Grabowski, F., Castello, A. & Brodersen, G. (2022). Diagnostische Gütekriterien bei Statustests. In M. Gebhardt, D. Scheer & M. Schurig (Hrsg.), *Handbuch der sonderpädagogischen Diagnostik. Grundlagen und Konzepte der Statusdiagnostik, Prozessdiagnostik und Förderplanung* (S. 247-258). Regensburg: Universitätsbibliothek. https://doi.org/10.5283/epub.53149

Diagnostische Gütekriterien bei Statustests

Friederike Grabowski, Armin Castello & Gunnar Brodersen

1 Einführung

Tests verschiedenster Arten begegnen uns nicht nur im wissenschaftlichen Kontext, sondern auch tagtäglich im Alltag. Alltägliche Tests, wie beispielsweise Produkttests von Haushaltsgeräten, unterscheiden sich jedoch sehr von den Tests, mit denen man es in der (pädagogisch-psychologischen) Diagnostik zu tun hat. Der größte Unterschied zwischen unwissenschaftlichen und wissenschaftlichen Tests liegt dabei im Vorhandensein von sogenannten *Testgütekriterien*, welche empirisch überprüft werden können. Testgütekriterien sind gute Orientierungshilfen dafür, ob eine diagnostische Methode gut und brauchbar ist für das Ziel, das damit angestrebt wird. Dabei unterscheidet man zwischen *Hauptgütekriterien* und *Nebengütekriterien*. Auch wenn die Begriffe dies suggerieren, sind die Hauptgütekriterien nicht wichtiger als die Nebengütekriterien. Für ein qualitativ hochwertiges Verfahren müssen alle Gütekriterien erfüllt sein, denn alle Gütekriterien zusammen stellen ein Instrument zur Qualitätsbeurteilung von (pädagogisch-psychologischen) Tests und anderen diagnostischen Methoden dar. Die Angaben zu den Gütekriterien für ein spezifisches Verfahren sollten in einem Manual oder Bericht zu finden sein.

Im folgenden Kapitel geht es um die drei Hauptgütekriterien Validität, Reliabilität und Objektivität sowie um die drei Nebengütekriterien Normierung, Ökonomie und Fairness.

2 Validität

Definition: »Ein Test gilt dann als valide (›gültig‹), wenn er das Merkmal, das er messen soll, auch wirklich misst und nicht irgendein anderes.« (Moosbrugger & Kelava, 2012)

Das Gütekriterium der Validität untersucht, inwieweit das vom Test gemessene Merkmal und das Merkmal, das man messen will, übereinstimmen. Validität kann in diesem Zusammenhang auch mit *Gültigkeit* übersetzt werden. Es geht hierbei immer um die Frage: *Wird mit dem Test das Merkmal gemessen, das gemessen werden soll?*

Beispiel: Stellen Sie sich vor Sie befinden sich im Mathematikunterricht und wollen die mathematischen Kompetenzen einer Schülerin erfassen. Dafür überlegen Sie sich verschiedene Textaufgaben, in die mathematische Operationen eingebaut sind, lassen die Schülerin diese

durchführen und bewerten anschließend das Ergebnis. Sie bemerken, dass die Schülerin in den Aufgaben sehr schlecht abschneidet. Dieses Ergebnis könnten Sie so bewerten, dass die Schülerin geringe mathematische Fähigkeiten hat. Jedoch haben Sie in diesem Beispiel Textaufgaben verwendet. Diese sind sprachbasiert, d.h. die mathematischen Aufgaben sind in einem sprachlichen Kontext verpackt. Hat die Schülerin zum Beispiel starke Leseschwierigkeiten, kann es sein, dass sie gar nicht in der Lage ist, die mathematische Aufgabe in der Textaufgabe herauszufiltern. Das Vorgehen wäre in diesem Beispiel daher nicht valide, da Sie nicht das Merkmal gemessen haben, welches Sie messen wollen (mathematische Kompetenzen), sondern ein anderes (Lesefähigkeit).

Die Validität ist insbesondere für den Praxiseinsatz diagnostischer Methoden relevant. Sie wird üblicherweise auf vier Arten überprüft, die im Folgenden näher beschrieben werden.

2.1 Inhaltliche Validität

Definition: »Unter Inhaltsvalidität versteht man, inwieweit ein Test oder ein Testitem das zu messende Merkmal repräsentativ erfasst.« (Moosbrugger & Kelava, 2012)

Bei der Inhaltsvalidität wird die Frage gestellt: *Sind die Testinhalte dem zu messenden Merkmal angemessen?* Dafür werden Einschätzungen von Experten und Expertinnen eingeholt, welche aufgrund »logischer und fachlicher Überlegungen« (Michel & Conrad, 1982) die Inhaltsvalidität bestimmen. Es wird somit kein bestimmter Kennwert erhoben. Die Experten und Expertinnen überprüfen, ob die Testinhalte adäquat sind, z. B. im schulischen Bereich, ob die Testaufgaben den Inhalten entsprechen, die laut Lehrplan vermittelt werden sollen. Inhaltsvalidität liegt beispielsweise dann vor, wenn Rechtschreibkenntnisse anhand eines Diktats überprüft werden.

2.2 Konstruktvalidität

Definition: »Ein Test weist Konstruktvalidität auf, wenn der Rückschluss vom Verhalten der Testperson innerhalb der Testsituation auf zugrundeliegende psychologische Persönlichkeitsmerkmale (›Konstrukte‹, ›latente Variablen‹, ›Traits‹) wie Fähigkeiten, Dispositionen, Charakterzüge, Einstellungen wissenschaftlich fundiert ist. Die Enge dieser Beziehung wird aufgrund von testtheoretischen Annahmen und Modellen geprüft.« (Moosbrugger & Kelava, 2012)

Im Zuge der Überprüfung der Konstruktvalidität soll die Frage beantwortet werden: *Wird das Merkmal (Konstrukt) entsprechend des theoretischen Modells gemessen?* Bei dieser Form der Validität geht es somit um die Abgrenzung von Konstrukten zueinander im Sinne der theoretischen Fundierung eines Tests. Bei einem Intelligenztest könnte beispielsweise überprüft werden, ob die Testaufgaben auch wirklich das Konstrukt »Intelligenz« erheben und nicht ein anderes Konstrukt wie zum Beispiel »Gewissenhaftigkeit« oder »Lesefähigkeit«.

2.3 Kriteriumsvalidität

Definition: »Ein Test weist Kriteriumsvalidität auf, wenn vom Verhalten der Testperson innerhalb der Testsituation erfolgreich auf ein ›Kriterium‹, nämlich auf ein Verhalten außerhalb der Testsituation, geschlossen werden kann. Die Enge der Beziehung ist das Ausmaß an Kriteriumsvalidität (Korrelationsschluss).« (Moosbrugger & Kelava, 2012)

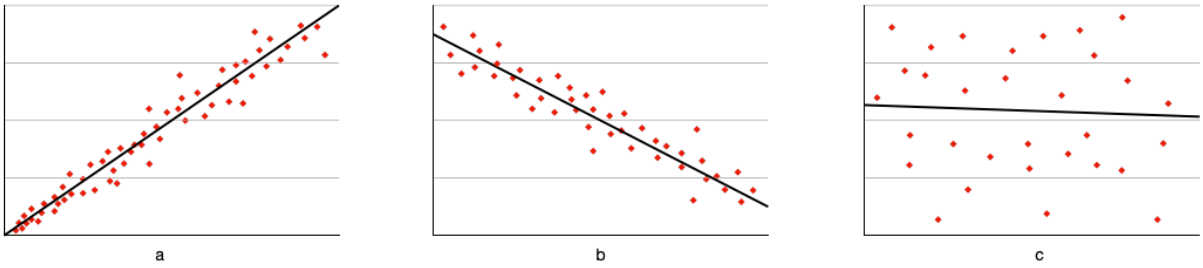

Abbildung 1: Grafische Darstellung verschiedener Korrelationen (a = positive Korrelation, b = negative Korrelation, c = keine Korrelation/unkorreliert).

Im Sinne der Kriteriumsvalidität wird folgende Frage beantwortet: *Stimmt das Testergebnis mit vergleichbaren Tests überein?* Es geht bei dieser Validitätsart somit darum, ob ein Test für die Vorhersage von Erleben und Verhalten der Testperson angewendet werden kann. Dabei kann die Kriteriumsvalidität, im Gegensatz zu den anderen Validitätsarten, statistisch überprüft werden. In der Praxis führt man zur Überprüfung der Kriteriumsvalidität das zu überprüfende Testverfahren, z. B. einen Intelligenztest, mit einer Gruppe von Personen durch und erhält ein Testergebnis. Dann führt man einen anderen, vergleichbaren Intelligenztest mit derselben Gruppe durch und erhält wiederum ein Testergebnis. Diese beiden Ergebnisse können dann statistisch verglichen (korreliert, siehe Exkurs: Korrelation) werden. Damit kann überprüft werden, ob beide Testverfahren zu ähnlichen Ergebnissen kommen. Ist die Korrelation angemessen hoch und positiv, ist der zu überprüfende Test kriteriumsvalide.

Im Schulkontext spielt die prognostische Validität, eine Variante der Kriteriumsvalidität, eine besonders wichtige Rolle. Sie wird auch als Vorhersagevalidität bezeichnet und beantwortet die Frage: *Sagt das Testergebnis ein bestimmtes Kriterium (z. B. schulische Leistung) vorher (z. B. durch Noten, Schulform)?* Diese ist beispielsweise dann von Interesse, wenn mit einem Schulleistungstest die zukünftige Leistung in der Schule vorhergesagt werden soll.

2.4 Exkurs: Korrelation

Korrelation ist das Maß für den Zusammenhang zweier Variablen. Korrelationen werden stets mit dem Korrelationskoeffizienten *r* angegeben. Dieser kann im Wertebereich von *r* = -1 bis *r* = 1 liegen. Dabei steht *r* = 1 für eine vollständige positive Korrelation. Das bedeutet, dass wenn Variable A hoch ausgeprägt ist, Variable B genauso hoch ausgeprägt ist. Umgekehrt steht *r* = -1 für eine vollständige negative Korrelation. In diesem Fall ist, wenn Variable A hoch ausgeprägt ist, Variable B niedrig ausgeprägt. In der Praxis wird man jedoch nie Korrelationen von *r* = -1 oder *r* = 1 finden, dies sind theoretische Annahmen. Korrelationseffizienten in der Praxis werden immer im Wertebereich dazwischen liegen. Korrelationen können grafisch dargestellt werden (Abbildung 1). In Abbildung 1 stellen die roten Punkte einzelne Messungen in einer Studie dar, welche den Zusammenhang zweier Variablen – dargestellt auf der X- und Y-Achse – zeigen. Die schwarze Gerade symbolisiert den Korrelationskoeffizienten. In Abbildung 1, Bild a sehen Sie ein Beispiel für eine positive Korrelation; Abbildung 1, Bild b zeigt eine negative Korrelation. Die untersuchten Variablen in Abbildung 1, Bild c weisen keinen Zusammenhang auf, d.h. sie sind unkorreliert.

Abbildung 2: Darstellung von zwei Zielscheiben

Wichtig bei der Betrachtung von Korrelation ist, dass diese keine einseitigen Kausalaussagen zulassen, da dies methodisch falsch ist. Bei korrelativen Ergebnissen werden immer zwei Variablen gleichzeitig erhoben. Daher ist die Aussage, dass eine Variable die andere beeinflusst oder die Ursache für diese ist, niemals zulässig. Um Kausalaussagen zum Zusammenhang zweier Variablen treffen zu können, muss immer ein experimentelles Vorgehen mit der bewussten Manipulation einer der beiden Variablen angewendet werden.

3 Reliabilität

Definition: »Ein Test ist dann als reliabel einzustufen, wenn er das Merkmal, das er misst, exakt, d.h. ohne Messfehler, misst.« (Moosbrugger & Kelava, 2012)

Mit dem Gütekriterium der Reliabilität wird die Messgenauigkeit eines Verfahrens überprüft, Reliabilität kann somit mit *Zuverlässigkeit* übersetzt werden. Es geht dabei um die Frage: *Wird das Merkmal zuverlässig, d.h. möglichst fehlerfrei, gemessen?* Messfehler treten bei jeder Testung mit jedem Testverfahren auf. Damit ein Test als reliabel einzustufen ist, sollte dieser Messfehler jedoch möglichst gering sein.

Beispiel: Es soll herausgefunden werden, wer von zwei Personen der bessere Bogenschütze ist. Dafür schießen beide einmal auf eine Zielscheibe (Abbildung 2, Bild a). Aus diesem einmaligen Schuss kann jedoch keine Aussage darüber abgeleitet werden, wer von beiden besser schießt. Beide Personen müssen mehrmals schießen, um ein zuverlässiges, daher reliables Ergebnis zu erhalten (Abbildung 2, Bild b). Durch das mehrmalige Durchführen zeigt sich, dass Schütze 2 der bessere Schütze ist.

Das Gütekriterium der Reliabilität kann auf verschiedene Arten überprüft werden, die häufigsten Vorgehensweisen werden im Folgenden näher erläutert.

3.1 Retest-Reliabilität

Retest bedeutet *Testwiederholung*. Bei dieser Reliabilitätsüberprüfung soll folgende Frage beantwortet werden: *Stimmen die Ergebnisse bei einer Messwiederholung überein?* Bei einem

Rechtschreibtest könnte die Retest-Reliabilität beispielsweise überprüft werden, indem der Test mit einer Klasse durchgeführt und nach 3-6 Wochen mit derselben Klasse wiederholt wird (unter der Annahme, dass es in dieser Zeitspanne zu keiner bzw. einer konstanten Lernentwicklung kommt). Die beiden Testergebnisse können miteinander korreliert und damit eine Aussage über die Retest-Reliabilität des Rechtschreibtests getroffen werden. Zu beachten bei dieser Form der Reliabilität ist jedoch, dass die ermittelte Korrelation je nach gewähltem Zeitintervall zwischen den Testungen variieren kann. Dies liegt darin begründet, dass Lern- oder Erinnerungseffekte auftreten können, welche sich auf das Testergebnis und somit die Reliabilität auswirken und diese verfälschen, d.h. überschätzen, können.

3.2 Split-Half-Reliabilität

Bei der Split-Half-Reliabilität wird der Frage nachgegangen: *Stimmen die Ergebnisse bei der Teilung des Tests in zwei Testhälften überein?* Hierfür wird ein zu überprüfender Test mit einer Gruppe durchgeführt und danach in zwei Hälften geteilt. Somit ergeben sich zwei Ergebnisse (Ergebnis Testhälfte A und Ergebnis Testhälfte B), welche miteinander korreliert werden können. Meist wird hierbei ein Korrekturfaktor angewendet, um die Halbierung zu berücksichtigen und den Reliabilitätskoeffizienten aufzuwerten (z. B. durch die Spearman-Brown-Formel) (Schermelleh-Engel & Werner, 2012).

3.3 Interne Konsistenz

Das Reliabilitätsmaß der internen Konsistenz ist eine Verallgemeinerung der Split-Half-Reliabilität. Hierbei wird jedes Item (jede Frage) als eigener Testteil angesehen und mit allen anderen Items korreliert. Es wird somit der *Zusammenhang zwischen den Testitems* untersucht. Das Ergebnis wird immer mit *Cronbach's α* angegeben, einem speziellen Reliabilitätskoeffizienten. Je stärker Items untereinander positiv korrelieren und je höher damit auch Cronbach´s α ist, desto höher ist die interne Konsistenz und desto reliabler ist das Testverfahren in dieser Hinsicht. Cronbach´s α kann einen Wert zwischen Null und Eins annehmen ($0 <$ Cronbach´s $\alpha < 1$) (Schermelleh-Engel & Werner, 2012). Ein Cronbach´s α von 1 drückt das Nichtvorhandensein von Messfehlern aus, ein Cronbach´s α von 0 bedeutet hingehen, dass das Testergebnis ausschließlich durch Messfehler zustande gekommen ist. Bei einem guten Test sollte Cronbach´s α über 0.7 liegen (Moosbrugger & Kelava, 2012).

3.4 Paralleltest-Reliabilität

Bei dieser Form wird die Reliabilität bestimmt, indem die *Ergebnisse zweier »paralleler« Testformen* korreliert werden. Hierfür wird mit inhaltlich ähnlichen Items (»Itemzwillingen«) ein zweiter Test entwickelt. Als parallel gelten die Tests dann, wenn sie trotz unterschiedlicher Items zu den gleichen wahren Werte und Varianzen der Testwerte führen (Moosbrugger & Kelava, 2012). Beide Tests werden mit einer Gruppe durchgeführt, um die Ergebnisse vergleichen zu können. Mit dieser Methode der Reliabilitätsbestimmung können u.a. Erinnerungs- und Übungseffekte vermieden werden.

4 Objektivität

Definition: »Ein Test ist dann objektiv, wenn er dasjenige Merkmal, das er misst, unabhängig von Testleiter [Testleiterin] und Testauswerter [Testauswerterin] misst. Außerdem müssen klare und anwenderunabhängige Regeln für die Ergebnisinterpretation vorliegen.« (Moosbrugger & Kelava, 2012)

Das Gütekriterium der Objektivität steht für die *unabhängigkeit* eines Testverfahrens. Dabei steht folgende Frage im Vordergrund: *Ist das Testergebnis unabhängig vom Testdurchführenden?* Mit der Objektivität wird somit die Vergleichbarkeit der Testergebnisse verschiedener Testpersonen gesichert. In der Praxis bedeutet dies, dass die Testdurchführenden und Testauswertenden in ihrem Verhalten keinen Spielraum bei der Durchführung, Auswertung und Interpretation eines Testverfahrens haben. Bei einer vollständigen Objektivität würden jede und jeder Testdurchführende sowie Testauswertende bei einer bestimmten Testperson zu einem gleichen Ergebnis und einer gleichen Interpretation gelangen. Das Gütekriterium der Objektivität gilt dann als erfüllt, wenn das Testverfahren in allen drei Bereichen (Durchführung, Auswertung, Interpretation) in einer Weise definiert ist, dass der Test unabhängig von Ort, Zeit, Testdurchführendem und Testauswertendem durchgeführt werden könnte und eine bestimmte Testperson dennoch dasselbe Ergebnis erreichen würde (Moosbrugger & Kelava, 2012).

Die drei genannten Objektivitätsbereiche werden im Folgenden genauer erläutert.

4.1 Durchführungsobjektivität

Ein Testverfahren ist dann objektiv bezüglich der Durchführung, wenn es für das Testergebnis irrelevant ist, welcher bzw. welche Testdurchführende den Test mit einer Testperson durchführt. Durchführungsobjektivität kann erreicht werden, indem *eindeutige Instruktionen und Anweisungen durch den bzw. die Testdurchführenden* gegeben werden. Besonders gut gelingt dies mit standardisierten Tests, in denen die Durchführungsbedingungen von den Testautoren und -autorinnen im Testmanual festgelegt sind. Diese Festlegungen können sich auf das Testmaterial, die mündlich und schriftlich dargebotenen Instruktionen sowie mögliche Zeitbegrenzungen der Aufgaben beziehen. Auch der Umgang mit Fragen oder das Einlegen von Pausen sollte festgelegt sein. Bezüglich der Instruktionen gibt es beispielsweise oft wortwörtlich vorgegebene Instruktionen, die vorgelesen werden. Das Einzige, was sich bei einem durchführungsobjektiven Verfahren somit unterscheiden sollte, ist das Verhalten der Testperson in der Testsituation.

Beispiel: Stellen Sie sich vor, Sie führen ein Testverfahren zur Erfassung schriftsprachlicher Kompetenzen in Ihrer Klasse und ein Kollege von Ihnen mit der Parallelklasse durch. Ein Schüler meldet sich und fragt: »Ich verstehe Aufgabe 6 nicht, was soll ich dort machen?« In diesem Fall ist es für Sie wichtig zu wissen, ob Sie dem Schüler helfen und die Aufgabe erklären dürfen oder ob Sie sagen müssen: »Tut mir leid, da kann ich dir nicht weiterhelfen. Versuch einfach dein Bestes!« Helfen Sie dem Schüler durch Erklären der Aufgabe, ihr Kollege hingegen reagiert auf die Nachfrage mit der Aussage »Versuch dein Bestes!«, gibt es ein unterschiedliches Verhalten der Testdurchführenden, welches sich auf die Testergebnisse auswirken kann. Der Test wäre in diesem Fall nicht mehr objektiv hinsichtlich der Durchführung.

4.2 Auswertungsobjektivität

Die Auswertungsobjektivität ist dann gegeben, wenn die Ergebnisse einer Testperson nicht von der testauswertenden Person abhängen. Dies ist zu erreichen, wenn *präzise Anweisungen und Schablonen* zur Auswertung vorliegen. Schablonen können zum Beispiel über Antwortbögen von Testpersonen gelegt werden, um die Auswertung möglichst fehlerfrei zu gestalten. Auswertungsobjektivität ist bei einigen Antwortformaten, z. B. Multiple Choice Aufgaben, meist ohne Probleme zu erreichen. Bei Antwortformaten wie offenen Antworten müssen jedoch klare Regeln zur Auswertung vorliegen, um diese objektiv zu gestalten. Je einheitlicher die Anweisungen und Schablonen von auswertenden Personen angewandt werden, desto auswertungsobjektiver ist das Testverfahren. Das Maß der Auswertungsobjektivität kann bestimmt werden, indem zwei oder mehr Testauswertende den Test einer bestimmten Testperson auswerten und die Übereinstimmung der Testergebnisse überprüft wird. Erreicht die Testperson bei jedem und jeder Testauswertenden das gleiche Testergebnis, ist die Auswertungsobjektivität optimal.

4.3 Interpretationsobjektivität

Bei standardisierten Testverfahren liegen auch bezüglich der Interpretation der Testleistungen *eindeutige Kriterien und Interpretationshinweise* vor. Interpretationsobjektivität ist dann gegeben, wenn verschiedene Testanwendende bei Testpersonen mit identischen Testergebnissen die gleichen Schlussfolgerungen treffen. Im Testmanual sollten somit klare Vorgaben dazu sein, wie bestimmte Testergebnisse zu interpretieren sind. Helfen kann hierbei die ausführliche Angabe von Normtabellen der Normstichprobe (vgl. Abschnitt »Normierung«), welche den Vergleich einer Testperson zu einer relevanten Bezugsgruppe ermöglicht (Goldhammer & Hartig, 2012).

5 Normierung

Definition: »Unter der Normierung (Eichung) eines Tests versteht man das Erstellen eines Bezugssystems, mit dessen Hilfe die Ergebnisse einer Testperson im Vergleich zu den Merkmalsausprägungen anderer Personen eindeutig eingeordnet und interpretiert werden können.« (Moosbrugger & Kelava, 2012)

Bei der Normierung eines Testverfahrens wird dieses mit einer großen, repräsentativen Gruppe von Personen (Normstichprobe) durchgeführt, um für individuelle Testungen *Vergleichswerte* zu erhalten. Dies ist relevant, damit erreichte Ergebnisse in diagnostischen Verfahren richtig eingeordnet werden können. Berichtet ein Schüler beispielsweise Zuhause, er hätte »20 Punkte im Deutschtest erreicht!« ist es wichtig zu wissen, ob dies im *Vergleich mit anderen* ein hohes oder niedriges Ergebnis darstellt. Mit dem Verfahren der Normierung wird sichergestellt, dass Testergebnisse eingeordnet und interpretiert werden können. Diese Normwerte liegen dem Testmanual bei. Führen Sie beispielsweise einen Test zur Erfassung mathematischer Kompetenzen mit einer 10-jährigen Schülerin durch, können Sie in den Normwerten nachgucken, wie Ihre Schülerin im Vergleich mit anderen 10-jährigen Mädchen abschneidet und das individuelle Testergebnis so einordnen. Ohne Normierung können Sie hingegen keine Aussage darüber treffen, ob ein individueller Testwert bezogen auf die Altersklasse oder soziale Bezugsnorm vom getesteten Kind durchschnittlich ist oder nicht.

Bei der Normierung muss der Geltungsbereich der entsprechenden Norm klar festgelegt werden. Dies bedeutet, dass die Normstichprobe *repräsentativ* für die Grundgesamtheit derjenigen Personen sein muss, für die der Test anwendbar ist (Moosbrugger & Kelava, 2012). Führen Sie einen Rechtschreibtest beispielsweise mit einer Klasse von 9-11-jährigen Schülerinnen und Schülern durch, müssen auch für 9-11-jährige Schülerinnen und Schüler Normwerte vorliegen. Zudem sollte sichergestellt sein, dass die Normierung aktuell ist. Laut DIN 33430 (Westhoff & Hagemeister, 2010) ist beispielsweise vorgesehen, dass bei Tests zur berufsbezogenen Eignungsbeurteilung nach spätestens 8 Jahren die Gültigkeit der Norm überprüft und gegebenenfalls eine Neunormierung durchgeführt werden muss. Eine Neunormierung ist dann angezeigt, wenn Lerneffekte in der Population auftreten oder sich Testergebnisse im Durchschnitt in der Testpopulation verändern.

Exkurs: Normalverteilung

In Testverfahren gemessene Merkmale können in einer Verteilung wiedergegeben werden, meistens der *Normalverteilung* (Abbildung 4), da diese als Informationen nur den Mittelwert (M) sowie die Standardabweichung (SD) benötigt. Die auch *Gaußsche Glockenkurve* genannte Verteilung zeigt an, wo der Mittelwert liegt. In Abbildung 4 ist dies in der Mitte beim z-Wert 0. Zudem ist die Standardabweichung eingezeichnet, dies sind in Abbildung 4 die Bereiche, in denen die Prozentzahlen eingetragen sind. In Abbildung 4 sind vom Mittelwert ausgehend bis zu drei Standardabweichungen oberhalb und unterhalb des Mittelwertes eingezeichnet. Jeder Wert, der innerhalb der ersten Standardabweichung oberhalb und unterhalb des Mittelwerts liegt, gilt als durchschnittlich (z. B. beim IQ-Wert der Bereich von 85-115, siehe Abbildung 4). In der Normalverteilung liegen daher 68,2 % im durchschnittlichen Bereich. Ein Testergebnis, das mehr als eine Standardabweichung unterhalb des Mittelwerts liegt (z. B. IQ < 85), gilt als unterdurchschnittlich. Bei mehr als zwei Standardabweichungen (z. B. IQ < 70) gilt das Ergebnis als weit unterdurchschnittlich. Im Gegensatz dazu gilt ein Testergebnis, welches mehr als eine Standardabweichung oberhalb des Standardwerts liegt (z. B. IQ > 115) als überdurchschnittlich und bei mehr als zwei Standardabweichungen (z. B. IQ > 130) oberhalb gilt es als weit überdurchschnittlich.

Werte innerhalb der Normalverteilung können in verschiedenen Skalen wiedergegeben werden, z. B. der IQ-Skala mit einem Mittelwert von 100 und einer Standardabweichung von 15. Sehr auffällig sind Testergebnisse dann, wenn diese mehr als zwei Standardabweichungen ober- oder unterhalb des Mittelwertes liegen. So deutet ein IQ-Wert kleiner als 70 beispielsweise auf eine geistige Behinderung hin und ein IQ-Wert höher als 130 spricht für eine besondere Begabung. Eine besondere Bedeutung haben zudem die *Prozenträngе* (siehe Abbildung 4, letzte Zeile). Fast jedes Testverfahren arbeitet mit Prozenträngen, durch die ersichtlich wird, wie eine individuelle Testperson im Vergleich zur Normstichprobe abschneidet.

Beispiel: Bei einem Testverfahren zur Erfassung der Rechtschreibleistung erreicht ein Schüler einen Prozentrang von 11. Dies bedeutet, dass nur 11 % der Vergleichsgruppe (Normstichprobe) einen Wert erreicht haben, der kleiner oder gleich dem Testergebnis des Schülers ist. D.h. 89 % der Normstichprobe haben besser abgeschnitten als der betreffende Schüler.

Bei den Prozenträngen ist sehr wichtig zu beachten, dass sich mit diesen keine Mittelwerte oder Durchschnitte berechnen lassen. Diese Berechnungen sind nur dann zulässig, wenn Skalen gleiche Abstände haben, d.h. Intervallniveau besitzen. Aus diesem Grund werden zusätzlich zu Prozenträngen oft *T-Werte* bei Testwerten angegeben. Auch diese sind normalverteilt (siehe

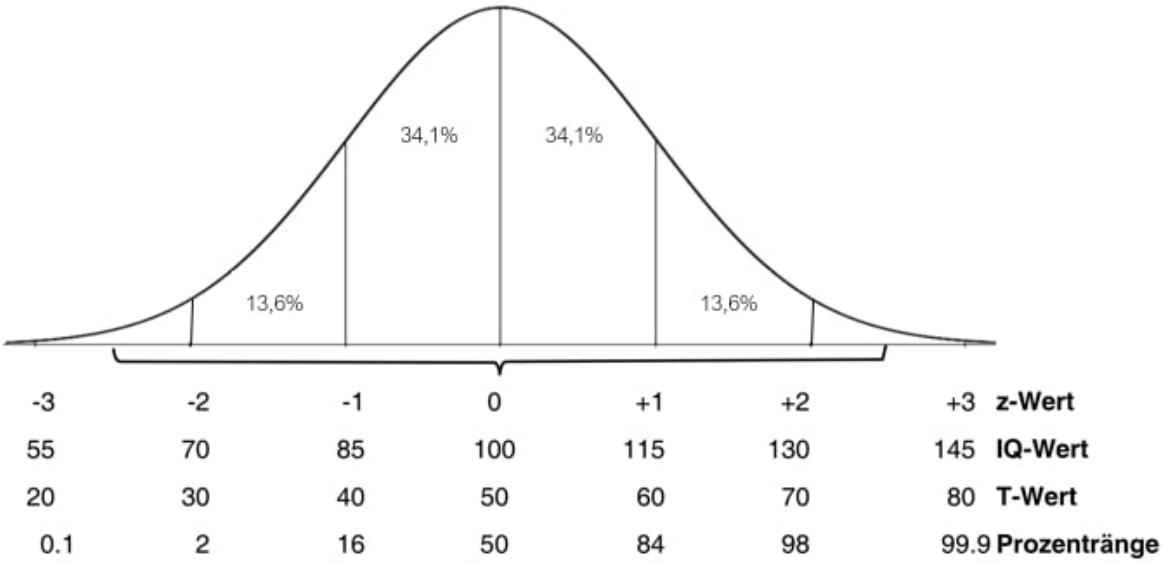

Abbildung 3: Darstellung der Normalverteilung

Abbildung 3) mit einem Mittelwert von 50 und einer Standardabweichung von 10. Mit den T-Werten ist somit bestimmbar, ob sich ein Testergebnis noch im Normalbereich oder im unter- bzw. überdurchschnittlichen Bereich befindet.

Bei vielen Testverfahren werden zusätzlich *Prozentrangbänder, T-Wert-Bänder* oder *Konfidenzintervalle* angegeben, da es bei jeder Testung zu Messfehlern kommt. Dies kommen durch unterschiedliche Umstände zustande, z. B. zu welcher Tageszeit getestet wird, wie müde die Testperson ist, usw. Deswegen hat es sich etabliert, dass zusätzlich zu »eindeutigen« Werten wie beispielsweise T-Werten auch T-Wert-Bänder angegeben werden (siehe Abbildung 4). Diese Bänder stellen dar, mit welcher Wahrscheinlichkeit (meistens 90 % oder 95 %) man darauf vertrauen kann, dass der wahre Wert (d. h. der Wert einer Testperson ohne Messfehler) in diesem Bereich liegt. Beispielsweise würde sich bei einem 95 %-igen Band/Intervall bei 100 Messungen der wahre Wert in 95 Fällen innerhalb des Bandes/Intervalles befinden. Die Angabe von T-Wert-Bändern bzw. Vertrauensintervallen ist daher sehr sinnvoll, um eine verlässlichere Aussage über die Ergebnisse von Testpersonen treffen zu können.

Beispiel: In Abbildung 4 ist der Ausschnitt einer Normtabelle zu finden. In der linken Spalte ist der Rohwert abzulesen. Die Rohwerte setzen sich meistens aus der Summe der Punkte eines Tests zusammen. In der zweiten Spalte kann man die jeweils zugehörigen Prozentränge finden und in der dritten Spalte die zugehörigen T-Werte. Das T-Wert Band ist in der letzten Spalte ersichtlich. Stellen Sie sich nun vor, Sie haben den Test durchgeführt und die Testperson hat einen Rohwert von 86. In der Normtabelle nachgeguckt, kann dem Rohwertbereich (82-86) ein Prozentrang von 2.4 sowie ein T-Wert von 28 zugeordnet werden. Das T-Wert-Band liegt bei 25-31 (siehe Abbildung 4). Mit diesen Angaben ist es möglich einzuschätzen, ob ein Rohwert von 86 eher gering oder eher hoch ist. In diesem Beispiel ist der Rohwert als gering zu beurteilen, da – vom Prozentrang ausgehend – 97.6 % der Normstichprobe ein besseres Ergebnis erreicht haben. Der T-Wert von 28 zeigt zudem, dass ein Rohwert von 86 weit unterdurchschnittlich ist, da dieser über zwei Standardabweichungen (T-Wert-Verteilung: M = 50, SD = 10) unterhalb des Mittelwertes liegt. Unter Berücksichtigung des 95 %-igen T-Wert-Bandes liegt der wahre Wert der Testperson zwischen 25-31.

Rohwert	Prozent-rang	T-Wert	T-Wert-Band
42–64	0.7	20	17–23
65–70	0.9	21	18–24
71–72	1.2	23	20–26
73–74	1.6	24	21–27
75–79	1.7	26	23–29
80–82	2.2	27	24–30
83–86	2.4	28	25–31
87–89	3.0	30	27–33
90–91	3.7	31	28–34
92–94	4.7	32	29–35
95–96	5.2	33	30–36
97–99	6.4	34	31–37

Abbildung 4: Ausschnitt aus einer Normtabelle

5.1 Ökonomie

Definition: »Ein Test erfüllt das Gütekriterium der Ökonomie, wenn er, gemessen am diagnostischen Erkenntnisgewinn, relativ wenig finanzielle und zeitliche Ressourcen beansprucht.« (Moosbrugger & Kelava, 2012)

Testökonomie ist dann gegeben, wenn das Verhältnis zwischen Nutzen und Aufwand eines Testverfahrens angemessen ist. Diese *Wirtschaftlichkeit* eines Tests bezieht sich sowohl auf finanzielle als auch auf zeitliche Aspekte. Bei der Lernverlaufsdiagnostik stellt die Ökonomie ein sehr wichtiges Gütekriterium dar, da man sehr häufig testet und ein Test lang genug sein muss, um ein aussagekräftiges, zuverlässiges und vergleichbares Ergebnis zu liefern, aber kurz genug, um im Unterricht eingesetzt werden zu können (Schurig et al., 2021). Finanziell verursacht ein Test Aufwand, in dem das Testverfahren angeschafft werden und verbrauchtes Testmaterial (z. B. Testhefte) immer wieder neu gekauft werden müssen. Zudem können bei computergestützten Verfahren Kosten für Software oder Lizenzgebühren entstehen. Der zeitliche Aufwand setzt sich nicht ausschließlich aus der Testdurchführung zusammen. Es muss zusätzlich Zeit für die Vorbereitung, Auswertung und Ergebnisbesprechung eingeplant werden. Um das Gütekriterium der Ökonomie zu erfüllen, muss der Erkenntnisgewinn durch die Testung somit höher sein als der entstehende finanzielle und zeitliche Aufwand (Moosbrugger & Kelava, 2012). Die Ökonomie eines Testverfahrens lässt sich oftmals nur im Vergleich zu anderen, ähnlichen Testverfahren bestimmen. Insbesondere computerbasierte Tests können dieses Gütekriterium vergleichsweise einfach erreichen. Es ist dabei wichtig zu erwähnen, dass eine höhere Ökonomie nicht zu Lasten anderer Gütekriterien fallen darf. So kann eine geringere Ökonomie bei einer konkreten Fragestellung dann in Kauf genommen werden, wenn aus Validitätsgründen nur ein bestimmtes Testverfahren für diese Fragestellung infrage kommt.

5.2 Fairness

Definition: »Ein Testverfahren erfüllt das Gütekriterium der Fairness, wenn die resultierenden Testwerte zu keiner systematischen Benachteiligung bestimmter Personen aufgrund ihrer

Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen führen.« (Moosbrugger & Kelava, 2012)

Das Gütekriterium der Fairness bezieht sich auf das Ausmaß, indem Testpersonen verschiedener Gruppen (z. B. Frauen vs. Männer) bei der Testdurchführung und -interpretation *fair, d.h. nicht diskriminierend*, behandelt werden. Die Fairness bezieht dabei zum einen auf die Inhalte der Testitems, bei denen ein »Itembias« vermieden werden sollte. Dieser liegt dann vor, wenn »die Aufgaben systematisch für verschiedene Personengruppen unterschiedlich schwierig sind« (Moosbrugger & Kelava, 2012). Zudem sollte der sonderpädagogische Förderbedarf kontrolliert werden, welches bei vielen Tests, vor allem in Hinblick auf die verschiedenen Förderschwerpunkte, nicht gegeben ist. Beispielsweise sind der Großteil der Tests nicht für besondere Gruppen wie Menschen mit Sehbeeinträchtigungen konzipiert (Capovilla & Kober, 2019). Der Intelligenztest CFT 1-R (Weiß & Osterland, 2013) ist ein Beispiel für einen Test, welcher auch bei Personen mit sonderpädagogischem Förderbedarf fair eingesetzt werden kann (Heine, Gebhardt, Schwab, Neumann, Gorges, & Wild, 2018). Zusätzlich dazu sollten auch die Testbedingungen für alle Testpersonen gleich sein und alle Testpersonen sollten die gleichen Voraussetzungen (z. B. Geschlecht, Muttersprache) haben. In diesem Zusammenhang kommen »Culture-Fair-Tests« eine besondere Bedeutung zu. Bei diesen Tests hängt das Lösen einer Aufgabe nicht oder nicht stark von den sprachlichen Kompetenzen der Testperson ab. Das bedeutet, dass die Aufgaben so gestaltet sind, dass das Verstehen der Instruktion und das Lösen der Aufgabe nicht davon abhängt, wie gut die Testperson die jeweilige Sprache beherrscht und auch nicht von anderen Fähigkeiten, die mit der Zugehörigkeit zu einer soziokulturellen Gruppe verbunden sind (Moosbrugger & Kelava, 2012). Der »Culture-Fair«-Ansatz stellt jedoch für viele Tests eher einen Grundgedanken als eine perfekte Umsetzung dar, da trotz der kulturfairen Intention meist etwas »Kultur-Konfundierung« bestehen bleibt (Süß, 2003). Ebenfalls im Zusammenhang mit der Fairness eines Tests steht die Testerfahrung und die Vertrautheit mit Testsituationen einer Testperson.

Zur Beurteilung der Fairness eines Testverfahrens gibt es keine festgelegten Maße oder Regeln. Daher ist jedes diagnostische Verfahren individuell hinsichtlich dieses Gütekriteriums einzuschätzen.

Literatur

Capovilla, D. & Kober, A. (2019). Intelligenzdiagnostik bei Kindern mit einer Sehbeeinträchtigung. Empirische Sonderpädagogik 11(1), 31-52. https://doi.org/10.25656/01:17769

Goldhammer, F. & Hartig, J. (2012). Interpretation von Testresultaten und Eichung. In H. Moosbrugger & A. Kelava (Hrsg.), Testtheorie und Fragebogenkonstruktion. Mit 66 Abbildungen und 41 Tabellen (Springer-Lehrbuch, 2., aktualisierte und überarbeitete Auflage, S. 173–201). Springer.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. Journal of Child Psychology and Psychiatry, and Allied Disciplines, 38(5), 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Heine, J. H., Gebhardt, M., Schwab, S., Neumann, P., Gorges, J., & Wild, E. (2018). Psychometric properties of the CFT 1 R for students with special educational needs. Psychological Test and Assessment Modeling, 60(1).

Michel, L. & Conrad, W. (1982). Theoretische Grundlagen psychometrischer Tests. In K.-J. Groffmann & L. Michel (Hrsg.), Enzyklopädie der Psychologie (Bd. 6, S. 19–70). Hogrefe.

Moosbrugger, H. & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), Testtheorie und Fragebogenkonstruktion (S. 7–26). Springer. https://doi.org/10.1007/978-3-642-20072-4_2

Schermelleh-Engel, K. & Werner, C. (2012). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), Testtheorie und Fragebogenkonstruktion (S. 119–141). Springer.

Schurig, M., Jungjohann, J. & Gebhardt, M. (2021). Minimization of a Short Computer-Based Test in Reading. Frontiers in Education, 6, Artikel 684595. https://doi.org/10.3389/feduc.2021.684595

Süß, H.-M. (2003). Culture fair. In K. D. Kubinger & R. S. Jäger (Hrsg.), Schlüsselbegriffe der Psychologischen Diagnostik (S. 82–86). Beltz.

Westhoff, K. & Hagemeister, C. (Hrsg.). (2010). Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430 (3., überarbeitete Auflage). Pabst.

Weiß, R.H. & Osterland, J. (2013). CFT 1-R. Grundintelligenztest Skala 1 – Revision. Hogrefe.

Dipl.-Psych. Friederike Grabowski ist Wissenschaftliche Mitarbeiterin in der Abteilung Sonderpädagogische Psychologie, Institut für Sonderpädagogik, Europa-Universität Flensburg. https://orcid.org/0000-0002-8559-3874

Dipl.-Psych. Prof. Dr. Armin Castello ist Professor für Psychologie und Diagnostik, Abteilung Sonderpädagogische Psychologie, Institut für Sonderpädagogik, Europa-Universität Flensburg. https://orcid.org/ 0000-0002-8994-7203

Dipl.-Psych. Dr. Gunnar Brodersen ist Dozent für pädagogisch-psychologische Diagnostik und Förderung in der Abteilung Sonderpädagogische Psychologie, Institut für Sonderpädagogik, Europa-Universität Flensburg. https://orcid.org/ 0000-0003-3029-7071