Potsdamer Zentrum für empirische Inklusionsforschung (ZEIF), 2019, Nr. 3

Schulpraktische Erhebungs- und Auswertungsmöglichkeiten von Einzelfalldaten

Markus Spilles und Tobias Hagen
Universität zu Köln

Zusammenfassung: Die Evaluation schulischer Fördermethoden stellt für Lehrerinnen und Lehrer eine zentrale Aufgabe dar. Zur Überprüfung individueller Fördererfolge bietet die Erhebung und Auswertung von Einzelfalldaten eine praktische Möglichkeit. Im aktuellen Beitrag werden verschiedene Optionen zur Erhebung von Einzelfalldaten benannt. Im Schwerpunkt geht es um Verfahren zur Effektstärkenberechnung und Signifikanzprüfung, die konkret und anschaulich beschrieben werden, um eine selbstständige Auswertung zu ermöglichen.

Schlagwörter: Einzelfallforschung, Datenerhebung, Effektstärkenberechnung, Mann-Whitney U-Test, Randomisierungstest

Abstract: The evaluation of school-based interventions is a key task for teachers. With regard to individual student development, the collection of single-case data seems to be promising for practical use. The present article gives a short overview of how to collect single-case data. Possibilities for data analysis are presented in greater detail.

Keywords: single-case research, data collection, effect size calculation, Mann-Whitney U-test, randomization test

Herr Maier ist Klassenlehrer einer dritten Grundschulklasse. Seinen Unterricht gestaltet er sehr gewissenhaft, sodass die meisten seiner Schülerinnen und Schüler hiervon angemessen profitieren. Bei David zeigen sich allerdings dennoch stärkere Probleme im Bereich seiner Lesekompetenzentwicklung. Janina verhält sich trotz des guten Classroom Managements von Herrn Maier im Unterricht oft sehr störend. Für beide Kinder hat er daher gemeinsam mit der Sonderpädagogin an seiner Schule Unterstützungsmöglichkeiten erarbeitet. Gerne würde er nun überprüfen, ob diese Maßnahmen auch erfolgreich sind.

Einleitung

Im Zuge der laufenden Umstrukturierung der deutschen Bildungslandschaft hin zu einem inklusiven Schulsystem werden diverse Ansätze zum Umgang mit einer zunehmend heterogenen Schülerschaft kritisch und kontrovers erörtert. Zwei Rahmenkonzepte, die häufig als Lösungen diskutiert werden, sind das Response-to-Intervention-Modell (Huber & Grosche, 2012) und der Schoolwide Positive Behavior Support (Sugai & Horner, 2006). Hierbei handelt es sich um mehrstufige Förderansätze, die Möglichkeiten aufzeigen, wie mit unterschiedlichen Unterstützungsbedarfen von Schülerinnen und Schülern umgegangen werden kann. Zentraler Bestandteil beider Konzeptionen ist u. a. der Einsatz von engmaschiger Lern- und Verhaltensverlaufsdiagnostik, um die Effektivität von Fördermaßnahmen zu überprüfen, individuelle Lernerfolge sichtbar zu machen und sozialen Vergleichsprozessen entgegenzuwirken. Im vorliegenden Beitrag werden vor diesem Hintergrund ausgewählte Möglichkeiten zur schulpraktischen Datenerhebung und -auswertung vorgestellt.

Möglichkeiten zur Datenerhebung

Zur Erhebung von Verlaufsdaten werden Instrumente benötigt, die zur formativen Evaluation von Fördermaßnahmen geeignet sind. Diagnostische Methoden, die diese Funktion erfüllen, sind bspw. Curriculumbasierte Messverfahren (Deno, 2003) und die Verhaltensverlaufsdiagnostik (Casale, Hennemann, Huber & Grosche, 2015).

Curriculumbasierte Messverfahren (CBM)

CBM fokussieren spezifische Inhaltsbereiche, wie bspw. die Leseflüssigkeit, das Leseverständnis oder mathematische Grundoperationen. Sie zeichnen sich dadurch aus, dass die Aufgabenschwierigkeit bei jeder Messung gleichbleibt. Die einzelnen Items werden verändert, prüfen aber prinzipiell gleiche Inhaltsbereiche mit identischen Schwierigkeitsmerkmalen. Zudem beanspruchen CBM wenig Zeit (1-5 Minuten) und lassen sich somit gut im Schulalltag implementieren. Konkrete Beispiele sind die Lernfortschrittsdiagnostik Lesen (Walter, 2009), die Verlaufsdiagnostik sinnerfassenden Lesens (Walter, 2013) und die Lernverlaufsdiagnostik – Mathematik für zweite bis vierte Klassen (Strathmann & Klauer, 2012).

Direkte Verhaltensbeurteilung (DVB)

Die DVB ist eine Form der Verhaltensverlaufsdiagnostik, bei der konkrete Verhaltensweisen

unmittelbar im Anschluss an eine Situation auf einer Rating-Skala beurteilt. Demnach handelt es sich um eine Kombination aus systematischer direkter Verhaltensbeobachtung und Verhaltensbeurteilung (Casale et al., 2015). Bei der Umsetzung wird zunächst eine Situation ausgewählt, in der ein Verhalten als besonders problematisch wahrgenommen wird. Die Häufigkeit, die Qualität oder der prozentuale Anteil des definierten Problemverhaltens im Beobachtungszeitraum wird direkt im Anschluss von der Lehrkraft eingeschätzt (Huber & Rietz, 2015). Wichtig ist, dass Verhaltensweisen ausgewählt werden, die gut beobachtbar sind. Außerdem müssen die Beurteilungen immer von der gleichen Person vorgenommen werden. Insbesondere störendes Verhalten (bspw. unerlaubtes aufstehen oder durch die Klasse rufen) und Probleme im lernbezogenen Verhalten (bspw. unvollständige Aufgabenbearbeitung oder unselbstständiges Arbeiten) können über die DVB zuverlässig eingeschätzt werden (Casale et al., 2015; Huber & Rietz, 2015).

Auch systematische Verhaltensbeobachtungen können zur Erhebung von Verlaufsdaten genutzt werden und liefern noch präzisere Ergebnisse (Casale et al., 2015). Allerdings ist die Umsetzung im Vergleich zur DVB wesentlich aufwändiger und im Schulalltag kaum umsetzbar. Aus diesem Grund wird hierauf nicht weiter eingegangen.

Praxisbeispiel

Zur Förderung der Lesekompetenzentwicklung planen Herr Maier und die Sonderpädagogin den Einsatz der tutoriellen Methode Lautlesetandems (LLT; Rosebrock, Nix, Rieckmann & Gold, 2011) im Klassenverband. Die beiden interessiert vor allem, ob David im Bereich seiner Leseflüssigkeit profitiert. Um dies zu überprüfen, entscheiden sie sich für den Einsatz der Lernfortschrittsdiagnostik Lesen (LDL; Walter, 2009). Die LLT sollen dreimal pro Woche während der Lesezeit für 15-20 Minuten umgesetzt werden. Jedes Mal im Anschluss wird mit David die LDL durchgeführt. Hierbei handelt es sich um einen 1-Minuten-Lesetest, der zur Ermittlung der richtig gelesenen Wörter pro Minute genutzt wird. Bevor die LLT in der Klasse eingeführt werden, wird die LDL zunächst über einen Zeitraum von zwei Wochen immer direkt nach der freien Lesezeit mit David durchgeführt. Erst in der dritten Woche beginnt die spezifische Leseförderung. Die ersten Messzeitpunkte werden zur Erfassung der Ausgangslage genutzt und dienen bei der Datenauswertung (s. u.) als Referenz.

Janina hat große Probleme während Einzelarbeitsphasen still an ihrem Platz zu sitzen und ruhig und konzentriert zu arbeiten. Aus diesem Grund soll zukünftig während Einzelarbeitsphasen ebenfalls dreimal pro Woche das KlasseKinderSpiel (KKS; Hillenbrand & Pütz, 2008) für zehn Minuten gespielt werden. Hierüber sollen störende Verhaltensweisen reduziert und das Lern- und Arbeitsverhalten verbessert werden. Herr Maier und seine Kollegin entscheiden sich dafür, die folgenden Verhaltensweisen immer direkt im Anschluss an die Spielphase mittels DVB für Janina einzuschätzen: 1) ruft in die Klasse, 2) bleibt nicht auf ihrem Platz sitzen, 3) macht Lärm, 4) führt unangemessene Seitengespräche. Hierzu

nutzen sie eine sechsstufige Rating-Skala: 0 (nie), 1 (selten), 2 (manchmal), 3 (oft), 4 (sehr oft), 5 (immer). Auch in diesem Fall beginnt Herr Maier bereits mehrere Tage vor Einführung des KKS mit der DVB. Er beurteilt Janinas Verhalten immer direkt im Anschluss an die Situationen, in denen das KKS später gespielt werden soll. Die Beurteilungen zu den vier Verhaltensweisen addiert er jeweils auf. Somit ergibt sich pro Messzeitpunkt ein Wert zwischen 0 und 20, der als Indikator für Janinas Störverhalten in diesen Unterrichtssituationen genutzt werden kann.

Möglichkeiten zur Datenauswertung

Die Grundlage für die Beurteilung von Fördererfolgen ist der Vergleich von unterschiedlichen Phasen. Hierzu wird, wie im Praxisbeispiel beschrieben, zunächst eine Grundrate (oder Baseline) erhoben, in der noch keine spezifische Förderung stattfindet, sondern bspw. der herkömmliche Unterricht. Die Wirkung einer anschließend eingesetzten Maßnahme kann anhand des Vergleichs der Daten aus Grundrate und Interventionsphase beurteilt werden[1]. Zur Auswertung von Einzelfalldaten werden im Folgenden drei Möglichkeiten vorgestellt: 1) die visuelle Inspektion, 2) die Berechnung von Effektstärken und 3) die Signifikanzprüfung.

Visuelle Inspektion

Der erste Schritt bei der Auswertung von Einzelfalldaten ist deren Visualisierung. In einem Verlaufsdiagramm markiert die y-Achse mögliche Messwerte (z. B. korrekt gelesene Wörter pro Minute) und die x-Achse die Messzeitpunkte. Zur Überprüfung von Interventionserfolgen wird empfohlen, die Visualisierung parallel zur Erhebung durchzuführen und besondere Ereignisse festzuhalten, wie Konflikte während der Pause oder Schwierigkeiten bei der Datenerhebung. Ebenfalls wird im Diagramm gekennzeichnet, ab wann eine Intervention eingesetzt oder verändert wurde, damit ein Phasenvergleich vorgenommen werden kann. Um die Wirksamkeit anhand der Verlaufsdarstellung zu beurteilen, können im Wesentlichen zwei Indikatoren betrachtet werden (Brunstein & Julius, 2014):

- Niveauunterschiede zwischen den Phasen
- Trends in beiden Phasen

Wurde z. B. eine spezifische Verhaltensförderung umgesetzt, kann ggf. anhand der visuellen Inspektion erkannt werden, ob das Verhalten während der Interventionsphase in etwa positiver beurteilt wurde, als während der Grundratenerhebung. Als Hilfestellung kann z. B. der Mittelwert der jeweiligen Phase als horizontale Linie in die Verlaufsdarstellung eingezeichnet werden. Außerdem sollte darauf geachtet werden, wie sich die Datentrends während der Grundrate und Interventionsphase entwickeln und ob sich Veränderungen andeuten. Zur Skizzierung von Trendlinien wird weiter unten ein Beispiel gegeben. Weiterhin können Ausreißer (auffallend hohe oder niedrige Werte) identifiziert und evtl. auf besondere Ereignisse zurückgeführt werden. Vorteilhaft an der

[1] Ein Überblick zu verschiedenen Versuchsplänen der Einzelfallforschung findet sich bei Jain & Spieß (2012).

visuellen Inspektion ist, dass offensichtliche Effekte schnell erkannt werden können. Bei weniger eindeutigen Datenverläufen ist eine Einschätzung alleine über visuelle Indikatoren jedoch nicht mehr valide (Börnert-Ringleb, Bosch & Wilbert, 2018). Für eine genauere Einschätzung bietet sich die Berechnung von Effektstärken an.

Effektstärken

In der Forschungsliteratur werden zahlreiche Möglichkeiten zur Berechnung von Effektstärken im Rahmen von Einzelfallauswertungen berichtet, die verschiedene Vor- und Nachteile aufweisen (Alresheed, Hott & Bano, 2013). Im aktuellen Kapitel werden vier ausgewählte und leicht zu berechnende Maße vorgestellt. Zur Veranschaulichung der jeweiligen Rechenvorgänge wird das folgende Beispiel betrachtet:

Baseline (A): 2, 3, 4, 4, 3

Intervention (B): 4, 5, 7, 6, 8

Alle Effektstärken werden ausgehend von einem erhofften Werteanstieg (bspw. Anstieg von lernbezogenem Verhalten oder von Lesekompetenzen) von Grundrate zu Intervention erläutert, berechnet und interpretiert. Bei erhofften Werteabstiegen, wie im Falle von Janinas Störverhalten, müssen die Berechnungen genau umgekehrt vorgenommen werden.

Percentage of data points exceeding the median (PEM). Der PEM (Ma, 2006) ermöglicht einen Vergleich der Daten aus Baseline und Intervention, indem diejenigen Werte während der Förderung identifiziert werden, die über dem Median[2] der Baseline liegen. Der PEM ist damit eher

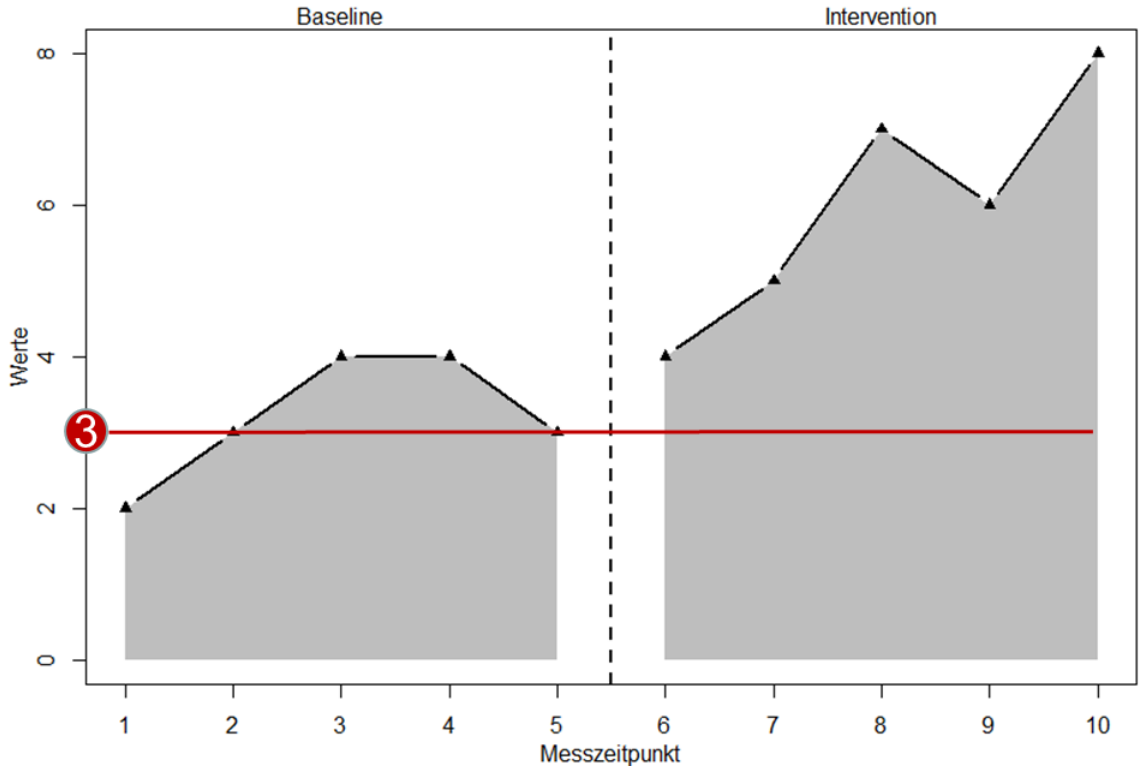

Abbildung 1. Veranschaulichung zum PEM

[2] Wenn Messwerte der Größe nach aufgereiht werden, findet sich der Median auf dem mittleren Rangplatz.

unempfindlich gegenüber Schwankungen in den Datenverläufen und bietet eine gute erste Orientierung für die Beurteilung von Einzelfallergebnissen. Zur Kalkulation wird in der Verlaufsdarstellung eine horizontale Linie ausgehend vom Median der Baseline in die Interventionsphase gezogen. Anschließend werden alle Punkte der Interventionsphase, die über dieser Linie liegen, gezählt und durch die Anzahl sämtlicher Punkte der Interventionsphase geteilt. Der PEM kann somit einen Wert zwischen 0 bis 1 bzw. umgerechnet zwischen 0 % bis 100 % annehmen, wobei ein Ergebnis ab 70 % auf einen moderaten Effekt und ein Ergebnis ab 90 % auf einen starken Effekt hindeutet (Alresheed et al., 2013).

Den Median der Baseline im Datenbeispiel markiert die Zahl 3 (Abbildung 1). Über diesem Wert liegen alle Werte der Interventionsphase (4, 5, 7, 6, 8). Der PEM beträgt in diesem Fall also 5/5 = 1 bzw. 100 % und weist auf einen starken Effekt der Intervention hin.

Percentage of all non-overlapping data (PAND). Der PAND (Parker, Hagan-Burke & Vannest, 2007) gibt den Anteil aller nicht überlappender Datenpunkte zwischen Baseline und Intervention an (Brunstein & Julius, 2014) und ist ebenfalls relativ robust gegenüber Ausreißern (Alresheed et al., 2013). Zur Berechnung wird mit dem Ziel der Überlappungsfreiheit der hierfür erforderliche minimale Anteil an Datenpunkten aus beiden Phasen entfernt (Brunstein & Julius, 2014). Anschließend wird die Anzahl der entfernten Punkte durch die Gesamtzahl aller Messzeitpunkte dividiert und von 1 subtrahiert. Der PAND kann also einen Wert von .50 bis 1 bzw. umgerechnet von 50 % bis 100 % annehmen. Ein Ergebnis im Bereich von 70 % bis 90 % deutet dabei

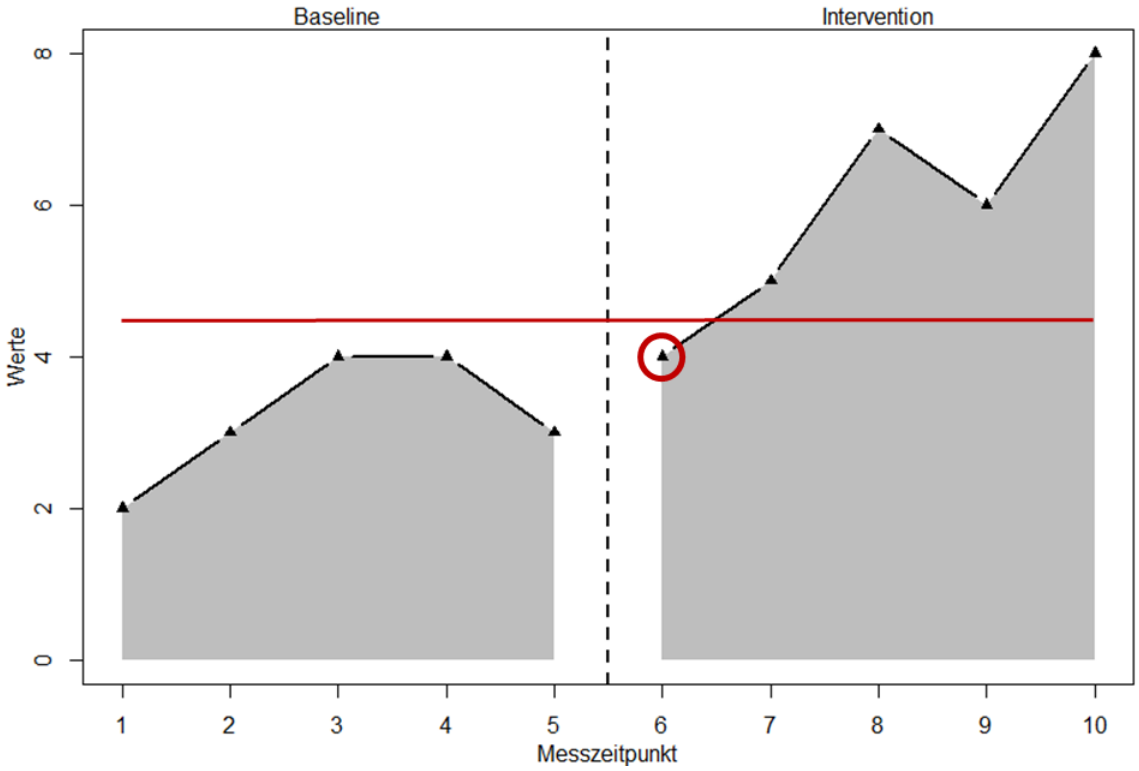

Abbildung 2. Veranschaulichung zum PAND

auf einen mittelstarken Effekt, ein Ergebnis über 90 % auf einen starken Effekt hin (ebd.).

Für unser Beispiel scheint es ausreichend, lediglich den ersten Datenpunkt (4) in der Interventionsphase zu entfernen, damit zwischen beiden Phasen keine Daten mehr überlappen[3] (Abbildung 2). Für den PAND ergibt sich somit ein Wert von 1 - 1/10 = .90 bzw. 90 % (mittelstarker bis starker Effekt).

Intervention

Baseline	4	5	7	6	8
2	+	+	+	+	+
3	+	+	+	+	+
4	0	+	+	+	+
4	0	+	+	+	+
3	+	+	+	+	+

Abbildung 3. Veranschaulichung zum NAP

Nonoverlap of all pairs (NAP). Der NAP (Parker & Vannest, 2009) stellt einen paarweisen Vergleich zwischen allen Datenpunkten aus Baseline und Interventionsphase an und errechnet sich wie folgt:

1) Bestimmung aller möglichen Vergleichspaare (v) der Messungen der Baselinephase (b) und der Interventionsphase (i): v = b*i.
2) Bestimmung aller Paare, die einen positiven (p) oder keinen (k) Trend im zeitlichen Verlauf aufweisen.
3) $NAP = \frac{p + \frac{1}{2}k}{v}$

Für den NAP ergeben sich somit für tendenziell ansteigende Datenverläufe Werte im Bereich von .50 bis 1 bzw. umgerechnet von 50 % bis 100 %. Mittlere Effekte finden sich zwischen 66 % und 92 %, starke Effekte ab 93 % (Parker & Vannest, 2009).

Bezogen auf das obige Beispiel ergeben sich insgesamt v = b*i = 5*5 = 25 Vergleichspaare (Abbildung 3). Hierbei lassen sich p = 23 positive („+“) und k = 2 neutrale („0“) Trends identifizieren. Die Berechnung des NAP ergibt somit einen Wert von $\frac{23 + \frac{1}{2}2}{25}$ = .96 bzw. 96 %, der auf einen starken Effekt der Förderung hindeutet. Die Berechnung kann praktischerweise auch über die Internetplattform *www.singlecasereasearch.org* vorgenommen werden.

Percentage of data exceeding a median trend (PEM-T). Der PEM-T (Wolery, Busick, Reichow & Barton, 2010) bezieht im Gegensatz zu allen bisher beschriebenen Maßen den Trend der Baseline als kritischen Faktor zur Beurteilung der Entwicklung während der Förderung mit in die Berechnung ein. Zur Bestimmung des PEM-T wird der Trend, der sich in der Baseline ergibt (ähnlich wie beim PEM), als Linie eingezeichnet und in die Interventionsphase verlängert. Im Falle eines erhofften Anstiegs werden anschließend alle Datenpunkte in der Interventionsphase, die über

[3] Alternativ hätten zur Herstellung der Überlappungsfreiheit auch der dritte (4) und vierte (4) Datenpunkt entfernt werden können. In diesem Falle wäre allerdings nicht der minimale Anteil an Datenpunkten entfernt worden.

dieser Linie liegen, gezählt und durch sämtliche Punkte der Interventionsphase geteilt. Wie der PEM kann auch der PEM-T Werte von 0 bis 1 bzw. umgerechnet von 0 % bis 100 % annehmen, wobei sich ein effektives Treatment ab einem Wert von 70 % und ein sehr effektives Treatment ab einem Wert von 90 % abzeichnet (ebd.).

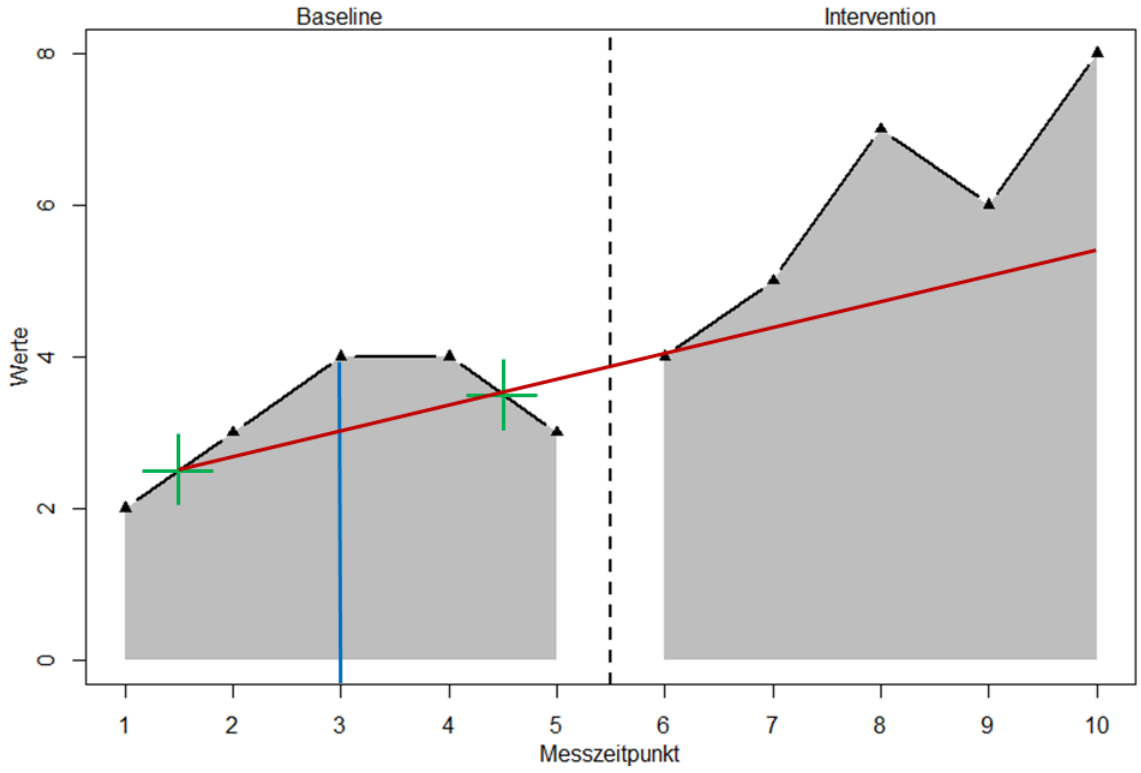

Abbildung 4. Veranschaulichung zur „split middle line" Technik

Die Kalkulation des Baseline-Trends kann dabei bspw. anhand der „split middle line" Technik nach White und Haring (1980) erfolgen (Abbildung 4), die sehr einfach umzusetzen ist. Hierfür wird die Baseline zunächst in zwei symmetrische Phasen unterteilt (blaue Linie). Bei einer ungeraden Anzahl an Datenpunkten liegt diese Linie genau auf dem mittleren Datenpunkt der Baseline, bei einer geraden Anzahl an Datenpunkten zwischen den beiden mittleren Datenpunkten. Anschließend wird die jeweilige Mitte der beiden Baselinephasen bestimmt (vertikale grüne Linien). Bei einer ungeraden Anzahl an Datenpunkten in der Baseline wird dabei der mittlere Punkt, durch den die blaue Linie verläuft, nicht berücksichtigt. Die jeweilige Phasenmitte befindet sich in der Mitte der verbleibenden Punkte. Nun wird der Median jeder Baselinephase eingezeichnet (horizontale grüne Linien – im Beispiel entspricht der Median dem Mittelwert). Die Schnittpunkte der grünen Linien markieren jetzt zwei Punkte, durch die die Trendlinie der Baseline verläuft.

Der Datenverlauf in der Baseline unseres Beispiels weist offensichtlich einen gewissen positiven Trend auf. Würde man diesen Trend in die grafische Darstellung der Daten einzeichnen und in die Interventionsphase verlängern, würden bis auf den ersten Wert der zweiten Phase (4) alle weiteren Punkte über dem Trend liegen. Für das Beispiel ergibt sich also ein Wert von 4/5 = .80

oder 80 %, der für einen mittleren Effekt des Treatments spricht.

Bei Vannest, Davis und Parker (2013) werden neben der Beschreibung der „split middle line“ Technik, die hier auch basierend auf den Ausführungen in diesem Buch dargestellt wurde, noch weitere Berechnungsmöglichkeiten für Trendlinien skizziert.

Signifikanzprüfung

Neben der Berechnung von Effektstärken können erhobene Daten auch auf Signifikanz hin geprüft werden. Gemeint ist hiermit, dass untersucht wird, ob der Daten An- oder Abstieg zwischen zwei Phasen mit einer hohen Wahrscheinlichkeit auf eine eingesetzte Intervention zurückgeführt werden kann, bzw. ob die Wahrscheinlichkeit, dass der Unterschied zwischen zwei Phasen rein zufällig entstanden ist, sehr gering ist. Konventionell wird in den Humanwissenschaften ein Ergebnis als signifikant bezeichnet, wenn die Wahrscheinlichkeit für letztere Aussage unter einem Niveau von 5 % liegt ($p < .05$).

Im Folgenden werden zwei Möglichkeiten vorgestellt, Signifikanztests bei Einzelfalldaten durchzuführen. Am praktikabelsten für die Schulpraxis ist sicherlich der Mann-Whitney U-Test, der im Kontext des NAP auch auf *www.singlecaseresearch.org* berechnet werden kann. Bei ohnehin tendenziell ansteigenden Datenverläufen erscheint der Randomisierungstest sinnvoller.

Mann-Whitney U-Test. Beim Mann-Whitney U-Test werden die Daten aus den zu vergleichenden Phasen A (Baseline) und B (Interventionsphase) in eine Reihenfolge vom niedrigsten bis zum höchsten Wert gebracht. Anschließend wird im Falle eines erhofften Werteanstiegs überprüft, ob die Werte der B-Phase im Schnitt höhere Rangplätze belegen, als die der A-Phase. Die praktische Umsetzung ist recht simpel und wird in Abbildung 5 skizziert.

Nachdem die Daten beider Phasen in die entsprechende Reihenfolge gebracht wurden, werden die Rangplätze vergeben. In unserem Beispiel belegen zufällig alle Datenpunkte aus Phase A (mit Ausnahme eines Datenpunkts aus Phase B) die niedrigsten sechs Ränge. Die höchsten vier Ränge werden ausschließlich an Messwerte aus Phase B vergeben. Weiterhin teilen sich einige Messzeitpunkte aufgrund gleicher Werte den gleichen, jeweils durchschnittlichen Rangplatz. Anschließend werden die Rangsummen der beiden Phasen berechnet. Mit Hilfe der Rangsumme von Phase A und der Anzahl der Messzeitpunkte beider Phasen werden nun drei Kennwerte (Rangplatzüberschreitungen: U, erwarteter U-Wert unter der Nullhypothese: μ_U, Streuung der U-Werte: σ_U) ermittelt, aus denen der *z*-Wert berechnet werden kann. Liegt dieser über einem kritischen Wert von 1.96 (bzw. im Falle eines erhofften Werteabstiegs unter -1.96), wird das Ergebnis als signifikant bezeichnet[4]. Bei Interesse können weitere Ausführungen zum

[4] Es sei angemerkt, dass bei n < 20 Messzeitpunkten pro Phase die Signifikanzprüfung mit Hilfe des *z*-Werts eigentlich nicht ausreichend exakt ist, da keine Normalverteilung der U-Werte mehr gegeben ist. Hier müsste der *p*-Wert theoretisch über die exakte U-Statistik ermittelt werden. Entsprechende Tabellen finden sich bei Rasch, Friese, Hofmann und Naumann (2014a).

Mann-Whitney U-Test bei Rasch, Friese, Hofmann und Naumann (2014b) nachgelesen werden. Das Datenbeispiel kann gerne auch auf *www.singlecaseresearch.org* nachgerechnet werden.

Tabelle 1: Signifikanzprüfung mit Hilfe des Mann-Whitney U-Tests

Mzp	Phase	Wert	Ränge (A)	Ränge (B)
1	A	2	1	
2	A	3	2,5	
5	A	3	2,5	
3	A	4	5	
4	A	4	5	
6	B	4		5
7	B	5		7
9	B	6		8
8	B	7		9
10	B	8		10
		Rangsummen:	16	39

$n(A) = 5$ (Anzahl der Mzp in Phase A)

$n(B) = 5$ (Anzahl der Mzp in Phase B)

$R(A) = 16$ (Rangsumme von Phase A)

$$U = n(A)n(B) + \frac{n(A)(n(A)+1)}{2} - R(A) = 24$$

$$\mu_U = \frac{n(A)n(B)}{2} = 12.5$$

$$\sigma_U = \sqrt{\frac{n(A)n(B)(n(A)+n(B)+1)}{12}} \approx 4.79$$

Signifikanzprüfung:

$$z = \frac{U - \mu_U}{\sigma_U} \approx 2.40 > \mathbf{1.96} \Rightarrow p < .05$$

Randomisierungstest. Kritisch anzumerken am Mann-Whitney U-Test ist, dass bei ohnehin tendenziell ansteigenden Verläufen sehr leicht signifikante Ergebnisse entstehen können. Bei der Erhebung der Leseflüssigkeit eines Grundschulkindes neigen die Daten einer nachfolgenden Phase allein aufgrund der natürlichen Entwicklung dazu, die Werte der ersten Phase zu übersteigen. Bei einem komplett linearen Datenverlauf über zwei Phasen hinweg käme der Mann-Whitney U-Test (bei ausreichend vielen Messzeitpunkten) also zu einem signifikanten Ergebnis, auch ohne dass eine Förderung innerhalb der zweiten Phase wirklich erfolgreich gewesen wäre.

Eine Alternative für solche Fälle stellen Randomisierungstests dar (Grünke, 2012). Hierbei wird vor der Datenerhebung festgelegt, wie viele Erhebungen in einem Experiment insgesamt stattfinden sollen und wie viele Messzeitpunkte in den jeweiligen Phasen A und B mindestens anzusetzen sind. In Abhängigkeit dieser Rahmenbedingungen kann die Intervention nun theoretisch zu verschiedenen Zeitpunkten einsetzen, wobei der tatsächliche Startpunkt per Zufall bestimmt wird. Nachdem alle Erhebungen stattgefunden haben, werden auf Grundlage der erhobenen Werte für alle potentiell möglichen Interventionsstartpunkte die Mittelwertdifferenzen zwischen Baseline und Intervention berechnet. Im Falle eines erwarteten Werteanstiegs werden diese dann von der höchsten bis zur niedrigsten Differenz aufgereiht. Anschließend wird der ermittelte Rang der tatsächlich gefundenen Mittelwertdifferenz durch die Anzahl aller Permutationen geteilt. Aus dem Quotient ergibt sich der zugehörige *p*-Wert.

Der Randomisierungstest ist somit ein kritischer Schätzer für den Interventionserfolg,

gerade auch bei Daten, die in der Baseline schon einen Trend aufweisen. Bei einem komplett linearen Datenverlauf über zwei Phasen hinweg käme der Randomisierungstest – im Gegensatz zum Mann-Whitney U-Test – zu keinem signifikanten Ergebnis, da alle möglichen Differenzen zwischen Baseline und Intervention gleich wären.

Für das aufgeführte Beispiel wurden insgesamt zehn und mindestens drei Messzeitpunkte für die Phasen A und B angesetzt. Somit ergeben sich fünf mögliche Permutationen für das Verhältnis von Baseline- und Interventionsmessungen: A/B = 3/7, 4/6, 5/5, 6/4, 7/3. Per Zufall wurde bestimmt, dass die Intervention ab dem sechsten Messzeitpunkt startet. Umgesetzt wurden also jeweils 5 Messungen in beiden Phasen. Die tatsächlich gefundene Mittelwertdifferenz beträgt auf Grundlage der erhobenen Daten für das Verhältnis 5/5 $M(B_5)$ - $M(A_5)$ = 2.80 und entspricht dem dritten Rang aller möglicher Permutationen (Abbildung 6). Der *p*-Wert liegt in diesem Fall bei $p = 3/5 = .60$. Die Tatsache, dass zwei der möglichen Differenzen unter dem realen Wert liegen, kann also lediglich mit einer Wahrscheinlichkeit von 40 % nicht allein durch den Zufall erklärt werden (vgl. Grünke, 2012) und wird somit als nicht-signifikantes Ergebnis eingestuft.

In der Praxis macht es aus empirischer und pädagogischer Perspektive Sinn, für Baseline und Interventionsphase mehr als nur drei Erhebungen anzusetzen, damit die Daten überhaupt eine gewisse Aussagekraft haben und Fördermaßnahmen nicht nur sehr kurz zum Einsatz kommen. Um ein signifikantes Ergebnis von $p < .05$ überhaupt erreichen zu können, müssen außerdem mindestens 21 Permutationen möglich sein. Es sollten also insgesamt – unter der Voraussetzung von bspw. minimal 5 Messzeitpunkten pro Phase – wenigstens 30 Erhebungen umgesetzt werden, damit im Rahmen des Randomisierungstests überhaupt ein signifikantes Ergebnis erzielt werden kann.

Händisch ist die Berechnung von Randomisierungstests sehr aufwändig. Für einzelne Entwicklungsverläufe ist die Kalkulation mit bspw. Excel durchaus möglich. Eine praktische Auswertung gerade auch über mehrere Fälle hinweg kann nach Einarbeitung auch z. B. über das Statistikprogramm „R“ mit Hilfe des „scan“-Packets von Wilbert (2016) vorgenommen werden.

Tabelle2: Signifikanzprüfung mit Hilfe des Randomisierungstests

A/B	M(A)	M(B)	M(B) – M(A)	Rang
7/3	3.57	7.00	3.43	1
6/4	3.33	6.50	3.17	2
5/5	**3.20**	**6.00**	**2.80**	**3**
3/7	3.00	5.29	2.29	4
4/6	3.25	5.50	2.25	5

A/B: potentielle Messzeitpunktver-hält-nisse beider Phasen (fett markiert: tatsächliche Umsetzung)

M(A): Mittelwert von Phase A

M(B): Mittelwert von Phase B

R: Rangplatz der tatsächlichen Mittelwertdifferenz im Vergleich mit allen potentiellen Permutationen

P: Anzahl aller möglicher Permutationen in Abhängigkeit der gesetzten Rahmen-bedingungen

Signifikanzprüfung: $p = \frac{R}{P} = \frac{3}{5} = .60$

Fazit

Brunstein und Julius (2014) verweisen darauf, dass aus rechtlichen wie aus ethischen Gründen Interventionen auf ihre Wirksamkeit hin überprüft werden sollten. Für den schulischen Bereich bietet die Erhebung und Auswertung von Einzelfalldaten eine praktische Möglichkeit, eingesetzte Fördermaßnahmen zu beurteilen. Im aktuellen Beitrag wurden verschiedene Erhebungs- und Auswertungsmethoden vorgestellt, die hierzu genutzt werden können. Letztendlich müssen Lehrkräfte selbstständig entscheiden, welche Verfahren fallbezogen am sinnvollsten erscheinen und wie detailliert die erhobenen Daten analysiert werden sollen.

Literaturverzeichnis

Alresheed, F., Hott, B. L. & Bano, C. (2013). Single subject research: A synthesis of analytic methods. *Journal of Special Education Apprenticeship, 2*, 1-18.

Börnert-Ringleb, M., Bosch, J. & Wilbert, J. (2018). Lernverlaufsdiagnostik. In M. Dziak-Mahler, T. Hennemann, S. Jaster, T. Leidig & J. Springob (Hrsg.), *Fachdidaktik inklusiv II - (Fach-)Unterricht inklusiv gestalten - Theoretische Annäherungen und praktische Umsetzungen* (S. 63-78). Köln: Waxmann.

Brunstein, J. C. & Julius, H. (2014). Evaluation von Interventionen durch Einzelfallstudien. In G. W. Lauth, M. Grünke & J. C. Brunstein (Hrsg.): *Interventionen bei Lernstörungen: Förderung, Training und Therapie in der Praxis* (S. 119-138). Göttingen: Hogrefe.

Casale, G., Hennemann, T., Huber, C. & Grosche, M. (2015). Testgütekriterien der Verlaufsdiagnostik von Schülerverhalten im Förderschwerpunkt Emotionale und soziale Entwicklung. *Heilpädagogische Forschung, 41*, 37-54.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, 37, 184-192.

Grünke, M. (2012). Auswertung von Daten aus kontrollierten Einzelfallstudien mit Hilfe von Randomisierungstests. *Empirische Sonderpädagogik, 4*, 247-264.

Hillenbrand, C. & Pütz, K. (2008). *KlasseKinderSpiel. Spielerisch Verhaltensregeln lernen*. Hamburg: Edition Körber Stiftung.

Huber, C. & Grosche, M. (2012). Das response-to-intervention-Modell als Grundlage für einen inklusiven Paradigmenwechsel in der Sonderpädagogik. *Zeitschrift für Heilpädagogik*, 63, 312-322.

Huber, C. & Rietz, C. (2015). Direct Behavior Rating (DBR) als Methode zur Verhaltensverlaufsdiagnostik in der Schule: Ein systematisches Review von Methodenstudien. *Empirische Sonderpädagogik*, 7, 75-98.

Jain, A. & Spieß, R. (2012). Versuchspläne der experimentellen Einzelfallforschung. *Empirische Sonderpädagogik, 4*, 211-245.

Ma, H.-H. (2006): An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. *Behavior Modification, 30*, 598-617.

Parker, R. I., Hagan-Burke, S. & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education, 40*, 194-204.

Parker, R. I. & Vannest, K. (2009): An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*, 357-367.

Rasch, B., Friese, M., Hofmann, W. J. & Naumann, E. (2014a). *Quantitative Methoden 1. Einführung in die Statistik für Psychologen und Sozialwissenschaftler* (4. Aufl.). Berlin: Springer.

Rasch, B., Friese, M., Hofmann, W. J. & Naumann, E. (2014b). *Quantitative Methoden 2. Einführung in die Statistik für Psychologen und Sozialwissenschaftler* (4. Aufl.). Berlin: Springer.

Rosebrock, C., Nix, D., Rieckmann, C. & Gold, A. (2011). *Leseflüssigkeit fördern. Lautleseverfahren für die Primar- und Sekundarstufe.* Seelze: Kallmeyer/Klett.

Strathmann, A.M. & Klauer, K.J. (2012). *LVD-M 2-4: Lernverlaufsdiagnostik – Mathematik für zweite bis vierte Klassen*. Göttingen: Hogrefe.

Sugai, G. & Horner, R.H. (2006). A promising approach for expanding and sustaining School-wide Positive Behavior Support. *School Psychology Review*, 35, 245-259.

Vannest, K. J., Davis, J. L. & Parker, R. I. (2013). *School-Based Practice in Action Series. Single Case Research in Schools: Practical Guidelines for School-Based Professionals*. New York, NY: Routledge/Taylor & Francis Group.

Walter, J. (2009). *LDL: Lernfortschrittsdiagnostik Lesen – Ein curriculumbasiertes Verfahren*. Göttingen: Hogrefe.

Walter, J. (2013). *VSL: Verlaufsdiagnostik sinnerfassenden Lesens*. Göttingen: Hogrefe.

White, O. R. & Haring, N. G. (1980). *Exceptional Teaching: A Multimedia Training Package*. Columbus, OH: Merrill.

Wilbert, J. (2016). *scan. Single case data analyses for AB-designs (Version 0.20)*. Potsdam: University of Potsdam. Abgerufen von cran.r-roject.org/web/packages/scan/index.html

Wolery, M., Busick, M., Reichow, B. & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education, 44*, 18-28.