »stimmt gar nicht« bis hin zum Wert 4 für »stimmt völlig«. Es wird eine Stichprobe von z. B. 80 Probanden befragt. Folglich sind für die gesamte Gruppe maximal 4×80=320 Punkte auf Item 10 erreichbar, sofern alle Probanden dem Item völlig zustimmen und minimal 0×80=0 Punkte, wenn alle Probanden die Kategorie »stimmt gar nicht« wählen. Addiert man nun die empirisch gefundenen Punktwerte für dieses Item, könnte sich z. B. ein Wert von 280 ergeben. Setzt man diese empirische Punktzahl mit der theoretisch maximal erreichbaren in Beziehung, ergibt sich ein Quotient von 280/320=0,875. Es handelt sich also um ein recht leichtes Item, dem – in der untersuchten Stichprobe – überwiegend zugestimmt wird.

Extrem schwierige Items, denen kaum jemand zustimmt, oder extrem leichte Items, die von fast allen Probanden gelöst werden, sind wenig informativ, da sie keine Personenunterschiede sichtbar machen. Damit ein Test Untersuchungsteilnehmer mit unterschiedlichen Fähigkeiten annähernd gleich gut differenziert, ist darauf zu achten, dass die Items eine möglichst breite Schwierigkeitsstreuung aufweisen. Im Allgemeinen werden Itemschwierigkeiten im mittleren Bereich (zwischen 0,2 und 0,8) bevorzugt. Zur Kennzeichnung eines Tests wird oftmals auch die durchschnittliche Itemschwierigkeit angegeben.

Trennschärfe. Die Trennschärfe bzw. der Trennschärfekoeffizient gibt an, wie gut ein einzelnes Item das Gesamtergebnis eines Tests repräsentiert. Die Trennschärfe wird für jedes Item eines Tests berechnet und ist definiert als die Korrelation der Beantwortung dieses Items mit dem Gesamttestwert. Da in den additiven Gesamttestwert auch das betrachtete Item selbst eingeht – was die Korrelation künstlich erhöht – werden üblicherweise sog. korrigierte Trennschärfekoeffizienten auf der Basis von Gesamttestwerten berechnet, die das aktuelle Item unberücksichtigt lassen (vgl. Fisseni, 1990, S. 40 f.; Lienert & Raatz, 1994).

Der zu berechnende Korrelationskoeffizient richtet sich nach dem Skalenniveau der Testwerte (vgl. Bortz, 2005, Tab. 6.11). Bei intervallskalierten Testscores wählt man als Trennschärfe (r_{it}) die Produkt-Moment-Korrelation zwischen den Punktwerten pro Item i und dem korrigierten Gesamttestwert t:

$$r_{it} = \frac{\mathrm{cov}(i,t)}{s_i \cdot s_t}$$

Der Begriff »Trennschärfe« ist so zu verstehen, dass Personen, die im Gesamtergebnis einen hohen Wert erreichen, auf einem trennscharfen Einzelitem ebenfalls eine hohe Punktzahl aufweisen. Umgekehrtes gilt für Personen mit niedrigem Testergebnis. Nach diesem Verständnis lässt sich an einem trennscharfen Einzelitem bereits ablesen, welche Personen bezüglich des betrachteten Konstrukts hohe oder niedrige Ausprägungen besitzen. Beide Gruppen werden durch das Item also gut voneinander »getrennt«.

Der Trennschärfe eines Items ist zu entnehmen, wie gut das gesamte Testergebnis aufgrund der Beantwortung eines einzelnen Items vorhersagbar ist.

Ursprünglich wurde statt des oben dargestellten Trennschärfekoeffizienten ein Trennschärfeindex verwendet, der sich aus der Mittelwertedifferenz der beiden Extremgruppen (Gruppe 1: 25% der Untersuchungsteilnehmer mit den höchsten Testwerten, Gruppe 2: 25% der Untersuchungsteilnehmer mit den niedrigsten Testwerten) berechnet und am Standardfehler relativiert wird (Schnell et al. 1999, S. 183 f.). Diese Formel entspricht genau dem t-Test für unabhängige Stichproben (vgl. Bortz, 2005, Kap. 5.1.2).

Grundsätzlich sind möglichst hohe Trennschärfen erstrebenswert: Beim Trennschärfekoeffizienten mit einem korrelationstypischen Wertebereich von –1 bis +1 sind positive Werte zwischen 0,3 und 0,5 mittelmäßig und Werte größer als 0,5 hoch (Weise, 1975, S. 219), während bei dem nach oben unbeschränkten Trennschärfeindex Werte größer als 1,65 zur Auswahl des Items führen (vgl. Schnell et al., 1999, S. 183). Items mit geringer Trennschärfe, die Informationen generieren, die nicht mit dem Gesamtergebnis übereinstimmen, sind als schlechte Indikatoren des angezielten Konstrukts zu betrachten und aus einem eindimensional angelegten Test zu entfernen. Es ist zu beachten, dass die Trennschärfe eines Items von seiner Schwierigkeit abhängt: Je extremer die Schwierigkeit, desto geringer die Trennschärfe. Bei sehr leichten und sehr schweren Items wird man deshalb Trennschärfeeinbußen in Kauf nehmen müssen. Items mit mittleren Schwierigkeiten besitzen die höchsten Trennschärfen.

Homogenität. Alle Items eines eindimensionalen Instruments stellen Operationalisierungen desselben Konstrukts dar. Entsprechend ist zu fordern, dass die Items untereinander korrelieren. Die Höhe dieser wechselseitigen Korrelationen nennt man Homogenität. (Die Auswahl des geeigneten Korrelationskoeffizienten hängt auch hier wiederum vom Skalenniveau der Items ab.) Korreliert man alle k Testitems paarweise miteinander, ergeben sich k(k–1)/2 Korrelationskoeffizienten ($r_{ii'}$), deren Durchschnitt ($\overline{r}_{ii'}$) die Homogenität des Tests quantifiziert (zur Berechnung einer durchschnittlichen Korrelation vgl. Bortz, 2005, S. 219). Mittelt man dagegen nur die Korrelationen eines Items mit allen anderen Items, erhält man die itemspezifische Homogenität. Bei der Homogenitätsberechnung werden die Autokorrelationen (Korrelation eines Items mit sich selbst) außer Acht gelassen.

Tab. 4.14. Iteminterkorrelationsmatrix eines Tests

	Item 1	Item 2	Item 3	Item 4	
Item 1	1,00	0,05	0,17	0,12	0,11
Item 2	0,05	1,00	0,42	0,37	0,28
Item 3	0,17	0,42	1,00	0,54	0,38
Item 3	0,12	0,37	0,54	1,00	0,34
Homogenitäten	0,11	0,28	0,38	0,34	**0,27**

Die Homogenität $\overline{r}_{ii'}$ gibt an, wie hoch die einzelnen Items eines Tests im Durchschnitt miteinander korrelieren. Bei hoher Homogenität erfassen die Items eines Tests ähnliche Informationen.

Beispiel: Die Homogenität eines aus vier Items bestehenden Tests soll ermittelt werden. Die Iteminterkorrelationen des Tests sind in einer symmetrischen Korrelationsmatrix (Tab. 4.14) darstellbar, d. h., oberhalb und unterhalb der Diagonale mit den Autokorrelationen befinden sich dieselben Elemente, nämlich hier die 4×3/2=6 Interkorrelationen der vier Testitems. Mittelt man diese Interkorrelationen spaltenweise oder zeilenweise, ergeben sich die itemspezifischen Homogenitäten. Der Mittelwert der Itemhomogenitäten ist die Testhomogenität, die mit 0,27 in diesem Beispiel eher gering ausfällt. Wie man sieht, weist Item 1 mit Abstand die geringste Homogenität auf (0,11), sodass es nahe liegt, Item 1 aus dem Test zu entfernen bzw. durch ein homogeneres Item zu ersetzen. Die Testhomogenität erhöht sich nach Entfernen von Item 1 auf 0,44. (Im Beispiel wurden zur Vereinfachung »normale« arithmetische Mittelwerte verwendet, die hier nur geringfügig von der für Korrelationskoeffizienten vorgesehenen Durchschnittsberechnung abweichen.)

Bei eindimensionalen Instrumenten sind hohe Homogenitäten erstrebenswert. Briggs und Cheek (1986, S. 115) schlagen zur Bewertung von Gesamttesthomogenitäten einen Akzeptanzbereich von 0,2 bis 0,4 vor. Innerhalb dieses Bereiches soll eine hinreichende Homogenität gewährleistet sein, ohne dass gleichzeitig die inhaltliche Bandbreite des gemessenen Konstrukts durch übermäßige Redundanz zu sehr eingeschränkt wird. Die mittlere Iteminterkorrelation geht in den zur Reliabilitätsschätzung verwendeten Alphakoeffizienten

von Cronbach ein (▶ S. 198 f. bzw. genauer Bortz, 2005, Gl. 15.84). Zuweilen wird deshalb der Alphakoeffizient auch als Homogenitätsindex bezeichnet. Es ist zu beachten, dass sich Alpha nicht nur mit wachsender Iteminterkorrelation, sondern auch mit steigender Itemzahl erhöht. Eine Homogenität von 0,5 produziert z. B. bei 10 Items ein Alpha von 0,9 (vgl. Schnell et al., 1999, S. 147).

Items, die wegen auffallend geringer itemspezifischer Homogenität offensichtlich etwas anderes messen als die übrigen Items, sollten aus dem Test entfernt werden. Lassen sich, evtl. unter Zuhilfenahme einer Clusteranalyse oder einer Faktorenanalyse über die Iteminterkorrelationen (▶ Anhang B), mehrere homogene Itemcluster identifizieren, die sich theoretisch klar interpretieren lassen, empfiehlt sich die Konstruktion eines Tests mit mehreren, aus diesen Items bestehenden Untertests. Die aus mehreren Subtests bestehende **Testbatterie** (bzw. Testsystem, mehrdimensionaler Test) führt dann nicht mehr zu *einem* Gesamtergebnis, sondern zu mehreren Testwerten eines Untersuchungsteilnehmers, die häufig grafisch als sog. Testprofil veranschaulicht werden.

Dimensionalität. Bei eindimensionalen Tests werden die Itemwerte in der Regel additiv zu einem Gesamtwert (bzw. Index, ▶ S. 143 ff.) gleich- oder ungleichgewichteter Items zusammengefasst (▶ auch ◘ Box 4.4). Welche dieser Vorgehensweisen gerechtfertigt ist, zeigt die Dimensionalitätsüberprüfung, die üblicherweise mit explorativen oder konfirmativen Faktorenanalysen durchgeführt wird (▶ S. 377 ff. und 516 bzw. ▶ Anhang B). Man achte hierbei darauf, dass eine repräsentative Stichprobe aus der jeweiligen Zielpopulation untersucht wird.

Faktoranalysen produzieren u. a. pro Faktor für jedes Item eine sog. Faktorladung. Eindimensionalität liegt vor, wenn die Item-Interkorrelationen auf einen Faktor (sog. Generalfaktor) reduziert werden können, auf dem sie hoch »laden« (d. h., mit dem sie hoch korrelieren). Der Faktor repräsentiert inhaltlich das »Gemeinsame«, das in allen Items ausgedrückt wird und steht für das zu messende Konstrukt. Sind die Faktorladungen homogen, d. h. sehr einheitlich, ist die Berechnung eines ungewichteten, additiven Gesamtwerts gerechtfertigt. Variieren die Faktorladungen innerhalb ihres theoretischen Wertebereiches von –1 bis +1 deutlich, so sind sie bei der Berechnung eines Gesamtwertes als Gewichte zu verwenden (▶ S. 145 ff.). Items mit geringen Faktorladungen (Faustregel: Beträge unter 0,6) sind aus dem Test bzw. Fragebogen zu entfernen (zum Problem »bedeutsamer« Faktorladungen vgl. Bortz, 2005, S. 551 f.; Briggs & Cheek, 1986; Fürntratt, 1969).

Eindimensional intendierte Tests erweisen sich nicht selten bei späteren empirischen Dimensionalitätsüberprüfungen als mehrdimensional. Wieviele Faktoren zu extrahieren und wie diese angemessen zu interpretieren sind, ist dabei jedoch keineswegs immer eindeutig, da die Technik der Faktorenanalyse erhebliche Interpretationsspielräume offenlässt (zur Überprüfung der Eindimensionalität vgl. auch Hattie, 1985). Die spätere Ausdifferenzierung eindimensionaler Tests hat in erster Linie explorativen Wert; sie dient der Verfeinerung theoretischer Annahmen über das Konstrukt und regt neue Testentwicklungen an. Eine methodisch saubere Konstruktion mehrdimensionaler Tests geht von einer theoretisch begründeten, genau festgelegten Zahl inhaltlich klar umrissener Teilkomponenten (Faktoren) des Zielkonstrukts aus, die als Subtests operationalisiert werden, d. h., für jeden Faktor wird ein separater (gewichteter oder ungewichteter) Testwert berechnet.

Die Dimensionalität eines Tests gibt an, ob er nur *ein* Merkmal bzw. Konstrukt erfasst (eindimen*sio*naler Test), oder ob mit den Testitems *mehrere* Konstrukte bzw. Teilkonstrukte operationalisiert werden (mehrdimensionaler Test).

Die klassische Testtheorie ist in der Konzeption ihrer Test- und Itemkennwerte auf eindimensionale Tests zugeschnitten. Bei der Übertragung dieser Kennwerte auf mehrdimensionale Tests oder Fragebögen bieten sich – sofern die Subtests genügend Items enthalten – separate Itemanalysen sowie Objektivitäts-, Reliabilitäts- und Validitätsbeurteilungen für die einzelnen Teiltests an. Gelegentlich interessieren bei der Itemanalyse auch die Reliabilitäten und Validitäten einzelner Items (vgl. Lienert & Raatz, 1994, Kap. 2.2). Ein Verfahren zur Bestimmung dieser Koeffizienten, das auch bei ordinalen Daten verwendbar ist, beschreibt Aiken (1980).

4.3.6 Testskalen

Während mit dem Begriff »Test« die Menge der Testitems und Antwortvorgaben samt Instruktion gemeint