

Entwicklungsförderung & Gewaltprävention (2):

Messung von Wirksamkeit und Umsetzungsqualität (Evaluation)

Friedrich Lösel, Nina Heinrichs & DFK-Sachverständige

Präventionsprogramme müssen vor ihrer Verbreitung evaluiert werden. Von zentraler Bedeutung ist dabei die **Wirkungsevaluation**. Durch sie wird mit wissenschaftlich anerkannten Methoden überprüft, inwieweit das jeweilige Programm seine Ziele erreicht, z. B. tatsächlich die kindliche Entwicklung positiv beeinflusst oder Jugendkriminalität vermindert. Nur wenn dies der Fall ist, kann ein Programm als „evidenzbasiert“ gelten. Es sollten nur evidenzbasierte Programme gefördert und ineffektive oder gar schädliche Maßnahmen vermieden werden.

Neben der Wirkungsevaluation ist auch eine **Prozessevaluation** erforderlich. Diese untersucht, wie das Programm faktisch durchgeführt wird. Z. B. wird geprüft, ob man tatsächlich die angestrebten Zielgruppen erreicht und die praktizierte Maßnahme dem zugrunde liegenden Konzept entspricht. Nur wenn dies der Fall ist, sind in der Wirkungsevaluation die angestrebten Effekte zu erwarten.

Im Folgenden werden einige Grundlagen der Prozess- und Wirkungsevaluation dargestellt. Ziel ist es, Praxis und Politik in knapper Form darüber zu informieren, worauf bei der Auswahl und Durchführung bestimmter Präventionsprogramme zu achten ist.

Grundsätzliches: Was ist eine Evaluation?

Durch eine Evaluation soll eine soziale Maßnahme auf möglichst objektive und transparente Weise mit wissenschaftlichen Methoden bewertet werden. Während man in der Medizin selbstverständlich wissen will, ob eine Behandlung tatsächlich wirksam ist und angemessen durchgeführt wird, begnügt man sich bei sozialen Programmen oft noch mit Behauptungen und plausiblen Vermutungen. Evaluationen sind hier auch manchmal mit negativen Assoziationen verbunden, weil man befürchtet, dass sich bei einer ungünstigen Bewertung negative Folgen für die betreffende Institution ergeben können. Tatsächlich sollen aber Evaluationen primär dazu beitragen, dass ein System etwas über sich selbst lernt und sich auf empirischer Basis weiterentwickelt. Bei der Evaluation von sozialen Interventionen fallen die Ergebnisse auch oft nicht pauschal positiv oder negativ aus, sondern haben differenzierte Ergebnismuster.

Aus diesen können wiederum Folgerungen für das weitere Vorgehen gezogen werden. Evaluation ist somit nicht ein einmaliger Vorgang, sondern ein kontinuierlicher Prozess.

Es sollten nur solche Präventionsmaßnahmen durchgeführt und finanziell gefördert werden, für die positive Wirkungen nachgewiesen sind und keine negativen Effekte festgestellt wurden. Manche Evaluationen zeigen, dass Letzteres selbst bei bester Absicht nicht immer auszuschließen ist. Bei neu entwickelten Programmen müssen zumindest empirisch fundierte Hypothesen zu der Frage bestehen, welche Effekte zu erwarten sind. Dies ist z. B. bei Programmen zur Förderung der sozialen Kompetenz von Kindern dann der Fall, wenn bewährte Theorien des sozialen Lernens und der Informationsverarbeitung zugrunde liegen. Nach einer Pilotphase sind die neuen Programme dann einer kontrollierten Wirkungsevaluation zu unterziehen.

In der Praxis lassen sich oft nicht alle Anforderungen an methodisch

fundierte Prozess- und Wirkungsevaluationen realisieren. Gleichwohl ist auch eine suboptimale Evaluation in jedem Fall besser als gar keine. Denn nur so kann im Laufe der Zeit systematisches Wissen über die Durchführung und Wirkung eines Präventionsprogramms aufgebaut werden.

Zu Recht wird seitens der Praxis vorgebracht, dass man selbst keine wissenschaftlich stichhaltige Wirkungsevaluation leisten kann. Bei kostspieligen Programmen sollten deshalb Programmträger oder Förderinstitutionen für eine angemessene Finanzierung von Evaluationsstudien sorgen (z. B. kann ein kleiner Prozentsatz der Programmkosten von vornherein für Evaluation reserviert werden). Darüber hinaus gibt es viele kostengünstige Möglichkeiten der Evaluation von Präventionsprogrammen durch Kooperation mit lokalen Hochschulen, an denen Studierende Daten für Qualifikationsarbeiten benötigen. Was die Prozessevaluation betrifft, sollte aber die Praxis selbst regelmäßig Daten sammeln, um die Qualität der Programm-Implementierung zu sichern.

Prozessevaluation: Wie werden Planung und Durchführung evaluiert?

Auch wenn ein Präventionsprogramm sehr sorgfältig geplant und entwickelt worden ist, kann es in der faktischen Durchführung Probleme geben. Deshalb ist durch eine Prozessevaluation zu erfassen, inwieweit das Programm zuverlässig umgesetzt wird (Integrität). Dies gilt insbesondere, wenn eine Maßnahme in größerem Umfang „in der Fläche“ eingesetzt wird. Empirische Untersuchungen haben gezeigt, dass entwicklungsorientierte Präventionsmaßnahmen nur dann die angestrebten Effekte haben, wenn die Durchführung programm-

gemäß erfolgt. Prozessevaluationen sollten routinemäßig erfolgen, damit bei eventuellen Qualitätsmängeln oder anderen Fehlentwicklungen rechtzeitig gegengesteuert werden kann. Um die Sammlung und Analyse von Prozessdaten in der Alltagspraxis sicherzustellen, ist ein Dokumentationssystem zu verwenden, das elektronisch oder in Papierform bearbeitet wird. Je einfacher diese Dokumentation durchzuführen ist, desto größer ist die Chance, dass die Mitarbeiter/-innen trotz ihrer Alltagsbelastung die erforderlichen Informationen vollständig eingeben.

Welche Daten in die Prozessevaluation einbezogen werden, hängt vom jeweiligen Präventionsprogramm und den vorhandenen Ressourcen ab. Wichtig sind z. B. einfache biografische Daten, an denen man abschätzen kann, ob die vorgesehenen Zielgruppen erreicht werden. Des Weiteren sollte erfasst werden, wie die Teilnehmer/-innen rekrutiert werden, ob sie regelmäßig anwesend sind, wie zufrieden sie mit den verschiedenen Teilen des Programms sind, inwieweit sie aktiv mitarbeiten, ob Probleme auftreten, welche Teilnehmer vorzeitig abbrechen (und warum) oder ob Programmänderungen vorgenommen werden. Während besondere Vorkommnisse in freier Form dokumentiert werden, empfiehlt sich bei den regelmäßig anfallenden Daten ein System, das auf einfache Weise Verläufe dokumentiert (z. B. zu- oder abnehmende Teilnehmerzufriedenheit, Entwicklung des Anteils von Familien mit Migrationshintergrund).

Falls möglich sind in die Prozessevaluation auch Aspekte der Schulung und Supervision des Personals einzubeziehen. Dazu gehören z. B. Rückmeldungen über die Auswahl und Qualifizierung, die Besprechung besonders gelungener oder schwieriger Programmelemente, Kooperationsstrukturen innerhalb und zwischen den Institutionen und gegebenenfalls auch besondere Vorkommnisse. Im Rahmen der Prozessevaluation kommen somit gleichermaßen sogenannte quantitative und qualitative Methoden zum Einsatz.

Wirkungsevaluation: Wie kann Wirksamkeit nachgewiesen werden?

Bei der Wirkungsevaluation ist der Untersuchungsplan bzw. das Design der Studie soweit wie möglich vor der

Durchführung des Programms festzulegen. Wirkungsevaluationen, die erst während oder nach der Durchführung geplant werden, sind in der Regel methodisch weniger stichhaltig. Entscheidungsträger sollten deshalb bei der Einführung neuer Programme von vornherein Expertise im Hinblick auf eine Evaluation einholen. Bei der Beurteilung bereits vorliegender Studien ist es im Zweifelsfall hilfreich, sich von einschlägig ausgewiesenen Wissenschaftlern beraten zu lassen.

Ein zentrales Kriterium für die Qualität einer Wirkungsevaluation ist es, ob ein *echt experimentelles Design* vorliegt (Randomized Controlled Trial; RCT). Hierbei erfolgt eine jeweils zufällige Zuweisung der Teilnehmer/-innen zu der Programmgruppe (PG) und einer Kontrollgruppe (KG), die das betreffende Programm nicht erhält. Dieses Design bietet die beste Gewähr dafür, dass PG und KG vergleichbar zusammengesetzt sind. Es ist bei knappen Ressourcen auch die fairste Regelung, weil alle potenziell Interessierten oder Geeigneten die gleiche Chance haben, am Programm teilzunehmen. Allerdings ist es bei ausreichenden Ressourcen oder bei Hochrisikogruppen manchmal nicht angezeigt, jemand von der Teilnahme auszuschließen. In solchen Fällen kann eine sogenannte Wartekontrollgruppe gebildet werden, die zunächst als Kontrollgruppe dient und das Programm dann zu einem späteren Zeitpunkt erhält. Soweit man mit einem Programm nicht völliges Neuland betritt, kann auch die PG mit einer KG verglichen werden, die eine andere oder eine weniger intensive Maßnahme erhält.

Wenn kein RCT möglich ist, gibt es eine Reihe von sogenannten *quasi-experimentellen Designs*, die unter suboptimalen Bedingungen zur Programmevaluation geeignet sind. Dazu gehören z. B. Designs, bei denen die PG und KG nach bestimmten Merkmalen ähnlich zusammengestellt sind (Parallelisierung) oder in denen äquivalente Jahrganggruppen verglichen werden (z. B. die Schüler/-innen eines Jahrgangs ohne Programm mit jenen nach Einführung der Maßnahme). Weil die Teilnehmer in PG und KG im statistischen Sinne nicht identisch sind, bleibt ein gewisses Maß an Unsicherheit, ob Unterschiede zwischen beiden Gruppen nach Durchführung der Präventionsmaßnahme tatsächlich auf der Maßnahme oder auf Unterschieden zwischen den Personen beruhen. Reine Vorher-Nachher-Messungen in-

nerhalb einer Gruppe ohne anderweitige Kontrolle liefern zumeist keinen stichhaltigen Wirkungsnachweis, da sich Kinder im Lauf der Entwicklung auch ohne Intervention verändern (z. B. sind sie normalerweise ab zwei bis drei Jahren seltener körperlich aggressiv).

Bei jeder Evaluationsstudie muss geprüft werden, ob die Ergebnisse alternativ erklärt werden können, also gar nicht auf den Einfluss des untersuchten Programms zurückzuführen sind. Z. B. kann eine Evaluation in der PG zu besseren Ergebnissen als in der KG geführt haben, weil die Teilnehmer der PG von vornherein motivierter oder kompetenter waren als jene der KG. Sind Verhaltensprobleme zeitweise besonders ausgeprägt, ergibt sich oft später ein natürlicher Rückgang zum Durchschnitt, der dann nicht durch das Präventionsprogramm bedingt ist. Gibt es keine plausible Alternativerklärung, besitzt die Untersuchung *interne Validität*.

Des Weiteren muss beurteilt werden, inwieweit die Ergebnisse einer Evaluation auf andere Personen (Gruppen), Zeitpunkte, Situationen, Wirkungskriterien und Kontexte verallgemeinert werden können (*externe Validität*). Z. B. kann ein vermeintlich positives Ergebnis eines Programms dadurch bedingt sein, dass die Wirkung nur mittels Selbstauskünften erfasst wurde und dabei die PG in sozial erwünschter Weise antwortete. Von solchen Ergebnissen kann man dann nicht verallgemeinern, dass das Programm auch wirksam ist, wenn die Teilnehmer/-innen nicht gefragt würden.

Eine aussagekräftige Evaluation erfordert auch eine gute Dokumentation der Programme, Stichproben, Methoden und Auswertungen (*deskriptive Validität*). Z. B. muss man beurteilen können, welche Alters- und Geschlechtsgruppen am Programm teilgenommen haben und ob die Datenauswertung sachgemäß ist. Bei kurzen zusammenfassenden Darstellungen zu Evaluationen von Präventionsprogrammen kann oftmals der Inhalt des Programms nicht näher dargestellt werden. Hier sind dann Angaben zu Manualen und zu eventuellen Prozessevaluationen der faktischen Durchführung hilfreich.

Effekte: Wodurch können Ergebnisse beeinflusst werden?

Die in Evaluationen festgestellten Effekte hängen keineswegs nur von

den durchgeführten Programmen ab. Z. B. zeigen sich in Demonstrations- oder Modellprojekten oft bessere Effekte als in der Alltagspraxis, weil beispielsweise das Personal am Anfang besonders motiviert ist und das Programm noch nicht der Alltagsroutine unterliegt. Dementsprechend wird im Englischen zwischen der „*Efficacy*“ unter günstigen experimentellen Umständen und der „*Effectiveness*“ in der Alltagsroutine unterschieden. Bei der Beurteilung von Präventionsprogrammen sollte deshalb auch berücksichtigt werden, inwieweit vorliegende Evaluationen unter realistischen Alltagsbedingungen durchgeführt wurden. Ein weiterer relevanter Faktor ist, ob Evaluationen nicht nur durch die Programmentwickler oder Durchführenden erfolgten, sondern auch durch unabhängige Experten. Eventuelle Interessenkonflikte sind offenzulegen.

Die genannten Faktoren sind Beispiele dafür, dass selbst bei sehr ähnlichen oder identischen Präventionsprogrammen unterschiedliche Ergebnisse auftreten können. Es gibt kaum eine evaluierte Maßnahme, bei der das nicht der Fall ist. Die Effekte hängen von zahlreichen Merkmalen des Programms und seiner Durchführung, des Kontexts der Implementierung, der teilnehmenden Gruppen

und nicht zuletzt der Methodik der Evaluation ab.

Zur Beurteilung der Wirksamkeit eines Programms oder Programmtyps sind deshalb systematische Bestandsaufnahmen aller vorliegenden Evaluationsstudien in Form von Meta-Analysen sehr hilfreich. Sie informieren darüber, wie die Ergebnisse insgesamt ausgefallen sind, wie groß die Effektstärken sind und welche Faktoren außer dem Inhalt des Programms einen Einfluss auf die Wirksamkeit haben. Bei der Entscheidung für ein bestimmtes Präventionsprogramm sollte mehr das Gesamtbild der Ergebnisse als eine einzelne Evaluationsstudie berücksichtigt werden.

Qualitätssicherung und Zertifizierung: Wie können gute Ergebnisse auf Dauer erzielt werden?

Auch bei positiv evaluierten Programmen muss darauf geachtet werden, dass sie durch Maßnahmen der Qualitätssicherung bei wiederholter Durchführung in der Alltagspraxis wirksam bleiben. Deshalb sind laufende Prozessevaluationen und wiederholte Wirkungsevaluationen wichtig. Programme, die sich empirisch bewährt haben, sollten zur besseren Orientierung auf unbürokratische

Weise akkreditiert bzw. zertifiziert werden. Dies muss anhand transparenter Kriterien durch unabhängige Experten erfolgen. Auch dabei handelt es sich um einen Bestandteil der kontinuierlichen Evaluation. Programme, die sich zu einem bestimmten Zeitpunkt nachweislich bewährt haben, können z. B. durch veränderte Umstände und Zielgruppen, Probleme der Qualitätssicherung oder neue wissenschaftliche Erkenntnisse nach etlichen Jahren „überholt“ sein. Deshalb ist es sinnvoll, Akkreditierungen nach einer bestimmten Zeit (z. B. fünf Jahren) zu überprüfen.

Die beim Landespräventionsrat Niedersachsen entwickelte öffentlich zugängliche „**Grüne Liste Prävention**“ (www.gruene-liste-praevention.de) liefert auf der Basis nachvollziehbarer Kriterien Informationen, welche evaluierten Präventionsprogramme in Deutschland verfügbar sind. In Stufe 3 „Effektivität nachgewiesen“ werden jene Programme eingeordnet, deren Evaluationsstudien den höchsten Standards einer Wirksamkeitsmessung entsprechen und somit eine hinreichende bis sehr starke Beweiskraft haben. Die vorliegenden Studien haben eine Zuweisung der Teilnehmer in die Kontroll- oder Interventionsbedingung nach dem Zufallsprinzip vorgenommen, oder es handelt es sich um gut kontrollierte „Quasi-Experimente“. Eine ausreichend große Anzahl an Interventions- und Vergleichsgruppen und Teilnehmern wurde gewählt. Eine „follow-up“-Messung mit positiven Ergebnissen, in der Regel sechs Monate oder später, wurde nach dem Ende der Maßnahme durchgeführt.

Merkmale im Zusammenhang mit der Qualität der Umsetzung werden in der „Grünen Liste Prävention“ noch nicht zur Bewertung herangezogen, können aber künftig einen wichtigen Unterschied in der Verbreitung von Programmen machen. Hier bringen einige Programme bessere Voraussetzungen mit als andere, d. h. sie verfügen über eine qualitätssichernde Implementierungsstrategie (z. B. zertifizierte Trainer bzw. ausgebildete Multiplikatoren). Eine weitere Frage ist, ob für eine wünschenswerte breitflächige Umsetzung genügend Kapazität in dieser Hinsicht vorhanden ist.

Der Artikel ist ein Auszug aus dem Leitfaden „Entwicklungsförderung & Gewaltprävention für junge Menschen“, Kapitel 5, im April 2013 beim DFK in Bonn erschienen, Bestellung: dfk@bmi.bund.de