# DISKUSSION ZUM SCHWERPUNKTTHEMA

DDS – Die Deutsche Schule
115. Jahrgang 2023, Heft 3, S. 247–261
https://doi.org/10.31244/dds.2023.03.08


Melinda Erdmann, Marcel Helbig & Irena Pietrzyk

## Plädoyer für eine neue methodische Übereinkunft

### Zum Potenzial von randomisiert-kontrollierten Studien in der Evaluation von Schulentwicklungsprogrammen

**Zusammenfassung**
*In diesem Beitrag plädieren wir für eine Verständigung zwischen verschiedenen Akteur*innen über methodische Standards bei der Evaluation von Schulentwicklungsprogrammen, die zur Reduzierung von Bildungsungleichheit eingesetzt werden. Ziel einer solchen Verständigung wäre es, geteiltes Wissen über methodische Fragen aufzubauen und eine breit akzeptierte Einigung darüber zu erreichen, welche Forschungsmethoden als sinnvoll und verlässlich für die Beurteilung von Schulentwicklungsprozessen gelten. Wir sind der Ansicht, dass ergänzend zu den bislang hauptsächlich durchgeführten Prozessevaluationen Wirkungsevaluationen, und hier insbesondere randomisiert-kontrollierte Studien, einen zentralen Baustein in der Bewertung von in der Schule angesiedelten Programmen darstellen sollten.*
*Schlüsselwörter: Schulentwicklung; Forschung; Evaluation; Feldexperiment; randomized controlled trial; Bildung; Ungleichheit*

**Pleading for a New Methodological Agreement**
On the Potential of Randomized Controlled Trials in the Evaluation of School Development Programs

**Abstract**
*In this paper, we argue for an understanding among different actors about methodological standards in the evaluation of school development programs that are implemented to reduce educational inequality. The aim of such an understanding would be to build shared knowledge about methodological issues and to reach a broadly accepted consensus on which research methods are considered useful and reliable for evaluating school development processes. We believe that, in addition to the process evaluations that have mainly been conducted to date, impact evaluations, and in particular randomized controlled trials, should be a central component in the evaluation of school-based programs.*

*Keywords: school development; research; evaluation; field experiment; randomized controlled trial; education; inequality*

## 1 Einleitung

Aktuell wird mit dem „Startchancen"-Programm das bislang größte Schulentwicklungsprogramm in Deutschland zur Bekämpfung von Bildungsungleichheit geplant. In der Politik, in den Schulen und in der Forschung besteht Einigkeit darüber, dass die im Rahmen dieses Programms eingesetzten Maßnahmen evidenzbasiert und auf Grundlage aktueller wissenschaftlicher Erkenntnisse ausgewählt sein sollten. Jedoch existiert im Feld bis heute keine Übereinkunft darüber, auf welche Weise Maßnahmen oder Programme bewertet werden sollten. Konkret besteht keine Verständigung unter Politiker*innen, Praktiker*innen und Wissenschaftler*innen darüber, welche Forschungsdesigns im Rahmen einer Evaluation dafür geeignet sind, verlässliche Aussagen über die Wirksamkeit von Maßnahmen zu treffen.

Erste Impulse für eine Verständigung über Forschungsansätze im Bildungsbereich hat vor Kurzem die Ständige Wissenschaftliche Kommission (SWK) der Kultusministerkonferenz (2022) gesetzt. Der vorliegende Beitrag will diesen Diskussionsprozess über Forschungsdesigns und ihre Praktikabilität im Feld der Schulentwicklung weiterführen. Einen solchen Dialog, der nach unserer Überzeugung keineswegs nur Wissenschaftler*innen, sondern alle an der Gestaltung und Umsetzung von Schulentwicklungsprogrammen beteiligten Akteur*innen einbeziehen sollte, erachten wir als dringend geboten, um einen allgemeinen Konsens darüber herbeizuführen, welche Forschungsmethoden geeignet sind, um empirisch belastbare Aussagen über die Wirksamkeit von Maßnahmen zu treffen.

Erörterungen über Forschungsdesigns mögen auf den ersten Blick trocken oder normativ gehaltlos erscheinen. *De facto* haben sie allerdings gewichtige Implikationen für die Ausgestaltung von Schulentwicklungsprogrammen und somit letztlich für den Bildungserfolg verschiedener Bevölkerungsgruppen. Angesichts der hohen Bedeutung wissenschaftlicher Forschungsdesigns für die Beurteilung von Interventionen haben sich in anderen Disziplinen, wie etwa in der Medizin oder in der Psychotherapie, Wissenschaftler*innen, Praktiker*innen und Politiker*innen bereits vor Jahrzehnten darauf verständigt, anhand welcher Forschungsmethoden (neue) Interventionsansätze beurteilt werden sollen. Die Evaluation von Schulentwicklungsmaßnahmen hängt, jedenfalls in Deutschland, diesem methodischen Diskussionsstand deutlich hinterher, obwohl die Schulbildung von herausgehobener Bedeutung für den weiteren Lebensverlauf ist. Wünschenswert wäre deshalb, dass sich auch die Beurteilung von Schulentwicklungsprogrammen an einer breiten methodischen Übereinkunft unter Politiker*innen, Praktiker*innen und Wissenschaftler*innen orientiert. Dies gilt vielleicht sogar in besonderem Maße für Maßnahmen, die auf den Abbau von Bildungsungleichheit zielen. Denn dieses Anliegen weist mit Blick auf die allseits diskutierten individuellen und gesamtgesellschaftlichen Folgen von Bildungsarmut einerseits eine ausgesprochen hohe gesellschaftspolitische Dringlichkeit auf; andererseits werden da-

für aktuell beträchtliche Summen zur Verfügung gestellt, an deren möglichst zielführender Verwendung alle Akteur*innen ein starkes Interesse haben müssen.

Aktuell lässt sich in aller Deutlichkeit beobachten, welche Folgen es haben kann, wenn eine Übereinkunft über die in Evaluationen anzustrebenden Forschungsansätze fehlt. Derzeit ist die gängige Forschung zu Schulentwicklungsprogrammen der sogenannten Prozessevaluation zuzuordnen. Vorliegende wissenschaftliche Begleituntersuchungen stellen typischerweise Aspekte wie die interne und externe Akzeptanz der Programme, das Schulklima, die Unterrichtsgestaltung und die Zusammenarbeit zwischen Lehrenden, Eltern und externen Partner*innen im Sozialraum in den Fokus und gehen dabei häufig auch der Frage nach, ob es verwaltungsrechtliche, kontextuelle oder schulinterne Hürden gibt, die eine effektive Umsetzung der Programme erschweren. Bei diesen Faktoren, die den Abbau von Bildungsungleichheit vermutlich befördern, handelt es sich indes um Prozessmerkmale, d. h. die bisherigen Begleituntersuchungen fokussieren vor allem den Prozess der Implementierung und Umsetzung der Programme. Auch wenn es natürlich bedeutsam ist sicherzustellen, dass die in einem Programm vorgesehenen Maßnahmen eine hohe Implementierungsqualität erreichen, ist diese keine hinreichende Bedingung dafür, dass ein konkretes Programm tatsächlich seine Ziele erfüllt, also etwa die Basiskompetenzen von sozioökonomisch benachteiligten Schüler*innen fördert und darüber vermittelt tatsächlich Bildungsungleichheiten abbaut.

Unserer Einschätzung nach werden in Ergänzung zu solchen Prozessevaluationen auch Wirkungsevaluationen dringend benötigt, um bewerten zu können, ob ein Programm seine Ziele letztlich erreicht. Wirkungsevaluationen, insbesondere solche mit randomisiert-kontrolliertem Design, das trotz einiger Probleme im Hinblick auf die Verallgemeinerbarkeit der Ergebnisse vielen Wissenschaftler*innen als „Goldstandard" gilt (z. B. Hariton & Locascio, 2018), liegen jedoch zur Beurteilung von Schulentwicklungsprogrammen in Deutschland unseres Wissens nicht vor. Folgerichtig konstatieren daher auch Marx und Maaz in diesem Band, dass mangels entsprechender Forschung gegenwärtig für Deutschland kaum belegt ist, dass die für Schulen in schwieriger sozialer Lage bestehenden Schulentwicklungsprogramme tatsächlich Bildungsungleichheiten reduzieren. Anstelle von Wirkungsevaluationen mit randomisiert-kontrolliertem Design wurden typischerweise Evaluationen durchgeführt, in denen etwaige Programmwirkungen allein über Einschätzungen von Schulleitungen und Lehrkräften erfasst werden (vgl. Braun & Pfänder, 2022). Auf welcher Grundlage aber ein umfassendes Programm wie das „Startchancen"-Programm zusammengestellt werden soll, wenn die empirische Evidenz zur Wirkung verschiedener Maßnahmen derart dünn ist, ist fraglich.

Wir sind der Ansicht, dass sich eine solch unbefriedigende Situation durch Bemühungen um eine methodische Übereinkunft hätte vermeiden lassen und zukünftig vermieden werden kann. Weil in unseren Augen Wirkungsevaluationen von essenzieller Bedeutung für die Beurteilung von Programmen sind und weil wir gerade im Feld der Schulentwicklungsmaßnahmen diesbezüglich eine Leerstelle wahrnehmen, stellen wir diesen Typus der Evaluation im vorliegenden Diskussionsbeitrag in den Vorder-

grund. Dabei liegt unsere Expertise in der experimentellen Evaluation von Bildungsprogrammen. Denn wir haben in der Vergangenheit als Wissenschaftler*innen Erfahrungen bei der randomisiert-kontrollierten Begleitung eines an Schulen angesiedelten Beratungsprogramms gesammelt, das auf die Reduktion von Bildungsungleichheit in der Studienaufnahme zielt. Wir sind hingegen keine ausgewiesenen Expert*innen für Schulentwicklungsmaßnahmen. Entsprechend können wir es nicht leisten, im Detail auszuformulieren, wie randomisiert-kontrollierte Studien im Feld der Schulentwicklung implementiert werden könnten. Stattdessen argumentieren wir für eine grundlegende forschungsmethodische Stoßrichtung, die den Stärken von randomisiert-kontrollierten Studien Rechnung trägt, und versuchen, einige Vorbehalte auszuräumen.

Dafür stellen wir im ersten Abschnitt dar, auf welchen Grundüberlegungen Wirkungsevaluationen fußen und warum sie bei der Beurteilung von Schulentwicklungsmaßnahmen einen zentralen Platz einnehmen sollten. Um in eine konstruktive Diskussion einzusteigen, thematisieren wir im zweiten Abschnitt Vorbehalte, die regelmäßig gegen Wirkungsevaluationen vorgebracht werden, formulieren Entgegnungen auf diese Vorbehalte und skizzieren Optionen für pragmatische Lösungen im Schulkontext. Im dritten Kapitel thematisieren wir, dass eine Wirkungsevaluation, die methodischen Standards genügt, stets eine klare Zieldefinition und eine angemessene Berücksichtigung des Zeithorizonts voraussetzt, wobei hier verschiedene Akteur*innen und nicht zuletzt die Bildungspolitik und -administration gefordert sind. Der Beitrag endet mit einem Ausblick auf die Möglichkeiten, die in der aktuellen Programmoffensive für die Entwicklung einer methodischen Übereinkunft, für evidenzbasierte Forschung und letztlich für einen Abbau von Bildungsungleichheiten liegen. Wir hoffen, dass er dazu beiträgt, die Diskussion um Forschungsmethoden im Bereich der Schulentwicklung zu intensivieren und einen Dialog unter den relevanten Akteur*innen zu befördern, innerhalb dessen weitere Vorbehalte, Fragen und Argumente ausgetauscht werden können (im Sinne einer Schulsystementwicklung; vgl. Berkemeyer & Hermstein in diesem Heft).

## 2 Stärken von Wirkungsevaluationen

Bei der sogenannten Wirkungsevaluation liegt der Fokus auf dem *Ergebnis* einer Intervention, also auf ihrer Wirkung – nicht auf der Implementierung und Durchführung von Maßnahmen, wie dies bei Prozessevaluationen der Fall ist (vgl. Döring & Bortz, 2016). Nachfolgend beschreiben wir für ein allgemeines Verständnis die grundlegende Herangehensweise von Wirkungsevaluationen und führen dann aus, warum es in unseren Augen dringend geboten ist, dieser Art der Evaluation bei der Beurteilung von Schulentwicklungsmaßnahmen, die auf die Reduktion von Bildungsungleichheiten zielen, einen besonderen Stellenwert einzuräumen.

Ganz prinzipiell unterliegen der Wirkungsevaluation kausalanalytische Überlegungen zu einem sogenannten *kontrafaktischen* Szenario. So haben in der Vergangenheit Forscher*innen darauf hingewiesen, dass ein kontrafaktisches Szenario wünschenswert wäre, um belastbare Aussagen darüber zu treffen, ob bzw. inwieweit eine bestimm-

te Maßnahme, z. B. ein Bildungsprogramm, ein bestimmtes Ergebnis, z. B. die Basiskompetenzen, beeinflusst (vgl. Rubin, 1974). Ein solches kontrafaktisches Szenario im buchstäblichen Sinne würde beispielsweise bedeuten, dass eine Gruppe von Schüler*innen an einer Maßnahme zur Kompetenzstärkung teilnimmt, während zugleich *dieselbe* Gruppe von Schüler*innen nicht an dieser Maßnahme teilnimmt. Wie stark sich die beiden Gruppen in den Basiskompetenzen, die man nach der Teilnahme an der Maßnahme messen würde, unterscheiden, wäre das Maß dafür, ob und wie stark das Programm die Basiskompetenzen fördert. Es versteht sich von selbst: Ein solches kontrafaktisches Szenario ist logisch nicht möglich. Mit dem sogenannten randomisiert-kontrollierten Verfahren existiert aber eine Methode, die versucht, sich an das kontrafaktische Szenario anzunähern.[1]

Bei randomisiert-kontrollierten Studien (*randomized controlled trials*) werden Untersuchungseinheiten, z. B. Schüler*innen oder Schulen, zufällig zwei Gruppen zugeordnet: einer Gruppe, die an einer Maßnahme teilnimmt (sog. Programmgruppe), und einer Gruppe, die nicht an der Maßnahme teilnimmt (sog. Kontrollgruppe). Die zufällige Zuordnung soll die Ähnlichkeit zwischen der Programm- und der Kontrollbedingung maximieren, um dem kontrafaktischen Szenario möglichst nahezukommen.[2] Werden ausreichend viele Untersuchungseinheiten (in diesem Fall also Schüler*innen oder Schulen) zufällig zugeordnet, kann mit einer sehr hohen Ähnlichkeit der Untersuchungsgruppen auf beobachteten und unbeobachteten Merkmalen vor Beginn der Maßnahme gerechnet werden. Programm- und Kontrollgruppe würden sich also im Idealfall nur in der Programmteilnahme unterscheiden. Kausale Rückschlüsse über die Wirkung einer Maßnahme erfolgen dann über den Gruppenvergleich nach Ende des Programms, da davon ausgegangen werden kann, dass sich Unterschiede zwischen den Gruppen (beispielsweise gemessene Kompetenzunterschiede) allein auf die Intervention zurückführen lassen. Dies ist der zentrale Grund dafür, dass sich randomisiert-kontrollierte Studien bereits vor vielen Jahrzehnten in der medizinischen und psychologischen Forschung durchgesetzt haben – auch wenn bekannt ist, dass die mittels randomisiert-kontrollierter Studien gewonnenen Ergebnisse auch Mängel aufweisen können, insbesondere bezüglich ihrer Verallgemeinerbarkeit.[3]

---

1 Darüber hinaus existieren weitere quantitative Verfahren, mittels derer versucht wird, Vergleichbarkeit zwischen Untersuchungsgruppen, die nicht randomisiert zugeordnet worden sind, herzustellen (z. B. Regressionsanalysen und *propensity score matching*).

2 Bei der zufälligen Zuordnung zur Programm- und Kontrollgruppe können verschiedene Verfahren zum Einsatz kommen, um eine Ähnlichkeit zwischen den Untersuchungsgruppen zu fördern. Sofern zum Beispiel Schulen randomisiert werden, können anhand von spezifischen Merkmalen (z. B. Schultyp, Schulgröße, herausfordernde soziale Lage) Paare von einander ähnlichen Schulen gebildet werden. Innerhalb dieser Paare wird eine Schule der Programm- und die andere Schule der Kontrollgruppe zufällig zugeordnet.

3 Wie beschrieben können mittels randomisiert-kontrollierter Studien relativ belastbare Aussagen darüber getroffen werden, ob und inwieweit ein Programm ein Ergebnis tatsächlich kausal beeinflusst (interne Validität). Als Nachteil dieser Studienart wird jedoch oftmals die unklare Verallgemeinerbarkeit genannt (externe Validität). Konkret ist oft nicht spezifiziert, wie stark die in der Studie gewonnenen Resultate auf reale Bedingungen verallgemeinerbar sind. Mögliche Abweichungen zwischen der Studie und den realen Bedingungen, die Verzerrungen verursachen könnten, betreffen unter anderem die Zusammensetzung der Teilnehmer*innen und die Programmdurchführung (vgl. Shadish et al., 2002). Entsprechend

Basierend auf der beschriebenen zufälligen Zuordnung minimieren randomisiert-kontrollierte Studien Verzerrungen, die andernfalls aufgrund des sogenannten Selektionsbias entstehen können. In der sozialen Realität bestimmt nicht der Zufall über eine Programmteilnahme. Vielmehr entscheiden Schüler*innen oder Schulleitungen nach bestimmten Kriterien, ob sie bzw. ihre Schulen an einer Maßnahme teilnehmen oder nicht – beispielsweise in Abhängigkeit von ihrer schulbezogenen Motivation, von ihren Ressourcen oder ihren Erwartungen bezüglich des Nutzens. Teils erfolgt eine Zuweisung auch durch die Bildungsadministration. Eine solche „Selbstselektion" oder „Fremdselektion" führt jedoch dazu, dass sich die Gruppe, die an einem Programm teilnimmt, bereits *vor* Implementation der Maßnahme von der Gruppe unterscheidet, die nicht an dem Programm teilnimmt. Unterschiede, die zwischen den Gruppen *nach* der Programmteilnahme beobachtet werden, können dann nicht mehr verlässlich auf die Wirkung des Programms zurückgeführt werden, da diese prinzipiell auch durch Besonderheiten der Programmgruppe bedingt sein können (beispielsweise eine höhere Motivation). Sofern man solche Selbst- oder Fremdselektionsprozesse nicht durch eine randomisierte Zuordnung systematisch auszuschließen versucht, ist das Risiko sehr hoch, dass in Bereichen von hoher gesellschaftlicher Tragweite am Ende falsche kausale Rückschlüsse gezogen werden (also etwa Kompetenzvorsprünge, die der Gruppenzusammensetzung geschuldet sind, fälschlicherweise dem Programm zugeschrieben werden).

Konkret finden in bestehenden Schulentwicklungsmaßnahmen oftmals nicht zufällige Selektionsprozesse statt, die teils sogar intendiert sind. Zum Bespiel wird laut Braun und Pfänder (2022) bei zwei Dritteln der Programme der Zugang zu Unterstützungsressourcen und Schulentwicklungsmaßnahmen über eine Schulbewerbung reguliert. Es ist recht wahrscheinlich, dass diese Praxis dazu führt, dass vorrangig solche Schulen am Programm teilnehmen, die ein hohes intrinsisches Interesse an Schulentwicklung haben und sich in der Lage sehen, ein solches Programm an ihrer Schule erfolgreich zu gestalten, während beispielsweise Schulen, die schon für das Bewerbungsprozedere keine (Personal-)Ressourcen entbehren können, weil etwa die Schulleitung nicht besetzt ist oder Lehrkräfte fehlen, so von der Programmteilnahme so gut wie ausgeschlossen werden. Im Extremfall würden in der Folge entwicklungsinteressierte oder ressourcenstarke Schulen in der Programmgruppe mit entwicklungsunmotivierten oder ressourcenschwachen Schulen in der Kontrollgruppe verglichen, was dann unter Umständen falsche Rückschlüsse bezüglich der Programmwirkung nach sich zöge. Bei randomisiert-kontrollierten Studien identifiziert hingegen in der Regel das Forschungsteam geeignete Studienschulen und kontaktiert diese. Je nach Studienanlage handelt es sich dabei beispielsweise ausschließlich um Schulen in schwieriger sozialer Lage. Nachdem diese Schulen ihre Studienbereitschaft freiwillig zum Ausdruck gebracht haben, entscheidet der Zufall darüber, welche der Studienschulen am Programm teilnehmen und welche nicht.

---

kann der mittels einer randomisiert-kontrollierten Studie geschätzte Effekt von dem tatsächlichen, unter realen Bedingungen gegebenen Programmeffekt abweichen. Studienbeschreibungen sollten daher detaillierte Angaben z. B. zu den Teilnehmer*innen und zum Programm umfassen.

Ist man nun – wie im Falle des „Startchancen"-Programms – daran interessiert, möglichst effektive Maßnahmen einzusetzen, wäre es also sinnvoll, Programme zu wählen, die sich im Rahmen von randomisiert-kontrollierten Studien als effektiv erwiesen haben. Hier wäre verhältnismäßig gut abgesichert, dass diese Programme tatsächlich eine positive Wirkung entfalten, z. B. auf die Entwicklung von Basiskompetenzen und darüber vermittelt auf die Veränderung von Bildungsungleichheit. Eine solche Wahl von Programmen und Maßnahmen ließe sich nach ökonomischen und normativen Kriterien rechtfertigen. Der Einsatz (knapper) Ressourcen wäre durch die nachweislich positive Wirkung der Programme besser begründbar. Im umgekehrten Fall würde es im Mindesten weiterer Rechtfertigung bedürfen, wenn für ein Programm, dessen Wirkung nicht als bestätigt gelten kann, finanzielle Mittel (in beträchtlichem Umfang) ausgegeben werden sollten – die dann an anderer Stelle möglicherweise fehlen. Auch aus einer normativen Perspektive, die der Förderung soziostrukturell benachteiligter Schüler*innen und dem Abbau von Bildungsungleichheit verpflichtet ist, wäre eine solche Auswahl von Programmen anhand aktueller wissenschaftlicher Standards wünschenswert. Denn es wäre wahrscheinlicher, dass diese Ziele – etwa durch das „Startchancen"-Programm – tatsächlich erreicht würden, wenn anhand von randomisiert-kontrollierten Studien positiv evaluierte Maßnahmen zum Einsatz kämen.

Faktisch aber geben die bisherige Forschungspraxis und hier insbesondere die bislang durchgeführten Prozessevaluationen in der Regel nicht verlässlich über die Programmwirkung Auskunft. Bei den wenigen Studien, die ein Vergleichsgruppendesign implementiert haben, erfolgte die Zuordnung zu den Gruppen nicht zufällig. Da jedoch eine zufällige Zuordnung für verlässliche kausale Rückschlüsse unerlässlich ist, sind Aussagen wie jene, dass Schulentwicklung soziale Ungleichheiten nicht verringern könne (z. B. Böttcher et al., 2022, S. 17), unseres Erachtens ebenso wenig gesichert wie optimistisch gestimmte Aussagen. Daher sollte der bisherige Fokus auf die Prozessqualität von Schulentwicklungsprogrammen in kommenden Forschungsprojekten unbedingt um die Wirkungsevaluation mittels randomisiert-kontrollierter Studien erweitert werden, da wir sonst auch in Zukunft nicht werden abschätzen können, welche Maßnahmen mit dem Ziel des Abbaus von Bildungsungleichheiten sinnvollerweise verstetigt werden sollten.

## 3 Vorbehalte gegen Wirkungsevaluationen

Obwohl Wirkungsevaluationen, und hier insbesondere randomisiert-kontrollierte Studien, es also ermöglichen, kausale Rückschlüsse verhältnismäßig verlässlich zu ziehen, kamen und kommen sie in Deutschland in der Forschung zu Schulentwicklungsmaßnahmen bislang unseres Wissens nach nicht zur Anwendung. Eine mögliche Ursache hierfür sind Vorbehalte, die sich speziell gegen randomisiert-kontrollierte Studien richten (vgl. Cook, 2002). Nachfolgend werden wir drei dieser Vorbehalte ansprechen und darlegen, warum diese in unseren Augen nicht so gewichtig sind, wie sie auf den ersten Blick möglicherweise erscheinen.

Der erste Vorbehalt bezieht sich auf den Umstand, dass im Rahmen von randomisiert-kontrollierten Studien eine Standardisierung des Programms bzw. der Maßnahme notwendig ist. Demgegenüber werden Schulentwicklungsmaßnahmen oft erst in den Schulen entwickelt und erprobt. Auch fokussieren sie oftmals die Einzelschule in ihrem komplexen sozialen Umfeld. Entsprechend kann aus Perspektive der Schulentwicklungsforschung die Standardisierung, die für randomisiert-kontrollierte Studien notwendig ist, als nicht kontextsensibel erscheinen und den Eindruck vermitteln, Wirkungszusammenhänge zu simplifizieren. In unseren Augen liegt jedoch gerade in der Standardisierung von Programmen ein erhebliches Potenzial – sie bietet nämlich die Möglichkeit zur Replikation. Wurden beispielsweise anhand einer Evaluation Erkenntnisse darüber gewonnen, dass eine Maßnahme tatsächlich Bildungsungleichheiten reduziert, ist die Standardisierung dieser Maßnahme eine Voraussetzung dafür, dass sie an anderen Schulen implementiert werden kann. Denn es ist die Standardisierung, die Handlungsanleitungen dafür liefert, wie ein Programm ausgestaltet ist und welche Inhalte und Prozesse es umfasst. Dabei bedeutet Standardisierung nicht notgedrungen, dass kontextspezifische Variationen nicht berücksichtigt werden können; Standardisierung heißt vielmehr, dass der Umgang mit Kontextspezifität systematisiert ist. Aufgrund dessen, dass Standardisierung Replizierbarkeit beinhaltet, ist sie eine Voraussetzung für Aussagen von allgemeinerem Gehalt. Da in unseren Augen das Erkenntnisinteresse von Evaluationen nicht nur darauf liegen sollte, was sich an einer spezifischen Einzelschule als sinnvoll erwiesen hat, sondern auf allgemeineren Aussagen, ist ein gewisser Grad der Standardisierung unseres Erachtens in der Evaluationspraxis unvermeidlich – was sich in gewisser Hinsicht in den Empfehlungen des Bundesministeriums für Bildung und Forschung (BMBF) zu einem Leitfaden in der Implementierung des „Startchancen"-Programms widerspiegelt (vgl. auch BMBF, 2023, S. 4). Im Dialog verschiedener Akteur*innen müsste sich erweisen, welcher konkrete Standardisierungsgrad für Schulentwicklungsmaßnahmen praktikabel ist.

Der zweite Vorbehalt betrifft die ethische Dimension von randomisiert-kontrollierten Studien. Hier wird in Zweifel gezogen, dass es ethisch legitim sei, der Kontrollgruppe ein Programm „vorzuenthalten". Bezugnehmend auf diesen Vorbehalt möchten wir festhalten, dass die Vorstellung oftmals gar nicht zutreffend ist, dass der Kontrollgruppe ein Programm aufgrund des Forschungsdesigns „vorenthalten" wird. Oftmals sind die Ressourcen vielmehr ohnehin knapp, was dazu führt, dass unabhängig vom Forschungsdesign einer Studie nicht alle Schulen bzw. nicht alle Schüler*innen, die potenziell von einem Programm profitieren könnten, auch tatsächlich die Möglichkeit erhalten, daran teilzunehmen. Bislang hatten die Mittel, die für Schulentwicklungsprogramme bereitgestellt wurden, regelmäßig vielmehr einen eher kleinen Umfang. Für diese kleinen Programme wäre es in der Vergangenheit relativ problemlos möglich gewesen, die gesamte Gruppe der potenziell förderungswürdigen Schulen zu bestimmen und darauf aufbauend randomisiert Schulen auszuwählen, die an den jeweiligen Programmen hätten teilnehmen können. So wäre die Anzahl der Schulen, die von einer Programmteilnahme ausgeschlossen wird, durch eine randomisiert-kontrollierte Wirkungsevaluation nicht „unnatürlich" vergrößert worden.

Im Unterschied zu den kleineren Programmen aus der Vergangenheit soll das aktuell geplante „Startchancen"-Programm allerdings ganze 4.000 Schulen umfassen. Es könnten somit potenziell alle Schulen teilnehmen, die besonders hohe Armutsquoten aufweisen (vgl. Helbig, 2023). Jedoch wäre auch hier ein Vergleichsgruppendesign denkbar, ohne dass einer Schule das Programm aufgrund des Forschungsdesigns vorenthalten würde. So wäre es konkret möglich, Ressourcen gestaffelt und randomisiert zur Verfügung zu stellen – dergestalt, dass einige Schulen in der ersten Phase zwar eine Mittelzusage bekämen, das Programm jedoch zeitlich verzögert einsetzen würde. Diese Schulen mit verzögerter Teilnahme könnten sodann als Kontrollgruppe fungieren. Erfahrungen mit großen Programmen, wie dem „DigitalPakt Schule", haben in der Vergangenheit gezeigt, dass es unrealistisch ist, dass alle zur Verfügung stehenden Mittel zeitgleich ausgeschöpft werden. Die Implementierung von neuen Programmen erfolgt vielmehr regelmäßig auch in der „natürlichen Umwelt" gestaffelt, sodass schrittweise neue Schulen in ein Programm aufgenommen werden oder dass an Schulen, die am Programm teilnehmen, schrittweise neue Klassen in eine Maßnahme inkludiert werden. Entsprechend ist es auch ohne wissenschaftliche Begleitung bereits jetzt gängige Praxis, dass manche Schulen bzw. manche Klassen zu einem konkreten Zeitpunkt (noch) nicht in den Genuss eines Programms kommen, während andere bereits teilnehmen. Diese in der „natürlichen Umwelt" auffindbare Staffelung kann prinzipiell im Design von randomisiert-kontrollierten Studien berücksichtigt werden. Ein solches Vorgehen erfordert natürlich eine intensive Kommunikation unter allen Akteur*innen in Politik, Praxis und Wissenschaft, Sensibilität für die jeweiligen Rationalitäten und die Bereitschaft, einander entgegenzukommen, soweit es in der eigenen Macht liegt. Auch dafür wäre es sinnvoll, wenn allen Akteur*innen aufgrund einer entsprechenden methodischen Übereinkunft der Sinn eines solchen Vorgehens bewusst wäre.[4]

Weiterhin denken wir, dass es oftmals fair ist, über die Zuweisung von „Programmplätzen" mittels Losverfahren zu entscheiden. Das Losverfahren ist bereits jetzt in zahlreichen Situationen als legitimes Mittel zur Verteilung begrenzter Ressourcen akzeptiert. So werden beispielsweise Schulplätze an immer mehr Orten (z. B. in Berlin, Köln, Erfurt) oder auch Studienplätze über das Losverfahren vergeben. Im Unterschied zu den meisten anderen Verteilungsprozeduren (wie etwa der Entscheidung über eine Bewerbung) ist das Losverfahren blind gegenüber den Ressourcen der Kandidat*innen oder ihrem sozialen Status. Auch unseres Erachtens wäre es daher grundsätzlich fairer, bei knappen Ressourcen die Teilnahme an einem Programm auf Individual- oder Schulebene durch ein Los zu entscheiden als über eine Bewerbung – dem häufig gewählten Weg zur Verteilung von Programmplätzen auf Schulebene.

4 Hierbei stellt es eine besondere Herausforderung dar, die beschriebene Staffelung mit dem teils langen Zeithorizont, in dem manche Maßnahmen ihre Wirkung entfalten, in Einklang zu bringen (vgl. auch Kap. 4). Im Detail müsste erörtert werden, ab wann die Kontrollschulen das Programm frühestens anbieten könnten, ohne dass die Schätzung innerhalb der randomisiert-kontrollierten Studie davon negativ berührt wäre. Bei Maßnahmen, die auf der Ebene der Klassen wirken, würde zum Beispiel ein Jahr Aufschub ausreichen. Bei Maßnahmen, die auf Schulebene gleichermaßen auf alle Schüler*innen wirken, würde ein Jahr Verzögerung hingegen nur im Falle einer sehr schnell sichtbaren Wirkung der Maßnahmen sinnvoll sein.

Als dritter Vorbehalt wird häufig formuliert, dass Wirkungsevaluationen sehr aufwändig und daher besonders kostenintensiv seien. Dieser Eindruck geht vermutlich darauf zurück, dass sich aus den längerfristigen Zielen zahlreicher Programme die Notwendigkeit einer langen wissenschaftlichen Begleitung ergibt (vgl. Rolff & Tillmann, 1980). Zudem muss im Rahmen randomisiert-kontrollierter Studien die Kontrollgruppe, die gar nicht am Programm teilnimmt, recht groß sein, um Unterschiede zwischen den Gruppen verlässlich absichern zu können. Auch dies mag den Eindruck eines (unnötig) hohen Aufwands verstärken.

Es muss jedoch gar nicht immer notwendig sein, viele neue Daten im Rahmen von randomisiert-kontrollierten Studien zu erheben. Sofern es gelingt, eine gut aufeinander abgestimmte Zusammenarbeit zwischen verschiedenen Akteur*innen zu etablieren, könnten bereits existierende Daten genutzt werden. Konkret gibt es zum Beispiel bereits jetzt für alle Schulen in Deutschland Daten zu Schulabgängen, Klassenwiederholungen, Fehltagen oder dem Unterrichtsausfall. Dies sind Indikatoren, die bereits in einigen wenigen vorliegenden Evaluationen in den Blick genommen werden, um die Wirkung von Programmen zu beurteilen. Zudem liegen in einigen Bundesländern im sogenannten individualstatistischen Kerndatensatz Informationen zu Schulverläufen aller Schüler*innen vor, die mit weiteren Informationen wie Angaben zum Migrationshintergrund angereichert sind. Darüber hinaus werden an Schulen in regelmäßigen Abständen bereits Kompetenzmessungen in Form von Lernstandserhebungen durchgeführt (z. B. VerA). Prinzipiell wäre es also denkbar und wünschenswert, diese Daten für die Wirkungsforschung zu nutzen und damit Doppelerhebungen zu umgehen – denn eine wissenschaftliche Nutzung dieser Daten findet bislang kaum statt.

Die Öffnung von bereits existierenden Datensätzen für wissenschaftliche Zwecke, von denen viele Forschungsprojekte von gesellschaftspolitischem, ökonomischem und normativem Gehalt profitieren könnten, würde es erfordern, dass Akteur*innen aus den Schulen, der Politik und der Wissenschaft intensiver als bislang in einen Austausch miteinander treten und ihre Tätigkeiten stärker als bisher auf einander abstimmen und koordinieren – mit dem übergeordneten Ziel, verlässliche Erkenntnisse zur Wirkung von Programmen und Maßnahmen zu generieren, um tatsächlich evidenzbasiert handeln zu können.

## 4 Herausforderungen von Wirkungsevaluationen: Zieldefinition und Zeithorizont

Auch wenn einige Vorbehalte gegen Wirkungsevaluationen in unseren Augen weniger gewichtig sind, als es auf den ersten Blick scheint, stellen sich bei dieser Form der Evaluation gleichwohl weitere spezifische Herausforderungen. Eine dieser Herausforderungen ist die konkrete und zeitlich stabile Zieldefinition. Denn im Unterschied zur wissenschaftlichen Grundlagenforschung haben Wirkungsevaluationen die Aufgabe, konkrete Programme zu bewerten. An welchen Kriterien diese Maßnahmen gemessen werden, hängt von der Zieldefinition dieser Programme ab. Eine weitere Herausforderung besteht darin, den Zeithorizont von durch die Maßnahme angestoßenen Ver-

änderungsprozessen im Forschungsdesign zu berücksichtigten. Wir werden im nachfolgenden Kapitel in drei Punkten darstellen, dass die Wirkungsevaluation nur dann glücken kann, wenn die Zieldefinition gelingt und der erforderliche Zeithorizont angemessen berücksichtigt wird. Damit möchten wir nicht zuletzt das Bewusstsein dafür schärfen, dass eine gelungene Wirkungsevaluation auch vom Agieren politischer Entscheidungsträger*innen abhängt. Die Aufnahme von Wirkungsevaluationen in eine methodische Übereinkunft zwischen verschiedenen Akteur*innen müsste daher auch ein deutliches Commitment von Politiker*innen für diese Form der Evaluation und ihre Anforderungen einschließen.

Die erste Herausforderung von Evaluationsstudien betrifft die Passung zwischen den Programmzielen und den Wirkungskriterien (Indikatoren), die zur Evaluation herangezogen werden (vgl. Döring & Bortz, 2016). Eine Passung kann unter zwei Bedingungen gewährleistet werden. Erstens müssen auf Seite der Programmverantwortlichen empirisch messbare Ziele formuliert werden. Demnach ist es notwendig, übergeordnete Ziele wie etwa „Chancengerechtigkeit" (BMBF, 2023, S. 2) in konkrete und messbare Ziele wie „Stärkung der Basiskompetenzen, d. h. […] [der] Kernkompetenzen in Lesen, Schreiben, Mathematik" (BMBF, 2023, S. 3) zu übersetzen. Zweitens müssen auf Seiten der Forscher*innen geeignete Verfahren gefunden werden, um diese Ziele zu messen. Sollte das „Startchancen"-Programm weiterhin darauf zielen, Basiskompetenzen zu stärken, wäre demnach eine Kompetenzmessung für die Beurteilung der Programmwirkung unabdinglich.

Als zweite Herausforderung ist zu nennen, dass die Zielsetzungen von Programmen und Maßnahmen für die Dauer der Evaluation konstant gehalten werden sollten und nicht fluide sein dürfen. Auf Seiten der Programmverantwortlichen besteht das Risiko einer Anpassung der Programmziele an verschiedene Faktoren – etwa aufgrund von (Zwischen-)Ergebnissen in Evaluationsprojekten. Im Unterschied zu solchen Anpassungen ist eine über die Dauer der Wirkungsevaluation konstante Zielbeschreibung für eine gute wissenschaftliche Praxis notwendig. Beispielsweise sollte im Rahmen des „Startchancen"-Programms an der Förderung der Basiskompetenzen als Zieldefinition festgehalten werden und nicht auf weichere Ziele wie die Schulmotivation oder die Zufriedenheit der Schüler*innen, die möglicherweise leichter positiv zu beeinflussen sind, ausgewichen werden. Denn erstens besteht nicht (immer) die Möglichkeit, das Design und das quantitative Instrumentarium an neu definierte Ziele anzupassen. Randomisiert-kontrollierte Studien sind in der Regel vom Beginn bis zum Ende durchgeplant. So existiert häufig bereits zu Beginn der Studie ein detaillierter Plan, welche Auswertungen am Ende des Forschungsprojekts die zentralen Ergebnisse liefern werden. Dabei sind auch die statistischen Auswertungsverfahren, die ungefähre Stichprobengröße und die Art der Messung des Outcomes (z. B. Basiskompetenzen) vordefiniert. Abweichungen von diesem Plan erschweren es, methodisch sauber zu arbeiten. Zweitens – und an dieser Stelle vielleicht bedeutsamer – würde eine Anpassung der Ziele an den Evaluationsverlauf das Anliegen von Wirkungsevaluationen, verlässliche Aussagen zu generieren, *ad absurdum* führen. Ändert man im Verlauf der Evaluation die Zieldefinition von Programmen oder Maßnahmen, wäre es möglich, eine gegebene Intervention durch das Ausweichen auf „weichere" Kriterien als erfolg-

reich zu deklarieren – obwohl diese keine Wirkung auf die ursprüngliche Zielvariable aufweist.

Als dritte Herausforderung ist der Zeithorizont der Wirkung zu berücksichtigen. Häufig werden mit Schulentwicklungsprogrammen längerfristige Ziele verfolgt, wie etwa die Förderung von Schulabschlüssen und die Vermeidung von Schulabbrüchen (für eine Übersicht von Zielen verschiedener Programme vgl. Tulowitzki et al., 2020, S. 52 ff.). Zudem benötigen einige programminduzierte Veränderungsprozesse einen ausgedehnten Zeithorizont, um eine Wirkung zu entfalten. Auch das „Startchancen"-Programm ist auf eine Laufzeit von zehn Jahren angelegt. Der (lange) Zeithorizont sollte sich im Forschungsdesign der Evaluationsprojekte wiederfinden. Ein zu kurzer Zeithorizont des Forschungsdesigns kann zu zwei verschiedenen Problemen führen. Erstens kann eine zeitlich zu beschränkte Betrachtung dazu führen, dass bestimmte Indikatoren nicht beobachtet werden können, weil sie außerhalb der beobachteten Zeitspanne liegen. Zweitens kann eine falsch bemessene Zeitspanne auch zu falschen Schlussfolgerungen führen. So zeigte sich während der Untersuchung (vgl. Erdmann et al., 2022) eines individuellen Beratungsprogramms zur Reduzierung von sozialer Ungleichheit beim Übergang von der Schule zur Hochschule, dass soziale Ungleichheit in der Studienaufnahme erst ein Jahr nach dem Abitur entsteht – weil viele Schüler*innen direkt nach dem Abitur zuerst ein *Gap Year*, z. B. in Form eines Freiwilligen Sozialen Jahres, aufnehmen. Eine zu frühe Beendigung der wissenschaftlichen Begleitung wäre zu dem falschen Ergebnis gekommen, dass die untersuchte Maßnahme keine positive Wirkung erzielt. Letztlich zeigte sich jedoch bei einer Betrachtung, die nach dem *Gap Year* angesetzt war, eine deutliche Reduzierung der sozialen Ungleichheit in der Studienaufnahme durch die Beratung.

Entsprechend ist es die Aufgabe der Programmverantwortlichen, bei Bedarf mit Unterstützung des Forschungsteams, im Rahmen von Wirkungsevaluationen Ziele zu benennen, die messbar sind und die während der Evaluation nicht verändert werden sollten, sowie dafür Sorge zu tragen, dass die Forschung ausreichend lange weitergeführt werden kann. Diese Aufgaben erfordern auf Seiten der Politiker*innen und der Bildungsadministrationen eine hohe Präzision und einen internen Verständigungsprozess darüber, unter welchen Bedingungen ein von der Politik aufgesetztes Programm tatsächlich als erfolgreich gelten kann – und wann es, im Umkehrschluss, als nicht erfolgreich deklariert werden muss.

## 5 Fazit

Dieser Beitrag versteht sich als Plädoyer dafür, dass Akteur*innen aus Politik, Praxis und Wissenschaft in einen Prozess der methodischen Verständigung im Hinblick auf Evaluationsstandards der Schulentwicklung eintreten. Ziel einer solchen Verständigung wäre es, geteiltes Wissen über methodische Fragen aufzubauen und eine breit akzeptierte Einigung darüber zu erreichen, welche Forschungsmethoden als sinnvoll und verlässlich für die Beurteilung von Schulentwicklungsprozessen gelten. Gemeinsam mit vielen anderen Wissenschaftler*innen (SWK, 2022) sind wir der Ansicht, dass

Wirkungsevaluationen, und hier insbesondere randomisiert-kontrollierte Studien, einen zentralen Baustein in der Bewertung von in der Schule angesiedelten Programmen darstellen sollten – und zwar nicht anstelle, sondern ergänzend zu den bislang hauptsächlich durchgeführten Prozessevaluationen, die ihrerseits zahlreiche Erkenntnisse zu den Bedingungen und Herausforderungen von Schulentwicklungsmaßnahmen zu Tage gefördert haben.

Wir sind der Ansicht, dass eine gelungene Evaluationspraxis eine Übereinkunft über die Rolle von Wirkungsevaluationen zur Voraussetzung hat. Wirkungsevaluationen sind in ihren Abläufen nicht trivial – auch wenn die Prozesse weniger kompliziert sein können als mitunter angenommen wird. Damit das Ziel, belastbare Aussagen über die kausale Wirkung von Programmen zu generieren, tatsächlich erreicht werden kann, müssen im Rahmen solcher Evaluationsprojekte die Aktivitäten zahlreicher Akteur*innen aufeinander abgestimmt werden, konkret etwa die Aktivitäten von Lehrkräften, von Schüler*innen, von Personen, die Programme durchführen, von Wissenschaftler*innen und von politischen Entscheidungsträger*innen und den ihnen unterstellten Bürokratien. Auch kann es notwendig werden, Informationen zu teilen, sich entgegenzukommen und nach kreativen Lösungen zu suchen – wie etwa das Beispiel der Integration einer gestaffelten Mittelvergabe in einem randomisiert-kontrollierten Design verdeutlicht. Nur so ist ein reibungsloser Ablauf der komplexen Prozesse im Sinne guter wissenschaftlicher Praxis möglich. Ein reibungsloser Ablauf bei hoher Komplexität lässt sich jedoch nur dann erreichen, wenn sich die Beteiligten den Zielen und Methoden der Evaluation verpflichtet fühlen – wenn sie selbst also das gesamte Unterfangen als sinnvoll erachten und damit auch bereit sind, ihr eigenes Handeln daran auszurichten.

In Auseinandersetzung mit den aktuellen bildungspolitischen Bemühungen, Bildungsungleichheiten mit dem bundesweiten „Startchancen"-Programm zu bekämpfen, wird besonders sichtbar, dass eine solche methodische Übereinkunft sich bislang nicht in der deutschen Forschungspraxis wiederfindet. Dennoch sind wir der Meinung, dass gerade die aktuelle Situation und ein klarer Blick auf Defizite bisher dominierender Evaluationspraxis einen besonderen Möglichkeitsraum eröffnen, um an einer methodischen Verständigung und Übereinkunft zwischen Wissenschaft, Politik und Schulentwicklungspraxis zu arbeiten. Wir verstehen diesen Diskussionsbeitrag als Schritt in einem größeren Verständigungsprozess, der zukünftig an verschiedenen Orten weitergeführt werden sollte. Zugleich bietet das „Startchancen"-Programm ganz konkret den Optionsraum, um verschiedene Elemente der Schulentwicklung auf ihre Wirkung hin zu untersuchen. Denn die methodischen Herausforderungen, die bei der Umsetzung von Wirkungsanalysen bestehen, lassen sich vor allem in breit angelegten Maßnahmen und groß angelegten Forschungsprojekten gut meistern. Das gegenwärtige Interesse an einem programmbezogenen Wissenstransfer, die inzwischen einsetzende Debatte über adäquate Forschungsmethoden sowie schließlich die – gerade auch im Rahmen des „Startchancen"-Programms – in Aussicht stehenden finanziellen und institutionellen Ressourcen für randomisiert-kontrollierte Wirkungsevaluationen stimmen uns optimistisch, dass es zukünftig möglich sein könnte, Programmpakete zur

Förderung soziostrukturell benachteiligter Schüler*innen und zur Bekämpfung von Bildungsungleichheit zu schnüren, die tatsächlich ihr Ziel erreichen.

Ob dieses Vorhaben tatsächlich gelingen kann, hängt jedoch nicht allein von Wissenschaftler*innen ab, sondern ganz maßgeblich von den Anstrengungen politischer Entscheidungsträger*innen. Denn letztlich sind sie es, die sowohl in einer besonderen Verantwortung stehen, einen Austausch zur methodischen Übereinkunft voranzutreiben, als auch Rahmenbedingung dafür zu schaffen, dass Programme anhand von randomisiert-kontrollierten Studien evaluiert werden können. Ob die von uns skizzierten Anregungen für eine neue methodische Übereinkunft tatsächlich Gehör finden und sich auf die gelebte Evaluationspraxis übertragen, hängt also entscheidend davon ab, welche politische Priorität die Bekämpfung von Bildungsungleichheit in den nächsten Jahren haben wird.

## Literatur und Internetquellen


BMBF (Bundesministerium für Bildung und Forschung). (2023, 22. Mai). *Eckpunkte zum Startchancen-Programm*. Entwurf des BMBF. https://table.media/bildung/wp-content/uploads/sites/15/2023/05/BMBF_Eckpunkte_Startchancen-Programm.pdf

Böttcher, W., Brockmann, L., Meierjohann, T., & Wiesweg, J. (2022). *Was brauchen Schulen in herausfordernden Lagen. Studie im Auftrag des Netzwerk Bildung*. Friedrich-Ebert-Stiftung. https://library.fes.de/pdf-files/a-p-b/19077.pdf

Braun, L., & Pfänder, H. (2022). *Unterstützung von Schulen in herausfordernden Lagen: Eine vergleichende Darstellung aktueller Programme*. Wübben-Stiftung. https://www.wuebben-stiftung.de/wp-content/uploads/2022/09/WS_UnterstuetzungvonSchuleninherausforderndenLagen_Expertise.pdf.

Cook, T. D. (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* (3), 175–199. https://doi.org/10.3102/01623737024003175

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer. https://doi.org/10.1007/978-3-642-41089-5

Erdmann, M., Helbig, M., Jacob, M., Pietrzyk, I., Schneider, J., & Allmendinger, J. (2022). Soziale Ungleichheit beim Hochschulzugang verringern. Intensive Beratung fördert die Passung zwischen Potenzialen und Bildungsentscheidungen. *WZBrief Bildung, 45*. https://bibliothek.wzb.eu/wzbrief-bildung/WZBriefBildung452022_erdmann_helbig_jacob_petrzyk_schneider_allmendinger.pdf

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials – the gold standard for effectiveness research. *BJOG – An International Journal of Obstretis and Gynaecology, 125* (13), 1716–1716. https://doi.org/10.1111/1471-0528.15199

Helbig, M. (2023). Eine „faire" Verteilung der Mittel aus dem Startchancenprogramm erfordert eine ungleiche Verteilung auf die Bundesländer. Eine Abschätzung der Mittelbedarfe für die deutschen Grundschulen anhand der Armutsquoten in den Sozialräumen. *WZB Disucssion Paper, P 2023–001*. https://bibliothek.wzb.eu/pdf/2023/p23-001.pdf

Rolff, H.-G., & Tillmann K.-J. (1980). Schulentwicklungsforschung: theoretischer Rahmen und Forschungsperspektive. *Jahrbuch der Schulentwicklung, 1*, 237– 264.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66* (5), 688–701. https://doi.org/10.1037/h0037350

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

SWK (Ständige Wissenschaftliche Kommission der Kultusministerkonferenz). (2022). *Entwicklung von Leitlinien für das Monitoring und die Evaluation von Förderprogrammen im Bildungsbereich*. https://www.kmk.org/fileadmin/Dateien/pdf/KMK/SWK/2022/SWK-2022-Impulspapier_Monitoring.pdf

Tulowitzki, P., Grigoleit, E., Haiges, J., & Hinzen, I. (2020). *Unterstützung von Schulen in herausfordernder Lage – Ein bundesweiter Überblick. Expertise im Auftrag der Wübben Stiftung.* Wübben Stiftung. https://doi.org/10.26041/FHNW-3412

*Melinda Erdmann*, Dr., wissenschaftliche Mitarbeiterin am Wissenschaftszentrum in Berlin für Sozialforschung in der Forschungsgruppe der Präsidentin.
E-Mail: melinda.erdmann@wzb.eu
Korrespondenzadresse: Wissenschaftszentrum Berlin für Sozialforschung, Reichpietschufer 50, 10785 Berlin

*Marcel Helbig*, Prof. Dr., Leiter des Arbeitsbereichs Strukturen und Systeme am Leibniz-Institut für Bildungsverläufe (LIfBi).
E-Mail: marcel.helbig@lifbi.de
Korrespondenzadresse: Leibniz-Institut für Bildungsverläufe, Wilhelmsplatz 3, 96047 Bamberg

*Irena Pietrzyk*, Dipl. Psych., M. A. Soz., wissenschaftliche Mitarbeiterin an der Universität zu Köln, Institut für Soziologie und Sozialpsychologie (ISS).
E-Mail: pietrzyk@wiso.uni-koeln.de
Korrespondenzadresse: Universität zu Köln, Institut für Soziologie und Sozialpsychologie (ISS), Albertus-Magnus-Platz, 50923 Köln