

Komplementäre Studiendesigns zur Evidenzbasierung in der Bildungswissenschaft

Pawel R. Kulawiak, Jürgen Wilbert

- 2.1 Evidenzbasierung in den Bildungswissenschaften – 18**
- 2.2 Experimentelle Forschung und Wirksamkeit – 20**
 - 2.2.1 Kontrafaktisches Modell – 21
 - 2.2.2 Interne Validität – 21
 - 2.2.3 Vorzüge und Limitationen der experimentellen Forschung – 22
 - 2.2.4 Externe Validität – 23
 - 2.2.5 Erforschung des Wirkmechanismus – 23
- 2.3 Komplementäre Forschungsdesigns zur Wirksamkeitsforschung – 24**
 - 2.3.1 Single-Case-Designs (Einzelfallforschung) – 25
 - 2.3.2 Nicht-experimentelle Forschungsdesigns – 28
- Literatur – 30**

2.1 Evidenzbasierung in den Bildungswissenschaften

Der vorliegende Beitrag möchte ausgehend von der Darlegung zweier bildungswissenschaftlicher Evidenzbeurteilungsrichtlinien die aktuelle Diskussion um den experimentellen Wirksamkeitsnachweis nachzeichnen. Dabei wird insbesondere auf die Bedeutung des Wirksamkeitsbegriffs eingegangen und verdeutlicht, dass mittels unterschiedlicher Forschungsansätze auch unterschiedliche sowie komplementäre Dimensionen (bzw. Qualitäten) der Wirksamkeit pädagogischer Handlungen erforscht werden können. Dabei kann aus pädagogischer Sicht der Begriff der evidenzbasierten Praxis wie folgt aufgefasst werden:

- » The term evidence-based practices represents a systematic approach to determining which research-based practices are supported by a sufficient number of research studies that (a) are of high methodological quality, (b) use appropriate research designs that allow for assessment of effectiveness, and (c) demonstrate meaningful effect sizes such that they merit educators' trust that the practice works.

(Cook et al. 2012, S. 495)

Der wissenschaftliche Nachweis über die Wirksamkeit (Effektivität) einer pädagogischen Handlung stellt demnach eine zentrale Handlungsmaxime der Evidenzbasierung dar. Die effektive Förderung eines Kindes (z. B. beim Erwerb des Schreibens oder dem Erlernen sozialer Kompetenzen) ist daher das Ziel einer evidenzbasierten pädagogischen Maßnahme (z. B. die Anwendung einer Unterrichtsmethode oder die Durchführung einer Lerntherapie). Dabei wird seit über einer Dekade die Kritik geäußert, dass pädagogische Handlungen vielfach eben nicht auf wissenschaftlichen Wirksamkeitsnachweisen, sondern auf tradierten Praktiken sowie ideologischen Überzeugungen basieren (Kavale und Mostert 2003; Slavin 2002). Wie schon deutlich früher in anderen Disziplinen (z. B. der Medizin), wurde daher auch in pädagogischen Handlungsfeldern der Ruf nach evidenzbasierten Handlungsempfehlungen deutlich (Slavin 2002). Befürworter der Evidenzbasierung plädieren

für eine stärkere empirische und vor allem experimentelle Ausrichtung der pädagogischen Forschung (Slavin 2002).

Hieraus ergibt sich die Frage, wie die Wirksamkeit einer pädagogischen Handlung festgestellt werden kann bzw. welche Kriterien zur Beurteilung der Wirksamkeit einer pädagogischen Handlung herangezogen werden sollten. Gemäß dem obigen Zitat können mehrere Anforderungen an den wissenschaftlichen Wirksamkeitsnachweis gestellt werden: die Anzahl empirischer Studien (die eine Wirksamkeit nachweisen), die methodologische Qualität der Studien, die Angemessenheit der Studiendesigns bezüglich der Aufklärung der Wirksamkeit sowie die Stärke des Wirksamkeitsnachweises in Form von statistischen Effektstärken.

Eine häufig angewandte Evidenzbeurteilungsrichtlinie, welche für den Einsatz in pädagogischen Handlungsfeldern entwickelt wurde, ist die *Best Evidence Encyclopedia* (BEE). Hinsichtlich der Beurteilung der empirischen Befunde weist die BEE-Richtlinie eine klare hierarchische Ordnung auf (von höchste bis limitierte Evidenzgüte):

- **„Strong Evidence of Effectiveness:** At least one large randomized or randomized quasi-experimental study and one additional large qualifying study, or multiple smaller studies, with a combined sample size of 500 and an overall weighted mean effect size of at least +0.20.
- **Moderate Evidence of Effectiveness:** Two large matched studies, or multiple smaller studies with a collective sample size of 500 students, with a weighted mean effect size of at least +0.20.
- **Limited Evidence of Effectiveness: Strong Evidence of Modest Effects:** Studies meet the criteria for 'Moderate Evidence of Effectiveness' except that the weighted mean effect size is +0.10 to +0.19.
- **Limited Evidence of Effectiveness: Weak Evidence with Notable Effect:** A weighted mean effect size of at least +0.20 based on one or more qualifying studies insufficient in number or sample size to meet the criteria for 'Moderate Evidence of Effectiveness.' (Best Evidence Encyclopedia 2017)

Diese Hierarchie der Wirksamkeitsevidenz ergibt sich, wie auch schon teilweise im Eingangszitat

dargelegt, aus der Art des Studiendesigns (vom randomisierten kontrollierten Experiment bis zum Quasiexperiment mit gematchten Untersuchungsgruppen), aus der Anzahl der durchgeführten Studien sowie aus der Anzahl der untersuchten Studienteilnehmer (über alle Studien hinweg) und aus der Größe der vorgefundenen Effektstärke (ebenfalls über alle Studien hinweg).

Eine zweite Evidenzbeurteilungsrichtlinie, welche ebenfalls für den Einsatz in pädagogischen Handlungsfeldern entwickelt worden ist, wurde vom Institute of Education Sciences vorgelegt: das What Works Clearinghouse (WWC). Das WWC formuliert hinsichtlich der Evidenzbeurteilung empirischer Befunde folgende Zielsetzung: „to be a central and trusted source of scientific evidence for what works in education“ (Institute of Education Sciences 2014, S. 1).

Somit wird, ebenso wie bei der BEE, die Wirksamkeitsfrage („what works“) in den Mittelpunkt der Evidenzbasierung gerückt. Das WWC sieht für die Beurteilung der Evidenzgüte eines Wirksamkeitsbefundes ebenfalls eine hierarchische Richtlinie vor. Im Fokus dieser Beurteilung steht unter anderem das Studiendesign (■ Abb. 2.1). Auch hier wird dem randomisierten kontrollierten Experiment die höchste Evidenzgüte zugeschrieben („without reservations“). Jedoch führt eine hohe Ausfallquote (z. B. aufgrund von systematischen Studienabbrüchen) zu einer Herabstufung der Evidenzgüte des randomisierten kontrollierten Experiments („with reservations“). Nicht-randomisierte Experimente mit vergleichbaren Untersuchungsgruppen (vergleichbar hinsichtlich der Baseline-Charakteristika) genießen ebenfalls diese eingeschränkte Evidenzgüte. Sofern in

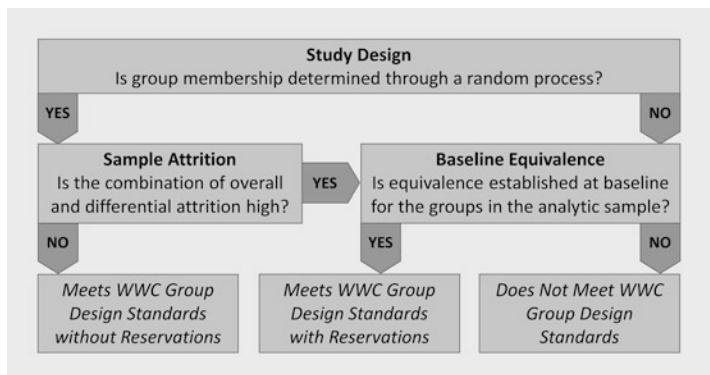
nicht-randomisierten Experimenten die Vergleichbarkeit der Untersuchungsgruppen nicht gegeben ist, können diese Studien auch nicht zur Evidenzbeurteilung herangezogen werden.

In der WWC-Richtlinie werden neben der Vergleichbarkeit der Untersuchungsgruppen und der Untersuchungsgruppenverzerrung aufgrund von systematischen Ausfällen auch weitere Evidenzbeurteilungskriterien berücksichtigt. Erhebliche Unterschiede zwischen den gruppenspezifischen Untersuchungsumständen („confounder“) sowie eine mangelnde Validität und Reliabilität hinsichtlich der Messung des Zielkriteriums können ebenfalls eine Herabstufung der Evidenzgüte nach sich ziehen. Erst die Gesamtschau aller Beurteilungskriterien ermöglicht die Zuschreibung der Wirksamkeitsevidenz zu einem konkreten empirischen Befund:

- » These ratings relate to the amount of confidence the WWC places in the ability of the study to generate an unbiased estimate of the causal relationship between the intervention and the outcomes of interest. (Institute of Education Sciences 2014)

Auch mit dieser Aussage wird die Wirksamkeitsfrage bzw. die Frage nach dem Kausalzusammenhang zwischen Intervention und Zielkriterium in den Fokus der Evidenzbasierung gerückt. Unter Berücksichtigung aller beurteilten Studien, der statistischen Signifikanz der Befunde sowie der Relevanz und Richtung (positiv/negativ) der Effektstärken wird die endgültige Wirksamkeitsbeurteilung der pädagogischen Maßnahme vorgenommen. Anhand der

■ **Abb. 2.1** Studienbeurteilung des WWC. (Aus Institute of Education Sciences 2014 mit freundlicher Genehmigung)



hierarchischen WWC-Richtlinie kann die Wirksamkeit einer pädagogischen Maßnahme nicht nur als positiv oder negativ (pädagogische Maßnahme ist der Vergleichsmaßnahme unterlegen), sondern auch als nicht erkennbar (weder positiv noch negativ) oder gemischt (sowohl positiv als auch negativ) bewertet werden. Die WWC-Richtlinie ist nachfolgend auszugswise wiedergegeben:

- **„Positive effects: Strong evidence of a positive effect with no overriding contrary evidence:** Two or more studies show statistically significant positive effects, at least one of which meets WWC group design standards without reservations, AND no studies show statistically significant or substantively important negative effects.
- **Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence:** At least one study shows statistically significant or substantively important positive effects, AND fewer or the same number of studies show indeterminate effects than show statistically significant or substantively important positive effects, AND no studies show statistically significant or substantively important negative effects.
- **No discernible effects: No affirmative evidence of effects:** None of the studies shows statistically significant or substantively important effects, either positive or negative.
- ...“
(Institute of Education Sciences 2014, S. 29)

Das Ausmaß der Evidenz („extent of evidence“) wird in der WWC-Richtlinie zusätzlich durch die Gesamtzahl der Studien und die Gesamtzahl der untersuchten Kinder (Quantität) sowie durch die Settingvariation (unterschiedliche Schulen oder Schulbezirke) über alle Studien hinweg festgelegt. Auch der Aspekt der Validität des Wirksamkeitsbefundes wird adressiert, wobei sowohl eine hohe interne als auch eine hohe externe Validität Indikatoren für eine hohe Evidenzgüte darstellen.

In der WWC-Richtlinie finden nebst den klassischen experimentellen und quasiexperimentellen Designs aber auch zwei weitere Studiendesigns „pilotartig“ Beachtung. Die Befunde dieser Studiendesigns fließen zurzeit allerdings nicht in die endgültige Wirksamkeitsbeurteilung mit ein. (I) Beim

Regressions-Diskontinuitäts-Design (Shadish et al. 2002) werden die Kinder nicht zufällig, sondern auf Basis eines Cutoff-Wertes (z. B. oberhalb/unterhalb der Durchschnittsleistung) den Interventionsgruppen zugeteilt. Dies führt zwangsläufig zu sehr unterschiedlichen Untersuchungsgruppen (Unterschiede hinsichtlich der Baseline-Charakteristika). Jedoch kann mit den experimentellen Daten anhand eines Bruchs der Regressionsgeraden (um den Cutoff-Wert) auf die Wirksamkeit einer Intervention geschlossen werden. (II) Single-Case-Designs/Einzelfall-Designs (Wilbert and Grünke 2015) beruhen nicht auf experimentellen Vergleichsgruppen. Stattdessen steht die intraindividuelle Entwicklung einzelner Kinder im Fokus der Einzelfallforschung. Das Zielkriterium wird bei den Kindern über einen bestimmten Zeitraum wiederholt (sowie häufig) gemessen (z. B. intraindividuelle Lernverläufe über ein Schulhalbjahr). Eine Intervention teilt eine intraindividuelle Messreihe in zwei oder mehrere Phasen (z. B. vor der Intervention/nach der Intervention). Die Phasen können nun dahingehend miteinander verglichen werden, ob mit der Intervention eine Veränderung des Zielkriteriums einhergehend ist. Anhand einer entsprechenden Veränderung kann auf die Wirksamkeit der Intervention geschlossen werden. Diesbezüglich empfiehlt die WWC-Richtlinie eine visuelle Inspektion der experimentellen Verlaufsdaten (mindestens 20 Einzelfälle über alle Studien hinweg).

Zwar wird innerhalb der deutschen Bildungswissenschaft das Paradigma der Evidenzbasierung kritisch diskutiert (Jornitz 2009; Schrader 2014; Hartmann et al. 2016), jedoch haben sich für das deutsche Bildungssystem bisher keine etablierten Evidenzbeurteilungsrichtlinien herauskristallisiert.

2.2 Experimentelle Forschung und Wirksamkeit

Gemäß den dargelegten Evidenzbeurteilungsrichtlinien ist die experimentelle Forschung (insbesondere das randomisierte kontrollierte Experiment) Dreh- und Angelpunkt des empirischen Wirksamkeitsnachweises. Hinsichtlich der Erforschung der Wirksamkeit (bzw. der Erforschung kausaler Zusammenhänge) gilt das randomisierte kontrollierte Experiment schließlich als der **Goldstandard**.

Die prominente Rolle des randomisierten kontrollierten Experiments in der Wirksamkeitsforschung äußert sich auch im folgenden Lehrbuchzitat:

- » Der einzige Weg, kausale Beziehungen aufzuzeigen, besteht in der experimentellen Methode, die sich definiert als eine Methode, in welcher der Forscher auf Zufallsbasis Teilnehmern Versuchsbedingungen zuteilt und sicherstellt, dass diese Situationen identisch sind außer der unabhängigen Variable (die Variable, von der angenommen wird, dass sie eine kausale Wirkung auf die Reaktionen der Menschen hat).¹
(Aronson et al. 2004, S. 43).

Im obigen Zitat wird die Potenz, kausale Zusammenhänge aufzuzeigen, ausschließlich dem randomisierten kontrollierten Experiment zugeschrieben. Verschiedene Autoren widersprechen dieser strikten Sichtweise und schreiben auch anderen Forschungsdesigns eine Aussagekraft hinsichtlich der Wirksamkeit einer pädagogischen Maßnahme zu (Berliner 2002; Briggs 2008; Cook et al. 2009). Evidenzbeurteilungsrichtlinien wie die zuvor dargestellte BEE bzw. das WWC sind dementsprechend mit Kritik behaftet (Briggs 2008; Cook et al. 2012): Einerseits bieten solche Systeme klare Standards zur Beurteilung der Evidenzgüte von Forschungsergebnissen, andererseits bilden sie ein striktes Regelwerk, das eine fragwürdige Hierarchie der verschiedenen Forschungsdesigns aufstellt. Diese Hierarchie vernachlässigt die Möglichkeit, dass unterschiedliche Forschungsdesigns komplementäre Informationen zur Generalisierbarkeit der Wirkung einer pädagogischen Maßnahme liefern können. Nachfolgend sollen daher die kontroverse Diskussion um den experimentellen Wirksamkeitsnachweis nachgezeichnet sowie die Vorzüge und Limitationen der experimentellen

Forschung für die Wirksamkeitsbeurteilung pädagogischer Maßnahmen dargestellt werden. Als Ausgangspunkte dieser Darstellung sollen hier das randomisierte kontrollierte Experiment sowie das kontrafaktische Modell dienen.

2.2.1 Kontrafaktisches Modell

Vier Studiendesignelemente bilden die Basis eines randomisierten kontrollierten Experiments (Bosch et al. 2016): **Manipulation** (Gestaltung von Untersuchungsbedingungen durch willentliche Veränderung der Untersuchungsumwelt; inkl. Intervention), **Randomisierung** (Zufallszuteilung der Kinder zu den Untersuchungsbedingungen), **Temporalität** (wiederholte Messung des Zielkriteriums; vor sowie nach der Intervention) und **Kontrolle** (Vergleich von Untersuchungsgruppen). Mittels experimenteller Forschung kann die Wirksamkeit einer Intervention auf Grundlage der Logik eines kontrafaktischen Schlusses nachgewiesen werden. Dabei werden die Messungen in der Interventionsgruppe und der Kontrollgruppe als alternative Wirklichkeiten aufeinander bezogen: Wenn die Kinder der Interventionsgruppe statt der Lerntherapie keine Intervention erfahren hätten (Kontrollgruppe), dann hätte sich die Lesekompetenz dieser Kinder (Interventionsgruppe) genauso entwickelt wie die Lesekompetenz der Kinder in der Kontrollgruppe (und vice versa).

Anhand dieser kontrafaktischen Aussage wird der Lesekompetenzunterschied zwischen der Interventionsgruppe und der Kontrollgruppe als ein auf die Lerntherapie zurückzuführender Kausaleffekt interpretiert. Bei einem vorliegenden Gruppenunterschied (zugunsten der Interventionsgruppe) ist die Lerntherapie demnach als wirksam zu erachten.

2.2.2 Interne Validität

Die Validität einer Wirksamkeitsaussage wird maßgeblich von den einzelnen Studiendesignelementen des Experiments determiniert. So verfolgt man beispielsweise mittels Randomisierung das Ziel, dass beobachtete sowie unbeobachtete Störvariablen (also Variablen, die selbst in einer Kausalbeziehung zum Zielkriterium stehen) über die

1 In Erweiterung des bis hierhin vorrangig verwendeten Begriffs der Wirksamkeit verwendet das Zitat den Begriff „Kausalität“. Die Termini Wirksamkeit und Kausalität sind in gewisser Weise redundant. Eine wirksame (effektive) Intervention setzt voraus, dass das Interventionsergebnis ursächlich auf die Intervention zurückzuführen ist, d. h., dass es einen Kausalzusammenhang zwischen Intervention (Ursache) und Interventionsergebnis (Wirkung) gibt.

Untersuchungsgruppen hinweg annähernd identisch verteilt sind. Zudem ist man bei der Durchführung eines Experiments um die Gestaltung standardisierter Untersuchungsbedingungen bemüht, z. B. durch einheitliche Instruktion, Ausschaltung oder Konstanthaltung von Störfaktoren.

Eine Gewissheit darüber, dass die Veränderung im Zielkriterium tatsächlich die Folge der Intervention ist, wird als **interne Validität** bezeichnet. Können einzelne Elemente des Designs eines Experimentes nicht berücksichtigt werden (z. B. die Randomisierung), geht dies mit einer verringerten internen Validität und somit mit einer geringeren Sicherheit bzgl. der Wirksamkeitsaussage einher.

2.2.3 Vorzüge und Limitationen der experimentellen Forschung

Randomisierte kontrollierte Experimente adressieren vielfach die Wirksamkeit einiger weniger Faktoren. In solchen Fällen werden zumeist **simple** und **präzise** Fragestellungen untersucht (Cook 2002), z. B.: Begünstigt Kaugummikauen die kognitive Leistungsfähigkeit? (Rost et al. 2010). Ein entsprechendes Experiment lässt sich verhältnismäßig leicht implementieren. Die Studienteilnehmer bearbeiten einen kognitiven Leistungsfähigkeitstest, wobei die Hälfte der Personen dabei Kaugummi kaut. Dem aus der experimentellen Untersuchung abgeleiteten Kausalzusammenhang kann aufgrund der Randomisierung, der Eindeutigkeit der Intervention, der Durchführung von Einzeltestungen in einer standardisierten Untersuchungsumgebung (in Form eines störungsarmen Untersuchungsraums) eine hohe interne Validität zugesprochen werden.

Die Studiendesignelemente des randomisierten kontrollierten Experiments können eine im hohen Maße standardisierte und somit artifizielle Untersuchungsumwelt erzeugen, die nicht zwangsläufig reale pädagogische Situationen widerspiegelt. Im realen Schulalltag befinden sich die Schüler und Schülerinnen stets in ökologischen Systemen (Schulklasse, Unterricht, Peergruppe) und interagieren mit diesen Systemen (Kind-Lehrkraft-Interaktion, Kind-Kind-Interaktion, Lehrkraft-Peergruppe-Interaktion) (Bronfenbrenner 1976). Auch die eigentlichen Lehr- und Lernprozesse sind in komplexe

ökologische Systeme eingebettet. Unterrichtsprozesse können daher nicht losgelöst von Eigenschaften der Kinder und der Lehrkräfte sowie deren Interaktionen untereinander betrachtet werden. Für die eigentliche Wirksamkeit einer pädagogischen Maßnahme kann daher eine nahezu unendliche Anzahl von Umweltfaktoren (z. B. Klassenkomposition, Lernmotivation, Verhaltensnormen, Einstellungen der Lehrkraft) eine Rolle spielen. Gerade aus diesem Grund ist man in einem Experiment darum bemüht, die Stör- und Umweltfaktoren auszuschalten oder konstant zu halten. Dies ermöglicht die Ableitung eines möglichst eindeutigen Kausalzusammenhangs.

Damit einhergehend stellt sich die Frage, ob ein aus der experimentellen Forschung hergeleiteter Kausalzusammenhang auch auf Situationen außerhalb der experimentellen Untersuchungsumwelt generalisierbar ist. Es ist möglich, dass sich eine im experimentellen Setting erfolgreich erprobte Methode in der pädagogischen Praxis als weniger effektiv erweist. Dies kann daran liegen, dass die im Experiment untersuchten Kinder nicht die Zielpopulation repräsentieren, d. h., nicht oder nur in Teilen mit der Schülerschaft von bestimmten Schulen vergleichbar sind. Insbesondere inklusive Schulklassen weisen ein hohes Maß an Heterogenität auf. Hingegen werden die Teilnehmer an Experimenten oft so ausgewählt, dass sie sich in ihren Eigenschaften eher ähneln. Diese Art der Teilnehmersauswahl kann den Einfluss von Störvariablen verringern. In solchen Fällen ist es möglich, dass die Effektivität (bzw. Ineffektivität) einer Intervention nur sehr eingeschränkt auf alle Schüler und Schülerinnen der Grundpopulation übertragen werden kann. Beispielsweise schlägt ein Verhaltenstraining nur bei Kindern mit einer sehr niedrigen Verhaltenskontrolle an, allerdings sind aufgrund eines Selektionsmechanismus nur Kinder mit einer mittleren bis hohen Verhaltenskontrolle in den Untersuchungsgruppen vertreten, und die Wirksamkeit des Trainings kann dementsprechend nicht aufgezeigt werden. Unter Beachtung dieser Aspekte sollte die bloße Wirksamkeitsfrage (Wirkt die Intervention?) erweitert werden (Kvernbeek 2016, S. 109): „What works for whom under what circumstances?“. Der Beantwortung dieser erweiterten Wirksamkeitsfrage sind unter der ausschließlichen Verwendung experimenteller Forschung Grenzen gesetzt.

2.2.4 Externe Validität

Die Aspekte der Generalisierbarkeit des Wirksamkeitsbefundes wird unter dem Konzept der externen Validität diskutiert (Shadish et al. 2002). Eine hohe externe Validität spricht dafür, dass ein vorgefundener Kausalzusammenhang auch auf Gegebenheiten außerhalb der Untersuchung übertragbar ist, also auf andere Personen (anderes Alter, andere Lernvoraussetzungen) und andere Settings (andere Lehrkräfte, andere Unterrichtsformen, andere Schulfächer). So ist beispielsweise nicht eindeutig klar, dass eine bestimmte Feedbackmethode die Leistungsmotivation fächerübergreifend bzw. aufgabenformatübergreifend steigern wird. Möglicherweise kommt die erhöhte Leistungsmotivation infolge sozialer Vergleichsprozesse lediglich bei einem bestimmten Aufgabenformat sowie nur bei Kindern mit einer hohen Leistungszielorientierung zum Vorschein. Je stärker die Untersuchungsumwelt sowie die Untersuchungstichprobe der tatsächlichen pädagogischen Realität entspricht, desto höher ist auch die externe Validität der Wirksamkeitsaussage, d. h., es besteht ein höheres Maß an Sicherheit darüber, dass der vorgefundene Kausalzusammenhang auch außerhalb der Untersuchungsumwelt Gültigkeit besitzt.

Aufgrund der standardisierten Umgebung innerhalb eines Experimentes ist die externe Validität der hier gefundenen Befunde bedroht. Interne und externe Validität stehen dabei in einem gegenläufigen Verhältnis zueinander. Maßnahmen, die die externe Validität einer Wirksamkeitsaussage steigern (Randomisierung, Standardisierung), können zu einer artifiziellen Untersuchungsumwelt und zu einer Minderung der externen Validität der Wirksamkeitsaussage führen. Wenn die Untersuchungsumwelt wiederum in einem hohen Maße einem realen pädagogischen Setting gleicht, dann besteht ein hohes Maß an Sicherheit darüber, dass der Kausalzusammenhang auf vergleichbare pädagogische Situationen übertragbar ist (hohe externe Validität), aber ein geringes Maß an Sicherheit darüber, dass der Kausalzusammenhang tatsächlich auf die Manipulation und nicht auf Umwelt- oder Störbedingungen zurückzuführen ist (geringe interne Validität).

Beide Validitätskriterien (intern und extern) lassen sich in einer individuellen Studie nur schwer maximieren. So sind beispielsweise die Lehrkräfte in

inklusive Schulklassen darauf angewiesen, binnendifferenziert zu unterrichten. Der standardisierten Umsetzung von Unterrichtseinheiten unter der Voraussetzung der Zufallszuteilung der Schulkinder zu Interventionsgruppen sind dadurch Grenzen gesetzt.

2.2.5 Erforschung des Wirkmechanismus

Neben der hohen internen und externen Validität des Wirksamkeitsbefundes ist ebenso relevant, dass eine Untersuchung zum Verständnis des zugrunde liegenden Mechanismus der Wirksamkeit beiträgt. Viele pädagogische Maßnahmen (z. B. eine bestimmte Unterrichtsmethode) setzen sich aus verschiedenen kleineren, oft miteinander verwobenen Elementen zusammen (z. B. Instruktionen, Art der sozialen Lernform, Aufgabenmaterial, Feedback). Mehrere Ursachen können demnach die gewünschte Wirkung herbeiführen. Die experimentelle Logik macht jedoch zunächst lediglich eine Aussage darüber, ob die Gesamtheit der Intervention eine Wirkung zeigt oder nicht. So ist anhand einer experimentellen Versuchsanordnung nur durch sehr komplexe mehrfaktorielle Designs differenzierbar, welche Elemente einer mehrdimensionalen Intervention die Wirkung tatsächlich hervorgerufen haben (einzelne Elemente, die Kombination einzelner Elemente oder die Wechselwirkung mehrerer Elemente). Dementsprechend ist ein der experimentellen Forschung anhaftender Kritikpunkt, dass sie lediglich **Blackbox**-Ergebnisse liefert (Howe 2004; Kvernbekk 2016; Morrison 2001), d. h., ausschließlich Input (Ursache) und Output (Wirkung) werden mittels einer experimentellen Versuchsanordnung beschrieben.

Die Wirksamkeitsbeurteilung kann dementsprechend nicht nur unter dem Aspekt der bloßen Wirksamkeitsfrage (Wirkt die Intervention?), sondern auch unter dem Aspekt der Frage nach dem zugrunde liegenden Wirkmechanismus (Wie wirkt die Intervention?) gesehen werden. Hinsichtlich der Einsicht in den Wirkmechanismus bietet jegliche empirische Untersuchung, nicht nur die experimentelle, immer nur einen bestimmten Grad der Auflösung. Der geringste Auflösungsgrad zeigt lediglich einen Zusammenhang zwischen Intervention und

Zielkriterium. Dies kann zu erheblichen Fehlinterpretationen des Kausalzusammenhangs führen.

Dies lässt sich auch am Beispiel des weiter oben erwähnten Kaugummi-Experiments erläutern. Dabei sollte die Verbesserung der kognitiven Leistungen aufgrund des Kaugummikauens nachgewiesen werden. Als Wirkmechanismus ließe sich annehmen, dass die gleichmäßige Kaubewegung eine Steigerung der Funktion des Arbeitsgedächtnisses bewirkt und dadurch die Aufmerksamkeitsfähigkeit erhöht. Diese Aufmerksamkeitssteigerung soll wiederum die Leistung im kognitiven Test fördern (vgl. hierzu Rost et al. 2010). Die experimentelle Versuchsanordnung per se macht jedoch noch keine Aussage darüber, ob der Wirkmechanismus tatsächlich in der hypothesierten Form zum Tragen gekommen ist. In Wirklichkeit kann ein ganz anderer Mechanismus die Wirkung herbeigerufen haben: So könnte beispielsweise der in vielen Kaugummis enthaltene Zucker die Konzentrationsleistung steigern. Erst eine differenziertere experimentelle Versuchsanordnung kann den Auflösungsgrad erhöhen und somit auch eine differenziertere Einsicht in den eigentlichen Wirkmechanismus bieten. Die Differenzierung der Versuchsanordnung kann unter anderem mittels einer Erweiterung der Untersuchungsbedingungen vollzogen werden, z. B. indem neben der Kontrollgruppe zwei Interventionsgruppen (Kaugummi mit und ohne Zuckergehalt) aufgestellt werden.

Ein Wirkmechanismus kann ein mentaler, kognitiver, sozialer, physischer, biochemischer Prozess oder auch eine Kombination mehrerer Prozesse sein. Dieser Wirkprozess bzw. Teilaspekte des Prozesses sind mittels empirischer Forschung immer nur zu einem bestimmten Auflösungsgrad erfahrbare. Ein tiefgründiges Verständnis für den eigentlichen Wirkprozess kann aber die Basis für die Entwicklung noch wirksamerer bzw. adressatenorientierterer pädagogischer Maßnahmen darstellen. So stellt sich beispielsweise die pädagogisch relevante Frage, wie ein bestimmtes Training die Problemlösungsfähigkeit eines Kindes (z. B. bei Matrizenaufgaben) verbessern kann: Hat das Kind gelernt, falsche Antwortoptionen durch den Vergleich mit dem Matrizenmuster zu eliminieren, oder leitet sich das Kind die richtige Antwort aus dem Matrizenmuster her und wählt dann die richtige Antwortoption? (vgl. Börnert und Wilbert 2015, 2016). Beide Lösungsstrategien führen auf unterschiedliche Weise zur richtigen Lösung.

2.3 Komplementäre Forschungsdesigns zur Wirksamkeitsforschung

Wie oben dargelegt, wird dem randomisierten kontrollierten Experiment innerhalb der bildungswissenschaftlichen Evidenzbeurteilungsrichtlinien BEE und WWC die höchste Evidenzgüte zugeschrieben. Diese Zuschreibung findet ihre begründete Berechtigung in der hohen internen Validität der experimentellen Wirksamkeitsaussage. Jedoch konnten wir auch Probleme der Generalisierbarkeit der experimentellen Wirksamkeitsaussage diskutieren. Das Genrealisierbarkeitsproblem ist auf die geringe externe Validität des experimentellen Wirksamkeitsbefundes zurückzuführen. Gemäß den Evidenzbeurteilungsrichtlinien ist eine Reduktion des randomisierten kontrollierten Experiments mit einer Herabstufung der Wirksamkeitsevidenz einhergehend. Ein reduziertes experimentelles Forschungsdesign (z. B. Verzicht auf Randomisierung) wird als Quasiexperiment bezeichnet. Die Herabstufung der Wirksamkeitsevidenz rührt von der geringen internen Validität der Wirksamkeitsaussage des quasiexperimentellen Designs her. Analog zu dieser Herabstufung der Wirksamkeitsevidenz müsste eine weitere Reduktion des Quasiexperiments (z. B. Verzicht auf aktive Manipulation oder Kontrolle) eine weitere Herabstufung der Wirksamkeitsevidenz nach sich ziehen.

Entsprechende nicht-experimentelle Studiendesigns (z. B. Kohortenstudien, Panelstudien, Fall-Kontroll-Studien, Korrelationsstudien) werden in den Evidenzbeurteilungsrichtlinien der medizinischen Forschung berücksichtigt und genießen dort eine eingeschränkte Evidenzgüte (Burns et al. 2011). Allerdings finden diese nicht-experimentellen Designs („observational studies“) keine Beachtung innerhalb der bildungswissenschaftlichen Evidenzbeurteilungsrichtlinien bzw. sind sogar explizit von diesen ausgeschlossen. Auch Experimentaldesigns, die nicht auf dem Gruppenvergleichsparadigma beruhen (Single-Case-Designs), oder Quasiexperimentaldesigns, die entgegen der experimentellen Idee Unterschiede zwischen den Untersuchungsgruppen forcieren (Regressions-Diskontinuitäts-Designs), werden in der BEE-Richtlinie gar nicht und in der WWC-Richtlinie nur peripher berücksichtigt.

Auch qualitative Forschungsansätze und Mixed-Methods-Studien stehen für die Wirksamkeitsbeurteilung einer pädagogischen Maßnahme zur Diskussion (Cook 2002; Cook et al. 2012; Howe 2004). Innerhalb der bildungswissenschaftlichen Evidenzbeurteilungsrichtlinien finden diese qualitativen Forschungsansätze jedoch keine Berücksichtigung. Dabei könnten beispielsweise „intensive qualitative case studies“ (Cook 2002) und „think-aloud studies“ (Börnert et al. 2016) einen substanziellen Beitrag zur explorativen Ergründung des eigentlichen Wirkmechanismus leisten. Diese qualitativen Forschungsansätze können Hinweise auf Implementationshürden, unerwünschte Nebenwirkungen sowie positive/negative Subgruppeneffekte der pädagogischen Maßnahme liefern und somit weiteren Aufschluss über die externe Validität der Wirksamkeitsaussage geben.

Die bildungswissenschaftlichen Evidenzbeurteilungsrichtlinien (BEE und WWC) bieten eine top-down-orientierte Evidenzhierarchie (von höchster bis niedrigster bzw. limitierter Evidenzgüte). Diese Hierarchie wird maßgeblich von der internen Validität der empirischen Wirksamkeitsaussage determiniert und demnach vom randomisierten kontrollierten Experiment angeführt. Die externe Validität der Wirksamkeitsaussage und das Wissen um den eigentlichen Wirkmechanismus sind jedoch nicht zu vernachlässigende Dimensionen eines umfangreichen Wirksamkeitsnachweises.

Allerdings lassen sich interne und externe Validität sowie die Einsicht in den eigentlichen Wirkmechanismus anhand eines einzigen Forschungsdesigns (z. B. Experiment) nur schwer maximieren. Dies kann auch als ein Argument für einen forschungsmethodischen Pluralismus, der die ganzheitliche Erforschung der Wirksamkeit anvisiert, verstanden werden. Die Evidenzbasierung in den Bildungswissenschaften kann dementsprechend auch als ein komplementäres System gedacht werden (Abb. 2.2). Die unterschiedlichen Forschungsdesigns stehen dabei nicht in einem Top-down-Hierarchieverhältnis zueinander, sondern fokussieren die unterschiedlichen Wirksamkeitsaspekte (interne und externe Validität sowie Einsicht in den Wirkmechanismus) aus sich ergänzenden Blickwinkeln. Erst die Gesamtschau dieser komplementären Forschungsbefunde beleuchtet die unterschiedlichen Facetten (bzw. Qualitäten) der Wirksamkeit pädagogischer Maßnahmen



Abb. 2.2 Evidenzbasierung als komplementäre Wirksamkeitsforschung

und kann somit einen ganzheitlichen bzw. umfangreichen Wirksamkeitsnachweis bieten.

Die Evidenz über die Wirksamkeit einer pädagogischen Maßnahme ist dann umso höher, je mehr Studien mit den aufgeführten unterschiedlichen Designs eine Wirksamkeit nachweisen konnten. Zusätzlich bildet die Replikation von Studien die Konsistenz des Kausalzusammenhangs über unterschiedliche Settings (z. B. unterschiedliche Schulen in unterschiedlichen Sozialmilieus) ab und kann somit Aufschluss über die externe Validität der Wirksamkeitsaussage geben.

Zum Abschluss dieses Beitrags möchten wir auf zwei im komplementären Modell der Evidenzbasierung verankerte Studiendesignstypen sowie auf ihre Bedeutung für die Erforschung der Wirksamkeit pädagogischer Maßnahmen eingehen: Single-Case-Designs (Einzelfallforschung) und nicht-experimentelle Designs.

2.3.1 Single-Case-Designs (Einzelfallforschung)

Bei der Einzelfallforschung werden wiederholt Daten an einer einzelnen Person erhoben. In der medizinischen Forschung wird von N-of-1 Trials gesprochen. Im Kern werden dabei während der Datenerhebung zwei Phasen unterschieden: das Merkmal

von Interesse vor dem Beginn einer Intervention (A-Phase) und die Messungen mit Beginn der Intervention (B-Phase) (■ Abb. 2.3). Die beiden Phasen lassen sich vereinfacht analog der Kontrollgruppe (A-Phase) und der Interventionsgruppe (B-Phase) aus der Gruppenstudie verstehen. Jede dieser Phasen hat demnach mehrere mögliche Messungen. Verglichen werden nun statistische Kennwerte der Daten der A- und der B-Phase, um daraus Rückschlüsse auf die Veränderung durch die Intervention zu ziehen. Zumeist wird der Mittelwert der A- und B-Phase verglichen, aber auch andere Werte, z. B. Steigungsparameter, Median oder Varianz, können sinnvoll herangezogen werden (für einen vertiefenden Überblick siehe Jain und Spieß 2012).

Einzelfallforschung hat eine lange Tradition in der psychologischen und pädagogischen Forschung. Tatsächlich stellt die Einzelfallforschung zu Beginn der experimentellen psychologischen Forschung im 19. und 20. Jahrhundert die vorherrschende Methode zur Gewinnung von Erkenntnissen dar. So geht ein Großteil der Theorien und deren Evidenzen von Wundt, Ebbinghaus, Pavlov und Skinner auf Untersuchungen mittels wiederholter Messung an einzelnen Fällen zurück (vgl. Grünke 2012). Seit der Mitte des 20. Jahrhunderts gewann die gruppenbasierte experimentelle Forschung an Bedeutung und entwickelte sich zum praktischen Standard in der Interventionsforschung. Im Bereich der klinisch-psychologischen und der sonderpädagogischen Forschung

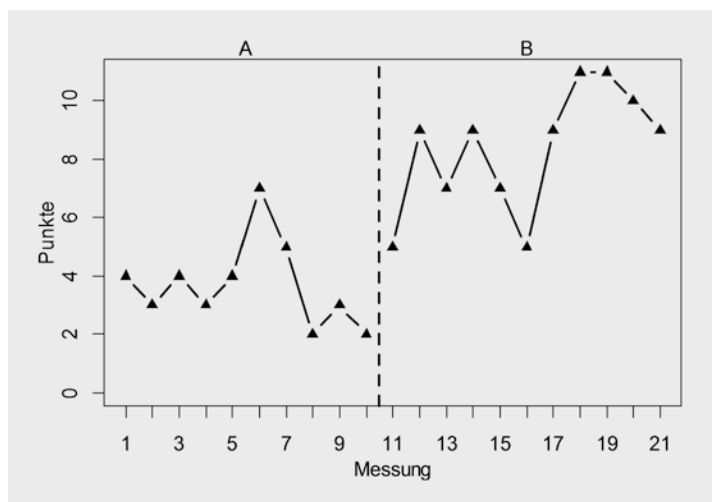
behielten Einzelfallstudien hingegen eine gewisse Bedeutung bei, was sich aufgrund des spezifischen Interesses dieser Disziplinen an dem Wirksamkeitsnachweis einer Intervention bei einem einzelnen Individuum erklären lässt.

Mit der Weiterentwicklung statistischer Verfahren gewann die Einzelfallforschung wieder zunehmend an Prominenz. Insbesondere wurden drei zentrale methodische Probleme überwunden oder zumindest methodisch adressiert.

Das erste dieser Probleme ist statistischer Natur und besteht in der fehlenden Unabhängigkeit der Messungen im Fall der wiederholten Messung an einer Person. Diese Abhängigkeit wird als Autokorrelation der Messwerte bezeichnet und verletzt eine wichtige Voraussetzung für die statistischen Analyseverfahren, die bei Gruppendesigns angewandt werden (z. B. T-Test und Varianzanalyse) (Huitema und McKean 1991). Durch die Anwendung allgemeiner linearer Schätzmodelle und die sich hierbei ergebenden Möglichkeiten, Autokorrelationen zu berücksichtigen, kann dieses Problem gelindert oder sogar aufgehoben werden (Huitema und McKean 2007).

Das zweite Problem entsteht aus der mangelnden Möglichkeit zur Kontrolle konfundierender Variablen, die parallel zum Beginn einer Intervention auftreten. So ist es möglich, dass mit dem Beginn einer Intervention zugleich auch zufällig ein weiteres Ereignis auftritt, das einen Einfluss auf die Zielvariable (abhängige Variable) hat (z. B. eine neue Klassenlehrerin;

■ Abb. 2.3 Visualisierung der Daten einer Einzelfallstudie im AB-Design



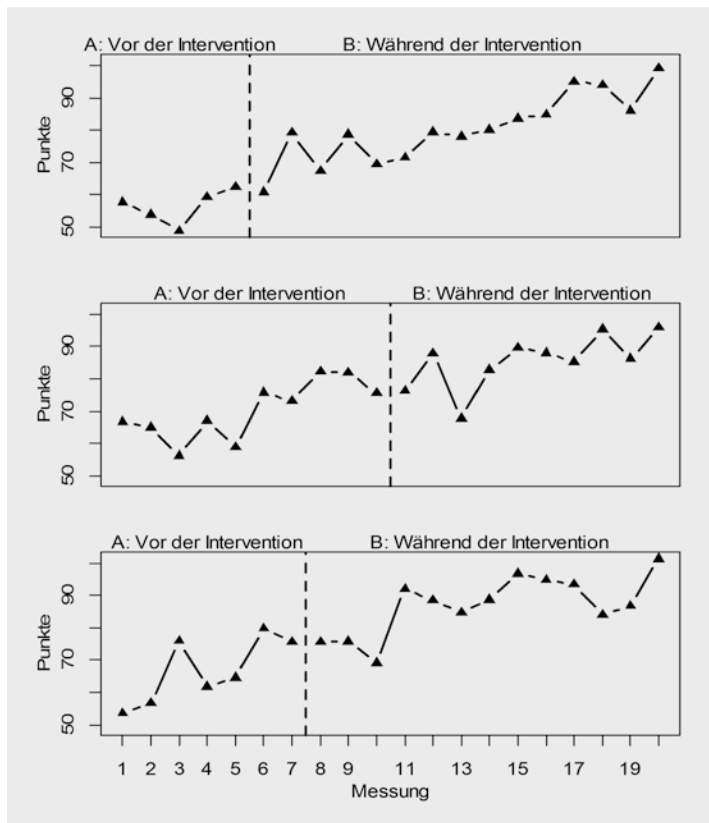
Veränderungen im familiären oder sozialen Umfeld des Kindes). Dieses Problem kommt dadurch zustande, dass die Messungen bei einem Einzelfall, anders als bei einem Gruppendesign, zeitlich geordnet sind. Bedroht ist hierdurch die interne Validität der Studie: Es ist nicht mehr eindeutig schließbar, dass eine Veränderung in dem Zielmerkmal das Ergebnis der Intervention ist. Um diesem Problem entgegenzuwirken, werden Multiple-Baseline-Designs angewandt (Jain und Spieß 2012). Bei diesen werden mehrere Einzelfälle zugleich in eine Studie implementiert. Zumeist ist dies die gleiche Intervention bei mehreren Personen. Möglich ist aber auch, die gleiche Person in drei verschiedenen Situationen zu untersuchen (z. B. könnte die Auswirkung eines verstärkerbasierten Verhaltenstrainings in drei Unterrichtsfächern erhoben werden). Wichtig ist hierbei, dass der Beginn der Intervention in den verschiedenen Fällen jeweils zufällig festgelegt wird (■ Abb. 2.4). Nun kann eine mit der Intervention nachweisbare Veränderung

in allen drei Fällen mit erhöhter Sicherheit ursächlich auf die Intervention zurückgeführt werden, was einer erhöhten internen Validität entspricht.

Eine weitere Möglichkeit, der Bedrohung der internen Validität entgegenzuwirken, besteht darin, innerhalb eines untersuchten Falls mehrere Phasen mit und ohne Intervention aneinanderzureihen (AB-AB-Design). Untersucht wird dann, ob sich die anhängige Variable mit jedem Übergang von A nach B in die erwartete Richtung verändert bzw. mit jedem Übergang von B nach A wieder zurück verändert. Diese Art des Designs ist für die meisten pädagogischen Kontexte nicht geeignet, da die Veränderungen hier zumeist auf Lernprozessen (akademisch, emotional oder auch sozial) beruhen und damit dauerhaft und nicht direkt reversibel sind.

Das dritte Problem der Einzelfallstudien betrifft das Studiendesign und ergibt sich aus der fehlenden Generalisierbarkeit eines einzelnen Falls auf die Grundpopulation. Es betrifft also die externe

■ **Abb. 2.4** Multiple-Baseline-Design zur Steigerung der internen Validität einer Einzelfallstudie



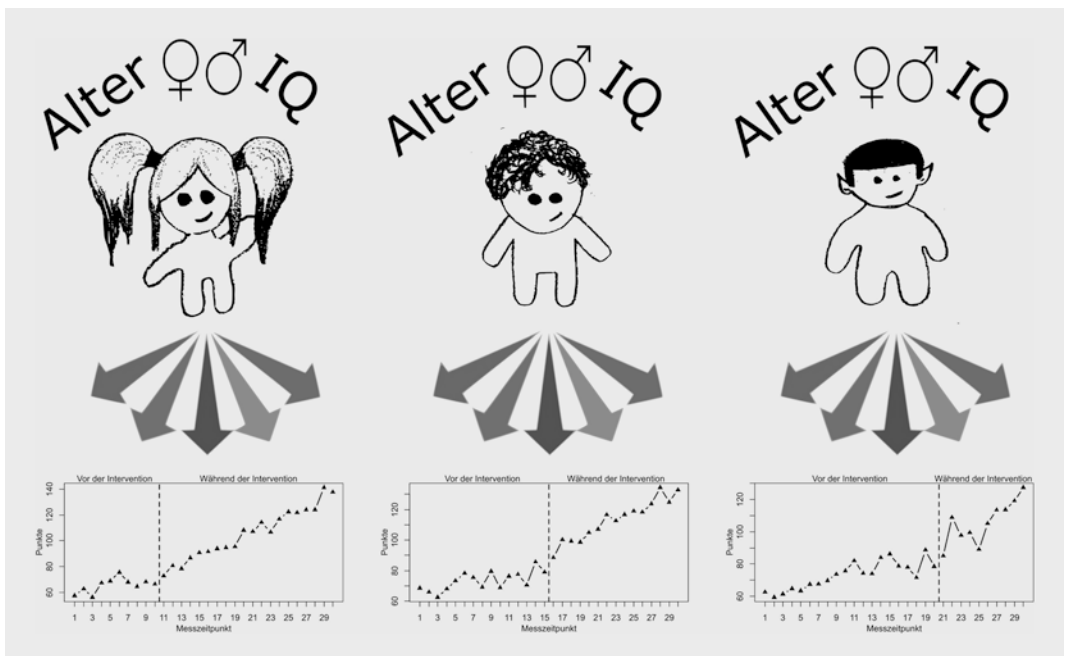
Validität der gefundenen Ergebnisse. Die einzige Möglichkeit, diesem Problem entgegenzuwirken, besteht darin, mehrere Einzelfälle in ein Studiendesign aufzunehmen oder alternativ mehrere Replikationsstudien mit der gleichen abhängigen und unabhängigen Variablen integriert zu analysieren. Erst in den vergangenen Jahren sind hierzu statistische Verfahren entwickelt worden, die dies ermöglichen. Besonders bewährt hat sich die Verknüpfung der Einzelfalldaten in Mehrebenenanalysen (Moeyaert et al. 2014; Van den Noortgate und Onghena 2008). Dabei werden die Messungen als genestete Daten betrachtet. D. h., wiederholte Messungen auf der Ebene 1 sind eindeutig Personen auf der Ebene 2 zugeordnet (■ Abb. 2.5). Dies erlaubt, die Wirkung einer Intervention auf Einzelfallniveau zu betrachten und dennoch zugleich auch die Verallgemeinerung („fixed effects“) und die Variabilität der Wirksamkeit („random effects“) zu analysieren.

Mit diesen Erweiterungen des Einzelfalldesigns und der damit einhergehenden erhöhten internen und externen Validität der hieraus gewonnenen Befunde eignen sich diese besonders für den Wirksamkeitsnachweis in pädagogischen Kontexten. Sie nehmen daher in einem komplementären

Evidenzbasierungsmodell neben dem randomisierten Kontrollgruppendesign eine notwendige methodische Position bzw. Relevanz ein.

2.3.2 Nicht-experimentelle Forschungsdesigns

Das Experimentieren, also die willentliche Manipulation (Veränderung) der Untersuchungsumwelt, ist ein notwendiger Bestandteil der experimentellen Forschung sowie ein zentrales Moment für die Ableitung einer Wirksamkeitsaussage. Eine experimentelle Manipulation der Untersuchungsumwelt ist aus ethischen, ökonomischen bzw. ökologischen Gründen nicht immer in das Forschungsdesign implementierbar. Diese Kontraindikationen der experimentellen Forschung und somit das Prinzip der nicht-experimentellen Forschung sollen exemplarisch an folgender Forschungsfrage erläutert werden (Kocaj et al. 2014): Hat die inklusive Regelbeschulung im Vergleich zur selektiven Beschulung einen positiven Effekt auf die Entwicklung der Schulkompetenzen von Kindern mit einem sonderpädagogischen Förderbedarf (SFB)?



■ Abb. 2.5 Analyse von Einzelfällen im Mehrebenenmodell

Die Wirksamkeit der inklusiven Regelbeschulung und somit die Frage, ob Kinder mit einem SFB hinsichtlich ihrer akademischen Entwicklung eher von einer inklusiven Regelbeschulung als von einer selektiven Beschulung (Förderschule) profitieren, stehen aktuell im Fokus des Forschungsinteresses. Die unterschiedlichen Schultypen (allgemeine Schule vs. Förderschule) können dabei als Untersuchungsbedingungen gedacht werden. Da die Entscheidungsgewalt bezüglich der Beschulungsart den Erziehungsberechtigten und/oder dem Schulamt obliegt, können die Kinder mit einem SFB den Untersuchungsbedingungen nicht per Zufall zugeteilt werden (ethische Kontraindikation). Die elterliche Entscheidung für eine Beschulungsart folgt bestimmten Gesetzmäßigkeiten. So kann beispielsweise angenommen werden, dass Eltern aufgrund der bisherigen Entwicklung ihrer Kinder entscheiden, ob sie ihren Kindern den Besuch einer inklusiven Schule zutrauen oder doch eher einen Förderschulbesuch für sinnvoller erachten. Ein derartiges Zuteilungsprinzip führt zwangsläufig zu sehr unterschiedlichen Untersuchungsgruppen, d. h., die Kinder in den allgemeinen Schulen und Förderschulen unterscheiden sich hinsichtlich ihrer Schulkompetenzen bereits vor Schulantritt.

Des Weiteren sind der experimentellen Umsetzung eines standardisierten inklusiven Schulunterrichts Grenzen gesetzt. Der Grad und die Art des inklusiven Unterrichts werden von Schule zu Schule und sogar von Klasse zu Klasse variieren. Dies liegt daran, dass es unterschiedliche Auffassungen und Konzepte von inklusiver Schulpraxis gibt (Grosche 2015). Um im Sinne einer experimentellen Manipulation den Grad und die Art des inklusiven Unterrichts zu steuern, wären regelmäßige und unter Umständen sehr ressourcenaufwändige (Geld, Zeit, Personal) Fortbildungen für die Lehrerkollegien notwendig (ökonomische Kontraindikation). Auch die Vermittlung der Fortbildungsinhalte ist keine Garantie dafür, dass die Lehrkräfte ihren Unterricht gänzlich an das vermittelte Inklusionskonzept anpassen. Jede Schule ist somit ein ökologisches System (Bronfenbrenner 1976) mit einem mehr oder weniger stark limitierten Zugang zur experimentellen Steuerung dieses Systems (ökologische Kontraindikation).

Der experimentellen Erforschung der Wirksamkeit inklusiver Schulpraxis sind demnach ethische,

ökonomische und ökologische Grenzen gesetzt. Dennoch ist die Ergründung des Kausalzusammenhangs zwischen der inklusiven Schulpraxis und der akademischen Entwicklung der Kinder mit einem SFB auch unter Verzicht auf die experimentellen Forschungsdesignelemente (Manipulation, Randomisierung, Standardisierung) möglich. Entsprechende nicht-experimentelle Studiendesigns werden gemeinhin als Beobachtungsstudien („observational studies“) bezeichnet (Rosenbaum 2010). Beobachtungsstudien sind in dem Sinne beobachtend, weil ausschließlich natürliche (nicht manipulierte) Gegebenheiten in der Schulumwelt beobachtet bzw. erfasst werden. Bezüglich der oben genannten Forschungsfrage stellen die Schultypen (allgemeine Schule vs. Förderschule) die natürlichen Untersuchungsbedingungen dar. In einer longitudinal angelegten Beobachtungsstudie werden die Zielkriterien (z. B. Schulkompetenzen wie das Lese- oder das Mengenverständnis) vor sowie nach der natürlichen Intervention gemessen (Temporalität), d. h., Datenerhebungen finden vor dem Eintritt in die allgemeine Schule/Förderschule sowie während der eigentlichen Schullaufbahn statt. Die Logik des kontrafaktischen Denkmodells ist auch im Falle einer longitudinalen Beobachtungsstudie anwendbar: Wenn die Kinder mit einem SFB statt einer inklusiven Schule die Förderschule besucht hätten, dann hätten sich die Schulkompetenzen dieser Kinder genauso entwickelt wie die Schulkompetenzen der Förderschulkinder.

Anhand eines Gruppenunterschieds in der Schulkompetenzentwicklung kann geschlussfolgert werden, dass beispielsweise die inklusive Beschulung für die Kinder mit einem SFB profitabler ist als die Förderbeschulung, d. h., dass hinsichtlich der akademischen Förderung der Kinder mit einem SFB die inklusive Schulpraxis wirksamer ist als die eigentliche Förderschulpraxis. Dieser Wirksamkeitsaussage kann nur dann eine hohe interne Validität zugesprochen werden, wenn die Kinder in den Untersuchungsgruppen mit Bezug auf möglichst viele Eigenschaften (potenzielle Störvariablen) einander ähnlich sind. Dies ist jedoch aufgrund des nicht-randomisierten Zuteilungsprinzips zu den Schultypen nicht der Fall, und die Kinder in den allgemeinen Schulen und Förderschulen unterscheiden sich somit bereits vor Schulantritt hinsichtlich ihrer Schulkompetenzen.

In der beobachtenden Forschung kann die Anwendung methodischer sowie statistischer Abhilfemaßnahmen die Untersuchungsanordnung eines Experiments imitieren und somit die Validität der Wirksamkeitsaussage steigern. Eine Alternative zur Randomisierung bilden z. B. Matchingverfahren (Rubin 2006): Förderschulkinder und Kinder mit einem SFB, die eine allgemeine Schule besuchen, werden hinsichtlich ihrer akademischen Entwicklung nur dann miteinander verglichen, wenn sie mit Bezug auf relevante Störvariablen (kognitive Grundfähigkeiten, sozioökonomischer Status, Bildungshintergrund der Eltern etc.) ein hohes Maß an Ähnlichkeit aufweisen.

Trotz der Anwendung eines Matchingverfahrens können jedoch unberücksichtigte Merkmale weitere Störvariablen darstellen. Das Matching bildet daher nur eine Annäherung an das Ziel der Randomisierung, also die annähernd identische Verteilung der erfassten sowie nicht erfassten Störvariablen in den beiden Untersuchungsgruppen. Zwar kann auf diese Weise eine Wirksamkeit der inklusiven Schulpraxis nachgewiesen werden, jedoch besteht aufgrund der mangelnden Standardisierung der inklusiven Schulpraxis keine präzise Erkenntnis darüber, welche inklusionspädagogischen Maßnahmen nutzbringend für die Kinder mit einem SFB sind. Sofern aber die inklusiven Unterrichtsmethoden erfasst worden sind (z. B. durch Lehrerbefragungen oder Unterrichtsbeobachtungen), können die Determinanten eines effektiven bzw. ineffektiven inklusiven Unterrichts im Rahmen einer longitudinalen Beobachtungsstudie genauer identifiziert werden.

Insbesondere aufgrund der Verankerung der nicht-experimentellen Studiendesigns in realen pädagogischen Settings genießen diese Designs eine hohe externe Validität und nehmen daher in einem komplementären Evidenzbasierungsmodell neben dem randomisierten Kontrollgruppendesign eine notwendige methodische Position bzw. Relevanz ein.

Literatur

Aronson E, Wilson TD, Akert RM (2004) Sozialpsychologie, 4. Aufl. Pearson Studium, München

Berliner DC (2002) Comment: Educational Research: The Hardest Science of All. *Educational Researcher* 31: 18–20. <https://doi.org/10.3102/0013189X031008018>

Best Evidence Encyclopedia (2017) About the Best Evidence Encyclopedia. <http://www.bestevidence.org/aboutbee.htm>

Börnert M, Grubert A, Wilbert J (2016) Lautes Denken als Methode zur Forschung und Diagnostik in inklusionspädagogischen Handlungsfeldern. In: Gebele D, Zepter AL (Hrsg) *Inklusion: Sprachdidaktische Perspektiven Theorie – Empirie – Praxis*. Gilles, Francke Verlag KG, Duisburg, S 165–186

Börnert M, Wilbert J (2015) Thinking-aloud protocols of Piagetian tasks: Insights into problem-solving processes of primary school students. *Insights on Learning Disabilities* 12: 19–34

Börnert M, Wilbert J (2016) Dynamisches Testen als neue Perspektive in der sonderpädagogischen Diagnostik – Theorie, Evidenzen, Impulse für Forschung und Praxis. *Zeitschrift für Heilpädagogik* 67: 156–167

Bosch J, Schaefer A, Kulawiak PR, Wilbert J (2016) Forschungsdesigns zur Untersuchung kausaler Beziehungen in den empirischen Bildungswissenschaften. In: Gebele D, Zepter AL (Hrsg) *Inklusion: Sprachdidaktische Perspektiven Theorie – Empirie – Praxis*. Gilles, Francke Verlag KG, Duisburg, S 138–164

Briggs DC (2008) Comments on Slavin: Synthesizing Causal Inferences. *Educational Researcher* 37: 15–22. <https://doi.org/10.3102/0013189X08314286>

Bronfenbrenner U (1976) The Experimental Ecology of Education. *Educational Researcher* 5: 5–15

Burns PB, Rohrich RJ, Chung KC (2011) The Levels of Evidence and Their Role in Evidence-Based Medicine. *Plastic and Reconstructive Surgery* 128: 305–310. <https://doi.org/10.1097/PRS.0b013e318219c171>

Cook BG, Smith GJ, Tankersley M (2012) Evidence-based practices in education. In: Harris KR, Graham S, Urdan T, McCormick CB, Sinatra GM, Sweller J (Hrsg) *APA educational psychology handbook, Vol 1: Theories, constructs, and critical issues*. American Psychological Association, Washington, S 495–527

Cook BG, Tankersley M, Landrum TJ (2009) Determining Evidence-Based Practices in Special Education. *Exceptional Children* 75: 365–383. <https://doi.org/10.1177/001440290907500306>

Cook TD (2002) Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for not Doing Them. *Educational Evaluation and Policy Analysis* 24: 175–199. <https://doi.org/10.3102/01623737024003175>

Grosche M (2015) Was ist Inklusion? In: Kuhl P, Stanat P, Lütjeh-Klose B, Gresch C, Pant HA, Prenzel M (Hrsg) *Inklusion von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen*. Springer Fachmedien Wiesbaden, Wiesbaden, S 17–39

Grünke M (2012) Editorial. Schwerpunktthema: Kontrollierte Einzelfallforschung. *Empirische Sonderpädagogik* 4: 207–209

Hartmann U, Decristan J, Klieme E (2016) Unterricht als Feld evidenzbasierter Bildungspraxis?: Herausforderungen

- und Potenziale für einen wechselseitigen Austausch von Wissenschaft und Schulpraxis. *Zeitschrift für Erziehungswissenschaft* 19: 179–199. <https://doi.org/10.1007/s11618-016-0712-4>
- Howe KR (2004) A Critique of Experimentalism. *Qualitative Inquiry* 10: 42–61. <https://doi.org/10.1177/1077800403259491>
- Huitema BE, McKean JW (1991) Autocorrelation estimation and inference with small samples. *Psychological Bulletin* 110: 291–304. <https://doi.org/10.1037/0033-2909.110.2.291>
- Huitema BE, McKean JW (2007) An improved portmanteau test for autocorrelated errors in interrupted time-series regression models. *Behavior Research Methods* 39: 343–349. <https://doi.org/10.3758/BF03193002>
- Institute of Education Sciences (2014) What Works Clearinghouse. Procedures and Standards Handbook Version 3.0. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf
- Jain A, Spieß R (2012) Versuchspläne der experimentellen Einzelfallforschung. *Empirische Sonderpädagogik* 4: 211–245
- Jornitz S (2009) Evidenzbasierte Bildungsforschung. *Pädagogische Korrespondenz* 68–75
- Kavale KA, Mostert MP (2003) River of Ideology, Islands of Evidence. *Exceptionality* 11: 191–208. https://doi.org/10.1207/S15327035EX1104_1
- Kocaj A, Kuhl P, Kroth AJ, Pant HA, Stanat P (2014) Wo lernen Kinder mit sonderpädagogischem Förderbedarf besser? Ein Vergleich schulischer Kompetenzen zwischen Regel- und Förderschulen in der Primarstufe. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie* 66: 165–191. <https://doi.org/10.1007/s11577-014-0253-x>
- Kvernbekk T (2016) Evidence-based practice in education: functions of evidence and causal presuppositions. Routledge, Taylor, Francis Group, London, New York
- Moeyaert M, Ferron JM, Beretvas SN, Van den Noortgate W (2014) From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology* 52: 191–211. <https://doi.org/10.1016/j.jsp.2013.11.003>
- Morrison K (2001) Randomised Controlled Trials for Evidence-based Education: Some Problems in Judging „What Works.“ *Evaluation, Research in Education* 15: 69–83. <https://doi.org/10.1080/09500790108666984>
- Rosenbaum PR (2010) Design of observational studies. Springer, New York
- Rost DH, Wirthwein L, Frey K, Becker E (2010) Steigert Kaugummikauen das kognitive Leistungsvermögen? *Zeitschrift für Pädagogische Psychologie* 24: 39–49. <https://doi.org/10.1024/1010-0652/a000003>
- Rubin DB (2006) Matched sampling for causal effects. Cambridge University Press, Cambridge, New York
- Schrader J (2014) Analyse und Förderung effektiver Lehr-Lernprozesse unter dem Anspruch evidenzbasierter Bildungsreform. *Zeitschrift für Erziehungswissenschaft* 17: 193–223. <https://doi.org/10.1007/s11618-014-0540-3>
- Shadish WR, Cook TD, Campbell DT (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin, Boston
- Slavin RE (2002) Evidence-Based Education Policies: Transforming Educational Practice and Research. *Educational Researcher* 31: 15–21. <https://doi.org/10.3102/0013189X031007015>
- Van den Noortgate W, Onghena P (2008) A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention* 2: 142–151. <https://doi.org/10.1080/17489530802505362>
- Wilbert J, Grünke M (2015) Kontrollierte Einzelfallforschung. In: Ellinger S, Koch K (Hrsg) *Forschungsmethoden in der Heil- und Sonderpädagogik*. Hogrefe, Göttingen, S 100–105

Evidenzbasierte Praxis in den Gesundheitsberufen
Chancen und Herausforderungen für Forschung und
Anwendung

Haring, R.; Siegmüller, J. (Hrsg.)

2018, XV, 219 S. 24 Abb., Softcover

ISBN: 978-3-662-55376-3