Zwischen bezugsgruppen- und kriteriumsorientierter Leistungsmessung

Lars Tischler
Medical School Hamburg

Die vorliegende Arbeit führt ein in die grundlegenden Unterschiede zwischen bezugsgruppen- und kriterienorientierter Leistungsmessung. Die teilweise Unvereinbarkeit beider Formen der Leistungsbewertung wird in seinen testtheoretischen Grundzügen mit Bezug auf Stichprobenparameter, Invarianzbedingung, Reliabilität, Validität und Itemanalyse dargestellt. Ergänzt wird um einen Exkurs zum Reliabilitäts-Validitäts-Dilemma der Veränderungsmessung. Als bedeutsam erweist sich die Arbeit für pädagogisch-psychologische Zuordnungsstrategien. Ausführungen zur Leistungsbewertung an Hochschulen stellen einen aktuellen juristischen Bezug her.

1. Bezugsnormen

Um eine Leistung, etwa das Ergebnis in einem Schulleistungsdiagnostikum zu einem bestimmten Messzeitpunkt, zu bewerten, stehen regelmäßig verschiedene Referenz-/Vergleichs-/*Bezugsgrößen* zur Verfügung.

Dies kann a) die Leistung desselben Schülers zu einem weiteren Messzeitpunkt sein. Dieser **individuelle Bezug** setzt die Leistung in Bezug zu der Leistung derselben Person zu einem oder mehreren anderen Messzeitpunkt/en. Der individuelle Bezug ermöglicht also den Leistungsvergleich im Sinne einer Verlaufs-/*Prozessdiagnostik* (Veränderungsmessung). Sie ermöglicht Aussagen darüber, ob sich die Leistung einer Schülerin in einem bestimmten Zeitraum verbessert oder verschlechtert hat, oder ob sie gleichgeblieben ist.

Individueller Bezug: Die Leistung wird relativiert an der Leistung derselben Person zu einem oder mehreren anderen Messzeitpunkt/en.
Sozialer Bezug: Die Leistung wird relativiert an der Leistung anderer Personen anhand eines relativen Lage- oder Streuungsmaßes.
Kriterialer Bezug: Die Leistung wird relativiert an einem vorher definierten, absoluten, Lehrziel.
Ipsativer Bezug: Die Partialleistung wird relativiert an der Total- oder einer anderen Partialleistung zum selben Messzeitpunkt (Profilanalyse).

Weiters kann die Leistung b) in Bezug gesetzt werden zu der Leistung anderer Personen (**sozialer Bezug**), die dasselbe Leistungsdiagnostikum durchgeführt haben. Diese *bezugsgruppenorientierte* Leistungsmessung erlaubt den Vergleich von Leistung verschiedener Individuen, sodass diese in eine *Rangreihe* gebracht werden können. Sie wird auch *sozialrelativierende* Leistungsmessung genannt. Ermöglicht werden so vergleichende Aussagen über die relative Position eines Schülers in der Rangreihe im Vergleich zu anderen Schülern (besser als, schlechter als).

Als statistische Kennwerte stehen hier Lage- und Streuungsmaße zur Verfügung. Verwendung finden regelmäßig der *Mittelwert*[1] als Maß der zentralen Tendenz und die *Standardabweichung*[2] als

[1] bei Intervallskalenniveau; $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

[2] $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}}$

durchschnittliche absolute[3] Abweichung vom Mittelwert. Mittelwert und Standardabweichung stehen in Relation zu und werden bestimmt durch die Stichprobenleistung. Sie erweisen sich entsprechend als *relative* Bezugsgrößen.

Darüber hinaus kann die Leistung c) in Bezug gesetzt werden zu einem vorher definierten Lehrziel (**kriterialer Bezug**). Ein solches *Kriterium* kann zum einen in einem finalen Lehrziel bestehen und zum anderen in einem beliebigen anderen, inhaltlich definierten, Punkt auf einem Kontinuum des Leistungs-/**Kompetenz**erwerbs. Entsprechend kann das Kriterium auch verstanden werden als Standard oder *Kompetenzstufe*[4]. Das Erreichen einer Kompetenzstufe zeigt sich in der Beherrschung eines bestimmten Lösungsverhaltens in der Testsituation. Im Gegensatz zur relativen bezugsgruppenorientierten Messung erweist sich das Kriterium als unabhängig von der Stichprobenleistung und somit als *absolut*. Zum Erreichen des Kriteriums erweist sich die Leistung der Bezugsgruppe entsprechend als völlig irrelevant[5].

Zuletzt kann die Leistung der einzelnen Testperson d) als Totalleistung in Bezug gesetzt werden zu den einzelnen, diese bildenden, Partialleistungen zum selben Messzeitpunkt derselben Testperson. Die (Teil-)Ergebnisse eines Diagnostikums werden dabei an den (Teil-)Ergebnissen desselben oder eines anderen Diagnostikums relativiert. Hierbei handelt es sich um den sogenannten **ipsativen**[6] **Bezug**. Dieser findet etwa Anwendung in der **Profilanalyse** bei komplexen Intelligenz- oder Rechendiagnostika sowie bei der Bestimmung der sogenannten *IQ-Diskrepanz* bei der Diagnostik von Umschriebenen Entwicklungsstörungen schulischer Fertigkeiten. Hier wird die Leistung in einem Schulleistungsdiagnostikum relativiert an der aufgrund der Intelligenz eigentlich zu erwartenden Leistung.

2. Probleme der bezugsgruppenorientierten Messung

Bei der bezugsgruppenorientierenden/sozialrelativierenden Leistungsmessung wird die Leistung des Einzelnen an der Leistung der Stichprobe relativiert. Dies kann etwa die Leistung der Mitschülerinnen und Kommilitonen bei Klassenarbeiten oder Klausuren sein. Bei Testverfahren stellt die Leistung der Eichstichprobe diese Referenzgröße dar. Wird also ein Testverfahren an einer leistungsschwachen

---

[3] Dies bezeichnet die gemeinsame Betrachtung von Abweichungen sowohl oberhalb als auch unterhalb des Mittelwerts. Andernfalls würden sich die Abweichungen gegenseitig zu Null aufsummieren. Aus diesem Grunde werden auch bei der Varianz s^2 die Abweichungen quadriert ($-x \cdot -x = +x^2$).

[4] Bei der internationalen Schulleistungsvergleichsstudie PISA (*programme for international student assessment*) etwa sind Kompetenzstufen formal „so definiert, dass Schülerinnen und Schüler auf dieser Stufe zugeordnete Schwellenitems mit einer bestimmten Wahrscheinlichkeit (62 Prozent) lösen. Aufgaben, die einer höheren Kompetenzstufe entsprechen, werden mit einer sehr viel geringeren Wahrscheinlichkeit gelöst“ (PISA-Konsortium Deutschland [Hrsg.]. [2005]. *PISA 2003. Der zweite Vergleich der Länder in Deutschland – Was wissen und können Jugendliche?* Münster: Waxmann).

[5] Eine mittelbare Relation zur Bezugsgruppe besteht gegebenenfalls in einer grundsätzlichen Ermöglichung der Lehrzielerreichung durch Teile der Bezugsgruppe. Andernfalls erübrigte sich die Definition eines—nunmehr unerreichbaren—Kriteriums.

[6] zu lat. *ipse* = selbst, eigen

Eichstichprobe normiert, erweisen sich bei normalgesunden Testandinnen und Testanden gute Testergebnisse als wahrscheinlicher denn bei einer leistungsstarken Eichstichprobe. Wird etwa ein Schulleistungsdiagnostikum an einer relativ leistungsschwachen Eichstichprobe normiert, führt die Leistung von Schülerinnen und Schülern beim fertigen Testverfahren dann regelmäßig zu besseren Ergebnissen.

Um derartige bezugsgruppenbezogene Verzerrungen zu vermeiden, wird eine für die sogenannte *Grundgesamtheit* repräsentative Stichprobe angestrebt. Dies kann etwa eine randomisierte Stichprobe von Schülerinnen und Schüler einer bestimmten Jahrgangsstufe aller Bundesländer sein. Wie verhält es sich allerdings bei einer nichtrepräsentativen Stichprobe, wie es etwa die Studierendenschaft eines Semesters einer einzelnen Hochschule darstellt?

Was ist messen?

Unter einer Messung versteht man die Überführung eines empirischen Relativs in ein numerisches Relativ. Das bedeutet, dass die zwischen messbaren empirischen Größen bestehenden Relationen anhand von Zahlen ausgedrückt werden. Die bestehenden Relationen/Proportionen bleiben dabei erhalten. Entsprechend bezeichnet man eine Messung auch als *homomorphe Abbildung* (*homo* = gleich, gleichartig; *-morph* = die Gestalt betreffend, -förmig).

Bei Klausuren steht als bezugsgruppenorientierte Referenzgröße allein der Leistungsdurchschnitt der *aktuellen* Klausurteilnehmerinnen und -teilnehmer zur Verfügung (hierbei ist angenommen, dass von Dozierenden regelmäßig neue Klausurfragen erarbeitet werden; vgl. auch Abschn. 5). Die vorliegenden Ergebnisse (empirisches Relativ) werden dann in Noten (numerisches Relativ) überführt, wobei immer das *gesamte* Notenspektrum von 1.0 bis 5 ausgeschöpft wird. Dies geschieht unabhängig davon, wie hoch oder niedrig das Leitungsniveau der entsprechenden Lerngruppe tatsächlich ist. In jedem Fall wird die Leistung eines bestimmten Anteils von Studierenden mit 5 (nicht bestanden) zu bewerten sein. Im Umkehrschluss bedeutet das, dass auch bei völlig unzureichender Leistung Prüfungsergebnisse mit 1.0 zu bewerten sind, wenn der Leistungsdurchschnitt entsprechend niedrig ausfällt.

Zwei Beispiele mögen dies verdeutlichen: Angenommen in einer Klausur mit 10 Teilnehmern und 100 zu vergebenden Punkten erreiche eine einzige Person 100 Punkte, und die weiteren Ergebnisse lauteten 1 · 97.5 Punkte, 6 · 95, 1 · 92.5 sowie 1 · 90 Punkte, dann betrüge der Leistungsdurchschnitt $\bar{x} = \frac{100 + 97.5 + 570 + 92.5 + 90}{10}$ = 95 Punkte. An dieser Punktzahl erfolgte nun die Relativierung der individuellen Leistungen von 1.0 bis 5. Das schlechteste Ergebnis betrüge 90 Punkte und würde mit 5 (*nicht bestanden*) bewertet. In einer weiteren Klausur mit ebenfalls 100 zu erreichenden Punkten und 10 Teilnehmern würden folgende Ergebnisse erreicht: 1 · 20 Punkte, 1 · 15, 6 · 10, 1 · 5 und 1 · 0 Punkte. Die durchschnittliche Leistung betrüge $\bar{x} = \frac{20 + 15 + 60 + 5 + 0}{10}$ = 10 Punkte. Die höchste erreichte Punktzahl 20 würde mit 1 (*sehr gut*) bewertet. Das bedeutet, dass allein in Abhängigkeit von der Gesamtleitung der Stichprobe ein Klausurergebnis von 20 Punkten mit *sehr gut*, eine deutlich bessere Leistung von 90 Punkten hingegen mit *nicht bestanden* bewertet würde. Wenig begabte Studierende können also gute

Noten erzielen, wenn sich ihre Stichprobe insgesamt als leistungsschwach erweist. Dieselben Studierenden würden in einer leistungsstarken Gruppe regelmäßig deutlich schlechtere Noten erhalten.

2.1 Die Bedeutung der Varianz in der bezugsgruppenorientierten Messung

Streuung, Varianz und Standardabweichung

Bei der bezugsgruppenorientierten Messung werden zur Relativierung von Messergebnissen Lage- und Streuungsmaße herangezogen. Eine sinnvolle Interpretation eines Testwerts kann ausschließlich bei einer angemessenen Verteilung der Merkmalsausprägungen erfolgen. Als ideal erweist sich gemäß der klassischen Testtheorie die sogenannte Normalverteilung[7].

Erhalten in einem Test mit 10 möglichen Punkten alle Testpersonen den gleichen Gesamtwert von beispielsweise 5 Punkten, beträgt der Mittelwert der Punkteverteilung 5 und die Standardabweichung—als durchschnittliche Abweichung vom Mittelwert—0. Eine sinnvolle Relativierung der Testergebnisse mittels Lage- und Streuungsmaßen kann entsprechend nicht erfolgen. Die Testwerte bilden keine Rangreihe und können daher nicht – beziehungsweise nur an sich selbst – relativiert werden.

Eine sinnvolle Relativierung der Ergebnisse kann also nur erfolgen, wenn die Testergebnisse ausreichend breit um den Mittelwert $\bar{x}$ streuen (Streuung). Statistisch ausgedrückt heißt das, dass die Ergebnisse über eine möglichst große Varianz s^2 (vgl. Fn 6) verfügen sollen (die Quadratwurzel der Varianz $\sqrt{s^2}$ stellt die Standardabweichung dar). Hierzu wird bei der Normierung von Testverfahren eine Auswahl geeigneter Items auf Grundlage der sogenannten **Itemanalyse** getroffen. Nur so kann eine angemessene Ausprägung der Gütekriterien eines Tests erreicht werden (s. u. *Varianz und Reliabilität*). Die Itemanalyse bezieht sich auf die Schwierigkeit, die Homogenität[8] und die Trennschärfe der einzelnen Aufgaben/Items. Im nächsten Abschnitt soll auf die Schwierigkeit und die Trennschärfe eingegangen werden.

Schwierigkeitsindex

Eine angemessene Varianz der Testitems wird erreicht über ihre inhaltliche[9] sowie eine Variation ihrer Schwierigkeit. Die entsprechende Lösungswahrscheinlichkeit einer Aufgabe wird ausgedrückt als **Schwierigkeitsindex** P (veralt. = Popularitätsindex). Wird eine Aufgabe von 20 von 100 Testpersonen

---

[7] Standardisiert zur Standardnormalverteilung (z-Transformation) verfügt sie über einen Mittelwert von $\bar{x} = 0$ und eine Standardabweichung (Quadratwurzel der Varianz = $\sqrt{s^2}$) von $s = 1$ (μ = 0 und σ = 1; die Verwendung griechischer Buchstaben kennzeichnet die Grundgesamtheit, während lateinische Buchstaben bei Maßzahlen für eine Stichprobe Verwendung finden).

[8] Homogenität besteht ähnlicher Schwierigkeit (P etwa .4 bis .6) und inhaltlicher Ähnlichkeit der Aufgaben. Dies zeigt sich in einer hohen Korrelation der Items untereinander. Bei inhaltlicher Itemhomogenität besteht bei komplexen Konstrukten jedoch regelmäßig keine angemessene **Konstruktrepräsentanz**.

[9] vgl. Konstruktrepräsentanz (Konstruktvalidität) und Inhaltsvalidität.

gelöst, beträgt der Schwierigkeitsindex $P = \frac{20}{100} = .2$.[10] Am ehesten eignen sich inhaltlich heterogene Items mit einen Schwierigkeitsindex um .5.[11] Werte mit extremen Schwierigkeitsindices sind in der **Itemselektion** zu verwerfen.

Der **Schwierigkeitsindex P** gibt die relative Anzahl richtiger Lösungen eines Items an (Lösungswahrscheinlichkeit einer Aufgabe). Diese Maßzahl kann einen **Wert von 0 bis 1** annehmen. Ein hoher Schwierigkeitsindex 1 steht hierbei für eine leichte Aufgabe, die von allen Testandinnen und Testanden gelöst werden kann, ein Schwierigkeitsindex von 0 entsprechend für eine Aufgabe, die von niemandem gelöst wird. Der Schwierigkeitsindex errechnet sich aus der relativen Anzahl von Personen, die eine Aufgabe lösen.

Items mit Schwierigkeitsindices von P nahe 1 erzeugen einen **Deckeneffekt**. Hierbei wird von vielen Testpersonen der vorgegebene Messbereich *über*schritten ($P = 1$, alle lösen und erreichen den maximalen Testwert [Decke]: keine Differenzierungsfähigkeit im oberen Leistungsspektrum, Test zu einfach). Items mit einem Schwierigkeitsindex nahe 0 erzeugen einen **Bodeneffekt**. Der vorgegebene Messbereich wird hier von vielen Testpersonen *unter*schritten ($P = 0$, niemand löst, alle erreichen den minimalen Testwert [Boden]: keine Differenzierungsfähigkeit im unteren Leistungsbereich, Test zu schwierig).

Hieraus folgt, dass Items mit mittlerem Schwierigkeitsindex am besten zwischen Lösern und Nicht-Lösern trennen können. Sie erweisen sich entsprechend als besonders *trennscharf*.

Trennschärfe

Die Trennschärfe r_{it} kann **Werte von -1 bis +1** annehmen. Sie gibt an, inwieweit die Beantwortung eines einzelnen Items kennzeichnend ist für das Gesamtergebnis einer Testperson. Es handelt sich also um ein Maß für die Korrelation r zwischen der Beantwortung eines Items i und dem Gesamttestergebnis t. Entsprechend kann davon ausgegangen werden, dass bei einer Trennschärfe von 1 das einzelne Item etwas Ähnliches misst wie der Gesamttest. **Als gut gelten Trennschärfen von .4 bis .7** (s. Moosbrugger & Kelava, 2008, S. 84).

Items mit einer Trennschärfe r_{it} von 0 lassen aufgrund fehlender Korrelation zwischen Item und Gesamtergebnis keine Aussagen für das Gesamtergebnis einer Person zu. Offensichtlich messen diese Items etwas anderes als der Gesamttest. Solche Items sind in der Itemselektion zu verwerfen. Items mit einer Trennschärfe von r_{it} nahe -1 verhalten sich entgegengesetzt zum Gesamttestergebnis. Personen, die diese Items lösen, schneiden im Gesamtergebnis schlecht ab, und Personen, die diese Items nicht lösen, schneiden im Gesamtergebnis gut ab. Items mit solch unvorteilhaften Trennschärfen sollten in der Itemselektion verworfen werden. Sie liefern keine dienlichen inhaltlichen Informationen und erweisen sich als zeitökonomisch von Nachteil.

---

[10] Da P maximal einen Wert von 1 annehmen kann, werden Werte unter 1 ohne Vorkommastelle angegeben (Bsp.: 0.5 wird dargestellt als .5).

[11] Am besten in Löser und Nichtlöser differenzieren können dichotome Items (ja/nein; 1/0) mit P = .5.

Zusammenfassend lässt sich feststellen, dass Items mit mittlerer Schwierigkeit ($P = .5$) über die größte Trennschärfe verfügen. Bei extremen Schwierigkeitsindices (1 und 0) beträgt die Varianz der Ergebnisse = 0, da entweder niemand ($P = 0$) oder alle ($P = 1$) die Aufgabe lösen. Entsprechend können solche Items nicht zwischen Lösern und Nicht-Lösern differenzieren und erweisen sich als nicht trennscharf.

Aus den vorangegangenen Ausführungen wird deutlich, dass sich auch Trennschärfe und Schwierigkeitsindex im gegebenen Rahmen bezugsgruppenorientierter Messungen immer auf die ursprüngliche Stichprobe beziehen. Ändert sich diese **Eichstichprobe** (Normstichprobe), so ändert sich auch die Verteilung von Merkmalsausprägungen, die als Referenz zur Leistungsbewertung in der Praxis herangezogen werden soll. Andersherum formuliert bedeutet dies, dass ein Testverfahren, das an einer klinischen Stichprobe normiert worden ist, sinnvoll ausschließlich zur Differenzierung im unteren Leistungsbereich Anwendung finden kann. Bei normalgesunden Testandinnen und Testanden käme es regelmäßig zu einem Deckeneffekt—viele Testpersonen würden alle Items lösen ($P = 1$) und ein etwaiger Förderbedarf bei schlechten Schülerinnen und Schülern würde unterschätzt. Umgekehrt verhielte es sich etwa mit einem Intelligenzdiagnostikum, das zur Hochbegabungsdiagnostik eingesetzt werden soll. Die grundlegende Eichstichprobe sollte Probandinnen und Probanden aufweisen, die tatsächlich über eine weitüberdurchschnittliche Intelligenz ($+ 2\,SD, IQ \leq 130$) verfügen, andernfalls träte ein Bodeneffekt auf—viele Testpersonen würden keine Aufgabe lösen ($P = 0$).

In beiden Fällen betrüge die Trennschärfe r_{it}= 1, da die Lösung eines einzelnen Items hoch mit dem Gesamtergebnis korrelierte. Wer das eine Item löst, löst alle Items, wer dieses Item nicht löst, löst kein Item. Selbstverständlich erweist sich die tatsächliche Aussagekraft der Trennschärfe hier als nichtig. Die Trennschärfe kann nur bei ausreichender Streuung der Testwerte sinnvoll interpretiert werden. Die zu messende Merkmalsausprägung (wahrer Wert τ) muss also über eine ausreichend hohe Varianz $s^2(\tau)$ verfügen. Anders ausgedrückt lässt sich die Trennschärfe ausschließlich bei einer ausreichend hohen Reliabilität $\frac{s^2(\tau)}{s^2(x)}$ der Messung sinnvoll interpretieren (s. nächster Abschnitt).

Der **wahre Wert** (true score) ist die wahre, tatsächliche Ausprägung eines untersuchten Merkmals. Eine idealpräzise Messung soll diesen Wert möglichst exakt numerisch abbilden. Gelingt eine 1:1 Abbildung, erweist sich die Messung als absolut reliabel (r_{tt} = 1). Gemäß KTT erweist sich der wahre Wert als unveränderlich (Invarianzbedingung; s. u.)

Varianz und Reliabilität

Die Reliabilität (Zuverlässigkeit) besteht in der formalen Präzision des Messvorgangs und der daraus resultierenden Exaktheit des Messergebnisses. Dabei wirkt sich der Messvorgang idealiter auch bei wiederholter Messung nicht auf das Messobjekt aus. Folglich erweist sich bereits eine einmalige

Messung als zuverlässig. Bei wiederholten Messungen resultierten die gleichen Messwerte. Das bedeutet auch, dass sich die Reliabilität nicht auf inhaltliche Aspekte des Messens bezieht.

Die Varianz s^2 der gemessenen Merkmalsausprägungen spielt für die Bestimmung der Reliabilität eine wesentliche Rolle. Denn rechnerisch lässt sich die Reliabilität als Korrelation r_{tt} zwischen der Varianz der (theoretischen) wahren Werte[12] (τ) und der Varianz der (tatsächlichen) diese möglichst exakt abbildenden Testwerte (x) darstellen[13]. Je exakter die gemessenen Testwerte also in der Lage sind, die angenommenen wahren Werte abzubilden, als desto reliabler erweist sich die Messung. Dies beinhaltet entsprechend nicht allein die unterschiedlichen wahren Werte und gemessenen Merkmalsausprägungen, sondern selbstverständlich auch deren Verteilung und die entsprechenden Relationen/Proportionen zwischen den Werten. Somit sollen Mittelwert und die Streuung – gemessen als Varianz – bei den wahren Werten mit den gemessenen Merkmalsausprägungen übereinstimmen. Kann eine Messung dies leisten, ergibt sich $r_{tt} = \frac{s^2(\tau)}{s^2(x)} = 1$. Zähler und Nenner—Varianz der wahren Werte $s^2(\tau)$ und Varianz der Testwerte $s^2(x)$—sind identisch. Die Messung erweist sich als frei von Messfehlern.

Messfehler

Da es sich bei der **Klassischen Testtheorie** (KTT) um eine *Messfehlertheorie* handelt, wird davon ausgegangen, dass eine Messung *ausnahmslos* fehlerbehaftet ist. Der einzelne Testwert (x) setzt sich also immer anteilig aus wahrem Wert (τ) und zufälligem Messfehler (ε)[14] zusammen. Dieses *Verknüpfungsaxiom* bezieht sich im Falle der Reliabilitätsberechnung auf die entsprechenden Varianzen. Die Varianz der Testwerte $s^2(x)$ im Nenner der Formel zur Reliabilitätsbestimmung $r_{tt} = \frac{s^2(\tau)}{s^2(x)}$ kann entsprechend ebenso aufgeteilt werden in wahre Varianz plus Fehlervarianz: $s^2(x) = s^2(\tau) + s^2(\varepsilon)$.

Je reliabler nun die Messung, desto geringer wird im Nenner der Anteil der Fehlervarianz $s^2(\varepsilon)$ an der gesamten Testwertvarianz $s^2(x) = s^2(\tau) + s^2(\varepsilon)$. Andersherum formuliert wird hierbei der Anteil der wahren Varianz $s^2(\tau)$ an der Testwertvarianz $s^2(x)$ immer größer. Damit nähern sich Testwertvarianz $s^2(x)$ im Nenner und wahre Varianz $s^2(\tau)$ im Zähler dem Verhältnis 1:1 und damit die Reliabilität r_{tt} dem Wert 1 an.

Es wird deutlich, dass die Reliabilität mit der relativen Abnahme der Fehlervarianz beziehungsweise mit der Zunahme der wahren Varianz steigt. Eine angemessene wahre Varianz kann anhand der Itemanalyse (s. o.) erreicht werden.

3. Reliabilität und Validität bei der kriterienorientierten Leistungsmessung

[12] Bei einer unendlichen Anzahl von Messungen entspricht der Mittelwert der Messergebnisse als sogenannter *Erwartungswert* tatsächlich dem wahren Wert (*Existenzaxiom* der Klassischen Testtheorie). Messfehler sind hierbei ausgemittelt.

[13] r_{tt} meint die Korrelation r zwischen t = true score (wahrer Wert) und t = test score (Testwert).

[14] ε = Error

Wie bereits ausgeführt, wird eine Leistung bei der bezugsgruppenorientierten Messung relativiert an der Leistung anderer Personen (Lage- und Streuungsmaße). Das bedeutet, dass eine Leistung *unabhängig* von der Erreichung eines inhaltlich definierten Lehrziels (Kriterium) bewertet wird. Dies mag sich zwar als sinnvoll erweisen, wenn durch vorangegangene Selektionsprozesse bereits ein einheitliches Leistungs- beziehungsweise **Kompetenz**niveau innerhalb einer Stichprobe vorausgesetzt werden kann. Dies ist jedoch häufig nicht der Fall, wie etwa die fortwährende Kontroverse um die Vergleichbarkeit von Schulnoten, Studienabschlüssen oder anderen Zugangsberechtigungen und Qualifikationen zeigt.

Während sich bei der bezugsgruppenorientierten Messung also die Verteilung—und somit die Varianz—von Merkmalsausprägungen für die Bewertung einer Leistung als entscheidend erweist, spielt gerade dies für die kriteriale Leistungsbewertung *keine* Rolle. Das zuvor inhaltlich definierte Lehrziel (Kriterium, Standard, Kompetenzstufe) gilt als erreicht, wenn ein bestimmtes Lösungs*verhalten* in der Testsituation zum gewünschten Ergebnis führt. Und dies gilt völlig unabhängig davon, ob andere Testandinnen und Testanden dieses Verhalten ebenfalls zeigen (das Lehrziel erreichen) oder nicht. Entsprechend erweist sich die Relativierung der Leistung auch als völlig unabhängig von Lage- und Streuungsmaßen—somit auch als völlig unabhängig von der Varianz.

Unter **Kompetenz** wird die Befähigung zu *bestimmten* Verhaltensweisen—etwa das Lösungsverhalten in einer Testsituation zur Erreichung einer inhaltlich definierten **Kompetenzstufe**—verstanden.
Damit weist die Kompetenz im Gegensatz etwa zur Intelligenz einen deutlichen Handlungsbezug auf. Da sich **Verhalten** im Gegensatz zu den transsituativ und -temporal stabilen Eigenschaften als variabel erweist, kommt Kompetenzen in der pädagogischen Verhaltensoptimierung eine besondere Bedeutung zu (vgl. Veränderungsmessung und Modifikationsdiagnostik).

Grenzen der Reliabilität

Da die Varianz der wahren $s^2(\tau)$ und der Testwerte $s^2(x)$ benötigt wird, um die Reliabilität $r_{tt} = \frac{s^2(\tau)}{s^2(x)}$ einer Messung zu bestimmen, stößt das Reliabilitätskonzept der Klassischen Testtheorie (KTT) hier an seine Grenzen. Ein Beispiel mag dies verdeutlichen: Das Ideal im Sinne des pädagogischen **Optimierungsgrundsatzes** stellt das Erreichen des Lehrziels durch *sämtliche* Testpersonen dar. Idealiter besteht also keinerlei Varianz in den Testergebnissen. Bei idealer Unterrichtung sind sämtliche Schülerinnen und Schüler einer Jahrgangstufe in der Lage, ein bestimmtes, vorher definiertes, Lösungsverhalten (Kompetenz) zu zeigen. Das heißt nichts anderes, als dass zwischen den einzelnen Testandinnen und Testanden keine offenbaren Leistungsunterschiede bestehen, die Varianz der wahren Werte $s^2(\tau)$ theoretisch also 0 beträgt. Gemäß der KTT als Messfehlertheorie ergibt sich zwar eine Fehlervarianz $s^2(\varepsilon)$, dies erweist sich für die Reliabilitätsbestimmung jedoch als irrelevant. Da gemäß der Formel $r_{tt} = \frac{s^2(\tau)}{s^2(x)}$ hier aufgrund fehlender wahrer Varianz ohnehin eine Null im Zähler steht, beträgt die Reliabilität unabhängig von der Fehlervarianz in jedem Falle 0. Die Messung erweist sich entsprechend als absolut unzuverlässig. Das Gütekriterium der Reliabilität lässt sich demnach, zumindest theoretisch und

tendenziell, *nicht* auf die kriterienorientierte Leistungsmessung anwenden. Die KTT stößt hier an ihre Grenzen.

Grenzen der Validität

Diese Unvereinbarkeit setzt sich mit Bezug auf die Validität fort. Für die kriteriumsorientierte Diagnostik erweist sich zwar zunächst die **Inhaltsvalidität** als bedeutsam. Inhaltsvalidität ist gegeben, wenn die Testitems eine möglichst repräsentative Stichprobe der Aufgabengrundgesamtheit—dem Aufgaben- oder *Itemuniversum*—darstellen: „Das wesentliche Charakteristikum eines kriteriumsbezogenen Tests ist eben der eindeutige Bezug zu einem genau definierten Verhaltensbereich" (Hilke, 1980, S. 55). Die Inhaltsvalidität ist regelmäßig augenscheinlich gegeben (Augenscheinvalidität) oder wird durch Experten beurteilt (Expertenvalidität).

Problematisch stellt sich aber der Umgang mit der sogenannten **Kriteriumsvalidität** dar. Sie wird auch als *prognostische* Validität oder *empirische* Validität bezeichnet. Lässt sich beispielsweise anhand des Ergebnisses in einem Rechentest eine verlässliche Aussage über die zukünftige Schulnote in Mathematik treffen, korrelieren Testergebnis und Kriterium[15] hoch miteinander—ein empirischer Zusammenhang wird belegt. Die Kriteriumsvalidität entspricht daher der Korrelation r zwischen einem Testergebnis t und einem Kriterium (criterion) c und wird dargestellt als r_{tc}. Bezeichnenderweise sind aber Prognosen, wie sie sich ausschließlich anhand transsituativ und transtemporal stabiler Eigenschaften als sinnvoll erweisen, in der pädagogischen Diagnostik gerade *nicht* von wesentlicher Bedeutung, da im Sinne der Optimierung von Lernverhalten grundsätzlich *Veränderungen* angestrebt werden. Veränderungen sind in der KTT nicht vorgesehen.

Die Kriteriumsvalidität wird rechnerisch bestimmt anhand des sogenannten **Reliabilitätsindexes**, der Quadratwurzel der Reliabilität $\sqrt{r_{tt}}$. Die Kriteriumsvalidität kann nicht größer werden als die Quadratwurzel der Reliabilität. Es gilt $r_{tc} \leq \sqrt{r_{tt}}$. Da nun bei der kriteriumsorientierten Leistungsmessung eine wahre Varianz $s^2(\tau) = 0$ angestrebt wird, erweist sich die rechnerische Bestimmung der Kriteriumsvalidität von vornherein als wenig sinnvoll. Sie wird im angestrebten Idealfall der Lehrzielerreichung durch sämtliche Schülerinnen und Schüler zwangsläufig ebenfalls $r_{tc} = 0$ betragen.

Invarianzbedingung

Mit Bezug auf den erwähnten Optimierungsgrundsatz ergibt sich eine weitere bedeutsame Unzulänglichkeit der KTT. Dies meint neben der bereits dargestellten Abhängigkeit der Maßzahlen/Testparameter von der Stichprobenzusammensetzung Probleme bei der **Veränderungsmessung**. Die KTT formuliert die sogenannte **Invarianzbedingung** der wahren Werte. „Nur weil der wahre Wert als zeit- und bedingungsinvarianter Parameter aufgefasst wird, können intraindividuelle Schwankungen

[15] Das (Außen-)Kriterium Mathematiknote ist nicht zu verwechseln mit dem Kriterium (Kompetenzstufe) eines Testverfahrens.

[Veränderungen] um einen Wert als meßfehlerbedingt interpretiert werden“ (Wiczerkowski & Schümann, 1978, S. 56). Damit misst ein auf der KTT basierender Test ausschließlich einen absolut stabilen Ist-Zustand (Status), anhand dessen eine diesen Status „fortschreibende (extrapolierende) Prognose“ (Pawlik, 1976, S. 24) gestellt wird. Eine Veränderung der wahren und resultierenden Testwerte stellt gemäß der KTT keinen erwünschten Behandlungserfolg—eine Optimierung—, sondern einen Messfehler dar. Die Veränderung von Merkmalsausprägungen ist nach KTT de facto nicht vorgesehen. Treten dennoch Veränderungen auf, kommt es aufgrund der unpassenden Axiomatik der KTT zum sogenannten *Reliabilitäts-Validitäts-Dilemma* (s. u., Exkurs).

4 Diagnostische Zielsetzung

4.1 Bezugsgruppenorientierte Leistungsmessung

Die Relativität der bezugsgruppenorientierten Leistungsmessung erweist sich jedoch nicht ausschließlich als problematisch. Ihr möglicher Nutzen wird durch die diagnostische Zielsetzung gemäß den bestehenden Rahmenbedingungen bestimmt. Da die bezugsgruppenorientierte Messung die unterschiedlichen Leistungen einer Stichprobe differenziert darzustellen beziehungsweise aufzuspreizen vermag, wird sie dazu benutzt, etwa Schülerinnen und Schüler in eine Rangreihe zu bringen. Hierzu können beispielsweise in Schulen *Screeningverfahren* eingesetzt werden. So können leistungsstarke und leistungsschwache Personen identifiziert und bei Bedarf einer spezifischen Behandlung zugeführt werden. Dies können etwa Förderkurse bei Lese- und Rechtschreibschwierigkeiten (vgl. Response-to-Intervention) sein oder auch spezielle Forderkurse mit erhöhtem Leistungsanspruch an Schülerinnen und Schüler mit weit überdurchschnittlichem Intelligenzniveau.

In der Personalauswahl findet die bezugsgruppenorientierte Leistungsmessung etwa statt, wenn weniger Bewerberinnen und Bewerber das zuvor definierte Kompetenzniveau erreicht haben, als für die Besetzung offener Stellen benötigt werden. Es können nun gewissermaßen die Geeignetsten unter den Ungeeigneten ermittelt werden. Diese durchlaufen dann zusätzliche Schulungsmaßnahmen, bevor sie in den regulären Ausbildungsbetrieb aufgenommen werden.

4.2 Kriterienorientierte Leistungsmessung

Die Zensurengebung an Schulen oder auch Universitäten hingegen sollte sich keinesfalls an der Stichprobenleistung orientieren, sondern ausschließlich an dem zuvor definierten Lehrziel. Demgemäß stellen die in Curricula ausformulierten **Anforderungen** das Kriterium dar. Die Freie Hansestadt Hamburg (2011) formuliert entsprechend in ihrem Bildungsplan Grundschule – Deutsch:

> Der Unterricht in den Fächern und Aufgabengebieten orientiert sich an den **Anforderungen**, die im jeweiligen Rahmenplan[16] beschrieben werden. Der Rahmenplan legt konkret fest, welche Regelanforderungen die Schülerinnen und Schüler zu bestimmten Zeitpunkten zu erfüllen haben und welche Inhalte in allen Grundschulen verbindlich sind, und nennt die Kriterien, nach denen Leistungen bewertet werden. Dabei ist zu beachten, dass die in diesem Rahmenplan tabellarisch aufgeführten Regelanforderungen **Kompetenzen** benennen, die von allen Schülerinnen und Schülern erreicht werden sollen. Durch definierte Regelanforderungen wird die Anschlussfähigkeit des schulischen Lernens gewährleistet und es wird eine Basis geschaffen, auf die sich die Schulen, Lehrerinnen und Lehrer, die Schülerinnen und Schüler, die Sorgeberechtigten sowie die weiterführenden Bildungs- und Ausbildungseinrichtungen verlassen können. (Freie Hansestadt Hamburg, 2011, S. 7)

Die gemäß den Regelanforderungen von den Schülerinnen und Schülern zu erreichenden „Standards" beziehungsweise „Kompetenzen" (MBJS Brandenburg, o. J., S. 15 [s. Fn. 7]) stellen die zuvor inhaltlich definierten Lehrziele (Kriterien) dar. „Durch die Erfassung und Analyse des jeweiligen *aktuellen Leistungsstandes* und der *Leistungsentwicklung* wird den Schülerinnen und Schülern rückgemeldet, welche Lernschritte als nächste erforderlich sind, um ein **Ziel** zu erreichen" (ebd., S. 34; Hervorh. LT). Hier wird das Kriterium als zuvor inhaltlich definierter Punkt auf einem Leistungskontinuum mit Bezug auf eine Modifikations- und Prozessdiagnostik dargestellt.

Die Kultusministerkonferenz hat bereits 1968 die kriteriale Leistungsmessung gemäß Anforderungen beschlossen. Tabelle 1 stellt die *Erläuterungen der Notenstufen bei Schulzeugnissen und Einzelergebnissen in staatlichen Prüfungszeugnissen* dar.

Tabelle 1

Erläuterungen der Notenstufen bei Schulzeugnissen und Einzelergebnissen in staatlichen Prüfungszeugnissen. Beschluss der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland vom 03.10.1968 (Beschluss 675)

Für die Bewertung der Leistungen sind die folgenden Notenstufen mit der angegebenen Wortbedeutung zu verwenden:	
Sehr gut (1)	wenn die Leistung den Anforderungen in besonderem Maß entspricht
Gut (2)	wenn die Leistung den Anforderungen voll entspricht
Befriedigend (3)	wenn die Leistung im Allgemeinen den Anforderungen entspricht
Ausreichend (4)	wenn die Leistung zwar Mängel aufweist, aber im Ganzen den Anforderungen noch entspricht
Mangelhaft (5)	wenn die Leistung den Anforderungen nicht entspricht, jedoch erkennen lässt, dass die notwendigen Grundkenntnisse vorhanden sind und die Mängel in absehbarer Zeit behoben werden können
Ungenügend (6)	wenn die Leistung den Anforderungen nicht entspricht und selbst die Grundkenntnisse so lückenhaft sind, dass die Mängel in absehbarer Zeit nicht behoben werden können

Der Begriff der Anforderung meint hierbei die „im Bildungsplan oder Lehrplan festgelegten Leitgedanken, Kompetenzen, Ziele und Inhalte, insbesondere auf den Umfang sowie auf die selbständige und richtige Anwendung der Kenntnisse, Fähigkeiten und Fertigkeiten sowie auf die Art der Darstellung" (KMK, 1983, § 5[3]).

[16] S. etwa Ministerium für Bildung, Jugend und Sport des Landes Brandenburg et al. (Hrsg). (o. J.). *Rahmenplan Grundschule Deutsch*. O. O.: Autor. Verfügbar unter https://www.bildung-mv.de/export/sites/bildungsserver/downloads/unterricht/Rahmenplaene/Rahmenplaene_allgemeinbildende_Schulen/Deutsch/rp-deutsch-gs.pdf

Die kriterienorientierte Leistungsmessung ist im Gegensatz zur bezugsgruppenorientierten Leistungsmessung jedoch nicht in der Lage, die Besten einer Stichprobe zu identifizieren. Sie differenziert ausschließlich in Löser und Nichtlöser. Eine Differenzierung oberhalb einer erreichten Kompetenzstufe beziehungsweise eines finalen Kriteriums ist nicht vorgesehen. Dennoch können beide Vorgehensweise miteinander verknüpft werden und schließen sich nicht gegenseitig aus.

Zusammenfassend lässt sich formulieren, dass die bezugsgruppenorientierte Leistungsmessung etwa geeignet ist, die Besten einer Stichprobe zu ermitteln, wobei ein inhaltlich definierter Mindeststandard jedoch unberücksichtigt bleibt. Daher kann eine beste bezugsgruppenorientierte Leistung zugleich eine kriterial ungenügende Leistung darstellen. Bei der kriterienorientierten Leistungsmessung hingegen findet keine Rangreihenbildung statt, sodass eine *relativ* beste Leistung nicht definiert ist.

Exkurs: Retest-Reliabilität, Reliabilität der Messwertdifferenzen und Reliabilitäts-Validitäts-Dilemma

Die Reliabilität eines Testverfahrens kann bestimmt werden als Retest-Reliabilität. Dabei wird an einer Stichprobe eine wiederholte Messung mit demselben Testverfahren durchgeführt. Bei gemäß der KTT angenommener Unveränderlichkeit des Messobjekts (**Invarianzbedingung**) sollten Pre- und Postmessung zu übereinstimmenden Ergebnissen kommen. Dies bezieht sich sowohl auf die einzelnen Testwerte x_1, x_2 als auch auf die Testwertvarianz $s^2(x_1), s^2x_2)$ innerhalb der Stichprobe. Gemäß der KTT als Messfehlertheorie wird jedoch jede Messung des wahren Wertes τ mit einem Messfehler ε behaftet sein. Der Messfehler erweist sich als zufallsabhängig, sodass allein die wahren Werte tatsächlich konstant sind. Ein Testverfahren erweist sich entsprechend dann als reliabel, wenn zwischen beiden Messungen t_1, t_2 sowohl die wahren Werte der Testandinnen und Testanden als auch Fehlereinflüsse unverändert bleiben. Es resultieren konstante Testwerte. Entsprechend bleiben auch die wahren, Fehler- und Testwertvarianzen von Pre- zu Postmessung unverändert. Die Retest-Reliabilität beträgt in einem solchen theoretischen, idealen Fall $r_{tt} = 1$. Doch wie verhält sich die Retest-Reliabilität, wenn sich die wahren Werte verändern? Abbildung 1a-d veranschaulicht die folgenden Ausführungen.

Zur Wiederholung: Die Reliabilität berechnet sich nach $r_{tt} = \frac{s^2(\tau)}{s^2(\tau) + s^2(\varepsilon)}$, daher sinkt die Reliabilität, je größer die Fehlervarianz $s^2(\varepsilon)$ bzw. je geringer der Anteil der wahren Varianz an der Testwertvarianz $s^2(x) = s^2(\tau) + s^2(\varepsilon)$. Die Reliabilität r_{tt} einer Messung steigt, je geringer die Fehlervarianz $s^2(\varepsilon)$ bzw. je höher der Anteil der wahren Varianz $s^2(\tau)$ an der Testwertvarianz $s^2(x)$.

Zu a): Bei erfüllter Invarianzbedingung und perfekt reliablem Messinstrument (oder konstantem Messfehler) stimmen die Testwerte in der Stichprobe zu beiden Messzeitpunkten überein (konstante Testwertvarianz bei konstanter Fehlervarianz). Der wahre Wert erweist sich ohnehin als unveränderlich (konstante wahre Varianz gemäß Invarianzbedingung).

In diesem idealen Fall entspricht die Korrelation der Testwerte x_1 und x_2 dem gesuchten Anteil der wahren Varianz $s^2(\tau)$ an der Testwertvarianz $s^2(x)$, hier ist dies $s^2(\tau) + s^2(\varepsilon) = s^2(\tau) + 0$. Daraus folgt $r_{x_1,x_2} = \frac{s^2(\tau)}{s^2(\tau)} = r_{tt} = 1$, die Korrelation zwischen den Testwerten zu t_1 und t_2 sowie die Retest-Reliabilität betragen 1.[17] (s. Formel u.). Aufgrund der hohen Korrelation der Testwerte ist davon auszugehen, dass zu beiden Messzeitpunkten t_1, t_2 dasselbe gemessen wurde – so wie es die Invarianzbedingung der KTT vorsieht. Die Validität der Messungen ist entsprechend hoch.

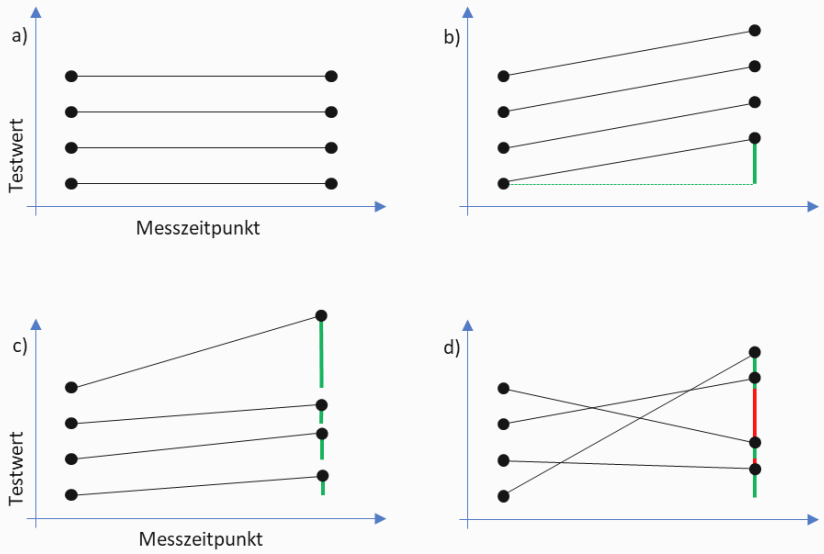

Abbildung 1: Veränderung des Retest-Reliabilitätskoeffizienten durch unterschiedliche Veränderungsmuster von Merkmalsausprägungen zwischen zwei Messungen t_1, t_2.

Zu b): Entgegen der Invarianzbedingung stimmen die Testwerte x_1, x_2 der beiden Messungen t_1, t_2 nicht überein. Allerdings hat hier eine systematische, völlig gleichförmige Veränderung des Merkmals/der wahren Werte über die gesamte Stichprobe hinweg stattgefunden. Entsprechend sind die Varianzen $s^2(\tau), s^2(x), s^2(\varepsilon)$ in beiden Messungen unverändert, und die Testwerte der Pre- und Postmessung korrelieren hoch miteinander. Die Retest-Reliabilität bleibt im Vergleich mit a) unverändert.

Zu c): Entgegen der Invarianzbedingung stimmen die Testwerte x_1, x_2 der beiden Messungen t_1, t_2 nicht überein. Es hat eine *un*systematische Veränderung stattgefunden. Die Korrelation zwischen den Testwerten x_1, x_2 ist im Gegensatz zu b) zwar niedriger, jedoch im Vergleich zu d) noch immer relativ hoch. Als Ursache für die Messwertdifferenzen kommen etwa Lern- und Wiederholungseffekte in Betracht. Die Retest-Reliabilität sinkt.

Zu d): Entgegen der Invarianzbedingung stimmen die Testwerte x_1, x_2 der beiden Messungen t_1, t_2 nicht überein. Das zu Messende Merkmal erweist sich als instabil. Die Korrelation zwischen Pre- und Posttest erweist sich als gering. Offenbar ist zu beiden Messzeitpunkten etwas anderes gemessen worden. Es ist keine Validität der Messungen gegeben. Die Retestreliabilität sinkt.

[17] $r(x_1, x_2) = \frac{Cov(x_1, x_2)}{s(x_1) \cdot s(x_2)} = \frac{Cov(\tau_1 + \varepsilon_1, \tau_2 + \varepsilon_2)}{s(x_1) \cdot s(x_2)} = \frac{Cov(\tau_1, \tau_2)}{s(x_1) \cdot s(x_2)} = \frac{s^2(\tau)}{s^2(x)} = r_x =$ Retestreliabilität (auch r_{tt})

Zusammenfassend lässt sich feststellen, dass die Korrelation der Testwerte x_1, x_2 die Höhe der Retest-Reliabilität bestimmt (vgl. Fn 17). Die Korrelation $r_{(x_1,x_2)}$ nähert sich 1 an, wenn bei erfüllter Invarianzbedingung die Fehlervarianz von der Pre- zur Postmessung konstant bleibt (s. Beispiel a) und b)). Dabei gibt eine hohe Korrelation zwischen zwei Testergebnissen auch an, dass offenbar tatsächlich dasselbe gemessen wurde, mithin eine valide Messung vorliegt. Bei unterschiedlichen oder sich verändernden Messobjekten hingegen variierten auch die individuellen Testwerte in der Pre- und Postmessung unterschiedlich. **Diese höhere Varianz der Messwertdifferenzen $s^2(x_1 - x_2)$ besteht also immer dann, wenn die Korrelation zwischen Pre- und Postmessung gering ist**.

Wird nun das retest-reliable Testverfahren in der Praxis eingesetzt, so werden bei wiederholten Messungen vergleichbare Ergebnisse resultieren. Die Testwerte x_1, x_2 werden über gleiche anteilige wahre und Fehlervarianz verfügen, mithin werden die Testergebnisse hoch miteinander korrelieren.

Weichen die Ergebnisse zweier Messungen t_1, t_2 jedoch wider Erwarten voneinander ab—besteht also eine geringe Korrelation zwischen den Testwerten x_1, x_2—, so bestehen im Rahmen der KTT zwei Interpretationsmöglichkeiten.

Zum einen können die Messwertdifferenzen durch unerwartete Messfehler aufgetreten sein. Da sich die wahren Werte gemäß Invarianzbedingung als transtemporal und transsituativ stabil erweisen, können Unterschiede in den Testwerten nur auf eine fehlerhafte Messung zurückgeführt werden. Dies betrifft die **Reliabilität** $r_{tt} = \frac{s^2(\tau)}{s^2(\tau)+s^2(\varepsilon)}$. Unabhängig davon was tatsächlich gemessen wird, ist das Messergebnis nicht verwertbar.

Zum anderen können Messwertdifferenzen dadurch erklärt werden, dass tatsächlich etwas Verschiedenes gemessen worden ist. Ein reliables Testverfahren kommt nur dann zu vergleichbaren Ergebnissen, wenn zu beiden Messzeitpunkten das gleiche Messobjekt gemessen wird. Ist dies nicht der Fall, ist die **Validität** der Messung nicht gegeben. Das Verfahren misst nicht das, was zu messen es konstruiert wurde/was es zu messen vorgibt. In einem solchen Fall erweist sich eine hohe Reliabilität als hinfällig.

Dieses Dilemma tritt als **Reliabilitäs-Validitäts-Dilemma der Veränderungsmessung**/Prozessdiagnostik auf. Die folgenden Ausführungen stellen dies genauer dar. Das Reliabilitäts-Validitäts-Dilemma der Veränderungsmessung (RVD) tritt bei tatsächlicher Veränderung bei Verwendung eines reliablen Testverfahrens auf. Es ist daher nicht als bloß theoretische Erklärung von Messungenauigkeit zu verstehen. Zu beachten ist, dass sich die Angaben zur Reliabilität hier auf die **Messwertdifferenzen** $x_1 - x_2$ zwischen zwei Messzeitpunkten beziehen, nicht jedoch auf die Reliabilität der einzelnen Messungen t_1, t_2. Auf diese beziehen sich hingegen Aussagen zur Validität. Die Messwertdifferenzen stellen bei einem reliablen Testverfahren schließlich die tatsächliche Veränderung des Messobjekts etwa als Therapierfolg dar.

Gemäß den Ausführungen zur Retest-Reliabilität kann bei einer hohen Korrelation zwischen zwei Messungen von einer hohen Validität ausgegangen werden. Dies trägt zur Retest-Reliabilität bei. Mit Bezug auf das RVD bedeutet eine hohe Korrelation der Testwerte x_1, x_2 jedoch, dass die Reliabilität der Messwertdifferenzen *sinkt*.

Die Reliabilität r_{dd} der Messwertdifferenzen bestimmt sich aus der Varianz wahren Differenzen und der Varianz der Fehlerdifferenzen gemäß $\frac{s^2(\tau_d)}{s^2(\tau_d) + s^2(\varepsilon_d)}$. Je valider die Messung, je höher also die Korrelation r_{x_1,x_2}, desto weniger differieren die Veränderungen (Messwertdifferenzen) der Testpersonen innerhalb der Stichprobe: Die Testpersonen verändern sich ähnlich. Daher sinkt der Anteil der Varianz wahrer Differenzen und steigt der Anteil der Varianz der Fehlerdifferenzen an der Varianz der Messwertdifferenzen. Die Reliabilität der Messwertdifferenzen sinkt. Anders ausgedrückt: Bei hoher Korrelation zwischen zwei Messungen haben diese einen großen Teil wahrer Varianz gemeinsam. Testwertunterschiede (Messwertdifferenzen) sind dann großenteils auf Messfehler zurückzuführen. Wenn die Messwertdifferenzen allerdings nahezu ausschließlich Messfehler darstellen, bleibt zu fragen, welche Art von Veränderungen überhaupt gemessen worden sind.

Andersherum bedeutet das RVD, dass bei geringer Korrelation r_{x_1,x_2} offensichtlich etwas anderes gemessen wurde, also tatsächliche eine Veränderung stattgefunden hat. In diesem Fall erweisen sich die Messungen als nicht valide. Die Reliabilität der Messwertdifferenzen hingegen steigt, da die Testpersonen sich individuell unterschiedlich verändern, und der Anteil der Varianz wahrer Differenzen an den Messwertdifferenzen zunimmt. Wenn jedoch nicht dasselbe gemessen wurde, erweist sich eine hohe Reliabilität der Messwertdifferenzen als nutzlos.

„Diese Zusammenhänge legen nah, dass das Reliabilitätskonzept der klassischen Testtheorie bei der Erfassung der Genauigkeit von Differenzwerten offenbar versagt“ (Bortz & Döring, 2002, S. 553).

Zusatzlich zu dem RVD vereinigen bei der Messung einer Veränderung von Messzeitpunkt t_1 zu Messzeitpunkt t_2 die resultierenden Differenzwerte die Messfehler beider Messungen auf sich („It comes from two fallible scores and the error variances from them summate“ [Guilford, 1954, S. 393]). Die Reliabilität der Messwertdifferenzen muss entsprechend zwangsläufig geringer ausfallen als die Reliabilität der sie konstituierenden Messwerte. „Ist die Zuverlässigkeit eines Tests zum Beispiel 0.80 und korrelieren beide Messungen zu 0.70 bei gleicher Varianz, so ist die Reliabilität der Differenzwerte 0.33“ (Wiczerkowski & Schümann, 1978, S. 57). „Ein Messintrument, das eine Reliabilität von beispielsweise r=0,90 aufweist ... [führt] zu Messwertdifferenzen mit einer Reliabiliät von 0,67 ..., wenn Pretest- und Posttestmessungen zu r=0,70 miteinander korellieren“ (Bortz & Döring, 2002, S. 552). Guilford (1954, S. 394) gibt die entsprechende Formel zur Berechnung für die Reliabilität von Messwertdifferenzen nach Mosier an:

$$r_{dd} = \frac{r_{jj} + r_{kk} - 2r_{jk}}{2(1 - r_{jk})}.$$

Setzt man die Werte von Wiczerkowski und Schümann (1978, S. 57) ein, so erhält man als Reliabilität der Messwertdifferenzen $r_{dd} = \frac{0.8 + 0.8 - 2 \cdot 0.7}{2 \cdot (1 - 0.7)} = \frac{1.6 - 1.4}{2 \cdot 0.3} = \frac{0.2}{0.6} = 0.33.$

Sinkt nun aufgrund einer höheren Varianz der Messwertdifferenzen die Korrelation r_{jk} zwischen beiden Messungen t_1 und t_2 auf beispielsweise lediglich r_{jk} = .3, so ergibt sich bei einem Test mit einer Reliabilität von $r_{tt} = .8$ eine rein *rechnerisch* deutlich höhere Reliabilität der Messwertdifferenzen von $r_{dd} = .71$. Es

erweist sich jedoch als unsinnig, dieses Ergebnis als tatsächliche Güte der Messung zu interpretieren, wenn gemäß Invarianzbedingung ganz offenbar ohnehin nicht dasselbe gemessen wurde (geringe Korrelation der Messergebnisse).

Leichner (1976, S. 50–57) erklärt das RVD anhand derselben Formel. Setzt man die Reliabilitäten r_{jj}, r_{kk} und auch die Korrelation r_{jk} gleich 1, fällt die Reliabilität der Differenzen auf 0.

$$r_{dd} = \frac{r_{jj} + r_{kk} - 2r_{jk}}{2(1 - r_{jk})} = \frac{1 + 1 - 2 \cdot 1}{2(1 - 1)} = \frac{2 - 2}{2 - 2} = 0$$

Setzt man die Reliabilitäten $r_{jj}, r_{kk} = 1$, die Korrelation r_{jk} hingegen $= 0$, steigt die Reliabilität der Differenzen auf 1.

$$r_{dd} = \frac{r_{jj} + r_{kk} - 2r_{jk}}{2(1 - r_{jk})} = \frac{1 + 1 - 2 \cdot 0}{2(1 - 0)} = \frac{2 - 0}{2 - 0} = 1$$

(nach Fisseni, 1997, S. 367)

Es lässt sich nun das Reliabilitäts-Validitäts-Dilemma der Veränderungsmessung knapp formulieren: **Bei hoher Korrelation und Validität der Messungen sinkt die Reliabilität der Messwertdifferenzen, bei geringer Korrelation und Validität der Messungen steigt die Reliabilität der Messwertdifferenzen.**

5 Praxis – Numerus Clausus

Numerus Clausus bezeichnet die Beschränkung der Studierendenanzahl für ein Hochschulstudium. Die Anzahl (lat. $numerus$) der zu einem Studium zugelassenen Personen ist also nicht nach oben hin offen, sondern beschränkt—der Zugang zum Studium bleibt für viele Interessierte verschlossen (lat. $clausus$). Diese Beschränkung wird neben Auswahlgesprächen und Eignungstests auch anhand des Abiturnotendurchschnitts vorgenommen.

Da die KMK (1968) eine kriteriale schulische Leistungsbewertung gemäß definierter Anforderungen vorsieht (s. 4.2), sind zur länderübergreifenden Vergleichbarkeit von Schulnoten bundeseinheitliche Anforderungen beziehungsweise **Bildungsstandards** anzustreben.

> Die Kultusministerkonferenz sieht es als zentrale Aufgabe an, ... die **Vergleichbarkeit** schulischer Abschlüsse ... zu sichern. **Bildungsstandards** sind hierbei von besonderer Bedeutung. ... Bildungsstandards beschreiben erwartete Lernergebnisse. Ihre Anwendung bietet Hinweise für notwendige Förderungs- und Unterstützungsmaßnahmen. Bildungsstandards greifen allgemeine **Bildungsziele** auf und benennen **Kompetenzen**, die Schülerinnen und Schüler bis zu einer bestimmten Jahrgangsstufe an zentralen Inhalten erworben haben sollen. ... Die Standards basieren auf fachspezifisch definierten **Kompetenzmodellen**, die aus der Erfahrung der Schulpraxis heraus entwickelt wurden. Sie beziehen international anerkannte Standardmodelle – u. a. theoretische Grundlagen der PISA-Studie und den Gemeinsamen europäischen Referenzrahmen für Sprachen – ein. ... Die Länder verpflichten sich, die Standards zu implementieren und anzuwenden. (Beschluss der KMK v. 04.12.2013)[18]

[18] Verfügbar unter http://db2.nibis.de/1db/cuvo/datei/bs_ms_kmk_mathematik.pdf (S. 3f)

Für die Gewährleistung der Vergleichbarkeit von Abiturnoten wird „auf Beschluss der Kultusministerkonferenz seit dem Jahr 2013 ein gemeinsamer Pools [sic] von standardbasierten Abiturprüfungsaufgaben aufgebaut, der kontinuierlich aufwächst [sic] und den Ländern seit dem Schuljahr 2016/2017 als Angebot für den *möglichen* Einsatz im Abitur zur Verfügung steht“ (KMK, 2018; Hervorh. LT).

Eine tatsächlich bundeseinheitliche Leistungsbewertung muss allerdings bisher als nicht erreicht gelten. Trotz den von der KMK vorgeschlagenen einheitlichen Prüfungsanforderungen (EPA) besteht in den Ländern Uneinheitlichkeit bei der abschließenden *Berechnung* der Abiturdurchschnittsnote (s. etwa Kramer & Márquez, 2015; Scholl, 2015). Das bedeutet, dass selbst wenn die Leistungsbewertung in den *einzelnen* Fächern einheitlich vollzogen würde, es bei gleichen Noten zu ungleichen Abiturnotendurchschnitten käme. Von einer gerechten Selektionsentscheidung kann bei der Studienplatzvergabe entsprechend keineswegs die Rede sein.

Grundsätzlich basiert die Verwendung des Abiturnotendurchschnitts auf der Annahme, dass dieser den besten Prädiktor für den Studienerfolg darstelle (s. etwa Formazin et al., 2008, S. 206). Das bedeutet, dass prinzipiell zunächst diejenigen Studienbewerberinnen und -bewerber einen Studienplatz zugesprochen bekommen, von denen angenommen wird, dass sie ihr Studium mit gutem Ergebnis beenden werden (Prognose). Dieses gute Ergebnis besteht im Sinne § 7 Hochschulrahmengesetz (HRG) darin, dass Studierende „zu wissenschaftlicher oder künstlerischer Arbeit und zu verantwortlichem Handeln in einem freiheitlichen, demokratischen und sozialen Rechtsstaat **befähigt**“ (Bundesministerium für Justiz und Verbraucherschutz, 1999; Hervorhebung LT) werden. Bedeutet dies jedoch auch, dass Personen, bei denen eine solche Zielerreichung weniger wahrscheinlich sei—also bei Personen, die über einen geringeren Abiturnotendurchschnitt verfügen—das Grundrecht auf Bildung vorübergehend oder dauerhaft verwehrt bleibt? Nein. Der Abiturnotendurchschnitt muss nicht nur länderübergreifend vergleichbar sein, sondern er darf auch nicht das einzige Zulassungskriterium sein. Es folgt, dass auch Auswahlgespräche und Eignungstests (in standardisierter Form) durchgeführt werden müssen.

Das Bundesverfassungsgericht (BVerfG) urteilte am 19.12.2017 im Zusammenhang mit der Studienplatzvergabe für das Fach Humanmedizin:

> 1. Nach Art. 12 Abs. 1 Satz 1 in Verbindung mit Art. 3 Abs. 1 GG haben jede Studienplatzbewerberin und jeder Studienplatzbewerber ein **Recht auf gleiche Teilhabe** an staatlichen Studienangeboten und damit auf **gleichheitsgerechte Zulassung zum Studium** ihrer Wahl.
>
> 2. Regeln für die Verteilung knapper Studienplätze haben sich grundsätzlich am **Kriterium der Eignung** zu orientieren. ... Die zur Vergabe knapper Studienplätze herangezogenen Kriterien müssen die Vielfalt der möglichen Anknüpfungspunkte zur Erfassung der Eignung abbilden. ...
>
> Verfassungswidrig sind die gesetzlichen Vorschriften zum Auswahlverfahren der Hochschulen insofern, ... als im Auswahlverfahren der Hochschulen die **Abiturnoten** berücksichtigt werden können, ohne einen Ausgleichsmechanismus für deren nur **eingeschränkte länderübergreifende Vergleichbarkeit** vorzusehen. (BVerfG, 2017)

Es wird deutlich, dass die Leistungsbewertung auch im Studium ausschließlich eine kriteriale sein kann, um die in § 7 HRG angestrebte **Befähigung** der Studierenden zu gewährleisten. Die Befähigung ist an den Anforderungen der angestrebten Tätigkeit zu relativieren, wie sie von den diese Tätigkeit Ausübenden definiert wird. Das Erreichen dieses Lehrziels kann nur von Expertinnen und Experten auf dem jeweiligen Fachgebiet bestätigt werden. Hierzu erweist es sich als notwendig, dass bundesweit einheitliche Anforderungen (durchschnittliche Anforderungen) für die einzelnen Fächer etabliert werden, ohne jedoch die Freiheit in Wissenschaft und Lehre zu berühren (s. Art. 5 Abs. 3 GG). Dies erweist sich als schwierig, da die inhaltliche Schwerpunktsetzung sich auch auf das Anforderungsniveau auswirkt—Lehrinhalte unterscheiden sich sowohl hinsichtlich ihres Umfangs als auch ihrer Komplexität. Da die universitäre Lehre jedoch eng verbunden ist mit den Forschungsschwerpunkten der Lehrenden, käme dies unter Umständen einer Entkoppelung von Forschung und Lehre gleich.

Die in diesem Zusammenhang zu formulierenden Anforderungen zur Erreichung eines Kriteriums können entsprechend nicht außerhalb der Forschung vorgeschrieben sein. Dennoch sind einheitliche Prüfungsanforderungen und Leistungsbewertungen *anzustreben*.

Zu der inhaltlichen Definition kriterialer Leistungsbewertung tritt hier die **durchschnittliche Anforderung**, die juristisch jedoch nicht definiert ist (vgl. zum juristischen Prüfungsrecht Zimmerling & Brehm, 2012, S. 266). Dabei wird der Bewertungsspielraum der Prüfenden selbst nicht eingeschränkt. Die Kenntnis durchschnittlicher Anforderungen kann gegebenenfalls jedoch „bei zukünftigen Bewertungen hilfreich ... und ein Weg zum Erreichen einer vergleichbaren Leistungsbewertung sein“ (ebd.). An anderer Stelle führen die Autoren aus: „Es ist allgemein anerkannt und entspricht der Lebenswirklichkeit, dass in die Bewertung auch relative Elemente einfließen [die durchschnittliche Leistung der Stichprobe; Anm. LT]. ... Mittelbar werden mithin die theoretisch nur absolut zu messenden Anforderungen im Rahmen einer Prüfung zumindest teilweise wieder relativiert“ (Zimmerling & Brehm, 2017, S. 52f).

> Allerdings hat im (theoretischen) Ausgangspunkt die Bewertung einer Prüfungsleistung nach einem objektiven/absoluten Maßstab zu erfolgen ohne Rücksicht darauf, wie andere Prüflinge des Durchgangs die Prüfungsfrage gelöst haben. Die Prüfer, die in der Regel aufgrund ihrer Prüfungserfahrungen eine Vielzahl von Prüfungsdurchgängen vor Augen haben, ordnen die jeweilige Aufgabenstellung und die Prüfungsleistungen in diesen übergeordneten Rahmen ein, ohne maßgeblich darauf abzustellen, ob es sich konkret um einen schwachen oder leistungs-starken Prüfungsdurchlauf handelt; denn zu beurteilen sind die individuellen Fähigkeiten bezogen auf den angestrebten Beruf, nicht dagegen sind die Leistungen der jeweiligen Prüfungsgruppe lediglich untereinander ins Verhältnis zu setzen. Der innerhalb einer insgesamt schwachen Prüfungsgruppe noch am besten abschneidende Mitprüfling kann also nicht allein deshalb schon eine hervorragende Bewertung erhalten, ebenso wenig wie die Leistung des schwächsten Mitglieds einer insgesamt starken Prüfungsgruppe allein deswegen als unzureichend bewertet werden darf. (Zimmerling & Brehm, 2017, S. 52; vgl. o. Abschn. 2)

Dipl.-Psych. Dr. Lars Tischler ist Wissenschaftlicher Mitarbeiter an der Medical School Hamburg, Dozent für Allgemeine Psychologie, Diagnostik in der Pädagogischen Psychologie sowie Intervention in der Pädagogischen Psychologie, wissenschaftlicher Beirat im Fachverband für Ganzheitliche Entwicklung und Ganzheitliche Therapie 4Kids2GET e. V., Dozent am VIGESCO-Institut für psychologisch-pädagogische Bildung und Heilpraktiker Psychotherapie.

Korrespondenzadresse
Dr. Lars Tischler, Medical School Hamburg, Am Kaiserkai 1, 20457 Hamburg

tischler@vigesco-institut.de
lars.tischler@medicalschool-hamburg.de
tischlerweb.wordpress.com

Literatur

Bortz, J. & Döring, N. (2002). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (3. Aufl.). Berlin: Soringer.

Bundesministerium für Justiz und Verbraucherschutz. (1999). *Hochschulrahmengesetz in der Fassung der Bekanntmachung vom 19. Januar 1999 (BGBl. I S. 18), das zuletzt durch Artikel 6 Absatz 2 des Gesetzes vom 23. Mai 2017 (BGBl. I S. 1228) geändert worden ist*. Zugriff am 17. Januar 2018. Verfügbar unter https://www.gesetze-im-internet.de/hrg/HRG.pdf

Bundesverfassungsgericht. (2017). *Urteil des Ersten Senats vom 19. Dezember 2017 - 1 BvL 3/14 - Rn. (1-253)*. Zugriff am 17. Januar 2018. Verfügbar unter http://www.bverfg.de/e/ls20171219_1bvl000314.html

Fisseni, H.-J. (1997). *Lehrbuch der psychologischen Diagnostik* (2., überarbeitete u. erweiterte Aufl.). Göttingen: Hogrefe.

Formazin, M., Wilhelm, O., Schroeders, U., Kunina, O., Hildebrandt, A. & Köller, O. (2008). Validitäts- und Nutzenüberlegungen zur Studierendenauswahl mit Präzisierungen für das Fach Psychologie. In H. Schuler & B. Hell (Hrsg.)., *Studierendenauswahl und Studienentscheidung* (S. 204–214). Göttingen: Hogrefe.

Freie Hansestadt Hamburg. Behörde für Schule und Bildung. (Hrsg.). *Bildungsplan Grundschule. Deutsch.* Hamburg: Autor. Zugriff am 17. Januar 2018. Verfügbar unter http://www.hamburg.de/contentblob/2481792/d180d66decd915caf50391ab07bdc51d/data/deutsch-gs.pdf;jsessinid=2F2422545A06216ECE0D05CED618B522.liveWorker2

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Hilke, R. (1980). *Grundlagen normorientierter und kriteriumorientierter Tests. Eine kritische Auseinandersetzung mit der klassischen Testtheorie und den logistischen Testmodellen*. Bern: Huber.

Kramer, B. & Márquez, A. (2015). *Notenrechner. Hätten Sie das Abi auch in Bayern bestanden?* SPIEGEL ONLINE. Zugriff am 17. Januar 2018. Verfügbar unter http://www.spiegel.de/lebenundlernen/schule /studium -und-nc-abiturnoten-sind-ungleich-in-deutschland-a-1044518.html

Leichner, R. (1979). *Psychologische Diagnostik. Grundlagen, Kontroversen, Praxisprobleme*. Weinheim: Beltz.

Ständige Kultusministerkonferenz der Länder. (2018). *Implementation der Bildungsstandards für die Allgemeine Hochschulreife*. Zugriff am 17. Januar 2018. Verfügbar unter

https://www.kmk.org/themen/qualitaetssicherung-in-schulen/bildungstandards/bildungsstandards-und-allgemeine -hochschulreife.html

Ständige Kultusministerkonferenz der Länder. (1983). *Verordnung des Kultusministeriums über die Notenbildung (Notenbildungsverordnung, NVO) vom 5. Mai 1983*.

Moosbrugger, H. & Kelava, A. (Hrsg.). (2008). *Testtheorie und Fragebogenkonstruktion*. Berlin: Springer.

Pawlik, K. (1976). Modell- und Praxisdimensionen psychologischer Diagnostik. In K. Pawlik (Hrsg.), *Zur Diagnose der Diagnostik. Beiträge zur Diskussion der psychologischen Diagnostik in der Verhaltensmodifikation* (S. 13–43). Stuttgart: Klett-Cotta.

Scholl, H. *(2015).* Wie Abiturprüflinge ungleich behandelt werden. *Frankfurter Allgemeine.* Zugriff am 17. Januar 2018. Verfügbar unter http://www.faz.net/aktuell/politik/inland/abitur-noten-ungerechtigkeit-in-der-schule-13655096.html

Wiczerkowski, W. & Schümann, M. (1978). Klassische Testtheorie. In K. J. Klauer (Hrsg.), *Handbuch der Pädagogischen Diagnostik* (Bd. 1, S. 41–58). Düsseldorf: Schwann.

Zimmerling, W. & Brehm, R. (2017). *Aktualisiertes Manuskript des Prüfungsrechtsseminars vom 22.05.2017 in Frankfurt/Main* (Stand: 30.09.2017). O. O.: Autor. Zugriff am 19. Januar 2018. Verfügbar unter www.zimmerling.de/downloads-149.html?file=files/zimmerling/.../Prüfungsrecht...pdf

Zimmerling, W. & Brehm, R. (2012). Kritisches zum juristischen Prüfungsrecht. *Deutsche Verwaltungsblätter, 1. März 2012*, 265–272.

Ebel, R. L. (1962). Content standard test scores. Educational and Psychological Measurement, 22, 15–25.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.

Hammock, J. (1960). Criterion measures: Instruction vs. selection research (paper read at the meetings of the American Psychological Association, September, 1960).

Hilke, R. (1980). *Grundlagen normorientierter und kriteriumorientierter Tests. Eine kritische Auseinandersetzung mit der klassischen Testtheorie und den logistischen Testmodellen*. Bern: Huber.

Klauer, K. J. (1987). *Kriteriumsorientierte Tests*. Göttingen: Hogrefe.
Pawlik, K. (1982). *Diagnose der Diagnostik*. Stuttgart: Klett-Cotta.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271–295.

Tent, L. & Stelz., I. (1993).
Pädagogisch-psychologische Diagnostik. Band 1 – Theoretische und methodische Grundlagen. Göttingen: Hogrefe.