Böttcher, Wolfgang; Hense, Jan

Evaluation im Bildungswesen - eine nicht ganz erfolgreiche Erfolgsgeschichte

Die Deutsche Schule 108 (2016) 2, S. 117-135

SCHULE UND EVALUATION

DDS – Die Deutsche Schule
108. Jahrgang 2016, Heft 2, S. 117-135

Wolfgang Böttcher/Jan Hense

Evaluation im Bildungswesen – eine nicht ganz erfolgreiche Erfolgsgeschichte

Zusammenfassung
Im Beitrag wird das Konzept Evaluation definiert; Typen der Evaluation sowie ihre unverzichtbaren Elemente werden vorgestellt. Evaluation hat sich international bewährt und als generische Methode für forschungsbasierte Entwicklung etabliert. Sie kann zudem die Besonderheiten verschiedener sozialer Handlungsfelder berücksichtigen (zum Beispiel Schule und Bildung) und kann somit zu durch solide Befunde informierten Entscheidungen in Politik und Praxis beitragen. Die Autoren kritisieren, dass der Begriff Evaluation in der deutschen Bildungsdebatte häufig auch für Konzepte benutzt werde, die gewisse Überschneidungen mit Evaluation aufweisen, jedoch das vollständige Konzept nicht abdecken. Dieser unkorrekte Gebrauch sei schädlich für Theorie, Praxis und das Ansehen der Evaluation.
Schlüsselwörter: Definition von Evaluation, Typen und Standards der Evaluation, evidenzbasierte Politik und Praxis, Evaluation in der deutschen Bildungsdebatte

Evaluation in the Educational System – A not quite Successful Story of Success

Summary
This article defines the concept of evaluation and describes types of evaluation with their essential features. Evaluation has succeeded internationally to establish itself as a generic research-based developmental method. It can also reflect the differences between various fields of social action (i.e. school and education) and can thus serve evidence-informed decision-making for policy and practice. The authors criticize that in the German education community the term evaluation frequently is applied to methods, which overlap with evaluation, but do not cover the full concept. This improper use is harmful for the theory, practice, and reputation of evaluation.
Keywords: definition of evaluation, types and standards of evaluation, evidence-informed policy and practice, evaluation in the German education community

Nach gut 40-jähriger Geschichte kann man feststellen, dass Evaluation eine große Karriere im deutschen Schulwesen gemacht hat. Schon 1972 hatte Christoph Wulf den Sammelband „Evaluation" herausgegeben. Er versammelte hier u.a. Beiträge der „Väter", und der Gegenstand der konzeptionellen und empirischen Beiträge von Scriven, Stake, Cronbach oder Stufflebeam ist Pädagogik und Bildung. Der Untertitel des Bandes: „Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen".

Man wird danach eine gewisse Pause in der Debatte konzedieren müssen, aber spätestens mit der Steuerungsidee von Autonomie und der dadurch bedingten Verpflichtung zur Rechenschaftslegung ist das Thema Evaluation wieder virulent. Mit der heutigen Selbstverständlichkeit der Verwendung des Begriffes Evaluation geht aber auch eine gewisse – und womöglich wachsende – Unschärfe einher.

Unser einleitender Beitrag in das Schwerpunktheft soll deshalb eine konzise, aber konturierte Fassung des Begriffs Evaluation liefern. Wir werden beschreiben, was Evaluation ist und welche Implikationen das hat. Wir sorgen dabei für die notwendige Abgrenzung gegenüber anderen Verfahren. Dabei beziehen wir uns einerseits auf den deutschsprachigen Diskurs zur Evaluation von und in Schulen; andererseits greifen wir punktuell auch auf die stärker international und transdisziplinär geprägte Literatur zur Evaluation zurück (vgl. z.B. Stufflebeam/Shinkfield 2007; Stockmann 2006; Rossi/Lipsey/Freeman 2004; Widmer/Beywl/Fabian 2009; Russ-Eft/Preskill 2009).

Im ersten Abschnitt werden wir deshalb einige unverzichtbare Merkmale von Evaluation beschreiben. Im nächsten Schritt werden wir skizzieren, wie erfolgreich Evaluation in den letzten Jahren war, nicht zuletzt, weil sie Antworten auf wichtige Fragen der Qualität von pädagogischem oder sozialpädagogischem Handeln zu geben vermag – jedenfalls dann, wenn sie „professionell" arbeitet. Eine – gerade auch für Pädagogik – grundlegende Frage ist die, ob Evaluation eher eine Aktivität von professionellen Evaluatoren und Evaluatorinnen ist oder ob die pädagogischen Akteure nicht auch selbst – als inhärente Kompetenz ihrer pädagogischen Profession – ihre eigene Arbeit evaluieren sollten und können. Zur Frage der „Selbstevaluation" werden wir uns im dritten Abschnitt positionieren.

Evaluation beschränkt sich nicht selten auf die Messung der Wirkungen von Bildungsmaßnahmen. Gerade für die Identifizierung von Verbesserungsmöglichkeiten ist oft aber ein umfassenderer Blick auf das Bildungsgeschehen notwendig. Hier hilft die Evaluation, das pädagogische Handeln oder den Aufbau pädagogischer Organisationen transparenter zu machen, indem versucht wird, deren „Logik" zu beschreiben. Darum geht es im vierten Abschnitt. Am besten, so denken wir, kann man beschreiben, was Evaluation bedeutet und was ihre spezifischen Ansprüche sind, indem Standards für ihre Güte definiert werden. Wir tun das im fünften Abschnitt.

Abschließend folgten dann, anknüpfend an die Eingangsfrage nach der Spezifik der Evaluation innerhalb von Verfahren der Bewertung, ein Plädoyer für „gute Evaluation“ und die Aufforderung zu konzeptioneller Genauigkeit. Und das ist nicht nur akademisch, sondern auch praktisch relevant.

1. Was ist Evaluation – und was nicht

Wenn man „Evaluation“ beschreiben oder gar definieren will, kann man sich überlegen, ob man das mittels der Anknüpfung an den Alltag tun will. In manch einer Einführung ins Thema findet man ein solches Vorgehen. Dort heißt es dann, dass jede und jeder -zig Male am Tag „evaluiert“, nämlich „wertet“ oder „bewertet“: Das Essen schmeckt gut, das Hemd steht einem nicht, man fährt lieber über die Landstraße nach Hause, weil der Verkehr auf der Autobahn zu dicht ist. Ganz falsch ist das wohl nicht. Aber auch nicht richtig. Und problematisch, denn hiermit werden tendenziell Besonderheiten der „Evaluation“ als spezifische Methode der Bewertung weggewischt. Und so kann man durchaus nachvollziehen, dass auch in der Bildung oder der Sozialen Arbeit die Tendenz besteht, fast alle Verfahren des Wertens und Bewertens als „Evaluation“ zu bezeichnen. Was immer dabei getan wird, mag notwendig und wichtig sein, aber ob es „Evaluation“ ist, das ist noch lange nicht ausgemacht. Es ist ein konstitutives Merkmal pädagogischer Arbeit, ihre Adressaten und Adressatinnen zu bewerten. In der Schule und der Hochschule werden Leistungen diagnostiziert, Noten vergeben, und es wird zwischen „bestanden“ und „nicht bestanden“ unterschieden. In der sozialen Arbeit werden Hilfepläne entwickelt, die selbstverständlich auch auf diagnostischen, also wertenden Verfahren beruhen. Kann man diese Tätigkeiten als „Evaluation“ bezeichnen?

1.1 Erhöhte Selbstständigkeit und Monitoring

Seit den 90er-Jahren des letzten Jahrhunderts hat sich im Bildungswesen das Konzept der Rechenschaftslegung als Antwort auf (vermeintlich) erhöhte Selbstständigkeit der Einrichtungen (vgl. Böttcher 2002) nach und nach durchgesetzt. Mit den internationalen Vergleichsstudien hat sich die Idee der „Aufsicht“ dramatisch gesteigert. Paradigmatisch ist hierfür das – mit einigen Modifikationen – gerade neu aufgelegte KMK-Programm des „Monitoring“ (vgl. KMK 2015). Folgen sind die Verstetigung der internationalen Leistungsmessungen, die Entwicklung von Leistungserwartungen mittels Bildungsstandards, der Einsatz von Vergleichsarbeiten, zentrale Abschlussprüfungen, Akkreditierungen oder Qualitätsanalysen der Schulen. Diese Verfahren werden häufig als externe Evaluation bezeichnet. Auch wurden vielfältige Angebote entwickelt, die Schulen und anderen Bildungseinrichtungen helfen sollen, eine „Selbstevaluation“ durchzuführen. Schließlich gibt es Instrumente für

Schulen, die Schülern und Schülerinnen die Möglichkeit zur Bewertung der Lehrerkompetenz bieten. In Hochschulen sind solche Rückmeldungen zur Lehre in aller Regel in Evaluationsordnungen sogar vorgeschrieben. Das sind erste Indizien dafür, dass Evaluation ein Erfolgsprogramm ist (siehe Abschnitt 2). Aber ist es immer und ohne Probleme gerechtfertigt, hier tatsächlich von Evaluation zu sprechen?

1.2 Information für Entscheidungen

In einer frühen Phase der Evaluation formuliert einer der Gründungsväter:

> „Im allgemeinen bedeutet Evaluation die Gewinnung von Informationen durch formale Mittel wie Kriterien, Messungen und statistische Verfahren mit dem Ziel, eine rationale Grundlage für das Fällen von Urteilen in Entscheidungssituationen zu erhalten" (Stufflebeam 1972, S. 124).

Zwei wesentliche Merkmale sind hiermit markiert: Evaluation ist – erstens – ein wissenschaftsbasiertes empirisches Verfahren; sie definiert die Indikatoren, die die Messung qualifizieren. Insofern also ist Evaluation empirische Sozialwissenschaft. Alle Verfahren, die diesem Anspruch nicht genügen (wollen), wären demnach nicht Evaluation. Das zweite Kriterium stellt auf Handlungsfolgen als Resultat von Evaluation ab. Die Informationen, die empirisch gewonnen werden, sollen Entscheidungen fundieren. Evaluation hilft also programmatisch, Urteile und sich daran anschließende Konsequenzen empirisch stützen zu können.

Urteile in Entscheidungssituationen, die (auch) auf Basis von Empfehlungen aus Evaluationen gefällt werden, können zum Beispiel zur Beendigung oder Weiterführung von Aktivitäten führen oder ihre Verbesserung bewirken. Beides ist typisch für Evaluation. Eine Funktion – und manchmal auch explizites Ziel – von Evaluation ist Kontrolle. Es kann durchaus darum gehen, eine Empfehlung für ein Ja oder Nein auszusprechen. Aber ein gleichberechtigtes, für manche Protagonisten in der aktuellen Debatte das vordringliche Ziel ist es, Informationen für die Verbesserung des Gegenstandes der Evaluation bereitzustellen. So oder so: Es geht hier um praktische Relevanz, um Nützlichkeit als Kernelement von Evaluation. Um einen weiteren Gründungsvater der Evaluation als Zeugen anzuführen, kann gesagt werden, dass in Evaluationen Urteilsdaten und Beschreibungsdaten gleichermaßen wichtig sind (vgl. Stake 1972, S. 98).

1.3 Gegenstände der Evaluation

Die Ergebnisse einer Evaluation liefern demnach Informationen, Daten, Befunde, die von denjenigen, die Einfluss auf die Gestaltung eines Evaluationsgegenstandes ha-

ben, dazu genutzt werden können, begründete Entscheidungen zu treffen. Wir werden zunächst darüber reden, was eigentlich evaluiert werden, was „Gegenstand" von Evaluation sein kann. Die schnelle Antwort heißt: Alles (vgl. Scriven 1991). Mithilfe von Beispielen aus dem Bildungsbereich dürfte dies deutlicher werden: Gegenstände können Organisationen (z.B. Hochschulen), Abteilungen von Organisationen (z.B. der Fachbereich Mathematik eines Gymnasiums), Maßnahmen oder Maßnahmenbündel (z.B. ein Training zur Entwicklung sozialer Kompetenzen von Schülern und Schülerinnen oder alle Maßnahmen einer Schule, die zur Stärkung sozialer Kompetenzen beitragen sollen), Projekte (z.B. ein zeitlich befristetes Experiment mit Blended Learning) oder Produkte (z.B. ein neues Lehrmittel) sein. Obwohl in der Evaluation auch kognitive Leistungsüberprüfungen oder psychometrische Tests eingesetzt werden, dienen sie nicht der Bewertung der einzelnen Personen. Solche Personendiagnosen oder Potenzialanalysen sind typischerweise nicht Gegenstand von Evaluation. Im Prinzip lassen sich zu evaluierende Gegenstände mittels eines analytischen Modells beschreiben, das wir in Abschnitt 4 skizzieren.

Hier soll zunächst noch darauf hingewiesen werden, dass die Rede von „Gegenständen" der Evaluation womöglich ein falsches Bild erzeugen könnte. In der neueren Generation der Evaluation wird ein großer Wert darauf gelegt, dass diejenigen Personen, die der Evaluation Daten liefern und/oder Akteure in den zu evaluierenden Maßnahmen oder Organisationen sind, als Mitwirkende betrachtet werden (vgl. Taut 2008). Diese partizipative Orientierung bezieht sich vor allem darauf, dass die Evaluierenden gemeinsam mit diesen Akteuren festlegen, was genau in der Evaluation getan werden soll. Diese komplexe kommunikative Situation sei kurz am Beispiel einer Maßnahme erläutert: Wenn ein Training zur Sozialkompetenz evaluiert werden soll, dann hat ein Evaluationsteam nicht ein festes Bild davon, wie eine solche Maßnahme strukturiert ist oder mit welchen Zielen agiert wird. Das Team ermittelt mit den Durchführenden der Maßnahme zum Beispiel, welche Lernergebnisse auf Seiten der Adressaten und Adressatinnen des Trainings angestrebt sind. Eine Evaluation wird nicht unabhängig von der spezifischen Maßnahme Lernergebnisse bewerten, selbst wenn dieses aus allgemeiner Sicht (hier: wissenschaftliche Erkenntnisse zur Entwicklung von Sozialkompetenz) gut begründet ist. Freilich kann eine Evaluation auch bewerten, ob das Training gemessen an wissenschaftlichen Erkenntnissen rational ist. Aber eine faire Bewertung der Wirkungen muss berücksichtigen, was diese spezifische Maßnahme bewirken will (vgl. Abschnitt 4). Eine Evaluation ist also keine Bewertung „von der Stange", sondern muss „maßgeschneidert" werden. Diese Gespräche mit Beteiligten sind wesentliches Element von Evaluation und erfüllen eine wichtige Aufgabe: Denn in der Regel führen sie dazu, dass alle an der Evaluation Beteiligten etwas über ihre eigene Arbeit lernen.

1.4 Bereitschaft zum Lernen

Wenn Evaluation sich, womöglich primär, dadurch kennzeichnen lässt, dass sie Entscheidungshilfen bietet, also auf Aktivitäten zielt, die als Ergebnis der Evaluation ergriffen werden, setzt das voraus, dass es Personen gibt, die an Entscheidungen und Veränderung interessiert sind. In aller Regel sind das die Auftraggeber von Evaluationen. Ein Evaluator bzw. ein Evaluationsteam hat also einen Auftraggeber. Um beim obigen Beispiel zu bleiben: Auftraggeber kann zum Beispiel ein Schulministerium sein, das ein Trainingsprogramm „Sozialkompetenz" zum Einsatz in seinen Schulen einkaufen möchte, aber nun Entscheidungshilfe benötigt, welches der auf dem Markt angebotenen Trainings es werden soll. Auch hier ist ein wesentlicher Teil der Evaluation die Diskussion, welche Kriterien und Zielwerte Maßstäbe für die Bewertung sein sollen. Auch kann die Organisation Auftraggeber sein, die das Training anbietet. Ihr wird es wahrscheinlich darum gehen, Wirksamkeit oder Durchführung zu verbessern. Beide Auftraggeber also sind also an Informationen als Basis für klar definierte Entscheidungen interessiert.

Allerdings kann eine Evaluation auch missbraucht werden. So könnte das Ministerium tatsächlich nur daran interessiert sein, Gründe zu finden, Ressourcen zu sparen und Trainings nicht zu erwerben. Der Anbieter des Trainings könnte lediglich daran interessiert sein, die Mitarbeiter und Mitarbeiterinnen zu identifizieren, von denen die Geschäftsführung annimmt, dass sie die Trainings „schlecht" durchführen. Dieses Thema kann hier nicht vertieft werden. Aber klar ist, dass wir in solchen Fällen allenfalls von missbräuchlichen Formen der Evaluation sprechen können (vgl. Christie/Alkin 1999).

1.5 Die Spezifika der Evaluation

Erste Schritte zum Verständnis der Besonderheiten von Evaluation sind hoffentlich gemacht. Hilfreich könnte auch eine Skizze von Verfahren sein, die mehr oder weniger große Überschneidungen aufweisen, aber eben nicht Evaluation sind. Zuallererst: Verfahren, die sich Evaluation nennen, müssen tatsächlich (auch) bewerten. Ohne Bewertung keine Evaluation! Bewertungsverfahren aber, die nicht auf dem Fundament empirischer Sozialforschung stehen, können sich nicht Evaluation nennen. Evaluationsmethoden müssen also valide, reliabel und intersubjektiv überprüfbar sein. Das heißt auch, dass Bewertungskriterien transparent sein müssen. In die Gruppe von „Nicht-Evaluation" fallen häufig Feedbacks, Selbstreflexion oder kollegialer Austausch. Auch Verfahren, die mit vorgefertigten, standardisierten Kriterien bewerten, entsprechen nicht den Ansprüchen einer Evaluation. Wenn also zum Beispiel eine Organisation nachweisen muss, dass sie bestimmte Verfahren, Normen oder Regeln einhält, dann wird es sich um ein Audit oder eine Prozedur des Qualitätsmanagements handeln. Manchmal sind solche Verfahren außerdem wissen-

schaftlich wenig robust. Verfahren, die allein zu Kontrollzwecken eingesetzt werden, sollen auch so heißen: Kontrolle oder Monitoring. Verfahren, die der Erklärung von Zusammenhängen dienen oder allgemeine Hypothesen prüfen, sind ebenfalls nicht Evaluation. Hiermit ist z.B. empirische Bildungsforschung gemeint, die, um beim obigen Beispiel zu bleiben, fragen würde, welche pädagogischen Arrangements besonders geeignet sind, soziale Kompetenzen zu entwickeln. Evaluation hingegen fragt danach, ob dieses spezifische Training diese Kompetenzen fördert. Freilich ist eine gute Evaluation auf genau solche allgemeinen Erkenntnisse angewiesen, wenn sie helfen will, dieses spezifische Training zu verbessern.

2. Eine kurze Skizze der Erfolgsgeschichte der Evaluation in Schule und Bildung

Nach Wahrnehmung vieler Beobachterinnen und Beobachter hat die Evaluation insgesamt in den unterschiedlichsten Handlungsfeldern einen bemerkenswerten Bedeutungszuwachs erfahren. Dass Evaluation nicht nur in Schulen eine wichtige Rolle spielt, sondern auch in vielen anderen Politikbereichen ein etabliertes Steuerungsinstrument geworden ist, belegen etwa ein Blick auf die entsprechenden Arbeitskreise der Gesellschaft für Evaluation (vgl. Böttcher et al. 2014) und der „Dreiländervergleich" zum Stand der Evaluation in Deutschland, Österreich und der Schweiz (vgl. Widmer/Beywl/Fabian 2009).

Im Bildungsbereich kam es in den 1970er-Jahren zu einer ersten „Hochphase" der Evaluation, die im Kontext von Bildungsreform und -expansion sowie als Begleiterscheinung entsprechender Modellversuchsprogramme gesehen werden kann. Nach einer Art „Winterschlaf" in den 1980er-Jahren, in denen Evaluation nur selten Thema war, kam es in den 1990er-Jahren zu einer Renaissance des Themas (vgl. ausführlich Hense 2006, Kap. 3). Diese lässt sich u.a. auf einen gewachsenen Kostendruck, ein allgemein gestiegenes Bewusstsein für die Rentabilität von Bildungsausgaben sowie internationale Einflüsse zurückführen. Zusätzliche Dynamik ergab sich in den 2000er-Jahren dann durch die PISA-Untersuchungen und weitere *Large Scale Assessments*, die den Blick stärker auf die Ergebnisse von Bildungsprozessen lenkten und zum gegenwärtigen Trend der Outcome-Steuerung von Bildung führten.

Evaluationsaktivitäten lassen sich heute auf allen Ebenen des Bildungssystems beobachten (vgl. Maag Merki 2009). Auf Systemebene adressiert sie v.a. Fragen nach den strukturellen Bedingungen von Bildung. Beispiele sind etwa die Frage nach der Leistungsfähigkeit von Gesamtschulen oder den Auswirkungen der Schulzeitverkürzung bei Einführung des G8. Auf Organisationsebene nimmt sie Einzelschulen in den Blick und steht hier im engen Zusammenhang mit den Themen Schulentwicklung und Schulprogrammarbeit (vgl. z.B. Rolff/Buchen 2009). Eine auf

dieser Ebene nach wie vor aktuelle Variante der Evaluation ist die Selbstevaluation (vgl. Abschnitt 3). Sie nimmt auch die dritte Ebene von Evaluationsaktivitäten, die Unterrichtsebene, in den Blick. Hier reicht das Spektrum von den „kleinen", prozessnahen und vor Ort verantworteten Selbstevaluationen bis hin zu den „großen" Schulleistungsuntersuchungen und Lernstanderhebungen, die Wirkungen von Unterricht bei den Schülerinnen und Schülern untersuchen. Letztere werden oft, auch von Maag Merki (2009), als Evaluationsverfahren bezeichnet, obwohl sie mit ihrer reinen Outcome-Orientierung vor allem das Ziel haben, Monitoring-Daten zur Verfügung zu stellen, und ihren „evaluierten" Gegenstand, den schulischen Unterricht, bei der Untersuchung nicht in den Fokus nehmen.

Evaluation hat sich also insgesamt auf verschiedenen Ebenen des Bildungssystems etabliert und wird in der Regel nicht mehr grundsätzlich hinterfragt. Trotz oder gerade wegen der Vielfältigkeit von Evaluationsbemühungen erscheint die „Evaluationslandschaft" derzeit aber äußerst heterogen. Natürlich erfordern unterschiedliche Kontexte und Zielsetzungen auch unterschiedliche Herangehensweisen. Blickt man aber alleine auf den Sprachgebrauch oder alleine auf die oft unklare Abgrenzung gegenüber verwandten Ansätzen der Qualitätsverbesserung von Schule und Unterricht, wird deutlich, dass Evaluation in Schulen immer noch ein relativ junges Tätigkeitsfeld mit nur wenig ausgeprägten Professionalisierungstendenzen ist.

Für das gesamte Feld der Evaluation lassen sich aber, auch in internationaler Perspektive, durchaus gewisse Tendenzen zur Herausbildung einer Evaluationsprofession erkennen (vgl. Böttcher/Hense 2015; Meyer 2015). Als sichtbare Anzeichen dafür lassen sich vor allem nennen: die Gründung von Fachgesellschaften für Evaluation – maßgeblich in Deutschland und Österreich ist die Gesellschaft für Evaluation (DeGEval); thematisch einschlägige Fachzeitschriften wie die Zeitschrift für Evaluation; die Etablierung von universitären Studiengängen wie etwa die Masterprogramme an den Universitäten Saarbrücken oder Bern; und vor allem die Entwicklung von fachlichen Standards guter Evaluation, die wir in Abschnitt 5 genauer vorstellen.

3. Fremd- und Selbstevaluation

Relativ früh in der jüngeren Erfolgsgeschichte der Evaluation im deutschsprachigen Raum haben sich Ansätze zur Selbstevaluation herausgebildet, die teils ergänzend, teils auch als Alternative zu Fremdevaluationen konzipiert sind. Die Wurzeln des Ansatzes liegen in der Sozialen Arbeit (vgl. Heiner 1988), aber auch im schulischen Bereich hat Selbstevaluation seit den späten 1990er-Jahren Verbreitung gefunden (vgl. Hense 2006).

Das kennzeichnende Merkmal von Selbstevaluation ist, dass jene (pädagogischen) Akteure, die für den evaluierten Gegenstand verantwortlich sind, gleichzeitig auch seine Evaluation verantworten (vgl. Altrichter 1999; Buhren 2007; Burkard 1995; Prell 2001). Wesentlich ist also die „Ownership" am Prozess der Evaluation, also die Frage, von wem über wesentliche „Weichenstellungen" einer Evaluation, wie etwa Fragestellungen, Methoden oder Ergebnisverwendung, entschieden wird. Im Falle der Selbstevaluation von Unterricht sind es also die Lehrkräfte, die evaluieren. Bezieht sich die Selbstevaluation auf Aspekte der gesamten Schule wie etwa das Schulprogramm, wird die Selbstevaluation üblicherweise von einer entsprechenden Arbeitsgruppe wahrgenommen, die sie im Auftrag der Schule durchführt, wobei die „Ownership" weiterhin bei der Schule liegt und an diese Gruppe delegiert wird. Auch wenn hier die Evaluation als „Nebentätigkeit" des pädagogischen Kerngeschäfts betrieben wird, gelten die Standards der Evaluation (vgl. Abschnitt 5) auch für den Bereich der Selbstevaluation (vgl. Müller-Kohlenberg/Beywl 2003).

Obwohl im schulischen Kontext der Begriff der Selbstevaluation häufig synonym mit dem der internen Evaluation gebraucht wird (vgl. z.B. Nevo 2001), sind dabei aus Sicht der allgemeineren Evaluationsliteratur die zwei Dimensionen Ort der Steuerung und Ort der Durchführung zu unterscheiden (vgl. Widmer/Rocchi 2012; Scriven 1991). Denn zumindest in größeren Organisationen sind auch interne Fremdevaluationen denkbar, wenn nämlich eine entsprechende Fachabteilung diese Funktion wahrnimmt und andere Teile der Organisation keine „Ownership" an der Evaluation haben. Das Begriffspaar Fremd-/Selbstevaluation bezieht sich also primär auf die Frage der „Ownership" am Evaluationsprozess, das Begriffspaar interne/externe Evaluation dagegen auf die Frage, ob die Evaluierenden inner- oder außerhalb der Organisation verortet sind (vgl. König 2000). Gerade in der schulischen Praxis ist diese analytische Trennung oft allerdings nicht in Reinform vorzunehmen. So kann etwa eine schulische Steuerungsgruppe zur Selbstevaluation in der Wahrnehmung von nicht beteiligten Kolleginnen und Kollegen eher als interne (Fremd-)Evaluation wahrgenommen werden.

Wie kam es zur Entwicklung und auch Verbreitung von Selbstevaluationsansätzen im schulischen Bereich? Verschiedene Einflüsse spielten dabei eine Rolle, die im größeren Kontext der allgemeinen Qualitätsdebatte im Bildungswesen zu betrachten sind (vgl. Hense 2006, Kap. 3; Fend 2000; Kuper 2002). Evaluation und an Outcomes orientierte Steuerungsansätze wurden dort seit den 1990er-Jahren immer wichtiger, wobei die großen Schulvergleichsstudien seit den 2000ern sicherlich noch einmal einen zusätzlichen Schub ausgelöst haben (vgl. Abschnitt 2). In vielen pädagogischen Institutionen wurden (Fremd-)Evaluationen und andere Qualitätssicherungsansätze zunächst aber eher als Fremdkörper und Zumutung wahrgenommen denn als Mittel der Qualitätssicherung und -verbesserung. Selbstevaluation war hierbei teilweise ein emanzipatorischer Ansatz, zumindest als Korrektiv und Ergänzung, die Qualitätsarbeit in die Hände der letztlich verantwortlichen Praktikerinnen und

Praktiker zu legen. Verbreitet hat sie sich vor allem in Verbindung mit Initiativen zur Schul-, Curriculum- und Unterrichtsentwicklung, wo sie häufig als unterstützendes Instrument im Entwicklungszyklus gesehen wurde (vgl. Buhren/Killus/Müller 2000; Radnitzky/Schratz 1999; Moser 1999).

Eine weitere Argumentationslinie, die vor allem aus internationalen Erfahrungen „importiert" wurde, sah Selbstevaluation als natürliches Korrektiv für eine angestrebte wachsende Autonomie pädagogischer Institutionen (vgl. Rürup 2007): Mehr Autonomie impliziere demnach mehr Rechenschaft, die u.a. durch Selbstevaluationen abgelegt werden könne. Schließlich gab es auch evaluationsimmanente Argumente für Selbstevaluationen, die an einer Kritik von Genauigkeit und Nützlichkeit von Fremdevaluationen ansetzten. Demnach seien Fremdevaluationen oft zu weit von der evaluierten Praxis entfernt, um für die Praxis valide, zeitnahe und nützliche Informationen zu generieren. Damit verbunden war die Erwartung, dass Selbstevaluation aufgrund der Personalunion von Evaluatoren/Evaluatorinnen und Praktikern/Praktikerinnen eher zu sichtbaren Konsequenzen führen könne. Eine weitere Erwartung war, dass Lehrkräfte ähnlich wie im Ansatz der „Empowerment Evaluation" (vgl. Fetterman 2001) durch die eigene Anwendung der evaluativen Handlungslogik die erforderlichen Kompetenzen erwerben, um Formen der externen Evaluation selbstbewusster und „auf Augenhöhe" gegenübertreten zu können.

Diesen Erwartungen und Hoffnungen sind natürlich auch kritische Stimmen gegenüberzustellen. Zu nennen sind vor allem Zweifel in Bezug auf die Glaubwürdigkeit und die Machbarkeit von Selbstevaluation. So liegt die Annahme nahe, dass Objektivität als eine unverzichtbare Bedingung für Evaluationen hier aufgrund von einer zu großen Nähe zum Gegenstand („Betriebsblindheit") nicht gegeben sein kann und der inhärente Interessenskonflikt eine zu große Versuchung mit sich bringt, Stärken zu schönen und Schwächen auszublenden (vgl. z.B. Scriven 1997; Döring/Bortz 2016; Müller-Kohlenberg/Beywl 2003). Es ist allerdings anzunehmen, dass je nach Evaluationskontext und -zwecken (z.B. Verbesserung vs. Rechenschaft) diese Probleme in unterschiedlichem Maße virulent werden. In jedem Fall aber stellt sich das Problem der Machbarkeit. Selbstevaluation ist durchaus voraussetzungsreich in Bezug auf die erforderlichen Ressourcen und Kompetenzen sowie weitere förderliche Bedingungen auf personeller und organisationaler Ebene sowie die Rahmenbedingungen (vgl. Hense 2006). Hinderliche Faktoren können etwa fehlende Kritikfähigkeit auf individueller Ebene, eine stark vom traditionellen Autonomie-Paritäts-Muster (vgl. Posch 1999) geprägte Schulkultur oder schlicht fehlende zeitliche Ressourcen bei den Rahmenbedingungen sein. Unabhängig von den jeweils im Einzelnen wirksamen Begründungszusammenhängen und Kritikpunkten etablierten sich vor allem in den 2000er-Jahren vielfältige Modellversuche, Initiativen und Angebote im Bereich der Selbstevaluation. Beispiele reichen von der europäischen Ebene wie dem Socrates-Projekt „Effective School Self-Evaluation" (vgl. SICI 2003) über bundesweite und länderübergreifende Initiativen wie das Instrumentarium

„Selbstevaluation in Schulen" (vgl. Viebahn/Brockhaus 2011) oder „Selbstevaluation für Schulleitungen", einem Bestandteil des BLK-Programms „Demokratie lernen & leben" (vgl. Schroeter/Kohle 2006), bis hin zur unterschiedlichen landesspezifischen Modellen unter Regie der jeweiligen Landesministerien. Sie nutzen dabei häufig etablierte und durch wissenschaftliche Expertise gestützte Instrumente der Selbstevaluation (vgl. z.B. Brägger/Posse 2007). Einen neueren Ansatz, der an der Initiative von Einzelschulen ansetzt, stellen Beywl und Balzer in diesem Heft (vgl. S. 191-204) vor.

4. Programmevaluation und ihre Logik

Wir versuchen nun, das Konzept Evaluation weiter zu schärfen, indem wir auf die Frage zurückkommen, was „Gegenstände" von Evaluationen sein können. Im Prinzip ist alles evaluierbar, z.B. Produkte, Ideen, Konzepte, Projekte, Interventionen, Organisationen (vgl. Scriven 2003). International dominant ist die sogenannte „Programmevaluation". Zwar bezieht sich dieser Begriff insbesondere auf Interventionen oder Bündel von Maßnahmen, aber die zugrunde liegende Denkfigur ist durchaus auch auf Gegenstände wie Produkte (z.B. Lehrbücher) oder Organisationen (z.B. Schulen als Handlungseinheiten) anwendbar. Hilfreich könnte sein, dass man jedem der oben gelisteten Objekte eine bestimmte Handlungslogik unterstellt. In der Bildung und in der Sozialen Arbeit haben wir es mit pädagogischen oder sozialen „Programmen" zu tun, Maßnahmen also, die beabsichtigen, Lernen und Kompetenzentwicklung zu sichern oder zu fördern, Beratung oder Hilfe zu verbessern oder bessere Bedingungen für diese Aktivitäten zu entwickeln. Auch Organisationen oder Produkte lassen sich durchaus mit Hilfe einer solchen Denkfigur beschreiben. Evaluationen bewerten – in diesem Sinne – Programme. Wir sprechen hier von Programmevaluation.

Ein Programm folgt – im Prinzip – einer bestimmten Logik. In der Geschichte der Evaluation gibt es vielfältige Versuche, den Aufbau von Programmen zu modellieren (vgl. z.B. Chen 2013; Funnell/Rogers 2011). Im Kern geht es um eine Beschreibung der „Wirklogik". In der internationalen Evaluationsliteratur hat sich dafür das Instrument „logisches Modell" („logical model") etabliert (vgl. Frechtling 2007; McLaughlin/Jordan 2010; W.K. Kellogg Foundation 2001). Es dient – allgemein gesprochen – der Veranschaulichung und Klärung des Ablaufs eines Programms und der von ihm intendierten Wirkungen. Es zielt vor allem darauf ab, die Realisierbarkeit eines Programms zu garantieren.

Der wohl prominenteste Versuch eines logischen Modells ist das CIPP-Modell von Stufflebeam (1972). Es unterscheidet *Context* (was soll getan werden?), *Input* (wie soll es getan werden?), *Process* (wird getan, was vorgesehen ist?) und *Product* (was sind

die Resultate?). Wir werden es ein wenig ausführen, ohne zu sehr ins Detail gehen zu können. Dazu benennen wir die wesentlichen logischen Schritte knapp und ergänzen sie in den Klammern durch kleine Hinweise auf konkrete Fragestellungen:

1. Das Programm definiert das *Problem*, auf das es reagieren will. Es beschreibt einen zu erreichenden Zustand (Ziele). (Schüler und Schülerinnen in der Schule X verhalten sich aggressiv und arbeiten gegeneinander. Ziel eines Trainings soll es sein, soziale Kompetenzen einer Zielgruppe zu entwickeln. Es wird genau beschrieben, wie der Ist-Zustand charakterisiert wird und woran der Erfolg des Programms gemessen werden soll.)
2. Das Programm muss berücksichtigen, dass es unter bestimmten Bedingungen umgesetzt werden muss. Diese *Kontexte* sind für die Realisierungsmöglichkeiten von großer Bedeutung. (Die aggressiven Schüler und Schülerinnen gelten als Idole. Insgesamt ist das Klima in der Schule gereizt. Ausgliederung in spezifische Programme gilt als Versagen.)
3. Wie sieht das *Konzept* aus, von dem theorie- und evidenzbasiert erwartet werden kann, dass der gewünschte Zustand mittels des Programms erreicht werden kann? (Was weiß man über die Wirkung solcher Trainings? Welche Methoden haben sich als wirksam erwiesen?)
4. Die *Ressourcen* werden bestimmt, die zur Umsetzung des Programms eingestellt werden. (Wie viel Geld steht zur Verfügung? Können Lehrkräfte einbezogen werden, die über einschlägige Erfahrungen verfügen? Kann man auf Kompetenzen der Adressaten zurückgreifen? Wie sehr unterstützen Lehrkräfte das Training?)
5. Sorgfältig müssen die *Prozesse* geplant werden, die für die Umsetzung des Programms nötig sind. (Wie genau kann die Umsetzung aussehen? Welche Zeitgefäße werden benötigt, wie kann die Mitarbeit der Adressaten gesichert werden? Enthält das Programm genaue Handlungsanweisungen?)
6. Was sind die geplanten *Outputs*? Wie viele Trainingsstunden sollen realisiert werden? (Wie viele Schüler und Schülerinnen sollen teilnehmen? An wie vielen Tests soll jeder bzw. jede mitwirken? Wie viele Materialien werden bearbeitet? Wie zufrieden sollen die Teilnehmer und Teilnehmerinnen sein?)
7. Was sind die geplanten *Outcomes*? Welche Wissenszuwächse kennzeichnen einen Erfolg, welcher Einstellungswandel, welche Verhaltensänderungen? Wie stabil sollen diese in zeitlicher Hinsicht sein? *Outcomes* bezeichnen nicht nur die angestrebten Ziele; auch unerwartete oder gar adverse Wirkungen müssen in den Blick der Evaluation kommen. Solche nicht intendierten Effekte sind prinzipiell von gleicher Bedeutung wie die beabsichtigten.

Der Aufbau der logischen Schritte suggeriert die Linearität der Modellierung. Freilich beeinflussen sich diese Elemente des Programms beständig und (leider auch) in kaum kalkulierbarer Art. So ist Programmtreue in der Umsetzung oft unsicher: Die Durchführenden fühlen sich dem Programm nicht wirklich verpflichtet (Stichwort: mangelnde Compliance). Auch sind die Ergebnisse in hohem Maße von der Interaktion zwischen Programmdurchführenden und -adressaten be-

stimmt (Stichwort: Koproduktion). Und um noch einen weiteren Aspekt zu nennen: Gerade im Prozess der Programmumsetzung spielen die möglicherweise unterschiedlichen, manchmal diametralen und oftmals nicht expliziten Interessen der Anspruchsgruppen (Stichwort: Stakeholder) eine bedeutende Rolle und können kaum kontrolliert werden.

Um aber evaluieren zu können, muss ein Evaluationsteam die Programmlogik abbilden können. Häufig haben die Akteure in Bildung und Sozialer Arbeit zwar gute Absichten; ihr Handeln genügt aber nicht diesem Programmaufbau. Evaluatoren und Evaluatorinnen müssen dann die Logik gemeinsam mit den relevanten Stakeholdern (re-)konstruieren. Das kann ein sehr schwieriger kommunikativer Prozess ein.

Im Kern kann diese Handlungslogik auch bei der Evaluation einer Organisation oder eines Produktes eingesetzt werden. Auch eine Organisation (eine Schule zum Beispiel) agiert wie ein Programm; auch ein Produkt (z.B. ein neues Lehrmittel) will ein Problem lösen bzw. ein Ziel verfolgen (z.B. kommunikative Kompetenz der Schüler und Schülerinnen im Sprachunterricht verbessern) und muss die Komponenten der Programmlogik kalkulieren, wenn es erfolgreich sein will.

5. Standards guter Evaluation

Auch ein schlechtes Buch ist immer noch ein Buch. Ähnlich verhält es sich mit der Evaluation: Nicht jede Evaluation ist gleich gut; es gibt gute und schlechte Evaluation und natürlich viele Zwischentöne. Aber woran erkennt man eine gute Evaluation? Diese Frage hat schon früh in der modernen Evaluationsgeschichte zu ausführlichen Selbstreflexionen geführt. Ein wichtiges Ergebnis war die Entwicklung und Verbreitung von Evaluationsstandards (vgl. Stufflebeam 2000).

Für Deutschland maßgeblich sind die Standards für Evaluation der Gesellschaft für Evaluation, die 2002 erstmals aufgelegt wurden (vgl. DeGEval 2002) und 2016 in einer revidierten Fassung erscheinen werden. Sie basieren auf den Vorarbeiten des Joint Comittee on Standards for Educational Evaluation (Joint Committee on Standards for Educational Evaluation/Sanders 2006) und definieren vier zentrale Merkmale, die gute Evaluationen auszeichnen: Nützlichkeit, Durchführbarkeit, Fairness und Genauigkeit. Die Nützlichkeitsstandards fordern, dass Evaluation sich immer daran messen lassen muss, inwieweit sie ihre intendierten Zwecke erfüllt und einen tatsächlichen Nutzen darstellt. Die Durchführbarkeitsstandards sollen garantieren, dass Evaluationsverfahren realistisch, kostenbewusst und diplomatisch geplant und durchgeführt werden. In der Gruppe der Fairnessstandards finden sich grundlegende Forderungen zu Aspekten wie Transparenz, Schutz von Persönlichkeitsrechten, Ganzheitlichkeit der Betrachtung und eine unparteiische Rolle. Schließlich umfas-

sen die Genauigkeitsstandards Forderungen, wie sie insgesamt an die sozialwissenschaftliche Forschung gestellt werden, aber auch den Imperativ der Metaevaluation, also die Forderung, dass auch Evaluation sich kritischen Fragen nach ihrer Güte oder Nützlichkeit stellen lassen muss. Konkret mit Leben gefüllt werden diese vier Standardbereiche durch eine jeweils unterschiedliche Anzahl von insgesamt 25 Einzelstandards.

So lautet etwa der Nützlichkeitsstandard „N7 Rechtzeitigkeit der Evaluation": „Evaluationsvorhaben sollen so rechtzeitig begonnen und abgeschlossen werden, dass ihre Ergebnisse in anstehende Entscheidungsprozesse bzw. Verbesserungsprozesse einfließen können." (DeGEval 2002, S. 9) Die publizierte Fassung der Standards für Evaluation enthält neben diesen Standards im eigentlichen Sinne jeweils noch Begründungen und Umsetzungshinweise sowie weitere Begleitmaterialien. Die Neuauflage 2016 wird zusätzlich u.a. ein Glossar zur Vereinheitlichung der Begrifflichkeiten enthalten.

Was ist der Nutzen von Standards? Die Standards für Evaluation sollen einerseits als Orientierung bei der Gestaltung von Evaluationen handlungsleitend sein. Sie lassen sich also als konkrete Leitlinien interpretieren, die in der Praxis berücksichtigt werden müssen. Wichtig ist dabei, dass sie sich nicht nur an die Evaluierenden richten, sondern an alle, die für eine Evaluation ganz oder teilweise Verantwortung tragen, wie z.B. auch jene, die Evaluationen in Auftrag geben. Andererseits stellen die Standards Kriterien für die Beurteilung der Qualität von Evaluationen zur Verfügung. Denn auch Evaluationen können mit Hilfe der Standards evaluiert werden. Für solche Evaluationen von Evaluationen hat sich der Begriff der Meta-Evaluation etabliert hat (vgl. Caspari 2015). Die Standards können also auch im Sinne von Bewertungskriterien verstanden werden, die bei der Beantwortung der Frage helfen, wie gut eine Evaluation oder ein Evaluationssystem ist.

Die Standards entstammen ursprünglich dem Bildungsbereich, sind aber universell für alle Evaluationsverfahren und -einsatzgebiete gültig. Auch im Bereich Schule sollten gute Evaluationen und Evaluationssysteme also nützlich, durchführbar, fair und genau sein. Es wäre sicherlich wert zu untersuchen, inwiefern die unterschiedlichen in Schulen etablierten Evaluationsverfahren, von der Notengebung der Lehrkräfte über schulische Selbstevaluationsverfahren bis hin zu Schulaufsicht oder Schulleistungsstudien, diesen Kriterien gerecht werden.

6. Einige Probleme der „Evaluation“ im Bildungswesen

Wir haben über Evaluation gesprochen und ihre Besonderheiten erläutert. Wir haben auch angesprochen, dass sie in Konkurrenz zu anderen Verfahren steht, die Maßnahmen oder Organisationen beschreiben, begleiten oder bewerten. Diese anderen Verfahren, die Überschneidungen mit Evaluation aufweisen, haben allesamt ihre Berechtigung: Aber sie sind nicht Evaluation. Die Abgrenzung von Evaluation zu anderen – nicht gänzlich unähnlichen Verfahren – ist nicht (nur) akademisch, sondern sie ist praktisch relevant (vgl. Widmer/Rocchi 2012).

Um zu wissen, was im Bildungswesen vor sich geht, benötigen bildungspolitische Entscheider und Entscheiderinnen sowie Praktiker und Praktikerinnen Daten. Daten aus Schulinspektionen, aus Vergleichsarbeiten, Adressatenbefragungen oder aus Bildungsberichten sind elementar wichtig. Aber im Bildungswesen geht es um mehr als das Datensammeln mittels Tests, Feedbacks, Audits, Monitoring und Statistik: Beschreiben und Bewerten gehören zur Evaluation wie die Orientierung an Nützlichkeit.

Es geht der Evaluation also immer auch um Nützlichkeit. Ihr Anspruch ist es, Informationen für die Verbesserung von Programmen oder Organisationen zu liefern. Die Schwächen der deutschen Bildungslandschaft sind durch Forschung und Berichte aus der Praxis gut belegt: Die Ressourcen sind in vielen Bereichen knapp; vorhandene Mittel werden ineffizient eingesetzt; eine zu große Gruppe bleibt ohne Bildungserfolge; Lehrkräfte verzweifeln an überbordenden Aufgaben; die Kompetenzen, die in der Lehrerbildung erworben werden, reichen nicht aus, die komplexen pädagogischen Aufgaben zu lösen; die formale Bildung kann benachteiligende Effekte der Herkunft der Kinder nicht kompensieren. Die Forschungen über die Verfahren des Bildungsmonitoring zeigen, dass, trotz hohen Aufwandes, Impulse für Reformen weitgehend ausblieben (vgl. Böttcher 2013). Weitere negative Diagnosen verstärken allenfalls den Druck auf die Bildungsarbeiter und -arbeiterinnen, die dringend Hilfe benötigen. Weitere Berichte über Defizite verführen die Politik dazu, weitere Ansprüche zu formulieren: nicht an sich selbst, sondern an die pädagogischen Akteure. Die Regierenden und Verwaltenden haben wenige konkrete Ideen, wie die Arbeit vor Ort besser gemacht werden könnte. Gute Evaluationen könnten ein erster Schritt sein, einschlägige Information zu liefern. Man muss sie aber hören wollen und Handlungsempfehlungen ernst nehmen – auch wenn sie womöglich mehr Geld kosten als das Monitoring.

Die überarbeitete Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring (vgl. KMK 2015) umfasst vier Maßnahmen, die wiederum vordringlich der Deskription dienen: Teilnahme an internationalen Schulleistungsstudien, Überprüfung und Umsetzung von Bildungsstandards, Verfahren zur Qualitätssicherung auf

Ebene der Schulen und Bildungsberichterstattung. Stärker als in der Ursprungsversion sollen hier aber auch „die Voraussetzungen verbessert werden, Entwicklungen nicht nur zu beschreiben, sondern auch zu erklären und dies mit Hinweisen zu verbinden, wie die festgestellten Probleme gelöst werden können" (S. 6). Ob dieses Versprechen nunmehr in den kommenden Jahren eingelöst wird, bleibt abzuwarten. Dauerhafte, aber uneingelöste Reformversprechen werden die Frustration vergrößern, die man in den Schulen (und auch anderen Bildungsorganisationen) beobachten kann. Vielleicht ist es ein zweitrangiges Problem, wenn sich die sich einer professionellen Evaluation verpflichtenden Akteure auch darüber sorgen, dass Evaluation nunmehr als Titel für jedwedes Beobachtungs- und Messverfahren herhalten muss, ohne Evaluation zu sein. Aber auch einem gelernten Koch kann es nicht gleichgültig sein, dass jeder, der vor einem Herd steht, sich Koch nennen darf.

Die in diesem einleitenden Beitrag vorgenommene Beschreibung der Programmatik „guter Evaluation" sollte aber nicht schließen, ohne wenigstens darauf hinzuweisen, dass die reale Tätigkeit von Evaluatoren und Evaluatorinnen oft vom hier entworfenen idealen Bild abweicht. Die Diskrepanz zwischen Idee und Wirklichkeit zu beleuchten, müsste Thema eines anderen Beitrags sein.

Literatur und Internetquelle

Altrichter, H. (1999): Selbstevaluation. Alle reden davon, wer macht sie? In: Rösner, E. (Hrsg.): Schulentwicklung und Schulqualität. Dortmund: IFS, S. 259-281.

Beywl, W./Balzer, L. (2016): Aufbau von Evaluationskompetenzen für interne Schulevaluation durch projektbezogene Fortbildung. In: Die Deutsche Schule 108, H. 2, S. 191-204.

Böttcher, W. (2002): Kann eine ökonomische Schule auch eine pädagogische sein? Schulentwicklung zwischen Neuer Steuerung, Organisation, Leistungsevaluation und Bildung. München/Weinheim: Juventa.

Böttcher, W. (2013): Das Monitoring-Paradigma – Eine Kritik der deutschen Schulreform. In: Empirische Pädagogik 27, H. 4, S. 497-509.

Böttcher, W./Hense, J. (2015): Professionelle Evaluation oder Evaluation als Profession? In: Hennefeld, V./Meyer, W./Silvestrini, S. (Hrsg.): Nachhaltige Evaluation? Auftragsforschung zwischen Praxis und Wissenschaft. Münster u.a.: Waxmann, S. 101-120.

Böttcher, W./Kerlen, C./Maats, P./Schwab, O./Sheikh, S. (Hrsg.) (2014): Evaluation in Deutschland und Österreich. Stand und Entwicklungsperspektiven in den Arbeitsfeldern der DeGEval – Gesellschaft für Evaluation. Münster u.a.: Waxmann.

Brägger, G./Posse, N. (2007): Instrumente für die Qualitätsentwicklung und Evaluation in Schulen (IQES). Wie Schulen durch eine integrierte Qualitäts- und Gesundheitsförderung besser werden können. Hrsg.: Landesprogramme Bildung und Gesundheit Nordrhein-Westfalen, Hessen und Schweiz. Bern: hep.

Buhren, C. (2007): Selbstevaluation in Schule und Unterricht. Ein Leitfaden für Lehrkräfte und Schulleitungen. Köln: LinkLuchterhand.

Buhren, C.-G./Killus, D./Müller, S. (2000): Implementation und Wirkung von Selbstevaluation in Schulen. In: Rolff, H.-G./Bos, W./Klemm, K./Pfeiffer, H./Schulz-Zander, R. (Hrsg.): Jahrbuch der Schulentwicklung, Bd. 11. Weinheim: Juventa, S. 327-364.

Burkard, C. (1995): Selbstevaluation. Ein Beitrag zur Qualitätsentwicklung von Einzelschulen? Bönen: Verlag für Schule und Weiterbildung, Kettler.

Caspari, A. (2015): Well done? Who knows … Ein Plädoyer für Meta-Evaluationen. In: Hennefeld, V./Meyer, W./Silvestrini, S. (Hrsg.): Nachhaltige Evaluation? Auftragsforschung zwischen Praxis und Wissenschaft. Münster u.a.: Waxmann, S. 143-166.

Chen, H.-T. ([2]2013): Theory-driven Evaluation: Current Views and Origins. In: Alkin, M.C. (Hrsg.): Evaluation Roots. A Wider Perspective of Theorists' Views and Influences. Los Angeles, CA: Sage, S. 132-152.

Christie, C.A./Alkin, M.C. (1999): Further Reflections on Evaluation Misutilization. In: Studies in Educational Evaluation 25, H. 1, S. 1-10.

DeGEval – Gesellschaft für Evaluation (2002): Standards für Evaluation. Köln: Geschäftsstelle DeGEval.

Döring, N./Bortz, J. ([5]2016): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Berlin: Springer.

Fend, H. (2000): Qualität und Qualitätssicherung im Bildungswesen. In: Helmke, A./Hornstein, W./Terhart, E. (Hrsg.): Qualität und Qualitätssicherung im Bildungsbereich: Schule, Sozialpädagogik, Hochschule. Zeitschrift für Pädagogik, 41. Beiheft. Weinheim: Beltz, S. 55-72.

Fetterman, D.M. (2001): Foundations of Empowerment Evaluation. Thousand Oaks, CA/London: Sage.

Frechtling, J.A. (2007): Logic Modeling Methods in Program Evaluation. San Francisco, CA: Jossey-Bass.

Funnell, S.C./Rogers, P.J. (2011): Purposeful Program Theory. Effective Use of Theories of Change and Logic Models. San Francisco, CA: Jossey-Bass.

Heiner, M. (Hrsg.) (1988): Selbstevaluation in der sozialen Arbeit. Freiburg i.Br.: Lambertus.

Hense, J.U. (2006): Selbstevaluation. Erfolgsfaktoren und Wirkungen eines Ansatzes zur selbstbestimmten Qualitätsentwicklung im schulischen Bereich. Frankfurt a.M.: Lang.

Joint Committee on Standards for Educational Evaluation/Sanders, J.R. ([3]2006): Handbuch der Evaluationsstandards. Die Standards des „Joint Committee on Standards for Educational Evaluation". Übersetzt und für die deutsche Ausgabe erweitert von Wolfgang Beywl und Thomas Widmer. Wiesbaden: VS.

KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland) (2015): Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring (Beschluss der 350. Kultusministerkonferenz vom 11.06.2015). Berlin: KMK.

König, J. (2000): Einführung in die Selbstevaluation. Freiburg i.Br.: Lambertus.

Kuper, H. (2002): Stichwort: Qualität im Bildungswesen. In: Zeitschrift für Erziehungswissenschaft 4, S. 533-511.

Maag Merki, K. (2009): Evaluation im Bildungsbereich Schule in Deutschland. In: Widmer, T./Beywl, W./Fabian, C. (Hrsg.): Evaluation. Ein systematisches Handbuch. Wiesbaden: VS, S. 157-162.

McLaughlin, J.A./Jordan, G.B. (2010): Using Logic Models. In: Wholey, J.S./Hatry, H.P./Newcomer, K.E. (Hrsg.): Handbook of Practical Program Evaluation. San Francisco, CA: Jossey-Bass, S. 55-80.

Meyer, W. (2015): Professionalisierung von Evaluation: ein globaler Blick. In: Zeitschrift für Evaluation 14, H. 2, S. 215-246.

Moser, H. (1999): Selbstevaluation und Schulentwicklung. In: PÄD Forum: unterrichten erziehen 3, S. 206-210.

Müller-Kohlenberg, H./Beywl, W. (2003): Standards der Selbstevaluation. In: Zeitschrift für Evaluation 2, S. 79-93.

Nevo, D. (2001): School Evaluation: Internal or External? In: Studies in Educational Evaluation 27, S. 95-106.
Posch, P. (1999): Interne Evaluation. In: Thonhauser, J./Patry, J.L. (Hrsg.): Evaluation im Bildungsbereich. Wissenschaft und Praxis im Dialog. Innsbruck: Studienverlag, S. 139-152.
Prell, S. (2001): Evaluation und Selbstevaluation in pädagogischen Feldern. In: Roth, L. (Hrsg.): Pädagogik. Ein Handbuch für Studium und Praxis. München: Ehrenwirth, S. 991-1003.
Radnitzky, E./Schratz, M. (Hrsg.) (1999): Der Blick in den Spiegel. Texte zur Praxis von Selbstevaluation und Schulentwicklung. Innsbruck: Studienverlag.
Rolff, H.-G./Buchen, H. (2009): Schulentwicklung, Schulprogramm und Steuergruppe. In: Dies. (Hrsg.): Professionswissen Schulleitung. Weinheim: Beltz, S. 296-364.
Rossi, P.H./Lipsey, M.W./Freeman, H.E. ([7]2004): Evaluation. A Systematic Approach. Thousand Oaks, CA, u.a.: Sage.
Rürup, M. (2007): Innovationswege im deutschen Bildungssystem. Die Verbreitung der Idee „Schulautonomie" im Ländervergleich. Wiesbaden: VS.
Russ-Eft, D.F./Preskill, H.S. ([2]2009): Evaluation in Organizations. A Systematic Approach to Enhancing Learning, Performance, and Change. New York: Basic Books.
Schroeter, K./Kohle, V. (2006): Selbstevaluation für Schulleitungen. Berlin: BLK.
Scriven, M. ([4]1991): Evaluation Thesaurus. Thousand Oaks, CA: Sage.
Scriven, M. (1997): Truth and Objectivity in Evaluation. In: Chelimsky, E./Shadish, W.R. (Hrsg.): Evaluation for the 21st Century. A Handbook. Thousand Oaks, CA: Sage, S. 477-500.
Scriven, M. ([4]2003): Evaluation Thesaurus. Newbury Park, CA, u.a.: Sage.
SICI (The Standing International Conference of Central and General Inspectorates of Education) (2003): Effective School Self-Evaluation. Project Report. URL: http://www.edubcn.cat/rcs_gene/extra/05_pla_de_formacio/direccions/primaria/bloc1/1_avaluacio/plugin-essereport.pdf; Zugriffsdatum: 11.04.2016.
Stake, R.E. (1972): Verschiedene Aspekte pädagogischer Evaluation. In: Wulf, C. (Hrsg.): Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München: Piper, S. 92-112.
Stockmann, R. (Hrsg.) ([3]2006): Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder. Münster u.a.: Waxmann.
Stufflebeam, D.L. (1972): Evaluation als Entscheidungshilfe. In: Wulf, C. (Hrsg.): Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München: Piper, S. 113-145.
Stufflebeam, D.L. (2000): Professional Standards and Principles for Evaluations. In: Stufflebeam, D.L./Madaus, G.F./Kellaghan, T. (Hrsg.): Evaluation Models. Viewpoints on Educational and Human Services Evaluation. Boston, MA: Kluwer, S. 440-454.
Stufflebeam, D.L./Shinkfield, A.J. (2007): Evaluation Theory, Models, and Applications. San Francisco, CA: Jossey-Bass.
Taut, S. (2008): What Have We Learned about Stakeholder Involvement in Program Evaluation? In: Studies in Educational Evaluation 34, H. 4, S. 224-230.
Viebahn, C. von/Brockhaus, U. (2011): Selbstevaluation in Schulen (SEIS). Bewährtes und Neues bei der Befragung. In: Schule NRW, H. 11, S. 594-596.
Widmer, T./Beywl, W./Fabian, C. (Hrsg.) (2009): Evaluation. Ein systematisches Handbuch. Wiesbaden: VS.
Widmer, T./Rocchi, T. de (2012): Evaluation. Grundlagen, Ansätze und Anwendungen. Zürich/Chur: Rüegger.

W.K. Kellogg Foundation (2001): Logic Model Development Guide. Using Logic Models to Bring Together Planning, Evaluation, & Action. Battle Creek, MI: W.K. Kellogg Foundation.

Wulf, C. (Hrsg.) (1972): Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München: Piper.

Wolfgang Böttcher, Prof. Dr. rer. pol., geb. 1953, Professor für Erziehungswissenschaft mit den Schwerpunkten Qualitätsentwicklung und Evaluation in Einrichtungen des Bildungs- und Sozialwesens an der Westfälischen Wilhelms-Universität Münster.
Anschrift: Westfälische Wilhelms-Universität, Institut für Erziehungswissenschaft, Georgskommende 33, 48143 Münster
E-Mail: wolfgang.boettcher@uni-muenster.de

Jan Hense, Prof. Dr., geb. 1970, Professor für Hochschuldidaktik und Evaluation an der Justus-Liebig-Universität Gießen.
Anschrift: Justus-Liebig-Universität Gießen, Otto-Behaghel-Str. 10F, 35394 Gießen
E-Mail: jan.hense@psychol.uni-giessen.de