Frank Herold

Was Sie schon immer über die statistischen Hintergründe von Verfahren zur Entwicklungseinschätzung wissen wollten

Falls Sie mit Beobachtungsbögen oder Screenings beziehungsweise Kurzverfahren zur Ressourcen- und Risikoeinschätzung von Kindern im Elementarbereich zu tun haben, dann sind Sie hier richtig: Um solche Verfahren besser beurteilen zu können, finden Sie hier statistische Grundlagen und Begriffe wie *Prozentrang* einfach dargestellt. Am Schluss des Textes finden Sie Quellen, in denen Expertinnen und Experten u. a. gängige Verfahren vorstellen.

1. Merkmale und Merkmalsausprägungen

Merkmale sind hier vor allem Eigenschaften von Kindern, z. B. Selbstständigkeit. Jedes Merkmal kann verschiedene Ausprägungen haben, z. B. große, mittlere oder geringe *Selbstständigkeit*.

2. Skala

Eine Skala ist der Maßstab, mit dem man Merkmalsausprägungen misst. Die wichtigsten Skalenarten sind – aufgezählt nach zunehmender Aussagekraft: Nominalskala, Ordinalskala, Intervallskala und Verhältnisskala (vgl. Clauß & Ebner 1978, S. 25–28, S. 29 Tab. 1.1). Im Folgenden geht es hauptsächlich um die Ordinalskala und die Intervallskala.

3. Intervallskala (Einheitenskala)

Weil die Intervallskala auch bei der Ordinalskala eine Rolle spielt, stelle ich sie Ihnen zuerst vor.
Eine bekannte Intervallskala ist die Temperaturskala in Grad Celsius. Ihre Merkmalsausprägungen oder Messwerte bilden eine natürliche Reihenfolge: 1 Grad, 2 Grad, 3 Grad usw. Typisch: Die Abstände zwischen benachbarten Messwerten sind gleich groß und sachlich begründet (vgl. Macha 2012). So ist der Abstand zwischen minus 15 Grad und 4 Grad genauso groß wie zwischen 14 und 33 Grad (vgl. Universität Zürich 2016). Intervallskalen haben aber keinen absoluten Nullpunkt. Den haben nur Verhältnisskalen, z. B. Null Gramm Gewicht bei der Gewichtsmessung. Das ermöglicht Verhältnisvergleiche: 20 Gramm sind tatsächlich doppelt so schwer wie 10 Gramm (vgl. Clauß, Finze & Partzsch 2011, S. 11). Dagegen haben Intervallskalen höchstens einen willkürlich festgelegten Nullpunkt. Bei der Celsius-Skala ist es der Gefrierpunkt des Wassers. Das verbietet Verhältnisvergleiche. Deshalb sind 20 Grad Celsius nicht doppelt so warm wie 10 Grad (vgl. Dieterich 1973, S. 73–74).
Die IQ-Skala (Intelligenzquotient) ist – wie viele Testskalen – als Intervallskala konstruiert. Man nimmt deshalb gleiche Abstände zwischen den IQ-Werten an (vgl. Clauß, Finze & Partzsch 2011, S. 11). Auch hier fehlt ein absoluter Nullpunkt. Deshalb ist ein Kind mit einem IQ von 140 nicht doppelt so intelligent wie ein Kind mit einem IQ von 70 (vgl. Dieterich 1973, S. 74). Das gilt auch für die EQ-Skala (Entwicklungsquotient).

4. Ordinalskala (Rangskala)

4.1. Schulnoten

Die Schulnotenskala ist eine Ordinalskala (vgl. Universität Zürich 2016). Ihre Ziffern 1, 2, 3, 4, 5 und 6 sind nur die traditionellen Abkürzungen für die sprachlichen Noten *sehr gut*, *gut*, *befriedigend*, *ausreichend*, *mangelhaft*, *ungenügend*. Man könnte sie auch durch die Bewertungsstufen A, B, C, D, E und F ersetzen (vgl. Eikenbusch & Leuders 2004, S. 26). Sie bilden wie bei einer Intervallskala eine natürliche Reihenfolge: Eine 1 ist besser als eine 2, eine 2 ist besser als eine 3 usw. Anders als bei einer Intervallskala sind die Abstände zwischen den Noten jedoch nicht festgelegt und können sehr unterschiedlich sein (vgl. Clauß, Finze & Partzsch 2011, S. 11). Man kann also nicht sagen, um wie viel eine 1 besser ist als eine 2 usw. Problematisch ist daher die Berechnung von Durchschnittsnoten als arithmetisches Mittel: „Man berechnet es als die Summe aller Werte, geteilt durch die Anzahl der Werte“ (Eikenbusch & Leuders 2004, S. 200).
Beispiel: Ein Kind bekam in fünf Aufsätzen die Noten 2, 2, 3, 3, 5.
Arithmetisches Mittel: (2 + 2 + 3 + 3 + 5) geteilt durch 5 = Durchschnittsnote 3.

Ein arithmetisches Mittel darf man allerdings nur berechnen, wenn wenigstens eine Intervallskala vorliegt (vgl. Wirtz & Nachtigall 2008, S. 73). Das kann man – streng genommen – bei der Schulnotenskala aber nicht annehmen (vgl. Eikenbusch & Leuders 2004, S. 23). Deshalb ist eine Durchschnittsnote nur ein ungenauer Anhaltspunkt für die Qualität von Schulleistungen (vgl. ebd., S. 26) und eine „Notlösung" für die Praxis (vgl. ebd., S. 23).

4.2. Ratingskala (Schätzskala)

Ratingskalen sind Antwortskalen für mündliche oder schriftliche Fragen oder Aussagen (vgl. Döring & Bortz 2016, S. 245).

Beispiel: Das Kind hält sich an die Gruppenregeln
1 = nie, 2 = selten, 3 = manchmal, 4 = oft, 5 = immer
Skalen mit sehr wenigen Stufen – z. B. drei oder vier – liefern unscharfe Urteile, die höchstens Ordinalskalenniveau besitzen (vgl. Matell & Jacoby 1971, zitiert nach Döring & Bortz 2016, S. 249). Auch grundsätzlich müssten Ratingskalen als Ordinalskalen betrachtet werden: Man kann nicht zwingend annehmen, dass Personen die Abstände als gleich wahrnehmen, wenn sie ihre Antworten ankreuzen. Trotzdem werden Ratingskalen oft als Intervallskalen verwendet, wenn es plausibel erscheint (vgl. Universität Zürich 2016). In Tests sind Ratingskalen aber meist nur Ordinalskalen (vgl. Bühner 2011, S. 220), ebenso in Beobachtungsbögen und Screenings. Berechnet man dennoch aus den angekreuzten Ziffern mehrerer Skalen ein arithmetisches Mittel als Durchschnittsantwort, dann ist das zumeist nur ein ungenauer Hilfswert – siehe Durchschnittsnote.

5. Grundgesamtheit und repräsentative Stichprobe

Man kann die Leistung eines Kindes mit einem Kriterium vergleichen. Ein Kriterium ist eine festgelegte Merkmalsausprägung, z. B. ein relativ altersunabhängiger Schritt beim Erwerb einer Zweitsprache (vgl. Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache 2013, S. 15).

Hier geht es aber um den Vergleich mit einer Bezugsgruppe: Man erfasst ein Merkmal eines Kindes im Vergleich zu anderen Kindern. Will man z. B. wissen, ob der Wortschatz eines Kindes altersgemäß ist, dann müsste man alle Kinder seines Alters in Deutschland untersuchen – die *Grundgesamtheit*. Das ist aber zu aufwendig. Deshalb wählt man eine Teilgruppe aus der Grundgesamtheit aus – die *Stichprobe* (Eich- oder Normstichprobe). Die Rückschlüsse von der Stichprobe auf die Grundgesamtheit sind aber immer fehlerbehaftet und nur mehr oder weniger genau (vgl. Clauß & Ebner 1978, S. 165). Damit sie überhaupt gültig sind, muss die Stichprobe repräsentativ sein für die Grundgesamtheit (vgl. ebd., S. 18) im Sinne eines verkleinerten, möglichst wahrheitsgetreuen Spiegelbildes (vgl. ebd., S. 179).

Das geht mit einer *Zufallsauswahl*: Jedes Mitglied der Grundgesamtheit hat die gleiche Chance, Teil der Stichprobe zu werden (vgl. ebd., S. 180). Bei einer ausreichend großen Zufallsstichprobe kann man annehmen, dass sie die Grundgesamtheit annähernd repräsentativ abbildet (vgl. Wirtz & Nachtigall 2008, S. 24). Doch selbst sie vertritt die Grundgesamtheit nur. Deshalb kann man nicht sicher sagen, welche Stellung ein Kind in der Grundgesamtheit hat (vgl. Eikenbusch & Leuders 2004, S. 38).

Die Erhebung einer repräsentativen Normstichprobe scheitert oft am Aufwand und an den Kosten. Zudem müssen Eltern um die Mitwirkung ihrer Kinder gebeten werden (vgl. Macha & Petermann 2013, S. 186). Besonders schwierig ist das z. B. bei Familien mit Migrationshintergrund und sprachlichen oder kulturellen Hürden (vgl. Schölmerich et al. 2008, zitiert nach Macha & Petermann 2013, S. 186).

6. Alter und Größe der Normstichprobe

Zum Alter: Heute sind die Lebensumstände von Kindern im Vergleich zu früher u. a. geprägt durch ein verändertes Spielverhalten, eine andere Erziehung und bessere frühkindliche Bildung. So hat sich im Durchschnitt die kognitive Leistungsfähigkeit erhöht und z.B. die motorische Leistungsfähigkeit von Kindergartenkindern eher verringert (vgl. Macha & Petermann 2017, S. 143). Allerdings beeinflusst der Zeitgeist erst Vorschulkinder langsam zunehmend (vgl. Largo, Kundu & Thun-Hohenstein 1993, zitiert nach Macha & Petermann 2013, S. 188). Deshalb empfiehlt sich schon für das Vorschulalter eine Neunormierung von Verfahren nach 10 Jahren (vgl. Macha & Petermann 2013, S. 188). Entwicklungsbezogene Altersangaben von Beobachtungsverfahren und Screenings sollten nur 10 bis 15 Jahre alt sein (vgl. Macha & Petermann 2017, S. 143). Doch auch ältere Entwicklungsangaben können noch gültig sein.
Zur Größe: Die Stichprobe verteilt sich auf alle normierten Altersgruppen eines Verfahrens, z.B. 4,0 – 4,5 Jahre usw. Sie sollte pro Altersgruppe mindestens 150 Kinder umfassen (vgl. Wyschkon & Esser 2015, S. 176, Abb. 1).
Viele Tests haben eine zu kleine oder zu alte Normstichprobe (vgl. Kreis Borken 2016, S. 4). Das gilt auch für einen Teil der Beobachtungsbögen und Screenings.

7. Verteilung

Erkennen Sie, was diese Tabelle darstellt?

Noten	**1**	**2**	**3**	**4**	**5**	**6**
Häufigkeiten	3	16	7	6	3	1

Richtig: Sie ist der Klassen- oder Notenspiegel zu einer Klassenarbeit von 36 Kindern und damit die Häufigkeitsverteilung der erzielten Schulnoten innerhalb der Klasse.
„Verteilungen geben also an, wie häufig bestimmte Werte vorkommen" (Eikenbusch & Leuders 2004, S. 31). Da die Noten 1–3 überwiegen, ist die Verteilung asymmetrisch bzw. linksschief. Bei Überwiegen der Noten 4–6 wäre sie rechtsschief.
Bei Verfahren zur Entwicklungseinschätzung ergibt sich eine Verteilung so: Man führt mit den Kindern der Normstichprobe die bei der Testentwicklung erstellten Aufgaben durch und erhält zunächst Merkmalsausprägungen als *Rohwerte*. Das sind in der Regel Punktwerte, z.B. die Anzahl bewältigter Aufgaben (vgl. Macha 2007).
Rohwerte allein sagen hier aber nichts aus. Erreicht ein Kind z.B. 40 Punkte, dann kann man dies erst bewerten, wenn man die Punktzahlen der anderen Kinder kennt. Deshalb rechnet man Rohwerte in Vergleichsnormen um (Normierung) und erstellt Normtabellen.

8 Normalverteilung und Standardnormen (Normierung)

Verteilungen von Rohwerten können sich dem Ideal einer völlig symmetrischen Glockenform annähern, der *Normalverteilung* (vgl. Kleber 1978, S. 46–47). Eine Abbildung finden Sie im Internet z.B. bei Macha (2007). Ist die Annäherung ausreichend, dann kann man Rohwerte in verschiedene *Standardnormen* umrechnen (vgl. Lienert 1961, S. 320 ff.). Mit ihnen kann man angeben, wie weit das Ergebnis einer Person vom Mittelwert der Normstichprobe nach unten oder oben abweicht (vgl. Stauche 2008, S. 2), also z.B. unterdurchschnittlich, durchschnittlich oder überdurchschnittlich ist.

Beispiel IQ-Skala:
Arithmetisches Mittel: IQ 100.
Breiter Durchschnittsbereich: IQ 85 bis 115. Hier liegen die mittleren ca. 68 Prozent der Messwerte. Darunter liegen die ca. 16 Prozent der unterdurchschnittlichen Werte, darüber die ca. 16 Prozent der überdurchschnittlichen Werte (vgl. Kleber 1978, S. 50, 51 Abb. 21). Die anderen Standardskalen sind im Prinzip ebenso aufgebaut. Sie unterscheiden sich im arithmetischen Mittel und in der Breite des Durchschnittsbereiches (Streuung).

Beispiel T-Wert-Skala, benannt nach dem Psychologen Terman:
Arithmetisches Mittel: T-Wert 50.
Breiter Durchschnittsbereich: T-Wert 40 bis 60. Auch hier liegen die mittleren 68 Prozent, die schwächsten 16 Prozent darunter, die besten 16 Prozent darüber (vgl. Macha 2007, Abb.).

9. Beliebige Verteilung und Prozentrangnormen (Normierung)

Da die Rohwerte von Entwicklungsmerkmalen in Beobachtungsbögen und Screenings oft nicht normalverteilt sind, ist der Prozentrang (PR) hier wichtig:
„Standardnormen können nur aus normalverteilten Rohwerten errechnet werden, Prozentrangnormen hingegen aus jeder beliebigen Verteilung" (Lienert 1961, S. 321).
Der Prozentrang zeigt, welchen Rang eine Person innerhalb einer Gruppe hat.

Beispiel: Erzielt eine Person einen Prozentrang von 45, dann weisen 45 Prozent der Personen der Normstichprobe eine niedrigere oder höchstens gleiche Ausprägung des gemessenen Merkmals auf (vgl. Stauche 2008, S. 5), 55 Prozent sind besser.
Dabei ist eine Eigenart der Prozentrangskala zu berücksichtigen.

Beispiel: Bei einem Verfahren werden nicht normalverteilte Rohwerte beziehungsweise Punkte in Prozentränge umgerechnet. Von möglichen 50 Punkten erreicht:

Anna	10 Punkte und Prozentrang 10,
Ben	15 Punkte und Prozentrang 15,
Clara	40 Punkte und Prozentrang 80,
Doris	45 Punkte und Prozentrang 95.

Der Leistungsunterschied zwischen Anna und Ben einerseits und Clara und Doris andererseits beträgt jeweils 5 Punkte. Dennoch ist der Prozentrangunterschied zwischen Clara und Doris mit 15 Prozenträngen deutlich größer als zwischen Anna und Ben mit 5 Prozenträngen (vgl. Bühner 2011, S. 264–265).
Schlussfolgerung: Prozentrangunterschiede kann man nicht direkt in Leistungsunterschiede übertragen (vgl. ebd., S. 265). Die Prozentrangskala ist eben eine Ordinalskala, bei der man nur sagen kann, „(...) ob jemand besser war als eine andere Person, aber nicht, um wie viel" (ebd., S. 265).
Deshalb darf man mit Prozenträngen nicht normal rechnen und z.B. kein arithmetisches Mittel aus Prozenträngen berechnen (vgl. Kleber 1978, S. 70).
Der *Mittelwert* der Verteilung ist hier der *Median*. Er teilt die Verteilung in zwei Hälften: 50 Prozent aller Werte sind kleiner oder gleich dem Median, 50 Prozent sind größer oder gleich (vgl. Wirtz & Nachtigall 2008, S. 72). Der Median liegt damit genau in der Mitte und entspricht dem Prozentrang 50, dem IQ 100 und dem T-Wert 50.

Enger Durchschnittsbereich: Prozentrang 25 bis 75 (vgl. Kleber 1978, S. 70), das entspricht IQ 90 bis 110; hier liegen die mittleren 50 Prozent der Messwerte (vgl. ebd., S. 51, S. 52 Abb. 22, 23). Dem *breiten Durchschnittsbereich* der mittleren 68 Prozent ent-

spricht der Prozentrangbereich 16 bis 84 (vgl. Macha 2007, Abb.). Diesen Durchschnittsbereich finden Sie am häufigsten.

10. Hauptgütekriterien

„Es ist grundsätzlich davon auszugehen, dass jede Testung mit einem Messfehler behaftet und somit das Messergebnis ungenau ist" (Macha 2012, *Standardmessfehler*).

Damit Messfehler möglichst klein bleiben, müssen Verfahren bestimmte Gütekriterien erfüllen. So soll ein guter Test objektiv, zuverlässig und gültig sein (vgl. Lienert 1961, S. 12–15). Das gilt auch für Screenings und Beobachtungsbögen, wird aber nicht immer berücksichtigt (vgl. z.B. Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache 2013, S. 15).

Objektivität: Sie ist gegeben, wenn das Ergebnis eines Verfahrens unabhängig davon ist, welche Fachkraft es durchführt, auswertet und interpretiert (vgl. Macha 2012). Das soll eine *Standardisierung* sicherstellen: Man schreibt Durchführung, Auswertung und Interpretation genau vor, z.B. Materialien, Situation, sprachliche Anweisungen und Regeln zur Bewertung und Auswertung (vgl. ebd.).

Bei vielen Aufgaben von Beobachtungsbögen und Screenings – z.B. zu sozialen Fähigkeiten – geht das aber nur schwer, da sie „(...) eine längerfristige Beobachtung des Kindes in verschiedenen Alltagssituationen (...)" erfordern (Flender 2005, S. 38). Es ist dann „(...) nicht immer möglich, genaue Kriterien zu Durchführung und Auswertung der Aufgaben zu benennen" (ebd., S. 120). Das begünstigt u.a. *Beurteilungsfehler*. Beispielsweise wird ein übergewichtiges Kind eher als motorisch ungeschickt eingeschätzt (vgl. Macha 2012, *Halo-Effekt)*.

Insgesamt haben Beobachtungsbögen und Screenings gegenüber vollstandardisierten Tests – die eine Momentaufnahme bieten – einen Vorteil und einen Nachteil: Einerseits können sie einen Zeitverlauf berücksichtigen, was ihre Beurteilungsgrundlage verbreitert; andererseits ist ihre Objektivität geringer (vgl. Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache 2013, S. 16–17).

Zuverlässigkeit (Reliabilität): Ein Verfahren ist zuverlässig, wenn es dasjenige Merkmal, welches es tatsächlich misst, genau misst (vgl. Lienert 1961, S. 13).
Häufig bestimmt man die Testwiederholungs-Zuverlässigkeit *(Re-Test-Zuverlässigkeit):* Sie gibt an, inwieweit die Wiederholung eines Verfahrens zu ähnlichen Ergebnissen führt (vgl. Kreis Borken 2016, S. 4). Beispielsweise schätzt jeweils eine Fachkraft ein bestimmtes Kind im Abstand von wenigen Wochen zweimal mit einem Beobachtungsbogen ein. Danach vergleicht man beide Ergebnisse.

Gültigkeit (Validität): Ein Verfahren ist gültig, wenn es das Merkmal, das es messen soll, auch zuverlässig misst. Es muss sich also zur genauen Untersuchung des Merkmals eignen (vgl. Lienert 1961, S. 14). Fachkräfte schätzen Kinder z.B. mit einem Beobachtungsbogen ein. Unabhängig davon untersuchen Experten dieselben Kinder mit einem Entwicklungstest. Danach vergleicht man die Ergebnisse *(Kriteriumsvalidität). Sensitivität* und *Spezifität*: Diese Begriffe stehen für die *Gültigkeit* von Screenings, die auffällige Kinder von normalen Kindern unterscheiden (vgl. Flender 2005, S.16), z.B. durch einen Grenzwert bzw. Cut-off-Wert (vgl. Macha 2012). Die Sensitivität besagt, inwieweit auffällige Kinder tatsächlich als auffällig erkannt werden. Die Spezifität besagt, inwieweit unauffällige Kinder nicht als auffällig eingestuft werden. Hierzu vergleicht man das Screening-Ergebnis mit dem Ergebnis eines anderen Verfahrens (vgl. Flender 2005, S. 16–17).

Ein Screening hat durch seine Kürze in der Regel eine niedrigere Zuverlässigkeit und Gültigkeit als ein Test (vgl. Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache 2013, S. 16).

Zu Gütekriterien wie Objektivität, Zuverlässigkeit, Gültigkeit usw. gibt es Kennwerte und ungefähre Beurteilungsrichtlinien (Fisseni 1997, zitiert nach Bühner 2011, S. 81, Abb. 2.12).

Informelle Verfahren haben keine Normstichprobe und keine geprüften Gütekriterien. Bei ihnen ist die kritische Würdigung der Ergebnisse durch einen Austausch im Team und mit den Eltern besonders wichtig, um Beurteilungsfehler klein zu halten.

11. Einige Möglichkeiten, sich über Verfahren zu informieren

Die Bedeutung eines Verfahrens ist nicht leicht zu beurteilen. Beispielsweise sagt ein Kennwert zur Gültigkeit allein wenig aus. Wichtig ist auch, was das Verfahren in welcher Situation zur Lösung einer Fragestellung beitragen kann (vgl. Fisseni 1997, zitiert nach Bühner 2011, S. 81).
Informationen zu einem Verfahren stehen im jeweiligen Handbuch (Manual).
Außerdem können Expertinnen und Experten Ihnen wichtige Hinweise geben:
Für Kindertageseinrichtungen geeignete Verfahren finden Sie gut dargestellt z. B. im „Praxishandbuch Kindergarten" (Petermann & Wiedebusch 2017, Kapitel 4, 6–9).
„Qualitätsmerkmale für Sprachstandsverfahren im Elementarbereich" bietet das Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache (2013).

Es gibt Testbesprechungen bzw. Testrezensionen, die Sie mit diesen Begriffen und dem Kurz- oder Langnamen eines Verfahrens im Internet finden können.
Es gibt auch ein Verzeichnis von Fachzeitschriften mit Rezensionen. Sie finden es auf folgender Seite des Leibniz-Zentrums für Psychologische Information und Dokumentation (ZPID): https://www.psyndex.de/pub/tests/verz_teil5.pdf

12. Schlussbemerkung

Beobachtungsbögen, Screenings und von diagnostisch geschulten Fachkräften durchgeführte Entwicklungstests sind trotz Einschränkungen wichtige Bausteine einer Entwicklungseinschätzung, wenn man ihre Stärken und Schwächen berücksichtigt.

Frank Herold
Herr Herold ist Diplom-Sozialpädagoge und Diplom-Heilpädagoge und arbeitete bis 2014 in der Beratungsstelle für Eltern, Jugendliche und Kinder der Stadt Hamm.

Literaturverzeichnis

Bühner, M. (2011): Einführung in die Test- und Fragebogenkonstruktion. München, Harlow, Amsterdam, Madrid, Boston, San Francisco, Don Mills, Mexico City, Sydney: Pearson, 3., aktualisierte und erweiterte Auflage.
Clauß, G., Ebner, H. (1978): Grundlagen der Statistik. Für Psychologen, Pädagogen und Soziologen. Berlin: Volk und Wissen, 6. Auflage.
Clauß, G., Finze, F.-R., Partzsch, L. (2011): Grundlagen der Statistik. Für Soziologen, Pädagogen, Psychologen und Mediziner. Frankfurt am Main: Harri Deutsch, 6., korrigierte Auflage.
Dieterich, R. (1973): Psychodiagnostik. Grundlagen und Probleme. München: Ernst Reinhardt.
Döring, N., Bortz, J. (2016): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Berlin, Heidelberg: Springer, 5. vollständig überarbeitete, aktualisierte und erweiterte Auflage.
Eikenbusch, G., Leuders, T. (Hrsg.) (2004): Lehrer-Kursbuch Statistik. Alles über Daten und Zahlen im Schulalltag. Berlin: Cornelsen.
Flender, J. (2005): Früherkennung von Entwicklungsstörungen durch Erzieherinnen: Überprüfung der Gütekriterien des Dortmunder Entwicklungsscreening für den Kindergarten (DESK 3-6). Dortmund: Dissertation.
Kleber, Ed. W. (1978): Lehrbuch der sonderpädagogischen Diagnostik. Berlin: Carl Marhold, 3., völlig neubearbeitete und erweiterte Auflage.
Kreis Borken, Regionale Schulberatungsstelle (Hrsg.) (2016): Psychologisches Testen. Informationen für Eltern und Lehrkräfte. Broschüre. Borken.
Lienert, G. A. (1961): Testaufbau und Testanalyse. Weinheim: Beltz.
Macha, T. (2007): Standardwerte - Entwicklungsdiagnostik.
http://entwicklungsdiagnostik.de/standardwerte.html (Zugriff am 27.02.2020).
Macha, T. (2012): Glossar. http://entwicklungsdiagnostik.de/glossar.html (Zugriff am 24.10.2019).
Macha, T., Petermann, F. (2013): Objektivität von Entwicklungstests. Zur Standardisierung der entwicklungsdiagnostischen Befunderhebung. In: Diagnostica, 59. Jg., Nr. 4, S. 183-191.
Macha, T., Petermann, F. (2017): Entwicklungsdiagnostische Verfahren:
Ressourcen- und Risikoerkennung. In: Petermann, F., Wiedebusch, S. (Hrsg.) (2017): Praxishandbuch Kindergarten. Entwicklung von Kindern verstehen und fördern. Göttingen: Hogrefe, 1. Auflage, S. 133-152.
Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache (Hrsg.) (2013): Qualitätsmerkmale für Sprachstandsverfahren im Elementarbereich. Ein Bewertungsrahmen für fundierte Sprachdiagnostik in der Kita. Köln.
https://www.mercator-institut-sprachfoerderung.de/fileadmin/user_upload/Mercator-Institut_Qualitaetsmerkmale_Sprachdiagnostik_Kita_Web_03.pdf (Zugriff am 01.04.2020)
Petermann, F., Wiedebusch, S. (Hrsg.) (2017): Praxishandbuch Kindergarten.
Entwicklung von Kindern verstehen und fördern. Göttingen: Hogrefe, 1. Auflage.
Stauche, H. (2008): Normwerte der Testdiagnostik - ihre Berechnungen und Umrechnungen ineinander. Institut für Erziehungswissenschaft. Universität Jena.
https://www.db-thueringen.de/servlets/MCRFileNodeServlet/dbt_derivate_00036055/normwerte.pdf (Zugriff am 28.01.2020)
Universität Zürich (2016): Methodenberatung. Skalenniveau.
https://www.methodenberatung.uzh.ch/de/skalenniveau.html (Zugriff am 17.02.2020)
Wirtz, M., Nachtigall, C. (2008): Deskriptive Statistik. Statistische Methoden für Psychologen Teil 1. Weinheim, München: Juventa, 5., überarbeitete Auflage.
Wyschkon, A., Esser, G. (2015): Testleiterfehler und Beurteilung von Testnormen: Empfehlungen für Testentwickler und -anwender. In: Esser, G., Hasselhorn, M., Schneider, W. (Hrsg.) (2015): Diagnostik im Vorschulalter. Göttingen, Bern, Wien, Paris, Oxford, Prag, Toronto, Boston, Amsterdam, Kopenhagen, Stockholm, Florenz, Helsinki: Hogrefe, S. 165 - 176.