Renner, G. (2022). Normtabellen analysieren und beurteilen II: Itemgradienten und Altersdifferenzierung. In M. Gebhardt, D. Scheer & M. Schurig (Hrsg.), *Handbuch der sonderpädagogischen Diagnostik. Grundlagen und Konzepte der Statusdiagnostik, Prozessdiagnostik und Förderplanung* (S. 291-304). Regensburg: Universitätsbibliothek. https://doi.org/10.5283/epub.53149

Normtabellen analysieren und beurteilen II: Itemgradienten und Altersdifferenzierung

Gerolf Renner

1 Vorbemerkung

Es wird in der Regel sinnvoll sein, diesen Text erst nach dem Kapitel zu »Bodeneffekten« (in diesem Band) zu lesen. Die Kenntnis von Begriffen wie Mittelwert, Standardabweichung, Wertpunkt, IQ-Wert, Prozentrang, Standardwert und Rohwert wird vorausgesetzt. Leser*innen sollten wissen, wie Normwerte in standardisierten Testverfahren berechnet werden, wie Normtabellen aufgebaut sind und wie man darin Standardwerte ablesen kann (vgl. Kap. »Normwerte«, in diesem Band).

2 Zum Einstieg ein kleines Diagnostik-Quiz

Es sei gleich verraten, dass es bei den folgenden Fragen nicht die eine richtige Antwort gibt. Aber welche Faktoren muss man im Blick haben, wenn man sich zu den kleinen Fallbeispielen eine Meinung bilden will?

- Frage 1: Ein Test, der aus acht Items besteht, wirbt mit der Aussage, dass er die »ökonomische« Erfassung visuell-räumlicher Leistungen ermöglicht. In der Tat haben Sie den Test sehr schnell durchgeführt und ausgewertet. Charlotte erhält einen Standardwert (M = 100, SD = 15) von 71 – also fast zwei Standardabweichungen unter dem Mittelwert. Dürfen Sie in ihrem Bericht mit gutem Gewissen schreiben, dass die Leistung eindeutig unterdurchschnittlich ausgefallen ist?
- Frage 2: Sie führen einen Test zur Erfassung der fluiden Intelligenz durch. Die untersuchten Kinder wählen die richtige Lösung jeweils aus vier vorgegebenen Alternativen. Sebastian erreicht einen Standardwert von 96 – eindeutig eine durchschnittliche Leistung. Können Sie sich auf dieses Ergebnis verlassen?
- Frage 3: Ein Zwillingspaar wird in der Frühförderung getestet. Peter erreicht drei Tage vor seinem fünften Geburtstag in einem Wortschatztest einen Standardwert von 96. Das ist nach den üblichen Kriterien ein durchschnittliches Ergebnis. Petra wird acht Tage später untersucht. Sie erzielt im gleichen Test einen Standardwert von 78, also ein unterdurchschnittliches Ergebnis. Die Reliabilität des Tests beträgt .90, die Differenz zwischen den

beiden Testwerten ist statistisch signifikant auf dem 1%-Niveau. Doch die Eltern wollen gar nicht glauben, dass sich der Wortschatz der beiden wesentlich unterscheidet. Wer hat recht? Der Test oder die Eltern?

3 Tückische Normtabellen

Standardisierte Tests sollen eine genaue Messung von pädagogischen und psychologischen Konstrukten ermöglichen. Nur weil man einen Standardwert in einer Normtabelle ablesen kann, ist noch nicht gesichert, dass dieser Wert tatsächlich die Leistung des untersuchten Kindes widerspiegelt. Ungünstige Itemgradienten und eine schlechte Altersdifferenzierung können – ebenso wie Bodeneffekte (vgl. Kap. »Bodeneffekte«, in diesem Band) – dazu führen, dass Fähigkeiten über- oder unterschätzt, Ergebnisse von Verlaufsmessungen verzerrt und Interpretationen von Testprofilen erschwert werden. Im Folgenden wird aufgezeigt, wie Diagnostiker*innen diese kritischen Qualitätsmerkmale von Testverfahren in Normtabellen erkennen und beurteilen können.

4 Itemgradienten

Auf eine Waage wird ein Gewicht von 4 Kilogramm gelegt. Der Zeiger verharrt auf der Null. Es folgt ein Gewicht von 7 kg. Der Zeiger springt auf die 10. Da bleibt er stehen, als 13 kg aufgelegt werden. Bei 15.5 kg springt er auf die 20.

Was soll man von so einer Waage halten? Diese Waage kann offensichtlich nur sehr grob messen: 0, 10, 20 – dazwischen liegende Werte zeigt sie nicht an. Was hat das mit Testdiagnostik zu tun?

Quantitative Testverfahren wollen die Ausprägung bestimmter Konstrukte (z. B. Intelligenz, Sprachverständnis, Lesekompetenz) auf einem Kontinuum erfassen. Theoretisch kann es unendlich viele Abstufungen dieses Kontinuums geben. Praktisch ist eine extrem differenzierte Messung psychologischer und pädagogischer Konstrukte jedoch nicht möglich. Jeder Test beinhaltet nur eine begrenzte Anzahl von Items. Dementsprechend ist auch der mögliche Bereich der Rohwerte eingeschränkt. Wenn ein Test 8 Items beinhaltet, die jeweils mit 0 (falsch) oder 1 (richtig) bewertet werden, können 0 bis 8 Rohwertpunkte erzielt werden, die für eine bestimmte Altersgruppe in 9 unterschiedliche Standardwerte umgerechnet werden können. Das gesamte Kontinuum des gemessenen Leistungsbereiches – von extrem unterdurchschnittlich bis extrem überdurchschnittlich – wird auf nur neun Abstufungen reduziert. Umfasst ein Test 30 Items, bei denen jeweils 0 bis 4 Rohwertpunkte vergeben werden, können die Rohwerte zwischen 0 und 120 variieren, es könnte also prinzipiell 121 unterschiedliche Standardwerte geben, was somit eine deutlich differenziertere Messung ermöglicht. Je weniger Abstufungen möglich sind, desto größer werden die Abstände zwischen den tatsächlich ermittelbaren Normwerten ausfallen.

Tabelle 1 entspricht einer typischen Normtabelle, mit der Rohwerte in Wertpunkte umgerechnet werden. Wertpunkte sind Standardwerte mit einem Mittelwert von 10 und einer Standardabweichung von 3. Üblicherweise werden die Wertpunkte in Testmanualen mit einem Wertebereich von 1 bis 19 tabelliert. Bei *Test 1* kommen bestimmte Standardwerte gar nicht vor. Es gibt z. B. keinen Rohwert, der 2 oder 4 Wertpunkten entspricht. Ein Kind, das einen Rohwert

Tabelle 1: Normwerttabellen mit unterschiedlichen Itemgradienten sowie Veränderung der Standardwerte bei einem Zuwachs des Rohwertes um 1

	Test 1			*Test 2*	
Wertpunkte	Rohwert	Veränderung des Standardwerts	Wertpunkte	Rohwert	Veränderung des Standardwerts
1	0		1	1-2	
2			2	3	+1
3	1	+2	3	4-5	+1
4			4	6	+1
5	2	+2	5	7	+1
6			6	8	+1
7			7	9-10	+1
8			8	11	+1
9	3	+4	9	12-13	+1
10			10	14-15	+1
11	4	+2	11	16-17	+1
12	5	+1	12	18	+1
13			13	19	+1
14			14	20	+1
15	6	+3	15	21	+1
16			16	22	+1
17	7	+2	17	23	+1
18			18	24	+1
19	8	+2	19	25	+1

von 2 erzielt, erhält 5 Wertpunkte. Das ist ein unterdurchschnittliches Ergebnis. Löst es eine einzige Aufgabe mehr, liegt der Standardwert bei 9 – dazwischen gibt es nichts. Eine Veränderung des Rohwerts um einen einzigen Punkt führt bei *Test 1* zu Veränderungen der Standardwerte von 2 bis 4 Wertpunkten. Bei *Test 2* verändern sich die Standardwerte durchgehend nur um einen Wertpunkt.

Der Begriff Itemgradient bezeichnet nun das Ausmaß, in dem sich Standardwerte verändern, wenn der Rohwert um die kleinstmögliche Einheit höher oder niedriger ausfällt. Bei fast allen Testverfahren sind Rohwerte ganze Zahlen, die kleinste Veränderung ist dann 1 Rohwertpunkt (in seltenen Fällen werden auch halbe Punkte oder andere Abstufungen verwendet). Itemgradienten gelten nach Bracken (1987) als kritisch oder ungünstig, wenn die Veränderung der Standardwerte mehr als 1/3 Standardabweichung beträgt. Dieser Wert entspricht bei gebräuchlichen Skalierungen von Normwerten z. B. 1 Wertpunkt, 5 IQ-Punkten oder 3.3 T-Werten. Bei der Wertpunktskala, die in Tabelle 1 verwendet wird, sollte also jedem Wertpunkt ein Rohwert zugeordnet sein. Dies ist bei *Test 2* der Fall. In der grafischen Darstellung (Abb. 1) wird der steile Anstieg der Normwerte bei *Test 1* deutlich, der typisch für Tests mit ungünstigen Itemgradienten ist.

Auch wenn bei *Test 1* verschiedene Standardwerte gar nicht auftreten können, gibt es natürlich Testpersonen, deren Leistung diesen nicht tabellierten Werten entspricht. Kinder, deren wahre

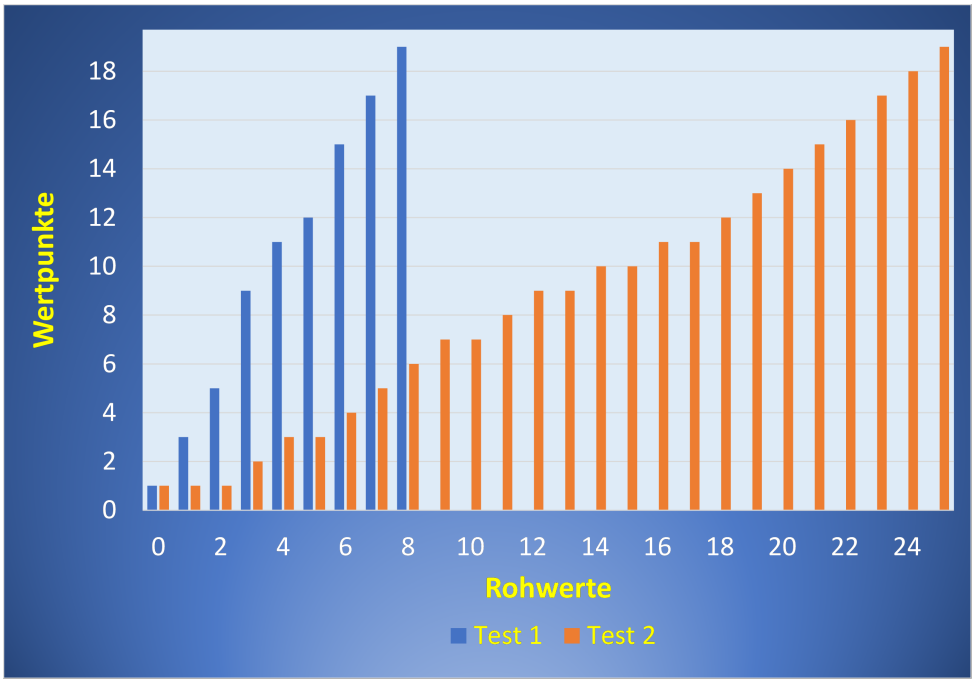

Abbildung 1: Grafische Darstellung der Itemgradienten für *Test 1* und *Test 2* aus Tabelle 1

Leistung 6, 7 oder 8 Wertpunkte beträgt, erhalten in *Test 1* trotzdem nur 5 Wertpunkte – ihre Leistung wird unterschätzt. Ein erster Effekt ungünstiger Itemgradienten ist also, dass etliche Testwerte nicht die wahren Fähigkeiten widerspiegeln. Ein zweiter Effekt ist eine erhöhte Anfälligkeit für Störfaktoren. Eine kurze Unaufmerksamkeit des Kindes oder eine sehr spezifische Wissenslücke kann dazu führen, dass nicht 3, sondern nur 2 Rohwertpunkte erzielt werden. Aus einem Durchschnittsergebnis wird übergangslos eine unterdurchschnittliche Leistung. Bei günstigeren Itemgradienten wird der Effekt solcher Störfaktoren geringer ausfallen.

Ungünstige Itemgradienten können in der Folge auch die Aussagekraft aller weiteren Auswertungen einschränken, die eine exakte Messung der Leistungen voraussetzen (z. B. intraindividuelle Vergleiche von Testwerten bei einer Profilanalyse oder längsschnittliche Vergleiche).

4.1 Gebrauchsanweisung: Itemgradienten erkennen und bewerten

Die Überprüfung der Itemgradienten muss immer erfolgen, wenn ein Test nur einen schmalen Rohwertbereich hat (vgl. *Test 1* in Tab. 1). Wenn es nur wenige Rohwertausprägungen gibt, kann es prinzipiell keine feine Abstufung der Normwerte geben. Ungünstige Itemgradienten können jedoch auch bei Tests auftreten, die theoretisch eine differenziertere Bewertung erlauben, z. B. wenn Decken- oder Bodeneffekte vorliegen. Testanwender*innen sollten daher immer überprüfen, ob die von ihnen eingesetzten Verfahren kritische Itemgradienten aufweisen. So gehen Sie vor:

- Schlagen Sie die Normtabelle auf.
- Beginnen Sie mit dem niedrigsten Rohwert, und lesen Sie den zugehörigen Standardwert ab.
- Lesen Sie beim nächsten Rohwert den zugehörigen Standardwert ab. Ermitteln Sie die Differenz zwischen den beiden Standardwerten.

- Wiederholen Sie diesen Schritt, bis Sie den höchst möglichen Rohwert erreicht haben.
- Stellen Sie nun fest, ob die von Ihnen ermittelten Differenzen mehr als 1/3 Standardabweichung betragen (s. o.). Falls ja, haben Sie ungünstige Itemgradienten entdeckt. Sie müssen dann davon ausgehen, dass der entsprechende Standardwert des Kindes mit einer zusätzlichen Ungenauigkeit behaftet ist, die über den reliabilitätsbedingten Messfehler hinausgeht.
- Wenn Sie sich einen umfassenden Überblick über die Itemgradienten eines Testverfahrens verschaffen wollen, wiederholen Sie diese Prozedur für alle Normgruppen.

In Normtabellen, die wie in Tabelle 1 aufgebaut sind, sind Sprünge in den Normwerten auch schnell an den Leerstellen in der Tabelle zu erkennen.

Will man Itemgradienten nicht für einen ganzen Test, sondern nur für ein einzelnes Testergebnis beurteilen, kostet dies nur wenige Sekunden:

- Lesen Sie den Standardwert ab, der dem erzielten Rohwert entspricht.
- Lesen Sie anschließend die Standardwerte ab, die mit dem nächsthöheren und dem nächstniedrigeren Rohwert erzielt worden wären.
- Fällt der Unterschied zu den so ermittelten Standardwerten größer als ein 1/3 Standardabweichung aus, bedenken Sie bei der Interpretation des Testergebnisses die Folgen ungünstiger Itemgradienten. Beachten Sie dabei vor allem die Möglichkeit, dass eine große Diskrepanz zum nächsthöheren Normwert zu einer Unterschätzung der Leistung führen kann.

Vereinzelt werden Sie in Tests Normwerte finden, die nur sehr grob abgestuft sind. So liegen bei einer Stanine-Skala mit einem Mittelwert von 5 und einer Standardabweichung von 2 immer ungünstige Itemgradienten vor, sofern nur ganzzahlige Werte ablesbar sind. Der Unterschied zwischen zwei benachbarten Standardwerten bei dieser Skala beträgt dann mindestens eine halbe Standardabweichung.

Es sei noch auf eine Besonderheit hingewiesen: In einigen Fällen (z. B. Raven's 2; NCS Pearson, 2019) werden Rohwerte erst in sogenannte Leistungs- oder Wachstumswerte überführt, zu denen dann Normwerte ermittelt werden. Hier muss bei der Überprüfung der Itemgradienten zuerst festgestellt werden, welche Leistungs-/Wachstumswerte den Rohwerten zugeordnet sind. Dann werden in den Normtabellen die zugehörigen Standardwerte abgelesen. Der Blick in die eigentlichen Normtabellen genügt nicht, da dort womöglich Wachstumswerte tabelliert sind, die effektiv gar nicht auftreten können. Das ist z. B. beim Raven's 2 der Fall. In der Normtabelle ist jedem IQ-Wert ein Wachstumswert zugeordnet. Das sieht auf den ersten Blick nach einer vorbildhaft feinen Abstufung der Standardwerte aus. In einzelnen Fällen sind jedoch durchaus ungünstige Itemgradienten festzustellen: Den Rohwerten 6 und 7 im Raven's 2 entsprechen die Leistungswerte 319 und 343 (dazwischen liegende Leistungswerte können nicht gemessen werden). Die zugehörigen IQ-Werte liegen z. B. bei 55 und 65 (Normgruppe 4;0 bis 4;3 Jahre) oder 44 und 55 (Normgruppe 6;0 bis 6;3 Jahre).

4.2 Übung

Sie können die Identifikation und Bewertung von Itemgradienten anhand von Tabelle 2 üben, in der ein weiterer gebräuchlicher Aufbau von Normtabellen verwendet wird. Die Normwerte

Tabelle 2: Übungstabelle zur Beurteilung von Itemgradienten

	Rohwerte															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Test 1	55	61	71	96	106	111	125	134	143							
Test 2	55	55	55	57	59	61	65	76	90	94	98	101	105	109	121	133
Test 3	64	79	95	104	116	120	124	131	135	137	139	144	145	145	145	145

entsprechen der sog. IQ-Skala mit einem Mittelwert von 100 und einer Standardabweichung von 15. Die IQ-Skala erlaubt prinzipiell eine feinere Abstufung als die Wertpunktskala. Sie garantiert aber nicht, dass alle möglichen Normwertausprägungen auch tatsächlich vorkommen. Am Ende dieses Textes finden Sie Hinweise, die Sie mit Ihrer Beurteilung von Tabelle 2 abgleichen können.

5 Altersdifferenzierung

Bei der Untersuchung von Kindern werden die Testleistungen natürlich vom Lebensalter beeinflusst. Die kindliche Entwicklung steht nicht still. Rohwerte, die für eine Altersstufe überragend sind, können wenige Jahre später einem weit unterdurchschnittlichen Ergebnis entsprechen.

Testverfahren, die bei Kindern und Jugendlichen eingesetzt werden, müssen daher Altersnormen bereitstellen. Diese Normen müssen so fein abgestuft sein, dass das Lebensalter keinen oder nur einen geringen Einfluss auf die Normwerte hat. Das gilt bei besonders bei kognitiven Leistungstests für die ersten Lebensjahre. Bei Schulleistungstests ist die Schulbesuchsdauer bedeutsam. Man kann die Lese- oder Rechenfertigkeiten von Kindern am Anfang und am Ende des ersten Schuljahres nicht mit ein und denselben Normwerten beurteilen. Die folgenden Hinweise zur Bewertung und Beurteilung der Altersdifferenzierung können auf die Analyse von Normtabellen, die nach Klassenstufen gegliedert sind, problemlos übertragen werden.

Ein Test mit einer günstigen Altersdifferenzierung zeichnet sich dadurch aus, dass im Vergleich benachbarter Altersgruppen keine deutlichen Diskrepanzen zwischen den Normwerten auftreten. Bei einem Test mit einer ungünstigen Altersdifferenzierung kann im Extremfall ein einziger Tag Unterschied bei der Testdurchführung zu deutlich abweichenden Ergebnissen führen. Hierzu ein Beispiel: Im Sprachentwicklungstest für zweijährige Kinder (SETK 2; Grimm, 2016) erhält ein Kind im Alter von 2 Jahren 5 Monaten und 31 Tagen im Untertest *Produktion II Sätze* bei einem Rohwert von 20 einen T-Wert von 48. Das ist ein durchschnittlicher Wert. Wird der Test nur einen Tag später, also im Alter von 2 Jahren 6 Monaten und 0 Tagen durchgeführt, liegt das Ergebnis bei einem T-Wert von 37. Das ist mehr als eine Standardabweichung niedriger und wird üblicherweise als unterdurchschnittlich bewertet.

Die Abbildungen 2 und 3 veranschaulichen die Auswirkungen von mehr oder weniger breiten Altersgruppen anhand fiktiver Daten. Dargestellt ist die durchschnittliche Testleistung in Abhängigkeit vom Alter der untersuchten Kinder (3;0 bis 6;11 Jahre). In Abbildung 2 wurden vier Altersgruppen gebildet, die jeweils ein Jahr umspannen. Die angegebenen Werte sind die Mittelwerte der Altersgruppen, die vertikalen Linien markieren eine Standardabweichung über und unter diesen Mittelwerten.

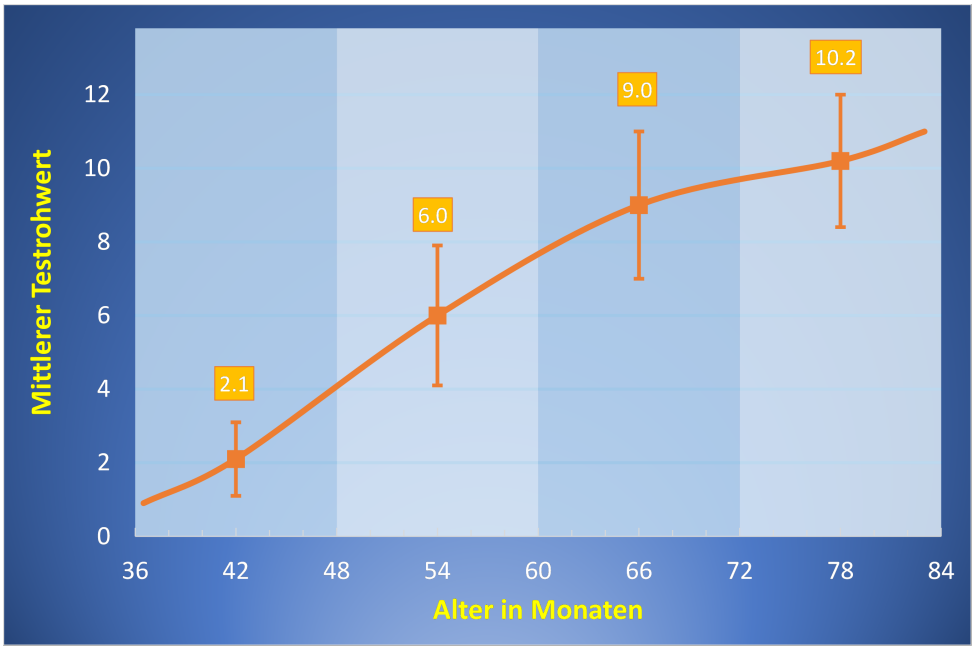

Abbildung 2: Beispielhafter Entwicklungsverlauf von Testrohwerten bei einer Einteilung in vier Altersgruppen

Bei dieser Einteilung finden sich recht deutliche Unterschiede zwischen den Altersgruppen. Ein Beispiel: Der durchschnittliche Rohwert steigt vom vierten zum fünften Lebensjahr um 3.9 Punkte. Kinder, die im vierten Lebensjahr mit ihrem Rohwert einen Durchschnittswert erreichen, fallen im fünften Lebensjahr (nächste Altersgruppe) mit dem gleichen Rohwert in den weit unterdurchschnittlichen Bereich. Je nachdem, ob ein Kind kurz vor oder kurz nach Überschreiten einer Altersgrenze getestet wird, werden sich in diesem Beispiel ganz erhebliche Unterschiede im Normwert ergeben.

Auch innerhalb der Altersgruppen sind Leistungsunterschiede zu erkennen: Mit 36 Monaten liegt der Mittelwert etwa bei 0.9, mit 47 Monaten knapp unter 4.0. Es werden also Kinder mit ganz unterschiedlichen Fähigkeiten mit ein und derselben Normtabelle bewertet. Praktisch bedeutsam ist eine solche ungünstige Altersdifferenzierung bei Kindern, deren Testalter am oberen (Gefahr der Überschätzung) oder unteren Rand (Gefahr der Unterschätzung) einer Normgruppe liegt.

In Abbildung 3 ist zu erkennen, dass die Problematik einer ungünstigen Altersdifferenzierung vermindert werden kann, wenn die Normgruppen nur noch vier Monate und nicht 12 Monate umfassen. Sowohl zwischen wie innerhalb der Gruppen sind die Diskrepanzen der Rohwerte geringer. Dementsprechend können sich so auch für Normwerte feinere Abstufungen ergeben.

Es gibt bei Testverfahren beträchtliche Unterschiede, wie groß die Altersspanne ist, die jeweils von einer Normgruppe abgedeckt wird. Hierzu einige Beispiele: Bei 5-Jährigen bietet die Kaufman Assessment Battery for Children – II (KABC -II; Melchers & Melchers, 2015) vier Normgruppen, die jeweils drei Monate umfassen. Im Sprachentwicklungstest für drei- bis fünfjährige Kinder (SETK 3-5; Grimm, Aktas & Frevert, 2001) gibt es dagegen für dieses Alter eine einzige Normgruppe. Die Leistung eines Kindes im Alter von 5;0 Jahren wird genau mit dem gleichen Maßstab bewertet wie die eines Kindes eines Alters von 5;11 Jahren. Im Verbalen Lern- und

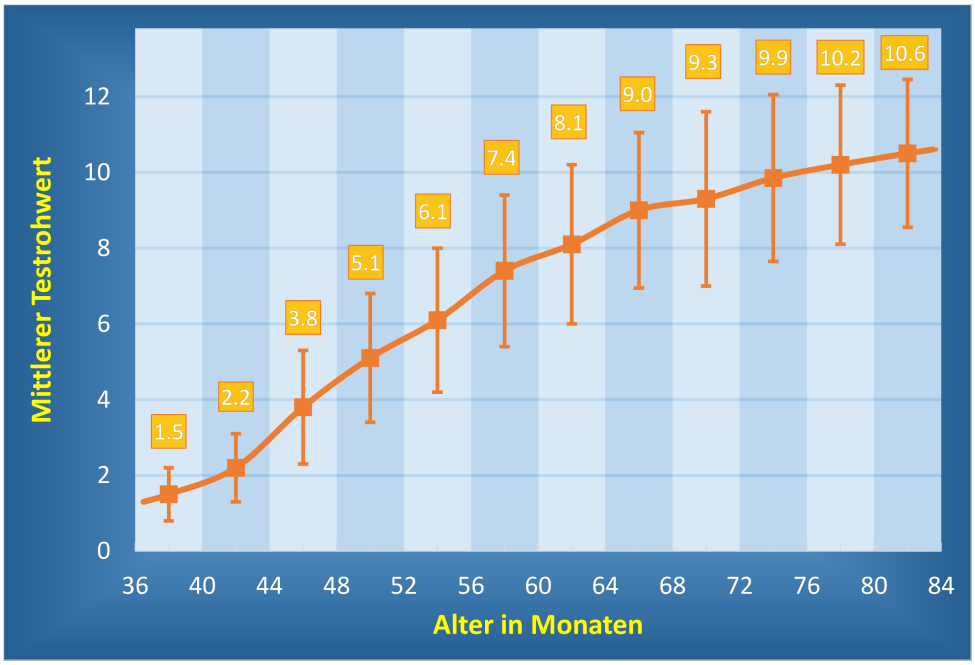

Abbildung 3: Beispielhafter Entwicklungsverlauf von Testrohwerten bei einer Einteilung in zwölf Altersgruppen

Merkfähigkeitstest (VLMT; Helmstaedter, Lendt & Lux, 2001) werden sogar Kinder im Alter von 6- bis 9 Jahren in einer einzigen Normgruppe zusammengefasst.

Pauschale Kriterien zur Bewertung der Altersspanne von Normgruppen sind nur bedingt sinnvoll. Entscheidend ist ja immer, wie sich die Testrohwerte in Abhängigkeit vom Alter typischerweise entwickeln. Dieser normale Entwicklungsverlauf wird sich für verschiedene Entwicklungsbereiche unterscheiden und ist zudem von der jeweiligen Itemzusammenstellung abhängig. Bracken (2000) hält im Vorschulalter einen Zeitraum von bis zu 3 Monaten für angemessen, bis zum Alter von zwei Jahren fordert er eine Normierung im Abstand von 1 bis 2 Monaten.

5.1 Gebrauchsanweisung: Altersdifferenzierung erkennen und bewerten

Die Altersdifferenzierung sollte immer überprüft werden, wenn Kinder im Vorschulalter untersucht werden und/oder wenn die Normgruppen einen Altersbereich umfassen, der breiter ausfällt, als die genannten Empfehlungen von Bracken (2000). Das Vorgehen ist im Prinzip einfach, kann sich aber je nach Gestaltung der Normtabelle als mühselig erweisen. Das hat damit zu tun, dass die Normwerte für unterschiedliche Altersgruppen oft in getrennten Normtabellen zu finden sind, die man nicht unmittelbar vergleichen kann.

- Schlagen Sie die Normtabelle auf. In Normtabellen, die wie in Tabelle 3 aufgebaut sind, ist der direkte Vergleich der Altersgruppen einfach möglich. Wenn Altersgruppen jedoch auf verschiedene Seiten des Manuals verteilt sind, ist es hilfreich, eine kleine Tabelle anzulegen, in der bei ausgewählten Rohwerten die Standardwerte für verschiedene Normgruppen eingetragen werden.
- Beginnen Sie die Bewertung der Altersdifferenzierung am besten mit der jüngsten oder ältesten Normgruppe. Das erleichtert ein systematisches Vorgehen.

Tabelle 3: Normtabelle mit Differenzen zwischen den Standardwerten benachbarter Altersgruppen

Rohwert	6;0-6; 11	*Differenz*	7;0-7;11	*Differenz*	8;0-8;11
0	69	*11*	58	*10*	48
1	72	*12*	60	*9*	51
2	75	*10*	65	*11*	54
3	79	*13*	66	*10*	56
4	83	*10*	73	*8*	65
5	87	*11*	76	*7*	69
6	90	*12*	78	*8*	70
7	93	*10*	83	*9*	74
8	96	*11*	85	*8*	77
9	100	*11*	89	*8*	81
10	103	*12*	91	*6*	85
11	105	*10*	95	*7*	88
12	108	*9*	99	*8*	91
13	112	*12*	100	*7*	93
14	117	*11*	106	*6*	100
15	121	*10*	111	*6*	105
16	128	*12*	116	*5*	111
17	133	*9*	124	*6*	118
18	139	*8*	131	*5*	126
19	145	*7*	138	*4*	134
20	145	*6*	139	*5*	134

- Ermitteln Sie für mehrere Rohwerte die zugehörigen Standardwerte. Da sich Probleme bei der Altersdifferenzierung auf einzelne Normwertbereiche beschränken können, sollten Sie wenigstens einen niedrigen, einen mittleren und einen hohen Rohwert wählen. Schauen Sie in den Bereichen der Normtabelle genauer hin, die für Ihre diagnostische Tätigkeit besonders relevant sind.
- Gehen Sie jetzt zur Normtabelle einer benachbarten Altersgruppe.
- Ermitteln Sie dort für dieselben Rohwerte die zugehörigen Standardwerte.
- Bilden Sie die Differenzen zwischen den jeweiligen Standardwerten in beiden Altersgruppen. Diese Differenzen zeigen an, welche Veränderungen sich bei Überschreiten der Altersgrenze ergeben.
- Wiederholen Sie dieses Vorgehen für verschiedene benachbarte Altersgruppen.
- Für die quantitative Beurteilung der Altersdifferenzierung gibt es keine verbindlichen Kriterien. Analog zur Bewertung von Itemgradienten können Sie beachten, ob eine oder mehrere der ermittelten Differenzen größer sind als 1/3 einer Standardabweichung. Wenn das Alter des untersuchten Kindes am oberen oder unteren Ende des normierten Altersbereichs liegt, wird der Standardwert des Kindes dann mit einer zusätzlichen Ungenauigkeit behaftet sein, die über den üblichen reliabilitätsbedingten Messfehler hinausgeht. Bei Kindern, deren Alter im mittleren Bereich der Normgruppe liegt, sind Verzerrungen kaum zu erwarten.

Für ein konkretes Testergebnis können Sie die Altersdifferenzierung mit weniger Aufwand überprüfen:

- Lesen Sie den Standardwert für den Rohwert des untersuchten Kindes ab.
- Wenn das Alter des Kindes an der oberen oder unteren Grenze der Normgruppe liegt, bestimmen Sie für diesen Rohwert den Standardwert in der benachbarten Normgruppe.
- Fällt der Unterschied zwischen den so ermittelten Standardwerten größer als 1/3 Standardabweichung aus, sollten Sie bei der Interpretation des Testergebnisses mögliche Folgen einer schlechten Altersdifferenzierung berücksichtigen. Überlegen Sie, ob sich Ihre Interpretation des Testergebnisses verändern würde, wenn sie das Kind etwas früher oder später getestet hätten.

In vermutlich sehr seltenen Fällen kann es auch vorkommen, dass die Unterschiede zwischen benachbarten Altersgruppen minimal oder nicht-existent sind. Bei sehr fein abgestuften Altersgruppen ist dies unkritisch. Wenn sich jedoch – das gilt wiederum besonders für das Vorschulalter – bei breiten Altersgruppen keine Unterschiede finden, sind Zweifel an der Testqualität angebracht. Ein Extrembeispiel ist der Subtest *Morphologische Regelbildung* im SETK 3-5. Hier finden sich in der Normtabelle für die Altersgruppe 4;6 bis 4;11 Jahre exakt die gleichen Normwerte wie in der Altersgruppe 5;0 bis 5;11 Jahre. Es ist höchst fragwürdig, dass sich grammatikalische Fähigkeiten in einem Zeitraum von 1½ Jahren nicht weiterentwickeln.

In neueren Testverfahren werden Normen für verschiedene Altersgruppen nicht selten auf Basis eines sog. kontinuierlichen Normierungsmodells (vgl. Wasserman & Bracken, 2013) exakt für das jeweilige Alter der Testperson berechnet. Solche Tests bieten in der Regel eine taggenaue computergestützte Testauswertung und/oder sehr fein abgestufte Normtabellen an (z. B. im Monatsabstand beim Nonverbalen Intelligenztest SON-R 2-8; Tellegen, Laros & Petermann, 2018). In einem solchen Fall braucht man sich dann keine Gedanken über die Altersdifferenzierung zu machen.

5.2 Übung

Tabelle 4 bietet die Gelegenheit, das beschriebene Vorgehen auszuprobieren. Die dort angegebenen Normwerte sind dieses Mal T-Werte mit einem Mittelwert von 50 und einer Standardabweichung von 10. Beurteilen Sie für die beiden Tests die Altersdifferenzierung beim Übergang von den 4- zu den 5-Jährigen. Ihre Ergebnisse können Sie wieder mit Hinweisen am Ende des Textes abgleichen.

Tabelle 4: Übungstabelle zur Beurteilung der Altersdifferenzierung

4 Jahre	Rohwerte															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Test 1	32	36	39	42	45	47	50	53	55	57	59	65	71	76	80	80
Test 2	36	38	41	45	48	51	53	55	58	60	62	65	67	71	76	80
5 Jahre	Rohwerte															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Test 1	27	30	33	37	40	43	46	49	52	54	56	61	67	72	76	80
Test 2	27	30	32	35	38	43	45	48	52	54	56	59	62	66	71	76

6 Zurück zum Diagnostik-Quiz

Bei allen Fragen lautet die richtige Antwort: »Das kann man nicht mit Sicherheit sagen«. Dafür kann es viele Gründe geben, z. B. Auswertungs- und Durchführungsfehler, Störfaktoren während der Testung oder eine geringe Reliabilität der eingesetzten Verfahren. Zusätzlich sollten Sie auch den möglichen Einfluss ungünstiger Itemgradienten und einer schlechten Altersdifferenzierung bedenken:

Frage 1: Eine Antwort ist erst möglich, wenn die Itemgradienten überprüft wurden. Charlotte hat einen Standardwert von 71 erreicht. Nehmen wir einen ausgesprochen ungünstigen Itemgradienten an, bei dem eine zusätzlich gelöste Aufgabe zu einem Standardwert von 96 geführt hätte (s. Tabelle 2, *Test 1*). Die Werte 72 bis 95 können in diesem Test prinzipiell nicht erreicht werden. Möglicherweise wird der wahre Leistungsstand unterschätzt, weil er in diesem Test nicht adäquat erfasst werden kann. Sie können dann nicht mit gutem Gewissen schreiben, dass Charlottes Leistung eindeutig unterdurchschnittlich ausgefallen ist.

Frage 2: Bei dem verwendeten Aufgabentyp (Auswahl einer Lösung aus mehreren Alternativen) besteht immer die Möglichkeit, dass korrekte Antworten zufällig erzielt werden. Bei vier Auswahlmöglichkeiten liegt die Wahrscheinlichkeit, eine Aufgabe allein durch Raten zu lösen, bei 25%. Die Kombination von Ratewahrscheinlichkeit und einem ungünstigen Itemgradienten kann zu deutlichen Verzerrungen von Testergebnissen führen. Sebastian hat sechs Aufgaben bearbeitet und drei Aufgaben richtig gelöst. Der Standardwert – abgelesen beim *Test 1* in Tabelle 2 – beträgt 96. Die Wahrscheinlichkeit, bei sechs Aufgaben wenigstens eine Lösung zufällig zu erzielen, liegt bei 82.2%. Es lohnt sich also einmal nachzuschauen, welcher Standardwert einem Rohwert von 2 entspricht. Aufgrund des extrem ungünstigen Itemgradienten liegt dieser Wert mit 71 erheblich niedriger als der ermittelte Standardwert von 96. Aufgrund des Testergebnisses kann also keine sichere Aussage über Sebastians Leistungsfähigkeit getroffen werden. Es ist allerdings nicht zulässig, den Wert einfach zu korrigieren. Zufallsergebnisse sind ja auch in die Normierung eingeflossen. Es wäre daher nicht gerechtfertigt, einen um die Ratewahrscheinlichkeit korrigierten Rohwert in Normwerte umzurechnen, die ohne diese Korrektur ermittelt wurden.

Frage 3: Grundsätzlich liefert die Erfassung sprachlicher und kognitiver Leistungen mit einem standardisierten Testverfahren zuverlässigere Ergebnisse als subjektive Einschätzungen durch Bezugspersonen. Allerdings ist bei *Frage 3* eine Überprüfung der Altersdifferenzierung angezeigt. Gibt es in dem durchgeführten Test möglicherweise eine große Diskrepanz zwischen den Normwerten in benachbarten Altersgruppen? Es sollte also überprüft werden, welchen Normwert Petra erhalten hätte, wenn sie wie Peter kurz vor dem 5. Geburtstag und nicht wenige Tage später untersucht worden wäre. Erst dann ist eine inhaltliche Interpretation des Testwerts sinnvoll.

7 Konsequenzen für die Auswahl und den Einsatz von Testverfahren

Testanwender*innen sollten die von ihnen eingesetzten Testverfahren und deren Eigenschaften gut kennen. Problematische Itemgradienten und/oder eine ungünstige Altersdifferenzierung werden in Testmanualen in der Regel nicht thematisiert. Um sie zu erkennen, bedarf es der Eigeninitiative.

Das hier dargestellte Vorgehen kann mit ein wenig Übung zu einer schnell durchführbaren Routine werden. Wenden Sie es bei allen Testverfahren an, die Sie einsetzen.

- Die Beurteilung von Itemgradienten ist besonders wichtig bei Testverfahren, in denen nur wenige Rohwertausprägungen möglich sind. Die Bewertung der Altersdifferenzierung hat ihre größte Bedeutung (a) bei Testverfahren für jüngere Kinder, (b) bei Kindern, deren Alter an den Grenzen der Normgruppe liegt und (c) bei Normgruppen, die einen breiten Altersbereich umfassen. Sie können sich auf einzelne Stichproben beschränken, wenn die o. g. Empfehlungen von Bracken (2000) zur Breite von Normgruppen von den Testautor*innen berücksichtigt wurden.
- Vermeiden Sie, wo immer möglich, den Einsatz von Testverfahren, deren Itemgradienten oder Altersdifferenzierung die Interpretation von individuellen Testergebnissen erschweren können.
- Fragen Sie vor Anschaffung eines Testverfahrens bei den Testverlagen, ob Sie ein Testmanual, das die Normtabellen enthält, zur Ansicht erhalten können. Vermeiden Sie die Anschaffung von Testverfahren, bei denen extrem ungünstige Itemgradienten und/oder eine sehr ungünstige Altersdifferenzierung zu fragwürdigen oder ganz nutzlosen Testbefunden führen können.
- Machen Sie bei der Darstellung von Testergebnissen deutlich, ob diese durch ungünstige Itemgradienten und/oder eine ungünstige Altersdifferenzierung beeinflusst sein könnten. Leser*innen von sonderpädagogischen Gutachten und klinisch-psychologischen Befundberichten werden sich sonst nur an den von Ihnen dokumentierten Testwerten orientieren.
- Denken Sie daran, dass Testwerte aus Fremdbefunden von den hier beschriebenen Phänomenen beeinflusst sein können. Wenn solche Testwerte in Ihre eigene diagnostische Einschätzung einfließen, sollten Sie sich über die Eigenschaften der verwendeten Tests informieren und idealerweise deren Normtabellen selbst einsehen.
- Bei Testwerten, die aus mehreren Untertests gebildet werden (z. B. Gesamtwerte von Intelligenztests), sind problematische Itemgradienten selten zu finden. Das oben geschilderte Vorgehen kann aber im Bedarfsfall analog angewendet werden. Statt der Itemrohwerte verwenden Sie dann diejenigen Werte (meist die Summe der Standardwerte der Untertests), die im jeweiligen Testverfahren der Bestimmung der Gesamtwerte zugrunde liegen.
- Bei der Testung von Jugendlichen wird eine ungünstige Altersdifferenzierung kaum eine Rolle spielen. Eine Überprüfung bietet sich trotzdem an, wenn es Grund zur Annahme gibt, dass auch im Jugendalter noch bedeutsame Veränderungen im untersuchten Leistungsbereich auftreten können. Auch wenn Normgruppen mehrere Jahre umfassen, kann bei Jugendlichen eine stichprobenartige Überprüfung sinnvoll sein.

8 Hinweise zu den Übungstabellen

8.1 Tabelle 2:

- Bei *Test 1* ist gleich klar, dass es ungünstige Itemgradienten geben muss. Es können nur neun Standardwerte abgelesen werden. Maximal ergibt sich zwischen benachbarten

Rohwerten eine Diskrepanz von 25 IQ-Punkten (Rohwert 2 vs. Rohwert 3). Nur in einem einzigen Fall (Rohwert 4 vs. Rohwert 5) liegt die Differenz zwischen den Normwerten bei 1/3 Standardabweichung, sonst immer deutlich darüber.

- *Test 2* könnte im Grunde feiner abgestufte Werte liefern. Immerhin sind 16 unterschiedliche Ausprägungen der Rohwerte möglich. Aber auch das lässt noch ungünstige Itemgradienten erwarten. Außerdem werden einigen Rohwerten (0, 1, 2) identische Standardwerte zugeordnet, so dass in der Tabelle effektiv nur 14 unterschiedliche Standardwerte zu finden sind. Im unteren Leistungsbereich zeigen sich erst einmal unkritische Übergänge. Der Sprung von 76 auf 90 (Rohwert 7 vs. Rohwert 8) ist jedoch deutlich zu groß, auch bei den höchsten Testwerten liegen kritische Itemgradienten vor.
- *Test 3* liefert nur 13 unterscheidbare Standardwerte. Im unteren und mittleren Wertebereich finden sich mehrere kritische Itemgradienten, am deutlichsten beim Übergang von Rohwert 1 zu Rohwert 2. Die Differenz der Standardwerte beträgt mit 16 IQ-Punkten mehr als eine Standardabweichung.

8.2 Tabelle 4:

Der Vergleich der Normwerte für die beiden Altersgruppen in Tabelle 4 war aufwändiger, da die Werte der beiden Altersgruppen nicht direkt nebeneinander aufgeführt waren. Tabelle 5 zeigt eine andere Anordnung, mit der Sie Ihre Ergebnisse zur Altersdifferenzierung kontrollieren können. Liegt das Testalter gerade an der Grenze der Altersgruppe, kann ein einziger Tag Unterschiede zwischen 0.3 und 0.6 Standardabweichungen bei *Test 1* und zwischen 0.6 und 1.0 Standardabweichungen bei *Test 2* bewirken.

Tabelle 5: Normtabelle mit Differenzen zwischen den Standardwerten benachbarter Altersgruppen

Test 1	Rohwerte															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4 Jahre	32	36	39	42	45	47	50	53	55	57	59	65	71	76	80	80
5 Jahre	27	30	33	37	40	43	46	49	52	54	56	61	67	72	76	80
Differenz	5	6	6	5	5	4	4	4	3	3	3	4	4	4	4	0
Test 2	Rohwerte															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4 Jahre	36	38	41	45	48	51	53	55	58	60	62	65	67	71	76	80
5 Jahre	27	30	32	35	38	43	45	48	52	54	56	59	62	66	71	76
Differenz	9	8	9	10	10	8	8	7	6	6	6	6	7	5	5	4

Literatur

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. Journal of Psychoeducational Assessment, 4, 313–326.

Bracken, B. A. (2000). Maximizing construct relevant assessment. The optimal preschool testing situation. In B. A. Bracken (Ed.), The psychoeducational assessment of preschool children (pp. 33–40). Boston: Allyn & Bacon.

Grimm, H. (2016). SETK-2. Sprachentwicklungstest für zweijährige Kinder (2., überarbeitete und neu normierte Auflage). Göttingen: Hogrefe.

Grimm, H., Aktas, M. & Frevert, S. (2001). Sprachentwicklungstest für drei- bis fünfjährige Kinder (SETK 3-5). Göttingen: Hogrefe.

Helmstaedter, C., Lendt, M. & Lux, S. (2001). Verbaler Lern- und Merkfähigkeitstest (VLMT). Göttingen: Hogrefe.

Melchers, P. & Melchers, M. (2015). KABC-II. Kaufman Assessment Battery for Children – II von Alan S. Kaufman & Nadeen L. Kaufman. Deutschsprachige Fassung. Frankfurt a. M.: Pearson.

NCS Pearson. (2019). Raven's 2. Progressive Matrices Clinical Edition. Deutsche Fassung. Frankfurt a. M.: Pearson.

Tellegen, P. J., Laros, J. A. & Petermann, F. (2018). SON-R 2-8. Non-verbaler Intelligenztest. Göttingen: Hogrefe.

Wasserman, J. D. & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In I. B. Weiner, J. R. Graham & J. A. Naglieri (Eds.), Assessment Psychology (Handbook of Psychology, vol. 10, 2nd ed., S. 50–81). Hoboken, NJ: Wiley.

Prof. Dr. Gerolf Renner ist Professor für Psychologie und Diagnostik im Förderschwerpunkt körperliche und motorische Entwicklung an der Pädagogischen Hochschule Ludwigsburg. https://orcid.org/0000-0003-4345-3619