DEUTSCHER PRÄVENTIONSTAG

Markt, Macht und Wissenschaft; Kritische Überlegungen zur deutschen Präventionsforschung

von

Manuel Eisner
Denis Ribeaud

Aus: Erich Marks & Wiebke Steffen (Hrsg.):
Starke Jugend – Starke Zukunft
Ausgewählte Beiträge des 12. Deutschen Präventionstages
Forum Verlag Godesberg, Mönchengladbach 2008, Seite 173-192

ISBN 3936999457 (Printausgabe)
ISBN 978-3936999457 (E-Book)

Manuel Eisner / Denis Ribeaud

Markt, Macht und Wissenschaft; Kritische Überlegungen zur deutschen Präventionsforschung

Wer sich auf manchen deutschsprachigen Internetseiten zur Gewaltprävention umsieht, kann den Eindruck gewinnen, heutige Präventionsprogramme seien geradezu Wunderwerke wirksamer Sozialtechnologie. Lehrer schreiben begeistert, wie phantastisch das Klassenklima nach Umsetzung eines Sozialkompetenzprogramms ist; Eltern werden zitiert, wie sich ihr verhaltensauffälliger Racker dank Erziehungskursen in ein Musterkind verwandelt hat; und man schmückt sich mit Presseberichten, in denen erklärt wird, ab sofort würden dank des neuen Programms die Kinder abgeklärt miteinander reden statt sich zu verprügeln. Selbstverständlich sagt solche als „Evaluation" kaschierte Eigenwerbung über die tatsächliche Wirksamkeit nicht mehr aus als begeisterte Lesermeinungen zu den Effekten von Wünschelruten, Kupferbändern oder Kristallkugeln, nämlich gar nichts.

Das gleiche gilt für reine Prozessevaluationen - also Einschätzungen darüber, wie das Projekt umgesetzt wurde und wie zufrieden die Benutzer mit dem Programm sind. Zwar wird gerade gegenüber der Praxis nicht selten der Eindruck geweckt, dass glückliche Projektteilnehmer ein Gradmesser für ein gutes Programm seien, oder dass man Wirksamkeit bestimmen könne indem man frage, ob die Umsetzenden das Programm für wirksam halten. Tatsächlich sind weder zufriedene Teilnehmer noch von einer Wirkung subjektiv überzeugte Umsetzende ein wissenschaftlich annehmbares Kriterium für die Frage, ob tatsächlich die gewünschten Änderungen erreicht wurden.

Vielmehr besteht unter Präventionsforschenden weit herum Einigkeit darüber, dass für einen wissenschaftlich abgestützten Nachweis von positiven Wirkungen höhere Ansprüche an die Forschungsanlage gestellt werden müssen. Worin diese Anforderungen bestehen, ist dank der Pionierarbeiten von Cochrane (1972), Campbell und Stanley (1966), Cook und Campbell (1979), sowie Shadish et al. (2002) weitgehend geklärt. Es handelt sich im Wesentlichen um methodisch sorgfältig angelegte Experimentalstudien, bei denen idealerweise die Teilnehmenden zufällig aus der Zielpopulation ausgewählt werden und die Zuweisung zu den Behandlungsbedingungen nach dem Zufallsprinzip erfolgt. Durch den Vergleich von Veränderungen zwischen Kontroll- und Interventionsgruppe erlauben sie, sofern keine Annahmen von Experimentaldesigns verletzt werden, Aussagen über die erzielten Wirkungen mit einer hohen Gültigkeit.

Um diese Erkenntnis herum ist in den vergangenen 20 Jahren im angelsächsischen Sprachraum eine wissenschaftliche Bewegung entstanden, welche sich dem Ziel einer *evidenzbasierten Prävention* verschrieben hat. Sie kann als das Bestreben definiert werden, Fehlschlüsse über die Wirkungen von Präventionsmaßnahmen zu vermeiden und Maßnahmen zur Verminderung oder Verhinderung von unerwünschten Verhaltensweisen möglichst weitgehend auf gesichertes empirisches Wissen abzustützen (für den Bereich der Kriminalprävention vgl.

Rössner, Bannenberg, & Landeshauptstadt Düsseldorf, 2002; Sherman, Farrington, Welsh, & MacKenzie, 2002). Evidenzbasierte Prävention beruht auf dem Grundsatz, dass die Wirksamkeit von Prävention durch gute empirische Forschung überprüft werden kann und dass durch den Zusammenzug der Forschungsergebnisse zuverlässige Kenntnisse darüber gewonnen werden können, welche Präventionsmaßnahmen wirksam, wirkungslos oder schädlich sind; wie Maßnahmen, welche sich in der Forschung als wirksam erwiesen haben, effektiv in die Praxis umgesetzt werden können; wie sich Maßnahmen an die Bedürfnisse unterschiedlicher Bevölkerungsgruppen anpassen lassen: und welche Aspekte der praktischen Umsetzung dafür verantwortlich sind, dass positive Wirkungen erzielt werden können.

Infolge des gestiegenen Interesses an besserer, evidenzgestützter Gewalt- und Kriminalprävention ist es in den letzten 10 Jahren zu einer deutlichen Zunahme von wissenschaftlichen Wirkungsevaluationen mit einem randomisierten Kontrollgruppendesign gekommen. Im Bereich der entwicklungsorientierten Gewaltprävention gehören hierzu im deutschsprachigen Raum beispielsweise die Studien von Heinrichs et al. (2006) über *Triple P*, von Schick und Cierpka (2005; 2006) zu *Faustlos*, von Asshauer und Hanewinkel (2000) über das Lebenskompetenztraining *Fit und Stark fürs Leben*, von Lösel et al. (2006) zu einem kombinierten Elterntraining und Sozialkompetenzprogramm *EFFEKT* sowie von Eisner et al. (2007) über eine weitere Kombination von Elterntraining und schulischen Sozialkompetenztraining. Mehrere größere Studien sind gegenwärtig im Gang oder haben noch keine Evaluationsergebnisse publiziert.

Es gibt zudem gegenwärtig in Politik und Praxis eine erfreulich große Bereitschaft, den Argumenten für eine evidenzbasierte Präventionspolitik Folge zu leisten und die damit einhergehenden zeitlichen und finanziellen Kosten auf sich zu nehmen. Allerdings kann diese Bereitschaft nur aufrecht gehalten bleiben, wenn die Wissenschaft ihre Funktionen so gut als möglich wahrnimmt. Hierzu gehört, dass sie Forschung so durchführt, dass die Ergebnisse unverzerrt die tatsächlichen Wirkungen eines Programms oder einer Maßnahme wiedergeben und dass sie die Ergebnis möglichst so kommuniziert, dass Außenstehende Dritte ein angemessenes Bild der Sachlage gewinnen.

Nun werden viele dieser Studien zur Wirkung von Präventionsmaßnahmen von Forschergruppen durchgeführt, welche das Produkt selbst entwickelt oder in Lizenz übernommen haben und es kommerziell vertreiben. Man kann daher auch von *Eigenevaluationen* sprechen. Dem stehen Fremdevaluation gegenüber, bei welchen die Durchführung der Studie, die Analyse der Daten sowie die Publikation der Ergebnisse und Folgerungen von Forschenden durchgeführt werden, die kein direktes Eigeninteresse am evaluierten Programm haben.

Es ist grundsätzlich zu begrüßen, wenn Programmentwickler und -vertreiber ihre Maßnahmen einer wissenschaftlichen Evaluation unterziehen. Dass hierbei Eigeninteressen bestehen, ist an sich nicht problematisch, sofern ausreichend wirksame Kontrollmechanismen bestehen, um die Qualität der Forschung zu garantieren. Es besteht aber kein Zweifel, dass diese Eigeninteressen nicht identisch sind mit dem Interesse von Öffentlichkeit und Fachpublikum, möglichst

unverzerrte und wirklichkeitsnahe Schätzwerte der tatsächlichen Effekte eines Präventionsprogramms zu erhalten. Und es wäre naiv zu glauben, dass dieser Interessenkonflikt in jedem Fall in eine unverzerrte Darstellung der erzielten Wirkungen mündet.

Dies dokumentiert eine wichtige Studie von Petrosino und Soydan (2005). Die Autoren haben im Rahmen einer Metaanalyse untersucht, wie stark die publizierten Wirkungen einer Evaluation davon abhängen, ob die Studie durch Programmentwickler und Programmvertreiber oder durch unabhängige Wissenschaftler durchgeführt wurde. Sie haben zu diesem Zweck 300 randomisierte Studien untersucht, welche die Wirkungen von Interventionsprogrammen auf die Rückfallwahrscheinlichkeit von Straftätern prüften. Alle Studien wurden anschließend bezüglich der Verbindung zwischen Programmentwicklern und Projektleitung bewertet. Die Auswertung zeigte für 137 Eigenevaluationen einen schwachen, aber beachtlichen positiver Effekt von Cohen's d = 0.16. Dem steht gegenüber, dass in den 124 Fremdevaluationen die durchschnittliche Effektstärke ziemlich genau bei Null lag (Cohens d = 0.02). Mit anderen Worten: Während man im Durchschnitt von Eigenevaluationen zum Schluss gelangt, dass heutige Interventionsprogramme zur Reduktion der Rückfallwahrscheinlichkeit eine gewisse Wirkung haben, führt eine Betrachtung der Fremdevaluationen zum Schluss, dass die Programme völlig wirkungslos sind.

Die Gründe für die Diskrepanz zwischen Eigenevaluationen und Fremdevaluationen werden unterschiedlich beurteilt. Petrosino und Soydan (2005) unterscheiden zwei Interpretation, die sie als Umsetzungsperspektive (*high fidelity view*) und als zynische Perspektive (*cynical view*) bezeichnen. Die positive Interpretation lautet, dass eine aktive Beteiligung der Programmvertreiber dazu führt, dass die Programme in besonders guter Qualität, mit Enthusiasmus und in großer Umsetzungstreue realisiert werden – und sich dies dann in einer positiven Wirkung niederschlägt. Das einzige Problem wäre dann, wie man all diese Komponenten eines idealen Modellversuchs auch in Bedingungen realisieren kann, unter denen die Programmentwickler nicht persönlich anwesend sind – also den Regelfall von Präventionsprojekten.

Die zynische Perspektive hingegen argumentiert, dass eine Reihe von subtilen und weniger subtilen Faktoren die Programmentwickler und –vertreiber unter Druck setzen, positive Ergebnisse zu produzieren. Dieses Eigeninteresse erstreckt sich auf die *wissenschaftliche Reputation*, da man mit einem Programm meist die eigenen theoretischen Annahmen prüft; auf die *zeitlichen Investitionen*, da Entwicklung und Test eines Programms bis zur Marktreife oft mehrere Jahre in Anspruch nehmen; auf *Forschungsgelder*, da positive Ergebnisse mit einer größeren Wahrscheinlichkeit weiterer Unterstützung einher gehen; auf *politischen Einfluss*, da Vertreiber von erfolgreichen Programmen in Verwaltung und Politik mehr Gehör finden; sowie direkte *finanzielle Eigeninteressen*, da für Entwicklung und Vertrieb eines Programms oft erhebliche Investitionen notwendig sind, die nur bei positiven Testresultaten wieder eingeholt werden können.

Ein verzerrtes Bild der Wirkungen von Präventionsprogrammen ist aus mehreren Gründen problematisch: Erstens wird eine optimistische Einschätzung von Interventionswirkungen in

der Regel dazu führen, dass politische Entscheidungsträger zusätzliche öffentliche Mittel in die entsprechende Maßnahme investieren. Wenn allerdings diese optimistische Einschätzung auf verzerrten Ergebnissen basiert, dann bedeutet das eine Fehlallokation von öffentlichen Ressourcen. Dies führt zu einem Entzug von Mitteln für andere Programme, die möglicherweise in Wirklichkeit ebenso wirksam oder gar wirksamer sind.

Zweitens leiten systematisch verzerrte Ergebnisse von Einzelstudien die wissenschaftliche Forschung auf Abwege und erschweren den Fortschritt in der Entwicklung besserer Präventionsmaßnahmen. Ein Beispiel sind Metaanalysen, in denen die Ergebnisse vieler Einzelstudien zusammengefasst und in statistischen Kennwerten ausgedrückt werden. Metaanalyse gilt als der beste Weg, ein von subjektiven Eindrücken unverzerrtes Bild des Wissensstandes zu einem Interventions- oder Präventionsbereich zu erhalten. Dies gilt aber nur, wenn die Einzelstudien selbst als unverzerrte Schätzungen der tatsächlichen Wirkungen einer Maßnahme gelten können. Wenn in eine Metaanalyse mehrere, systematisch in die gleiche (meistens zu optimistische) Richtung verzerrte Einzelstudien eingehen, dann führen sie zu einer irrtümlichen Bilanz des aktuellen Forschungsstandes.

Drittens schließlich bewirken systematisch verzerrte Ergebnisse von Evaluationsforschungen einen Vertrauensverlust der Öffentlichkeit in die wissenschaftliche Wirkungsevaluation. Beispielsweise wird es auf die Dauer außenstehenden Beobachtern nicht verborgen bleiben, wenn die Effekte von Evaluationsstudien immer das vom Studienleiter vertretene Präventionsprogramm stützen oder wenn sich nach einer Weile herausstellt, dass die Umsetzung eines Programms im Alltag nicht die wissenschaftlich versprochenen Wirkungen erbringt. Hierdurch kann es geschehen, dass sich öffentliche Hand und Wissenschaftsförderung enttäuscht von den meist kostspieligen experimentellen Evaluationsprojekten abwenden, da sie das Versprechen einer unverzerrten Beurteilung von Interventionswirkungen nicht einzulösen vermögen.

Empirische Illustrationen

Der folgende Abschnitt diskutiert an ausgewählten Beispielen drei Probleme, die mit Eigenevaluationen einher gehen können: Die Schwierigkeit, Befunde aus Eigenevaluationen zur Wirksamkeit von Präventionsprogrammen in Fremdevaluationen zu replizieren; das Problem, dass im Forschungsprozess mehrere methodische Entscheide so gefällt werden, dass sie das gewünschte Ergebnis begünstigen und die Befunde daher nicht valide sind; und die Frage, ob Forschungsbefunde gegenüber Praxis und Öffentlichkeit fair dargestellt werden.

Neben Beispielen aus der US-amerikanischen Forschung werden hierbei auch Beispiele aus der deutschsprachigen Forschung aufgeführt. Die Auswahl reflektiert die Interessen und Kompetenzen des Autors. Sie soll nicht als Stellungsnahme für oder gegen die diskutierten Programme interpretiert werden, sondern dient einzig der Illustration von Konflikten, die in der Struktur von Eigenevaluationen angelegt sind.

Ergebnisse aus Eigenevaluationen können in Fremdevaluationen nicht repliziert werden

Experimentelle Wirkungsevaluationen werden in der Regel durchgeführt, um aufgrund der gemessenen Effekte Aussagen über die zu erwartenden Wirkungen auch außerhalb der Untersuchung machen zu können. Man spricht hier von *externer Validität* (vgl. z.B. Shadish et al., 2002). Damit aber diese Folgerung zulässig ist, müssen die gemessenen Effekte *unverzerrte Schätzwerte* der tatsächlich erzielbaren Wirkungen sein – sie sollen also die effektive Wirkung weder systematisch über- noch unterschätzen. Nur dann gilt, dass die Studienergebnisse auch auf die Welt außerhalb der besonderen Bedingungen einer experimentellen Studie generalisiert werden können. Allerdings scheint dies häufig nicht zuzutreffen, wie zwei Beispiele illustrieren.

Beim ersten Beispiel handelt es sich um das Suchtpräventionsprogramm *ALERT*. ALERT ist ein schulbasiertes Programm zur Förderung von Lebenskompetenzen, das in den USA weit verbreitet ist und als wissenschaftlich gut abgestützt gilt (vgl. die Webseite www.projectalert.best.org). Es besteht aus insgesamt 14 Lektionen im siebten und achten Schuljahr. Gemäß der Webseite der Betreiber wird das Programm auf sieben Empfehlungslisten von US-amerikanischen Behörden - darunter dem Bildungsministerium, dem Gesundheitsministerium und dem Justizministerium - als evidenzbasiertes Modellprogramm empfohlen. Die Vertreiber von ALERT werben damit, dass das Programm beispielsweise den Marijuana-Gebrauch um 60 Prozent reduziert, den Nikotinmissbrauch um 35-55 Prozent vermindert und den Alkoholmissbrauch signifikant senkt (Ellickson, 1998; Ellickson & Bell, 1990; Ellickson, Bell, & Harrison, 1993; Ellickson, Bell, & McGuigan, 1993).

Aufgrund diese positiven Erfolgsmeldungen unterwarfen St. Pierre et al (2006) das Programm erstmals einer unabhängigen Evaluation. Mit über 1600 Schülern und einem randomisierten Kontrollgruppendesign handelte es sich um eine sorgfältig angelegte Längsschnittstudie, die höchsten Qualitätsanforderungen an den Wirkungsnachweis von Interventionen genügt.

Die Ergebnisse zeigten für *keine* der Variablen zur Messung von Veränderungen des Substanzkonsums einen positiven Effekt. Hingegen finden sie einen möglichen negativen (d.h. unerwünschten) Effekt des Programms auf den Substanzkonsum, wobei sie aber darauf hinweisen, dass dieser angesichts der vielen gemessenen Zielgrößen möglicherweise zufällig zustande gekommen ist. Ebenso wenig konnte eine systematische Wirkung auf die Mediatoren (d.h. direkt durch das Programm angesprochenen Zielgrößen wie beispielsweise Einstellungen zu Drogen) gezeigt werden; die wenigen Effekte verteilten sich gleichermaßen auf positive und negative Wirkungen.

Ein analoges Problem kann im deutschsprachigen Raum für *Triple P* beobachtet werden. Triple P ist ein vom Matthew Sanders an der Queensland Universität in Australien entwickeltes Elterntraining. Es fußt auf verhaltenstheoretischen Grundlagen und will dem Problemverhalten von Kindern und Jugendlichen vorbeugen (Sanders, 1999; Sanders, Lynch, & Markie-Dadds, 1994). Die internationale Webseite von Triple P (www.triplep.net) wirbt mit dem un-

bescheidenen Slogan „parenting now comes with an instruction manual“. Das Programm wurde seit seiner Entwicklung in über 30 Studien, von denen allerdings viele auf einer sehr kleinen Zahl von Teilnehmern basieren, auf seine Wirksamkeit hin überprüft. Dabei kommen ausnahmslos alle publizierten Studien der Programmentwickler zum Befund, dass Triple P gute Wirkungen auf das Erziehungsverhalten und das Problemverhalten der Kinder zeige. Es findet sich daher auch auf Empfehlungslisten etwa der WHO und des Europarates.

Seit 2001 wird Triple P auch in Deutschland und der Schweiz vertrieben. Im Jahr 2006, mehrere Jahre nachdem der Vertrieb bereits angelaufen war, publizierten die Vertreiber der deutschen Version Ergebnisse ihrer Evaluationsstudie (Heinrichs et al., 2006). In dieser Studie wurde Triple P als universelle Maßnahme den Eltern von Kindergärtnern in Braunschweig angeboten. Die Studie berichtet von durchwegs positiven Effekten. Insbesondere kommt sie zum Schluss, dass durch Triple P das Problemverhalten der Kinder signifikant zurückgegangen sei. Die Forschergruppe empfahl daher in den Folgerungen „eine breitflächige Dissemination dieses Trainings“ (Heinrichs et al., 2006: 94).

Triple P wurde im Rahmen des Zürcher Projektes zum sozialen Verhalten von Kindern (z-proso) erstmals einer unabhängigen Evaluation unterzogen (Eisner et al., 2007). Die Studie ist mit rund 1300 teilnehmenden Kindern und Eltern weltweit eines der größten Projekte zur Evaluation der Wirkung von Elterntrainings. Das Studiendesign wurde in Zusammenarbeit mit renommierten Experten der Evaluationsforschung als Längsschnittstudie angelegt, in die ein randomisiertes Kontrollgruppendesign eingebettet ist. Das Verhalten der Kinder wurde in einem Mehrinformandenansatz aus der Perspektive der Eltern, der Lehrpersonen sowie der Kinder selbst untersucht. Der Kurs selbst wurde von erfahrenen und gemeinsam mit Triple P Schweiz ausgewählten Trainerinnen vermittelt. Rund 31 Prozent der Familien, welche in der Interventionsgruppe an der Längsschnittstudie teilnehmen, absolvierten den Kurs. Die Zufriedenheit unter den Kursteilnehmern war hoch.

Die Teilnahmerate und die Zufriedenheit der Eltern mit dem Kurs entsprechen weitgehend jenen der Braunschweiger Studie. Dies gilt allerdings nicht für die Ergebnisse zu den Wirkungen: Eine so genannte Intention-to-treat Analyse ergab weder für die sieben Teilaspekte von elterlichem Erziehungsverhalten noch für die neun Teilindikatoren von kindlichem Problemverhalten irgendwelche signifikanten positiven Effekte. Beschränkt man die Analyse nur auf jene Eltern, welche tatsächlich den größten Teil des Programms absolviert hatten, dann ergaben sich einige schwach positive Effekte auf das Erziehungsverhalten. Dem steht gegenüber, dass ein statistisch signifikanter negativer (d.h. unerwünschter) Effekt für nichtaggressives externalisierendes Problemverhalten des Kindes auf der Perspektive der Lehrperson gefunden wurde.

Für beide Beispiele ergibt sich: Studien, welche als Eigenevaluationen durchgeführt werden, berichten positive Wirkungen. Unabhängige Studien hingegen finden nichts. Solche Ergebnisse sind in der Forschung zu evidenzbasierter Prävention keine Einzelfälle. Ähnliche Ergebnisse werden auch in anderen Präventionsbereichen berichtet. In der Metaanalyse von 84

Studien zur Wirkung von *Sozialkompetenzprogrammen* durch Lösel und Beelmann (2003) beispielsweise lag die mittlere Effektstärke bei Eigenevaluationen mit Cohens d=0.49 fast doppelt so hoch wie bei jenen Programmen, die von unabhängigen Forschenden realisiert wurden. Eine Metaanalyse von Borman et al. (2003: 37) für *Schulreformprogramme* fand, dass "studies performed by the developer yielded considerably stronger effects than studies performed by others". Und eine kürzlich publizierte Studie, welche die fünf meistempfohlenen *Drogenpräventionsprogramme* der USA unter die Lupe genommen hat (Gandhi, Murphy-Graham, Petrosino, Chrismer, & Weiss, 2007) kommt zum Schluss, dass die publizierten Ergebnisse zur Wirksamkeit der Programme selektiv nur die besten Befunde rauspicken, dass unabhängige Evaluationen weitgehend fehlen und dass dort, wo Replikationen realisiert wurden, die Wirkungen oft nicht bestätigt werden konnten.

Positive Ergebnisse können das Ergebnis einer Kette von methodisch problematischen Entscheiden sein

Es kann mehrere Gründe dafür geben, weshalb Ergebnisse von Selbstevaluationen in Fremdevaluationen nicht repliziert werden können. Dabei kann sicher eine Rolle spielen, dass die Umsetzungsqualität in Fremdevaluationen nicht in gleichem Masse garantiert ist wie bei Eigenevaluationen. Allerdings gibt es aus verschiedenen Studien auch Hinweise darauf, dass die positiven Ergebnisse von Selbstevaluationen damit zusammenhängen können, dass in jeder Evaluationsstudie eine Vielzahl von aufeinander folgenden methodischen Entscheiden getroffen werden muss. Wenn jeder dieser Entscheide nur geringfügig durch das gewünschte Ergebnis eingefärbt ist, kann hieraus im Endergebnis ein problematisches Gesamtbild entstehen.

Eine systematische Analyse von Littell (2005) hat kürzlich das Problem in grellem Licht aufscheinen lassen. Sie untersuchte im Auftrag der *Campbell Collaboration* systematisch den Wissensstand bezüglich der Wirkungen von multisystemischer Therapie (MST) bei der Behandlung von verhaltensauffälligen und delinquenten Jugendlichen (Littell, 2005). Multisystemische Therapie (Henggeler et al., 1996; Henggeler, Melton, & Smith, 1992) gilt als eines der erfolgreichsten Programme zur Reduktion von Verhaltensproblemen bei Hochrisikopopulationen. Es steht als Modellprogramm auf vielen Empfehlungslisten US-amerikanischer Behörden und wird in den USA und in Europa bei jährlich rund 10,000 jugendlichen Straftätern eingesetzt. Anfangs der 2000er Jahre erhielt die Gruppe um die Programmentwickler jährlich über 20 Millionen US-Dollar an Forschungsaufträgen.

Dieser Erfolg gründet fast ausschließlich auf wissenschaftlichen Publikationen der Programmentwickler, welche durchwegs über positive Effekte von MST berichten (für eine Übersicht aller Publikationen zu Experimentalstudien mit MST vgl. Littell, Popa, & Forsythe, 2005). Die systematische Review von Littell (2005) nahm diese Effekte genauer unter die Lupe, indem sie neben den publizierten Studien der Programmentwickler auch unpublizierte Studien und unabhängige Studien berücksichtigte. Zudem zog sie alle empirisch getesteten (und nicht nur die positiven) Effektgrößen in Betracht und unterzog die Verfahren bei der Zuordnung zu den Behandlungsgruppen einer genauen Prüfung.

Sie kam zum Schluss, dass zwar keine Hinweise auf eine schädliche Wirkung von MST bestünden, dass aber die bisherigen Ergebnisse keine Unterstützung für die Hypothese liefern, MST sei wirksamer als herkömmliche Therapieverfahren. Die sehr viel positivere Selbstdarstellung von MST erklärt sie als Folge einer Kombination von systematischen Verzerrungen. Hierzu gehören: Studien mit unerwünschten Ergebnissen wurden nicht publiziert und verzerren daher den Eindruck bei den publizierten Arbeiten; in gewissen Studien mit positiven Ergebnissen ist das Vorgehen bei der Zuteilung zu Interventions- und Experimentalgruppe unklar; zwischen den anfänglichen unveröffentlichten wissenschaftlichen Schlussberichten und den publizierten Ergebnissen bestehen ungeklärte Diskrepanzen, welche sich zu Gunsten der Behandlung auswirken; was als Behandlungsabschluss gilt, wurde subjektiv definiert und ist damit möglichen unbewussten Manipulationen ausgesetzt.[1]

Vor diesem Hintergrund lohnt es sich, die oben skizzierte Diskrepanz zwischen den sehr positiven Ergebnissen der Eigenevaluation von Triple P in Braunschweig (Heinrichs et al., 2006) und den deutlich weniger erfreulichen Befunden der Fremdevaluation von Eisner et al. (2008) nochmals unter die Lupe zu nehmen. Eine solche Betrachtung ergibt, dass analog zu den Befunden von Littell (2005) in der Studie von Heinrichs et al. (2006) mehrere methodisch problematische Entscheide gefällt wurden, welche allesamt zum Effekt haben, dass die Interventionseffekte überschätzt werden (vgl. auch Eisner & Ribeaud, 2008). Sie seien hier kurz zusammengefasst:

Erstens haben in Braunschweig knapp 70 Prozent der angesprochenen Eltern die Teilnahme an der Studie völlig verweigert, so dass über diese Eltern keinerlei Daten vorliegen. Es handelt sich hierbei überwiegend um bildungsferne und sozial unterprivilegierte Eltern. Eine solch tiefe Teilnahmerate bewirkt, dass eine inferenzstatistische Absicherung (d.h. der Schluss von den erhobenen Daten auf die Grundbevölkerung) der Befunde kaum möglich ist. Vielmehr ähneln Rückschlüsse auf die Gesamtpopulation, aus der die Stichprobe gezogen wurde, eher einer Lotterie als einer wissenschaftlichen Analyse. Dies gilt besonders, wenn beansprucht wird, eine so genannte Intention-to-Treat Analyse durchzuführen. Solche Analysen sind immer Analysen *aller* Personen, welche einer Behandlungsbedingung zugewiesen wurden. Wenn für sieben von zehn potentiellen Teilnehmern aber keine Daten vorliegen, dann ist auch eine entsprechende Analyse unmöglich.

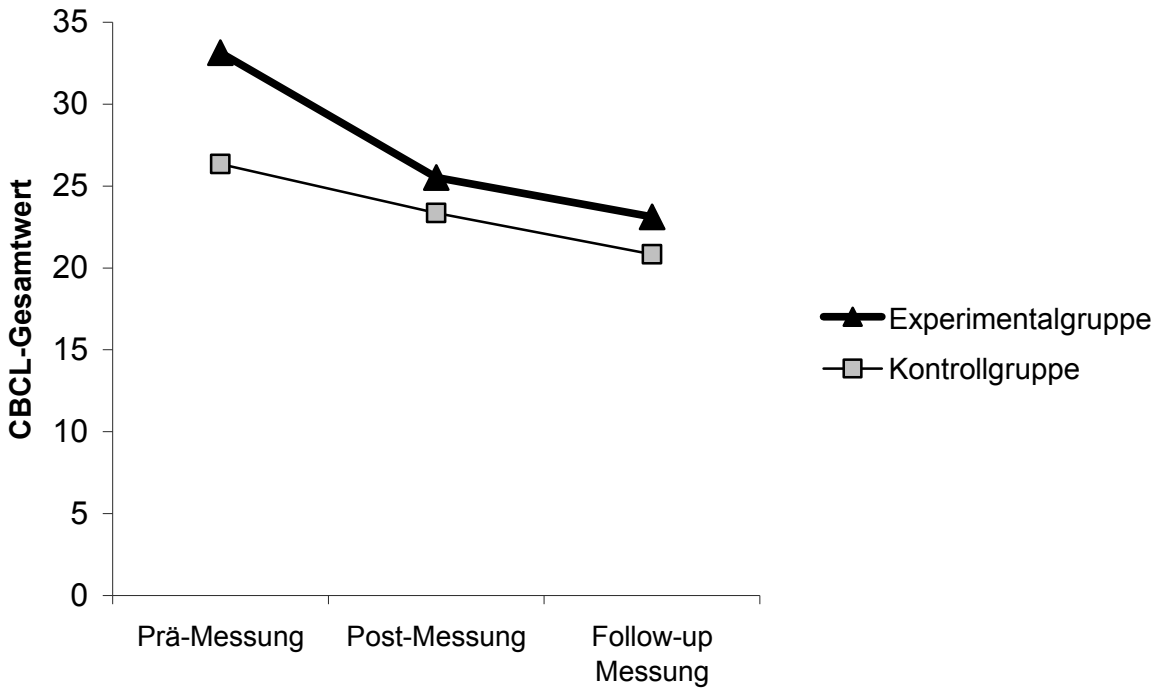
Zweitens wurden Projektteilnehmer für die Datenanalyse nachträglich aus der Behandlungsbedingung in die Kontrollbedingung umgeteilt. So heißt es in Heinrichs et al (2006: 86), dass „Familien, die zur Kontrollbedingung randomisiert wurden (N = 62) und Triple P Ablehner (N = 28) [….] für die folgenden Auswertungen zusammengefasst wurden“. Mit anderen Worten: Jene Familien, die für eine Teilnahme an Triple P vorgesehen waren aber das Programm nicht nutzen wollten, wurden im Nachhinein der Kontrollgruppe zugewiesen. Diese Umtei-

[1] Seit der Publikation der Kritik von Littell sind mehrere neue Arbeiten zu multisystemischer Therapie erschienen, die allerdings ein eher positives Bild der Wirkungen zeichnen. Unabhängige Evaluationen von Timmons-Mitchell et al. (2006) in den USA sowie von Ogden und Hagen (2006) in Norwegen fanden positive Effekte; die ebenfalls unabhängige Evaluation von Olsson und Sundell (2008) in Schweden hingegen fand keine Effekte.

lung ist eine grobe Verletzung der methodischen Prinzipien von randomisierten Forschungsdesigns. Es ist unverständlich, wie die Autoren ihre Studie dennoch als randomisiertes Kontrollgruppendesign darstellen können.

Drittens unterscheiden sich die Ausgangswerte (d.h. vor der Intervention) zwischen Kontrollgruppe und Behandlungsgruppe bei mehreren zentralen Zielvariablen massiv. Dies verletzt eine zentrale Annahme von randomisierten Experimentalstudien, nämlich dass die verglichenen Gruppen hinsichtlich der Kriteriumswerte vor der Intervention nicht unterscheiden. Die Auswirkungen der Verletzung dieser methodischen Anforderung sind in Abbildung 1 illustriert, welche den Verlauf des Gesamtwertes von externalisierendem Problemverhalten in der Kontrollgruppe und der Interventionsgruppe zwischen der Prä-Messung, der Post-Messung und der Follow-up Messung zeigen. Die Daten sind der Publikation von Heinrichs et al. (2006: 89) entnommen.

Abbildung 1 Ein Beleg für die Wirksamkeit von Triple P? Mittelwerte von kindlichem Problemverhalten in Experimental- und Kontrollgruppe gemäß Heinrichs et al. (2006: 89)



Hinweis: Mittelwerte der Gesamtscores für das Problemverhalten des Kindes (Child Behavior Checklist) nach Einschätzung der Mutter. Die Originaldaten wurden aus Heinrichs et al. (2006: 89) übernommen.

Die Autoren interpretieren diesen Datenverlauf als Beleg für die Wirksamkeit von Triple P, da der Rückgang in der Triple P Gruppe größer ist als in der Kontrollgruppe. Dem steht ge-

genüber, dass ein Datenverlauf, wie er in Abbildung 1 gezeigt ist, geradezu als Textbuchbeispiel für einen Regressionseffekt und damit ein Methodenartefakt interpretiert werden kann (für eine Diskussion des Problems vgl. Nachtigall & Suhl, 2002: 7). Denn er zeigt im Wesentlichen, dass sich Gruppen mit hohen anfänglichen Unterschieden im Verlauf der Zeit einander angleichen. Wichtig ist in diesem Zusammenhang, dass die Werte der Triple P Gruppe zu keinem Messzeitpunkt nach der Intervention besser werden als jene in der Kontrollgruppe.

Die hier geschilderten Probleme unterstellen nicht, dass die Daten absichtlich manipuliert wurden, um positive Ergebnisse zu erzielen. Sie dokumentieren aber, dass die von Petrosino und Soydan (2005) „zynisch" genannte Interpretation der Diskrepanzen zwischen Eigenevaluationen und Fremdevaluationen nicht leichtfertig vom Tisch gefegt werden sollte: Vielmehr wäre es erstaunlich, wenn es bei Eigenevaluationen nie zu Situationen käme, in denen Evaluatorenteams mit einem Eigeninteresse an den Ergebnissen Gefahr laufen, problematische methodische Entscheide und Interpretationen so miteinander zu verknüpfen, dass die publizierten Ergebnisse nicht als unverfälschte Schätzungen der tatsächlich erzielbaren Wirkungen interpretiert werden können.

Diskrepanzen zwischen der öffentlichen Darstellung und den tatsächlich nachweisbaren Effekten von Programmen

Drittens soll ein Problem aufgegriffen werden, das in der bisherigen Fachdiskussion zur Qualitätssicherung von Evaluationsstudien nur am Rande gewürdigt wurde. Es bezieht sich auf die Frage, ob die Ergebnisse von Eigenevaluationen *gegenüber der Öffentlichkeit* fair dargestellt werden.

Dabei müssen zwei Interessen gegeneinander abgewogen werden: So ist es einerseits völlig legitim, dass Programmbetreiber auf Webseiten und in Prospekten die Vorteile ihrer Produkte herausstreichen, positive Ergebnisse von wissenschaftlichen Studien betonen und die gemessenen Wirkungen in eine einfache Sprache übersetzen, soweit sich die Darstellung im Rahmen einer populären Interpretation der tatsächlich erzielten Effekte bewegt. Dem steht entgegen, dass Praxis und Laienpublikum ein Interesse daran haben, ein wirklichkeitsnahes Bild des Wissensstandes zu erhalten und zuverlässig über das Wirkungspotential eines Präventionsprogramms unterrichtet zu werden.

Das Problem soll am Beispiel des Gewaltpräventionsprogramms *Faustlos* illustriert werden: *Faustlos* ist ein Programm zur Förderung sozialer Kompetenzen und zur Reduktion von Gewalt und Aggression, dass sich in der Bundesrepublik Deutschland seit einigen Jahren sehr großer Beliebtheit erfreut. Es basiert auf dem Sozialkompetenzprogramm *Second Step*, welches in den 1980er Jahren von der US-amerikanischen gemeinnützigen Organisation *Committee for Children* entwickelt worden war. Es ist in den USA weit verbreitet (Grossman et al., 1997; Grossman et al., 1987). Nach Einschätzung der deutschen Vertreiber wird auch *Faustlos* inzwischen „in tausenden von Kindergärten und Schulen mit grossem Erfolg einge-

setzt“. Hierzu gehört beispielsweise, dass die LBS Bayern im Jahr 2004 *Faustlos*-Patenschaften in über 1000 Schulen und Kindergärten Bayerns übernommen hat und diese Patenschaften in den Jahren 2005 und 2006 erneuert hat.

Die Internetseite von *Faustlos* gibt sich bezüglich der erzielten Wirkungen nicht bescheiden. Auf der Frontseite des Internetauftritts wird etwa damit geworben, dass *Faustlos* als *Best-Practice* Projekt in Deutschland gilt. An prominenter Stelle steht dort: *Aufgrund seiner Effektivität und "überregional beispielhaften Qualität" wurde Faustlos im Rahmen der vom Bundesministerium für Bildung und Forschung in Auftrag gegebenen Bestandsaufnahme zu demokratiepolitischen und gewaltpräventiven Potenzialen in Schule und Jugendhilfe als "Best-Practice-Projekt" ausgezeichnet* (siehe www.faustlos.de). Eine Internet-Recherche zeigt, dass dieser Text vielfach übernommen und als Beleg für die Wirksamkeit von Faustlos interpretiert wurde.

Es lohnt sich daher, dieser Auszeichnung genauer nachzugehen. Sie erfolgte im Rahmen einer Expertise und Ausstellung zum Thema *Demokratie lernen in Schule und Gemeinde – demokratiepolitische und gewaltpräventive Potenziale in Schule und Jugendhilfe* (Beutel et al. , 2001). Liest man diesen Bericht, so findet man, dass die Auswahl von insgesamt 35 „best practice“ Projekten nicht auf einem Wirkungsnachweis basierte, sondern sehr pragmatisch aufgrund eines *explorierenden Vorgehens* erfolgte (Beutel et al., 2001: 5). Der Bericht nennt Lernqualität, institutionelle Qualität, Aktualität und Originalität als Kriterien.

Allerdings werden diese Kriterien nicht genauer definiert und empirisch überprüfte Effektivität wird nirgends als Auswahlkriterium erwähnt. In der Kurzbeschreibung von *Faustlos*, die man im Anhang des erwähnten Berichtes findet (Beutel et al., 2001: 65-66) wird zudem ausdrücklich darauf hingewiesen, dass zur Wirksamkeit keine Aussagen gemacht werden können, weil entsprechende Auswertungen zum Zeitpunkt der Auszeichnung noch gar nicht vorlagen. Mit anderen Worten: Die Behauptung, die Auszeichnung als „Best-Practice Projekt“ sei aufgrund der empirisch nachgewiesenen Wirkungen erfolgt, entspricht nicht der tatsächlichen Sachlage.

Die Programmbetreiber haben inzwischen in zwei deutschen Studien die Wirkungen von *Faustlos* untersucht und die Ergebnisse hierzu auch publiziert (Schick & Cierpka, 2003, 2005). Das Problem bilden hierbei weniger die wissenschaftlichen Publikationen und die dort berichteten Ergebnisse. Problematisch ist die Art und Weise, wie die Ergebnisse öffentlich weiter verbreitet werden. Beispielsweise ist im Faltblatt zu *Faustlos* folgendes zu lesen: „Mit Faustlos liegt ein deutschsprachiges Curriculum vor, das die zentralen gewaltpräventiven Kompetenzen Empathie, Impulskontrolle und den Umgang mit heftigen Gefühlen bei Kindern und Jugendlichen gezielt fördert. Die Effektivität des Programms wurde durch zahlreiche Studien belegt“ (Heidelberger Präventionszentrum, 2005). Laien, Politiker und Praktiker müssen aufgrund eines solchen Textes den Eindruck haben, es handle sich bei *Faustlos* um ein Programm mit einem wissenschaftlich abgestützten Nachweis von gewaltpräventiven Wir-

kungen. Tatsächlich zeigen die Publikationen von Schick und Cierpka (Schick & Cierpka, 2003, 2005) aber etwas völlig Anderes.

Die erste Studie basiert auf einer kontrollierten Experimentalstudie, die zwischen 2001 und 2003 in 44 Grundschulklassen im Raum Heidelberg/Mannheim durchgeführt wurde.[2] Aus der Perspektive der Kinder wurden in dieser Studie sieben Zielgrößen überprüft. Als *wirkungslos* erwies sich *Faustlos* in Bezug auf die Zielgrößen Empathie, Akzeptanz bei anderen Kindern, Selbstvertrauen, Selbstwertgefühl, Angst vor Verletzungen, Angst vor schlimmen Dingen, sowie aggressivem Verhalten. Es konnte ein positiver Effekt auf Angst vor Kontrollverlust gefunden werden.

Die Angaben der Eltern wurden verwendet, um zwölf Zielgrößen zu prüfen. *Keine präventiven Wirkungen* ergaben sich für folgende Masse: sozialer Rückzug, körperliche Beschwerden, soziale Probleme, schizoid/zwanghaftes Verhalten, Aufmerksamkeitsstörungen, delinquentes Verhalten, aggressives Verhalten, Selbstkontrolle, Selbstbehauptung, Perspektivenübernahme und Kooperation/soziale Regeln. Einzig auf Angst/Depressivität wurde ein knapp signifikanter positiver Effekt gefunden.

Schließlich wurde die Entwicklung der Kinder nach Einschätzung der Lehrpersonen untersucht. Hierbei wurden sechs Zielgrößen analysiert, nämlich: Das Ausmaß von Bandenbildung; die Bereitschaft, anderen zu helfen; Aggression gegen Klassenmitglieder; Diskriminierung gegen Klassenmitglieder; Zusammenhalt zwischen Klassenmitgliedern; Rivalität zwischen Klassenmitgliedern. Für *keine* dieser Zielgrößen berichten die Autoren über eine statistisch abgesicherte Wirkung von *Faustlos*.

Zusammengefasst bedeutet dies, dass für 23 der 25 überprüften Zielgrößen *kein statistisch abgesicherter Effekt* gefunden werden konnte. Hierzu gehören alle Zielgrößen, die sich entweder auf Gewalt selbst oder auf Vorläufer von Gewalt beziehen. Die beiden schwach signifikanten Effekte hingegen wurden in Bereichen beobachtet, die angesichts der Ziele von *Faustlos* marginal sind. Hinzu kommt, dass bei 25 überprüften Effekten rein zufällig ein bis zwei signifikante Wirkungen zu erwarten sind.

In einer zweiten Wirkungsstudie wurde die Version von *Faustlos* an Kindergärten überprüft (Schick & Cierpka, 2004, 2006). An der Studie nahmen 124 Kinder in 14 Kindergärten teil, die nach einem nicht näher erläuterten Verfahren den Experimentalbedingungen zugewiesen wurden. Die Ergebnisse dieser zweiten Studie fallen für die Daten, die direkt bei den Kindern erhoben wurden, positiv aus:[3] Die Autoren können erwünschte Effekte bei unmittelbaren In-

[2] Die Ergebnisse sind in den Tabellen 2, 3 und 4 von Schick und Cierpka (2003) dargestellt.

[3] Es ist an dieser Stelle anzufügen, dass die Aussagekraft der Studie durch mehrere methodisch problematische Merkmale eingeschränkt ist. Hiervon seien zwei namentlich erwähnt. Erstens nehmen die Autoren bei der Prüfung der Äquivalenz von Kontroll- und Experimentalgruppe aus nicht näher erläuterten Gründen eine methodisch fragwürdige Korrektur der Signifikanzschwellen vor (Schick und Cierpka, 2004: 14-16). Lässt man diese Korrektur unberücksichtig, dann zeigen die Daten, dass die Kinder in der Experimentalgruppe (im Vergleich zur Kontrollgruppe) vor der Intervention statistisch signifikant jünger waren, dass sie tiefere soziale

dikatoren wie der Emotionserkennung, den sozial kompetenten Reaktionen und dem Einsatz von Beruhigungstechniken zeigen, wenn sie aus der Sicht der Kinder selbst beurteilt werden (Schick & Cierpka, 2004: 19). Aggressives Verhalten wurde allerdings aus der Sicht der Kinder nicht gemessen.

Die Autoren finden jedoch keine Wirkungen von *Faustlos*, wenn dieselben Dimensionen aus der Sicht der Eltern beobachtet werden. Die Eltern selbst nehmen keine Verbesserung der emotional-sozialen Kompetenzen des Kindes wahr. Es gibt aus der Sicht der Eltern auch keine Effekte von Faustlos auf Aggressivität oder Ängstlichkeit (Schick & Cierpka, 2004: 20f).

Ebenso wenig hat *Faustlos* aus der Perspektive der Erzieherinnen messbare Effekte auf die Kompetenzen und das Verhalten der Kinder. Dies gilt für alle neun geprüften Verhaltensdimensionen einschließlich Aggressivität (Schick & Cierpka, 2004: 21). In den Verhaltensbeobachtungen zeigt sich kein Effekt von Faustlos auf verbales Verhalten, nonverbale Kompetenz, emotionale Kompetenz und körperliche Aggression. Hingegen finden die Autoren einen kleinen positiven Effekt auf verbale Aggression (Schick & Cierpka, 2004: 22).[4]

Dass bei der Messung von Interventionseffekten durch mehrere Informanden widersprüchliche Ergebnisse gefunden werden, ist leider in der Präventionsforschung eher die Regel als die Ausnahme. Auch erweist es sich in allen Studien zu universeller Gewaltprävention als überaus schwierig, Effekte auf der Ebene des tatsächlichen Kindsverhaltens nachzuweisen. Die Studien zu Faustlos bilden hier keine Ausnahme.

Allerdings: Wenn in einem Text der Praktikerzeitschrift „Schulverwaltung“ zu lesen ist, das Gewaltpräventionsprogramm Faustlos habe „gerade für ein Präventionsprogramm bemerkenswert große Effekte“ (Schick, 2004) erzielt, dann kann der Eindruck nicht völlig von der Hand gewiesen werden, hier werde ein insgesamt doch recht zwiespältiges wissenschaftliches Ergebnis über Gebühr strapaziert und der Präventionspraxis in Bild vermittelt, das durch die Forschung nicht gedeckt ist.

Folgerungen und Empfehlungen

Die wissenschaftspraktische Bewegung der evidenzbasierten Prävention ist in den letzten 20 Jahren mit dem Anspruch angetreten, für Praxis und Politik zuverlässiges Wissen darüber aufbereiten zu können, welche Maßnahmen wirken, welche nicht wirken, und welche schädlich sind. Es ist erfreulich, dass diese Bewegung in den letzten 10 Jahren auch im deutsch-

Kompetenzen hatten und höhere Werte von Problemverhalten nach Eltern- und Lehrereinschätzung aufwiesen. Es kann daher nicht ausgeschlossen werden, dass die Befunde durch Regressionseffekte verzerrt sind. Zweitens wurde bei der Analyse der Effekte die Klumpenrandomisierung nicht berücksichtigt. Eine Klumpenrandomisierung verlangt eine Anpassung der inferenzstatistischen Absicherung. In der vorliegenden Studie dürften daher alle Signifikanzwerte der Interventionseffekte systematisch überschätzt sein.

[4] Allerdings ist darauf hinzuweisen, dass diese Variable vermutlich eine starke Schiefverteilung aufweist, was sich in einer doppelt so hohen Standardabweichung im Vergleich zum Mittelwert ausdrückt (M = 0.015; SD = 0.032 bei einer Skala von 0-1). Es dürfte daher fraglich sein, ob eine konventionelle Varianzanalyse angemessen ist.

sprachigen Raum Fuß gefasst hat und neue Anstöße für die Präventionspolitik ausgelöst hat. Es ist ebenfalls zu begrüßen, dass sich Wissenschaftler zunehmend für die Entwicklung von Präventions- und Interventionsprogrammen interessieren und sich der Aufgabe einer Wirkungsevaluation mit Hilfe experimenteller Designs stellen. Angesichts der manchmal sehr hohen Erwartungen der Öffentlichkeit ist allerdings zu beachten, dass dieses Ziel nur dann dauerhaft erreicht werden kann, wenn die Wissenschaft unvoreingenommen und mit der nötigen Vorsicht die möglichst sorgfältig ermittelten Ergebnisse kommuniziert und hierbei keine unrealistischen Erwartungen weckt, welche auf Dauer nicht eingelöst werden können.

Die im vorangehenden Abschnitt gezeigten Einzelbeispiele dokumentieren, dass die Überscheidung von Forschungstätigkeit, Vertrieb von Präventionsprogrammen sowie politischen Beratungsfunktionen nicht unproblematisch ist. Es kommt dabei notwendigerweise zu Interessenkonflikten, die etwa der Situation ähnlich sind, wo ein Jugendlicher Richter über sein eigenes Verhalten sein müsste. Abschließend sollen daher einige Möglichkeiten vorgestellt werden, welche diesen Interessenkonflikt entschärfen können.

Verbindliche Richtlinien für Durchführung und Publikation von Wirkungsevaluationen
Das Programmentwickler ihre eigenen Programme wissenschaftlich evaluieren, ist grundsätzlich zu begrüßen. In dem Ausmaß, in dem in Zukunft mehr standardisierte und in Zusammenarbeit zwischen Forschung und Wissenschaft erarbeitete Präventionsprogramme auf den Markt kommen.

Das Problem der verzerrenden, ungenauen und möglicherweise durch Interessenkonflikte beeinflussten Berichterstattung über Evaluationsstudien wurde in der Medizin deutlich früher als in der Sozialwissenschaft erkannt. Daher wurden in diesem Forschungsbereich schon seit längerem Richtlinien über die Durchführung und Publikationen von Experimentalstudien verfasst, die von wissenschaftlichen Zeitschriften und Fachverbänden als verbindlich betrachtet werden. Was die rein wissenschaftlichen Anforderungen anbelangt, könnten sich Forschende beispielsweise problemlos an den CONSORT (Consolidated Standards for Reporting Trials) orientieren (Altman, 1996). Es enthält alle notwendigen Anweisungen, welche für eine wissenschaftliche Beurteilung notwendig sind. Außerdem existieren spezifisch für die kriminologische Präventionsforschung einschlägige Publikationen, welche Forschenden detaillierte Qualitätsstandards an die Hand geben, welche verschiedene Aspekte der Validität umfassen (Farrington, 2003; Lösel & Koferl, 1989). Besonders hilfreich in dieser Hinsicht ist die Liste von Qualitätskriterien für Evaluationsprojekte, welche Farrington (2003) zusammengestellt hat. Es wäre sinnvoll, wenn auch im deutschsprachigen Raum solche Qualitätsstandards in der praxisorientierten Forschung eine möglichst große Verbreitung finden würden. Fachvereinigungen, Institutionen der Forschungsförderung und öffentliche Auftraggeber könnten beispielsweise Checklisten der Informationen erstellen, die in Forschungsberichten zwingend enthalten sein müssen.

In den USA, wo standardisierte und kommerziell vertriebene Präventionsprogramme schon seit längerem existieren und ein entsprechend vielfältiges Angebot besteht, wurde schon vor

gut 10 Jahren die Notwendigkeit erkannt, Praktikern Leitfäden zur Beurteilung von Präventionsprogrammen an die Hand zu geben. Dabei wurde auch erkannt, dass solche Beurteilungen in einem transparenten Reviewverfahren durch unabhängige wissenschaftliche Gremien mit einer hohen Vertrauenswürdigkeit vorgenommen werden sollten.

Modellcharakter in dieser Hinsicht haben die Blueprints of Violence Prevention des Center for the Study and Prevention of Violence an der Universität von Colorado (Elliott & Mihalic, 2004). Diese 1996 begonnene Initiative evaluiert einzelne Präventionsprogramme mit besonderem Blick auf die Frage, ob Programme den Nachweis der Wirksamkeit erbracht haben und sie lokalen Entscheidungsträgern zur Umsetzung empfohlen werden können. Bisher wurden 600 Programme in einem rigiden Verfahren geprüft, hiervon wurden bisher 11 als Modellprogramme eingestuft, weitere 28 gelten gegenwärtig als viel versprechend. Die Programme werden durch ein Gremium von sieben hervorragend qualifizierten Spezialisten der Evaluationsforschung nach standardisierten Kriterien evaluiert. Die wichtigsten dieser Kriterien sind nachweislich positive Effekte in einem rigiden Forschungsdesign, Nachhaltigkeit der Effekte über das Ende der Intervention hinaus sowie Replikation der positiven Effekte in mindestens zwei weiteren, unabhängig durchgeführten Studien.

Angesichts der zunehmenden Popularität von evidenzbasierter Prävention – und der steigenden Zahl von Programmanbietern – wäre es sinnvoll, im deutschen Sprachraum ähnliche Mechanismen der Qualitätskontrolle und praxisnahen Aufbereitung des Wissensstandes zu schaffen. Gegenwärtig gibt es aber weder in der Schweiz noch in Deutschland solche unabhängigen Reviews, welche im Sinne einer Qualitätskontrolle wirken würden.

Unabhängige Evaluationen fördern

Es gibt viele Gründe, warum Evaluationen mit einer starken aktiven Beteiligung der Programmentwickler oder –vertreiber auch in Zukunft die Regel bleiben werden. Hierzu gehört etwa das Interesse an einer wissenschaftlichen Validierung. Es wird in dem Maße weiter zunehmen, als die öffentlichen Nachfrager nach Prävention einen wissenschaftlichen Leistungsausweis zur Bedingung für die Übernahme von Programmen machen. Hinzu kommt, dass der externe Finanzierungsbedarf bei Eigenevaluationen in der Regel sehr viel geringer ist als bei Evaluationen durch Dritte, weil die Programmentwickler oft einen beträchtlichen Teil der Kosten selbst übernehmen.

Dennoch sollten Wissenschaftsräte und Stiftungen gezielt auch unabhängige Evaluationen fördern. Es gibt drei wesentliche Gründe, weshalb unabhängige Evaluationen ein wichtiges Korrektiv zu den Befunden von Selbstevaluationen sind. Erstens ist anzunehmen, dass Programmentwickler besonders sorgfältig auf eine erstklassige Umsetzung, hohe Motivation und überdurchschnittliche Betreuung des Projektes achten. Das ist legitim. Allerdings sind solche Bedingungen in der Alltagspraxis kaum je gegeben. Daher ist es notwendig, Programme auch so zu evaluieren, wie sie auf dem Markt angeboten und im Alltag tatsächlich realisiert werden. Zweitens besteht vermutlich generell in der Präventionsforschung die Tendenz, er-

wünschte Ergebnisse eher zu publizieren als Nullbefunde oder gar unerwünschte Effekte. Diese auch als Publikationsbias oder „Schubladenproblem“ bekannt Phänomen (Rosenthal, 1979) dürfte besonders dann ein Problem sein, wenn für eine Forschergruppe materielle oder immaterielle Interessen mit dem Evaluationsausgang verbunden sind. Drittens dürften Forschende, welche ein Eigeninteresse am geprüften Programm haben, besonders in Versuchung sein, die publizierten Ergebnisse in die gewünschte Richtung zu schönen indem unerwünschte Teilbefunde unterschlagen oder die Ergebnisse durch problematische Manipulationen zurechtgebogen werden.

Bessere Ausbildung von Nutzern in Praxis und Verwaltung
Häufig verfügen die Nutzer von sozialwissenschaftlichen Wirkungsevaluationen nur über sehr begrenztes Wissen, um an eine Studie die richtigen Fragen zu stellen und die Ergebnisse kritisch beurteilen zu können. Es scheint wichtig, dass die verantwortlichen Nutzer von Wirkungsstudien in Politik und Verwaltung über das Fachwissen verfügen, um den wissenschaftlichen Output beurteilen zu können und die richtigen Fragen zu stellen. Es könnte daher sinnvoll sein, ein entsprechendes passives Fachwissen durch Weiterbildungsveranstaltungen und Kurse zu fördern. Fachpersonen in der Praxis sollten verstehen, welche Kriterien die Qualität einer Wirkungsevaluation beeinflussen; welche Forschungsdesigns am besten auf eine Forschungsfrage angemessen sind; welche Informationen sie von einem wissenschaftlichen Bericht erwarten dürfen; und woher sie allenfalls Hilfe bei der Beurteilung der berichteten Effekte einholen können.

Literatur

Altman, D. G. (1996). Better Reporting of Randomised Controlled Trials: the CONSORT Statement. *BMJ, 313*(7057), 570-571.

Asshauer, M., & Hanewinkel, R. (2000). Lebenskompetenztraining für Erst- und Zweitklässler: Ergebnisse einer Interventionsstudie. *Zeitschrift für Klinische Kinderpsychologie, 9*, 251-263.

Beutel, W., Schnurre, S., Senge, K., Thöne, A., & Fauser, P. (2001). *Demokratie lernen in Schule und Gemeinde - Demokratiepolitische und gewaltpräventive Potenziel in Schule und Jugendhilfe*. Berlin: Bundesministerium für Bildung und Forschung.

Borman, G. D., Hewes, G. M., Overman, L. M., & Brown, S. (2003). Comprehensive School Reform and Achievement: A Meta-Analysis. *Review of Educational Research, 73*(2), 125-230.

Campbell, D., & Stanley, J. C. (1966). *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.

Cochrane, A. (1972). *Effectiveness and Efficiency: Random Reflections on Health Services*. London: Nuffield Provincial Hostpitals Trust.

Cook, T. D., & Campbell, D. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

Eisner, M., & Ribeaud, D. (2008). Kritischer Kommentar zu Nina Heinrichs, Kurt Halweg, Heike Bertram, Annett Kuschel, Sebastian Naumann, Sylvia Harstick: Die langfristige Wirksamkeit eines Elterntrainings zur universellen Prävention kindlicher Verhaltensstörungen: Ergebnisse aus der Sicht der Mütter und der Väter. Pädagogisches Institut der Universität Zürich.

Eisner, M., Ribeaud, D., Jünger, R., & Meidert, U. (2007). *Frühprävention von Gewalt und Aggression; Ergebnisse des Zürcher Interventions- und Präventionsprojektes an Schulen*. Zürich: Rüegger.

Ellickson, P. L. (1998). Preventing Adolescent Substance Abuse: Lessons from the Project ALERT Program. In J. Crane (Ed.), *Social Programs That Work* (pp. 201-257). New York: Russell Sage.

Ellickson, P. L., & Bell, R. M. (1990). Drug Prevention in Junior High: A Multi-siteLongitudinal Test. *Science, 247*, 1299-1305.

Ellickson, P. L., Bell, R. M., & Harrison, E. R. (1993). Changing Adolescent Propensities to Use Drugs: Results from Project ALERT. *Health Education Quarterly, 20*, 227-242.

Ellickson, P. L., Bell, R. M., & McGuigan, K. (1993). Preventing Adolescent Drug Use: Long-Term Results of a Junior High Program. *American Journal of Public Health, 93*(11), 1830-1836.

Elliott, D., & Mihalic, S. (2004). Issues in Disseminating and Replicating Effective Prevention Programs. *Prevention Science, 5*(1), 47-53.

Farrington, D. P. (2003). Methodological Quality Standards for Evaluation Research. *Annals of the American Academy of Political and Social Sciences, 587*, 49-68.

Gandhi, A. G., Murphy-Graham, E., Petrosino, A., Chrismer, S. S., & Weiss, C. H. (2007). The Devil is in the Details: Examining the Evidence for "Proven" School-Based Drug Abuse Preventiopn Programs. *Evaluation Review, 31*(1), 43-74.

Grossman, D. C., Neckerman, H. J., Koepsell, T. D., Asher, K. N., Beland, K., Frey, K., et al. (1997). Effectiveness of a Violence Prevention Curriculum Among Children in Elementary School: A Randomized Controlled Trial. *Journal of the American Medical Association, 277*(20), 1605-1611.

Grossman, D. C., Neckerman, H. J., Koepsell, T. D., Liu, P.-Y., Adher, K. N., Beland, K., et al. (1987). Effectiveness of a Violence Prevention Curriculum among Children in Elementary School - A Randomized Controlled Trial. *Journal of the American Medical Association, 277*(20), 1605-1611.

Heidelberger Präventionszentrum. (2005). *Faustlos - Gewaltprävention durch Förderung sozial-emotionaler Kompetenzen*. Heidelberg: Heidelberger Präventionszentrum.

Heinrichs, N., Hahlweg, K., Bertram, H., Kuschel, A., Naumann, S., & Harstick, S. (2006). Die langfristige Wirksamkeit eines Elterntrainings zur universellen Prävention kindlicher Verhaltensstörungen: Ergebnisse aus Sicht der Mütter und Väter. *Zeitschrift für klinische Psychologie und Psychotherapie, 35*(2).

Henggeler, S. W., Cunningham, P. B., Pickrel, S. G., Schoenwald, S. K., & Brondino, M. J. (1996). Multisystemic Therapy: an Effective Violence Prevention Approach for Serious Juvenile Offenders. *Journal of Adolescence, 19*(1), 47-61.

Henggeler, S. W., Melton, G. B., & Smith, L. A. (1992). Family Preservation using Multisystemic Threapy: An Effective Alternative to Incarceration. *Journal of Consulting and Clinical Psychology, 60*(6), 953-961.

Littell, J. (2005). Lessons from a Systematic Review of Effects of Multisystemic Therapy. *Children and Youth Services Review, 27*(4), 445-463.

Littell, J., Popa, M., & Forsythe, B. (2005). Multisystemic Therapy for Social, Emotional, and Behavioral Problems in Youth Aged 10-17 (report for the Campbell Collaboration) [Electronic Version]. Retrieved 14 January 2008 from http://www.sfi.dk/graphics/Campbell/Dokumenter/MST_Review/MULTISYSTEMIC%20TERAPHY%20-%20REVIEW.pdf.

Lösel, F., & Beelmann, A. (2003). Effects of Child Skills Training in Preventing Antisocial Behavior: A Systematic Review of Randomized Evaluations. *The ANNALS of the American Academy of Political and Social Science, 587*(1), 84-109.

Lösel, F., Beelmann, A., Stemmler, M., & Jaursch, S. (2006). Probleme des Sozialverhaltens im Vorschulalter: Evaluation des Eltern- und Kindertrainings EFFEKT. *Zeitschrift für klinische Psychologie und Psychotherapie, 35*(2), 127-139.

Lösel, F., & Koferl, P. (1989). Evaluation Research on Correctional Treatment in West Germany: A Meta-Analysis. In H. Wegener, F. Lösel & J. Haisch (Eds.), *Criminal Behavior and the Justice System: Psychological Perspectives*. New York: Springer.

Nachtigall, C., & Suhl, U. (2002). Der Regressionseffekt - Mythos und Wirklichkeit. *methevalreport http://www.metheval.uni-jena.de/materialien/reports/report_2002_02.pdf, zuletzt aufgerufen am 11.1.2008), 4*(2).

Ogden, T., & Hagen, K. A. (2006). Multisystemic Treatment of Serious Behaviour Problems in Youth: Sustainability of Effectiveness Two Years after Intake. *Child and Adolescent Mental Health, 11*(3), 142-149.

Olsson, T., & Sundell, K. (2008). The Transportability of MST to Sweden: Short-term Results from a Randomized Trial of Conduct Disordered Youth (Manuskript, zur Publikation eingereicht).

Petrosino, A., & Soydan, H. (2005). The Impact of Program Developers as Evaluators on Criminal Recidivism: Results from Meta-analyses of Experimental and Quasi-experimental Research. *Journal of Experimental Criminology, 1*(4), 435-450.

Rosenthal, R. (1979). The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin, 86*(3), 638-641.

Rössner, D., Bannenberg, B., & Landeshauptstadt Düsseldorf. (2002). *Düsseldorfer Gutachten: Empirisch gesicherte Erkenntnisse über kriminalpräventive Wirkungen*. Düsseldorf: Landeshauptstadt Düsseldorf.

Sanders, M. R. (1999). Triple P-Positive Parenting Program: Towards an Empirically Validated Multilevel Parenting and Family Support Strategy for the Prevention of Behaviour and Emotional Problems in Children. *Clinical Child and Family Psychology Review, 2*(2), 71–89.

Sanders, M. R., Lynch, M. E., & Markie-Dadds, C. (1994). *Every Parent's Workbook: A Practical Guide to Positive Parenting*. Brisbane: Australian Academic Press.

Schick, A. (2004). Inhalte, Implementation und Effektivität eines Gewaltpräventions-Curriculums. *Schulverwaltung Spezial*(3), 22-24.

Schick, A., & Cierpka, M. (2003). Faustlos - Evaluation eines Curriculums zur Förderung sozial/emotionaler Kompetenzen und zur Gewaltpraevention in der Grundschule. *Kindheit und Entwicklung, 12*(2), 100-110.

Schick, A., & Cierpka, M. (2004). *Evaluation des Faustlos-Curriculums für den Kindergarten (Schriftenreihe der Landesstiftung Baden-Württemberg, Nr 7)*. Retrieved 14 January 2008,from http://www.landesstiftungbw.de/publikationen/files/schr_eval_faustkinder.pdf.

Schick, A., & Cierpka, M. (2005). Faustlos: Evaluation of a Curriculum to Prevent Violence in Elementary Schools. *Applied and Preventive Psychology, 11*(1), 157-165.

Schick, A., & Cierpka, M. (2006). Evaluation des Faustlos-Curriculums für den Kindergarten. *Praxis der Kinderpsychologie und Kinderpsychiatrie, 55*(6), 157-165.

Shadish, W. R., Cook, T. D., & Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Sherman, L. W., Farrington, D. P., Welsh, B. C., & MacKenzie, D. L. (Eds.). (2002). *Evidence-Based Crime Prevention*. London: Routledge.

St Pierre, T. L., Osgood, D. W., Mincemoyer, C. C., Kaltreider, D. L., & Kauh, T. J. (2006). Results of an Independent Evaluation of Project ALERT Delivered in Schools by Cooperative Extension. *Prevention Science, 6*(4), 305-317.

Timmons-Mitchell, J., Bender, M. B., Kishna, M. A., & Mitchell, C. C. (2006). An Independent Effectiveness Trial of Multisystemic Therapy With Juvenile Justice Youth. *Journal of Clinical Child & Adolescent Psychology, 35*(2), 227-236.

Inhalt

III. Überblick zum 12. Deutschen Präventionstag