Interpretation von Testresultaten und Testeichung

Chapter · January 2008

DOI: 10.1007/978-3-540-71635-8_8

2 authors:

Frank Goldhammer
DIPF | Leibniz Institute for Research and Information in Education, Centre for Inter…
148 PUBLICATIONS **3,075** CITATIONS
SEE PROFILE

Johannes Hartig
DIPF - Leibniz Institute for Research and Information in Education
177 PUBLICATIONS **4,991** CITATIONS
SEE PROFILE

8 Interpretation von Testresultaten und Testeichung

Frank Goldhammer, Johannes Hartig

Wendet man einen psychologischen Test an, so erhält man in der Regel ein numerisches Testresultat, das Auskunft über die Merkmalsausprägung der Testperson geben soll. Fragt man sich, was dieser Testwert hinsichtlich der Merkmalsausprägung aussagt, dann lässt sich diese Frage in zweierlei Weise sinnvoll beantworten: einerseits dadurch, dass der Testwert durch den Vergleich mit den Testwerten einer Bezugsgruppe interpretiert wird (normorientierte Interpretation) oder dass andererseits eine genaue theoretische Vorstellung darüber besteht, wie der erzielte Testwert mit einem inhaltlich-psychologisch definierten Kriterium in Beziehung steht (kriteriumsorientierte Interpretation).

8.1 Testwertbildung und Testwertinterpretation

Bevor ein **Testwert**, d.h. das numerische Testresultat einer Testperson, interpretiert werden kann, muss er gemäß definierter Regeln gebildet werden (vgl. Kelava &

Moosbrugger, 2007, Kap. 4.7 in diesem Band). Regeln zur Testwertbildung sind in Abhängigkeit von dem im psychologischen Test geforderten Antwortverhalten unterschiedlich komplex. Im Falle von frei formulierten Verbalantworten ist eine ausführliche Anleitung nötig, um den Antworten in differenzierter Weise Punktwerte zuzuweisen, welche zusammengenommen den Testwert ergeben. Bei Wahlaufgaben genügt dagegen ein Antwortschlüssel, anhand dessen einer gewählten Antwort ein bestimmter Punktwert zugeordnet wird (z.B. multiple-choice Persönlichkeitsfragebogen) oder aber es wird das Verhältnis von Arbeitsmenge und Bearbeitungszeit zur Ermittlung des Testwertes herangezogen (z.B. Leistungstests mit Zeitbegrenzung). Der Testwert wird auch als **Rohwert** bezeichnet, da er sich unmittelbar aus den registrierten Antworten ergibt und nicht weitergehend verarbeitet ist.

Obwohl also der Testwert das Antwortverhalten der Testperson widerspiegelt, führt, wie im nachfolgenden Beispiel 8.1 gezeigt wird, die Interpretation des Testwertes ohne die Hinzunahme weiterer Informationen leicht zu Fehlinterpretationen.

Beispiel 8.1

Mangelnde Aussagekraft von Rohwerten.

Der Schüler Peter kann in einer Rechenprobe 18 von 20 Aufgaben richtig lösen, d.h. 90% der Aufgaben. In der Vokabelprobe schafft er 21 von 30 Aufgaben, d.h. nur 70%. Seine Eltern freuen sich über die Rechenleistung ihres Sohnes, mit der Vokabelleistung sind sie weniger zufrieden. Zu Recht wendet Peter ein, dass viele seiner Mitschüler nur eine deutlich geringere Anzahl von Vokabelaufgaben richtig lösen konnten, d.h. im Vergleich zu den anderen Schülern habe er sehr gut abgeschnitten. Außerdem verteidigt er sich damit, dass in der Probe einige Vokabeln abgefragt wurden, die der Lehrer im Unterricht kaum vorbereitet hatte, d.h. die Vokabelprobe war sehr schwierig. Auch wenn Peter mit dieser Argumentation seine Leistung in der Vokabelprobe zu seinen Gunsten relativieren kann, stellt sich analog die Frage, ob Peter seine gute Rechenleistung der Auswahl einfacher Aufgaben verdankt, d.h. ob viele seiner Mitschüler ebenfalls mit Leistungen im oberen Punktebereich abgeschnitten haben.

Das Beispiel macht deutlich, dass der Testwert (hier z.B. der Anteil gelöster Aufgaben) per se nicht aussagekräftig ist, insofern die Höhe des Testwertes nicht nur von dem inte-

ressierenden Merkmal (z.B. Rechenfähigkeit) abhängt, sondern auch wesentlich durch die Aufgabenauswahl bzw. die Aufgabenschwierigkeit mitbestimmt wird.

Um die aus der Aufgabenauswahl resultierende Uneindeutigkeit von Testwerten zu vermeiden, werden Testwerte anhand eines Vergleichsmaßstabes interpretiert.

Vergleichsmaßstab zur Testwertinterpretation

Um eine eindeutige Aussage über die individuelle Merkmalsausprägung treffen zu können, wird zusätzlich zum Testwert ein **Vergleichsmaßstab** benötigt, anhand dessen der Testwert eingeordnet wird. Als Vergleichsmaßstäbe können entweder Merkmalsverteilungen von Bezugsgruppen (*normorientierte Testwertinterpretation*) oder psychologisch-inhaltliche Beschreibungen, welche die Testwertausprägungen genau charakterisieren (*kriteriumsorientierte Testwertinterpretation*), herangezogen werden.

Bei der normorientierten Testwertinterpretation (s. Kap. 8.2) erhält man zum Einen Informationen über die Merkmalsausprägung relativ zur Bezugsgruppe, zum Anderen wird damit das Problem der Abhängigkeit des Testwertes von der Aufgabenauswahl behandelt, insofern sich die Information aus der Testnorm nur mehr auf die Stellung der Testperson innerhalb der Bezugsgruppe bezieht und nicht mehr direkt von der Aufgabenauswahl abhängt.

Bei der kriteriumsorientierten Testwertinterpretation (s. Kap. 8.3) wird die unkontrollierte Abhängigkeit des Testwertes von der Aufgabenauswahl von vornherein vermieden, insofern eine genaue theoretische Vorstellung darüber besteht, wie das Beantworten bestimmter Testaufgaben und somit der jeweilige Testwert mit einem genau definierten psychologisch-inhaltlichen Kriterium in Beziehung steht (s. auch Moosbrugger, 2007a, Kap. 10 in diesem Band).

8.2 Normorientierte Testwertinterpretation

Die normorientierte Testwertinterpretation besteht darin, dass zu einem individuellen Testwert ein **Normwert** bestimmt wird, anhand dessen die Testperson hinsichtlich der erfassten Merkmalsausprägung innerhalb der Bezugs- bzw. Referenzgruppe positioniert wird.

Im Rahmen normorientierter Testwertinterpretation lassen sich grundsätzlich zwei Vorgehensweisen unterscheiden, wie für einen bestimmten Testwert ein Normwert ermittelt

wird, nämlich eine nicht-lineare Transformation des Testwertes zur Gewinnung von *Prozenträngen* sowie eine lineare Transformation des Testwertes, welche zu standardisierten *z-Normwerten* führt.

8.2.1 Bildung von Prozentrangnormen durch nicht-lineare Testwerttransformation

Die gebräuchlichste **nicht-lineare Testwerttransformation** ist die Transformation des Testwertes in einen Prozentrang (PR). Nicht-Linearität bedeutet hierbei, dass der Prozentrang durch eine *Transformation der Häufigkeitsverteilung* der Bezugsgruppe gewonnen wird. Dazu wird die relative Position des Testwertes x_v in der aufsteigend geordneten Rangreihe der Testwerte in der Bezugsgruppe ermittelt.

Definition

Ein **Prozentrang** gibt an, wie viel Prozent der Bezugsgruppe bzw. Normierungsstichprobe einen Testwert erzielten, der niedriger oder maximal ebenso hoch ist, wie der Testwert x_v der Testperson v. Der Prozentrang entspricht somit dem prozentualen Flächenanteil der Häufigkeitsverteilung der Bezugsgruppe, der am unteren Skalenende beginnt und nach oben hin durch den Testwert x_v begrenzt wird.

Um die Zuordnung eines Testwertes x_v zu seinem korrespondierenden Prozentrang einfach vornehmen zu können, wird im Rahmen der Testkonstruktion aus den Testwerten x_v der Normierungsstichprobe eine tabellarische Prozentrangnorm gebildet.

Zur Erstellung einer **Prozentrangnorm** werden für alle Testwerte x_v der Normierungsstichprobe die zugehörigen Prozentränge PR_v folgendermaßen bestimmt:

- die Testwerte x_v der Normierungsstichprobe vom Umfang N werden in eine aufsteigende Rangordnung gebracht,
- die Häufigkeiten $\text{freq}(x_v)$ der einzelnen Testwertausprägungen werden erfasst,
- die kumulierten Häufigkeiten $\text{freq}_{\text{cum}}(x_v)$ bis einschließlich des jeweiligen Testwertes x_v werden bestimmt,
- die kumulierten Häufigkeiten werden durch den Umfang N der Normierungsstichprobe dividiert und mit dem Faktor 100 multipliziert

$$PR_v = 100 \cdot \frac{freq_{cum}(x_v)}{N} \quad .$$

Aus den so gewonnenen Prozenträngen PR_v wird eine **Normtabelle** gebildet, in der jedem Prozentrang zwischen 1 und 100 der jeweils zugehörige Testwert x_v zugeordnet ist. Dabei ist es möglich, dass im Bereich geringer Testwertdichte einem Intervall von unterschiedlichen Testwerten derselbe Prozentrang zugeordnet wird; im Bereich hoher Testwertdichte kann umgekehrt aufeinander folgenden Prozenträngen derselbe Testwert zugeordnet werden (s.u. das Beispiel 8.2). Liegt für ein psychologisches Testverfahren eine Prozentrangnormierung vor, kann der Testwert x_v einer Testperson dadurch interpretiert werden, dass anhand der Normtabelle der dem Testwert x_v entsprechende Prozentrang abgelesen wird.

Während der Prozentrang die relative Position der Merkmalsausprägung einer Testperson in der Normierungsstichprobe beschreibt, bezeichnet das **Perzentil** jenen Testwert x_v, der einem bestimmten Prozentrang in der Normierungsstichprobe entspricht. Das heißt z.B., dass derjenige Testwert, welcher von 30% der Testwerte unterschritten bzw. höchstens erreicht wird, 30. Perzentil genannt wird. Die grobstufigeren **Quartile** entsprechen dem 25., 50. bzw. 75. Perzentil. Als 1. Quartil (Q1) wird demnach derjenige Testwert bezeichnet, der von 25% der Testwerte unterschritten oder erreicht wird; das 2. Quartil (Q2) ist derjenige Testwert, den 50% der Testwerte unterschreiten oder erreichen, d.h. Q2 entspricht dem Median Das 3. Quartil (Q3) schließlich ist derjenige Testwert, den 75% der Testwerte unterschreiten oder erreichen.

Die Bedeutung von Prozenträngen und Quartilen ist anhand einer schiefen Häufigkeitsverteilung von Testwerten x_v in Abbildung 8.1 veranschaulicht.

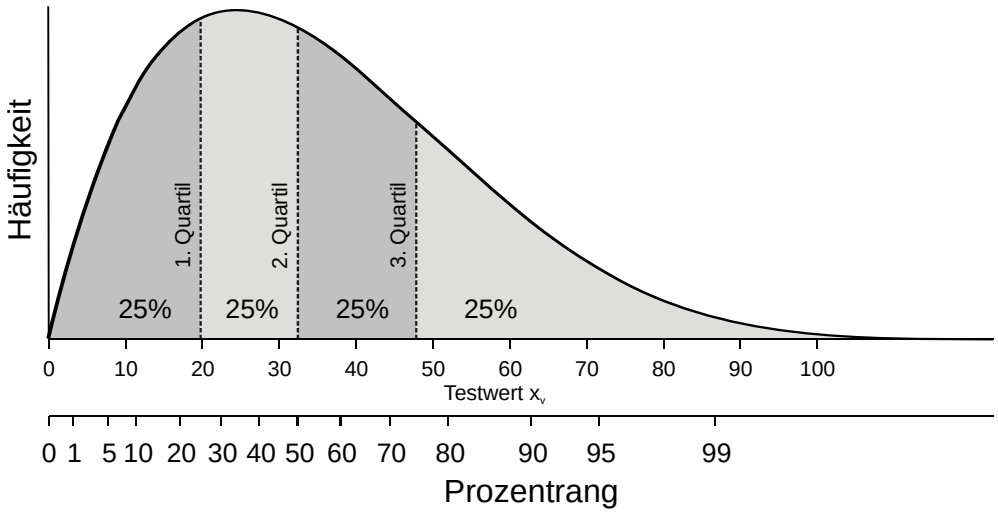

Abbildung 8.1: Prozentränge und Quartile bei einer schiefen Häufigkeitsverteilung.

Beispiel 8.2

Prozentrangnorm

Für den computerbasierten Frankfurter Adaptiven Konzentrationsleistungs-Test FAKT-II (Moosbrugger & Goldhammer, 2007) liegen Prozentrangnormen vor, die nach Testform, Durchführungsart und Testungszahl differenziert sind (s. dazu auch Kap. 8.5). Auf dem Ergebnisbogen wird u.a. für den Testwert Konzentrations-Leistung KL1 automatisch ein Prozentrang ausgegeben. Beispielweise hat ein Proband, der bei erstmaliger Bearbeitung des FAKT mit Testform E und der Durchführungsart „Standardtestzeit von 6 Minuten“ für die Konzentrations-Leistung KL1 einen Testwert von $x_v = 146$ erzielt, den Prozentrang $PR_v = 76$. Der Prozentrang von 76 zeigt an, dass 76% der Normierungsstichprobe eine geringere oder gleich hohe Leistung erbracht haben; die Leistung des Probanden liegt oberhalb von Q3 und er zählt somit zum leistungsstärksten Viertel der Bezugsgruppe.

Zur Ermittlung eines Prozentrangs sucht das Computerprogramm in der entsprechenden Normentabelle automatisch denjenigen Prozentrang, welcher dem von der Testperson erzielten Testwert entspricht bzw. bei fehlender Entsprechung den nächst höheren Prozentrang.

Testwert KL1	Prozentrang
…	…
145	74
146	75
146	**76**
147	77
…	…

Zu beachten ist, dass in der Normierungsstichprobe der Testwert 146 relativ häufig auftrat (hohe Testwertdichte), sodass dieser sowohl als 76. als auch als 75. Perzentil zu bezeichnen ist. In solch einem Fall wird gemäß obiger Prozentrangdefinition dem Probanden der jeweils höhere Prozentrang zugewiesen.

Prozentrangnormen stellen verteilungsunabhängige Normen dar, d.h. bei der Erstellung von Prozentrangnormen wird keine bestimmter Verteilungsform, wie z.B. die Normalverteilung der Testwerte in der Normierungsstichprobe, vorausgesetzt. Für die Bildung von Prozentrangnormen ist Ordinalskalenniveau (vgl. Kelava & Moosbrugger, 2007, Kap. 4.2 in diesem Band) der Testwerte ausreichend, da an den Testwerten keine Lineartransformation sondern lediglich eine monotone Transformation vorgenommen wird. Falls für eine Testwertvariable nur Ordinalskalenniveau angenommen werden kann, kommt also lediglich die Bildung von Prozentrangnormen in Frage (Rettler, 1999).

Selbst bei intervallskalierten Testwerten können Prozentrangnormen nicht als intervallskaliert aufgefasst werden, da durch die Flächentransformation die Differenzen zwischen je zwei Testwerten im Bereich geringer Testwertdichte verkleinert, im Bereich hoher Testwertdichte vergrößert werden (vgl. Abb. 8.1). Dies bedeutet, dass Prozentränge im Bereich hoher Testwertdichte Unterschiede zwischen Merkmalsausprägungen in einer Weise hervortreten lassen, wie sie empirisch gar nicht bestehen, wohingegen im Bereich geringer Testwertdichte tatsächlich bestehende Unterschiede weitgehend nivelliert werden. Trotz der Verteilungsunabhängigkeit von Prozentrangnormen ist also die Kenntnis der Verteilungsform für den Testanwender beim Vergleich von Prozenträngen von Bedeutung. Liegt beispielsweise eine Normalverteilung vor, wie z.B. für das Merkmal Intelligenz, dann werden durch die Anwendung von Prozentrangnormen kleine Testwertunterschiede im mittleren Skalenbereich überbetont, da bei normalverteilten Merkmalen im mittleren Skalenbereich eine hohe Testwertdichte besteht; liegt hingegen eine linkssteile Verteilung vor, wie z.B. für das Merkmal Reaktionsschnelligkeit, dann muss der Testanwender mit dieser Überbetonung im unteren Skalenbereich rechnen, da hier eine hohe Testwertdichte besteht. Aus der fehlenden Intervallskalierung der Prozentrangnormen folgt auch, dass Prozentränge nicht für Vergleiche herangezogen werden dürfen, denn eine Prozentrangdifferenz von 50-40 = 10 hat bezogen auf die untersuchte Merkmalsausprägung eine ganz andere Bedeutung als eine numerisch gleiche Prozentrangdifferenz von 90-80 = 10, wovon man sich in Abbildung 8.1 leicht überzeugen kann. Vergleiche, die sich nicht auf die Merkmalsausprägung, sondern auf die Fläche der Merkmalsverteilung beziehen, sind hingegen zulässig, wie z.B. dass sich eine Person A mit dem Prozentrang von 90 von einer Person B mit Prozentrang 70 darin unterscheidet, dass in der Normierungsstichprobe 20% der Personen einen höheren

Testwert erzielten als Person B und zugleich den Testwert von Person A unterschritten oder erreichten.

8.2.2 Bildung von z_v-Normen durch lineare Testwerttransformation

Wie die nicht-lineare Prozentrang-Transformation dient auch die **lineare z_v-Transformation von Testwerten** dazu, die relative Position des Testwertes x_v in der Verteilung der Bezugsgruppe anzugeben. Im Gegensatz zu Prozentrangnormen muss bei z_v-Normen für die Testwertvariable Intervallskalenniveau (vgl. Kelava & Moosbrugger, 2007, Kap. 4.2 in diesem Band) angenommen werden können, da die Position des Testwertes x_v einer Testperson als Abstand bzw. Differenz zum arithmetischen Mittelwert $\bar{x}$ der Verteilung der Bezugsgruppe ausgedrückt wird. Durch die Differenzbildung kann der Testwert x_v als unter- oder überdurchschnittlich interpretiert werden. Um die Vergleichbarkeit von Testwerten aus Tests mit verschiedenen Testwertstreuungen und Skalenbereichen zu ermöglichen, wird bei der z_v-Transformation die Differenz $x_v - \bar{x}$ an der Standardabweichung s_x der Testwerte x_v relativiert.

Definition

Der **z_v-Normwert** gibt an, wie stark der Testwert x_v einer Testperson v vom Mittelwert $\bar{x}$ der Verteilung der Bezugsgruppe in Einheiten der Standardabweichung s_x der Testwerte x_v abweicht. Der z_v-Normwert von Testperson v wird folgendermaßen berechnet:

$$z_v = \frac{x_v - \bar{x}}{s_x}$$

z_v-Normwerte haben eine Mittelwert von $\bar{z} = 0$ und eine Standardabweichung von $s_z = 1$.

Im Gegensatz zu Prozenträngen können auf die z_v-Normwerte dieselben algebraischen Operationen angewendet werden wie auf die Testwerte. Die Differenz zwischen zwei z_v-Normwerten ist proportional zu derjenigen der entsprechenden Testwerte. Bei Statistiken, welche invariant gegenüber Lineartransformationen sind, z.B. die Produkt-Moment-Korrelation, resultiert bei Verwendung von Normwerten das gleiche Ergebnis wie bei Verwendung der Testwerte.

Die Berechnung des z_v-Normwertes ist prinzipiell verteilungsunabhängig, d.h. die Testwerte x_v können beispielsweise auch bei fehlender Normalverteilung durch Berech-

nung des z_v-Normwertes interpretiert werden. Wenn jedoch der Testwert x normalverteilt ist, nimmt der interpretative Gehalt des z_v-Normwertes beträchtlich zu. Der z_v-Normwert heißt dann **Standardwert** und es besteht die Möglichkeit, die prozentuale Häufigkeit der Standardwerte innerhalb beliebiger Wertebereiche über die Verteilungsfunktion der Standardnormalverteilung zu bestimmen (vgl. Rettler, 1999). Liegt dagegen keine normalverteilte Testwertvariable vor, ist diese Interpretation falsch und daher unzulässig[1].

Da mit der Bildung von z_v-Normen negative Vorzeichen und Dezimalstellen einhergehen, ist ihre Verwendung eher unüblich. Vielmehr wird der z_v-Normwert weiteren Lineartransformation unterzogen, um Normwerte mit positivem Vorzeichen sowie möglichst ganzzahliger Abstufung zu erhalten. Auf diese Weise lassen sich Normen mit unterschiedlicher Metrik erzeugen, am Prinzip der Testwertinterpretation ändert sich dadurch jedoch nichts (s. folgendes Beispiel 8.3). Beispielsweise wird für Intelligenzmessungen der Standardwert z_v oft noch durch Multiplikation mit dem Faktor 15 und Addition einer Konstante von 100 in den Intelligenz-Quotienten (IQ) umgeformt, dessen Mittelwert 100 und Standardabweichung 15 beträgt[2]. Für die im Rahmen der PISA-Studien (Programme for International Student Assessment) berichteten Schülerleistungen wurde eine normierte Skala gebildet, bei der der Leistungsmittelwert über alle teilnehmenden OECD-Staaten 500 und die Standardabweichung 100 Punkte beträgt (z.B. OECD 2001, 2004).

Beispiel 8.3

z_v-Normwert

[1] Bei fehlender Normalverteilung müsste für jeden spezifischen Verteilungsfall die Verteilungsfunktion bestimmt bzw. tabellarisiert werden. Im Prinzip wäre dies die Vorgehensweise bei der Bestimmung von Prozentrangnormen für beliebige Verteilungsformen (s. Kap. 8.2.1).

[2] Die Wahl der additiven Konstante von 100 und des Faktors von 15 erfolgte mit dem Ziel, den auf dem z_v-Normwert basierenden Intelligenz-Quotienten mit dem klassischen Intelligenzquotienten vergleichen zu können bzw. die gebräuchliche Metrik beizubehalten. Der frühere IQ berechnet sich nämlich aus dem Verhältnis von Intelligenz- und Lebensalter multipliziert mit 100, d.h. er beträgt im Mittel 100, zudem ergab sich für ihn empirisch eine Standardabweichung von 15.

Eine Testperson habe in einem Intelligenztest mit dem Mittelwert von $\bar{x} = 31$ und der Standardabweichung von $s_x = 12$ einen Testwert von $x_v = 27$ erzielt. Der z_v-Normwert ergibt sich folgendermaßen:

$$z_v = \frac{27 - 31}{12} = -0.33 .$$

Der Testwert liegt also um ein Drittel der Standardabweichung unter der durchschnittlichen Testleistung. An dieser Interpretation ändert sich nichts, wenn aus dem z_v-Normwert der Intelligenz-Quotient wie folgt bestimmt wird:

$$IQ_v = 100 + 15 \cdot z_v = 100 + 15 \cdot (-0.33) = 95 ,$$

d.h. auch in dieser Metrik liegt die Intelligenzleistung der Testperson um ein Drittel der Standardabweichung bzw. um 5 IQ-Punkte unter dem Mittelwert von 100.

Ist die Testwertvariable x normalverteilt, kann anhand von z_v der entsprechende Prozentrang aus der tabellarisierten Verteilungsfunktion der Standardnormalverteilung abgelesen werden (s. z.B. Bortz, 2006). Für das vorliegende Beispiel ist der PR = 36. Liegt dagegen keine Normalverteilung, sondern z.B. eine schiefe Verteilung vor, lässt sich mit Hilfe der Standardnormalverteilung *keine* Aussage darüber treffen, wie hoch die Wahrscheinlichkeit für einen Testwert ist, der niedriger oder maximal ebenso hoch ist wie z_v. Grobe Abschätzungen erlauben die Ungleichungen nach Tchebycheff (s. z.B. Bortz, 2006)

Liegt eine normalverteilte Testwertvariable x vor, dann wird die z-Norm als **Standardnorm** bezeichnet. Unter Annahme der Normalverteilung verschafft Abbildung 8.2 einen Überblick über die Standardnormen IQ-, T-Werte und die PISA-Skala mit der jeweils zugehörigen Vorschrift zur Transformation von z_v, ihren Mittelwerten und Standardabweichungen. Zusätzlich ist hier die so genannte Stanine-Norm dargestellt, bei der anhand der Werteverteilung neun Abschnitte mit Häufigkeiten von 4%, 7%, 12%, 17%, 20%, 17%, 12%, 7% und 4% gebildet werden – diese Unterteilung ergibt für einen normalverteilten Messwert Intervalle von der Breite einer halben Standardabweichung, wobei bezüglich Stanine 1 und 9 vereinfachend angenommen wird, dass Stanine 1 bei z = -2.25 beginnt und Stanine 9 bei z = +2.25 endet. Weiterhin sind in Abbildung 8.2 die Prozentränge für eine normalverteilte Variable eingetragen.

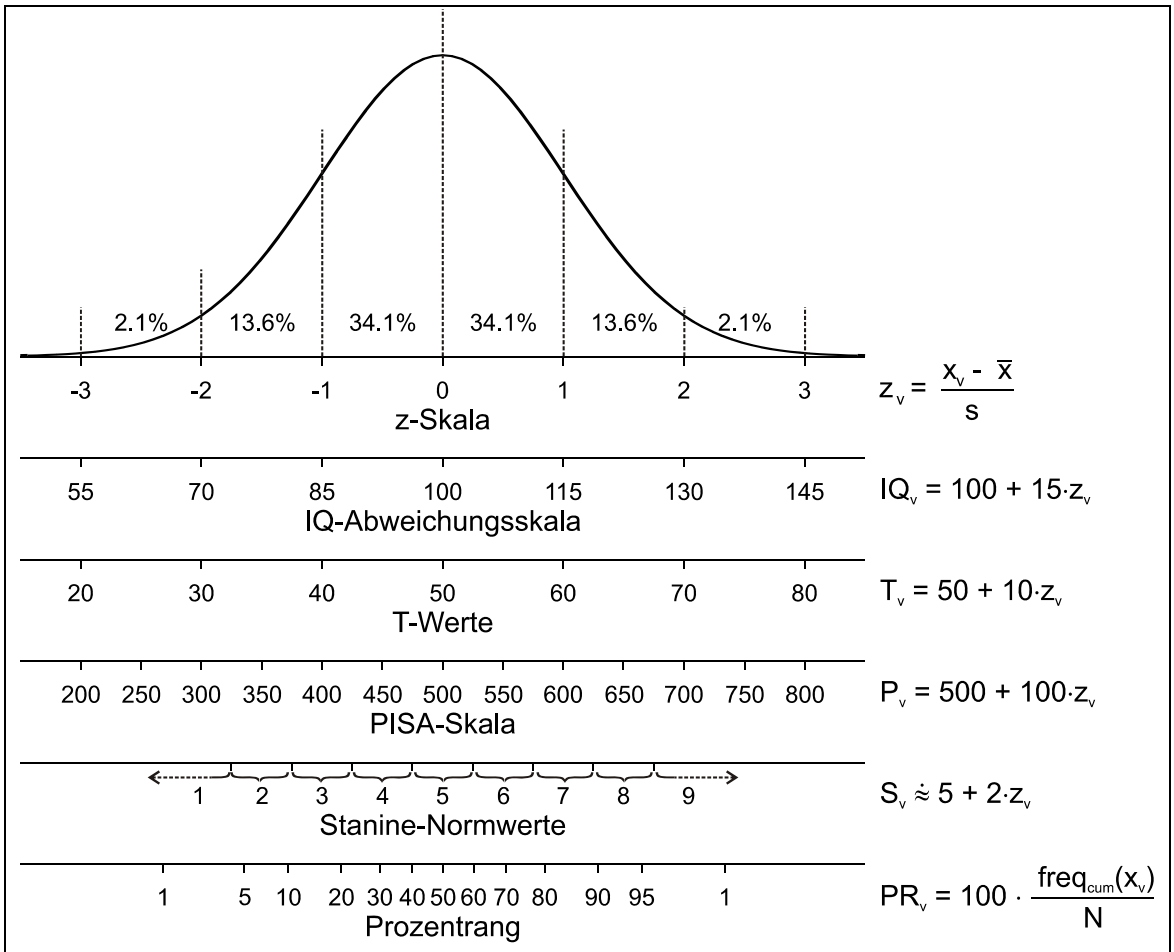

Abbildung 8.2: Gebräuchliche Standardnormen, Stanine-Norm und Prozentrangnorm unter Annahme normalverteilter Testwerte.

Falls die Testwertvariable x nicht normalverteilt ist, kann die Testwerteverteilung über eine Flächentransformation in eine Normalverteilung umgewandelt werden (Normalisierung, s. Kelava & Moosbrugger, 2007, Kap. 4.8.2 in diesem Band). Allerdings sollte hierfür plausibel gemacht werden können, weshalb sich im konkreten Fall empirisch keine Normalverteilung gezeigt hat, und warum die Annahme einer Normalverteilung für das jeweils untersuchte Merkmal dennoch begründbar ist.

Die normorientierte Interpretation von Testresultaten bezieht sich in der Regel auf psychologische Testverfahren, welche nach der klassischen Testtheorie (s. Moosbrugger, 2007b, Kap. 5 in diesem Band) konstruiert wurden. Doch auch bei probabilistischen Testmodellen der Item-Response-Theorie (s. Moosbrugger, 2007a, Kap. 10 in diesem Band) lassen sich die latenten Personenparameter ξ_v normorientiert interpretieren, wenn sie so normiert werden, dass ihre Summe gleich Null ist (Rost, 2004). In diesem Fall geben das Vorzeichen und der Betrag des Personenparameters an, wieweit sich die

Testperson über oder unter dem mittlerem Personenparameter von 0 in der Bezugsgruppe befindet.

8.3 Kriteriumsorientierte Testwertinterpretation

Bei der kriteriumsorientierten Testwertinterpretation erfolgt die Interpretation des Testwertes nicht in Bezug zur Testwerteverteilung einer Bezugsgruppe, sondern in Bezug auf ein spezifisches inhaltliches Kriterium (s. folgendes Beispiel 8.4). Ein derartiges Kriterium besteht in bestimmten diagnostischen Aussagen, die auf Basis des Testwertes über die getesteten Personen gemacht werden sollen. Bei der kriteriumsorientierten Testwertinterpretation interessiert nicht, wie viele Personen das Kriterium erfüllen. Theoretisch könnten alle getesteten Personen ein Kriterium erreichen oder aber keine einzige.

Beispiel 8.4

Diagnostische Aussagen bezogen auf ein spezifisches Kriterium

– Ein Patient in einer psychotherapeutischen Ambulanz wird mit einem Depressivitäts-Fragebogen untersucht. Aufgrund des Testresultats soll entschieden werden, ob eine genauere Diagnostik und Therapie hinsichtlich einer ausgeprägten Depression (*Major Depression*) angezeigt erscheint. Die Höhe des Testwertes soll also Auskunft darüber geben, ob der Patient zu jener Gruppe von Patienten gehört, die das Kriterium für eine Major Depression erfüllen, oder nicht. Das Kriterium, anhand dessen eine Interpretation des Testergebnisses erfolgen soll, ist hier also das Vorliegen einer Major Depression.

– In einer Schulklasse werden die Schüler zum Ende des Schuljahres mit einem Vokabeltest in Englisch getestet. Es soll geprüft werden, wie viele Schüler in der Klasse das im Lehrplan gesetzte Leistungsniveau erreicht haben und das entsprechende Vokabular schriftlich beherrschen. Das Kriterium ist in diesem Fall also das Erreichen eines bestimmten, sachlich begründeten Leistungsniveaus.

Um eine kriteriumsorientierte Interpretation eines Testwertes vorzunehmen, werden in der Regel vorab bestimmte *Schwellenwerte* definiert, ab denen ein Kriterium als zutreffend angenommen wird. Es wird zum Beispiel ab einem Wert von 19 Punkten im Depressivitäts-Fragebogen das Vorliegen einer Major Depression oder ab 30 Punkten im Vokabeltest das Erreichen des im Lehrplan definierten Leistungszieles angenommen.

Solche Schwellenwerte können auf zwei unterschiedliche Weisen ermittelt werden. Zum einen kann der Testwert in eigens dafür durchgeführten Untersuchungen in Bezug zu einem *externen Kriterium* gesetzt werden (s. Kap. 8.3.1). Das Vorliegen des externen Kriteriums muss in diesen Untersuchungen zusätzlich zu den individuellen Testwerten erfasst werden. Zum anderen können die *Inhalte der Testaufgaben* herangezogen werden, um die Testwerte inhaltlich zu beschreiben (s. Kap. 8.3.2).

8.3.1 Bezug des Testwertes auf ein externes Kriterium

Ein einfaches mögliches Verfahren, einen Schwellenwert zur Unterscheidung von zwei Gruppen anhand eines externen Kriteriums zu ermitteln, ist die *Receiver-Operating-Characteristics-Analyse*, kurz **ROC-Analyse**. Dieses Verfahren stammt aus der Signalentdeckungstheorie der Psychophysik (Green & Swets, 1966). In der diagnostischen Anwendungspraxis eignet sich die ROC-Analyse für Situationen, in denen ein Teil der Fälle ein Kriterium erfüllt und ein Teil nicht, wobei die Zuordnung der Fälle hinsichtlich dieses Kriteriums anhand eines Testresultats vorgenommen werden soll. Es soll also zum Beispiel anhand des Depressivitäts-Fragebogens beurteilt werden, ob ein Patient tatsächlich das Kriterium der Major Depression erfüllt. Eine derartige Entscheidung kann in der diagnostischen Praxis in vier verschiedene Kategorien fallen:

- **Treffer:** Der Fall erfüllt das Kriterium Depression und wird korrekt als depressiv klassifiziert (*richtig positiv, RP*)
- **Verpasser:** Der Fall erfüllt das Kriterium Depression, er wird aber fälschlicherweise als nicht depressiv diagnostiziert (*falsch negativ, FN*)
- **Falscher Alarm:** Der Fall erfüllt das Kriterium Depression nicht, er wird aber fälschlicherweise als depressiv diagnostiziert (*falsch positiv, FP*)
- **Korrekte Ablehnung:** Der Fall erfüllt das Kriterium Depression nicht und wird korrekt als nicht depressiv klassifiziert (*richtig negativ, RN*)

		Klassifikation +	Klassifikation -
Kriterium	+	RP	FN
	-	FP	RN

-	negativ
+	positiv
RP	richtig positiv
FN	falsch negativ
FP	falsch positiv
RN	richtig negativ

Wenn eine Klassifikationsentscheidung anhand eines einzelnen Testwertes getroffen werden soll, verändert sich die Wahrscheinlichkeit für eine spezifische Klassifikation in Abhängigkeit davon, ob der Schwellenwert für die Klassifikation höher oder niedriger angesetzt wird. Die Genauigkeit der Entscheidungen in Abhängigkeit vom Schwellenwert lässt sich anhand der Maße **Sensitivität** und **Spezifität**[3] ausdrücken. Sensitivität oder *Trefferquote* bezeichnet die Wahrscheinlichkeit für die Entscheidung „RP", d.h. dafür, dass ein Fall, der das Kriterium erfüllt, auch entsprechend als positiv klassifiziert wird. Aus dem Komplement 1 – Sensitivität ergibt sich die *Verpasserquote*, welche die Wahrscheinlichkeit für die Entscheidung „FN" angibt, d.h. dafür, dass ein Fall, der das Kriterium erfüllt, fälschlicherweise als negativ klassifiziert wird. Spezifität oder die *Quote korrekter Ablehnungen* bezeichnet hingegen die Wahrscheinlichkeit für die Entscheidung „RN", d.h. dafür, dass ein Fall, der das Kriterium nicht erfüllt, auch entsprechend als negativ klassifiziert wird. Aus dem Komplement 1 – Spezifität ergibt sich die *Quote falscher Alarme*, welche die Wahrscheinlichkeit für die Entscheidung „FP" angibt, d.h. dafür, dass ein Fall, der das Kriterium nicht erfüllt, fälschlicherweise als positiv klassifiziert wird.

Definition: Sensitivität und Spezifität

$$\text{Sensitivität} = \frac{RP}{FN + RP} \quad \text{(Trefferquote)}$$

$$1 - \text{Sensitivität} = \frac{FN}{FN + RP} \quad \text{(Verpasserquote)}$$

$$\text{Spezifität} = \frac{RN}{RN + FP} \quad \text{(Quote korrekter Ablehnungen)}$$

$$1 - \text{Spezifität} = \frac{FP}{FP + RN} \quad \text{(Quote falscher Alarme)}$$

[3] Der Begriff „Spezifität", wie er im Rahmen der ROC-Analyse gebraucht wird, ist nicht zu verwechseln mit dem Spezifitätsbegriff in der Latent State-Trait-Theorie (s. Kelava & Schermelleh-Engel, 2007, Kap. 15 in diesem Band).

In Abbildung 8.3 ist die grundsätzliche Voraussetzung für eine sinnvolle Klassifikation auf Basis des Testwertes veranschaulicht, nämlich dass sich hinsichtlich der Testwertausprägung die Fälle, die das Kriterium erfüllen („Positive"), deutlich von den übrigen Fällen („Negative") unterscheiden. Im Beispiel weist die Gruppe von Personen mit Major Depression (rechte Verteilung, „Depressive") in einem Depressivitätsfragebogen einen höheren Mittelwert (und zudem eine geringere Streuung) auf als die Vergleichsgruppe ohne Major Depression (linke Verteilung, „Gesunde").

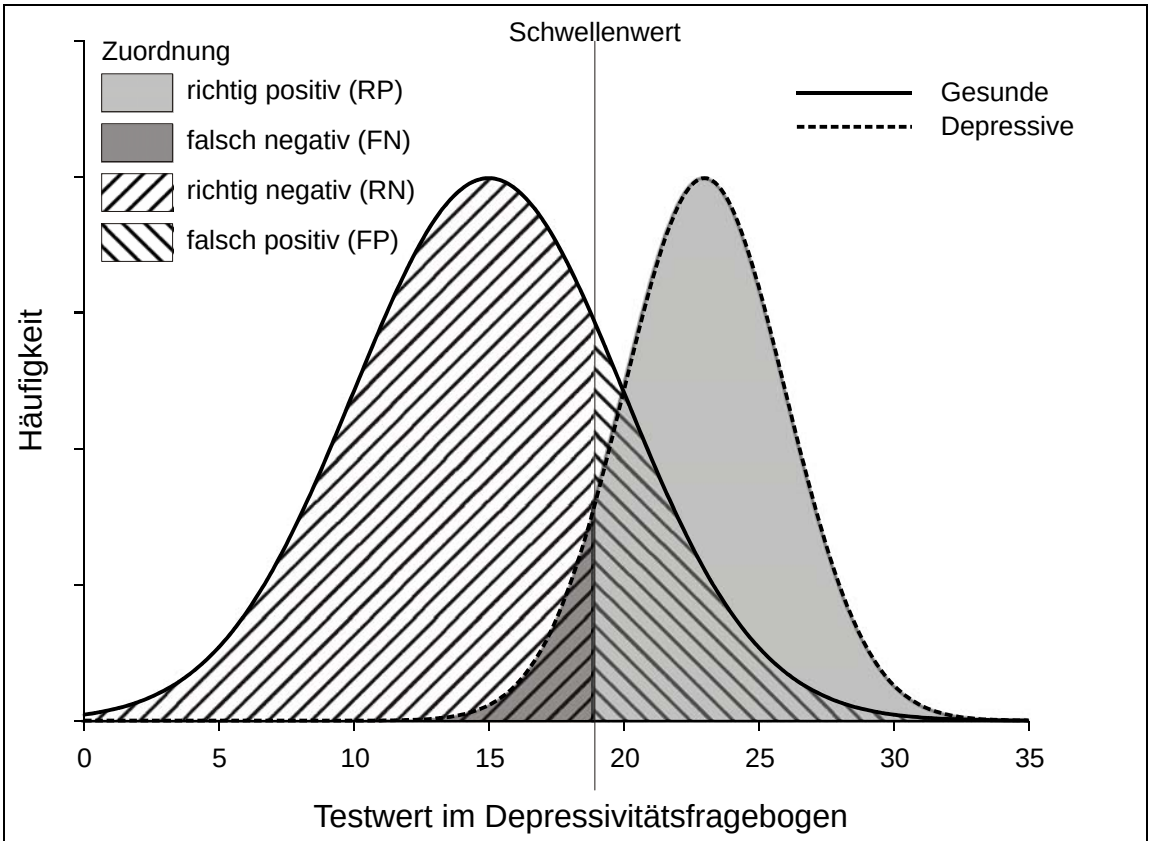

Abbildung 8.3: Verteilung der Testwerte in der Gruppe mit Major Depression (rechts, „Depressive") und in der Gruppe ohne Major Depression (links, „Gesunde"). Der Schwellenwert zur Klassifikation von Testpersonen liegt bei 18.5 Punkten.

Wie aus Abbildung 8.3 ersichtlich, sind Sensitivität und Spezifität vom Schwellenwert und voneinander abhängig. Wenn der Schwellenwert im Depressivitätsfragebogen sehr hoch gesetzt wird (z.B. 25), ist die Wahrscheinlichkeit für die Entscheidung „FP" klein und es wird nur selten fälschlicherweise eine Major Depression angenommen werden, die Spezifität ist also hoch. Gleichzeitig werden jedoch viele Patienten, die eine Major Depression haben, fälschlicherweise als nicht depressiv klassifiziert, die Sensitivität ist also niedrig. Ein niedriger Schwellenwert (z.B. 15) führt umgekehrt dazu, dass die

Wahrscheinlichkeit für die Entscheidung „FN“ sehr klein ist und dass fast alle Patienten mit Major Depression als richtig positiv klassifiziert werden; allerdings würde auch fast die Hälfte der Nicht-Depressiven fälschlicherweise als positiv klassifiziert. In diesem Fall lägen eine hohe Sensitivität und eine niedrige Spezifität vor. In Abbildung 8.4 ist graphisch dargestellt, wie sich die Maße für die Genauigkeit der Klassifikation, d.h. Sensitivität und Spezifität, mit der Verschiebung des Schwellenwertes gegenläufig verändern.

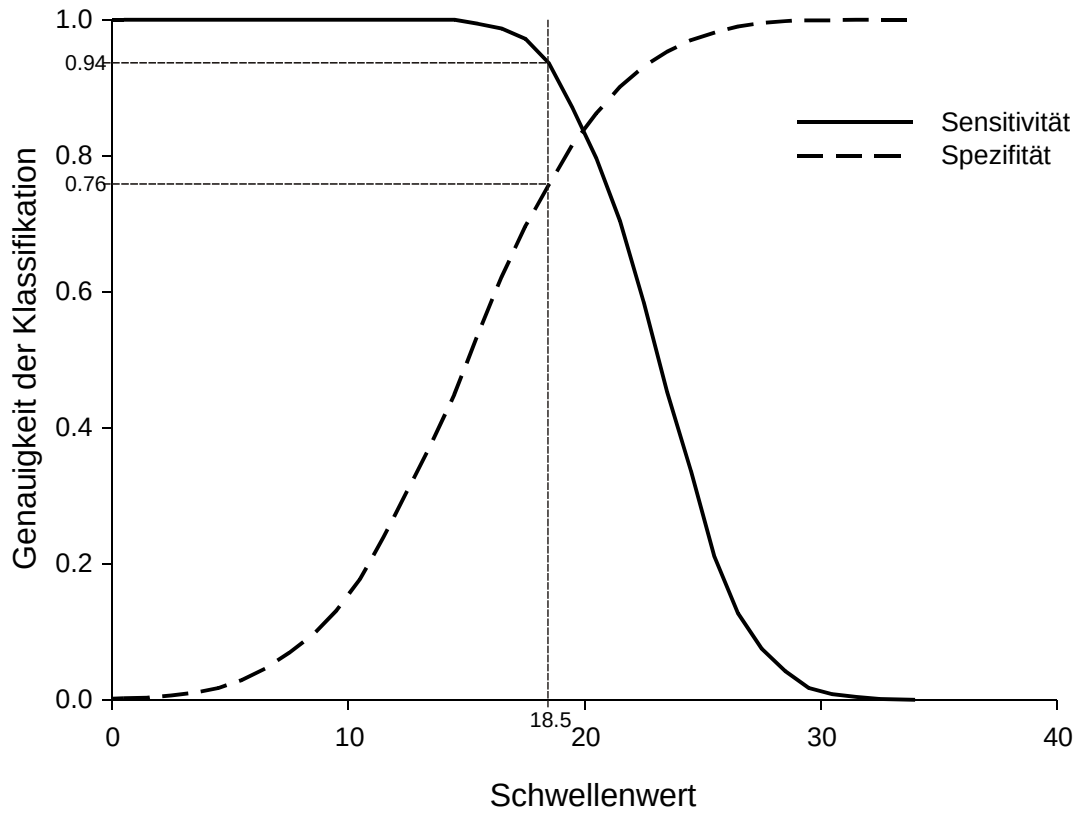

Abbildung 8.4: Sensitivität und Spezifität bei der Klassifikation depressiver und gesunder Patienten in Abhängigkeit vom Schwellenwert. Für das untenstehende Zahlenbeispiel 8.5 ist der optimale Schwellenwert von 18.5 eingetragen, welcher mit einer Sensitivität von 0.94 und einer Spezifität von 0.76 die Gruppe der Gesunden und die Gruppe der Depressiven optimal voneinander trennt, da hier die Summe aus Sensitivität und Spezifität am größten wird.

Bei der ROC-Analyse wird nach jenem Schwellenwert gesucht, der ein optimales Gleichgewicht zwischen Sensitivität und Spezifität herstellt. Voraussetzung dazu ist eine Untersuchung, in der die Personen den Test bearbeitet haben und das Kriterium

(z.B. die Diagnose) bekannt ist. Das Vorliegen der Major Depression im Beispiel kann etwa über ein ausführliches klinisches Interview ermittelt worden sein. In der ROC-Analyse wird nun für jeden der potentiellen Schwellenwerte (Testwerte) die Sensitivität und Spezifität berechnet, die sich ergeben würde, wenn man diesen Wert als Schwellenwert verwenden würde. Dann werden die jeweils zueinander gehörigen Werte für Sensitivität und 1-Spezifität (d.h. Quote falscher Alarme) grafisch gegeneinander abgetragen. Diese Darstellung wird als **ROC-Kurve** bezeichnet (s. Abb. 8.5). Sie veranschaulicht die Verschiebung der Sensitivität zugunsten der Spezifität und gibt zugleich Aufschluss darüber, wie gut der Test überhaupt geeignet ist, zwischen Fällen, die das Kriterium, erfüllen und den übrigen Fällen zu trennen. Wenn der Test nicht zwischen den beiden Gruppen trennt, verläuft die empirische ROC-Kurve nahe der Hauptdiagonalen, d.h. Sensitivität und 1 – Spezifität sind für alle Schwellenwerte gleich groß - in diesem Fall würden die Verteilungen in Abbildung 8.3 übereinander liegen. Haben die Fälle, bei denen das Kriterium vorliegt, im Mittel höhere Testwerte als die übrigen, verläuft die Kurve oberhalb der Diagonalen. Der Schwellenwert, an dem die Summe von Sensitivität und Spezifität am größten ist, entspricht demjenigen Punkt in der ROC-Kurve, an dem das Lot auf die Hauptdiagonale den größten Abstand anzeigt. Rechnerisch lässt sich dieser Punkt auch über den **Youden-Index** (Youden, 1950) bestimmen, der als Sensitivität + Spezifität – 1 so gebildet wird, dass er Werte zwischen 0 und 1 annimmt. Mit demjenigen Schwellenwert, für den die Summe von Sensitivität und Spezifität und somit der Youden-Index am größten wird, gelingt die Trennung der beiden Gruppen am besten (vgl. nachfolgendes Beispiel 8.5). Dieser Punkt in der ROC-Kurve ist genau derjenige Punkt, an dem die Tangente parallel zur Hauptdiagonalen verläuft. Bis zu diesem Punkt nimmt der Gewinn an Sensitivität durch das Absenken des Schwellenwertes stark zu, während die Quote falscher Alarme vergleichsweise wenig ansteigt. Jenseits dieses Punktes steigt jedoch die Quote falscher Alarme schneller als die Sensitivität zunimmt, es lohnt sich also nicht, den Schwellenwert noch niedriger zu wählen.

Beispiel 8.5

Bestimmung des optimalen Schwellenwertes

Mit höherem Schwellenwert sinkt die Sensitivität, während die Spezifität ansteigt (vgl. Abb. 8.4). Die ROC-Kurve (Abb. 8.5) zeigt, dass der Test gut zwischen beiden Gruppen trennt, die Kurve liegt deutlich oberhalb der Hauptdiagonalen. Als optimaler Schwellenwert lässt sich im Beispiel ein Testwert von 18.5 ermitteln, der mit 0.69 den

höchsten Youden-Index aufweist (vgl. Tab. 8.1); der zugehörige Punkt auf der ROC-Kurve weist den maximalen Abstand zur Hauptdiagonalen auf (s. Abb. 8.5). Personen mit einem Testwert von 18 oder weniger Punkten werden also als nicht depressiv klassifiziert, bei Personen mit einem Wert von 19 oder mehr Punkten besteht der dringende Verdacht einer Major Depression. Anhand des mit der ROC-Analyse ermittelten Schwellenwertes kann der Testwert im Depressivitätsfragebogen also kriteriumsorientiert interpretiert werden. Wie aus Tabelle 8.1 hervorgeht, wird im Beispiel mit dem Schwellenwert von 18.5 eine Sensitivität von 0.94 erreicht, d.h. 94% der Patienten mit Major Depression werden korrekt klassifiziert. Zugleich werden 76% der nicht depressiven Personen korrekt klassifiziert (Spezifität) bzw. 24% der nicht depressiven Personen fälschlicherweise als depressiv diagnostiziert (1-Spezifität). Würde man den Schwellenwert z.B. auf 20.5 Punkte heraufsetzen, würden nur noch 14% der nicht depressiven Personen fälschlicherweise als depressiv diagnostiziert werden. Zugleich würden aber nur noch 80% der Patienten mit Major Depression korrekt als solche erkannt (vgl. Tab. 8.1).

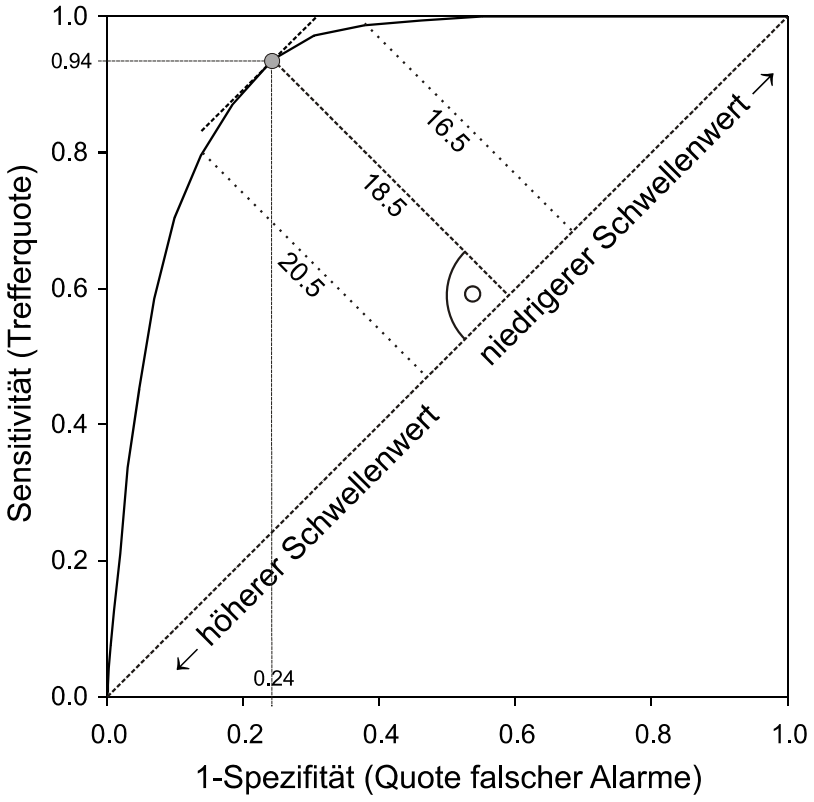

Abbildung 8.5: ROC-Kurve. Bei einem Schwellenwert von 18.5 wird der maximale Abstand zur Hauptdiagonalen erreicht, zur Illustration sind zusätzlich die Projektio-

nen eingezeichnet, die sich jeweils für einen niedrigeren (16.5) und einen höheren (20.5) Schwellenwert ergeben.

Tab. 8.1: Koordinaten der ROC-Kurve (Ausschnitt).

Schwellenwert	Sensitivität	1-Spezifität	Spezifität	Youden-Index
14.5	1.00	0.55	0.45	0.449
15.5	0.99	0.46	0.54	0.532
16.5	0.99	0.38	0.62	0.610
17.5	0.97	0.30	0.70	0.668
18.5	0.94	0.24	0.76	**0.693**
19.5	0.87	0.18	0.82	0.686
20.5	0.80	0.14	0.86	0.659
21.5	0.70	0.10	0.90	0.605
22.5	0.59	0.07	0.93	0.515
23.5	0.45	0.05	0.95	0.406
24.5	0.34	0.03	0.97	0.307

Die grafische ROC-Analyse ist ein verteilungsfreies Verfahren, d.h. die Testwerteverteilungen in den Gruppen („Positive“ und „Negative“) müssen keiner bestimmten Verteilungsfunktion folgen. Es ist für die Untersuchung auch unerheblich, wie groß die beiden anhand des Kriteriums gebildeten Gruppen im Verhältnis zueinander sind. Zur Bestimmung des Schwellenwertes können zum Beispiel gleich große Gruppen von Depressiven und Nicht-Depressiven untersucht werden, auch wenn die tatsächliche Quote De-

pressiver in der diagnostischen Praxis kleiner als 50% ist. Die beiden Gruppen müssen allerdings möglichst repräsentativ für die Populationen sein (vgl. Abschnitt 8.6), die das Kriterium erfüllen bzw. nicht erfüllen, damit der gewonnene Schwellenwert in der Praxis für die Klassifikation von neuen Fällen verwendet werden kann. Im Beispiel sollten also die Personen ohne Major Depression diejenige Population repräsentieren, aus der sich auch in der klinischen Praxis die interessierende „Vergleichsgruppe" rekrutiert – zum Beispiel die Gesamtheit der Patienten, die wegen Beschwerden in einer psychotherapeutischen Ambulanz untersucht werden, jedoch nicht das Kriterium der Major Depression erfüllen.

Anhand der ROC-Analyse können auch Schwellenwerte festgelegt werden, die ein anderes Optimierungsverhältnis ergeben als das aus dem Youden-Index resultierende. Wenn etwa die Konsequenzen einer falschen negativen Diagnose („Verpasser") schwerwiegender sind als diejenigen einer falschen positiven („falscher Alarm"), kann ein niedrigerer Schwellenwert mit einer höheren Sensitivität und niedrigeren Spezifität zweckmäßiger sein als der Wert mit der besten Balance zwischen beiden Kriterien. Wenn zum Beispiel die Konsequenz aus dem Testergebnis des Depressivitätsfragebogens lediglich in einer intensiveren Diagnostik besteht (und nicht in einer therapeutischen Entscheidung), ist ein „falscher Alarm" nicht sonderlich schwerwiegend. Einen Patienten mit Major Depression hingegen nicht als solchen zu erkennen („Verpasser"), könnte wesentlich schwerwiegendere Folgen haben, etwa das Nichtbeachten einer bestehenden Suizidgefährdung. In diesem Fall könnte es angemessener sein, nicht den „optimalen" Schwellenwert zu nehmen, sondern einen niedrigeren von zum Beispiel 17.5, womit eine Sensitivität von 97% erreicht werden würde.

8.3.2 Bezug des Testwertes auf Aufgabeninhalte

Eine zweite Möglichkeit, die Testwerte aus einem Test bezogen auf inhaltliche Kriterien zu interpretieren, ist der Bezug auf die Test- bzw. Aufgabeninhalte. Dieses Vorgehen ist dann möglich, wenn a priori eine genaue inhaltliche Vorstellung von der *theoretischen Grundgesamtheit* der für das interessierende Konstrukt relevanten Aufgaben existiert und die im Test verwendeten Aufgaben eine *Stichprobe* aus diesen möglichen Aufgaben darstellen. Die Fähigkeiten, die Schüler am Ende ihrer schulischen Ausbildung in der ersten Fremdsprache erreichen sollen, werden zum Beispiel in den Bildungsstandards der Kultusministerkonferenz beschrieben und anhand von Beispielaufgaben illustriert (Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland, 2004,

2005). Ein Vokabeltest kann so konstruiert werden, dass die Testaufgaben die in den Bildungsstandards definierten Lernziele repräsentieren. Die abgefragten Vokabeln können zum Beispiel eine zufällige Stichprobe aus der Menge der Wörter darstellen, die bei einem bestimmten Fähigkeitsniveau in einer Fremdsprache beherrscht werden sollen. Für einen solchen Test ist es möglich, den Testwert im Sinne des Ausmaßes der *Erfüllung eines Lernziels* interpretieren. Dieser Testwert wird dann als der Anteil gelöster Aufgaben interpretiert, den ein Schüler gelöst hätte, wenn man ihm alle theoretisch möglichen Aufgaben des interessierenden Fähigkeitsbereiches vorgelegt hätte (Klauer, 1987a, 1987b).

Diese Form von kriteriumsorientierter Testwertinterpretation setzt voraus, dass eine Definition der theoretischen Grundgesamtheit von Aufgaben vorgenommen werden kann. Die wesentliche Anforderung an die Aufgabenzusammenstellung ist hierbei, dass die im Test verwendeten Aufgaben eine Schwierigkeitsverteilung aufweisen, die derjenigen in der Grundgesamtheit der Aufgaben entspricht. Im Falle von Lernzielen in pädagogisch-psychologischen Kontexten ist dies besonders gut möglich, wenn auch mit einem nicht unerheblichen Aufwand verbunden. So müssen Lernziele hinreichend (z.B. bildungspolitisch) legitimiert sein und es muss unter geeigneten Experten Einigkeit über die Definition der relevanten Aufgaben bestehen.

Während die kriteriumsorientierte Interpretation auf Basis der Aufgabeninhalte also vor allem bei Lernziel- oder Leistungstests möglich ist, wenn die Grundgesamtheit möglicher Aufgaben und deren Schwierigkeitsverteilung definiert werden können, ist im Falle von Fragebögen dieses Vorgehen in der Regel nicht möglich. Dies liegt daran, dass die Schwierigkeiten der Items eines Fragebogens nicht nur durch die Inhalte, sondern auch durch die verbale Formulierung beeinflusst werden. Es ist daher ein häufig gemachter Fehler, den Testwert aus einem Fragebogen auf den „theoretischen Wertebereich“ zu beziehen, der sich aus der Antwortskala ergibt, und hieraus eine kriteriumsorientierte Interpretation abzuleiten.

Beispielsweise kann ein Fragebogen zur Erfassung depressiver Symptome eines der in Tabelle 8.2 aufgelisteten Items enthalten.

Tabelle 8.2: Zwei unterschiedlich schwierige Fragebogenitems zur Erfassung derselben depressiven Symptomatik mit einer vierstufigen Antwortskala.

	trifft nicht zu	*trifft eher nicht zu*	*trifft eher zu*	*trifft zu*
1. Ich fühle mich manchmal grundlos traurig.	①	②	③	④
2. Mich überkommt oft ohne Anlass eine tiefe Traurigkeit.	①	②	③	④

In beiden Items wird nach dem gleichen Inhalt gefragt, nämlich nach Traurigkeit ohne äußeren Anlass. Das erste Item ist jedoch so formuliert, dass ihm wahrscheinlich auch einige Personen ohne eine ausgeprägte Depression zustimmen können. Das zweite Item ist hingegen deutlich schwieriger, ihm würden wahrscheinlich deutlich weniger Personen zustimmen. Wenn nun die Antworten auf diese Items mit 1 bis 4 Punkten bewertet würden, würde sich beim ersten Item bei denselben Personen ein höherer Mittelwert ergeben als beim zweiten. Dieser Unterschied würde jedoch nichts über eine unterschiedliche Depressivität aussagen, sondern wäre allein auf die verbale Itemformulierung zurückzuführen.

Im Beispiel könnte sich etwa für das erste Item ein Mittelwert von 2.8 Punkten ergeben. Es wäre unzulässig, diesen Wert dahingehend zu interpretieren, dass die getesteten Personen „eher depressiv" seien, da die Personen dem Item „eher zugestimmt haben" und der Mittelwert über der „theoretischen Mitte" des Wertebereichs von 2.5 Punkten liegt. Wenn dieselben Personen anstelle des ersten das zweite Item vorgelegt bekämen, würde der Mittelwert niedriger liegen (zum Beispiel bei 1.6 Punkten), ohne dass die Personen deswegen „weniger depressiv" wären. Eine kriteriumsorientierte Interpretation einer Fragebogenskala anhand des möglichen Wertebereichs ist also nicht sinnvoll, es müsste in diesem Fall ein externes Kriterium (s. Kap. 8.3.1) herangezogen werden, um zu beurteilen, welche Testwerte tatsächlich mit einer depressiven Störung einhergehen.

Weitaus differenziertere Möglichkeiten zur kriteriumsorientierten Testwertinterpretation auf Basis der Aufgabeninhalte bieten Auswertungen von Tests, die mit Modellen der Item-Response-Theorie (IRT) konstruiert wurden. Diese Modelle erlauben es, die Schwierigkeiten der Aufgaben und die Messwerte der Personen auf einer gemeinsamen Skala darzustellen. Die hierdurch mögliche kriteriumsorientierte Testwertinterpretation wird in Kapitel 10.5 (Dahl & Hartig, 2007, in diesem Band) erörtert.

8.4 Integration von norm- und kriteriumsorientierter Testwertinterpretation

Norm- und kriteriumsorientierte Testwertinterpretation stellen grundsätzlich keine Gegensätze dar. Je nach diagnostischer Fragestellung kann einer der beiden Interpretationsansätze angemessener sein; sie können sich aber auch ergänzen, indem das Testresultat aus unterschiedlichen Perspektiven bewertet wird. Physikalische Maße, wie z.B. die Größe bzw. Länge eines Objektes, können hierbei als Analogie dienen. Beispielsweise wird bei einem Angelwettbewerb derjenige Angler gewinnen, der den größten oder schwersten Fisch fängt (normorientierte Diagnostik). In den gesetzlichen Regelungen, ab welcher Mindestgröße ein Fisch überhaupt von einem Angler aus dem Wasser entnommen werden darf („Schonmaß"), geht es hingegen um das theoretisch-inhaltlich definierte Kriterium, dass ein Fisch eine hinreichende Größe erreicht haben soll (kriteriumsorientierte Diagnostik), um sich wenigstens einmal im Leben fortzupflanzen zu können.

In gleicher Weise ermöglicht die Beachtung sowohl des normorientierten als auch des kriteriumsorientierten Vergleichsmaßstabes eine differenziertere Bewertung von psychologischen Merkmalsausprägungen. Beispielsweise zeigen sich Peters Eltern erfreut über die Vokabelleistung ihres Sohnes nachdem sie erfahren haben, dass die meisten Klassenkameraden von Peter weniger als 70% der Aufgaben richtig gelöst haben (normorientierte Interpretation). Allerdings wurde ihre Freude wieder etwas getrübt als sie erfuhren, dass das Lehrziel eine Lösungsrate von 90% war (kriteriumsorientierte Interpretation).

Dieses Beispiel deutet bereits an, dass ein psychologisch-inhaltliches Kriterium oft nicht völlig unabhängig von der empirischen Verteilung der Merkmalsausprägungen erfolgen kann bzw. auch mit normorientierten Überlegungen in Zusammenhang steht. So sind die Aufgaben für die Vokabelprobe derart zusammengestellt worden, dass eine Lösungsrate von 90% in der entsprechenden Schülerpopulation überhaupt erzielt werden kann.

Obwohl die Integration norm- und kriteriumsorientierter Testwertinterpretationen die angemessenste Form diagnostischer Informationsverarbeitung darstellt (Rettler, 1999), können norm- und kriteriumsorientierte Testwertinterpretation bei konfligierenden Interessenlagen zu teilweise unvereinbaren Zielsetzungen führen. Cronbach (1990) unterscheidet in Zusammenhang mit der Festlegung von Standards für Auswahlprozesse empirische, politische und entscheidungsbezogene Aspekte.

- Aus empirischer Sicht sind solche Personen z.B. für einen Studienplatz oder eine freie Stelle auszuwählen, welche mit ihrem Testwert ein psychologisch-inhaltlich gesetztes Kriterium erreichen (kriteriumsorientierter Standard) und somit wahrscheinlich einen hohen Studien- oder Berufserfolg haben werden.
- Hinzu kommen jedoch auch politische Einflüsse, welche die Anhebung oder Absenkung von Kriterien zum Ziel haben können. Aus Arbeitgebersicht mag es z.B. von Interesse sein, dass strengere Standards zur Erreichung eines Schulabschlusses gesetzt werden, um mit größerer Sicherheit geeignetes Personal einzustellen. Durch höhere Standards entstehen jedoch auch höhere soziale Kosten (z.B. durch Klassenwiederholung, Arbeitslosigkeit), sodass von Politikern ein normorientierter Standard eingefordert werden könnte, z.B. dass unabhängig vom Erreichen eines psychologisch-inhaltlich definierten Kriteriums nur die 10% Schwächsten der Population in der Abschlussprüfung durchfallen dürfen, um gesellschaftliche Folgekosten zu begrenzen (normorientierter Standard).
- Dieses Spannungsfeld von unterschiedlichen Interessen und Einflüssen führt zu einem komplexen Entscheidungsprozess, in dem Kosten und Nutzen abgewogen werden müssen sowie politische Machbarkeit beachtet werden muss. Cronbach (1990) stellt zusammenfassend fest, „standards must be set by negotiations rather than by technical analysis“ (S. 98).

8.5 Normdifferenzierung

Das Problem der **Normdifferenzierung** bezieht sich im Rahmen normorientierter Testwertinterpretation auf die für Testentwickler und Testanwender gleichermaßen bedeutsame Frage, wie spezifisch eine Vergleichs- bzw. Referenzgruppe zusammengesetzt sein soll. Im Beispiel zur Prozentrangnormierung in Kapitel 8.2.1 wurde deutlich gemacht, dass bei erstmaliger Erfassung der Konzentrationsleistung einer Testperson die Vergleichsgruppe nur solche Testpersonen beinhalten sollte, welche ebenfalls nur einmal getestet wurden, d.h. die Vergleichsgruppe stimmt hinsichtlich des Übungsgrades mit dem der Testperson überein (vgl. Moosbrugger & Goldhammer, 2007). Würde die Vergleichsgruppe aus Personen bestehen, welche schon zweimal getestet wurden, würde die Testwertinterpretation aufgrund des übungsbedingt höheren Leistungsniveaus der Vergleichsgruppe zu einer Unterschätzung der Konzentrationsleistung der Testperson

führen. Eine Differenzierung von Normen ist dann erforderlich, wenn wesentliche, d.h. mit dem Untersuchungsmerkmal korrelierte Hintergrundfaktoren der Testpersonen nicht mit jenen der Vergleichsgruppe übereinstimmen. Wird für relevante Ausprägungen auf dem Hintergrundfaktor jeweils eine eigene Norm gebildet, kann der Einfluss des Faktors auf die Testwertinterpretation kontrolliert werden. Beispielsweise kann der Übungsgrad dadurch kontrolliert werden, dass eine Normdifferenzierung nach der Testungszahl erfolgt, d.h. danach, ob es sich um die erste, zweite oder wievielte Testung handelt. Normdifferenzierungen werden nicht nur zum Ausgleich von Übungseffekten, sondern häufig auch zum Ausgleich von Alters-, Geschlechts- oder Bildungseffekten vorgenommen, d.h. es werden getrennte Normen nach Alter, Geschlecht, Bildung etc. gebildet (vgl. folgendes Beispiel 8.6).

Beispiel 8.6 (nach Cronbach, 1990)

Vorteile und Nachteile einer geschlechtsbezogenen Normdifferenzierung

Thomas erzielt in einem Test zur Erfassung „Mechanical Reasoning“ einen Testwert von 50. Nach der geschlechtsunspezifischen Prozentrangnormentabelle entspricht dem Testwert ein Prozentrang von 65. Anhand dieser Information den Erfolg von Thomas im Ausbildungsprogramm einzuschätzen, wäre allerdings verzerrend, da die meisten Auszubildenden in mechanischen Berufen männlich sind. Realistischer wird die Erfolgsaussicht also durch den Vergleich mit der männlichen Bezugsgruppe bestimmt. Es stellt sich heraus, dass der Prozentrang von 65 auf den weniger Erfolg versprechenden Wert von 50 abfällt, weil Männer im Test zur Erfassung von „Mechanical Reasoning“ im Allgemeinen höhere Testwerte erzielen als Frauen.

Clara erzielt im selben Test ebenfalls einen Testwert von 50. Während der Vergleich mit gemischtgeschlechtlichen Bezugsgruppe zu einem Prozentrang von 65 führt, steigt er bei einem Vergleich mit der weiblichen Bezugsgruppe auf 80. Steht Clara mit anderen Frauen um die Aufnahme in das Ausbildungsprogramm im Wettbewerb, hat sie somit sehr gute Erfolgsaussichten. Sofern ihre Mitbewerber im Ausbildungsprogramm hauptsächlich Männer sein werden, ist zur Abschätzung ihres Ausbildungserfolges aber ein Vergleich ihrer Testleistung mit der männlichen Bezugsgruppe angezeigt. Wie Thomas läge sie demnach nur noch im Durchschnittsbereich.

Am Beispiel von Cronbach (1990) wird deutlich, dass in Wettbewerbssituationen nicht immer der Vergleich mit der Gruppe, welche bestmöglich mit der Testperson übereinstimmt, diagnostisch am sinnvollsten ist. Vielmehr ist entscheidend, dass ein Vergleich mit der Gruppe der tatsächlichen Konkurrenten bzw. Mitbewerber vorgenommen wird, denn dadurch können Erfolgaussichten realistisch eingeschätzt und negative Auswirkungen, wie z.B. Frustrationen, vermieden werden.

Während differenzierte Normen also auf der einen Seite zu diagnostisch sinnvollen Entscheidungen verhelfen können, besteht auf der anderen Seite die Gefahr, dass durch eine zu starke Anpassung der Testnorm an bestimmte Teilpopulationen die Bedeutung des normierten Testresultats an Aussagekraft verliert und Fehleinschätzungen vorgenommen werden (*overadjustment*, Cronbach, 1990). Da beispielsweise der Testwert in Intelligenztests auch mit dem soziokulturellen Hintergrund zusammenhängt, mag man es bei Selektionsprozessen als fairer ansehen, nur Bewerber mit vergleichbarem soziokulturellen Hintergrund miteinander zu vergleichen, d.h. Testnormen nach soziokulturellem Hintergrund zu differenzieren. Dies entspricht der Grundidee des Fairnessmodells der proportionalen Repräsentation (Quotenmodell, s. z.B. Amelang & Schmidt-Atzert, 2006). Demnach gilt ein Selektionskriterium dann als fair, wenn in der Gruppe der ausgewählten Bewerber die Proportion unterschiedener Teilgruppen dieselbe ist wie in der Bewerberpopulation. Auf diese Weise werden jedoch tatsächlich vorhandene Unterschiede zwischen Bewerbern aus unterschiedlichen Teilpopulationen nivelliert. Im Extremfall kann die systematische Berücksichtigung von Unterscheidungsmerkmalen beispielsweise die äußerst bedenkliche Folge haben, dass ein 40jähriger Alkoholkranker für eine verantwortungsvolle Tätigkeit (z.B. Überwachung eines Produktionsprozesses) eingesetzt wird, sofern er in der Teilpopulation der 40jährigen Alkoholkranken sehr gute Testresultate erzielte.

Ein Kompromiss kann darin bestehen, dass durch gesellschaftspolitische Wertvorstellungen motivierte Normdifferenzierungen vorgenommen werden, gleichzeitig muss jedoch eine tatsächliche Bewältigung der Anforderungen erwartet werden können. Das obige Beispiel lässt sich in diesem Sinne auffassen, da Clara zwar auf der einen Seite durch Anwendung einer frauenspezifischen Norm leichteren Zugang zum Ausbildungsprogramm erhält als männliche Bewerber, auf der anderen Seite jedoch gleichzeitig ihre

tatsächlichen Erfolgsaussichten im Ausbildungsprogramm anhand der männerspezifischen Norm abgeschätzt werden müssen.

Eine weitere Facette des *overadjustment*-Problems besteht nach Cronbach (1990) darin, dass defizitäre Zustände in einer Teilpopulation als „normal“ bewertet werden. Das bedeutet beispielsweise, dass bei der mangelhaften Integration von Immigrantenkindern in das Schulsystem der Vergleich des Testwertes eines dieser Kinder mit der Teilpopulation der Immigrantenkinder zu der Annahme führt, dass ein „normales“ Leistungsniveau vorliegt. Dabei wird durch den inadäquaten Vergleichsmaßstab das eigentliche Problem, nämlich die sehr niedrige Schulleistung von Immigrantenkindern, unkenntlich gemacht und erforderliche Interventionen werden aufgrund des scheinbar „normalen“ Leistungsniveaus nicht initiiert.

8.6 Testeichung

Die **Testeichung** stellt den letzten Schritt einer Testkonstruktion dar und dient dazu, einen Vergleichsmaßstab bzw. Normwerte zur normorientierten Testwertinterpretation (vgl. Kap. 8.2) zu gewinnen. Dazu wird das zu normierende psychologische Testverfahren an Personen einer Normierungsstichprobe, welche hinsichtlich einer definierten Bezugsgruppe repräsentativ ist, durchgeführt.

In der DIN 33430 (s. DIN, 2002; Westhoff et al., 2005) werden zur Qualitätssicherung von diagnostischen Beurteilungen hinsichtlich der Normwerte eine Reihe von Anforderungen formuliert: „[Norm] values must correspond to the research question and the reference group [..] of the candidates. The appropriateness of the norm values is to be evaluated at least every eight years.“ (Hornke, 2005, S. 262). In den International Guidelines for Test Use (International Test Commission, ITC, 2000, S. 14) ist analog folgende Richtlinie enthalten: “Ensure that invalid conclusions are not drawn from comparisons of scores with norms that are not relevant to the people being tested or are outdated.” (vgl. auch Moosbrugger & Höfling, 2007, Kap. 9 in diesem Band).

Für die Testeichung bedeutet die für die Testanwendung geforderte Entsprechung zwischen Normwerten und Forschungsfrage, dass zu Beginn der Testeichung unter Berücksichtigung von Anwenderinteressen vom Testautor zu entscheiden ist, welche Fragen auf Basis der zu bildenden Normen beantwortet werden sollen. Beispielsweise kann die Frage nach der Rechenleistung eines Jungen in der dritten Klasse im Vergleich zu seinen Klassenkameraden nur dann beantwortet werden, wenn eine aktuelle schulspezifi-

sche Testnorm, welche auf der Bezugsgruppe der Drittklässler und nicht etwa der Viertklässler basiert, zur Verfügung steht. Das heißt, es ist zu Beginn der Testeichung genau zu definieren und entsprechend im Testmanual zu dokumentieren, für welche Bezugsgruppe bzw. **Zielpopulation**die zu erstellenden Testnormen gelten sollen. Im Rahmen der Festlegung der Zielpopulation(en) ist auch die Frage zu klären, ob eine Normdifferenzierung (vgl. Kap. 8.5) vorgenommen werden soll.

Um sicherzustellen, dass, wie in der DIN 33430 gefordert, die Testnorm der Merkmalsverteilung der Bezugsgruppe entspricht, ist bei der Erhebung der Normierungsstichprobe darauf zu achten, dass die Normierungsstichprobe bezüglich der Bezugsgruppe der Testperson **repräsentativ** ist. Hierbei ist zwischen globaler und spezifischer Repräsentativität zu unterscheiden.

Globale Repräsentativität und spezifische Repräsentativitäteiner Stichprobe

Eine repräsentative Stichprobe liegt dann vor, wenn die Stichprobe hinsichtlich ihrer Zusammensetzung die jeweilige Zielpopulation möglichst genau abbildet. Dies bedeutet, dass sich Repräsentativität immer auf eine *bestimmte* Zielpopulation bezieht bzw. dass eine Stichprobe repräsentativ bezüglich einer vorher definierten und keiner beliebigen anderen Population ist.

Eine Stichprobe wird *global repräsentativ* genannt, wenn ihre Zusammensetzung hinsichtlich aller möglichen Faktoren mit der Populationszusammensetzung übereinstimmt – dies ist nur durch Ziehen einer echten **Zufallsstichprobe** aus einer definierten Population zu erreichen. Dagegen gilt eine Stichprobe als *spezifisch repräsentativ*, wenn sie lediglich hinsichtlich derjenigen Faktoren der Populationszusammensetzung repräsentativ ist, die mit dem Untersuchungsmerkmal bzw. dem Testwert in irgendeiner Weise zusammenhängen, wobei für die Bildung von Testnormen insbesondere Geschlecht, Alter und Bildungsgrad oder Beruf von Bedeutung sind. Es sind also genau solche Faktoren, die auch Anlass zu einer Normdifferenzierung geben könnten (vgl. das o.a. Beispiel 8.6 in Kap. 8.5 zur geschlechtsspezifischen Normierung des „Mechanical Reasoning“-Tests). Mangelnde Repräsentativität kann durch einen größeren Stichprobenumfang nicht kompensiert werden, d.h. eine kleine repräsentative Stichprobe ist nützlicher als eine große, jedoch nicht repräsentative Stichprobe.

Liegen Kenntnisse vor, welche Faktoren mit dem Untersuchungsmerkmal zusammenhängen, bieten sich zur Testeichung spezifisch repräsentative **geschichtete Stichproben** bzw. **Quotenstichproben** an (vgl. Bortz & Döring, 2001). Durch diese Art der Stichprobenziehung wird erreicht, dass die prozentuale Verteilung der Ausprägungen auf merkmalsrelevanten Faktoren in der Stichprobe mit der Verteilung in der Population identisch ist. Hierfür muss jedoch in Erfahrung gebracht werden können, wie die Ausprägungen auf den Faktoren, welche mit dem Untersuchungsmerkmal zusammenhängen, in der Population verteilt sind. Beispielsweise dürften für den Faktor „Geschlecht" die Ausprägungen „weiblich" und „männlich" in vielen Zielpopulationen gleichverteilt sein bzw. mit einer Häufigkeit von jeweils 50% auftreten. Wird eine Person aus der Menge von Personen mit gleichen Ausprägungen auf merkmalsrelevanten Faktoren zufällig ausgewählt, spricht man von einer *geschichteten (stratifizierten) Stichprobe*, bei einer nicht zufälligen Auswahl dagegen von einer *Quotenstichprobe*. Soll beispielsweise ein persönlichkeitspsychologisches Testverfahren für Jugendliche normiert werden, so ist davon auszugehen, dass der Testwert mit den Faktoren familiäres Milieu und Arbeitslosigkeit zusammenhängt. Wenn aus demographischen Erhebungen bekannt ist, dass von den Jugendlichen in der Population 30% Arbeiterfamilien, 15% Unternehmerfamilien, 15% Beamtenfamilien und 40% Angestelltenfamilien angehören und der Anteil der arbeitslosen Jugendlichen je nach familiärer Herkunft gemäß o.a. Reihenfolge 5%, 1%, 2% und 3% beträgt, kann ein entsprechender Erhebungsplan nach diesen Quoten erstellt werden. Beispielsweise beträgt die Quote Jugendlicher aus Arbeiterfamilien, die arbeitslos sind, 1.5% (5% von 30%), wohingegen 28.5% der Jugendlichen aus Arbeiterfamilien beschäftigt sind. Für die Normierungsstichprobe werden Personen so ausgewählt, dass sich für die kombinierten Ausprägungen auf merkmalsrelevanten Faktoren eine prozentuale Verteilung gemäß Erhebungsplan ergibt.

Liegen zur Gewinnung von geschichteten bzw. Quotenstichproben keine Informationen über die mit dem Untersuchungsmerkmal korrelierten Faktoren vor, ist eine **Zufallsstichprobe** zu ziehen, welche zu globaler Repräsentativität führt. Idealtypisch bedeutet dies für die Erhebung der Normierungsstichprobe, dass aus der Menge aller Personen der Zielpopulation nach dem Zufallsprinzip eine bestimmte Menge von Personen ausgewählt wird. In der Praxis wird dieses Vorgehen jedoch höchstens im Falle sehr spezifischer Populationen realisierbar sein, da mit echten Zufallsstichproben ein erheblicher finanzieller, organisatorischer und personeller Aufwand sowie datenschutzrechtliche Hürden verbunden wären.

Wird eine anfallende Stichprobe oder **ad-hoc-Stichprobe** zur Bildung von Testnormen verwendet, bedeutet dies, dass nicht gezielt versucht wird, eine repräsentative Stichprobe hinsichtlich einer bestimmten Zielpopulation zu ziehen. Stattdessen werden sich bietende Gelegenheiten genutzt, um Testdaten für die Normierungsstichprobe zu sammeln. Anfallende Stichproben können an Wert gewinnen, wenn sich aus ihnen in Anlehnung an die oben beschriebene Vorgehensweise nachträglich eine Quotenstichprobe für eine bestimmte Zielpopulation bilden lässt.

Der erforderliche **Umfang der Normierungsstichprobe** hängt von verschiedenen Faktoren ab. Allgemein gilt, dass eine gebildete Norm aus empirischer Sicht im Prinzip nur so fein zwischen Testpersonen differenzieren kann, wie vorher in der Normierungsstichprobe unterschiedliche Testwerte angefallen sind, wobei eine größere Anzahl unterschiedlicher Testwerte durch eine größere repräsentative Normierungsstichprobe gewonnen werden kann. Möchte man also fein abgestufte Normwerte bilden, wie z.B. in einer Prozentrangnorm oder Standardnorm, ist eine umfangreichere Stichprobe nötig, als wenn grobstufige Normen, wie z.B. Quartile, erstellt werden sollen. Feinstufige Normen sollten jedoch nur dann angestrebt werden, wenn der Test eine hohe Reliabilität (s. Moosbrugger, 2007b, Kap. 5.5 in diesem Band) aufweist. Im Falle einer niedrigen Reliabilität würde eine feinstufige Norm zum Vergleich von Testpersonen eine hohe Genauigkeit vortäuschen, die sich aufgrund des hohen Standardmessfehlers aber als nicht gerechtfertigt erweist. Es resultieren nämlich breite Konfidenzintervalle, die zahlreiche Normwerte ober- und unterhalb des individuellen Normwertes einer Testperson einschließen. Da jedoch Werte innerhalb eines Konfidenzintervalls nicht unterschieden werden können, wäre die feine Abstufung der Normwerte bei niedriger Reliabilität ohne Nutzen.

Bei der Planung des Stichprobenumfangs ist ausgehend von der oben angeführten Überlegung weiter zu beachten, dass die potentiell mögliche Anzahl unterschiedlicher Testwerte in einer Zielpopulation von deren jeweiligen Zusammensetzung abhängt. Handelt es sich um eine relativ homogene Zielpopulation, wie z.B. Gymnasialabsolventen in einem bestimmten Bundesland, ist die Spannweite der Testwerte, die vom Bildungsgrad abhängen, begrenzter als in einer sehr heterogen zusammengesetzten Zielpopulation, wie z.B. alle 18jährigen Jugendlichen in Deutschland. Das bedeutet, dass zur Erreichung desselben Abstufungsgrades in der Testnorm bei einer heterogeneren Zielpopulation bzw. bei einem Test mit weitem Geltungsbereich eine größere Normierungsstich-

probe gezogen werden muss, als bei einer homogenen Zielpopulation bzw. bei einem Test mit engem Geltungsbereich.

Liegen die an der Normierungsstichprobe gewonnenen Daten vor, sind zunächst deren **Verteilungseigenschaften** zu überprüfen. Insbesondere ist von Interesse, ob das Merkmal normalverteilt ist, was z_v-Normen bzw. Standardnormen (s. Kap. 8.2.2) nahe legen würde. Bei fehlender Normalverteilung lassen sich Prozentrangnormen (s. Kap. 8.2.1) realisieren. Unter bestimmten Bedingungen können die Daten auch normalisiert werden (s. Kelava & Moosbrugger, 2007, Kap. 4.8.2 in diesem Band). Gegebenenfalls sind die Normen nach bestimmten Faktoren, z.B. Geschlecht, Alter, Bildungsgrad oder Beruf, zu differenzieren (s. Kap. 8.5) und entsprechend gruppenspezifische Normen zu bilden, wobei auch hier wieder die jeweiligen Verteilungseigenschaften zu untersuchen und insbesondere die Mittelwertunterschiede, welche Anlass für die Normdifferenzierung geben, auf Signifikanz und Relevanz zu überprüfen sind. Die Darstellung der gebildeten Testnormen erfolgt in der Regel tabellarisch, sodass bei Papier-Bleistift-Tests vom Testanwender der Normwert in einer Tabelle aufgesucht werden muss. Bei computerbasierten Tests oder Auswertungsprogrammen (z.B. FAIR, Moosbrugger & Goldhammer, 2005) wird der Normwert automatisch ausgegeben.

Damit es dem Testanwender möglich ist, die Angemessenheit der Normen hinsichtlich seiner Fragestellung beurteilen zu können, sind die erstellten Normen im Testmanual hinsichtlich folgender Gesichtspunkte zu dokumentieren (s. z.B. Häcker, Leutner & Amelang, 1998; s. auch Cronbach, 1990; Moosbrugger & Höfling, 2007, Kap. 9 in diesem Band):

- Geltungsbereich der Normen, d.h. Definition der Zielpopulation(en), welche diejenige(n) sein sollte, mit denen ein Anwender die Testpersonen in der Regel vergleichen will
- Erhebungsdesign bzw. Grad der Repräsentativität hinsichtlich der Zielpopulation
- Stichprobenumfang und -zusammensetzung
- Deskriptivstatistiken
- Jahr der Datenerhebung

Der letzte Dokumentationsgesichtspunkt hebt darauf ab, dass Normen über mehrere Jahre hinweg veralten können und somit Ihre Eignung als aktuell gültigen Vergleichsmaßstab verlieren (s. z.B. Beispiel 8.7 zum Flynn-Effekt, Flynn, 1999). Die in der DIN

33430 geforderte **Überprüfung der Gültigkeit von Normen** nach spätestens acht Jahren (Hornke, 2005) ist als Richtwert zu verstehen. Sofern empirische Evidenzen schon vor Ablauf der acht Jahre auf eine Änderung der Merkmalsverteilung in der Bezugsgruppe hinweisen, ist eine frühere **Normaktualisierung** angezeigt.

Beispiel 8.7

Flynn-Effekt

Der Flynn-Effekt ist ein eindrucksvolles Beispiel dafür, dass Testnormen über einen längeren Zeitraum ihre Gültigkeit verlieren können. Flynn (1999) hat gezeigt, dass in den westlichen Industrienationen der mittlere IQ über einen Zeitraum von mehreren Jahren hinweg ansteigt. Das Phänomen wurde von ihm zufällig entdeckt, als er Testmanuale studierte, in denen die Intelligenz einer Testperson mit der ersten und zudem mit der revidierten Version eines Intelligenztests bestimmt wurde. Unter Verwendung der für die jeweilige Version gebildeten Testnorm zeigte sich, dass eine Person im älteren Test einen deutlich besseren Normwert bzw. IQ erzielte als im zeitnah normierten Test. Die normorientierte Testwertinterpretation auf Basis der veralteten Norm führt also unter der Bedingung eines längsschnittlichen Intelligenzanstiegs zu einer Überschätzung der individuellen Intelligenzausprägung, da der Vergleich mit der früheren Population dem Vergleich mit einer insgesamt leistungsschwächeren Population entspricht. Nach Flynn stieg unter weißen Amerikanern zwischen 1932 und 1978 der IQ um 14 Punkte an, was einer Rate von etwa 1/3 IQ-Punkten pro Jahr entspricht.

Zur Vermeidung von Fehlern bei der Testwertinterpretation muss also die Gültigkeit von Normen regelmäßig überprüft werden. Bei einer geänderten Merkmalsverteilung in der Bezugsgruppe sollte eine Normaktualisierung bzw. eine erneute Testeichung erfolgen.

Literatur

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4. Aufl.) unter Mitarbeit von Th. Fydrich und H. Moosbrugger. Berlin: Springer.

Bortz, J. (2006). *Statistik für Human- und Sozialwissenschaftler* (6. Auflage). Berlin: Springer.

Bortz, J., Döring, N. (2001). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (3., überarbeitete Auflage). Berlin: Springer.

Cronbach, L.J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper and Row.

DIN (2002). DIN 33430: *Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen.* Berlin: Beuth.

Dorans, N.J. (1990). Scaling and equating. In H. Wainer, N.J. Dorans, R. Flaugher, B. F. Green, R.J. Mislevy, L. Steinberg & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 137-160). Hillsdale, NJ: Lawrence Erlbaum.

Flynn, J.R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54*, 5-20.

Frey, A. (2007). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.). *Testtheorie: Test- und Fragebogenkonstruktion.* Heidelberg: Springer.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley and Sons.

Häcker, H., Leutner, D. & Amelang, M. (1998). *Standards für pädagogisches und psychologisches Testen.* Göttingen: Hogrefe.

Hornke, L.F. (2005). *Die englische Fassung der DIN 33430*. In K. Westhoff, L.J. Hellfritsch, L.F. Hornke, K.D. Kubinger, F. Lang, H. Moosbrugger, A. Püschel & G. Reimann (Hrsg.), *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (2., überarb. Aufl.) (S. 255-283). Lengerich: Pabst.

International Test Commission (ITC) (2000). *International Guidelines for Test Use.* (Version 2000). USA: Author.

Kelava, A. & Moosbrugger, H. (2007). Deskriptivstatistische Itemanalyse. In H. Moosbrugger & A. Kelava (Hrsg.). *Testtheorie: Test- und Fragebogenkonstruktion.* Heidelberg: Springer.

Klauer, K.J. (1987a). *Kriteriumsorientiertes Tests.* Göttingen Hogrefe.

Klauer, K.C. (1987b). Kriteriumsorientiertes Testen: Der Schluß auf den Itempool. *Zeitschrift für differentielle und diagnostische Psychologie, 8*, 141-147.

Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.) (2004). *Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Mittleren Schulabschluss.* Neuwied: Luchterhand.

Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.) (2005). *Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Hauptschulabschluss.* Neuwied: Luchterhand.

Moosbrugger, H. (2007a). Item-Response-Theorie. In H. Moosbrugger & A. Kelava (Hrsg.). *Testtheorie: Test- und Fragebogenkonstruktion.* Heidelberg: Springer.

Moosbrugger, H. (2007b). Klassische Testtheorie: Testtheoretische Grundlagen. In H. Moosbrugger & A. Kelava (Hrsg.). *Testtheorie: Test- und Fragebogenkonstruktion.* Heidelberg: Springer.

Moosbrugger, H. & Goldhammer, F. (2005). *Computerprogramm zur computergestützten Testauswertung des Frankfurter Aufmerksamkeits-Inventar FAIR. Aktualisierte Fassung von 2005.* Göttingen: Hogrefe.

Moosbrugger, H. & Goldhammer, F. (2007). *FAKT-II. Frankfurter Adaptiver Konzentrationsleistungs-Test. Grundlegend neu bearbeitete und neu normierte 2. Auflage des FAKT von Moosbrugger und Heyden (1997).* Testmanual. Bern: Huber.

Moosbrugger, H. & Höfling, V. (2007). Standards für psychologisches Testen. In H. Moosbrugger & A. Kelava (Hrsg.). *Testtheorie: Test- und Fragebogenkonstruktion.* Heidelberg: Springer.

OECD (2001). *Knowledge and Skills for Life. First Results from the OECD Programme for International Student Assessment (PISA) 2000.* Paris: OECD.

OECD (2004). *Learning for Tomorrow's World – First Results from PISA 2003.* Paris: OECD.

Rettler, H. (1999). Normorientierte Diagnostik. In R. Jäger& F. Petermann (Hrsg.), *Psychologische Diagnostik: Ein Lehrbuch* (4. Aufl.) (S. 221-226). Weinheim: Psychologie Verlags Union.

Rost, J. (2004): *Lehrbuch Testtheorie, Testkonstruktion*. Bern: Huber.

Westhoff, K., Hellfritsch, L.J., Hornke, L.F., Kubinger, K.D., Lang, F., Moosbrugger, H., Püschel, A. & Reimann, G. (Hrsg.) (2005), *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (2., überarb. Aufl.). Lengerich: Pabst.

Youden, W. (1950). Index rating for Diagnostic Test. *Cancer, 3,* 32-35.

Zusammenfassung mit Anwendungsempfehlungen

Ob ein Testwert normorientiert (s. Kap. 8.2) oder kriteriumsorientiert (s. Kap. 8.3) interpretiert werden können soll, d.h. ob eine Idealnorm in Form eines Kriteriums (z.B. Lernziel) oder eine Realnorm in Form einer Bezugsgruppe (z.B. eine Prozentrangnorm) angelegt wird, hängt von den diagnostischen Zielsetzungen ab, für die ein Test geeignet sein soll.

Vor der Bildung einer Bezugsgruppennorm (s. Kap. 8.2, z.B. Prozentrangnorm) muss die Zielpopulation (s. Kap. 8.6) definiert werden, d.h. diejenige Population, mit der ein Testanwender den Testwert einer Testperson in der Regel vergleichen will. Um eine repräsentative Stichprobe aus der Zielpopulation zu gewinnen, kann eine Quoten- bzw. geschichtete Stichproben oder eine Zufallsstichprobe gezogen werden. Liegt zunächst nur eine ad-hoc-Stichprobe vor, sollte nachträglich die Bildung einer Quotenstichprobe für eine bestimmte Zielpopulation angestrebt werden.

Falls die Testwertvariable nicht intervallskaliert ist, kommt für eine normorientierte Testwertinterpretation nur die Bildung einer Prozentrangnorm (s. Kap. 8.2.1) in Frage, welche die relative Position eines Testwertes in der aufsteigend geordneten Rangreihe der Testwerte in der Bezugsgruppe angibt. Falls hingegen eine intervallskalierte Testwertvariable vorliegt, ist auch die Bildung einer z_v-Norm (s. Kap. 8.2.2) möglich, welche für einen Testwert seinen Abstand zum Mittelwert der Bezugsgruppe in Einheiten der Standardabweichung angibt. Wenn eine intervallskalierte Testwertvariable die Voraussetzung der Normalverteilung erfüllt, ist die z_v-Norm insbesondere von Vorteil, da anhand der tabellarisierten Standardnormalverteilung die prozentuale Häufigkeit der z-Werte innerhalb beliebiger Wertebereiche bestimmt werden kann.

Eine kriteriumsorientierte Interpretation eines Testwertes (s. Kap. 8.3) kann dadurch vorgenommen werden, dass anhand eines zusätzlich zu erhebenden externen Kriteriums auf der Testwertskala ein Schwellenwert bestimmt wird, dessen Überschreitung anzeigt, dass das Kriterium erfüllt ist (s. Kap. 8.3.1). Die ROC-Analyse stellt eine Möglichkeit dar, einen Schwellenwert empirisch zu definieren. Alternativ kann eine kriteriumsorientierte Interpretation anhand der Aufgabeninhalte erfolgen (s. Kap. 8.3.2). Dieses Vorgehen stellt jedoch deutlich höhere Anforderungen an die Aufgabenkonstruktion, da eine genaue inhaltliche Vorstellung von der Grundgesamtheit der für das zu erfassende Merkmal relevanten Aufgaben bestehen und die Testaufgaben eine repräsentative Stichprobe aus der Aufgabengrundgesamtheit darstellen müssen. Der Testwert stellt in

diesem Fall unmittelbar einen Indikator für die Merkmalsausprägung dar, da von der Leistung in der Aufgabenstichprobe auf die Leistung in der Aufgabengrundgesamtheit geschlossen werden darf. Verfahren zur Generierung repräsentativer Aufgabenstichproben werden von Klauer (1987a) beschrieben.

Mit einschlägiger Statistiksoftware können Perzentilwerte, z-Werte und ROC-Kurven inklusive der Koordinatenpunkte Sensitivität und 1-Spezifität einfach berechnet werden.

Glossar

Testwert

Der Testwert (= Rohwert) ist das individuelle numerische Testresultat und wird aus den registrierten Antworten einer Testperson durch Anwendung definierter Regeln unmittelbar gebildet.

Normorientierte Testwertinterpretation

Die normorientierte Testwertinterpretation besteht darin, dass zu einem individuellen Testwert ein Normwert bestimmt wird, anhand dessen die Testperson hinsichtlich der erfassten Merkmalsausprägung innerhalb der Bezugsgruppe positioniert wird.

Normwert

Ein Normwert (z.B. Prozentrang, z_v-Wert) ermöglicht es, den Testwert einer Testperson durch Bestimmung seiner Position in der Testwerteverteilung einer bestimmten Bezugsgruppe zu interpretieren.

Prozentrang

Ein Prozentrang gibt an, wie viel Prozent der Bezugsgruppe bzw. Normierungsstichprobe einen Testwert erzielten, der niedriger oder maximal ebenso hoch ist, wie der Testwert x_v der Testperson v.

Perzentil

Das Perzentil bezeichnet jenen Testwert x_v, der einem bestimmten Prozentrang in der Normierungsstichprobe entspricht. Beispielswiese wird derjenige Testwert, welcher von 30% der Testwerte unterschritten bzw. höchstens erreicht wird, als 30. Perzentil bezeichnet.

Quartil

Das erste, zweite bzw. dritte Quartil (Q1, Q2, Q3) ist jener Testwert x_v, der von 25%, 50% bzw. 75% der Testwerte unterschritten bzw. höchstens erreicht wird.

z_v-Normwert

Der z_v-Normwert gibt an, wie stark der Testwert x_v einer Testperson v vom Mittelwert $\bar{x}$ der Verteilung der Bezugsgruppe in Einheiten der Standardabweichung s_x abweicht.

Standardnormen

Im Falle einer normalverteilten Testwertvariablen x, werden die z-Norm und die durch Lineartransformationen der z-Norm gewonnenen Normen (z.B. IQ- oder T-Norm) als Standardnormen bezeichnet.

Kriteriumsorientierte Testwertinterpretation

Bei der kriteriumsorientierten Testwertinterpretation erfolgt die Interpretation des Testwertes in Bezug zu einem externen Kriterium oder aber dadurch, dass die Inhalte der Testaufgaben herangezogen werden, um die Testwerte inhaltlich zu beschreiben.

Schwellenwert

Im Rahmen kriteriumsorientierter Testwertinterpretation bezeichnet ein Schwellenwert jenen Testwert, ab dem ein Kriterium als zutreffend angenommen wird. Schwellenwerte können z.B. mittels ROC-Analyse empirisch bestimmt werden.

ROC-Analyse

Die Receiver-Operating-Characteristics (ROC)-Analyse ermöglicht für eine binäre Klassifikation (z.B. krank vs. nicht krank) den zur Fallunterscheidung verwendeten Schwellenwert optimal in der Weise festzulegen, dass Trefferquote und Quote korrekter Ablehnungen in der Summe maximiert werden.

Repräsentative Aufgabenstichprobe

Eine repräsentative Aufgabenstichprobe stimmt hinsichtlich der Schwierigkeitsverteilung mit der Grundgesamtheit aller merkmalsrelevanten Aufgaben überein und erlaubt somit eine kriteriumsorientierte Testwertinterpretation in Bezug auf die Aufgabeninhalte.

Normdifferenzierung

Unter Normdifferenzierung versteht man die Bildung von separaten Normen für die einzelnen Stufen eines mit dem Untersuchungsmerkmal korrelierten Hintergrundfaktors (z.B. separate Normen für Männer und Frauen).

Testeichung

Die Testeichung dient dazu, Normwerte zur normorientierten Testwertinterpretation zu gewinnen. Dazu wird der Test an Personen einer Normierungsstichprobe, welche hinsichtlich einer definierten Bezugsgruppe repräsentativ ist, durchgeführt.

Repräsentativität

Eine Stichprobe ist dann repräsentativ, wenn sie hinsichtlich ihrer Zusammensetzung die jeweilige Zielpopulation möglichst genau abbildet.

Zielpopulation

Die im Rahmen der Testeichung zu definierende Zielpopulation ist diejenige Bezugsgruppe, für die zu erstellenden Testnormen gelten soll und aus der entsprechend die Normierungsstichprobe zu ziehen ist.

Normaktualisierung

Unter Normaktualisierung versteht man die erneute Testeichung, sobald die empirische Überprüfung der Gültigkeit von Normen ergeben hat, dass sich die Merkmalsverteilung in der Bezugsgruppe seit der vorherigen Testeichung bedeutsam verändert hat.