

## Abstract

„Evaluation“ ist ein Kind prosozialer und demokratischer Bestrebungen, so wie es auch die grossen frühpädagogischen Förderprogramme sind. Der Kern von Evaluation war immer schon Teil dieser Förderprogramme und der Frühkindlichen Bildung. Evaluation dient Programmen wie der institutionalisierten Frühpädagogik zur Vergewisserung über ihre Wirkungen und zur Reflexion ihrer Weiterentwicklung. Sie stellt auch ihrerseits Ansprüche an die Bildungseinrichtungen und Programme hinsichtlich Zielklarheit, Organisiertheit, Wirkungsbewusstheit und vor allem hinsichtlich Kooperation. Dieser wechselseitige Anspruch ist noch lange nicht erfüllt, da sich beide Felder, die frühkindliche Bildung wie der ihr zugewandte Teil der Evaluation jeweils noch in einer stürmischen Entwicklung befinden. Für die Untersuchung der methodischen Entwicklung der Evaluation erscheint daher eine Modellierung der Bewegung beider Professionen aufeinander zu sinnvoll. Wie diese Arbeitsbeziehung begann, wie der Stand aktuell ist und wohin die Entwicklung zu weichen scheint, davon handelt dieser Beitrag.<sup>1</sup>

## Definition von Evaluation und begriffliche Eingrenzung

Evaluation hilft grossen soziotechnischen Entwicklungsvorhaben, ihre Arbeiten und ihre Ressourcennutzung zielwirksamer, also effizienter und effektiver zu gestalten, indem sie für diese vorübergehenden, komplexen Projekte ein Zielführungs-, Feedback- und Wissensmanagementsystem schafft.<sup>2</sup> Wirksame, wissens- und evidenzbasierte Evaluation ist an allen Phasen pädagogischer Entwicklungsvorhaben beteiligt: Planung, Durchführung, Wirkungsüberprüfung und Wirksamkeitsbewertung und folgt damit dem Dewey'schen „Plan-Do-Check-Act-Handlungsreflexionszyklus“ (vgl. Carle, 2000, S. 391). Ihre Unterstützungsleistung erbringt Evaluation, indem sie bei der Schaffung eines Systems zur Indizierung von Gestaltungswirkungen in komplexen Entwicklungsvorhaben zielführungsrelevante Prozessmerkmale (Daten) sammelt, aufbereitet und fragenbezogen wie auch entscheidungsgerecht interpretiert. Die Methoden dieses Zielführungs-, Feedback- und Wissensmanagementsystems sind entsprechend vielseitig. Sie beinhalten auch, aber bei weitem nicht nur Methoden der Sozialforschung.

Auch der Begriff der *Evaluationsmethodik* muss geklärt werden, da er auf drei unterschiedlichen Handlungsebenen verwendet wird: Auf der höchsten, der *institutionellen Verantwortungsebene* meint „Evaluationsmethodik“ das grundsätzliche Verhältnis der Evaluation zum Entwicklungsfeld, d. h. die strategische Ausrichtung, die organisatorische Einbettung der Evaluation in den Programmentwicklungsprozess und den Einfluss der Evaluation auf die Programmm Zukunft. Beispiele sind die „Nutzenorientierte Evaluation“, die „Wirkungsorientierte Evaluation“, die „Realistische Evaluation“, die „Entwicklungsorientierte Evaluation“ oder die „Partizipative Evaluation“. Je nach Entwicklungsstand des Programms finden sich auf dieser Ebene auch „explorative“, „formativ“ oder abschliessend bewertende, „summative“ Methodiken.

Auf der mittleren, der *erhebungsglogischen Durchführungsebene* meint „Evaluationsmethodik“ die logische Schrittfolge der Evaluation, beispielsweise die Folge, die Form und die Verbindung von Zielklärung, Erfolgskriterienentwicklung, Erhebungsdesignbestimmung bis hin zur Berichtsformwahl. Beispiele sind experimentelle, qualitative, quantitative, dokumentarische oder Mehrebenen-Evaluationen.

Auf der untersten, der *operativ-instrumentellen Ebene* meint „Evaluationsmethodik“ die Anlage (das Design) und das Instrumentarium der Erhebung, beispielsweise ein experimentelles Erhebungsdesign, eine teilnehmende Beobachtung oder ein statistisches Auswertungsinstrument. Alle drei Ebenen, die der Wirkungsorientierung, der Durchführung und der Instrumentierung sind miteinander verschränkt.

Im Folgenden sollen einige weitere Begriffe, die gewöhnlich mit Evaluation verbunden werden, von dieser abgegrenzt werden (vgl. Alkin, 2011, S. 10). Der Begriff Evaluation wird in der Früh- und Schulpädagogik auch im Sinne von *Qualitätsentwicklungs- und Selbstbewertungsverfahren* verwendet (z. B. Carle, 2010). Hierfür – wie auch für „Large Scale Assessment“ á la PISA oder *Organisationsbewertung* nach dem EFQM-Verfahren – wird im Folgenden wegen des fehlenden Entwicklungsprojektcharakters beim Bewerteten nicht von Evaluation, sondern von *Qualitätsentwicklung* bzw. *Assessment* gesprochen. Evaluation, Qualitätsentwicklung und Assessment im pädagogischen Bereich besitzen historisch gesehen zwar gemeinsame Quellen, haben sich aber zu jeweils eigenständigen Professionen auseinanderentwickelt. Das gilt auch für das *Monitoring*. Monitoring. Monitoring meint die ständige Beobachtung und Protokollierung des Verlaufs interessierender Entwicklungen, etwa das „Sozio-ökonomische Panel (SOEP)“, aus dem z. B. Studien zum „Volkswirtschaftlichen Nutzen langfristiger

<sup>1</sup> Weiterführende und vertiefende Analysen zur wechselseitigen professionellen Entwicklung von Evaluation und Pädagogik finden sich bei Carle/Metzen, 2012.

<sup>2</sup> Zur Meisterung der Komplexität von Entwicklungsvorhaben siehe Patton 2011, S. 111ff; zur komplexitätsreduzierenden Verwendung der Strukturen und Techniken des Wissensmanagements in der Evaluation siehe Hense/Mandl, 2011.

Bildungseffekte bei Krippenkindern“ (Fritschi/Oesch, 2008) oder zur „Wirkung der Kindergartenteilnahme auf die Schulkarriere“ (Spiess u.a., 2003) hervorgehen.

Bleibt die Frage, was sind *Meta-Analysen* im Vergleich zu Evaluationen? Sind sie Über-Evaluationen? - Ja, wenn die Meta-Analyse hermeneutischer Natur ist und die Interpretationen der Evaluationsstudien kritisch reflektiert und theoretisch ordnet (z. B. Burger, 2010). Sie sind keine Über-Evaluationen, wenn die Meta-Analysen rein statistischer Natur sind („Effektstärken-Evaluation“). Meta-Analysen z.B. zu den (Neben-) Wirkungen neuer Medikamente in der Medizin können eine aussagefähige Basis bieten, wenn sich die Einzelanalysen auf sehr ähnliche Massnahmen (z. B. gleiches Medikament in ähnlicher Dosierung) und auf ähnliche Ergebnisse (Symptomlinderung in gut definierbarem Grade) beziehen.

Ein weiterer verwandter Begriff, die *Evaluationsforschung*, wird doppelsinnig gebraucht, für Forschung über Evaluation (Metaanalysen, Metaevaluationen, Methodologische Studien etc.), oder zur Abgrenzung sozialwissenschaftlich arbeitender pädagogischer Wirkungsuntersuchungen zu nicht sozialwissenschaftlichen. (z. B. Prengel u.a., 2008, S. 189).

## Entwicklungsphasen der Evaluationsprofession

Bronfenbrenner konstatierte nach einer sehr fundierten, methodenkritischen Analyse von sieben frühpädagogischen Förderprogrammen in den USA Anfang der 1970er Jahre, dass die univariable Intervention, wie sie dem deterministisch-experimentellen Effektstärkedenken zugrunde liegt, der Wirkfaktorenkonstellation realer Lebens- und Lernwelten nicht gerecht wird. Damit stellt sich für ihn weniger die Frage, wie sehr diese Programme für welche Kinder mit welcher Wirkung und wie lange im Laufe ihrer Schullaufbahn wirken, sondern eher, ob die Anlage dieser Frühförderprogramme den realen lebensweltlichen Problemen dieser Kinder tatsächlich gerecht wird. Seine Antwort ist die Forderung nach einer *Ökologischen Intervention* („ecological intervention“), der Herstellung des sprichwörtlichen Dorfes, dessen die Entwicklung eines Kindes ja wohl bedarf. Und er resümierte 1974: Die Ökologische Intervention ist bis jetzt eine weitgehend ungenutzte Vorgehensweise, in unserer Wissenschaft und unserer Gesellschaft (ebd., S. 296 - freie Übersetzung d. A.). Sie ist es bis heute cum grano salis geblieben. Wie so eine „ökologische Intervention“ wirken kann, belegt die jüngste Sure Start Studie aus England (National Evaluation of Sure Start (NESS) Team, 2012).

Eine erste Charakterisierung der aus ihrer Sicht wichtigsten Entwicklungsphasen der Evaluationsprofession unternahmen Egon G. Guba und Yvonna S. Lincoln 1989 in ihrem Buch „Fourth generation evaluation“. Dabei unterscheiden sie vier paradigmatische Phasen, die sie durch eine vorherrschende, die vorangegangenen Schwerpunkte erweiternde Evaluationsstrategie kennzeichnen (Guba/Lincoln, 1989, S. 22 ff):

1. „*Measurement Generation*“ - vorherrschend ist die Messung verschiedener Eigenschaften und Schulleistungen von Schulkindern; die Definition der Massstäbe (Tests) wird qua fachlicher Autorität vorgegeben; Evaluation ist hauptsächlich Messtechnik (Metrologie)<sup>3</sup>.
2. „*Description Generation*“ - die Kybernetik 1. Ordnung erobert das Forschen: beschrieben wird die Differenz des SOLL-IST-Vergleichs, der Grad der Zielerreichung; die Massstäbe hierfür ergeben sich aus den Indikatoren für die Zielerreichung von Massnahmen; EvaluatorInnen engagieren sich vorwiegend als Programm-InterpretatorInnen und Ziel-OperationalisierInnen; Messen wird zu einem unter mehreren Erhebungswerkzeugen; in dieser Zeit entwickelt sich die „Programm-Evaluation“<sup>4</sup>.
3. „*Judgement Generation*“ - der Evaluation wurde von den Verantwortlichen zunehmend die Nutzenbewertung sozialpolitischer Grossprogramme abverlangt; Evaluation sollte die politische Entscheidungsfindung unterstützen; zur Überprüfung der Programmwirkung sollte die Einschätzung des Programmwirkungsnutzens hinzu kommen; mit dieser Herausforderung geraten die WirkungsadressatInnen stärker in den Fokus der Evaluation.<sup>5</sup>
4. „*Constructivistic Evaluation Generation*“ - die Beteiligung der Programmakteure und der ProgrammnutzerInnen, ihre Forderungen, Anliegen und Fragen avancieren zu inhaltlich-methodischen Schwerpunkten der Evaluation; dabei bleiben das Messen, die Zielerreichung und die Nutzenbewertung weiterhin zentrale Leistungen der Evaluation, auch wenn sich diese erheblich erweitern hin zu Multimethoden-

<sup>3</sup> So wurde der gut gemeinte „Kampf gegen die Armut“ in Form des nur notdürftig geplanten Head Start-Programms mit einem Sammelsurium an Evaluationstests „überprüft“ (siehe dazu Zigler/Styco 2010, S. 70 ff; zur Detailkritik an der Studie siehe Besharov et al. 2011; zu weiteren Hintergründen siehe Carle/Metzen, 2012, S. 15 ff).

<sup>4</sup> Für den Bereich der Pädagogik findet sich eine aufschlussreiche Gegenüberstellung des alten *experimentell-positivistischen Objektivismus* der frühen Evaluationszeit mit dem modernen nutzen- und serviceorientierten multi-methodischen *Design-based Research* der heutigen Evaluation bei Cook und Gorard (2007).

<sup>5</sup> Das Denken in Nutzenkategorien betrachtet Wirkung (Effektivität) und Aufwand (Effizienz). So entwickelten sich in den 1970er Jahren erste Schritte in Richtung Prozessdenken; zu einigen Details und Quellen siehe Carle/Metzen 2012, S. 18 ff.

ansätzen mit deutlich erhöhter Ziel- und Nutzen-Diversität. Beides erlaubt vertiefte Einsichten in individuelle, kollektive und organisationale Entwicklungsprozesse von pädagogischen Förderprogrammen.<sup>6</sup>

Die von Guba und Lincoln skizzierten Fortschritte in der Evaluationsmethodik sind sehr wichtig, reichen aber nach unserer Erfahrung nicht aus, wenn Evaluation in der Praxis wirksam werden soll. Insbesondere erscheint uns die Kooperation zwischen Programmgestaltern, Programmumsetzern und Programmumsetzungsunterstützern - darunter den Evaluatoren - noch unzureichend entwickelt. Gerade diesbezüglich hat sich seit den 1980er Jahren sehr viel getan (siehe Carle/Metzen, 2012, S. S. 18 ff<sup>7</sup>). Wir fügen daher den vier Entwicklungsphasen von Guba und Lincoln zwei weitere hinzu:

- Eine Phase der sich *vertiefenden Kooperation* zwischen Evaluation und Programmgestaltung. Hierbei steht die Entdeckung und die methodische wie organisationale Berücksichtigung der Nichttrivialität pädagogischer Entwicklungsvorhaben und ihrer Evaluation als wechselwirkender, komplexer, entwicklungsöffener, mehrschichtiger Designprozess im Mittelpunkt.
- Die derzeit anstehende Phase ist gekennzeichnet durch die *beginnende Koevolution* von entwicklungsförderlicher Evaluation und kontrollierter (transparenter und reflektierter) frühpädagogischer Programmentwicklung. Evaluation macht sich auf dem Weg hin zu einem prozessintegrierten und prozessunterstützenden Entwicklungsmoment pädagogischer Entwicklungsvorhaben („Systemische Institutionen-Begleitforschung“). Die Koevolution erscheint als förderliche Bedingung zur Autonomisierung von ProgrammgestalterInnen und EvaluaterInnen.

Diese Kennzeichnung der Entwicklungsphasen bildet eine doppelte Abstraktion: Sie sind Idealtypen, d. h. sie verzichten auf die Darstellung der in der Realität immer vorfindlichen, partiellen Gleichzeitigkeit von Vorbereitung, Verbreitung, Mainstreaming, Professionalisierung. Sie abstrahieren zudem von der je dominierenden gesellschaftlichen Entwicklung, die die Autonomisierung der Frühpädagogik wie auch die ihr zugewandte wissenschaftliche Politik- und Praxisberatung fördern oder behindern kann. Die von uns hinzugefügten und im Folgenden näher betrachteten Phasen folgen auch einem praktisch wichtigen Punkt: Wir gehen nämlich davon aus, dass die Fachleute für Evaluation nur im Zusammenwirken mit den Programmgestaltern in Planung, Umsetzung und Umsetzungsunterstützung zu brauchbaren Ergebnissen kommen können.

Die Verbesserung und Vertiefung der Kollaboration hin zur Kooperation setzt die von Guba und Lincoln in ihrem Entwurf einer „konstruktivistischen“ Evaluation (1989) geforderte Gleichwertigkeit und Gleichberechtigung voraus, wohingegen Ungleichheit vor allem zu Abgrenzungen und damit zu massiven „organisationalen“ Wissensverlusten führt (dazu Cook/Brown, 1999). Deutlich wird das am Fall ‚Head Start‘. Er ist ein klassisches Beispiel für die Interaktionsentwicklungsstufe eines konkurrenten „Antagonismus“. Obwohl auch im Umfeld des Head Start-Projekts entstanden, bietet das Sesame Street-Projekt aus evaluationsmethodischer Sicht dazu einen deutlichen Kontrast. Das Fernseh-Programm selbst wurde von ausgewiesenen Profis ihres Metiers konzipiert, realisiert und optimiert. Die „formativen“ WissenschaftlerInnen des Educational Testing Service aus Princeton waren von Beginn an der Planung des Fernsehprogramms beteiligt. Bereits im Evaluationsbericht über das erste halbe Jahr bescheinigten sie dem TV-Programm eine deutlich positive Wirkung auf die Testleistungen (Klassifizieren, Buchstaben erkennen usw.). Dieser Befund wurde im Abschlussbericht bestätigt (Bogatz/Ball, 1971).

Die „summative“ Evaluation wurde nachträglich fachfremd als Meta-Evaluation von Cook und Mitarbeitern im Auftrag der Russell Sage Foundation<sup>8</sup> auf Basis der Daten der formativen Evaluation durchgeführt. Cook errechnete daraus einen starken Effekt des wöchentlichen Besuchs der ForschungsassistentInnen in den Familien und keinen wesentlichen Effekt des reinen Zuschauens sowie eine verstärkende Wirkung auf den Lerngewinn der nicht benachteiligten gegenüber den stark benachteiligten Kindern (Cook/Conner, 1976). Diese Kritik ist typisch für die unpädagogische Orientierung der instrumentalistischen Evaluationsschule. Erstens war die von Cook und anderen geprüfte, kompensatorische Wirkung des Sesame Street-Programms weder vom Programm noch von seinen EvaluaterInnen geplant. Zweitens ist das, dem statistischen Auseinanderrechnen von Programmnutzungswirkung und familiärer Kontextwirkung zugrunde liegende Wirkungsmodell mechanistisch und

---

<sup>6</sup> Die vorausgegangene Phase der beginnenden Nutzenorientierung litt noch unter einem grundsätzlichen Mangel. Die beteiligten Expertisen dachten immer noch vorrangig an ihren eigenen Nutzen und an den Nutzen ihrer unmittelbaren AuftraggeberInnen. Dieses Denken gewann auch bei den Evaluierten in den pädagogischen Programmen zunehmend an Boden. Die „konstruktivistische Sicht“ auf Programmentwicklung und Evaluation versucht dieser Perspektivenvielfalt strategisch und methodisch-instrumentell durch eine grundsätzliche Bejahung von Pluralität und damit durch einen Verzicht auf Machtpositionen und Methodenmonismen im Evaluationsgeschehen gerecht zu werden. Die unterschiedlichen Werthaltungen und Interessen der Beteiligten werden in die Wirkungs- und Nutzenreflexionen einbezogen (Guba/Lincoln 1989, S. 184 ff).

<sup>7</sup> Dort findet sich auch sehr viel mehr weiterführende Literatur als in diesem Rahmen möglich ist.

<sup>8</sup> Die Russell Sage Foundation förderte seit den 20er Jahren des 20. Jahrhunderts die wissenschaftliche Fundierung der Sozialpolitik: [www.russellsage.org/about](http://www.russellsage.org/about).

entspricht einem konservativen, a-sozialökologischen Menschenbild - kein Kind lernt isoliert. Ein nicht sozial eingebundener Lernanlass ist weder vorstellbar noch herausrechenbar. Deutlich wurde: Auf jeden Fall reicht das einfache, testbasierte, experimentelle Vorher-Nachher-Experimental-Kontrollgruppen-Erhebungsdesign nicht zur lehrreichen Evaluation relevanter pädagogischer Programme. Auch dann nicht, wenn die Evaluation eines Programms schon bei seiner Konzeptionierung mit angedacht wird<sup>9</sup>.

### **Vertiefung der Kooperation: Entdeckung der Nichttrivialität pädagogischer Entwicklungsvorhaben und ihrer Evaluation („Gestaltungsorientierte Evaluation“)**

Was allen Programmen in dieser Zeit abging, war der Gedanke, dass man das Wirkungssystem und seine Methoden Schritt für Schritt verbessern kann. Die Frage nach der Effektivität eines Programms träte dann zurück gegenüber der Frage nach seiner Sensitivität für die Weiterentwicklung. Doch erst einmal übertrugen die EvaluationspraktikerInnen und -wissenschaftlerInnen die ihnen angestammten Theorien, Methoden und Instrumente auf die Arbeit am Evaluationsgegenstand. Damit entwickelte sich die „Programm-Evaluation“ als praxisorientierte und von den Belangen der Programmentwicklungspraxis vorangetriebene Forschungsdisziplin (siehe Rossi u.a., 1979; vgl. Rossi u.a., 2004, 7. Aufl.).

In USA hatten sich schon Evaluations-Fachgesellschaften gebildet<sup>10</sup>. 1974 konstituierte sich „Joint Committee on Standards for Educational Evaluation (JCSEE)“ als freiwilliges Gremium, das auch heute noch für das „American National Standard Institute (ANSI)“ arbeitet und von diesem akkreditiert wurde. Diese ursprünglich für die Verwendung von Tests in der Evaluation gegründete Arbeitsgruppe veröffentlichte 1981 die „Standards for Evaluations of Educational Programs, Projects, and Materials“. Dabei handelte es sich um eine Überarbeitung und Erweiterung der „Standards for Educational and Psychological Tests and Manuals“, 1966 herausgegeben von der „American Psychological Association (APA)“. 1978 erschien die erste Ausgabe der Zeitschrift „New Directions for Program Evaluation“<sup>11</sup>, in der die „Evaluation Research Society (ERS)“ 1982 die „Standards for Program Evaluation“ veröffentlichte (Rossi, 1982, S. 1-97). Bemerkenswerterweise enthielten diese stark instrumentell orientierten Standards noch keinen Hinweis auf die von den Programmen angestrebten und ihnen zugrunde liegenden Normen und Werte. 1986 schlossen sich die ERS mit dem 1975 gegründeten „Evaluation Network (ENet)“ zur „American Evaluation Association (AEA)“ zusammen. Damit schlossen sich die Mehrzahl der eher instrumentenorientierten („methodenorientierten“) in der ERS mit der Minderzahl der eher relevanzorientierten (Multimethodiker) in der ENet zusammen. Eine glückliche, wenn auch nicht konfliktarme Liaison, da es Aufgabe der Evaluationsprofession war und ist, diese Polarisierung zu überwinden. Als ein wichtiges Ergebnis legte die AEA 1994 die vom Joint Committee überarbeiteten „Program Evaluation Standards“ vor, in der praktische Relevanz („Nützlichkeit“) und Wertorientierung eine hervorgehobene Rolle spielen (zur Geschichte: Mertens, 2000, S. 46 ff; zu den Standards: Yarbrough et al./Joint Committee, 2011/1981).

Die vorwiegend aus der Sozialforschung stammende Evaluationsprofession bewegt sich also in den 1970er/1980er Jahren mit grossen Schritten auf ihre pädagogischen Gegenstände zu und damit z. T. auch weg von deren methodologischer Monokultur einer experimentell-testmetrologischen Methodik. Das äussert sich in der zunehmenden Verwendung „qualitativer“ Zugänge zum Evaluationsfeld (Weiss, 1998, S. 252 ff; Alkin, 2011, S. 109 ff). Über diese rein methodische Anerkennung Eigenständigkeit bzw. der Besonderheiten des Evaluationsfeldes hinaus fordern jedoch immer mehr Evaluationswissenschaftler die Unterordnung des Evaluationsprozesses unter die Entwicklungsaufgaben der Programme und die realen Gegebenheiten des Entwicklungsfeldes. Michael Quinn Patton spitzt seine Botschaft zu: Evaluation, die den Programmen nicht nützt ist wertlos, egal wie gut sie methodisch angelegt ist (Patton, 2008/1978).

Die vierte „Evaluationsgeneration“ von Guba und Lincoln, die „Konstruktivistische Evaluation“, rückt die Forderungen, Anliegen und Fragen der Programmbeteiligten in den Mittelpunkt des Evaluationsinteresses (Guba / Lincoln, 1989, S. 50). Sie unterscheiden dabei nicht nur zwischen den Programm- bzw. Evaluationsauftraggebern, Programmumsetzern und den Programmnutzern und bei letzteren wiederum zwischen Programmwirkungsgewinnern und -verlierern (ebd., S. 40 f). Konstruktivistisch oder systemisch ist an ihrem Ansatz vor allem die Öffnung des theoretischen Wirkungsbildes der Evaluationsprofession für die Vielfältigkeit und Vielschichtigkeit des Evaluationsfeldes, die sich nicht mehr einfach durch eine experimentelle Labortestsituation einfangen lässt. Guba und Lincoln sehen ihren Beitrag weniger in einer Alternative zum experimentellen Positivismus, sondern mehr in einer methodischen (multimethodischen) Erweiterung, die der wachsenden Verantwortung entspricht,

<sup>9</sup> Bezeichnenderweise beginnt Margrit Stamm die Einleitung ihre Untersuchung zu den Nutzungsbedingungen von Evaluationen im pädagogischen Bereich mit dem provokanten Titel: „Wirkungslos - undurchsichtig - ambivalent“ (dies., 2003, S. 23).

<sup>10</sup> In Deutschland geschah dies erst 1998 (siehe [www.degeval.de](http://www.degeval.de)).

<sup>11</sup> Die „New Directions for Program Evaluation“ heissen seit dem Herbst 1995 „New Directions for Evaluation“.

die viele Evaluationswissenschaftler für das Gelingen des Programmerfolgs zu übernehmen bereit sind (ebd., S. 38 ff).

Der methodische Monismus und die experimentelle Distanziertheit gegenüber den Förderprogrammen und prospektiven Nutzniessern scheinen also mit Beginn der 1990er Jahre zumindest prinzipiell überwunden. Dennoch stehen Programmverantwortliche, Programmumsetzer, Programmnutzer und Evaluatoren immer noch vor zwei fast unüberwindlichen Problemen. Die situative Anpassung an den realen Verlauf der Programmumsetzung ändert nichts an der Fülle von Unwägbarkeiten, denen sich die Programmumsetzung angesichts nur skizzenhaft entwickelter Programme und angesichts kaum bekannter Zielgruppenverhältnisse gegenübersteht. Letztere hatte Bronfenbrenner bereits in den 1970er Jahren eingehender analysiert und war zu dem Schluss gekommen, dass die kindliche Entwicklung alleine durch individuelle Merkmale nur unzureichend erklärbar ist (Bronfenbrenner, 1974, S. 295 ff). Sein sozialökologisches Entwicklungsmodell ergänzte Bronfenbrenner 1986 durch ein chronosystemisches Theorem, nach dem nicht nur das soziale Umfeld einen bedeutsamen Einfluss auf die soziale, körperliche und intellektuelle Persönlichkeitsentwicklung von Kindern hat, sondern zusätzlich vom bisherigen Entwicklungspfad und von der aktuellen Zeitströmung mit beeinflusst wird („*Chronosystem*“ - ders., 1986). Wirksame pädagogische Förderprogramme müssen also um den Preis des Scheiterns die individuellen wie die überindividuellen *sozialen und zeitlichen Umfeldbedingungen* mit beeinflussen.

Eine erste konzeptionelle Zerschlagung dieses gordischen Knotens gelang einer Kombination aus Lernpsychologie und technisch-organisatorischen Errungenschaften (Designwissenschaft, Künstliche Intelligenz und Computerunterstütztes Lernen). 1992, mehr als zwanzig Jahre nach Herbert A. Simon's bahnbrechendem Buch zur Grundlegung einer *Designwissenschaft* in Abgrenzung zur Naturwissenschaft (1996/1969, deutsch 1994) veröffentlichten die im Umfeld des „Situiereten Lernansatzes“ forschende Lernpsychologin Ann L. Brown und der Kognitionswissenschaftler Allan M. Collins je einen Artikel zum Entwurf einer Theorie der designorientierten Erforschung des Pädagogischen (Brown, 1992; Collins, 1992). Beide orientierten sich am Lernansatz des „*Situiereten Lernens*“, den eine Universitätskollegin von Ann Brown an der University of California in Berkeley in einem Buch zur Diskussion gestellt hatte (Lave/Wenger, 1991, S. 13). Mit der „praxisintegrierten Feldforschung“, die sich stärker am Vorrang der ökologischen Validität orientierte, war man zwar jetzt der Praxis sehr viel näher. Dafür hatte man sich aber gleichzeitig von den damals vorherrschenden methodischen Standards der positivistischen, „kontrollierten“ Sozialforschung entfernt (Stark/Mandl, 2007, S. 250).

Die Unterrichtsforscherin Ann Brown schildert ihre Verzweiflung angesichts der Unkontrollierbarkeit didaktischer Experimente in realen Unterrichtssituationen. Sie sieht sich als Forscherin mit einem besonderen Interesse für *unterrichtliche Lerngemeinschaften* gefangen im Dilemma zwischen praxisrelevanten Untersuchungen in vivo aber unter kaum kontrollierbaren Wirkungsbedingungen auf der einen Seite und dem Gegenteil, der „sauberen Forschung“ unter kontrollierbaren Laborbedingungen aber ohne unterrichtsrelevante Forschungsbefunde auf der anderen Seite. Als Ausweg erschien ihr die gerade erschienene Konzeptskizze des Kognitionswissenschaftlers Allan M. Collins' für eine designorientierte Gestaltung des pädagogischen Experimentierens und Forschens. 2004 editierten die Kognitionswissenschaftler und Erziehungstechnologen Sasha Barab und Kurt Squire (2004) im „*Journal of the Learning Sciences*“ ein Sonderheft zum Design-based Research. Auch im deutschsprachigen Europa gibt es eine WissenschaftlerInnen, die sich mit dem neuen Forschungs- und Entwicklungsparadigma „Design-based Research“ auseinandersetzen (Reinmann, 2005; DGfE-Kommission Schulforschung und Didaktik (KSD), 2007; Stark / Mandl, 2007; Krüger, 2010; Reusser, 2011; Vohle / Reinmann, 2012). Wir werden im Folgenden als Synonym für Design-based Research auch die Übersetzung „Gestaltungsorientierte Forschung“ bzw. im evaluativen Zusammenhang „*Gestaltungsorientierte Evaluation*“ verwenden.

### **Beginnende Koevolution: Evaluation als prozessintegriertes und prozessunterstützendes Entwicklungsmoment („Systemische Institutionen-Begleitforschung“)**

1998 veröffentlichte Maja Heiner die wichtigsten Ergebnisse einer Fachtagung des Bundesministeriums für Familie, Senioren, Frauen und Jugend zu organisationalem Lernen und Qualitätsentwicklung in der Jugendhilfe unter dem Titel „Experimentierende Evaluation“ (Heiner, 1998). In ihrem einleitenden Beitrag plädiert die Sozialpädagogin der Universität Tübingen für eine lern- und innovationsunterstützende Evaluation, die Evaluieren als heuristisches Experiment von Forschung und Praxis (ebd., S. 25) einsetzt und dabei Organisationsveränderung (Organisationsentwicklung) durch Kopplung von Praxisentwicklung und Praxisuntersuchung (ebd., S. 32) unterstützt. Nach der summativen Evaluation der ersten Entwicklungsphase und der formativen Evaluationsorientierung aus der Zeit der stärkeren Nutzenorientierung steht die Evaluation nun mitten im Prozessgeschehen. Die formative Evaluation beschränkte sich noch auf programmbegleitende Zwischenuntersuchungen, quasi portionierte phasen-summative Evaluationen. Das ergibt prozessnähere Zwischenbefunde und erlaubt so rechtzeitiges Nachsteuern und Korrigieren. Wie aber wird dann die Rückwirkung dieser Korrekturen auf die ursprünglichen Projektziele erhoben? Manche Autoren versuchten darauf eine planerische Antwort: Modellierung und Pla-

nung des Entwicklungsvorhabens müssen verbessert werden (so etwa bei Wottawa/Thierau, 1998, 83 ff), Evaluation und Erhebung müssen projektspezifischer arbeiten. Das Lernen wollen aus dem realen Projektverlauf wird dadurch zwar erheblich verbessert, aber auch erheblich aufwendiger.

Kluge Entwicklungsgestaltung geht einen anderen, ökonomischeren und lernoffeneren Weg der hierarchisch-sequentuellen Planung und Umsetzung (Carle, 2000, 418 ff; Patton, 2011, S. 305 ff). Zuerst wird das übergeordnete Entwicklungsziel skizziert. Dabei sollten sehr ehrgeizige Ziele gesetzt werden, damit durch diese Apriori-Ausweitung des Zielhorizontes später unerwartet auftauchende Zusatzwirkungen verortet werden können. Entsprechend dieser übergeordneten, langfristigen, qualitativ definierten Ziele werden grobe Indikatoren formuliert - quasi das Basismodell der „summativen“ Systementwicklungsevaluation. Detaillierter geplant, vorbereitet und organisiert wird dann im Hinblick auf die übergeordneten Ziele die erste überschaubare Entwicklungsphase. Hierfür können nun realistische Ziel-, Wirkungs-, Handlungs- und Indikatorenmodelle ausgearbeitet werden. Die Möglichkeiten zur realitätstüchtigeren Planung dieser Phase definiert die Entwicklungsspanne, nicht abstrakte Zeitziele. Für diese erste Phase werden nun die einzelnen Entwicklungsbedingungen erhoben und erst auf dieser Basis die detaillierteren Entwicklungsarbeiten und Evaluationsmassnahmen geplant. Zu den vorgesehenen Indikatoren werden die passenden Erhebungsinstrumente ausgesucht bzw. entwickelt sowie ein Nutzenbestimmungs- und Entscheidungsmodell für die Handlungskonsequenzen aus erwarteten bzw. unerwarteten Befunden („Bewertungsmodell“) ausgearbeitet. Die Planung der jeweils folgenden Phasen erfolgt erst dann, wenn genügend phasenspezifisches Systemwissen vorhanden ist. Dieses Wissen über die nächste Phase wird parallel zur Durchführung der vorherigen Phase erhoben.

Ergibt sich aufgrund unerwarteter positiver wie negativer Ereignisse / Gegebenheiten die Notwendigkeit, muss auch die Evaluationsstrategie geändert werden. Dazu eignet sich eine „Minievaluation“. Sie kann plötzlich auftauchende Handlungs- und Entscheidungsschwierigkeiten der Entwicklungsarbeiter lösen helfen<sup>12</sup>. Im Mittelpunkt der Kurzevaluationen steht natürlich nicht die Bewertung der Produkte eines Programms, sondern die kurzfristige Veränderung im Verhalten, in den Beziehungen, in den Handlungen der Menschen und der Organisation. Dieses kurzfristige Eingehen auf alltägliche Entwicklungsprobleme ist allerdings nur dann sinnvoll, wenn das Entwicklungsvorhaben auch der grundsätzlichen Strategie der „Kontinuierlichen Verbesserung“ folgt (siehe Bösenberg/Metzen, 1995, S. 108 ff), wenn also jede Möglichkeit des Lernens in einer Optimierung der Arbeitsbedingungen und einer Verbreiterung der Wissensbasis aller Beteiligten im Entwicklungsprojekt mündet.

Die Energie zum Wandel weht aus Richtung der vitalen Eigeninteressen der Strukturebenen (z. B. zwischen Kindergärten, Trägern und der Politik). Zielführungs- und Unterstützungssysteme einschliesslich der Evaluation können diese Interessendifferenzen intelligent nutzen. Dafür steht die von uns präferierte *Integration von interner und externer Evaluation* (Details siehe Carle/Metzen, 2012, S. 29 f; zum Konzept der serviceorientierten Evaluation siehe Carle/Metzen, 2003). Bei dieser gewagten Position sehen wir uns auch von Michael Quinn Patton bestärkt (2011, S. 64 ff). Diese Integration von interner und externer Evaluation auf den wichtigsten Strukturebenen von Sozialprojekten lässt sich anhand eines kybernetischen Mehrebenenmodells theoretisch und praktisch erläutern. Dieses Mehrebenenmodell geht davon aus, dass sich jede Entwicklung in einer mitgestaltenden Umgebung („Ökosystem“) vollzieht und umgekehrt bezieht sich jede selbstgesteuerte Entwicklung auf diese mitwirkenden Umfelder.

Die interne Evaluation<sup>13</sup> dient der Selbstvergewisserung der Ebenen-Akteure, die externe Evaluation der Beratung der Zielvereinbarungserreichung - auch Patton (2011, S. 23ff) plädiert für ihre Verschränkung. Die gemeinsam erarbeiteten Evaluationsziele leiten sich aus den Zielvereinbarungen zwischen den Ebenen ab. Durch diese evaluativen Verknüpfungen und Separierungen gewinnen die „Kunden“ der Evaluation auf den verschiedenen Strukturebenen ein Eigeninteresse an der Konzertierung (Abstimmung und Ko-Produktion) von Entwicklungsprozessen, Projektmanagement und Evaluationen, wobei der Projektleitung die bedeutsame Rolle zukommt, das langfristige Projektziel immer präsent zu halten (siehe Carle/Metzen, 2009, S. 131 ff). Welche Veränderungen auf welcher Ebene besonders betrachtet werden, hängt nicht nur vom Interesse der Evaluationskunden ab, sondern auch von der Entwicklungszeit, die erfasst werden kann und soll. Die schnellsten Veränderungen sind auf der operativen Ebene feststellbar. Veränderungen auf der strukturellen Ebene bedürfen in Bildungseinrichtungen - je nach Strukturebene - Monate bis Jahre. Auf der übergeordneten normativen Ebene braucht es sogar Generationen für nachhaltige und signifikante Veränderungen. Diese Überlegungen fliessen in die Methoden- und Instrumentenwahl ein. Das war früher einfacher: Am Anfang der Evaluation stand ja das Gutachten, das seinerseits

<sup>12</sup> Hier einige Ansätze zur Durchführung von Kurzevaluationen: Brinkerhoff, 2003 Success Case Method; Earl/Carden /Smutylo, 2002 Outcome Mapping; Widmer/Rothmayr/Serdült, 1996, Kurzevaluationen.

<sup>13</sup> „Interne Evaluation ist der in unserem Evaluationskontext treffendere Ausdruck für das, was auch unter dem Label „Selbstevaluation“ firmiert, auch wenn die interne Evaluation nicht alleine von den Evaluierten getragen wird, sondern durch ihre Kooperation mit den Evaluierenden; siehe hierzu DeGEval, 2004 (Standardsadaptation zur Selbstevaluation).

einem von „Oben“ vorgegebenen Bewertungsmodell folgte. Was in der Komplexität des gutachterlichen Sachverständnisses noch intuitiv verankert war, bedarf in einem kollektiven oder gar organisationalen Lernprozess der expliziten Zielfindungs- und Entscheidungslogik.

Wo aber bleiben nach all diesen Empfehlungen für eine wirksame Unterstützung pädagogischer Entwicklungsvorhaben durch Evaluation die Wissenschaft und die ihr wichtigen neuen Erkenntnisse? Noch immer wird in der Evaluationsliteratur die Fahne der instrumentellen Exzellenz hoch gehalten, zumal die AutorInnen zumeist aus den entsprechenden Hochschulbereichen stammen. Aus ihrer Sicht ist das Eingehen auf die Gegebenheiten des zu evaluierenden Programms oder Produkts ein Kompromiss zwischen den hohen forschungsmethodischen Ansprüchen und den realen Einschränkungen hinsichtlich Feldzugang, Untersuchungszeit und personell-sächlichen Ressourcen. Wir plädieren für eine andere Sicht auf diese Situation: Die wahrgenommenen Beschränkungen können auch als Zugangsvoraussetzungen zu einem hoch interessanten Forschungsfeld gesehen werden. Die methodologischen Restriktionen eröffnen die Chance zur Entwicklung prozessnaher Erhebungsmöglichkeiten. Der Zwang zur evaluativen Unterstützung der Entwicklungsvorhaben bietet die Gelegenheit zur reflexiven Beeinflussung der real wirkenden Entwicklungsmomente.

Die dieser Auffassung zugrunde liegende „Generative Forschungslogik (Epistemologie)“ geht davon aus, dass die „Dinge“ am intensivsten „mit uns reden“, wenn wir an ihrer Veränderung unterstützend beteiligt sind. Dieser Logik folgend, schöpft prozessintegrierte und engagierte Evaluation aus prinzipiell reichhaltigeren Quellen als distanzierte Laborforschung. So wie ein Chirurg bei der Operation der Anatomie des Lebendigen näher ist als bei seiner Arbeit am Seziertisch. Diese Nähe geschieht natürlich vordergründig auf Kosten der reflexiven Distanz. Der Laborvorteil, mit seinen Zusatzressourcen Zeitfülle, instrumentelle Variabilität und Wiederholbarkeit, kann aber durch einen Kunstgriff partiell auf die Felderhebungssituation übertragen werden: Die Evaluationserhebung wird um ein mit wissenschaftlichen Zusatzfragen gefülltes Surplus erweitert. Dies ist bei Expeditionen gang und gäbe. Raumfahrtexpeditionen oder Meinungsumfragen werden häufig um weitere, periphere Erkundungsaufträge ergänzt, die den Hauptauftrag nicht gefährden aber die günstige Gelegenheit kostengünstig mitnutzen. Ähnlich kann Evaluation verfahren und die Evaluationsfragen übersteigende oder vertiefende Forschungsfragen in ihre Erhebungen einplanen. Dazu werden die Planung, Vorbereitung, Durchführung, Auswertung und Präsentation der Evaluation mit diesem wissenschaftlich besonders interessierenden Surplus ausgestattet und nebenher bzw. im Nachhinein vertiefend untersucht. Das ist ohne weiteres praktikierbar und äusserst ergiebig (z. B. Berthold, 2011). Natürlich müssen die Verwendungsinteressen, die Inhalte und die Aufwände des Erhebungssurplus mit den Auftraggebern, den Evaluierten und dem Unterstützungssystem abgestimmt werden. Die so fundierte "Metaanalyse" hat dann ein verlässliches Fundament und muss ihre Interpretationen nicht auf ungesicherte Annahmen über Ähnlichkeiten bei Fragestellung und Datenqualität bauen.

Dieser Vorschlag zur Verbindung von Gestaltungsarbeit und Grundlagenforschung für wissenschaftlich engagierte EvaluatorInnen erscheint erst einmal beschwerlicher als das „reine“ Forschung im Labor. Vielleicht hilft den WissenschaftlerInnen, die sich diesen schwierigeren Weg zumuten wollen die Ironie der Wissenschaftsgeschichte, indem sich ihnen auf diesem Weg der unmittelbarste Zugang zu relevantem (früh-) pädagogischem Wissen eröffnet. Daran glaubt auch kein geringerer als Michael Scriven in seinem Entwurf einer „Transdisziplinären Evaluation“ (2003, S. 19 ff):

*"Applied social science will divide into the progressive, evaluation-enriched school, and the conservative, evaluation-impaired school. The evaluation-enriched group will become the winner in nearly all bids for contracts aimed at separating solutions from non-solutions of social / educational problems. The evaluation-impaired branch, following in the tracks of typical applied social science departments today, will gradually wither on the vine, with its aging adherents exchanging stories about the good old days."*

Wichtiger als alle methodischen Bedenken ist das evaluatorische Engagement in allen frühpädagogischen Feldern, die es verdienen, untersucht zu werden, deren Weiterentwicklung aber noch nicht der evaluatorischen Unterstützung wert erachtet wurden (Scriven, 2012, S. 169).

## **Zusammenfassung**

Dieser Beitrag betrachtet die widersprüchliche Entwicklung der Evaluationsprofession in den rund 50 Jahren ihrer öffentlichen Anwendung auf grosse frühpädagogische Förderprogramme. Sie hat sich zu einer weltweit beachtlichen Profession entwickelt und zeigt dennoch untrügliche Anzeichen weiteren Entwicklungsbedarfs: Nach Innen wird sie von über hundert unterschiedlichsten Ansätzen geprägt, die von grossen und kleinen Lehns Herren (Scriven: „Fiefdoms“) eifersüchtig verteidigt werden. Als Unterscheidungsmerkmale dienen vor allem methodische bzw. Nutzenargumente. Dabei sehen Letztere nicht sehr überzeugend aus. Produziert die Evaluation

doch überwiegend Analysen, die den Beurteilten mehr oder weniger Wirkungslosigkeit attestieren (Zusammenstellung siehe Carle/Metzen, 2012, S. 35 ff).

Nach unserer Analyse haben die Evaluation wie ihr Wirkungsfeld, hier die frühpädagogischen Förder- und Entwicklungsprogramme noch nicht die institutionelle und professionelle Autonomie erreicht, die sie erreichen könnten. Der Evaluation mangelt es überwiegend noch an der Fähigkeit, sich als Teilfunktion eines Unterstützungssystems für diese Förderprogramme zu etablieren, ohne ihre wissenschaftlichen Standards und Erkenntnisinteressen aufzugeben. Auch die Frühpädagogik sieht sich bis dato kaum imstande, der Evaluation auf Augenhöhe zu begegnen und sie als notwendige Entwicklungsbedingung zu verstehen (Stichwort: „Klientenprofessionalisierung“). Beide Disziplinen stehen also vor diesem letzten Schritt hin zu einem beginnenden koevolutiven Verhältnis. Dieser letzte Schritt verlangt von der wissenschaftlichen Evaluation die Überwindung überkommener naturwissenschaftlich-medizinischer Forschungsstrategien, nicht die Aufgabe der dieser Tradition zugrunde liegenden methodischen Strenge und wissenschaftlichen Generalisierungsansprüche. Die phasenweise Unterordnung der Evaluation unter den Entwicklungsauftrag der Förderprogramme (Perspektivenwechsel von „Attribution“ hin zu „Contribution“) würde sie mit Hilfe eines zweigleisigen Vorgehens in wissenschaftlicher Hinsicht näher an die Bedingungsfaktoren gelingender Programmentwicklung heran bringen.

Die so mögliche Koevolution hat zwei Seiten, eine innere und eine äussere. Die innere, der methodische Austausch wurde in diesem Artikel eingehender betrachtet - auch unter einem für die Pädagogik wie für die Sozialwissenschaften wenig vertrauten Aspekt des professionellen Organisierens. Die äussere Seite, das Verhältnis des Förderprogramm-Evaluations-Soziotops zum gesellschaftlichen Umfeld, verdiente ebenfalls eine vertiefte Betrachtung. Schliesslich bestimmt das gesellschaftliche Interesse an beiden Disziplinen, ob ihnen weitere Entwicklungsressourcen zur Verfügung gestellt werden<sup>14</sup>. Allerdings ist dieses Äussere auch aus der historischen Perspektive durchaus zwiespältig zu sehen. So kam insbesondere der Evaluation bislang in Bezug auf den Prozess der demokratischen und prosozialen Verantwortungsübernahme eine nicht selten äusserst ambivalente Rolle zu. Sie wurde sowohl als Hilfsmittel eingesetzt, um aus Erfolgen wie aus Misserfolgen der Förderprogrammentwicklung zu lernen. Sie wurde aber auch missbraucht, um das „Gelernte“ als Argument in entwicklungsfeindliche politische Entscheidungsprozesse einfließen zu lassen. Auch hiergegen würde eine vertiefte Kooperation wappnen.

## Literatur

Alkin, M. C. (2011). *Evaluation essentials from A to Z*. New York: Guilford Press.

Barab, S. & Squire, K. (Hrsg.) (2004). Design-Based Research. Clarifying the Terms. Introduction to the Learning Sciences Methodology Strand. *The Journal of the Learning Sciences* 13(1) Special Issue, 1-128.

Besharov, D., J., Germanis, P., Higney, C. A. & Call, D. M. (2011). *Ohio University Report on Head Start, 1969. Meta-Analysis of Twenty-Six Early Childhood Evaluations*, Westinghouse Learning Center, Bd. 27. College Park, MD, USA: Welfare Reform Academy, School of Public Policy, University of Maryland

Bösenberg, D. & Metzen, H. (1995). *Lean Management. Vorsprung durch schlanke Konzepte* (5. Auflage). Landsberg am Lech: Moderne Industrie.

Bogatz, G. A. & Ball, S. (1971). *The Second Year of Sesame Street. A Continuing Evaluation*. Vol. 1(ERIC Document Number ED122800); Vol. 2 (ED122801). Princeton, NJ, USA: Educational Testing Service (ETS).

Brinkerhoff, R. O. (2003). *The success case method. Find out quickly what's working and what's not*. San Francisco, CA, USA: Berrett-Koehler.

Bronfenbrenner, U. (1986). Ecology of the Family as a Context for Human Development. *Research Perspectives. Developmental Psychology* 22(6), 723-742.

Bronfenbrenner, U. (1974). Is Early Intervention Effective? *Teachers College Record* 76(3), 279-303.

---

<sup>14</sup> Das gesellschaftliche Interesse scheint vorzuliegen. So fordern die Washingtoner „National Association for the Education of Young Children (NAEYC)“ und die „National Association of Early Childhood Specialists in State Departments of Education (NAECS/SDE)“ in einem nun schon 10 Jahre alten gemeinsamen Positionspapier die Schaffung eines umfassenden Systems der integrierten Curriculumentwicklung, der Leistungsdokumentation und der Programmevaluation zur angemessenen Unterstützung frühkindlicher Förderung und Entwicklung.



Brown, A. L. (1992). Design Experiments. Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *The Journal of the Learning Sciences* 2(2), 141–178.

Burger, K. (2010). How does early childhood care and education affect cognitive development? An international review of the effects of early interventions for children from different social backgrounds. In *Early Childhood Research Quarterly* 25(2), 140–165.

Carle, U. (2010). Pädagogische Diagnostik als forschende Tätigkeit. In B. Friebertshäuser, A. Langer, A. Prengel & S. Richter (Hrsg.), *Handbuch qualitative Forschungsmethoden in der Erziehungswissenschaft* (3. Aktualisierte Auflage), (S. , 831–844). Weinheim: Juventa-Verlag.

Carle, U. (2000). *Was bewegt die Schule? Internationale Bilanz, praktische Erfahrungen, neue systemische Möglichkeiten für Schulreform, Lehrerbildung, Schulentwicklung und Qualitätssteigerung*. Baltmannsweiler: Schneider Hohengehren.

Carle, U. & Metzen, H. (2012). Phasen der wechselseitigen professionellen Entwicklung von Evaluation und Pädagogik. Historisch begründetes Plädoyer für den Ausbau der Kooperation von pädagogischer Praxis und Evaluation. Arbeitsgebiet Elementar- und Grundschulpädagogik, Universität Bremen. Bremen. Download am 18.08.2012 von [www.grundschulpaedagogik.uni-bremen.de/archiv/Carle/2012/CAMZ2012Entw.Eval+Paed.pdf](http://www.grundschulpaedagogik.uni-bremen.de/archiv/Carle/2012/CAMZ2012Entw.Eval+Paed.pdf).

Carle, U. & Metzen, H. (2009). Die Schuleingangsphase lohnt sich! Erfolgsmomente für die bestmögliche Entwicklung des landesweiten Schulentwicklungsvorhabens 'Begleitete Schuleingangsphase (BeSTe)' in Thüringen. Bericht der wissenschaftlichen Begleitung nach zweieinhalb Jahren 'BeSTe' (2005-2008). Bremen: Arbeitsgebiet Elementar- und Grundschulpädagogik, Universität Bremen. Download am 18.08.2012 von [www.grundschulpaedagogik.uni-bremen.de/forschung/thueringen/index.html#PAB](http://www.grundschulpaedagogik.uni-bremen.de/forschung/thueringen/index.html#PAB).

Carle, U. & Metzen, H. (2003). Evaluation und Evaluationsforschung – ein neues Paradigma: Entwicklungsservice statt Werkzeugkiste. Bremen: Arbeitsgebiet Elementar- und Grundschulpädagogik, Universität Bremen. Download am 18.08.2012 von [www.grundschulpaedagogik.uni-bremen.de/archiv/Carle/2003/camz2003evaluation\\_service.pdf](http://www.grundschulpaedagogik.uni-bremen.de/archiv/Carle/2003/camz2003evaluation_service.pdf).

Collins, A. (1992). Towards a design science of education. In E. Scanlon und T. O'Shea (Hrsg.), *New directions in educational technology. Proceedings of the NATO Advanced Research Workshop on New Directions in Advanced Educational Technology, Milton Keynes, GBR, 10.-13.11.1988* (S. 15-22). Berlin: Springer-Verlag (NATO ASI series, Series F, Computer and system sciences, 96).

Cook, S. D. N. & Brown, J. S. (1999). Bridging Epistemologies. The Generative Dance Between Organizational Knowledge and Organizational Knowing. *Organization Science* 10(4), S. 381–400.

Cook, T. D. & Conner, Ross F. (1976). The Educational Impact. Secondary Evaluation of seven studies of the influence of Sesame Street. *Sesame Street Around the World* 26(2), S. 155–164.

Cook, T. D. & Gorard, S. (2007). What Counts and What Should Count as Evidence. Chapter 2. In Organization for economic co-operation and development (OECD) Centre for Educational Research and Innovation (CERI) (Hrsg.), *Evidence in Education. Linking research and policy* (S. 33–49). Paris: Organization for economic co-operation and development (OECD).

DeGEval – Gesellschaft für Evaluation (Hrsg.) (2004). Empfehlungen zur Anwendung der Standards für Evaluation im Handlungsfeld der Selbstevaluation. Alfter: DeGEval Download am 20.07.2012 von <http://www.alt.degeval.de/calimero/tools/proxy.php?id=24059>.

DGfE-Kommission Schulforschung und Didaktik (KSD) (Hrsg.) (2007). *Allgemeine Didaktik trifft Lehr-Lernforschung*. Theorietagung der Kommission Schulforschung / Didaktik der Sektion Schulpädagogik der DGfE am 08.-09.03.2007 an der Universität Hildesheim. Universität Hildesheim, Centrum für Bildungs- und Institut für Erziehungswissenschaft Unterrichtsforschung (CeBU). Hildesheim: CeBU. Download am 18.08.2012 von [www.uni-hildesheim.de/media/forschung/cebu/PDFs/DGfE-KSD-Tagung-2007.pdf](http://www.uni-hildesheim.de/media/forschung/cebu/PDFs/DGfE-KSD-Tagung-2007.pdf).

Earl, S., Carden, F. & Smutylo, T. (2002). *Outcome mapping. Building learning and reflection into development programs*. Foreword by M. Q. Patton. Ottawa: International Development Research Centre (IDRC).

Fritsch, T. & Oesch, T. (2008). *Volkswirtschaftlicher Nutzen von frühkindlicher Bildung in Deutschland. Eine ökonomische Bewertung langfristiger Bildungseffekte bei Krippenkindern*. Gütersloh: Bertelsmann Stiftung.

Guba, E. G. & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA, USA: Sage Publications.

Heiner, M. (1998). Lernende Organisation und Experimentierende Evaluation. Verheißungen Lernender Organisation. In M. Heiner (Hrsg.), *Experimentierende Evaluation. Ansätze zur Entwicklung lernender Organisationen* (S. 11–53). Weinheim: Juventa.

Hense, J. & Mandl, H. (2011). *Wissensmanagement und Evaluation*. Zeitschrift für Evaluation 10(2), 267–301.

Joint Committee on Standards for Educational Evaluation, Sanders, J. R. (Hrsg.) (2006), *Handbuch der Evaluationsstandards. Die Standards des „Joint Committee on Standards for Educational Evaluation“*. Aus dem Amerikanischen übersetzt von W. Beywl & Th. Widmer. DeGEval – Gesellschaft für Evaluation e. V. (3., erweiterte und aktualisierte Auflage). Wiesbaden: VS Verlag für Sozialwissenschaften.

Krüger, M. (2010). *Das Lernszenario VideoLern. Selbstgesteuertes und kooperatives Lernen mit Vorlesungsaufzeichnungen. Eine Design-Based-Research Studie*. Dissertation der Universität der Bundeswehr in München, Fakultät für Pädagogik. Neubiberg: Universität der Bundeswehr in München, Fakultät für Pädagogik, Professur für Lehren und Lernen mit Medien. Download am 18.08.2012 von <http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:706-2333>.

Lave, J. & Wenger, E. (1991). *Situated learning. Legitimate peripheral participation*. Cambridge, GBR: Cambridge University Press.

Mertens, D. M. (2000). Institutionalizing Evaluation in the United States of America. In R. Stockmann (Hrsg.), *Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder*. Opladen: Leske + Budrich, 39–56.

National Association for the Education of Young Children (NAEYC) (Hrsg.) (2003). *Early Childhood Curriculum, Assessment, and Program Evaluation. Building an Effective, Accountable System in Programs for Children Birth through Age 8*. NAEYC Position Statement with expanded Resources. Washington, D.C., USA: NAEYC. Download am 18.08.2012 von [www.naeyc.org/files/naeyc/file/positions/CAPEexpand.pdf](http://www.naeyc.org/files/naeyc/file/positions/CAPEexpand.pdf).

National Evaluation of Sure Start (NESS) Team (2012). *The impact of Sure Start Local Programs on seven year olds and their families*. Department for Education Research Report DFE-EE220. London: Department for Education. Download am 18.08.2012 von [www.education.gov.uk/publications/eOrderingDownload/DFE-RR220.pdf](http://www.education.gov.uk/publications/eOrderingDownload/DFE-RR220.pdf).

Patton, M. Q. (2011). *Developmental evaluation. Applying complexity concepts to enhance innovation and use*. New York: Guilford Press.

Patton, M. Q. (2008). *Utilization-focused evaluation*. (4. Auflage, 1. Auflage 1978). Los Angeles: Sage Publications.

Phi Delta Kappa (PDK) National Study Committee on Evaluation, Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Marriman, H. O. & Provus, M. M. (1974). *Educational evaluation and decision making* (4. Druckauflage, 1. Auflage 1971). Itasca, IL, USA: F.E. Peacock Publishers.

Prenzel, A., Heinzel, F. & Carle, U. (2008). Methoden der Handlungs-, Praxis- und Evaluationsforschung. In W. Helsper & J. Böhme (Hrsg.), *Handbuch der Schulforschung* (S. 181–197). (2., durchgesehene und erweiterte Auflage). Wiesbaden: VS Verlag für Sozialwissenschaften.

Reinmann, G. (2005). Innovation ohne Forschung? Ein Plädoyer für den Design-Based Research-Ansatz in der Lehr-Lernforschung. Unterrichtswissenschaft. Zeitschrift für Lernforschung 33(1), 52–69.

Reusser, K. (2011). Von der Unterrichtsforschung zur Unterrichtsentwicklung - Probleme, Strategien, Werkzeuge. In W. Einsiedler (Hrsg.), *Unterrichtsentwicklung und Didaktische Entwicklungsforschung* (S. 11–40). Bad Heilbrunn: Klinkhardt.

- Rossi, P. H., Freeman, H. E. & Wright, S. R. (1979). *Evaluation. A systematic approach*. Beverly Hills: Sage Publications.
- Rossi, P. H., Lipsey, M. W. & Freeman, H. E. (2004). *Evaluation. A systematic approach* (7. Auflage, 1. Auflage 1979). Thousand Oaks, CA, USA: Sage Publications.
- Rossi, P. H. (Hrsg.) (1982). Standards for Evaluation Practice. Special Issue. *New Directions for Program Evaluation* 1982(15), 1-97.
- Scriven, M. (2012). Conceptual Revolutions in Evaluation: Past, Present, and Future. Chapter 12. In M. C. Alkin (Hrsg.), *Evaluation roots. A wider perspective of theorists' views and influences* (S. 167–179) (2. Auflage). Los Angeles, CA: Sage.
- Scriven, M. (2003). Evaluation in the New Millennium. The Transdisciplinary Vision. In Stewart I. Donaldson und Michael Scriven (Hrsg.), *Evaluating social programs and problems. Visions for the new millennium* (S. 11–4). New York: Psychology Press.
- Simon, H. A. (1996). *The sciences of the artificial* (3. Auflage, 1. Auflage 1969). Cambridge, MA, USA: MIT Press.
- Spiess, C. K., Büchel, F. & Wagner, G. G. (2003). Children's school placement in Germany. Does Kindergarten attendance matter? *Early Childhood Research Quarterly* 18(2), 255–270.
- Stamm, M. (2003). *Evaluation und ihre Folgen für die Bildung. Eine unterschätzte pädagogische Herausforderung. Zugleich Habilitationsschrift an der Universität Fribourg, CHE, 2002*. Münster: Waxmann.
- Stark, R. & Mandl, H. (2007). Bridging the gap between basic and applied research by an integrative research approach. *Educational Research and Evaluation* 13(3), 249–261.
- Vohle, F. & Reinmann, G. (2012). Förderung professioneller Unterrichtskompetenz mit digitalen Medien. Lehren lernen durch Videoannotation. In R. Schulz-Zander, B. Eickelmann, H. Moser, H. Niesyto & P. Grell (Hrsg.), *Jahrbuch Medienpädagogik 9*. Wiesbaden: VS Verlag für Sozialwissenschaften, 413-429.
- Weiss, C. H. (1998). *Evaluation. Methods for studying programs and policies* (2. Auflage). Upper Saddle River, N.J., USA: Prentice Hall.
- Widmer, Th., Rathmayr, Ich. & Serdal, U. (1996). *Kurz und gut? Qualität und Effizienz von Kurzevaluationen*. Vorwort von U. Klöti (1. Auflage). Zürich: Rüegger.
- Wottawa, H. & Thierau, H. (2003). *Lehrbuch Evaluation* (3., korrigierte Auflage, 1. Auflage 1990). Bern: Huber.
- Yarbrough, D. B., Shulha, Lyn M., Hopson, R. K. & Caruthers, F. A. / Joint Committee on Standards for Educational Evaluation (JCSEE) (2011). *The program evaluation standards. A guide for evaluators and evaluation users. Joint Committee on Standards for Educational Evaluation (JCSEE)* (3. Auflage, 1. Auflage 1981). Thousand Oaks, CA, USA: Sage Publications.
- Zigler, E. & Styfco, S. J. (2010). *The hidden history of Head Start*. New York: Oxford University Press.