Stiftung Deutsches Forum für Kriminalprävention

Entwicklungsförderung & Gewaltprävention 2015 / 2016

Aktuelle Beiträge aus Wissenschaft und Praxis

Heute für ein besseres Morgen.

Der buddY E.V. und seine Programme

Der buddY E.V. versteht sich als Forum für eine Neue Lernkultur. Der Verein gestaltet und verändert mit innovativen Konzepten die Bildungslandschaft. Im Verbund der zentralen Bildungsinstitutionen KITA - SCHULE - FAMILIE - HOCHSCHULE entwickelte er das buddY-Programm, das familY-Programm, das studY-Programm und aktuell mYkita.

Das buddY-Programm hat die Zielsetzung, Schulen in potenzialentfaltenden Reformprozessen zu begleiten und zu unterstützen.

Im familY-Programm werden Eltern in ihrer Rolle als Bildungsbegleiter und Potenzialentfalter gestärkt.

Das neue mYkita-Programm unterstützt pädagogische Fachkräfte in der Kooperation mit Eltern.

An der Universität Duisburg-Essen werden im studY-Programm Lehramtsstudierende durch innovative Lehre und die Erfahrung zukunftsweisender Schulpraxis in der Entwicklung ihrer professionellen Rolle unterstützt.

Methodische Beurteilung von Evaluationsstudien im Bereich der Gewalt- & Kriminalitätsprävention: Beschreibung und Begründung eines Methodenprofils

Andreas Beelmann & Judith Hercher
Friedrich-Schiller-Universität Jena

Der Einsatz von Präventionsprogrammen und Präventionsstrategien wird unter anderem mit dem wissenschaftlichen Nachweis legitimiert, dass vorab intendierte Effekte tatsächlich auftreten. Eine wissenschaftliche Fundierung ist daher vor allem mit der Notwendigkeit verbunden, die Evidenz mit Ergebnissen aus aussagekräftigen Untersuchungen zur Wirksamkeit zu belegen (Beelmann, 2015). Darauf beruht auch der Anspruch wissenschaftlich fundierter bzw. evidenzbasierter Maßnahmen, anderen Handlungsstrategien überlegen zu sein.

1. Zur Bewertung der forschungsmethodischen Qualität von Evaluationsstudien

Der wissenschaftliche Anspruch wird allerdings nur eingelöst, wenn entsprechende Untersuchungen nach wissenschaftlichen Kriterien durchgeführt und ausgewertet werden. Eine Bewertung der forschungsmethodischen Qualität von Evaluationsstudien ist somit unerlässlich, will man zu einer rationalen und reflektierten Entscheidung über die Einsatzmöglichkeiten von Fördermaßnahmen und Präventionsprogrammen kommen. Dafür sprechen unter anderen Befunde über den Einfluss der Forschungsmethodik auf die erzielten Ergebnisse (Wilson & Lipsey, 2001). Die gewählte Forschungsmethodik beeinflusst danach nicht nur grundsätzlich die Aussagekraft einer Untersuchung. Methodisch schwache Studien können zudem zu systematischen Ergebnisverzerrungen führen, die möglicherweise Fehleinschätzungen oder den Einsatz unwirksamer Programme nach sich ziehen. Im Hinblick auf eine Praxisempfehlung und Routine-Implementation eines Präventionsprogramms ist es nicht nur wichtig, ob und mit welchen Ergebnissen ein Präventionsprogramm evaluiert wurde, sondern auch, mit welcher Forschungsmethodik diese Befunde erzielt wurden.

Nun können wissenschaftlich fundierte Evaluationen sehr unterschiedlich aufgebaut und durchgeführt werden (vgl. kurze Einführung in DFK, 2013). In der empirisch orientierten Sozialforschung hat sich eine Position durchgesetzt, die der experimentellen Überprüfung und der Erfassung quantitativer Daten einen hohen Stellenwert zumisst. Im Rahmen dieser Festlegung können unter Forschungsmethodik alle Aspekte einer wissenschaftlichen Untersuchung verstanden werden, die mit der Erfassung sowie Verarbeitung von Befunden und Ergebnissen im Rahmen einer Wirksamkeitsstudie zu tun

Methodisch schwache Studien können zu systematischen Ergebnis-verzerrungen führen.

Die Vorteile eines Globalratings bestehen in der relativ einfachen und zuverlässigen Beurteilung.

haben, insbesondere die Festlegung der Untersuchungsstrategie (dem Forschungsdesign), die Auswahl der Untersuchungsstichprobe, die Datenerhebung (Feststellung von Veränderungen bei der Untersuchungsgruppe) und die statistische Auswertung der Untersuchungsergebnisse.

Zur Bewertung der forschungsmethodischen Qualität von Evaluationsstudien können grob zwei Ansätze unterschieden werden. Ein erster Ansatz versucht, die Studienqualität in einem methodischen Globalrating zu erfassen. Dabei werden Qualitätsangaben zumeist aus dem Studiendesign der Untersuchung abgeleitet. Ein zweiter Ansatz besteht darin, eine mehr oder weniger detaillierte Liste von methodischen Einzelkriterien zu bewerten und ggf. im Anschluss zu einem zusammenfassenden Urteil zu kommen. Die Vorteile eines Globalratings bestehen in der relativ einfachen und damit i.d.R. zuverlässigen Beurteilung, die auch für methodische Laien leicht nachvollziehbar ist. Die Nachteile liegen in der wenig differenzierten Auseinandersetzung mit forschungsmethodischen Fragen, die unter Umständen zu erheblichen Fehleinschätzungen führen können.

2. Das randomisierte Experiment zwischen „Gold-Standard“ und Praxisferne

Gegen diese Logik wird allerdings eingewandt, dass mit der Nutzung bestimmter Forschungsdesigns (z.B. einer randomisierten Kontrollgruppenstudie) zumindest ein gewisser methodischer Standard gesetzt wird, der nur dann gefährdet ist, wenn erhebliche zusätzliche Probleme bei der Untersuchung auftreten. Anhänger dieser Sichtweise plädieren daher dafür, das randomisierte Experiment als sogenannten „Gold-Standard“ in der Evaluationsforschung zu verwenden, zumindest wenn es sich um die Bewertung von abgegrenzten Interventionsprogrammen handelt und nicht etwa die Wirkungen von populationsbasierten Interventionen mit vollständiger Erfassung der Zielgruppen (z.B. Auswirkungen einer neuen Gesetzgebung).

Das randomisierte Experiment zeichnet sich in seiner einfachsten Form dadurch aus, dass mindestens zwei Versuchsgruppen gebildet werden (i.d.R. eine Gruppe, die an einem Präventionsprogramm teilnimmt, und eine Vergleichs- oder Kontrollgruppe, die nicht an dem Programm teilnimmt), die mindestens einmal vor und einmal nach der Durchführung der Maßnahme einer systematischen Datenerhebung (die mit den Zielen der Maßnahme

korrespondiert) unterzogen wird. Die Aufteilung der Untersuchungsgruppe in die Programm- und Kontrollbedingung erfolgt dabei zufällig (randomisiert), d.h. jeder Studienteilnehmer hat die gleiche Chance in eine der beiden Gruppen gelost zu werden. Mit dieser Zufallsaufteilung auf unterschiedliche Gruppen soll garantiert werden, dass sich die Vergleichsgruppen möglichst gleichen und keine interventionsrelevanten Unterschiede (etwa im Alter, der Geschlechteraufteilung oder in der Motivation, am Programm teilzunehmen) entstehen. Damit sollen Ergebnisverzerrungen vermieden und irrelevante Einflüsse sogenannter Störvariablen kontrolliert werden. Dies geschieht allerdings nur, wenn die Randomisierungsannahmen nicht durch andere Probleme konterkariert werden. Dies kann z.B. der Fall sein, wenn die Stichprobengröße gering (z.B. unter 30) oder die Ausfallrate (d.h. Anzahl der Personen, die aus der Studie aussteigen) während der Studie hoch ist. Darüber hinaus wird die Praktikabilität von randomisierten Experimenten bei Untersuchungen in realen Settings (Feldstudien) durchaus angezweifelt (z.B. in Schulen, wo zumeist eine klassenweise Förderung vorgenommen wird und eben nicht jedes Kind die gleiche Chance hat, in die Präventions- oder Vergleichsgruppe gelost zu werden). Zudem unterlaufen unkalkulierbare Vorkommnisse nicht selten die Randomisierungsprozedur, wenn etwa Personen aus welchen Gründen auch immer nicht mehr an der Maßnahme oder der Studie teilnehmen wollen. Aus diesen Gründen wird vorgeschlagen, zur methodischen Bewertung von Evaluationsstudien weitere methodische Einzelkriterien zu erfassen. Dazu haben etwa Cook und Campbell (1979) eine umfängliche Liste von methodischen Gefährdungen der Validität (d.h. der Gültigkeit von Forschungsergebnissen) vorgelegt. Die Grundidee dieses Modells besteht darin, anhand methodischer Überlegungen Alternativerklärungen und Interpretationsprobleme auszuschließen, um damit zu einer besonders gültigen und wahren Aussage über die Wirkung eines Programms zu kommen.

Die Praktikabilität von randomisierten Experimenten wird bei Untersuchungen in realen Settings (Feldstudien) durchaus angezweifelt.

Andere Forscher haben dieses Modell der Validitätsgefährdungen aufgegriffen und schlagen zum Beispiel eine Bewertung von relativ breiten methodischen Unteraspekten vor (z.B. Eindeutigkeit des kausalen Schlusses, Repräsentativität der Studie), die je nach Bedarf

differenzierter gestaltet werden können (z. B. Valentine & Cooper, 2008), oder wählen Aspekte je nach ihrer Bedeutung im betrachteten Forschungsfeld aus (z.B. MacLeod & Weisz, 2004). Unabhängig vom konkreten Vorgehen liegen die Vorteile eines derartigen Vorgehens in der detaillierten Bewertung einer Evaluationsstudie. Die Nachteile betreffen die zum Teil schwierigen Beurteilungsprozesse, die unter Umständen auch dadurch erschwert werden, dass relevante Informationen in den Studienberichten fehlen oder nur indirekt zu erschließen sind.

3. Vorschlag für ein einheitliches Bewertungssystem von Evaluationsstudien

Vor diesem Hintergrund verfolgt das im Folgenden dargestellte Methodenprofil den Zweck, ein einheitliches forschungsmethodisches Bewertungssystem von Evaluationsstudien bereitzustellen. Dabei haben wir versucht, die unterschiedlichen Beurteilungsmethoden für die Einschätzung von Präventionsstudien miteinander zu kombinieren und schlagen vor, einerseits ein methodisches Globalrating anzuwenden und anderseits eine relativ detaillierte Einschätzung methodischer Einzeleffekte vorzunehmen. Selbstverständlich lassen sich über die ausgewählten Aspekte wie auch über die Bewertungsaspekte selbst aus forschungsmethodischer und wissenschaftstheoretischer Sicht sehr kontroverse Diskussionen führen. Zudem kann die folgende Darstellung nicht als Einführung in die Evaluationsforschung angesehen werden, in der alle verwendeten Kriterien und Einschätzungen ausreichend begründet werden. Interessierte Leser seien hier an zwei ausgezeichnete (allerdings englischsprachige) Standardwerke verwiesen (Rossi, Lipsey & Freeman, 2004; Shadish, Cook & Campbell, 2002). Das nachfolgende Bewertungssystem versteht sich somit als eine mögliche Variante der forschungsmethodischen Begutachtung, die allerdings auf intensiven Erfahrungen im Bereich der Evaluationsforschung und auf einer Vielzahl von Evaluationsbewertungen im Rahmen von Meta-Analysen und integrativen Forschungszusammenfassungen basiert.

Das Bewertungssystem versteht sich als eine mögliche Variante der forschungsmethodischen Begutachtung.

4. Ein Methodenprofil für Präventionsstudien

Das vorgeschlagene Methodenprofil besteht aus insgesamt zehn Aspekten, die bei einer konkreten Wirksamkeitsstudie eingeschätzt werden, um zu einer einheitlichen Bewertung der forschungsmethodischen Qualität zu kommen. Darin enthalten sind ein Globalrating zum Studiendesign sowie neun ausgewählte Einzelaspekte. Die Auswahl multipler Beurteilungskriterien trägt dem Umstand Rechnung, dass die forschungsmethodische Qualität einer empirischen Studie in der Regel mehrdimensionale Einschätzungen verlangt und in einem einzigen Globalrating nur unzureichend abgebildet werden kann. Selbstverständlich beeinflusst etwa die Qualität des Forschungsdesigns im Globalrating in nicht unwesentlichem Maße mehrere der nachstehenden Einzelkriterien, wie auch diese untereinander eine gewisse Abhängigkeit aufweisen (z.B. die Qualität der statistischen Auswertungen hängt i.d.R. auch mit der mit Qualität der Erfolgsmessung zusammen). Diese Abhängigkeit wird aber zugunsten einer detaillierteren Beurteilung in Kauf genommen, die unseres Erachtens der Forschungspraxis gerechter wird als alleinige Globalbeurteilungen. Im Einzelnen werden folgende zehn Aspekte (A1 - A10) zu einer Wirksamkeitsstudie eingeschätzt:

A1	Ein Globalrating zum Studiendesign
A2	Die Qualität des Vergleichsmaßstabs
A3	Die Kontrolle weiterer Einflussfaktoren (z.B. konfundierende Ereignisse)
A4	Die Qualität des statistischen Materials und der statistischen Auswertungen
A5	Die Qualität der Erfolgsmessung (Breite, Kriterienqualität)
A6	Die Qualität der Implementation der Maßnahme
A7	Die Repräsentativität für die Praxis (Zielgruppe, Setting, Durchführungsbedingungen)
A8	Die Angemessenheit der Ergebnisinterpretation
A9	Die Qualität der Studiendokumentation
A10	Die Kontrolle von Interessenkonflikten

Das Globalrating zum Studiendesign (A1) wird sechsstufig vorgenommen (s.u.) und zu den Aspekten A2 bis A10 werden jeweils dreistufige Qualitätseinschätzungen vergeben:

- mehrere oder gravierende Qualitätseinschränkungen (geringe Aussagekraft)
- höchstens eine bedeutsame oder mehrere moderate Qualitätseinschränkungen (mittlere Aussagekraft)
- keine oder nur geringfügige Qualitätseinschränkungen (hohe Aussagekraft)

Zudem ist die Einschätzung *nb = nicht beurteilbar* in den Fällen vorgesehen, in denen eine Bewertung nicht zuverlässig vorgenommen werden kann (wenn etwa wichtige Aspekte in den Studienberichten nicht ausreichend ausführlich erläutert werden).

A1 Globalrating zum Studiendesign

Zunächst werden Studien anhand ihrer Untersuchungsstrategie (ihres Forschungsdesigns) eingeschätzt. Dabei werden sechs Qualitätsstufen angewandt, die einer modifizierten bzw. ausdifferenzierten Form der *Maryland-Scale of Scientific Rigor* entsprechen (vgl. Farrington, 2003), die international bereits seit einiger Zeit für die Qualitätseinschätzung von Evaluationsstudien eingesetzt wird. Danach können sechs Qualitätseinstufungen vorgenommen werden, die sich im Wesentlichen auf drei Studienmerkmale beziehen, nämlich (1) wie oft eine Datenerhebung erfolgte, (2) inwieweit Vergleichsgruppen vorliegen und (3) inwieweit diese Vergleichsgruppen nach wichtigen interventionsrelevanten Merkmalen vergleichbar sind bzw. nach welchen Prinzipien Studienteilnehmer in verschiedene Vergleichsgruppen aufgeteilt wurden.

Globalrating zum Studiendesign (A1)	
1	**Korrelative Designs:** Die Untersuchung besteht ausschließlich aus korrelativen Daten, d.h. die Teilnahme an einem Programm wird mit dem aktuellen Verhalten, mit der eingeschätzten Veränderung des Verhaltens oder mit Daten zur Zufriedenheit der Teilnehmer mit dem Programm in Verbindung gebracht. Daten zur Wirksamkeit werden bei diesen Untersuchungsdesigns vorwiegend aus Angaben von Programmteilnehmern nach einer Maßnahme erfasst (z.B. Zufriedenheitsangaben, Einschätzungen von Veränderungen oder Verbesserungen durch die Intervention).
2	**Vorher-Nachher-Design ohne angemessene Vergleichsgruppe:** Es werden Vergleiche zwischen den Daten vor und nach einer Intervention angestellt, diese Ergebnisse aber nicht mit einer angemessenen Vergleichsgruppe kontrastiert. Dazu gehören beispielsweise auch retrospektive Designs (Vorher-Daten werden rückwirkend erfasst), Designs mit Eigenwartegruppe (Vergleich wird über den Verlauf der Interventionsgruppe vor der Intervention hergestellt) und Zeitreihen ohne Vergleichsgruppe oder Untersuchungsdesigns, in denen Teilnehmer mit Aussteigern einer Intervention verglichen oder vorhandene normative Daten (sogenannte generische Kontrollen) zum Vergleich herangezogen werden.

3	**Quasi-Experiment mit angenommener Vergleichbarkeit der Versuchsgruppen:** Ein Quasi-Experiment verlangt mindestens zwei Versuchsgruppen (z.B. mit und ohne Intervention) und mindestens zwei Erhebungszeitpunkte (vor und nach der Intervention). Als Vergleichsgruppe wird eine sogenannte nicht-äquivalente Gruppe herangezogen, die nach wichtigen Variablen vergleichbar sein sollte (z.B. zwei Schulklassen einer Schulstufe). Es liegen jedoch keine expliziten Nachweise zur Vergleichbarkeit hinsichtlich bedeutsamer und präventionsrelevanter Merkmale vor. Weitere Designs, die diesem Typus zugeordnet werden können sind unterbrochene Zeitreihen mit nicht-äquivalenter Kontrollgruppe sowie das Regression-Diskontinuitäts-Design.
4	**Quasi-Experiment mit demonstrierter Vergleichbarkeit oder mit Verwendung bestimmter Kontrolltechniken:** Designs dieser Qualität entsprechen dem Rating 3, nur wird die Vergleichbarkeit der Versuchsgruppen entweder explizit hergestellt oder aposterio überprüft. Techniken zur Konstruktion von vergleichbaren Kontrollgruppen sind z.B. bestimmte Matching Prozeduren (Parallelisierung, Propensity Score Matching). Aposterio Techniken beziehen sich auf statistische Verfahren wie etwa der statistische Vergleich hinsichtlich relevanter Parameter, Kovarianzanalysen oder regressionsbasierte Auswertungsverfahren.
5	**Randomisiertes Experiment mit Problemen in der Vergleichbarkeit der Versuchsgruppen oder anderen Validitätseinschränkungen:** Es wird mindestens ein Design wie unter 3 angewandt (zwei Messzeitpunkte, zwei Gruppen) und zusätzlich werden die Personen zufällig (randomisiert) auf die Versuchsgruppen aufgeteilt. Es liegen aber Einschränkungen der Aussagekraft vor, da trotz Randomisierung wichtige Unterschiede zwischen den Vergleichsgruppen (z.B. zufällige Unterschiede bei geringer Stichprobengröße) existieren bzw. Validitätsprobleme wie etwa differentieller Dropout (mehr als 10%) aufgetreten sind. Dabei werden unterschiedliche Randomisierungsstrategien (individuelle, Block-, Clusterrandomisierung) berücksichtigt.
6	**Randomisiertes Experiment mit Vergleichbarkeit/ohne Einschränkungen:** Dieses Rating bezieht sich auf Designs wie unter 3 mit einer Randomisierung wie unter 5, jedoch mit zusätzlich demonstrierter Vergleichbarkeit der Versuchsgruppen und dem Ausschluss wichtiger Validitätsgefährdungen.

A2 Qualität des Vergleichsmaßstabs

Um fundierte Aussagen hinsichtlich der Wirksamkeit einer Intervention treffen zu können, ist ein zuverlässiger Vergleichsmaßstab von zentraler Bedeutung. Dieser Vergleichsmaßstab wird i.d.R. durch eine vergleichbare Gruppe von Untersuchungsteilnehmern erzeugt, die nicht an einem Programm teilnehmen und deren Ergebnisse mit denen einer geförderten Gruppe verglichen werden. Das Ausmaß der Vergleichbarkeit dieser (im einfachsten Fall zwei) Versuchsgruppen (Interventions- und Kontrollgruppe) ist insofern ein wichtiges methodisches Qualitätskriterium, als aus diesem Vergleich unmittelbar auf die Wirksamkeit der Maßnahme geschlossen wird. Liegt kein oder nur ein unzureichender Vergleichsmaßstab vor, können massive Ergebnisverzerrungen oder Fehleinschätzungen resultieren, weil nicht zweifelsfrei entschieden werden kann, ob erzielte Wirkungen durch die Programmteilnahme oder aus einem anderen Grund entstanden sind.

Die Vergleichbarkeit zweier Gruppen bezieht sich dabei auf interventionsrelevante Merkmale, d.h. auf alle Faktoren, die die Erfolgskriterien potentiell beeinflussen können. Hierzu zählen sowohl demographische (z.B. Alter, Geschlecht, Bildungsstand) und andere personengebundene Faktoren (z.B. Persönlichkeit, Motivation), als auch die Ausprägung der Erfolgskriterien vor der Intervention (d.h. zur Vorher-Messung) selbst. Von einer hohen Vergleichbarkeit wird bei einer randomisierten Zuweisung der Studienteilnehmer zu den Versuchsgruppen ausgegangen (vgl. Globalrating unter Punkt 1). Mittels der Randomisierung wird (bei ausreichend großer Stichprobe, d.h. mindestens 30 Personen pro Versuchsgruppe) mit hoher Wahrscheinlichkeit eine Gleichverteilung aller beobachteten und nicht-beobachteten Faktoren vor der Intervention erreicht. Weitere Möglichkeiten vergleichbare Versuchsgruppen zu erhalten, bieten sogenannte Matching-Verfahren. Dabei wird versucht, die Versuchsgruppe vorab vergleichbar zu gestalten, indem z.B. eine Gleichverteilung personenbezogener Merkmale (Alter, Geschlecht) bewusst hergestellt wird oder nach unterschiedlichen Merkmalen vergleichbare Personen zusammengestellt werden (sogenannte Parallelisierung). Auch bei sehr guten Designs, die eigentlich eine Gleichverteilung der Merkmale garantieren (Globalratings von 3 bis 6), sollte eine Überprüfung der Vergleichbarkeit der Versuchsgruppen hinsichtlich relevanter Variablen erfolgen. Dies setzt jedoch entsprechende Vorüberlegungen (welche Merkmale sind wichtig?) sowie die Erfassung potentiell bedeutsamer Einflussgrößen voraus. Weiterhin kann die Vergleichbarkeit von Versuchsgruppen auch während der Studie beeinträchtigt werden, wenn etwa ein differentieller Ausfall auftritt (d.h. in den Versuchsgruppen scheiden unterschiedlich viele und auch jeweils andere Studienteilnehmer aus). Liegen systematische

Unterschiede vor, können statistische Kontrollverfahren wie die Kovarianzanalyse oder regressionsanalytische Auswertungen eingesetzt werden, die diese Unterschiede bei den Auswertungen mit berücksichtigen. Solche Techniken erlauben trotz der aufgetretenen Unterschiede vor der Maßnahme eine einigermaßen unverzerrte Interpretation der Ergebnisse.

Einschätzungen zur Qualität des Vergleichsmaßstabes (A2)	
●	Eingeschränkte Vergleichbarkeit der Versuchsgruppen ohne Anwendung angemessener statistischer Kontrollverfahren oder kein geeigneter Vergleichsmaßstab.
●	Quasi-experimentelles Design oder eingeschränkte Vergleichbarkeit trotz Randomisierung, mit (nachträglicher) statistischer Kontrolle.
●	Randomisierte Zuweisung bei hinreichender Stichprobengröße (≥ 30 pro Gruppe) und demonstrierte Vergleichbarkeit (keine signifikanten Vortestunterschiede zwischen den Versuchsgruppen sowie keine Beeinträchtigungen der Vergleichbarkeit im Untersuchungsverlauf).

A3 Kontrolle weiterer Einflussfaktoren

Um eindeutige kausale Schlüsse hinsichtlich der Effekte eines Präventionsprogramms ziehen zu können, muss eine Konfundierung der Interventionswirkung mit möglichen Störvariablen (d.h. Einflussfaktoren, die neben der Intervention einen Einfluss auf die Erfolgskriterien haben) reduziert werden. Nur so kann mit hoher Sicherheit auf die Wirkung eines Interventionsprogramms geschlossen werden. Nicht-programmbezogene Einflüsse liegen z.B. vor, wenn zwischenzeitliche Ereignisse (z.B. die öffentliche Diskussion eines Amoklaufs während eines Gewaltpräventionsprogramms), die natürliche Entwicklung (z.B. Schwankungen im Aggressionsniveau mit dem Alter) oder wiederholte Erhebungen (sogenannte Test-Effekte) mit Veränderungen in den Erfolgskriterien einhergehen. Diese Einflüsse können i.d.R. durch sehr gut vergleichbare Kontrollgruppen bei ausreichend großer Stichprobe kontrolliert werden. Zusätzliche Probleme ergeben sich z.B. durch Interventionen in der Kontrollgruppe (Teilnehmer werden auch von anderer Seite gefördert), Kontakte zwischen den Versuchsgruppen (z.B. Interventions- und Kontrollklassen treffen sich in Schulpausen und tauschen Inhalte des Programms aus) oder bestimmten Reaktionen auf die Gruppenzuweisung (ausgleichende Rivalität oder Demoralisierung in der Kontrollgruppe). Diese Faktoren können zu Effekten führen, die fälschlicherweise der Intervention zugeschrieben werden und sind somit bei der Untersuchungsplanung und -durchführung zu beachten und ggf. zu dokumentieren.

Einschätzungen zur Kontrolle weiterer Einflussfaktoren (A3)	
●	Design ohne Kontrollgruppe oder Vorliegen einer Kontrollgruppe ohne demonstrierte Vergleichbarkeit oder anderweitige Indikationen für Störfaktoren oder Konfundierungen aufgrund des Designs und der Kontextbedingungen.
●	Vorliegen einer Kontrollgruppe mit eingeschränkter Vergleichbarkeit oder anderweitige Indikationen für Störfaktoren oder Konfundierungen aufgrund des Designs und der Kontextbedingungen.
●	Vorliegen einer Kontrollgruppe mit demonstrierter Vergleichbarkeit; keine oder nur geringfügige Indikationen für Störfaktoren oder Konfundierungen aufgrund des Designs und der Kontextbedingungen.

A4 Qualität des statistischen Materials und der statistischen Auswertungen

Die Aussagekraft des statistischen Materials ist zunächst von einer vollständigen, korrekten und ausführlichen Berichterstattung der statistischen Befunde abhängig. Darüber hinaus hängt die statistische Qualität mit der Auswahl adäquater Auswertungsmethoden einschließlich der zu überprüfenden Anwendungsvoraussetzungen zusammen. Im Hinblick auf die Schätzgenauigkeit der statistischen Parameter sowie die Identifikation von (in der Präventionsforschung häufig anzutreffenden) eher kleinen Effekten sollte eine ausreichend große Stichprobe zugrunde gelegt werden (sogenannte statistische Power). Die Frage der Stichprobengröße ist von maßgeblicher Bedeutung für die Anwendung statistischer Auswertungsprozeduren. Die notwendige Stichprobengröße kann stark variieren, mindestens sollte jedoch eine Teilnehmerzahl von 30 pro Versuchsbedingung vorliegen. Weiterhin sind Auswirkungen multipler statistischer Vergleiche zu beachten und mittels multivariater Verfahren oder einer Korrektur des kumulierten Alphafehlers zu adjustieren. Zudem sollte das Ausmaß fehlender Werte (missing data) und eine entsprechende Schätzung der Daten (Datenimputation) genausten beschrieben oder weiterführende Vergleichsanalysen (z.B. ein Vergleich von Teilnehmern vs. ausgeschiedenen Teilnehmern) durchgeführt werden. Dies gilt vor allem für differentielle Ausfälle von mehr als 10 % des Stichprobenumfangs.

Einschätzungen zur Kontrolle weiterer Einflussfaktoren (A4)	
●	Gravierende Einschränkungen der statistischen Aussagekraft aufgrund mehrfacher Nichteinhaltung von Auswertungsstandards (geringe Power, Fehlen von Anwendungsvoraussetzungen, fehlerhafte Anwendung statistische Verfahren).
●	Adäquate statistische Analysen samt Adjustierung multipler Vergleiche; eingeschränkte statistische Aussagekraft aufgrund eines unzureichenden Stichprobenumfangs oder nennenswerte Probleme mit fehlenden Daten.
●	Adäquate statistische Analysen bei ausreichend großer statistischer Power; ggf. Anwendung erforderlicher statistischer Korrektur- und Kontrollmaßnahmen.

A5 Qualität der Erfolgsmessung

Die Auswahl geeigneter Erfolgsmaße setzt eine theoretisch und wissenschaftlich fundierte Operationalisierung der Erfolgskriterien (abhängigen Variablen) entsprechend der Fragestellung und Programmziele voraus. Zudem sollten vorrangig etablierte Erhebungsverfahren, die bestimmten psychometrischen Gütekriterien (Objektivität, Reliabilität, Validität) genügen, sowie alltagsrelevante (augenscheinvalide) Kriterien (z.B. Polizeikontakte) eingesetzt werden. Von der Verwendung sogenannter Ad-hoc-Instrumente (z.B. selbstkonstruierten Fragebögen) ohne Nachweis seiner psychometrischen Qualitäten ist abzuraten. Um die Befunde zuverlässig interpretieren und generalisieren zu können, ist eine möglichst breite Erfolgsmessung zu empfehlen. Wünschenswert wäre die Erfassung mehrerer Inhaltsbereiche (multi-modale Messung bestehend z.B. aus der Erfassung von sozialer Kompetenz und aggressivem Verhalten), die Erhebung mit unterschiedlichen Erhebungsmethoden (multi-methodale Messung bestehend z.B. aus Daten aus Fragebögen und Verhaltensbeobachtung) sowie die Erfassung von Daten mehrerer Informationsquellen (multi-informante Erhebung bestehend z.B. aus Selbsteinschätzungen und Befragungen von Eltern oder Lehrern). Zudem wäre eine Ergebnismessung unter Verwendung proximaler (bezogen auf unmittelbare Programmziele) und distaler (bezogen auf langfristige Ziele) Erfolgskriterien optimal.

Einschätzungen zur Qualität der Erfolgsmessung (A5)	
●	Gravierende Einschränkungen in der Qualität der Erfolgsmessung aufgrund von Operationalisierungsdefiziten, einseitigen Erhebungsstrategien oder der Verwendung von Instrumenten geringer oder nicht dokumentierter psychometrischer Qualität.
●	Gewisse Einschränkungen in der Qualität der Erfolgsmessung aufgrund unzureichender Operationalisierungen, einseitigen Erhebungsstrategien oder der Verwendung von Instrumenten geringer oder nicht dokumentierter psychometrischer Qualität.
●	Keine oder nur geringfügige Einschränkungen in der Qualität der Erfolgsmessung; theoriegeleitete Auswahl und Operationalisierung der abhängigen Variable(n); Anwendung zuverlässiger Instrumente; umfassende und multiple Erfassung relevanter Erfolgskriterien.

A6 Qualität der Implementation

Neben der wissenschaftlichen Fundierung der Inhalte einer Intervention zeigt sich auch deren qualitativ hochwertige Umsetzung als wirkungsrelevant (Beelmann & Karing, 2014; Durlak & DuPre, 2008). Hierbei ergeben sich unterschiedliche Aspekte, die sich mit der Frage befassen, ob das Programm zuverlässig und wie geplant durchgeführt werden konnte. Probleme der Implementation können sich auf verschiedenen Ebenen ergeben: bei der Auswahl der Zielgruppe (z.B. intendierte Zielgruppe konnte nicht erreicht werden), bei der Durchführung des Programms (z.B. die Inhalte werden nur zum Teil umgesetzt oder das Programm willkürlich verändert), im Hinblick auf die Mitarbeit und Motivation der Zielgruppe (z.B. Ausscheiden von Personen, mangelhafte Mitarbeit), bei der Auswahl und der professionellen Kompetenz der Administratoren (z.B. geringes Engagement, negative Einstellungen zum Programm) sowie durch finanzielle sowie institutionelle Rahmenbedingungen (z.B. notwendiges Zeitbuget wird durch die Schule nicht bereit gestellt). Maßnahmen zur Verbesserung der Implementation beziehen sich beispielsweise auf die Manualisierung der Intervention sowie die Ausbildung und Supervision der Programmadministratoren. Eine ausführliche Dokumentation der Durchführung einschließlich möglicher Veränderungen der Intervention/des Programm-Manuals ist ebenso unabdingbar wie Verfahren der Implementationskontrolle (Monitoring, Erfassung der Mitarbeit, Abbruchrate etc.). Unterschiede in der Interventionswirkung aufgrund systematischer Variationen der Implementationsqualität (z.B. bei unterschiedlichen Administratoren) sollten nach Möglichkeit überprüft werden.

Einschätzungen zur Qualität der Implementation (A6)	
🔴	Hinweise auf gravierende Einschränkungen der Implementationsqualität; mangelnde Implementationskontrolle.
🟡	Hinweise auf eine eingeschränkte Realisierung der intendierten Programmimplementation (z.B. Haltequote in der Maßnahme unter 90%); Maßnahmen zur Implementationskontrolle werden aber angewandt.
🟢	Keine oder nur geringfügige Probleme hinsichtlich der Programmrealisierung; von konzepttreuer Durchführung kann ausgegangen werden (z.B. Vorliegen eines detaillierten Manuals); Maßnahmen zur Sicherung der Implementationsqualität sowie Implementationskontrollen sind dokumentiert.

A7 Repräsentativität für die Praxis

Aussagen bezüglich der Wirksamkeit einer Intervention unter natürlichen Bedingungen erfordern die Repräsentativität der Stichprobe und der Kontextbedingungen einer Studie. Dies bedeutet: (a) die Erfassung einer möglichst repräsentativen Stichprobe, die die Zielpopulation einer Maßnahme ausreichend gut abbildet, und (b) die Auswahl eines Settings sowie Rahmen- und Durchführungsbedingungen entsprechend der realen Versorgungspraxis (z.B. Durchführung im Rahmen des regulären Schulunterrichts). Probleme der Repräsentativität treten etwa auf, wenn nur bestimmte Zielgruppen (z.B. aus städtischen Kontexten) ausgewählt wurden oder besonders günstige Rahmenbedingungen vorlagen (z.B. deutlich mehr Personal als es normalerweise in der Praxis verfügbar wäre, um die Maßnahmen durchzuführen).

Einschätzungen zur Repräsentativität für die Praxis (A7)	
🔴	Gravierende Einschränkungen hinsichtlich der Repräsentativität in mindestens zwei Bereichen (Stichprobe, Setting).
🟡	Gravierende Einschränkungen hinsichtlich der Repräsentativität in einem Bereich (Stichprobe, Setting).
🟢	Repräsentativität der Stichprobe und der Rahmenbedingungen; Generalisierbarkeit auf die vorgesehene Zielgruppe und die reale Versorgungspraxis möglich.

A8 Angemessenheit der Ergebnisinterpretation

Die Ergebnisse werden entsprechend der Befundlage und der Fragestellung sowie unter Berücksichtigung methodischer Limitationen (Validitätseinschränkungen) interpretiert. Dabei erfolgt eine ausführliche Berichtlegung hypothesenkonformer als auch nicht-konformer Resultate unter Bezugnahme auf den aktuellen Forschungsstand sowie eine reflektierte Diskussion der Untersuchung und möglicher validitätseinschränkender Aspekte oder Ergebnisverzerrungen. Bei geringer forschungsmethodischer Qualität kann sowohl eine Überschätzung der Programm-Effekte (sogenannter Alpha-Fehler) oder auch eine Unterschätzung (Beta-Fehler) resultieren.

Einschätzungen zur Angemessenheit der Ergebnisinterpretation (A8)	
●	Unangemessene, verzerrte Interpretation der Ergebnisse; fehlende kritische Auseinandersetzung mit Limitationen und validitätseinschränkenden Aspekten der Studie.
●	Angemessene, reflektierte und sachliche Interpretation der Ergebnisse; unzureichende Diskussion möglicher Limitationen und Validitätsgefährdungen.
●	Angemessene, reflektierte und sachliche Interpretation der Ergebnisse; kritische Diskussion möglicher Limitationen und Validitätsgefährdungen.

A9 Qualität der Studiendokumentation

Für eine angemessene Interpretation der Programmeffekte, die Replikation der Untersuchung und den Vergleich verschiedener Studien ist eine ausführliche, nachvollziehbare und vollständige Dokumentation erforderlich. Dies bezieht sich sowohl auf die Beschreibung von Fragestellung und Zielsetzung als auch auf die Schilderung der gesamten Untersuchungsmethodik inklusive Selektions- und Zuweisungsmodalitäten, Stichprobencharakteristika, Programminhalte, Rahmenbedingungen der Untersuchung, Erfolgsmaße und experimentelle Mortalität. Weiterhin sollte eine detaillierte Berichtlegung der statistischen Analysen und der Befundlage erfolgen, die relevante Kennwerte aller Erfolgskriterien umfasst (Primär- und Teststatistiken, Signifikanzniveaus, Effektgrößen). Tabellen, Graphiken und Flussdiagramme dienen der Veranschaulichung.

Einschätzungen zur Qualität der Studiendokumentation (A9)	
●	Gravierende Einschränkungen in der Studiendokumentation; Unstimmigkeiten oder Auslassungen in der Berichtslegung.
●	Partielle Einschränkungen in der Studiendokumentation; teilweise unzureichende Berichtslegung.
●	Keine oder nur geringfügige Einschränkungen in der Studiendokumentation; vollständige und stringente Berichtslegung in allen Bereichen.

A10 Kontrolle von Interessenkonflikten

Meta-Analysen konnten wiederholt zeigen, dass die Wirksamkeit eines Interventionsprogramms auch mit dem Status der Studienautoren variiert. Sind Autoren zugleich auch die Programmentwickler (oder sind mit ihnen assoziiert) werden in der Regel höhere Effekte berichtet. Dies ist möglicherweise Folge einer besseren Implementation der Intervention durch die Programm-Autoren. Eine alternative Erklärung für diesen Zusammenhang bilden Interessenkonflikte und damit einhergehende intentionale als auch unbewusste Verzerrungen der Ergebnismessung, -darstellung, -interpretation sowie -publikation. Interessenkonflikte treten vor allem in Situationen auf, in denen professionelle Kriterien (z.B. Wissenschaftsstandards) mit Eigeninteressen der durchführenden Personen konkurrieren. Wesentliche Quellen dieses Konflikts bilden einerseits finanzielle (z.B. Urheberrechte, Honorare, Lizenzen, Fördergelder) und andererseits immaterielle Interessen (z.B. Ansehen, erhöhte Wahrscheinlichkeit weiterer erfolgreicher Publikationen, Vertreten eines persönlichen Standpunktes). Die Wahrscheinlichkeit eines potentiellen Interessenkonflikts wird hierbei von drei Variablen beeinflusst: (a) Es besteht eine enge Verbindung zwischen Programmentwicklern und Evaluatoren, (b) es erfolgt eine kommerzielle Verbreitung des gesamten Programms oder einzelner Elemente, (c) es wird eine gewinnorientierte Vermarktung des Programms angestrebt (vgl. Eisner & Humphreys, 2011).

Einschätzungen zu Interessenkonflikten (A10)	
●	Studienautor ist Programmentwickler bzw. Mitarbeiter; Programm wurde kommerziell verbreitet; gewinnorientierte Vermarktung.
●	Studienautor ist Programmentwickler bzw. Mitarbeiter; Programm wurde (bisher) nicht kommerziell verbreitet; keine gewinnorientierte Vermarktung.
●	Unabhängige Evaluation, d.h. keiner der Studienautoren ist Programmentwickler, Mitarbeiter des Programmentwicklers oder Lizenzhalter.

5. Fazit

Mit den genannten Einschätzungen sollte eine hinreichend zuverlässige wie umfängliche Einschätzung der Aussagekraft möglich sein, die in der Gesamtschau einen raschen Überblick zu den jeweiligen Einschätzungen erlaubt. Die folgende Tabelle 1 zeigt ein exemplarisches Methodenprofil. Es ist geplant, entsprechende Methodenprofile der deutschsprachigen Evaluationsstudien zu Programmen im Bereich „Entwicklungsförderung und Gewaltprävention“ (vgl. Beelmann, Pfost & Schmitt, 2014) im Webportal des DFK (■ www.wegweiser-praevention.de) zu publizieren. In einheitlich strukturierten Steckbriefen sollen zudem die Studienergebnisse dokumentiert und – auch für vergleichende Einschätzungen - zur Verfügung gestellt werden.

Methodenkriterium		Einschätzung
A1	Globalrating zum Studiendesign	4 / 6 *
A2	Qualität des Vergleichsmaßstabs	●
A3	Kontrolle weiterer Einflussfaktoren	●
A4	Qualität des statistischen Materials/der Auswertungen	●
A5	Implementationsqualität	●
A6	Repräsentativität für die Praxis	●
A7	Qualität der Erfolgsmessung	●
A8	Angemessenheit der Ergebnisinterpretation	●
A9	Qualität der Studiendokumentation	●
A10	Kontrolle von Interessenkonflikten	nb
Gesamteinschätzung		●

Tabelle 1: Exemplarische Zusammenfassung der Einschätzungen im Methodenprofil
Anmerkungen: ● = geringe, ● = mittlere, ● = hohe Aussagekraft, nb = nicht beurteilbar
* 4 / 6 = Qualitätsstufe 4 von möglichen 6 Stufen erreicht

Literatur

Beelmann, A. (2015). Konstruktion und Entwicklung von Interventionsmaßnahmen. In W. Melzer, D. Hermann, U. Sandfuchs, M. Schäfer, W. Schubarth & P. Daschner (Hrsg.), Handbuch Aggression, Gewalt und Kriminalität bei Kindern und Jugendlichen (S. 340-346). Bad Heilbrunn: Klinkhardt.

Beelmann, A. & Karing, C. (2014). Implementationsfaktoren und –prozesse in der Präventionsforschung: Strategien, Probleme, Ergebnisse, Perspektiven. Psychologische Rundschau, 65, 129-139.

Beelmann, A., Pfost, M. & Schmitt, C. (2014). Prävention und Gesundheitsförderung bei Kindern und Jugendlichen. Eine Meta-Analyse der deutschsprachigen Wirksamkeitsforschung. Zeitschrift für Gesundheitspsychologie, 22, 1-14.

Cook, T. D. & Campbell, D. T. (1979). Quasi-experimentation. Design and analysis issues for field settings. Bosten: Houghton Mifflin.

Durlak, J. A. & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. American Journal of Community Psychology, 41, 327-350.

Eisner, M., & Humphreys, D. (2011). Measuring conflict of interest in prevention and intervention research. A feasibility study. In T. Bliesener, A. Beelmann, & M. Stemmler (Hrsg.), Antisocial behavior and crime. Contributions of developmental and evaluation research to prevention and intervention (S. 165-180). Cambridge: Hogrefe.

Farrington, D. P. (2003). Methodological quality standards for evaluation research. Annals of the American Academy of Political and Social Science, 587, 49-68.

MacLeod, B. D. & Weisz, J. R. (2004). Using dissertations to examine potential bias in child and adolescent clinical trials. Journal of Consulting and Clinical Psychology, 72, 235-251.

Rossi, P. H., Lipsey, M. W. & Freeman, H. E. (2004). Evaluation. A systematic approach (7. Aufl.). Thousand Oakes: Sage.

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

Stiftung Deutsches Forum für Kriminalprävention (Hrsg.). (2013). Entwicklungsförderung und Gewaltprävention für junge Menschen. Impulse des DFK-Sachverständigenrates für die Auswahl und Durchführung wirksamer Programme. Ein Leitfaden für die Praxis. Bonn: DFK.

Valentine, J. C. & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). Psychological Methods, 13, 130-149.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. Psychological Methods, 6, 413-429.

Autorenverzeichnis

Marion Altenburg–van Dieken,
Projekt des Hessischen Kultusministers “Gewaltprävention und Demokratielernen“ (GuD) Frankfurt, Landeskoordinatorin Hessen-Süd
Email: marion.altenburg@kultus.hessen.de

Andreas Beelmann, Professor Dr.
Friedrich-Schiller-Universität Jena, Institut für Psychologie, Abteilung für Forschungssynthese, Intervention, Evaluation
Email: andreas.beelmann@uni-jena.de

Judith Hercher, M.Sc. Psych.
Wissenschaftliche Mitarbeiterin Friedrich-Schiller-Universität Jena, Institut für Psychologie, Abteilung für Forschungssynthese, Intervention, Evaluation

Wolfgang Kahl, Dipl. Kfm.
Projekt- und Redaktionsleiter bei der Stiftung Deutsches Forum für Kriminalprävention
Email: wolfgang.kahl@bmi.bund.de

Saskia Niproschke, M.A.,
Wissenschaftliche Mitarbeiterin an der Universität Potsdam, Professur für Erziehungs- und Sozialisationstheorie
Email: niprosch@uni-potsdam.de

Helmolt Rademacher, Dipl.Päd.
Projekt des Hessischen Kultusministers “Gewaltprävention und Demokratielernen“ (GuD) Frankfurt, Projektleiter
Email: helmolt.rademacher@kultus.hessen.de

Herbert Scheithauer, Professor Dr.,
Freie Universität Berlin, Professur für Entwicklungspsychologie und Klinische Psychologie, Arbeitsbereich Entwicklungswissenschaft & Angewandte Entwicklungspsychologie, Fachbereich Erziehungswissenschaft und Psychologie, Wissenschaftsbereich Psychologie
Email: herbert.scheithauer@fu-berlin.de

Wilfried Schubarth, Professor Dr.,
Universität Potsdam, Professur für Erziehungs- und Sozialisationstheorie
Email: wilschub@uni-potsdam.de

Sarah Ulrich, Dipl. Psych.,
Fachstelle Wirkungsorientierung beim buddY E.V.
Email: sarah.ulrich@buddy-ev.de

Sebastian Wachs, Dipl.-Päd.,
Wissenschaftlicher Mitarbeiter an der Universität Potsdam, Professur für Erziehungs- und Sozialisationstheorie
Email: wachs@uni-potsdam.de

Jutta Wedemann, Dr.
Wissenschaftliche Mitarbeiterin im Institut für Bildungswissenschaft der Leuphana Universität Lüneburg
Email: jutta.wedemann@leuphana.de

Layoutgestaltung: www.jmp-media.de,
Druck und Weiterverarbeitung: www.druckerei-paffenholz.de