Renner, G. (2022). Normtabellen analysieren und beurteilen I: Bodeneffekte erkennen und verstehen. In M. Gebhardt, D. Scheer & M. Schurig (Hrsg.), *Handbuch der sonderpädagogischen Diagnostik. Grundlagen und Konzepte der Statusdiagnostik, Prozessdiagnostik und Förderplanung* (S. 275-290). Regensburg: Universitätsbibliothek. https://doi.org/10.5283/epub.53149

Normtabellen analysieren und beurteilen I: Bodeneffekte erkennen und verstehen

Gerolf Renner

1 Einleitung

Zum Einstieg ein kleines Diagnostik-Quiz:

- Ein 4;1-jähriges Kind wird mit einem Intelligenztest, der auch zwei Wortschatztests enthält, untersucht. Im Untertest *Wortschatz Aktiv* erreicht es einen Standardwert (Mittelwert = 100, Standardabweichung = 15) von 80, im Untertest *Wortschatz Passiv* einen Standardwert von 58. In welchem Untertest hat es besser abgeschnitten?
- Das gleiche Kind erreicht bei einer Testwiederholung im Alter von 6;0 Jahren im Untertest *Wortschatz Aktiv* einen Standardwert von 56. Bei der Erstuntersuchung lag der Standardwert bei 80. Hat sich seine Leistung verschlechtert?
- In einem Bericht über eine psychologische Untersuchung steht, dass ein Kind in einem Untertest einen Prozentrang von 79 erreicht hat. Liegt dieser Testwert im durchschnittlichen Bereich?

Wie beantworten Sie diese Fragen?

2 Bodeneffekte und ihre Folgen

Anwender*innen von Testverfahren erwarten, dass die Testergebnisse etwas über die Leistungen des untersuchten Kindes aussagen. Warum sonst sollte man sich auch der Mühe einer Testung unterziehen? Leider wird diese Erwartung nicht immer erfüllt. *Ein* Grund hierfür kann sein, dass ein Test Bodeneffekte aufweist (Bracken, 1988). Ein Bodeneffekt liegt vor, wenn ein Test keine Items im unteren Schwierigkeitsbereich enthält. Bei starken Bodeneffekten können altersgemäße Ergebnisse nicht von unterdurchschnittlichen Leistungen abgegrenzt werden. Bodeneffekte sind daher hoch relevant, wenn Tests bei Menschen eingesetzt werden, deren Fähigkeiten im unterdurchschnittlichen oder weit unterdurchschnittlichen Bereich liegen.

Die Fähigkeit, Bodeneffekte in normierten Testverfahren zu erkennen und zu bewerten, ist eine unabdingbare Voraussetzung für deren verantwortungsvollen Einsatz. Ausgeprägte Bodeneffekte können gravierende Folgen haben:

- Testergebnisse spiegeln nicht die wahre Leistungsfähigkeit der Kinder wider. Die Leistungen von Kindern mit deutlichen Beeinträchtigungen werden überschätzt.
- Bei Verlaufsmessungen kann es zu vermeintlichen Leistungsverschlechterungen kommen, auch wenn ein Kind Entwicklungsfortschritte gemacht hat.
- In Testprofilen zeigen sich Diskrepanzen zwischen einzelnen Leistungsbereichen, auch wenn sich die wahren Leistungen des Kindes nicht wesentlich unterscheiden.
- Bei Kindern mit deutlichen Entwicklungsstörungen und Behinderungen kann die Testung zu einem frustrierenden und demotivierenden, möglicherweise sogar psychisch belastenden Erlebnis werden: Sie werden nur sehr wenige Aufgaben lösen und erleben erheblich mehr Misserfolgs- als Erfolgserlebnisse.

Dieser Beitrag setzt im Folgenden voraus, dass Leser*innen 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]). Begriffe wie Mittelwert, Standardabweichung, Wertpunkt, IQ-Wert, Prozentrang, Standardwert und Rohwert sollten bekannt sein.

In diesem Text verwende ich durchgehend Beispiele, die realen Normtabellen nachempfunden sind. Auch wenn es den Leser*innen an einigen Stellen unglaubwürdig erscheinen mag: Ich kann versichern, dass ich an keiner Stelle Bodeneffekte darstellen werde, die extremer ausfallen als in real existierenden diagnostischen Instrumenten.

3 Wie kommen Bodeneffekte zustande?

Ausgangspunkt bei der Berechnung von Standardwerten sind die in einer Normierungsstichprobe erzielten Rohwerte. Testautor*innen haben sich nach mehr oder weniger umfangreichen Vorstudien für eine bestimmte Itemauswahl entschieden, mit der dann die Daten für die Normierung erhoben werden. Diese Items weisen typischerweise unterschiedliche Schwierigkeiten auf. Die Itemschwierigkeit bezieht sich bei Items, die zweistufig als »gelöst« oder »nicht gelöst« kodiert werden, auf den Anteil der Testpersonen in der Normstichprobe, die ein bestimmtes Item korrekt bearbeitet haben. Je schwieriger die einfachsten Items innerhalb eines Tests sind, desto mehr Kinder werden keine einzige Aufgabe lösen und einen Testrohwert von 0 erzielen.

Dies soll im Folgenden ausgehend von drei unterschiedlichen Rohwertverteilungen (Abb. 1) veranschaulicht werden. Die – fiktiven – Tests haben 20 Items, die jeweils mit 1 (gelöst) oder 0 (nicht gelöst) gewertet werden. Die Normierungsstichprobe umfasste 500 Kinder. In *Test 1* finden sich sowohl sehr einfache als auch sehr schwierige Items. Alle Kinder haben mindestens ein Item korrekt gelöst, dementsprechend hat kein Kind einen Testrohwert von 0 erzielt. Bei *Test 2* konnten 20 Kinder keine einzige Aufgabe bewältigen. Bei Test 3 waren es sogar 100 Kinder, die auch mit den einfachsten Aufgaben überfordert waren.

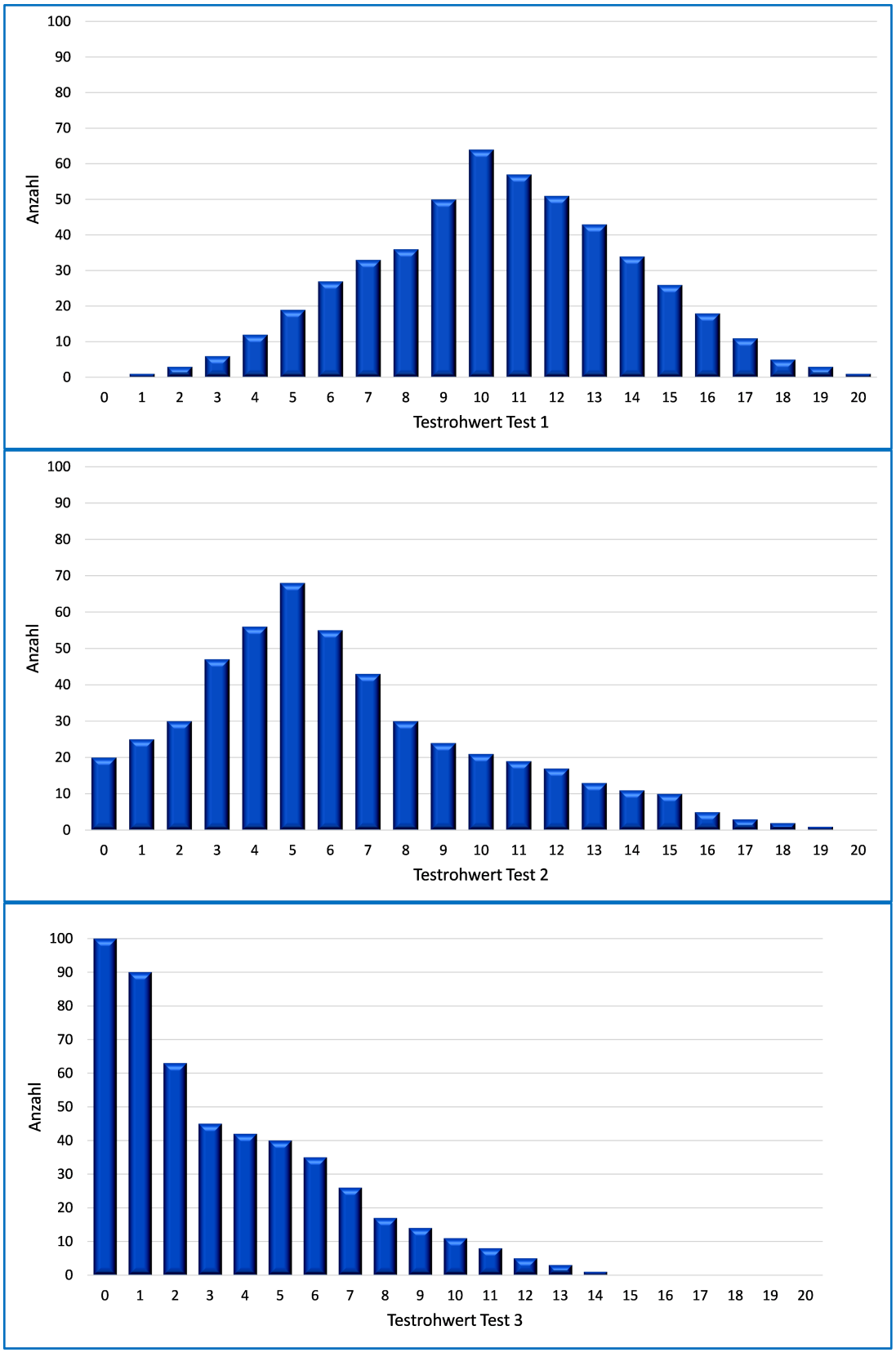

Abbildung 1: Beispiele für Verteilungen von Testrohwerten

Tabelle 1: Beispiele für Normtabellen

	Test 1	Test 2	Test 3
Wertpunkte	Rohwert	Rohwert	Rohwert
1	0-1		
2	2		
3	3		
4	4	0	
5	5	1	
6	6	2	0
7	7	3	
8	8	4	1
9	9	5	
10	10-11	6	2-3
11	12	7	4
12	13	8-9	5-6
13	14	10-11	7
14	15	12-13	8-9
15	16	14	10
16	17	15-16	11
17	18	17	12
18	19	18	13
19	20	19	14-20

Berechnet man auf Basis dieser Rohwertverteilungen Standardwerte (hier Wertpunkte mit einem Mittelwert von 10 und einer Standardabweichung von 3) ergeben sich die Normwerte in Tabelle 1. Bevor Sie weiterlesen, könnten sich einmal diese drei Normtabellen anschauen und überlegen, welche Besonderheiten Ihnen auffallen. Da die Rohwertverteilungen von *Test 2* und *Test 3* von einer Normalverteilung abweichen, wurden die Wertpunkte einheitlich auf Basis der Prozentränge mittels der Flächentransformation nach McCall berechnet (vgl. Lienert & Raatz, 1994).

4 Gebrauchsanweisung: Bodeneffekte erkennen und bewerten

Ein erster und einfacher Schritt bei der Ermittlung von Bodeneffekten ist es, in den Normtabellen nachzuschauen, welcher Wertebereich überhaupt tabelliert ist. Bei vielen deutschsprachigen Testverfahren wird bei Untertests die Wertpunktskala verwendet. Typischerweise umfassen die Normtabellen einen Wertebereich von 1 (3 Standardabweichungen unter dem Mittelwert, entsprechend einem Prozentrang von 0,1) bis 19 (3 Standardabweichungen über dem Mittelwert, entsprechend einem Prozentrang von 99,9). Damit wird ein sehr breiter Bereich abgedeckt. In den Extrembereichen ist jedoch eine nähere Differenzierung nicht möglich. Alle Testpersonen, deren wahrer Testwert bei oder unter einem Wertpunkt liegt, erhalten denselben Standardwert.

Bei Intelligenztests, deren Standardwerte (IQ-Werte) einen Mittelwert von 100 und eine Standardabweichung von 15 haben, finden sich unterschiedliche Wertebereiche, z. B. bei der Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V; Wechsler, 2017) und der Kaufman Assessment Battery for Children – II (KABC-II; Melchers & Melchers, 2015) von 40 bis 160 (± 4 Standardabweichungen), bei den kognitiven Untertests der Intelligence and Development Scales – Preschool (IDS-P; Grob, Reimann, Gut & Frischknecht, 2013) und beim Nonverbalen Intelligenztest (SON-R 2-8; Tellegen, Laros & Petermann, 2018) von 55 bis 145 (± 3 Standardabweichungen).

Mit dieser einfachen Inspektion der Normtabellen kann schon einmal festgestellt werden, welche niedrigsten und höchsten Testwerte überhaupt bestimmt werden können. Damit ist jedoch noch nicht klar, ob dieser Wertebereich tatsächlich ausgeschöpft wird.

In Testmanualen wird leider nur selten auf vorliegende Bodeneffekte hingewiesen. Testanwender*innen sind also gefordert, sich ein eigenes Bild zu verschaffen. Die folgende Anleitung enthält alles, was Sie wissen müssen, um Bodeneffekte festzustellen. Mit etwas Übung werden Sie in der Lage sein, Bodeneffekte mit einem minimalen Zeitaufwand zu erkennen. Zuerst geht es um Bodeneffekte in Untertests bzw. eindimensionalen (also nicht in Untertests gegliederten) Testverfahren.

4.1 Bodeneffekte in Untertests und eindimensionalen Verfahren

- Schlagen Sie die Normtabelle der Untertests oder des Gesamtwerts bei einem eindimensionalen Testverfahren auf.
- Beginnen Sie die Bewertung der Bodeneffekte mit der jüngsten Altersgruppe (das erleichtert ein systematisches Vorgehen erheblich). In vielen Testverfahren ändert sich die Testzusammenstellung mit dem Alter des Kindes. Wenn ein Untertest erst ab einem höheren Alter eingesetzt werden, beginnen Sie die Prozedur in der Altersgruppe, in der der Untertest zuerst eingesetzt wird.
- Suchen Sie in der Normtabelle den Standardwert, der bei einem von Rohwert 1 vergeben wird. In Tabelle 1 ist das 1 Wertpunkt bei *Test 1*, bei *Test 2* sind es 5 Wertpunkte und bei *Test 3* sogar 8 Wertpunkte.
- Wenn der sehr seltene Fall auftritt, dass Rohwerte zwischen 0 und 1 vorkommen können (z. B. ein Rohwert von 0.5 im Untertest *Gedächtnis Räumlich-Visuell* der IDS-P), lesen Sie den Standardwert für den niedrigsten Rohwert ab, der größer ist als 0.
- Führen Sie die gleiche Prozedur solange für jeweils nächsthöhere Altersgruppe durch, bis der Rohwert 1 dem niedrigsten prinzipiell möglichen Standardwert entspricht. Bei der Wertpunktskala mit einem Wertebereich von 1 bis 19 sollte also der Rohwert 1 einem Wertpunkt entsprechen.

Sie haben jetzt die niedrigsten eindeutig interpretierbaren Standardwerte ermittelt, die Sie bei der Bewertung der Bodeneffekte zugrunde legen. Jetzt ist klar, dass mit *Test 2* keine weit unterdurchschnittlichen und mit *Test 3* nicht einmal leicht unterdurchschnittliche Fähigkeiten exakt gemessen werden können. Bei der Planung einer Untersuchung macht ein Blick in die Normtabelle von *Test 3* in wenigen Sekunden deutlich, dass der Einsatz des Verfahrens bei Verdacht auf eine Entwicklungsstörung völlig sinnlos ist – und zwar auch dann, wenn allen anderen Gütekriterien für eine hohe Testqualität sprechen.

Tabelle 2: Bewertung von Bodeneffekten (Renner, 2017)

Bei einem Testrohwert von 1 werden vergeben …	
$\leq$ WP 1, IQ 55, T 20, PR 0,1	kein Bodeneffekt
WP 2-4, IQ 56-70, T 21-30, PR 0,2-2	leichter Bodeneffekt
WP 5-7, IQ 71-85, T 31-40, PR 3-16	deutlicher Bodeneffekt
$\geq$ WP 8, IQ 86, T 41, PR 17	extremer Bodeneffekt

Für die Bewertung von Bodeneffekten gibt es keine allgemein anerkannten Regeln. Renner (2017) schlägt die in Tabelle 2 dargestellten Kriterien vor. Flanagan, Ortiz, Alfonso und Mascolo (2006) legen einen weniger strengen Maßstab an und bezeichnen Tests als frei von Bodeneffekten, wenn der Rohwert 1 einen Standardwert ergibt, der mindestens zwei Standardabweichungen unter dem Mittelwert liegt (z. B. 4 Wertpunkte, IQ-Wert von 70).

Auch wenn nur 0.1% aller wahren Testwerte unter einem Wertpunkt liegen, kann es bei (Unter-) Tests, die in die Kategorie »kein Bodeneffekt« fallen, zu Überforderungen kommen. Dies betrifft zum einen Kinder mit Leistungen im extrem unterdurchschnittlichen Bereich (z. B. Intelligenzdiagnostik bei Kindern mit mittelgradigen Intelligenzminderungen). Zum anderen kann sich aufgrund der üblichen Regeln zum Testabbruch (z. B. nach vier Fehlern in Folge) auch bei Kindern, die in den meisten Untertests ein oder zwei Items richtig lösen, ein ungünstiges Verhältnis zwischen Erfolgs- und Misserfolgserlebnissen ergeben, das die Testmotivation oder die Befindlichkeit des Kindes beeinträchtigen kann.

4.2 Bodeneffekte in Skalen und Gesamtwerten von mehrdimensionalen Testverfahren

Dieser Abschnitt beginnt mit einer erfreulichen Botschaft: Werden bei Tests, die mehrere Untertests beinhalten, Skalenwerte oder Gesamtwerte gebildet, werden sich Bodeneffekte weniger deutlich zeigen. Je mehr Untertests durchgeführt werden, umso wahrscheinlicher wird es, dass Testpersonen eine oder mehrere Aufgaben lösen. Umgekehrt wird es mit jedem zusätzlichen Untertests immer unwahrscheinlicher, dass Testpersonen gar keine Aufgaben korrekt bearbeiten. Dementsprechend werden Rohwerten von Null oder Eins niedrigere Standardwerte zugeordnet als in den einzelnen Untertests.

Diese allgemeine Tendenz schließt aber nicht aus, dass Gesamtwerte verzerrt werden, wenn Bodeneffekte in einzelnen Untertests vorliegen. Jeder einzelne Untertestwert, der durch einen Bodeneffekt höher ausfällt als die wahre Leistungsfähigkeit der untersuchten Person, wird zu einem etwas höheren Gesamtwert führen als angemessen. Eine eigenständige Überprüfung durch die Testanwender*innen ist daher sehr zu empfehlen, denn das Vorliegen starker Bodeneffekte in mehreren Untertests wird sich auch auf die Ebene der Skalen- und Gesamtwerte auswirken. Dabei kann man sich auf die Skalen beschränken, deren zugeordnete Untertests in bestimmten Altersgruppen Bodeneffekte aufweisen.

- Bestimmen Sie wie oben beschrieben den Testboden (bei Rohwert 1) für jeden Untertest, der in den Skalen-/Gesamtwert einfließt.
- Verrechnen Sie die entsprechenden Standardwerte gemäß der Vorgabe des Testmanuals.
- Lesen Sie für die so erhaltenen Werte die Normwerte der Skalen/des Gesamttests ab.

Tabelle 3: Ermittlung von Bodeneffekten bei Skalen und Gesamtwerten – Schritt 1

	Wertpunkte bei Testrohwert 1	
Skala / Untertest	Alter 2;9 Jahre	Alter 5;3 Jahre
Handlungsskala		
Puzzles	6	2
Zeichenmuster	6	1
Mosaike	7	1
Wertpunktsumme	19	4
Denkskala		
Kategorien	4	2
Situationen	2	1
Analogien	4	1
Wertpunktsumme	11	4
Gesamtsumme	30	8

Tabelle 4: Ermittlung von Bodeneffekten bei Skalen und Gesamtwerten – Schritt 2

Ergebniswert	IQ-Werte bei Testrohwerten von 1	
	2;9 Jahre	5;3 Jahre
Handlungsskala	77	≤ 55
Denkskala	59	≤ 55
Gesamt-IQ	66	≤ 55

Hierzu ein Beispiel (Tabelle 3) für zwei Altersgruppen in einem Intelligenztest (SON-R 2-8), der sechs Untertests umfasst, die zu zwei Skalen zusammengefasst werden. Außerdem kann ein Gesamtwert berechnet werden. Die Standardwerte der Untertests sind Wertpunkte. In dem Beispiel wurden mittels des PC-Auswertungsprogramms des SON-R 2-8 die Wertpunkte für den Rohwert 1 bei einem Kind im Alter von 2;9 Jahren ermittelt. Nach den Kriterien in Tabelle 2 finden sich zwei leichte (*Situationen*, *Analogien*) und vier deutliche Bodeneffekte. Im Alter von 5;4 Jahren bestehen noch zwei leichte Bodeneffekte.

Als Zwischenschritt zur Bestimmung der IQ-Werte werden im SON-R 2-8 die Wertpunkte für die Skalen und den Gesamtwert aufsummiert. In den Normtabellen kann dann abgelesen, welchem Standardwert (hier: IQ-Skala) die Wertpunktsumme entspricht. Dies Ergebnis ist in Tabelle 4 dargestellt.

Der allgemeine Testboden liegt in diesem Verfahren bei 55, niedrigere IQ-Werte können prinzipiell nicht erreicht werden. Bei den 2;9-Jährigen liegt der tatsächliche Testboden, ermittelt über Testrohwerte von 1 in allen Untertests, jedoch deutlich höher.

Warum sollte man diese umständliche Prozedur durchführen? Der Testboden müsste sich doch wie bei den Untertest leicht in den Normtabellen identifizieren lassen. Doch dies ist leider nicht der Fall. In Tabellen, die zur Umrechnung von Wertpunktsummen in Skalenwerte dienen, sind oft auch Werte aufgeführt, die in bestimmten Altersgruppen gar nicht vorkommen können. Schaut man sich beispielsweise die Normtabelle D.2 der Altersgruppe 3;0 bis 3;11 Jahre in der KABC-II an, liegt auf den ersten Blick die Vermutung nahe, dass der Fluid-Kristallin-Index (FKI)

einen Testboden bei einer Wertpunktsumme von 7 und dem dazugehörigen Standardwert (IQ-Wert) von 40 hat. Wertet man die KABC-II jedoch für ein 3;1-jähriges Kind aus, das in allen Untertests einen Rohwert von 0 erzielt hat, liegt die Wertpunktsumme bei 14 und ergibt einen FKI bei 47. Sinnvoll interpretierbar ist dieser Wert natürlich nicht. Geht man von eindeutig interpretierbaren Testrohwerten von 1 in allen Untertests aus, erhält man eine Wertpunktsumme von 25 und einen FKI von 57. Die in der Tabelle aufgeführten Werte unter einer Wertpunktsumme von 7 können definitiv nicht vorkommen, und Wertpunktsummen unter 25 sind nur möglich, wenn mindestens ein Untertest einen Rohwert von 0 aufweist.

Kasten 1: Bodeneffekte können zu Überschätzungen der Leistungen von Testpersonen führen

Bei einem Kind im Alter von 3;1 Jahren wurde die Kaufman Assessment Battery for Children – II (KABC-II; Melchers & Melchers, 2015) durchgeführt. Beim Subtest *Zahlen nachsprechen* liest die Testleiterin einen Standardwert von 6 Wertpunkten in der Normtabelle ab. In ihrem Befund interpretiert sie das Ergebnis so: »Das Ergebnis im Subtest *Zahlen nachsprechen* entspricht einer unterdurchschnittlichen Leistung«. Gemäß den üblichen Konventionen ist an dieser Aussage erst einmal nichts zu bemängeln. Ein genauer Blick in die Normtabelle zeigt jedoch, dass 6 Wertpunkte bei einem Testrohwert von 0 vergeben wurden.

Testrohwerte von 0 sind jedoch nicht eindeutig interpretierbar, wie eine Analogie schnell deutlich macht: Es sollen die Hochsprungleistungen von Kindern gemessen. Die Einstiegshöhe beträgt 40 cm. Was wissen wir über die Leistungsfähigkeit eines Kindes, dass diese Höhe nicht bewältigt? Wir wissen nichts weiter, als dass seine Hochsprungleistung irgendwo unter 40 cm liegt. Beim Hochspringen würde man jetzt die Latte tiefer legen, um ein genaueres Ergebnis zu erhalten. Die Möglichkeit, leichtere Aufgaben hinzuzufügen, wenn Kinder schon die Anfangsanforderungen nicht bewältigen können, haben Anwender*innen eines standardisierten Tests jedoch nicht.

Der wahre Testwert eines 3;1-jährigen Kindes beim *Zahlen nachsprechen* kann also durchaus 6 Wertpunkte betragen, aber auch beliebig niedriger liegen. Es gibt für diese Altersgruppe keine sehr einfachen Aufgaben, mit denen weit unterdurchschnittliche Leistungen gemessen werden können. Die angemessene verbale Interpretation des Testwerts wäre daher: »Das Ergebnis im Subtest *Zahlen nachsprechen* entspricht einer unterdurchschnittlichen bis weit unterdurchschnittlichen Leistung.« Eine genauere Aussage ist nicht möglich.

5 Testrohwert Null: Was ist zu beachten?

Ergänzend ist es noch sinnvoll, sich einen Überblick zu verschaffen, zu welchen Standardwerten ein Rohwert von 0 führt. Standardwerte, die auf Rohwerten von 0 basieren, sind nicht eindeutig interpretierbar (vgl. Kasten 1). Die Leistung des Kindes kann dem in der Normtabelle abgelesenen oder einem *beliebig niedrigeren* Standardwert entsprechen.

Suchen Sie dazu in der Normtabelle den Standardwert, der bei einem von Rohwert 0 vergeben wird. In Tabelle 1 finden Sie bei *Test 1* den Wert 1, bei *Test 2* den Wert 4 und bei *Test 3* den Wert 6.

Tabelle 5: Wie Bodeneffekte in Normtabellen meist dargestellt werden und wie sie dargestellt werden könnten

	Übliche Darstellung	*Transparente Darstellung*
Wertpunkte	Rohwert	Rohwert
1		0
2		0
3		0
4		0
5		0
6	0	0
7		
8	1	1
9		
10	2-3	2-3
11	4	4
12	5-6	5-6
13	7	7
14	8-9	8-9
15	10	10
16	11	11
17	12	12
18	13	13
19	14-20	14-20

In Befundberichten sollten diese Standardwerte *unbedingt* mit einem vorangestellten <= (kleiner gleich) angegeben werden. In nahezu allen mir bekannten Tests werden die Testanwender*innen dazu verführt, auf das <= zu verzichten, da dem Rohwert 0 in den Normtabellen genau ein Wert zugeordnet wird. Tabelle 5 zeigt am Beispiel von *Test 3* eine übliche Darstellung und eine alternative Darstellung, mit der dieses Phänomen leicht kenntlich gemacht werden könnte.

Vorgaben in den Testmanualen zur Verwertung von Untertests mit Rohwert 0 sind zu beachten. Gesamtwerte sollten in der Regel nicht bestimmt werden, wenn mehr als ein Untertest einen Rohwert von 0 aufweist. Es besteht immer die Gefahr, dass der Gesamtwert zumindest leicht überschätzt wird, sobald ein Testrohwert von 0 aufgetreten ist.

Trotz der genannten Interpretationsprobleme können Rohwerte von 0 je nach Fragestellung durchaus zu relevanten Ergebnissen führen. Hierzu ein Beispiel: Eine Untersuchung findet statt mit der Fragestellung, ob eine Beeinträchtigung im Sprachverständnis vorliegt. Das Kind löst kein Item korrekt, in der Normtabelle lässt sich für den Rohwert 0 ein Standardwert (IQ-Skala) von 64 ablesen. Auch wenn dieser Wert nicht eindeutig interpretierbar ist und auch niedriger liegen könnte, ist damit doch klar, dass das Kind nicht über ein altersgemäßes Sprachverständnis verfügt.

6 Zurück zum Diagnostik-Quiz

Bei allen drei Fragen lautet die richtige Antwort: »Keine Ahnung«. Präziser formuliert: Die Fragen lassen sich nicht beantworten, wenn man nicht weiß, ob die Testergebnisse durch Bodeneffekte beeinflusst wurden.

Bei Frage 1 kann es sein, dass *Wortschatz Aktiv* einen starken Bodeneffekt aufweist. Beruht der Standardwert von 80 vielleicht auf einem Testrohwert von 0? Bevor dies nicht überprüft wurde, kann die Diskrepanz zwischen den beiden Testwerten nicht sinnvoll interpretiert werden (vgl. Kasten 2).

Kasten 2: Bodeneffekte können inhaltlich bedeutungslose Profilunterschiede erzeugen oder inhaltlich bedeutungsvolle Unterschiede maskieren

In den Intelligence and Development Scales – Preschool (IDS-P) erzielt ein 3;3-jähriges Kind im Subtest *Aufmerksamkeit Selektiv* 2 Wertpunkte und im Untertest *Denken Bildlich* 6 Wertpunkte. Die Diskrepanz beträgt 4 Wertpunkte, entsprechend 1.3 Standardabweichungen. Die Prüfung auf statistische Signifikanz auf Basis der im Manual angegebenen Reliabilitätswerte ergibt, dass der kritische Wert von 3.6 (für $p < .01$) überschritten wird. Spricht dies dafür, dass die selektive Aufmerksamkeit des Kindes deutlich stärker beeinträchtigt ist als seine räumlich-konstruktiven Leistungen?

Der Standardwert von 6 in *Denken Bildlich* ist durch einen deutlichen Bodeneffekt bedingt (es sei angemerkt, dass die IDS-P zu den wenigen Verfahren gehören, die in den Normtabellen auf das Vorliegen von Bodeneffekten hinweisen). Das Kind hatte nur einen Testrohwert von 0 erreicht, der auch einem beliebig niedrigeren Standardwert entsprechen kann. Die Signifikanzprüfung kann also kein sinnvolles Ergebnis liefern.

Wenn das Kind in *Aufmerksamkeit Selektiv* ebenfalls 6 Wertpunkte erhalten hätte, wären die Ergebnisse beider Untertests identisch ausgefallen. Trotzdem könnte ein bedeutsamer Unterschied vorliegen. Vielleicht beträgt die Differenz tatsächlich 0, aber es können auch größere Differenzen vorliegen, die maskiert werden, weil in *Denken Bildlich* Werte unter 6 prinzipiell nicht gemessen werden können.

Auch Frage 2 kann erst beantwortet werden, wenn die Überprüfung auf einen Bodeneffekt erfolgt ist. Mit steigendem Alter werden Kinder mehr und mehr Aufgabe korrekt lösen, nach und nach verschwinden die Bodeneffekte. Wenn bei *Wortschatz Aktiv* im Alter von 4;1 Jahren ein extremer Bodeneffekt besteht, wissen wir nur, dass der wahre Wert des Kindes bei 80 oder beliebig niedriger liegt. Im Alter von 6;0 Jahren können dagegen deutlich niedrige Werte ermittelt werden. Die scheinbare Verschlechterung des Testwerts würde dann keinen Rückschluss auf die Entwicklung des Kindes erlauben, sondern schlicht und einfach einen Qualitätsmangel des eingesetzten Verfahrens widerspiegeln.

Die Antwort auf Frage 3 scheint völlig klar zu sein. Ein Prozentrang von 79 muss doch ein eindeutiger Hinweis auf eine altersgemäße Entwicklung sein. Allerdings nicht, wenn es sich um den Untertest *Wortflüssigkeit* des Kognitiven Entwicklungstest für das Kindergartenalter (KET-KID; Daseking & Petermann, 2009) handelt. In der Altersgruppe 3;0-3;5 Jahre wird dieses Ergebnis bei einem Testrohwert von 0 erzielt. 79% der Kinder konnten keine einzige Aufgabe lösen. Vielleicht entspricht die wahre Leistung des Kindes dem Prozentrang 79, vielleicht aber auch dem Prozentrang 0.1 oder irgendeinem Wert dazwischen. Hier gilt wiederum: Erst die Prüfung auf einen Bodeneffekt macht offenkundig, dass es sich um einen völlig nichtssagenden Testwert

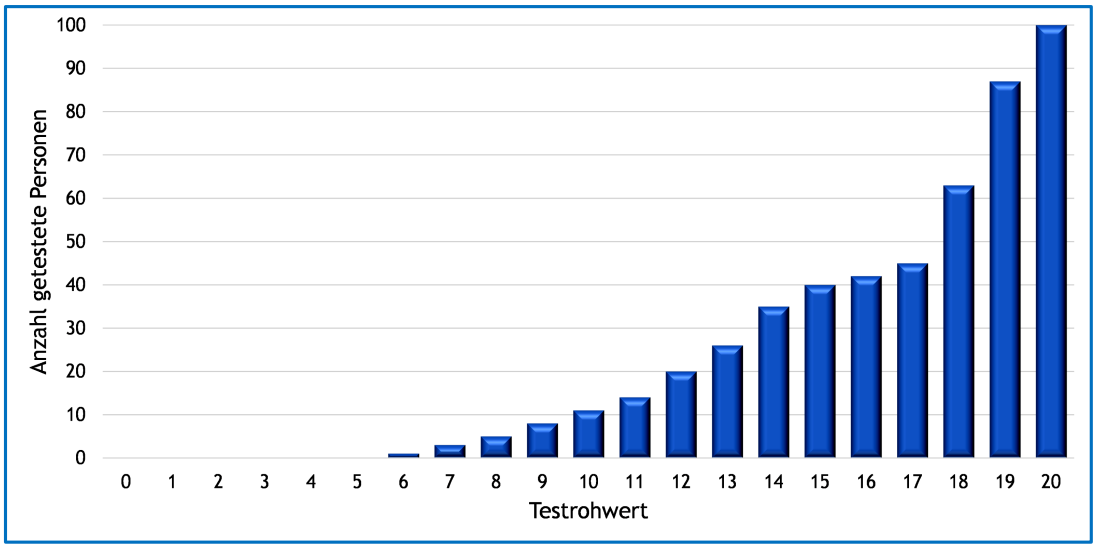

Abbildung 2: Beispiele für eine Verteilung von Testrohwerten mit einem Deckeneffekt

handelt und dass es prinzipiell unmöglich ist, durch Einsatz dieses Untertests etwas über die Leistungsfähigkeit von Kindern im unterdurchschnittlichen Bereich zu erfahren. In der Normtabelle hätte man das kenntlich machen können, wenn bei einem Testrohwert von 0 nicht der vermeintlich exakte Prozentrang 79, sondern ein Prozentrangbereich von 0 bis 79 eingetragen worden wäre. Im Manual des KET-KID ist das nicht der Fall, naive Testanwender*innen werden die Angabe PR = 79 in ihren Befundbericht übernehmen und so in vielen Fällen ein völlig falsches Bild von den Fähigkeiten der getesteten Kinder vermitteln.

7 Das andere Ende der Normtabelle: Deckeneffekte

In diesem Text war bisher immer von Bodeneffekten die Rede, obwohl man die gleichen Überlegungen auch für Testwerte im weit überdurchschnittlichen Bereich anstellen kann. In der sonderpädagogischen Diagnostik wird sich selten die Frage stellen, ob Schülerinnen und Schüler über eine weit oder extrem überdurchschnittliche Leistungsfähigkeit verfügen. In bestimmten Fällen – z. B. bei Schulleistungsproblemen, die auf einer kognitiven Unterforderung basieren – wird es dennoch wichtig sein, auch im oberen Leistungsbereich differenziert messen zu können.

Das Pendant zu Bodeneffekten sind Deckeneffekte. Ein Deckeneffekt liegt vor, wenn ein Test zu wenig schwierige Items enthält. In den Normtabellen können dann weit überdurchschnittliche Werte nicht abgelesen werden. Auch hier gilt: Das Fähigkeitsniveau der Kinder kann nicht adäquat bewertet werden, es können Pseudo-Testprofile entstehen (wenn z. B. ein Untertest eine Testdecke von 13 Wertpunkten, ein anderer dagegen eine Testdecke von 19 Wertpunkten hat) und Verlaufsmessungen können verfälscht werden (Testwerte sinken scheinbar, weil bei älteren Testpersonen höhere Testwerte gar nicht mehr erzielt werden können). Abbildung 2 zeigt, wie eine Rohwertverteilung aussehen kann, die zu einem Deckeneffekt führt. Die daraus abgeleiteten Standardwerte sind in Tabelle 6 zu finden. Mehr als 14 Wertpunkte können nicht erreicht. Dieser Wert ist nicht eindeutig interpretierbar: Die wahre Leistungsfähigkeit der Person kann 14 Wertpunkten entsprechen, aber auch höher liegen, was sich mit diesem Test aber

Tabelle 6: Beispiel für eine Normtabelle mit einem Deckeneffekt

Wertpunkte	Rohwert
1	0-6
2	7
3	8
4	9
5	10
6	11-12
7	13
8	13-14
9	15
10	17-18
11	
12	19
13	
14	20
15	
16	
17	
18	
19	

nicht sicher feststellen lässt. Dies sollte in Befundberichten deutlich gemacht werden durch ein dem Testwert vorangestelltes $\geq$ (größer gleich).

8 Jetzt kann geübt werden!

In den Tabellen 7 und 8 finden sich teilweise Boden- und/oder Deckeneffekte. Tabelle 7 ist aufgebaut wie in den bisherigen Beispielen, in Tabelle 8 sind diesmal die Rohwerte in der linken Spalte eingetragen und die Standardwerte (Prozentränge) in den folgenden Spalten. In Testmanualen sind Normtabellen häufig wie in Tabelle 7 gestaltet, aber auch die Variante in Tabelle 8 ist gebräuchlich. Wo sich in diesen Tabellen Boden- und Deckeneffekte finden, wird am Ende dieses Beitrags aufgelöst.

9 Konsequenzen für die Auswahl und den Einsatz von Testverfahren

- Die wichtigste Konsequenz ist, dass Testanwender*innen in der Lage sein müssen, Boden- und ggf. auch Deckeneffekte der von ihnen eingesetzten Verfahren zu erkennen. Das hier dargestellte Verfahren kann mit ein wenig Übung zu einer schnell durchführbaren Routine werden.

Tabelle 7: Übungstabellen zur Ermittlung von Boden- und Deckeneffekten

	Test A	Test B	Test C
Wertpunkte	Rohwert	Rohwert	Rohwert
1		0-4	
2		5-6	
3		7-9	0
4		10-11	1
5		12-13	2
6	0	14-16	3
7	1	17-18	4
8	2	19-21	5
9	3	22-24	6-7
10	4	25-29	8-9
11	5-6	30-31	10-11
12	7	32-33	12
13	8	34-35	13
14	9	36	14
15	10	37	
16	11	38	
17	12		
18	13		
19	14		

Tabelle 8: Übungstabellen zur Ermittlung von Boden- und Deckeneffekten

	Test D (20 Items)	Test E (12 Items)	Test F (16 Items)
Rohwert	Prozentrang	Prozentrang	Prozentrang
0	0.0	0.0	19.7
1	0.1	0.1	45.3
2	0.5	0.9	56.1
3	1.4	2.4	65.5
4	3.2	5.1	74.0
5	6.3	8.4	81.3
6	10.9	12.3	86.9
7	16.9	16.9	91.4
8	23.8	22.5	95.1
9	32.4	29.8	97.6
10	43.8	38.8	99.1
11	55.9	49.8	99.8
12	66.7	78.0	100.0
13	76.1		100.0
14	83.8		100.0
15	89.8		100.0
16	94.2		100.0
17	97.1		
18	98.7		
19	99.5		
20	99.9		

- Verschaffen Sie sich einen systematischen Überblick über Bodeneffekte in allen Testverfahren, die Sie regelmäßig einsetzen. Besonders wichtig ist die Suche nach Bodeneffekten bei Testverfahren für das Vorschulalter. Die Entwicklung von Items, die bei sehr jungen Kindern mit Entwicklungsstörungen eingesetzt werden können, stellt eine große Herausforderung dar, die nicht in allen gängigen Testverfahren bewältigt wurde.
- Wenn Sie vor einer Untersuchung erste Informationen oder Eindrücke über das Leistungsniveau der zu testenden Kinder haben, wählen Sie Verfahren aus, die den Testpersonen mit hoher Wahrscheinlichkeit einige Erfolgserlebnisse erlauben. Für das getestete Kind ist es andernfalls eine ausgesprochen frustrierende Erfahrung, wenn es reihenweise mit Aufgaben konfrontiert wird, die es nicht lösen kann, und Testleiter*innen werden nur erfahren, was das Kind nicht kann.
- Schaffen Sie keine Tests mit Bodeneffekten an, wenn Sie überwiegend Kinder untersuchen, deren Leistungen im unterdurchschnittlichen oder weit bis extrem unterdurchschnittlichen Bereich liegen. Fragen Sie vor Anschaffung eines Testverfahrens bei den Testverlagen, ob Sie ein Testmanual, das die Normtabellen enthält, zur Ansicht erhalten können und beurteilen Sie Bodeneffekte nach dem oben dargestellten Vorgehen.
- Schaffen Sie keine Testverfahren an, die Ihnen eine Überprüfung von Bodeneffekten verwehren. Bei einer rein computerisierten Testauswertung, bei der Normtabellen nicht mehr eingesehen werden können, ist eine eigenverantwortliche Bewertung von Bodeneffekten nicht möglich. Man kann sich leider nicht darauf verlassen, dass Bodeneffekte in den Testmanualen umfassend und transparent dargestellt und bewertet werden.
- Machen Sie in sonderpädagogischen Gutachten oder anderen Dokumentationen transparent, ob die Testergebnisse von Bodeneffekten beeinflusst sein könnten. Machen Sie auch dadurch bedingte Grenzen der Interpretierbarkeit von Testwerten deutlich. Leser*innen von Befundberichten haben keinerlei Chance, die mit Bodeneffekten verbundenen Fallstricke zu erkennen, wenn sie keinen Zugang zu den Normtabellen haben. Insbesondere machen Sie immer kenntlich, welche Standardwerte auf einem nicht eindeutig interpretierbaren Testrohwert von 0 oder – bei Deckeneffekten – auf einem nicht eindeutig interpretierbaren maximalen Testwert beruhen.
- Denken Sie auch bei der Lektüre von Fremdbefunden daran, dass die Testergebnisse von Bodeneffekten beeinflusst sein könnten. Wenn Sie dies nicht ausschließen können und Zugriff auf die Manuale der eingesetzten Verfahren haben, nehmen Sie Ihre eigene Bewertung vor.
- Verzichten Sie auf die Interpretation von Testprofilen, wenn sich bei einem Testrohwert von 0 vermeintliche Leistungsstärken in Untertests oder Skalen zeigen, die stark von Bodeneffekten betroffen sind.
- Bei Wiederholungsmessungen interpretieren Sie Leistungsverschlechterungen nicht, ehe Sie die Frage geklärt haben, ob die höheren Testwerte bei einer früheren Untersuchung auf Bodeneffekte zurückgehen könnten.
- Wenn Sie doch einmal erst während einer Untersuchung feststellen, dass ein Kind mit den einfachsten Items überfordert ist (so was kann und wird passieren), beharren Sie nicht darauf, eine erkennbar unergiebige und das Kind belastende Testung weiterzuführen. Beenden Sie die Testung und wechseln Sie, wenn möglich, zu einem Verfahren mit leichte-

ren Anfangsanforderungen oder gewinnen Sie erste diagnostische Eindrücke durch ein gemeinsames Spiel oder informelle Verfahren.

10 Lösung zu den Übungstabellen 7 und 8

10.1 Tabelle 7:

- *Test A* ist durch einen deutlichen Bodeneffekt gekennzeichnet. Dem Rohwert 1 sind 7 Wertpunkte zugeordnet. Selbst wenn keine einzige Aufgabe gelöst wurde, werden noch 6 Wertpunkte vergeben. Die Testdecke liegt bei 19 Wertpunkten, es können also weit überdurchschnittliche Leistungen differenziert gemessen werden.
- In *Test B* können maximal 16 Wertpunkte erreicht werden, es liegt also ein Deckeneffekt vor.
- In *Test C* haben die Testentwickler*innen sowohl sehr leichte als auch sehr schwierige Items vergessen. Bei einem Rohwert von 1 werden 4 Wertpunkte vergeben, das ist ein leichter Bodeneffekt. Die Testdecke liegt bei 14 Wertpunkten, damit können überdurchschnittliche Leistungen nicht differenziert erfasst werden.

10.2 Tabelle 8:

- *Test D* zeigt weder einen Boden- noch einen Deckeneffekt. Der Rohwert 1 entspricht einem Prozentrang von 0.1. Der maximal mögliche Prozentrang beträgt 99.9.
- *Test E* zeigt keinen Bodeneffekt (beim Rohwert 1 beträgt der Prozentrang 0.1), aber einen deutlichen Deckeneffekt. Maximal kann ein Prozentrang von 78 erreicht werden. In der Rohwertverteilung, auf der diese Tabelle basiert, hatten 44% aller Kinder in der Normstichprobe alle Aufgaben richtig bearbeitet. Fast die Hälfte der Stichprobe erhielt somit ein und dasselbe quantitative Testergebnis.
- *Test F* zeigt einen extremen Bodeneffekt. Schon bei einer einzigen gelösten Aufgabe wird ein Prozentrang von 45.3 erzielt.

Literatur

Bracken, B. A. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology, 26*, 155–166.

Flanagan, D. P., Ortiz, S. O., Alfonso, V. C. & Mascolo, J. T. (2006). *The Achievement Test Desk Reference* (2nd. ed.). Hoboken, NJ: Wiley.

Grob, A., Reimann, G., Gut, J. & Frischknecht, M. C. (2013). *IDS-P. Intelligence and Development Scales – Preschool*. Bern: Huber.

Lienert, G. A. & Raatz, U. (1994). *Testaufbau und Testanalyse* (5., völlig neubearb. und erw. Aufl). Weinheim: Beltz PVU.

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.

Renner, G. (2017). Chancen und Fallstricke der Intelligenzdiagnostik bei kognitiven Entwicklungsstörungen. In V. Mall, F. Voigt & N. Jung (Hrsg.), *Entwicklungsstörungen und chronische Erkrankungen. Diagnose, Behandlungsplanung und Familienbegleitung* (Aktuelle Fragen der Sozialpädiatrie, Bd. 2, S. 76–91). Lübeck: Schmidt-Römhild.

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

Wechsler, D. (2017). *WISC-V. Wechsler Intelligence Scale for Children – Fifth Edition. Deutsche Bearbeitung Franz Petermann*. Frankfurt: Pearson Assessment.

Prof. Dr. Gerolf Renner Pädagogische Hochschule Ludwigsburg, Fakultät für Sonderpädagogik, Förderschwerpunkt körperliche und motorische Entwicklung https://orcid.org/0000-0003-4345-3619