Empirische Sonderpädagogik, 2012, Nr. 3/4, S. 247–264

Auswertung von Daten aus kontrollierten Einzelfallstudien mit Hilfe von Randomisierungstests

Matthias Grünke

Universität zu Köln

Zusammenfassung
Zahlreiche Forschungsfragen innerhalb der Sonderpädagogik lassen sich am besten mittels kontrollierter Einzelfallanalysen bearbeiten. Derartige Ansätze haben innerhalb dieser Disziplin eine lange Tradition. Die Frage nach geeigneten Methoden zur Auswertung von Daten aus entsprechenden Studien wird jedoch bis heute sehr kontrovers diskutiert. Sowohl visuelle Inspektionen als auch Effektstärkeberechnungen sind oftmals problembehaftet. Inferenzstatistische Auswertungen mittels üblicher parametrischer Tests kommen aufgrund massiver Verletzungen ihrer Voraussetzungen in der Regel nicht in Frage. Randomisierungstests könnten hier eine brauchbare Alternative darstellen, allerdings gilt ihr Einsatz als enorm aufwändig und unhandlich. Sie spielen deswegen bei der Auswertung von Daten aus kontrollierten Einzelfallstudien in der sonderpädagogischen Forschungspraxis bislang keine Rolle. Neuerdings ist es allerdings möglich, die Analyse unter Verwendung effizienter Makros bzw. Syntaxen mit herkömmlichen PC-Programmen (Microsoft Excel oder IBM SPSS Statistics) relativ ökonomisch durchzuführen. Zwar weisen auch Randomisierungstests ihre eigenen methodischen Probleme auf. Außerdem ist ihre Verwendung nur im Zusammenhang mit bestimmten Fragestellungen angebracht. Insgesamt handelt es sich bei Randomisierungstests jedoch um sehr vielversprechende Ansätze, die sicherlich zu einer merklichen Steigerung der Aussagekraft vieler kontrollierter Einzelfallanalysen beitragen können.

Schlüsselwörter: Einzelfallstudie, Randomisierungstest, inferentielle Datenauswertung, visuelle Inspektion, Effektstärkeberechnung

Analyzing data from single-subject designs using randomization tests

Abstract
Many research questions in the field of special education are best tackled using single-subject studies. These approaches have a long tradition within this discipline. However, the question of how to best analyze data from those designs remains controversial. Visual inspections and effect size calculations are often problematic. Using common parametric tests to inferentially analyze the data is often unsuitable due to severe violations of their assumptions. Randomization tests could present a useful alternative in this regard, but they are considered to be extremely intricate and unmanageable. This explains why these methods have not yet played a role in analyzing data from single-subject designs in practical research within the scope of special education. More recently it has been made possible to compute respective calculations rather economically by using efficient macros or syntaxes for certain familiar statistical packages (Microsoft Excel and IBM SPSS Statistics). It has to be noted that randomization tests render their own methodological problems. In addition, they are only suitable for certain research questions. But taken as a whole, randomization tests are certainly very promising concepts that can substantially contribute to improve the validity of many single-subject studies.

Keywords: Single-subject study, randomization test, inferential data analysis, visual inspection, effect size calculation

Zur Bedeutung von Einzelfallstudien in der Sonderpädagogik

Der Nutzen von kontrollierten Einzelanalysen bei der Wirksamkeitsprüfung von Interventionen gilt in den angewandten Humanwissenschaften (Sozialarbeit, Psychotherapie, Neurorehabilitation, Sonderpädagogik, …) als unbestritten. Besonders in Situationen, in denen die relevanten Populationen vergleichsweise heterogen sind oder die Rekrutierung von größeren Versuchsgruppen ein Problem darstellt, lassen sich derartige Methoden aus der Forschungspraxis nicht mehr wegdenken. Vor diesem Hintergrund verwundert es nicht, dass die zentrale Bedeutung kontrollierter Einzelfallanalysen für den Erkenntnisgewinn in der Sonderpädagogik seit Jahrzehnten in der einschlägigen Literatur auffallend oft hervorgehoben wird (z. B. Barnett, Daly, Jones & Lentz, 2004; Horner, Carr, Halle, McGee, Odom&Wolery, 2005; Julius, Schlosser & Goetze, 2000; Kern, 1996; 1997a; 1997b; Wember, 1994). Denn gerade diese humanwissenschaftliche Teildisziplin zeichnet sich dadurch aus, dass empirische Studien vielfach in Schulen, Heimen oder anderen Einrichtungen durchgeführt werden, in denen oft nur recht kleine und inhomogene Stichproben zur Verfügung stehen (Schindele, 1985).

Gängige Methode zur Auswertung von Einzelfallstudien und ihre Probleme

Visuelle Inspektion

Kontrollierte Einzelfallanalysen basieren auf einer sehr häufigen Messung eines relevanten Zielverhaltens (z. B. Lesegeschwindigkeit, prosoziale Äußerungen, konzentriertes Arbeiten) bei Einzelpersonen oder sehr wenigen Menschen. Die multiplen Beobachtungen erfolgen hierbei vor, während und manchmal auch nach einer Intervention. Aus dem Verlauf der Daten über die verschiedenen Phasen hinweg werden letztendlich Schlussfolgerungen über die Wirksamkeit einer untersuchten Maßnahme gezogen. Der mit Abstand üblichste Weg bei der Auswertung ist die graphische Darstellung der Ergebnisse anhand eines einfachen Liniendiagramms und einer anschließenden „visuellen Inspektion" (Busk & Marascuilo, 1992; Horner et al., 2005). Hierbei werden die Datenpunkte der im Fokus stehenden Zielvariablen auf der Y-Achse und die Zeiteinheiten auf der X-Achse abgetragen. Eine Beurteilung der Effektivität beruht bei dieser Methode auf dem Eindruck, den eine Person beim Betrachten der Graphik gewinnt. Zuweilen ist der Nutzen einer Maßnahme mit Blick auf ein entsprechendes Diagramm so offensichtlich, dass sich jegliche Diskussion erübrigt. Edwards, Lindman und Savage (1963) bringen solche Situationen wie folgt auf den Punkt: „… you know what the data mean when the conclusion hits you between the eyes" (S. 217). In einem Großteil der Fälle stellt sich die Sachlage jedoch alles andere als eindeutig dar. Die Auffassung, dass die visuelle Dateninspektion bei kontrollierten Einzelfallanalysen im Allgemeinen als eine relativ konservative Auswertungsoption anzusehen ist (z. B. Parsonson & Baer, 1986) oder dass geschulte Expertinnen und Experten bei der Bewertung des Nutzens einer Intervention relativ hohe Übereinstimmungen erzielen (z. B. Huitema, 1986), lässt sich durch die Empirie nicht stützen. Einschlägige Studien bieten hier vielmehr ein eher ernüchterndes Bild: So fassen etwa Brossart, Parker, Olson und Mahadevan (2006) den Wissenstand zu dieser Fragestellung zusammen und präsentieren verschiedene Untersuchungen, in denen die Interrater-Reliabilitäten lediglich zwischen 0,40 und 0,60 schwanken. Selbst ein vorheriges ausgiebiges Einweisen in das Vorgehen bei der visuellen Inspektion führt

nicht zu einer nennenswert besseren Übereinstimmungsquote (Ottenbacher, 1993). Der offensichtliche Grund hierfür liegt darin, dass bei diesem Ansatz auf keine allgemein akzeptierten und gut objektivierbaren Bewertungskriterien zurückgegriffen wird. Ohne das Vorhandensein konkreter Beurteilungsmaßstäbe lassen sich diese natürlich auch nicht vermitteln. Zwar stellen Gast und Spriggs (2010) in ihrem viel zitierten Artikel „Visual Analysis of Graphic Data" vergleichsweise „handfeste" Richtlinien für die Abwägung der Wirksamkeit von Interventionen vor, allerdings kann ihr Konzept kaum noch als visuelle Inspektion bezeichnet werden. Vielmehr handelt es sich hierbei bereits um eine statistische Analyse.

Effektstärkeberechnung

Vor dem Hintergrund der offensichtlichen Nachteile einer visuellen Inspektion werden in der Literatur als Ergänzung bei der Auswertung von Daten aus kontrollierten Einzelfallanalysen oft verschiedene Effektstärkemaße vorgeschlagen. Sie stellen einen quantitativen Indikator für die Größe des Behandlungserfolgs dar. Als die häufigste Methode gilt in diesem Zusammenhang der Prozentsatz nicht überlappender Daten („Percent of Non-Overlapping Data", PND), bei dem „… die Anzahl der Datenpunkte während der Interventionsphase, die nicht mit den dazugehörigen Grundratendaten überlappen, … durch die Gesamtzahl der Datenpunkte in der Interventionsphase … dividiert und dann mit 100 multipliziert [wird]" (Kern, 1997b, S. 162). Von einer mittelmäßig erfolgreichen Behandlung spricht man, wenn die Quote zwischen 70 und 90% liegt. Geht sie darüber hinaus, so gilt sie als hoch bzw. als sehr hoch (Scruggs, Mastropieri&Casto, 1987). Die große Verbreitung dieser Art von Effektstärkeberechnung dürfte jedoch mehr auf die Einfachheit ihrer Durchführung und weniger auf die solide Aussagekraft der Kennwerte zurückzuführen sein. Indices sind nämlich dann wenig valide, „… wenn sich bereits in der Grundratenphase ein Datentrend in gewünschter Richtung der Interventionsphase zeigt, oder wenn in der Grundratenphase Datenwerte um Null erhoben wurden" (Julius, Schlosser & Goetze, 2000, S. 138). In solchen Fällen kommt es nämlich zu einer groben Über- oder Unterschätzung der Wirksamkeit. Als Alternativen zum PND finden sich in der Literatur zahlreiche Vorschläge: „Percent of Zero Data" (PZD; Scotti, Evans, Meyer & Walker, 1991), „Percent of Data Exceeding the Median of Baseline" (PEM; Ma, 2006), „Mean Baseline Difference" (MBD; Campell, 2003), „Standardized Mean Difference" (SMD; Busk & Serlin, 1992), „Regression-Based Standardized Mean Difference" (dREG; Allison & Gorman, 1993), „Hierarchical Linear Modeling" (HLM, Van den Noorgate & Onghena, 2003), … Allein die Vielzahl der Vorschläge deutet an, dass eine Ergänzung der visuellen Inspektion als Auswertungsmethode um ein Effektstärkemaß gemeinhin als notwendig erachtet wird, dass im Hinblick auf die beste Alternative in der Fachwelt jedoch wenig Einigkeit besteht. Campell und Herzinger (2010) schreiben hierzu: „… little consensus exists regarding the appropriate calculation of effect sizes for single-case designs" (S. 440). Sie führen weiter aus: „… the current state of knowledge prohibits recommendation of a single subject effect size for the purpose of quantitative synthesis" (S. 447f.).

Inferenzstatistische Analyse

Untersuchungen zur Überprüfung der Wirksamkeit einer Intervention beinhalten normalerweise nicht nur eine graphische Aufbereitung der Beobachtungen und eine Ermittlung von Effektstärken, sondern auch eine inferentielle Auswertung der Daten. Selbst im Zusammenhang mit kontrollierten

Einzelfallanalysen ist dies trotz der äußerst geringen Stichprobenzahl von oft nur $N = 1$ immer wieder vorgeschlagen und umgesetzt worden. Hier ist allerdings eine ganz besondere Herausforderung zu konfrontieren, die bei Gruppenstudien im Allgemeinen eine weitaus geringere Rolle spielt: der Umgang mit der seriellen Abhängigkeit der Residuen (also mit deren Autokorrelation). Unter Residuen sind Abweichungen von den exakten Ergebnissen zu verstehen – sie repräsentieren also die „Fehler", die bei einer Messung ungewollt mit erhoben werden. Die üblichen parametrischen Verfahren (z. B. t-Test, ANOVA) sind gegenüber den meisten Verletzungen ihrer Voraussetzungen (Normalverteilung der abhängigen Variablen, verschieden große Stichproben, Homogenität der Varianzen) relativ robust. Allerdings muss die Bedingung, dass die Residuen in ausreichendem Maße voneinander unabhängig sind, in jedem Fall gegeben sein (Borckardt & Nash, 2002; Manolov, Arnau, Solanas &Bono, 2010). Bei großen Studien mit vielen Personen, einer Zufallszuweisung zu den Versuchsbedingungen und relativ wenigen Messungen, stellt dies in der Regel kein Problem dar. Es gibt meist keinen plausiblen Grund, warum die Residuen bei den einzelnen Probandinnen und Probanden innerhalb einer Gruppe systematisch voneinander abhängen sollten. Deswegen wird dieses Thema in der einschlägigen Literatur oftmals nicht einmal angesprochen.

Bei kontrollierten Einzelfallstudien verhält sich die Sachlage jedoch anders. Möchte man die üblichen parametrischen Verfahren bei der Analyse von Daten aus diesen Forschungsdesigns verwenden, müsste man die verschiedenen Untersuchungsphasen in der Auswertung wie Gruppen und die Beobachtungen wie Probandinnen bzw. Probanden behandeln. Wenn jedoch bei ein und derselben Person multiple Erhebungen durchgeführt werden, besteht eine erhebliche Gefahr, dass der Verlauf der Daten nicht nur von der Entwicklung der jeweils im Fokus stehenden (abhängigen) Variable, sondern in systematischer Weise auch von weiteren Faktoren (beispielsweise von Gewöhnungseffekten) abhängt. Sind die Residuen in positiver Weise miteinander autokorreliert, reduziert sich der Standardfehler bei den meisten parametrischen Tests, was zu einer Erhöhung der jeweiligen Prüfgröße (t, F) führt. Dadurch steigt die Gefahr, einen Fehler 1. Art zu begehen. Man unterliegt somit einem erhöhten Risiko, die Nullhypothese zurückzuweisen, obwohl sie in Wirklichkeit zutrifft (Suen & Ary, 1987). Beträgt die serielle Abhängigkeit der Residuen $r = 0{,}30$, kann sich diese Gefahr u. U. sogar verdreifachen (Scheffé, 1959). Bei einer negativen Autokorrelation sinkt die jeweilige Prüfgröße und es steigt das Risiko, einen Fehler 2. Art zu begehen (also die Nullhypothese beizubehalten, obwohl sie falsch ist) (Parker & Brossart, 2003).

Huitema (1985), Kazdin (1984) und andere Autorinnen bzw. Autoren schlagen zwar trotzdem vor, zur Auswertung von Daten aus kontrollierten Einzelfallstudien t-Tests oder Varianzanalysen zu verwenden, bei denen die Werte einer einzigen Person aus den einzelnen Phasen (Baseline, Intervention, Maintenance) miteinander verrechnet werden. Allerdings lässt sich dies in Anbetracht des besprochenen Problems der seriellen Abhängigkeiten nicht rechtfertigen (Busk & Marascuilo, 1992).

Eine potenzielle Möglichkeit, um den Einfluss von Autokorrelationen zu kontrollieren, stellt der Einsatz regressionsanalytischer Verfahren dar. Mit Hilfe dieser Methoden sollen Veränderungen im Level oder in der Steigung von Messwertreihen aufgespürt werden (Brossart et al., 2006). So lässt sich anhand der Daten aus den A-Phasen der Verlauf der abhängigen Variable während der B-Phasen vorhersagen. Kommt es zu bedeutsamen Abweichungen – werden beispielsweise im Zuge einer Förderung einfacher Additionsfertigkeiten deutlich

bessere Leistungen erzielt, als dies auf Grundlage der Entwicklung der Messwerte während der Baseline zu erwarten gewesen wäre – so liegt ein Indiz für die Wirksamkeit der Maßnahme vor. Manolov et al. (2010) konnten jedoch zeigen, dass zu hohe (positive) Autokorrelationen der Residuen selbst dann die Gefahr erhöhen, einen Fehler 1. Art zu begehen, wenn vorab Maßnahmen ergriffen wurden, um die Effekte dieser seriellen Abhängigkeiten herauszupartialisieren. Campell und Herzinger (2010) weisen auf ein weiteres potenzielles Problem regressionsanalytischer Ansätze hin: Der Verlauf der Messungen in der Baselinephase könnte u. U. den Anschein erwecken, als ob ein linearer Trend vorläge, obwohl dies mitunter gar nicht möglich wäre, da eine Prognose irgendwann unrealistische Werte ergeben würde. Der Autor und die Autorin führen zur Erläuterung ein Beispiel an, in dem es um eine Förderung zur Erhöhung sozial erwünschter Interaktionen geht. Da die Anzahl der Minuten, in denen sich die Versuchsperson im jeweiligen Beobachtungszeitraum in erhoffter Weise verhält, während der Baselinephase abnimmt, werden für die Interventionsphase ab einem bestimmten Zeitpunkt negative Zeiten vorhergesagt.

Im Hinblick auf die Kontrolle serieller Abhängigkeiten im Rahmen von Zeitreihenanalysen hat Köhler (2008) ein stimmiges Konzept vorgelegt: Nach der Erhebung der Daten ist zunächst die Höhe der Autokorrelationen bei den Residuen festzustellen. Liegt sie bei über $r = 0{,}20$ (bzw. bei unter $r = -0{,}20$), so ist nur jeder zweite (oder gar nur jeder dritte) Wert in die inferenzstatistische Analyse mittels üblicher parametrischer Verfahren (v. a. mittels des t-Tests) miteinzubeziehen. Der offensichtliche Nachteil eines solchen Vorgehens liegt darin, dass eine relativ (und oft unpraktikabel) hohe Anzahl an Daten nötig ist, um sinnvolle Auswertungen vornehmen zu können.

Als Alternative zu den bekannten parametrischen Tests sind in der Literatur wiederholt verteilungsfreie Verfahren (wie etwa spezielle Rangordnungstests) vorgeschlagen worden (vgl. Julius, Schlosser & Goetze, 2000). In Abhängigkeit von der Beschaffenheit der Messwerte lassen sich die negativen Auswirkungen eines zu hohen Zusammenhangs von aufeinander folgenden Residuen auf die Validität eines Tests jedoch auch bei diesen Methoden in vielen Fällen nicht ausreichend kontrollieren. Doch selbst wenn sich dieses Problem irgendwie in den Griff bringen ließe, bleiben andere Herausforderungen bestehen: Bei verteilungsfreien Verfahren werden alle Werte auf ein Rang- oder gar auf ein Nominalniveau reduziert, wodurch ein Teil der in den Daten enthaltenen Informationen verloren geht und sich die Teststärke (Power) reduziert. Dadurch erhöht sich wiederum die Gefahr, einen Fehler 2. Art zu begehen (Bortz, Lienert & Boehnke, 2008).

Neben den eben angesprochenen und relativ oft vorgeschlagenen Ansätzen zur inferenzstatistischen Analyse von Daten aus kontrollierten Einzelfallstudien finden sich in der Literatur noch zahlreiche weitere Alternativen (z. B. Bloom & Fisher, 1982; Orme & Cox, 2001; Tyron, 1982), die teilweise ganz erhebliche Nachteile aufweisen und deswegen nie Gegenstand einer breiten fachwissenschaftlichen Diskussion waren.

Zur Logik von Randomisierungstests

Um das Wesen von Randomisierungstests plausibel zu machen, erscheint es sinnvoll, sich die Merkmale von „klassischen" inferentiellen Analysen ins Gedächtnis zu rufen, von denen sich die Charakteristika der in diesem Artikel vorgestellten Methodengruppe unterscheiden. Da die üblichen Auswertungsverfahren im Rahmen von Gruppenstudien zum Einsatz kommen, be-

ziehen sich die folgenden Ausführungen auf diesen Anwendungsbereich. Das „traditionelle“ Prozedere bei inferenzstatistischen Analysen basiert primär auf der Theorie von Neyman und Pearson (1933). Bei den in diesem Zusammenhang am häufigsten eingesetzten Verfahren geht es um die Beantwortung der Frage, mit welcher Wahrscheinlichkeit Unterschiede im Hinblick auf relevante arithmetische Mittel (oder Mediane) in untersuchten Gruppen zufällig zustande gekommen sein können (allerdings lassen sich natürlich auch Differenzen von Standardabweichungen, Prozentwerten oder Häufigkeitsverteilungen sowie Korrelations- oder Regressionskoeffizienten auf ihre Signifikanz hin überprüfen). Möchte man beispielsweise unter Verwendung eines so genannten Proaktionsplans (vgl. Grünke & Masendorf, 2000) herausfinden, ob ein bestimmtes Training zur Verbesserung von basalen Additions-, Subtraktions-, Multiplikations- und Divisionskompetenzen bei lernschwachen Kindern tatsächlich sein Ziel erreicht, so vergleicht man die durchschnittlichen Rechenergebnisse einer geförderten Gruppe von Mädchen und Jungen mit denen einer ungeförderten. Als geeignetes Auswertungsverfahren bietet sich in diesem Fall ein t-Test für unabhängige Stichproben an. Mit Hilfe der relevanten Prüfgröße (t) wird nun festgestellt, ob die beiden arithmetischen Mittel zur gleichen Grundgesamtheit gehören oder nicht. Dieser eben kurz angedeutete Gedankengang stellt den Kern der Vorgehensweise auf der Basis der Theorie von Neyman und Pearson (1933) dar: Man bestätigt oder verwirft die Nullhypothese. Es geht hierbei also um die Frage, ob ein bestimmtes Ergebnis bei Gültigkeit der (meistens parametrischen Prüfverteilung) mit einer vorab festgelegten Wahrscheinlichkeit zufällig zustande gekommen ist. Ausgegangen wird hierbei davon, dass es in der Grundgesamtheit aller Kinder, die mit Blick auf die evaluierte Maßnahme relevant erscheinen, eine bestimmte Verteilung der Rechenleistungen gibt. Diese spiegelt sich (vermeintlich) in den Eigenschaften der Daten aus der Kontrollgruppe wider. Ob eventuelle Abweichungen der Werte der Mädchen und Jungen aus der Trainingsgruppe von den Verteilungsparametern der Grundgesamtheit noch mit einer vertretbaren Wahrscheinlichkeit mit dem Zufall erklärbar sind, ergibt sich eben durch einen Vergleich der beiden Teilstichproben mittels des erwähnten t-Tests.

Ein solches Herangehen an Forschungsfragen wirft allerdings eine Reihe von Problemen auf, die in der Literatur seit Jahrzehnten diskutiert werden: Zu den wichtigsten gehört zweifellos der Umstand, dass die Nullhypothese (und nicht die Alternativhypothese) den Ausgangspunkt einer inferentiellen Analyse darstellt. Somit liefert die Auswertung häufig eine Antwort auf eine Frage, die überhaupt nicht im Zentrum des Interesses steht. Todman und Dugard (2001) bringen dies wie folgt auf den Punkt: „What you really want is to be able to make a probability statement about the hypothesis given your data. What you get is a probability statement about your data given the null hypothesis, which is usually not even the one you are interested in” (S. 28f.). Ein weiterer heikler Aspekt dieses Vorgehens besteht darin, dass Stichproben in aller Regel nicht per Zufall aus einer Population gezogen werden, obwohl eine „echte“ Zufallsstichprobe die Grundlage jedes inferenzstatistischen Schlusses ist.

Den Randomisierungstests liegen hingegen andere Prämissen zugrunde als den üblichen Methoden der Inferenzstatistik. Es handelt sich hierbei um eine Subgruppe der so genannten Resamplingmethoden, zu denen auch das Bootstrapping- und das Jackknife-Verfahren gehören (vgl. Efron & Tibshirani, 1993). Bei Randomisierungstests geht es nicht darum, einen Wert (z. B. t- oder F-Wert) aus einer wahrscheinlichkeitstheoretisch ermittelten Tabelle mit einem konkreten empirisch errechneten Wert

zu vergleichen, um dadurch eine Nullhypothese zu bestätigen oder zu verwerfen. Stattdessen wird davon ausgegangen, dass die konkret vorliegenden Rohdaten die einzige Basis zur Abschätzung einer Verteilung darstellen. Man vergleicht in diesem Zusammenhang eine für die eigene Forschungsfrage relevante Teststatistik mit den Indices einer Liste, die sich aus einer wiederholten Neuanordnung der Messungen ergibt. Je nachdem, wie viel Prozent der Werte aus dieser Liste größer oder kleiner als die besagte Teststatistik sind, gilt die im Rahmen einer Studie zu überprüfende Annahme als bestätigt oder als widerlegt (Saint-Mont, 2011). Letztendlich wird davon ausgegangen, dass die Stichprobe der Population entspricht. Man ermittelt eine empirische Stichprobenkennwerteverteilung durch alle möglichen Permutationen, die dann durch die Beantwortung der folgenden Frage ein ähnliches Schließen wie bei den üblichen inferenzstatistischen Verfahren erlaubt: Wie wahrscheinlich ist es, ein bestimmtes empirisches Ergebnis unter Berücksichtigung aller „möglichen" Werte zu erhalten?

Die ersten konkreten und speziellen Grundlagen zur Entwicklung von Randomisierungstests stammen von Fisher (1926; 1934). Das Bekanntwerden dieser Methoden ist v. a. Edgington zu verdanken, der sie 1969 mit seinem viel beachteten Lehrbuch „Statistical inference: The distribution-free approach" systematisch darstellte und damit in die breite Fachdiskussion einführte. Es folgten zwei weitere einschlägige Werke des Autors (z. B. Edgington, 1980; Edgington & Onghena, 2007), die zweifellos zur Standardliteratur gehören. Hersen und Barlow veröffentlichten 1976 ein erstes Lehrbuch, in dem auf die Möglichkeiten der Auswertung von Daten aus kontrollierten Einzelfallstudien mittels Randomisierungstests eingegangen wird. Tawney und Gast (1984) gehen in ihrem Werk ebenfalls auf diese Thematik ein und stellen hierbei insbesondere den Nutzen für die sonderpädagogischen Forschung heraus.

Das Prinzip der Randomisierungstests beruht bei ihrer Anwendung im Bereich der kontrollierten Einzelfallstudien auf den folgenden fünf Schritten (vgl. Sierra, Solanas & Quera, 2005):

(1) Bei dem gewählten Design (z. B. ABA-Plan, multipler Grundratenversuchsplan, Kriterien-Veränderungsplan) muss mindestens ein Wechsel zwischen zwei Phasen (Nicht-Behandlung vs. Behandlung) innerhalb definierter Grenzen zufällig bestimmt werden. So wäre es bei einem ABA-Plan etwa denkbar, dass man vorhat, die abhängige Variable im Verlauf des Versuchs über einen Zeitraum von sechs Wochen einmal täglich (also insgesamt 42 Mal) zu erfassen. Die Behandlung soll sich in diesem Beispiel aufgrund inhaltlicher Notwendigkeiten auf mindestens zehn Tage erstrecken. Für die zweite A-Phase, in der das relevante Zielmerkmal nach dem Absetzen der Intervention noch weiter erfasst wird, sind genau fünf Messzeitpunkte (so genannte „Probes") vorgesehen, in der ersten A-Phase (in der so genannten Baseline) sollen zumindest nicht weniger als fünf Datenerhebungen stattfinden. Innerhalb dieser Grenzen ergeben sich nun Spielräume, in denen die Behandlung beginnen kann. Gäbe es beispielsweise fünf Probes in der ersten A-Phase, dann blieben 42 (Anzahl der Gesamtmessungen) – 5 (Anzahl der Messungen in der ersten A-Phase) – 5 (Anzahl der Messungen in der zweiten A-Phase) = 32 Datenerhebungen für die B-Phase übrig. Damit wären alle Bedingungen erfüllt. Allerdings könnte die erste A-Phase auch sechs Messzeitpunkte umfassen und die B-Phase infolgedessen aus 42 – 6 – 5 = 31 Probes bestehen. Spielt man alle Möglichkeiten durch, kommt man in diesem Fall auf 23 Permutationen. Eine davon muss

nun per Zufall bestimmt und im Zuge der Studie umgesetzt werden.

(2) Der statistische Wert, um den es bei der jeweiligen Untersuchung geht, ist anhand der erhobenen Daten zu ermitteln. In den allermeisten Fällen ist hier eine Mittelwertsdifferenz relevant. Realisiert man beispielsweise einen einfachen AB-Plan, so ist die wesentliche statistische Maßzahl der Unterschied zwischen den durchschnittlichen Werten aus der A-Phase und der B-Phase. Bei komplexeren Designs steht eine Summe aus verschiedenen Mittelwertsdifferenzen im Fokus. Im Falle einer Erweiterung eines AB-Plans im Sinne eines multiplen Grundratenversuchsplans über Probandinnen bzw. Probanden müssten die jeweiligen Unterschiede zwischen den durchschnittlichen Werten aus den verschiedenen Phasen aufaddiert werden.

(3) Im Anschluss ist diese relevante statistische Maßzahl für alle Optionen zu ermitteln, die unter Einhaltung der vorab festgelegten Bedingungen denkbar sind. Hierbei verwendet man die tatsächlich erhobenen Daten und stellt fest, welche statistischen Werte sich ergeben, wenn man die Grenzen des Beginns bzw. des Endes der Förderung innerhalb des definierten Rahmens verschiebt.

(4) Die sich bei diesem Unterfangen ergebenden Maßzahlen werden ihrer Größe nach in eine Reihenfolge gebracht.

(5) Man identifiziert die Position desjenigen statistischen Wertes in der Liste, der unter Berücksichtigung der Grenzen errechnet wurde, die während der Untersuchung tatsächlich den Beginn bzw. das Ende der Förderung definiert hatten. Je nachdem, wo sich dieser Index in der Rangreihe befindet, lässt sich nun eine Wahrscheinlichkeit ermitteln, mit der die Unterschiede zwischen den durchschnittlichen Werten in den Phasen nicht zufällig zustande gekommen sein können. Stellt die relevante Maßzahl beispielsweise von hundert Optionen die drittgrößte dar, so beträgt diese Chance 3/100 = 0,03.

Diese kurze Präsentation des grundsätzlichen Vorgehens bei der Durchführung von Randomisierungstests verdeutlicht, dass diese Methode nicht nur eine Form der Datenauswertung, sondern auch einen zentralen Aspekt der Untersuchungsplanung darstellt. Die Länge der einzelnen Phasen sowie die Anzahl der währenddessen durchgeführten Messungen werden bei Einzelfallstudien ansonsten entweder vorab festgelegt oder vom Verlauf der Daten abhängig gemacht. So könnte man eine Grundrate beispielsweise von vornherein auf zehn Probes beschränken oder aber warten, bis sich die Werte „eingependelt" und eine gewisse Konstanz erreicht haben. Bei Randomisierungstests wird hingegen versucht, die interne Validität des Designs zu erhöhen, indem die Wechsel von einer Phase zur jeweils nächsten innerhalb festgelegter Grenzen zufällig zustande kommen (Todman & Dugard, 2001).

Beispiel für die Anwendung eines einfachen Randomisierungstests

Das Prinzip der Randomisierungstests soll nun anhand eines einfachen AB-Designs demonstriert werden. Die Aussagekraft derartiger Versuchspläne gilt als relativ gering. Zu Demonstrationszwecken bietet sich dieses Design jedoch eher an als komplexere Alternativen. In der hier beschriebenen hypothetischen Studie geht es um einen neunjährigen Jungen, der die vierte Klasse einer Grundschule besucht. Seine Leistungen sind insgesamt als durchschnittlich zu bezeichnen, nur seine Orthographie ist weit unter dem Niveau seiner Klasse angesiedelt. Im „Deutschen Rechtschreibtest für das dritte und vierte Schuljahr" (Stock & Schneider, 2008) erreicht er einen T-Wert von 34 (was einem Prozentrang von ungefähr 5 entspricht). Die Durchführung der

„Oldenburger Fehleranalyse“ (Thomé & Thomé, 2009) offenbart, dass Probleme mit den Lautnachbarschaften „St“ und „Sp“ für mehr als ein Drittel seiner unkorrekt geschriebenen Worte verantwortlich sind. Dem Jungen ist augenscheinlich nicht bewusst, dass ein gesprochenes „Scht“ bzw. „Schp“ am Wort- oder Silbenanfang „St“ bzw. „Sp“ geschrieben wird (z. B. Stab, an|stecken, Spaß, ver|sprechen). Im Zuge der sechswöchigen Studie werden dem Kind an jedem Schultag (also insgesamt 30 mal) zehn zufällig ausgewählte Worte diktiert, die zu gleichen Teilen aus einem Pool der 50 häufigsten Worte mit „St“ und aus einem Pool der 50 häufigsten Worte mit „Sp“ stammen (Universität Leipzig, 2012). Die zu evaluierende Intervention besteht aus einer direktiven Vermittlung der noch nicht verinnerlichten Rechtschreibregeln und einer passgenauen Förderung mit Hilfe des PC-Programms „GUT 1“ (Grund, 2012). Vorab wird festgelegt, dass sich die Baseline der Leistungserhebung anhand der Diktate zu Beginn der Untersuchung über mindestens fünf Sitzungen erstreckt, und dass die Behandlung mindestens fünf Einheiten umfassen muss. Berücksichtigt man diese Vorgabe, so kann die Förderung frühestens am sechsten, und spätestens am 26. Tag beginnen. Es sind im Hinblick auf den Zeitraum der Intervention innerhalb der festgesetzten 30 Erhebungszeitpunkte theoretisch also insgesamt 21 Konstellationen möglich (6-30, 7-30, 8-30, … 26-30). Die sonderpädagogische Lehrkraft, welche die Studie durchführt, schreibt diese Möglichkeiten jeweils auf einen Zettel und zieht per Zufall die Zusammenstellung „14-30“. Somit liegen 13 Messzeitpunkte vor der Intervention und 17 in der Behandlungsphase. Tabelle 1 gibt die Anzahl der korrekt geschriebenen Worte während der Baseline und während der Förderung wieder.

Die Differenz zwischen den Durchschnittswerten aus den beiden Phasen beträgt 6,94 – 2,23 = 4,71. Nun gilt es zu ermitteln, wie groß die Unterschiede zwischen den Mittelwerten gewesen wären, wenn die Intervention am 6., 7., 8., … oder 26. Tag begonnen hätte. Tabelle 2 gibt einen Überblick über alle Indices bei den insgesamt 21 Konstellationen.

Im Falle einer wirksamen Intervention sollte man im Durschnitt relativ niedrige Werte während der Baseline, und relativ hohe Werte während der Förderung erwarten. Dementsprechend ist bei einer effektiven Behandlung von einer vergleichsweise großen Differenz zwischen den Mittelwerten auszugehen. Wie erwähnt beträgt der Unterschied zwischen M2 und M1 bei der von der Lehrerin per Zufall ausgewählten Konstellation 4,71. Es handelt sich mit Blick auf die Werte in Tabelle 2 um die größte Differenz, die im Rahmen der definierten Grenzen bei variierenden Zeitpunkten des Interventionsbeginns ermittelt werden kann. Die Chance, dass der deutlichste Unterschied tatsächlich bei der Konstellation zu Tage tritt, der mit dem per Los bestimmten Beginn der Förderung einhergeht, liegt in diesem Fall bei 1/21 = 0,048. Somit kann der Umstand, dass keine der 20 alternativen Differenzen den Wert von 4,71 überschreitet, mit ca. 95%iger Wahrscheinlichkeit nicht mit dem Zufall erklärt werden.

Tabelle 1: Ergebnisse der hypothetischen Interventionsstudie für die Konstellation 14-30.

Phase	***N***	**Rohwerte**	***M***
Baseline	13	2; 1; 3; 0; 2; 3; 2; 3; 4; 2; 1; 3; 3;	2,23
Förderung	17	5; 6; 7; 5; 8; 7; 7; 8; 6; 8; 9; 7; 9; 5; 7; 6; 8;	6,94

Rechnungen mit Daten aus relativ simplen Untersuchungen mit wenigen Permutationen (wie in dem eben präsentierten Beispiel) lassen sich im Allgemeinen recht

Tabelle 2: Auflistung aller möglichen Permutationen im Rahmen der vorab definierten Grenzen.

Konstellation	M Baseline (M1)	M Förderung (M2)	M2-M1
6-30	1,60	5,56	3,96
7-30	1,83	5,67	3,84
8-30	1,86	5,83	3,97
9-30	2,00	5,95	3,95
10-30	2,22	6,05	3,83
11-30	2,20	6,25	4,05
12-30	2,09	6,53	4,44
13-30	2,17	6,72	4,55
14-30	2,23	6,94	4,71
15-30	2,43	7,06	4,63
16-30	2,67	7,13	4,46
17-30	2,94	7,14	4,20
18-30	3,06	7,31	4,25
19-30	3,33	7,25	3,92
20-30	3,53	7,27	3,74
21-30	3,70	7,30	3,60
22-30	3,90	7,22	3,32
23-30	4,00	7,38	3,38
24-30	4,17	7,29	3,12
25-30	4,38	7,00	2,62
26-30	4,48	7,00	2,52

leicht auch „per Hand" durchführen. Viele kontrollierte Einzelfallstudien sind jedoch deutlich komplexer. Müsste man beispielsweise in einem multiplen Grundratenversuchsplan mit fünf Probandinnen und Probanden pro Versuchsperson 30 Permutationen berücksichtigen, so wären 30^5 = 24.300.000 Werte zu ermitteln. Nach Dugard, File und Todman (2012) stellt der enorme Aufwand, der im Zusammenhang mit der Analyse von Daten aus kontrollierten Einzelfallstudien im Rahmen von Randomisierungstests aufgewendet werden muss, den Hauptgrund dafür dar, warum diese Methodengruppe in der sonderpädagogischen Forschung bislang kaum Berücksichtigung gefunden hat. Zum Zeitpunkt der Erstellung dieses Artikels lagen erst zwei empirische Untersuchungen vor, die sich in Zeitschriften finden, welche in den Datenbanken PSYNDEX oder PsycINFO inventarisiert sind, und in denen dieses Konzept in diesem Kontext eingesetzt worden ist (T. E. Scruggs, persönliche Mitteilung, 08.06.2012). Es handelt sich hierbei um die Arbeiten von Mastropieri et al. (2009) sowie von Regan, Mastropieri und Scruggs (2005). Demgegenüber existieren jedoch etliche Dutzend Aufsätze, in denen die Anwendung von Randomisierungstests bei kontrollierten Einzelfallstudien beschrieben und meistens sogar „wärmstens" empfohlen wird (z. B. Edgington, 1996; Haardörfer & Gagné, 2010; Ferron & Sentovich, 2002; Ferron & Ware, 1994; Levin & Wampold, 1999; Manolov & Solanas, 2008; Manolov, Solanas, Bulté & Onghena, 2010).

Das neue Lehrbuch „Single-case and Small-*n* Experimental Design" von Dugard, File und Todman (2012) schließt vor diesem Hintergrund eine wichtige Lücke, weil es aufzeigt, wie diese Methodengruppe in der Praxis relativ problemlos zum Einsatz kommen kann. Die beiden Autorinnen und der Autor haben für das PC-Programm Microsoft Excel spezielle Makros entwickelt, mit deren Hilfe sich die Daten auch mittels eines einfachen Computers auswerten lassen. Auf der Homepage des Werkes finden sich Links zum Download dieser Befehlsfolgen (http://www.routledge.com/books/details/9780415886932/). Die Zeit, die ein herkömmlicher PC benötigt, um die Daten nach deren Eingabe zu analysieren, kann zwar je nach Komplexität der Untersuchung weit mehr als eine Stunde in Anspruch nehmen, allerdings hält sich der Auswertungsaufwand für die meisten Anwenderinnen und Anwender jetzt neuerdings zumindest in einem erträglichen Rahmen. Auf der besagten Internetseite sind darüber hinaus Links zu Syntaxen für IBM SPSS Statistics hinterlegt, so dass sich auch mit diesem Programm arbeiten lässt.

Bei dem Buch von Dugard, File und Todman (2012) handelt es sich um eine grundlegend überarbeitete Neuauflage des Werkes von Todman und Dugard (2001). Schon die erste Edition enthielt Werkzeuge zur Auswertung von Messwerten aus Einzelfallstudien mittels Randomisierungstests unter Rückgriff auf verschiedene PC-Programme (Minitab, RANDIBM, SAS, SCRT, StatXact, …). Die Möglichkeiten der Datenanalyse sind mit den neuen Makros und Syntaxen für Microsoft Excel bzw. IBM SPSS Statistics jedoch deutlich vielfältiger, komfortabler und ökonomischer als bei den vorherigen Optionen. Forscherinnen und Forscher können dadurch auch ohne tiefgreifende Spezialkenntnisse mit der ihnen vertrauten Software schnell und unkompliziert Messwerte aus Einzelfallstudien inferentiell auswerten.

Einwände gegen den Einsatz von Randomisierungstests bei der Auswertung von Einzelfallstudien

Unerheblichkeit der Ergebnisse im Hinblick auf die klinische Signifikanz von Behandlungserfolgen

Die Bedenken, die man gegen eine Anwendung von Randomisierungstests bei kontrollierten Einzelfallstudien vorbringen könnte, sind teils allgemeiner, teils spezieller Natur. Zunächst einmal wäre es möglich, den Sinn inferenzstatistischer Verfahren in diesem Zusammenhang ganz generell in Frage zu stellen. Kazdin (1984) tut dies aus der Perspektive eines Praktikers. Er unterscheidet zwischen klinischer und statistischer Signifikanz. Das erste Kriterium sollte bei Evaluationen nach Maßgabe der kontrollierten Einzelfallforschung seiner Auffassung nach als entscheidend angesehen werden. In der täglichen sonderpädagogischen Arbeit spiele es keine Rolle, ob die Veränderung eines Verhaltens statistisch bedeutsame Ausmaße erreichten oder nicht. Wesentlich seien vielmehr „soziale Standards". Es gehe um die Frage, wie Eltern, Lehrkräfte, Gleichaltrige oder andere jeweils relevante Personen ein im Fokus stehendes Verhalten im Verlauf bzw. nach einer Intervention bewerten würden. Allerdings spricht dieses Argument im Grunde nicht speziell gegen die Verwendung von inferenzstatistischen Methoden bzw. von Randomisierungstests in der kontrollierten Einzelfallforschung, sondern gegen den Nutzen empirischer Untersuchungen per se. Denn worin läge der Sinn, ein ausgeklügeltes Vorgehen zu planen, Variablen zu operationalisieren und in mehr oder weniger aufwändiger Weise Daten zu erheben, wenn als Kriterium für den Erfolg einer Maßnahme vornehmlich das subjektive Empfinden des Umfeldes dienen würde? Möchte man Behandlungserfolge möglichst gut objektivieren, kommt man um quantitative Forschungsmethoden (wie kontrollierte Einzelfallstudien) nicht herum.

Mangelnde Verallgemeinerungsmöglichkeiten der Befunde

Auch der nächste Einwand ist auf einer allgemeinen Ebene angesiedelt. Im Falle von Randomisierungstests ist er allerdings ganz besonders naheliegend: Inferenzstatistik bezeichnet man auch als schließende Statistik, weil ihr Hauptanliegen darin besteht, von Stichproben auf Populationen zu schließen. Man möchte also mittels eines geeigneten Tests abschätzen, ob Aussagen, die für eine untersuchte Untermenge aus einer Gesamtmenge gültig sind, eben auch auf diese Gesamtmenge zutreffen. Bei Randomisierungstests werden alle in den Daten vorhandenen Informationen verwertet. Am Ende erhält man dann einen exakten *p*-Wert. Diese Präzision erscheint im ersten Moment möglicherweise unangemessen oder gar sinnlos, weil auch mit den exaktesten Methoden auf Basis eines Einzelfalls natürlich keine Rückschlüsse auf eine Population zu ziehen sind. Es gilt hierbei jedoch zu bedenken, dass inferentielle Verfahren zwei Anliegen verfolgen: (1) Sie sollen eine Generalisierbarkeit von einer Stichprobe auf die Gesamtpopulation sicherstellen, die sich durch den Geltungsbereich bestimmt, den eine jeweilige Hypothese für sich in Anspruch nimmt und (2) sie sollen möglichst ausschließen, dass die beobachteten Unterschiede (oder Zusammenhänge) zufällig zustande gekommen sind. Das zweite Ziel lässt sich mittels Randomisierungstests zweifellos erreichen – beim ersten sind Grenzen gesetzt. Allerdings geht es hier um ein Problem, das auch bei den allermeisten Gruppenstudien eine Rolle spielt, in denen inferenzstatistische Verfahren zum Einsatz kommen. Beim Gros der Untersuchungen im sozialwissenschaftlichen Bereich werden so genannte anfallen-

de Stichproben verwendet. Das heißt, dass man diejenigen Versuchspersonen in die eigene Forschungsarbeit mit einbezieht, die gerade zur Verfügung stehen (z. B. Kinder aus bestimmten Tagesstätten, Mädchen und Jungen aus verschiedenen Schulen einer Region, Jugendliche aus speziellen Einrichtungen zur beruflichen Rehabilitation). In den seltensten Fällen ist es möglich, auf eine Zufallsstichprobe zurückzugreifen, bei der jede Teilnehmerin und jeder Teilnehmer die exakt gleichen Chancen gehabt hatte, aus einer meist unüberschaubaren Gesamtpopulation (alle Drittklässlerinnen und Drittklässler mit einer Lese-Rechtschreibstörung in Deutschland, alle hiesigen Förderschülerinnen und Förderschüler, alle Jugendlichen in einer beruflichen Vorbereitungsmaßnahme der Bundesagentur für Arbeit) ausgewählt zu werden. Dadurch ist eine Verallgemeinerung der Untersuchungsergebnisse schwierig. Um dieses Problem zu relativieren, behilft man sich in der Regel damit, eine Studie in verschiedenen Settings und mit verschiedenen Stichproben mehrfach zu replizieren.

Dieser Weg ist in der Forschung allerdings nicht nur Gruppenuntersuchungen vorenthalten. In den von Chambless et al. (1998) entwickelten Standards für empirisch-validierte Interventionsansätze ist festgelegt, dass eine Behandlungsmethode erst dann als wirksam bezeichnet werden darf, wenn ihre Effizienz durch mehrere anspruchsvolle und in hochklassigen Fachzeitschriften veröffentlichte Studien belegt ist. Handelt es sich bei diesen Arbeiten um Gruppenuntersuchungen, so sind mindestens zwei Publikationen nötig, im Falle von kontrollierten Einzelfallanalysen sind es mindestens neun. Im Übrigen ist an dieser Stelle darauf zu verweisen, dass bei Studien aus der zuletzt genannten Kategorie von Forschungsansätzen oftmals nicht nur jeweils eine einzige Probandin bzw. ein einziger Proband ins Blickfeld genommen wird. Bei den häufig eingesetzten multiplen Grundratenversuchsplänen über Personen liegt die empfohlene Mindestzahl an Teilnehmerinnen und Teilnehmern bei vier (Barlow & Hersen, 1984). Durch eine Serie von Replikation können die Möglichkeiten einer Verallgemeinerung der Befunde bei kontrollierten Einzelfallanalysen also leicht genauso groß sein wie bei Gruppenuntersuchungen.

Für Edgington und Onghena (2007) stellt der Umstand, dass es bei Randomisierungstets nicht darum geht, über eine Ziehung einer Zufallsstichprobe aus einer Grundgesamtheit allgemeine Rückschlüsse zu ziehen, keinen Nachteil dar – ganz im Gegenteil:

> … randomization tests are the ultimate nonparametric tests. To say that they are free from parametric assumptions is a gross understatement of their freedom from questionable assumptions – they are free from the most conspicuously incorrect assumptions of all, which is the assumption that subjects … were randomly drawn from a population (S. 1).

Nach Auffassung der beiden Autoren basieren Randomisierungstests also auf einem viel „ehrlicheren" und realitätsangemesseneren Ansatz als statistische Tests auf Basis der Theorie von Neyman und Pearson (1933).

Negative Auswirkungen einer Autokorrelation der Residualwerte auf die Teststärke

Ein weiterer (potenzieller) Kritikpunkt betrifft die Gefahr eines zu hohen Zusammenhangs zwischen den mit den Werten einer Zeitreihe verbundenen Residuen. Dieses Phänomen ist bereits weiter oben thematisiert und als wesentlicher Grund dafür angeführt worden, warum v. a. parametrische Verfahren für die Auswertung von Daten aus kontrollierten Einzelfallstudien in den allermeisten Situationen ungeeignet sind. Ab wann eine Autokorrelation

als „zu hoch“ gilt, ist nicht verbindlich festgelegt. Liegen die Werte jedoch außerhalb der Grenzen zwischen $r = –0,3$ und $r = 0,3$, so werden sie in vielen einschlägigen Veröffentlichungen als heikel angesehen. Campbell und Herzinger (2010) stellen zu Recht heraus, dass das Problem der seriellen Abhängigkeit von Residuen bei so gut wie allen inferenzstatistischen Verfahren eine Rolle spielt – also möglicherweise auch bei der Auswertung von Daten aus Einzelfallstudien mit Hilfe von Randomisierungstests. Nach Sierra, Solanas und Quera (2005) besteht diese potenzielle Gefahr v. a. dann, wenn die Anzahl der Messungen pro Phase unter 12 liegt. Inwiefern sich eine hohe Autokorrelation nun jedoch tatsächlich auch bei Randomisierungstests in nicht mehr vertretbarem Ausmaß negativ auf deren Power auswirken kann, wurde im Rahmen verschiedener Monte Carlo Simulationsstudien überprüft.

Eine in diesem Zusammenhang zentrale Untersuchung stammt von Ferron und Ware (1995). Die beiden Autoren verwendeten in ihrer Arbeit vier verschiedene Formen von Randomisierungstests zur Auswertung von Werten aus einer hypothetischen Studie (multipler Grundratenversuchsplan AB) mit vier Probandinnen bzw. Probanden, 15 Messungen und einem Minimum von fünf Probes pro Phase. Im Ergebnis zeigte sich, dass die Power bei einer moderaten bis hohen Autokorrelation dann in beträchtlichem Maße leiden kann, wenn die Behandlungserfolge relativ gering ausfallen. Allerdings ist an dieser Stelle darauf hinzuweisen, dass die negativen Auswirkungen serieller Abhängigkeiten bei vielen multiplen Grundratenversuchsplänen insgesamt als relativ unerheblich zu bezeichnen sind (Ferron & Sentovich, 2002).

Sierra, Solanas und Quera (2005) überprüften im Rahmen ihrer Simulationsstudien die Validität eines von Levin, Marascuilo und Hubert (1978) entwickelten Randomisierungstests für ABAB-Designs und deren Abwandlungen. Auch hier wurde deutlich, dass die Power durchaus relativ stark beeinträchtigt sein kann, falls zwischen nebeneinanderliegenden Residualgrößen mittlere bis hohe serielle Abhängigkeiten bestehen. Bei manchen Designs (wie etwa bei einem simplen ABAB- oder BABA-Plan) ist es jedoch auch möglich, dass Autokorrelationen Größenordnungen von –0,6 bzw. 0,6 erreichen, ohne dass die Teststärke in bedenklicher Weise darunter leidet.

Insgesamt ist zu konstatieren, dass der von Kratochwill und Levin (1980) in einem viel zitierten Aufsatz vor über dreißig Jahren ausgestellte „Freifahrtsschein“ für Randomisierungstests, nach dem diese Verfahren auch bei hohen seriellen Abhängigkeiten der Daten problemlos einsetzbar sind, mit Blick auf die Ergebnisse mittlerweile vorliegender Simulationsstudien heute nicht uneingeschränkt „genehmigt“ werden kann. Es ist somit stets zu überprüfen, ob die Autokorrelationen der Residuen innerhalb einzelner Phasen außerhalb von noch tolerierbaren Grenzen liegen. Ist das nicht der Fall, so muss dies noch keine gravierend negativen Effekte auf die Power eines Randomisierungstests ausüben. Bei durchschnittlichen Leistungsverbesserungen von ca. 1,4 (Ferron & Ware, 1995) bzw. 2,0 Standardabweichungen (Ferron & Sentovich, 2002) richteten selbst serielle Abhängigkeiten von bis zu $r = \pm 0,5$ hier keinen nennenswerten „Schaden“ an. Auch mit Hilfe einer Verwendung bestimmter Designs oder eines Einbezugs von Mittelwerten als Teststatistiken lässt sich die Problematik in der Regel recht gut kontrollieren. Dennoch sind Randomisierungstests nicht grundsätzlich völlig frei von der Gefahr, durch zu hohe Autokorrelationen in ihrer Präzision beeinträchtigt zu werden.

Beschränkte Einsatzmöglichkeiten

Ein letzter relevanter Einwand betrifft die begrenzten Einsatzmöglichkeiten dieser

Verfahren im Hinblick auf geeignete Interventionen und Zielvariablen. Damit mit Hilfe eines Randomisierungstests ein statistisch bedeutsamer Unterschied zwischen verschiedenen Phasen ausgewiesen werden kann, ist es in aller Regel notwendig, dass ein Behandlungserfolg unmittelbar eintritt (bzw. abklingt). Wenn sich der Nutzen einer Förderung erst allmählich offenbart oder ein eventueller Zugewinn nicht direkt gemessen werden kann, wirft dies ernsthafte Schwierigkeiten auf. Derartige Fälle stellen bei der Evaluation sonderpädagogischer Maßnahmen allerdings keine Seltenheit dar. Möchte man beispielsweise die induktive Denkfähigkeit, die expressive Schreibkompetenz oder die Fertigkeit zum Lösen komplexer Sach- und Textaufgaben bei Kindern oder Jugendlichen verbessern, so erfordert dies Zeit. Man kann nicht erwarten, dass sich hier unmittelbar im Anschluss an die erste, zweite oder dritte Fördersitzung markante und schnell messbare Erfolge einstellen. Hier wäre es u. U. sinnvoll, bei der Auswertung der Daten von Randomisierungstests abzusehen.

Fazit

Die kontrollierte Einzelfallforschung scheint nach dem Ausklingen ihrer Hochphase in den 1980er Jahren nun in der jüngeren Vergangenheit eine kleine Renaissance zu erleben. Durchforstet man die Datenbank PsycINFO nach Zeitschriftenartikeln, in denen im Titel entweder der Ausdruck „single-case“ oder „single-subject“ vorkommt, so wird deutlich, dass allein in den vergangenen fünf Jahren (zwischen 2007 und 2011) ca. 80% so viele einschlägige Arbeiten veröffentlicht wurden wie in den zehn Jahren zuvor (also zwischen 1997 und 2006). Seit 2010 sind auffallend viele Standardlehrbücher zur Einzelfallforschung erschienen, die auch (bzw. ganz besonders) auf den sonderpädagogischen Bereich Bezug nehmen (z. B. Dugard, File & Todman, 2012; Gast, 2010; Huitema, 2011; Kazdin, 2010; O’Neill, McDonnell, Billingsley & Jenson, 2010). Der Frage nach geeigneten Auswertungsoptionen für Daten aus entsprechenden Untersuchungsdesigns wird in den allermeisten dieser Publikationen relativ viel Platz eingeräumt. Auch Randomisierungstests finden in diesem Zusammenhang oftmals lobend Erwähnung, während andere statistische Methoden teilweise sehr stark in der Kritik stehen. Umso mehr verwundert es, dass der Einsatz dieser Verfahren in der Forschungspraxis bislang so gut wie keine Rolle spielt. Natürlich sind Randomisierungstests nicht über jeden Zweifel erhaben: Auch sie betrifft die generelle Kritik am Sinn einer inferentiellen Analyse von Einzelfällen, auch ihre Power kann bei einer zu hohen Autokorrelation der Residualwerte u. U. nennenswert beeinträchtigt sein und auch sie eignen sich selbstverständlich nicht für alle Problemstellungen. Bewertet man Randomisierungstests allerdings vor dem Hintergrund der Vor- und Nachteile anderer Auswertungsoptionen, so schneiden sie insgesamt zweifellos gut bis sehr gut ab. Es ist somit zu wünschen, dass diese Gruppe von Methoden allmählich ihren Weg in die sonderpädagogische Forschungspraxis findet, um dadurch die Qualität und die Aussagekraft so mancher Einzelfallstudie zu erhöhen. Denn gerade im Hinblick auf die Güte entsprechender Arbeiten, die sich mit der wichtigen Frage befassen, was genau während einer Intervention geschieht und wie sie bei welchen Personen wirkt, liegt derzeit noch einiges im Argen.

Literatur

Allison, D.B. & Gorman, B.S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31*, 621–631.

Barlow, D.H. &Hersen, M. (1984). *Single case experimental designs: Strategies for studying*

behavior change. Needham Heights, MA: Allyn & Bacon.

Barnett, D.W., Daly, E.J. III, Jones, K.M.& Lentz, F.E. (2004). Response to intervention: Empirically based special service decisions from single-case designs of increasing and decreasing intensity. *The Journal of Special Education, 28*, 66–79.

Bloom, M. & Fisher, J. (1982).*Evaluating practice: Guidelines for the accountable professional*. Englewood Cliffs, NJ: Prentice Hall.

Borckardt, J.J. & Nash, M.R. (2002). How practitioners (and others) can make scientifically viable contributions to clinical-outcome research using the single-case time-series design. *International Journal of Clinical and Experimental Hypnosis, 50*, 114–148.

Bortz, J., Lienert, G.A. & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik*. Berlin: Springer.

Brossart, D.F., Parker, R.I., Olson, E.A. & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*, 531–563.

Busk, P.L. & Marascuilo, L.A. (1992). Statistical analysis in single-case research: Issues, procedures, and recommendations, with applications to multiple behaviors. In T.R. Kratochwill & J.R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 159–185). Hillsdale, NJ: Lawrence Erlbaum.

Busk, P.L. & Serlin, R. (1992). Meta-analysis for single case research. In T.R. Kratochwill & J.R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Lawrence Erlbaum.

Campbell, J. M. (2003). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: A quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 24*,120–138.

Campbell, J.M. &Herzinger, C.V. (2010). Statistics and single subject research methodology. D.L. Gast (Ed.), *Single subject research design in behavioral sciences* (pp. 417–453). Hillsdale, NJ: Lawrence Erlbaum.

Chambless, D.L., Baker, M.J., Baucom, D.H., Beutler, L.E., Calhoun, K.S., Crits-Christoph, P. et al. (1998). Update on empirically validated therapies. *The Clinical Psychologist, 51*, 3–16.

Dugard, P., File, P. & Todman, J. (2012). *Single-case and small-n experimental designs: A practical guide to randomization tests*. New York: Routledge.

Edigington, E.S. (1969). *Statistical inference: The distribution-free approach*. New York: McGraw-Hill.

Edgington, E.S. (1980). *Randomization tests*. New York: Marcel Dekker

Edgington, E.S. (1996). Randomized single-subject experimental designs.*Behaviour Research and Therapy, 34*, 567–574.

Edgington, E.S. & Onghena, P. (2007). *Randomization tests*. Boca Raton, FL: Chapman & Hall.

Edwards, W., Lindman, H. & Savage, L.J. (1963). Bayesian statistical inference for psychological research.*Psychological Review, 70*, 193–242.

Efron, B. & Tibshirani, R.J. (1993). An introduction to the bootstrap.New York: Chapman and Hall.

Ferron, J. & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *Journal of Experimental Education, 70*, 165–178.

Ferron, J. & Ware, W. (1994). Using randomization tests with responsive single-case designs. *Behaviour Research and Therapy, 32*, 787–791.

Ferron, J., & Ware, W. (1995). Analyzing single-case data: The power of randomization tests. *The Journal of Experimental Education, 63*, 167–178.

Fisher, R.A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture for Great Britain, 33*, 503–513.

Fisher, R.A. (1934). The logic of inductive inference. *Journal of the Royal Statistical Society, 98*, 39–54.

Gast, D.L. (2010). *Single subject research methodology in behavioral sciences*. New York: Routledge.

Gast, D.L. & Spriggs, A.D. (2010). Visual analysis of graphic data. In D.L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 199–233). New York: Routledge.

Grünke, M. & Masendorf, F. (2000). Experimentelle Interventionsforschung in Gruppen. In J. Borchert (Hrsg.), *Handbuch der Sonderpä-*

dagogischen Psychologie (S. 974–986). Göttingen: Hogrefe.

Grund, M. (2012). *GUT 1: Rechtschreibtraining für Klasse 2–6*. Baden-Baden: GUT.

Haardörfer, R. & Gagné, P. (2010). The use of randomization tests in single-subject research. *Focus on Autism and Other Developmental Disabilities, 25*, 47–54.

Hersen, M. & Barlow, D.H. (1976). *Single-case experimental designs: Strategies for studying behavior change*. Oxford, UK: Pergamon.

Horner, R.H., Carr, E.G., Halle, J., McGee, G., Odom, S. & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165–179.

Huitema, B.E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7*, 107–118.

Huitema, B.E. (1986). Statistical analysis and single-subject designs: Some misunderstandings. In A. Poling & R.W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 209–232). New York: Plenum.

Huitema, B.E. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies*. New York: Wiley.

Julius, H., Schlosser, R.W. & Goetze, H. (2000). *Kontrollierte Einzelfallstudien: Eine Alternative für die sonderpädagogische und klinische Forschung*. Göttingen: Hogrefe.

Kazdin, A. (1984). Statistical analysis for single-case experimental designs. In M. Hersen & D.H. Barlow (Eds.), *Single case experimental designs: Strategies for studying behavior change* (pp. 265–317). New York: Pergamon.

Kazdin, A. (2010). *Single-case research designs: Methods for clinical and applied settings*. Oxford, UK: Oxford University Press.

Kern, H.J. (1996) Einzelfallforschung in der (Sonder) Pädagogik: Multiple-Grundraten-Versuchspläne. *Heilpädagogische Forschung, 22*, 131–141

Kern, H.J. (1997a) Einzelfallforschung: Versuchsplan-Kombinationen für die (Sonder)Pädagogik. *Vierteljahreszeitschrift für Heilpädagogik und ihre Nachbargebiete, 66*, 325–336

Kern, H.J. (1997b). *Einzelfallforschung: Eine Einführung für Studierende und Praktiker*. Weinheim: Beltz.

Köhler, T. (2008). *Statistische Einzelfallanalyse: Eine Einführung mit Rechenbeispielen*. Weinheim: Beltz.

Kratochwill, T.R. & Levin, J.R. (1980).On the applicability of various data analysis procedures to the simultaneous and alternating treatment designs in behavior therapy research.*Behavioral Assessment, 2*, 353–360.

Levin, J.R., Marascuilo, L.A. & Hubert, L.J. (1978). N = Nonparametric randomization tests. In T.R. Kratochwill (Ed.), *Single-subject research: Strategies for evaluating change* (pp. 167–196). New York: Academic Press.

Levin, J.R. &Wampold, B.E. (1999).Generalized single-case randomization tests: Flexible analyses for a variety of situations. *School Psychology Quarterly, 14*, 59–93.

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

Manolov, R., Arnau, J., Solanas, S. & Bono, R. (2010).Regression-based techniques for statistical decision making in single-case designs. *Psicothema, 22*, 1026–1032.

Manolov, R. & Solanas, A. (2008).Randomization tests for ABAB designs: Comparing data-division-specific and common distributions. *Psicothema, 20*, 291–297.

Mastropieri; M.A., Scruggs, T.E., Mills, S., Irby Cerar, N., Cuenca-Sanchez, Y., Allen-Bronaugh, D. et al. (2009). Persuading students with emotional disabilities to write fluently. *Behavioral Disorders, 35*, 19–40.

Neyman, J. & Pearson, E.S. (1933).On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, 231*, 289–337.

O'Neill, R.E., McDonnell, J.J., Billingsley, F. & Jenson, W. (2010). *Single case research designs in educational and community settings*. Upper Saddle River, NJ: Prentice Hall.

Orme, J.G. & Cox, M.E. (2001). Analyzing single-subject design data using statistical process control charts. *Social Work Research, 25*, 115–127.

Ottenbacher, K.J. (1993). Interrater agreement of visual analysis in single-subject decisions: Quantitative review and analysis. *American Journal of Mental Retardation. 98*, 135–142.

Parker, R.I. & Brossart, D.F. (2003). Evaluating single-case research data: A comparison of

seven statistical methods. *Behavior Therapy, 34*, 189–211.
Parsonson, B.S. & Baer, D. M. (1986).The graphic analysis of data. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 157–186). New York: Plenum.
Regan, K.S., Mastropieri, M.A. & Scruggs, T.E. (2005).Promoting expressive writing among students with emotional and behavioral disturbance via dialogue journals. *Behavioral Disorders, 31*, 33–50.
Saint-Mont, U. (2011). *Statistik im Forschungsprozess: Eine Philosophie der Statistik als Baustein einer integrativen Wissenschaftstheorie*. Berlin: Springer.
Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.
Schindele, R.A. (1985). Research methodology in special education: A framework approach to special problems and solutions. In S. Hegarzy & P. Evans (Eds.), *Research and evaluation methods in special education* (pp. 3–24). Windsor: NFER-Nelson.
Scotti, J.R., Evans, I.M., Meyer, L.H. & Walker, P. (1991). A meta-analysis of intervention research with problem behavior: Treatment validity and standards of practice. *American Journal on Mental Retardation, 96*, 233–256.
Scruggs, T.E., Mastropieri, M.A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*, 24–33.
Sierra, V., Solanas, S. & Quera, V. (2005). Randomization tests for systematic single-case designs are not always appropriate. *The Journal of Experimental Education, 73*, 140–160.
Stock, C. & Schneider, W. (2008). *Deutscher Rechtschreibtest für das dritte und vierte Schuljahr (DERET 3-4+)*. Göttingen: Hogrefe.
Suen, H. K., & Ary, D. (1987). Autocorrelation in applied behavior analysis: Myth or reality? *Behavioral Assessment, 9*, 125–130.
Thomé, G. & Thomé, D. (2009). *OLFA Oldenburger Fehleranalyse: Instrument und Handbuch zur Ermittlung der orthographischen Kompetenz aus freien Texten ab Klasse 3 und zur Qualitätssicherung von Fördermaßnahmen*. Oldenburg: Igel.
Todman, J.B. & Dugard, P. (2001). *Single-case and small-n experimental designs*. New York: Routledge.
Tyron, W.W. (1982). A simplified time-series analysis for evaluating treatment interventions. *Journal of Applied Behavior Analysis, 15*, 423–429.
Universität Leipzig (2012). Wortlisten. URL: http://wortschatz.uni-leipzig.de/html/wliste.html (Stand: 18.09.2012).
Van den Noortgate, W., & Onghena, P. (2003). Combining single-case experimental data using hierarchical linear models.*School Psychology Quarterly, 18*, 325–346.
Wember, F.B. (1994). Möglichkeiten und Grenzen der empirischen Evaluation sonderpädagogischer Interventionen in quasi-experimentellen Einzelfallstudien. *Heilpädagogische Forschung, 20*, 99–117.

Anschrift des Autors

Prof. Dr. Matthias Grünke
Universität zu Köln
Department Heilpädagogik & Rehabilitation
Klosterstraße 79b
50931 Köln
matthias.gruenke@uni-koeln.de