108. Jahrgang, Heft 2, 2016

Die Deutsche Schule

Zeitschrift für Erziehungswissenschaft Bildungspolitik und pädagogische Praxis

DDS

Herausgegeben von der Gewerkschaft Erziehung und Wissenschaft

2

2016

Schule und Evaluation

Berichte zum Schwerpunktthema

WAXMANN

Die Deutsche Schule
Zeitschrift für Erziehungswissenschaft, Bildungspolitik und pädagogische Praxis

Herausgeber: Gewerkschaft Erziehung und Wissenschaft im DGB
in Zusammenarbeit mit der Max-Traeger-Stiftung

Beitragseinreichung und Double-blind Peer Review: Manuskripte (nur Originalbeiträge) werden als Word-Datei an die Geschäftsführung (redaktion@dds-home.de) erbeten. Bitte beachten Sie die Hinweise zur Manuskriptgestaltung (www.dds-home.de). Seit dem 103. Jahrgang (2011) durchlaufen alle Fachartikel in der DDS (Texte zum Themenschwerpunkt und für die Rubrik „Weitere Beiträge") ein externes Review-Verfahren. Nach einer redaktionellen Prüfung der eingereichten Aufsätze im Hinblick auf ihre grundsätzliche Eignung für die DDS schließt sich eine Begutachtung im Doppelblindverfahren durch ehrenamtlich tätige Gutachter/innen an.

Die Deutsche Schule erscheint vierteljährlich (Februar/Mai/August/November). Zusätzlich zu den vier Heften pro Jahrgang können Beihefte erscheinen, die den Abonnenten außerhalb des Abonnements zu einem ermäßigten Preis mit Rückgaberecht geliefert werden. Unter www.waxmann.com und www.dds-home.de finden Sie weitere Informationen.

Preise und Bezugsbedingungen: Jahresabonnement 53,00 €, für GEW-Mitglieder/Studierende 43,00 €, inkl. Online-Zugang für Privatpersonen. Campuslizenz auf Anfrage. Die Preise verstehen sich zzgl. Versandkosten. Ein Einzelheft kostet 16,50 € inkl. Versandkosten. Abbestellungen spätestens 6 Wochen vor Ablauf des Jahresabonnements.

ISSN 0012-0731

Steinfurter Straße 555, 48159 Münster, Telefon: 02 51/2 65 04 0, Fax: 02 51/2 65 04 26,
Internet: www.waxmann.com, E-Mail: info@waxmann.com

Anzeigenverwaltung: Waxmann Verlag GmbH, Martina Kaluza: kaluza@waxmann.com
Druck: Buschmann GmbH, Münster
Satz: Stoddart Satz- und Layoutservice, Münster

DDS – Die Deutsche Schule
Zeitschrift für Erziehungswissenschaft,
Bildungspolitik und pädagogische Praxis
108. Jahrgang 2016 / Heft 2

INHALT

Die Deutsche Schule

Zeitschrift für Erziehungswissenschaft Bildungspolitik und pädagogische Praxis

DDS

Herausgegeben von der Gewerkschaft Erziehung und Wissenschaft

3
2016

Vorschau

Themenschwerpunkt: Querschnittsaufgaben von Schule

„Querschnittsaufgabe“ ist derzeit eine Art Zauberwort, auch wenn Querschnittsaufgaben in der Sache nichts Neues sind. Der Schule sind stets Aufgaben zugeschrieben worden, die sich nicht ausschließlich einzelnen Fächern zuordnen lassen, sondern die über ein Fach hinausweisen und darüber hinaus für die Gestaltung des Schullebens insgesamt relevant sind. Schulhistorisch war der Auftrag der Nationalerziehung eine der wirksamsten Querschnittsaufgaben, und verschiedene der aktuellen können u.a. als Korrektur derselben verstanden werden, so z.B. interkulturelle Erziehung oder sprachliche Bildung im Kontext von Mehrsprachigkeit oder Demokratieerziehung (in der Migrationsgesellschaft).

Eine Sichtung der aktuellen bildungspolitischen Dokumente hat ergeben, dass derzeit mehr als zehn, teilweise einander überschneidende Querschnittaufgaben diskutiert und z.T. auch implementiert werden. Wie sich dies in Lehrplänen der Bundesländer abbildet, wird in einem Bericht vorgestellt; in einem weiteren geht es um die Frage, wie überfachliche Ziele im Bildungsraum der Schule gefördert werden können.

Für den Schwerpunktteil des Heftes sind vier Querschnittsaufgaben ausgewählt worden: Demokratieerziehung, Gewaltprävention, Gesundheitsförderung und Nachhaltigkeit. Es sind vier Aufgaben, die z.T. eng miteinander zusammenhängen und im Kontext der Schulentwicklung auch zusammengeführt werden müssen. Inhaltlich kann auf Fachwissen zurückgegriffen werden; es kann gezeigt werden, dass und wie Wissen und Kompetenzen aus dem Unterricht in einem Fach für andere Fächer relevant sind und wie dieses Wissen und diese Kompetenzen dazu beitragen können, die Schule als Institution – unter Beteiligung der Lernenden und Lehrenden – weiterzuentwickeln.

Heft 3/2016 erscheint im August 2016.

Waxmann • Steinfurter Str. 555 • 48159 Münster • www.waxmann.com

DDS – Die Deutsche Schule
Zeitschrift für Erziehungswissenschaft,
Bildungspolitik und pädagogische Praxis
108. Jahrgang 2016 / Heft 2

CONTENTS

Die Deutsche Schule

Zeitschrift für Erziehungswissenschaft Bildungspolitik und pädagogische Praxis

DDS

Herausgegeben von der Gewerkschaft Erziehung und Wissenschaft

3
2016

Preview

Focus Topic: Cross-sectional Tasks of Schools

Currently, "cross-sectional task" is kind of a magic word, although this type of task is not new. Tasks that do not refer to particular subjects but transcend them and that are relevant for school life as a whole have always been allocated to schools. Regarding school history, the duty of national education was one of the most effective tasks, and several topical ones may be understood as kind of a revision, e.g. intercultural education, or linguistic education in the context of multilingualism, or education in democracy (in a migration society).

A screening of topical education-political documents has resulted in the depiction of more than ten cross-sectional, partly overlapping tasks, which are currently discussed and are implemented to some extent. It will be shown in a report how this situation is reflected in the curricula of the various Federal States; another report deals with the question how cross-sectional aims can be promoted in the educational field of the school.

Four cross-sectional tasks have been chosen for the focus topic: education in democracy, prevention of violence, health promotion, and education for sustainable development. These four tasks are partly interwoven and must be merged in the context of school development. With regard to content, it is possible to refer to specialized knowledge; it can be shown that knowledge and competencies from one school subject are relevant for other ones, and how this knowledge and these competences can make a contribution to enhance school as an institution, involving both learners and teachers.

Issue 3/2016 will be out in August 2016.

Waxmann • Steinfurter Str. 555 • 48159 Münster • www.waxmann.com

EDITORIAL

Editorial zum Schwerpunktthema: Schule und Evaluation

Editorial to the Focus Topic: School and Evaluation

Manchmal erweist es sich als besonders schwierig, sich über Selbstverständlichkeiten zu verständigen. In Bezug auf das Thema „Schule und Evaluation" scheint gerade dies zuzutreffen: Vermittelt über die unterschiedlichsten Reforminstrumente, ausgehend von Schulprogrammarbeit über Schulinspektion bis hin zu Vergleichsarbeiten auf der Systemebene, hat ein „evaluationsorientiertes Denken" in der Schulorganisation in den letzten Jahren eine solche Bedeutung gewonnen, dass es fast schon naiv anmutet, die schlichte Frage zu stellen, was denn eigentlich Evaluation sei.

Während diese Frage vor zwanzig Jahren im Schulkontext noch als legitim gegolten hätte, erweckt der oder die Fragende heutzutage beim Gegenüber den Eindruck, dass die letzten zwei Dekaden Schulentwicklungsdiskurs an ihm oder ihr vorübergegangen sein müssen. Gleichwohl – so unsere Überlegungen zum vorliegenden Themenschwerpunkt – erscheint uns die Klärung oder Reflexion dieser Frage nicht nur legitim, sondern notwendig.

Womöglich verhält es sich mit dem Evaluationsbegriff und unserem Denken über Evaluation nicht sehr viel anders als mit dem Qualitätsbegriff: Für diesen gilt in besonderem Maße, dass er nur als relationaler Begriff überhaupt Sinn macht; zugleich wird er aber von vielen Akteuren mit einer so großen Selbstverständlichkeit verwendet, dass vollkommen aus dem Blick gerät, dass eigentlich in gar keiner Weise geklärt ist, was es denn mit dem Begriff der „Qualität von Schule" auf sich hat. Möglicherweise ist es auch nicht zufällig, dass dieses Schicksal auf beide „Modebegriffe", den der „Evaluation" und den der „Qualität", zutrifft, die aus dem gemeinsamen semantischen Feld stammen, das insgesamt die Schulentwicklungsdiskussion in den letzten Jahren so stark geprägt hat.

Dies aber ist genau der entscheidende Punkt, denn die Tatsache, dass sich auf den verschiedenen Ebenen der Handlungskoordination im Mehrebenensystem „Schule" eine hohe Ausdifferenzierung der Praxen und Praktiken der Evaluation und Qualitätssicherung findet und diese kaum noch infrage gestellt werden, erzeugt eine Pfadabhängigkeit im Denken und Handeln, die es kritisch zu reflektieren gilt.

Ein aktuelles Beispiel für den Hochschulbereich ist das Urteil des Bundesverfassungsgerichts vom 18.03.2016, demzufolge die Akkreditierung von Studiengängen in Nordrhein-Westfalen durch externe Qualitätssicherungsagenturen in der bislang praktizierten Form für unzureichend erklärt wurde. Begreift man dies, d.h. die Akkreditierung von Studiengängen, als ein zentrales Merkmal der universitären Qualitätssicherung und die Begehungen durch die Peers als Evaluationsmoment, dann wird deutlich, wie viel Ungeklärtes in unserem Denken über Evaluation mitschwingt – bis hin zu der besorgten Nachfrage, wer denn nunmehr an den Universitäten für die Qualitätssicherung zuständig bzw. wie nach einer solch harschen Kritik Qualitätssicherung neu zu denken sei – und ob es hier durch externe Agenturen tatsächlich notwendig zu einer Qualitätssteigerung kommt. Aus dem Blick geraten dabei beispielsweise die Praktiken der Qualitätssicherung durch die Wissenschaftlerinnen und Wissenschaftler selbst, die in den Jahren vor der Etablierung der Akkreditierungsagenturen gepflegt wurden und nunmehr durch den vermehrten Ruf nach Systemakkreditierung wieder aufleben. Das Forschen und Lehren an deutschen Universitäten wurde ja nicht erst durch Bologna qualitätsvoll. Überträgt man diese Irritationen aus dem Hochschulsektor auf das Schulsystem, so wird deutlich, dass wir uns immer wieder einmal zentraler Grundbegriffe der Schulentwicklung kritisch-reflexiv vergewissern müssen, um hier nicht blind zu werden, wo es doch gerade eine der entscheidenden Aufgaben von Evaluationen selbst sein sollte, blinde Flecken aufzudecken.

Im einführenden Beitrag von *Wolfgang Böttcher* und *Jan Hense* wird dementsprechend die Erfolgsgeschichte der Evaluation kritisch in den Blick genommen, um sich zentrale Elemente wie die Bedeutung formativer und summativer Evaluation oder auch die Implikationen der Differenzen zwischen Selbst- und Fremdevaluation zu vergegenwärtigen sowie die Bedeutung von Evaluation im Feld und die damit notwendig werdende Qualitätssicherung für Evaluation selbst, die in den letzten Jahren in der öffentlichen Diskussion über Standards der Evaluation ihren Niederschlag gefunden hat.

Welche unterschiedlichen Vorstellungen von Evaluation von den verschiedenen Akteuren im Feld vertreten werden, dokumentiert der Beitrag von *Jutta Wolff*. Zugleich konzentriert sie sich hierbei auf ein bedeutsames Missverständnis im Kontext von Evaluationen: die überhöhten Hoffnungen („Goldstandard") auf Programmevaluationen im Sinne von Wirksamkeitsstudien, die mit (quasi-)experimentellen Designs den Eindruck erzeugen, Kausalbeziehungen im Feld nachweisen zu können. Bedeutsam ist dieses Missverständnis in zweierlei Hinsicht: Erstens zeigt es die übersteigerten Ansprüche, die Auftraggeber an Evaluationen erheben, ohne dass diese forschungsmethodisch und forschungspraktisch erfüllt werden könnten, und zweitens die starke Fokussierung auf Wirksamkeitsstudien, die mit ihren Kausalimplikationen Technologisierbarkeitsphantasien schüren und somit unerfüllbare Hoffnungen gegenüber dem sozialwissenschaftlichen Instrument Evaluation auslösen.

Wie man demgegenüber reflektierter mit der Idee von Programmevaluation umgehen kann, illustriert *Susanne Giel* anhand eines fiktiven Beispiels von Willkommensklassen. Sie verdeutlicht, wie gleichzeitig wirkungsorientiert und praxisorientiert operiert werden kann, so dass vermittelt über eine konkrete Programmtheorie keine unerfüllbaren Ansprüche gegenüber Evaluationsmaßnahmen formuliert werden.

Eine besondere Form des Praxisbezugs stellen *Sylvia Rahn*, *Sabine Gruehn*, *Miriam Keune* und *Christoph Fuhrmann* vor. Sie nehmen eben jene Akteursgruppe in den Blick, die sonst oftmals vergessen wird: die Schülerinnen und Schüler. Hierbei ist die Zwiespältigkeit zu bearbeiten, dass Schulqualität sich nicht nur durch Zufriedenheitsabfragen bei Schülerinnen und Schülern erheben lässt, gleichwohl aber umgekehrt deren Perspektive für eine „professionelle Feedbackkultur" in der Schule unentbehrlich ist.

Wie komplex das Zusammenspiel unterschiedlicher Daten hierbei – auch auf Systemebene – gedacht werden muss, illustrieren *Klaudia Schulte, Detlef Fickermann* und *Markus Lücken* in ihrem Bericht zum Hamburger Prozessmodell zur datengestützten Schulentwicklung, in dem sie aufzeigen, wie Daten unterschiedlicher Aggregationsebenen in ein Gesamtsystem sinnvoll integriert werden müssen, um Effekte für die Schulentwicklung auslösen zu können.

Dass allerdings die Orchestrierung eines Datenmanagements auf Systemebene für Schulentwicklung zwar eine wahrscheinlich notwendige, aber noch nicht hinreichende Bedingung ist, zeigen *Wolfgang Beywl* und *Lars Balzer* in ihrem Beitrag zum „Aufbau von Evaluationskompetenzen für interne Schulevaluation durch projektbezogene Fortbildung". Die Autoren entwerfen praxisnah ein „Kompetenzportfolio für schulinterne Evaluationen", um auch auf schulischer Seite die Voraussetzungen für positive Effekte evaluativen Denkens und Handelns zu schaffen.

Der Durchgang durch Strukturprobleme und Anwendungsbezüge von Evaluation im schulischen Feld zeigt das hohe Maß an Ausdifferenziertheit des Evaluationsdiskurses, das sich – weltweit – nicht zuletzt in der Existenz der zahlreichen Fachgesellschaften für Evaluation dokumentiert. Hier werden Standards für Evaluationen formuliert, die auch für alle anderen sozialen Felder jenseits der Schule gelten, wie bspw. das Gesundheits- oder das Sozialwesen. Für den deutschsprachigen Raum existieren hier die Standards der DeGEval (vgl. dazu ausführlich den Beitrag von *Wolfgang Böttcher* und *Jan Hense* im vorliegenden Heft).

Demgegenüber hat sich in der schul(-system-)bezogenen empirischen Bildungsforschung ein spezifisches Verständnis von „Evaluation" durchgesetzt, das sich nicht zuletzt durch die bildungspolitischen Legitimationszwänge des Monitorings der öffentlichen Institution Schule ergibt und damit immer auch in der Gefahr steht, eine stärkere Nähe zum Kontroll- und weniger zum Entwicklungsparadigma zu haben.

Dieses spezifische Evaluationsverständnis der schul(-system-)bezogenen empirischen Bildungsforschung wird im abschließenden Bericht des Themenschwerpunkts entfaltet. In seinem Beitrag nimmt *Hans Merkens* dementsprechend Begriffsbestimmungen vor, systematisiert historische Theorielinien und entwirft einen Orientierungsrahmen für die unterschiedlichen Bezugspunkte von Evaluation in ihren Kontexten der Neuen Steuerung, der Bildungsstandards und anderer Bereiche des Schulsystems.

Die Beiträge dieses Themenheftes dokumentieren damit die mehrfachen Interessenslagen und Bedürfnisse gegenüber dem Evaluationsbegriff und seinen Anwendungen im Schulbereich: sowohl die kritische Reflexion von Evaluation in ihren Anwendungskontexten als auch konkrete Varianten der Operationalisierung, des Datenmanagements und der Beteiligung unterschiedlichster Akteursgruppen sowie letztendlich wiederum das Bedürfnis nach einer orientierenden Begriffsklärung – auch wenn deutlich geworden sein dürfte, dass solche Vereindeutigungen nie für alle paradigmatisch divergierenden Diskurse Geltung beanspruchen können. Die schulbezogene Evaluationsforschung stellt eben nur einen Teil des gesamtgesellschaftlich bedeutsamen Evaluationsdiskurses dar. Die Beiträge in diesem Heft belegen, dass es Sinn macht, über die vermeintliche Selbstverständlichkeit von Evaluation im Schulkontext nochmals nachzudenken, oder wie *Hans Merkens* in seinem Schlusssatz resümierend zum Evaluationsbegriff festhält: „Die Selbstverständlichkeit, mit der der Begriff verwendet wird, trügt."

Kathrin Dedering/Martin Heinrich

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

DDS – Die Deutsche Schule
108. Jahrgang 2016, Heft 2, S. 136-148

Jutta Wolff

Das evaluieren wir (mal eben)

Was Auftraggebende über Wirksamkeitsnachweise wissen sollten

Zusammenfassung

Auftraggebende verbinden mit dem Inauftraggeben einer Evaluation oft hohe Erwartungen, wobei die Notwendigkeit eigener vorbereitender Arbeiten teilweise unterschätzt und das Potenzial von Evaluationen, die gewünschten Erkenntnisse zu generieren, vielfach überschätzt werden. So soll mit Evaluationen häufig die Wirksamkeit eines Programms nachgewiesen werden. Verbreitet ist deshalb der Ruf nach (quasi-)experimentellen Designs, die beim Nachweis kausaler Beziehungen oftmals als „Goldstandard" gelten. Im Beitrag wird dargestellt, welchen Begrenzungen diese Designvarianten gerade im pädagogischen Bereich unterliegen.
Schlüsselwörter: Bildungsevaluation, Evaluationsplanung, Evaluationsprozess, Evidenzbasierung, (quasi-)experimentaldesigngesteuerte Evaluation, Wirksamkeit

We Evaluate That (just Quickly)

What Customers Should Know about the Proof of Effectiveness

Summary

Customers often have high expectations when commissioning an evaluation, whereby the own planning needs are partly underestimated and the potential of evaluations to generate the desired knowledge is overestimated. Often an evaluation is intended to demonstrate the effectiveness of a program. Then the call for (quasi-)experimental designs, which are viewed as a "gold standard" in the detection of causal relationships, is widespread. The article illustrates the limits of these design variations in the field of education
Keywords: educational evaluation, evaluation planning, evaluation process, evidence-based, (quasi-)experimental design driven evaluation, effectiveness

1. Einleitung

Evaluationen können verschiedenste Gegenstände in den Blick nehmen wie z.B. Organisationen, Curricula, Projekte, Leistungen. Der Beitrag fokussiert auf *Programmevaluationen*, worunter Beywl und Niestroj (2009, S. 83) folgend die „Evaluation eines Bündels von Maßnahmen (Programm), das basierend auf einem Set von Ressourcen aus einer Folge von Interventionen besteht", verstanden wird, das „auf bestimmte, in der Regel bei bezeichneten Zielgruppen zu erreichende Resultate gerichtet ist."

Nicht immer ist allen Beteiligten bewusst, dass es sich bei Evaluationen um sehr voraussetzungsvolle Prozesse handelt, in denen zahlreiche Entscheidungen zu fällen sind, die den Erkenntnisgewinn der Evaluation und die Nutzung der Ergebnisse beeinflussen: Auftraggebende[1] von Evaluationen nehmen bisweilen an, ihre Aufgabe sei mit der Beauftragung von professionellen Evaluierenden weitestgehend abgeschlossen, und sie könnten nach einigen Monaten oder Jahren den Ergebnisbericht interessiert entgegennehmen. Doch nur, wenn Auftraggebende und ggf. weitere Beteiligte intensiv an der Planung der Evaluation mitwirken, können nützliche und potenziell genutzte Evaluationsergebnisse generiert werden.

Entsprechend wurden in den letzten Jahren Leitfäden für Auftraggebende entwickelt, um diese im Vorfeld einer Evaluation über die auf sie zukommenden Aufgaben und zu klärenden Aspekte zu informieren (vgl. Deutsches Jugendinstitut 2010; DeGEval 2012). Unter anderem wird darauf aufmerksam gemacht, dass Auftraggebende (ggf. mit Unterstützung der Evaluierenden)

- *ihren Eigenaufwand und den weiterer Beteiligter realistisch einschätzen sollten* (z.B. zeitliche und personelle Ressourcen für Informationsweitergabe, Absprache von Evaluationseckpunkten und ggf. Erhebungsinstrumenten, Datenerhebung),
- *eine Evaluation rechtzeitig in Auftrag geben sollten* (z.B. um Dokumentationsverfahren eines Programms so zu konzipieren, dass sie gleichzeitig als Datenquelle für die Evaluation genutzt werden können, oder um Erhebungen der Ausgangslage zu ermöglichen),
- *prüfen sollten, ob und in welchem Umfang wichtige Akteure in den Evaluationsprozess eingebunden werden sollten, und diese ggf. einbeziehen* (z.B. um die Qualität der Evaluation durch fachliche Beratung und Akzeptanzsteigerung zu erhöhen),
- *den Evaluationszweck und die Fragestellungen präzisieren sollten* (u.a. um die Evaluation auf diese fokussieren zu können),
- das Verhältnis von Aufwand und Nutzen abwägen sollten.

1 Mit „Auftraggebenden" sind im Folgenden neben Geldgebern v.a. Projektverantwortliche oder andere Personen gemeint, die als Ansprechpartner für die Evaluierenden fungieren.

„Das evaluieren wir (mal eben)" betrifft jedoch nicht nur die Verkennung der notwendigen Aktivitäten der Auftraggebenden im Planungsprozess einer Evaluation, sondern auch – und das soll hier im Fokus des Beitrags stehen – die oftmals überhöhten Erwartungen an die Aussagekraft von Evaluationsergebnissen bezüglich der Wirksamkeit eines Programms.

Werden neue pädagogische Maßnahmen implementiert, interessiert es sowohl Auftraggebende als auch Verantwortliche von Konzeption und Durchführung, ob die Intervention so, wie gedacht, funktioniert („what works"), d.h., ob eine Maßnahme die erhofften Wirkungen entfaltet. Dabei sollen in der Regel die Annahmen über einen Ursache-Wirkungs-Zusammenhang geprüft werden, um die gemessenen Wirkungen *kausal* auf eine Intervention zurückzuführen (= Wirksamkeit).

Doch wie lässt sich solch ein Wirksamkeitsnachweis führen? In der evidenzbasierten Medizin hat sich ein Stufenmodell des Wirksamkeitsnachweises etabliert. In dieser Evidenzhierarchie gelten Experimente (= randomisiertes Kontrollgruppendesign) als „Goldstandard". Ursprünglich in Studien zur Wirksamkeit von Medikamenten eingesetzt, wurde das randomisierte Kontrollgruppendesign als Forschungsmethode auf weitere Gebiete der Medizin und schließlich auch auf den sozialen Bereich ausgeweitet. Verstärkt wurde dies unter anderem durch die Neue Steuerung im Bildungswesen, in deren Zuge der Generierung von empirischer Evidenz eine immer größere Bedeutung zugeschrieben wurde (vgl. Hammersley 2013; Pant 2014).

Die Annahme einer generellen Überlegenheit des experimentellen Designs gegenüber anderen Designvarianten führt in der Evaluationspraxis häufig dazu, dass bei der Frage nach einem Wirksamkeitsnachweis reflexartig ein (quasi-)experimentelles Design angestrebt wird, z.T. ohne Kenntnis der Implikationen eines solchen Designs.[2] Im Folgenden sollen deshalb Begrenzungen einer (quasi-)experimentaldesigngesteuerten Evaluation[3] angesprochen werden, um Auftraggebenden (und Evaluierenden) eine Abschätzung von Chancen und Risiken einer solchen Designvariante zu ermöglichen. Die „Checkliste" im Anhang ist ein erster Versuch, überblicksartig Aspekte zu benennen, die bei der Erwägung einer (quasi-)experimentaldesigngesteuerten Evaluation bedeutsam sind.

2. Begrenzungen (quasi-)experimentaldesigngesteuerter Evaluation

Ziel des experimentellen Designs ist es, einen Kausalzusammenhang zwischen einer Intervention (unabhängige Variable) und einer Wirkung (abhängige Variable) zwei-

2 Ausgangspunkt der intensiven Beschäftigung mit dieser Thematik waren eigene Erfahrungen bei der Umsetzung eines quasi-experimentellen Designs bei der Evaluation eines Qualifizierungsprogramms für Schulen und ihre Lehrkräfte (vgl. Wolff 2015).

3 Diesen Begriff habe ich Beywl (2006) entnommen.

felsfrei zu belegen (= interne Validität). Hierbei wird das Prinzip des Vergleichs genutzt: Es werden Zielgrößen für die am Programm teilnehmende Gruppe (Experimentalgruppe) gemessen und mit den Werten einer Kontrollgruppe, die nicht an der Intervention teilgenommen hat, verglichen.[4] Die Zuweisung zu Experimental- bzw. Kontrollgruppe erfolgt per Zufall, d.h. durch *Randomisierung*. Dahinter steht die Annahme, dass bei genügend großen Versuchsgruppen Personenvariablen, die ggf. Einfluss auf das Ergebnis haben („Störfaktoren"), ausgemittelt werden und Unterschiede der abhängigen Variablen ausschließlich auf die Intervention zurückgeführt werden können.

Zentrale Komponenten des Experiments sind demnach: (1) Randomisierung: Zuordnung zu Interventions- und Kontrollgruppe per Zufall, (2) eine Intervention, die bestimmte Wirkfaktoren (unabhängige Variablen) umfasst, und (3) die Messung von abhängigen Variablen, d.h. der angestrebten Wirkungen.

Im Folgenden werden Schwierigkeiten bei der Umsetzung dieser zentralen Komponenten im pädagogischen Bereich und der zu erwartende Erkenntnisgewinn angesprochen.[5]

2.1 Probleme der Randomisierung

Der relativ unbestrittene Nutzen des Experiments liegt darin, dass durch die Randomisierung die Gefahr des Selektionsbias reduziert wird: Durch die Zufallszuweisung wird sichergestellt, dass sich die Objekte in den Versuchsgruppen (z.B. Schulen, Lehrkräfte, Schüler und Schülerinnen) nicht systematisch unterscheiden und gemessene Unterschiede tatsächlich auf die Intervention zurückzuführen sind. Aber gerade dieser Vorteil des Experiments kann im pädagogischen Kontext oft nicht zum Tragen kommen, da die Zufallszuordnung oft nicht möglich ist. So stellt es ein ethisches Problem dar, bestimmten Personen eine Intervention zu versagen oder andere mit einer Intervention „zwangszubeglücken".

Stattdessen ist es in der Praxis die Regel, dass die Teilnehmenden der Interventionsgruppe bereits festgelegt sind, u.a. durch Auswahlkriterien des Programms (z.B. Kinder mit erheblichen Sprachdefiziten nehmen am Sprachförderprogramm teil) oder durch das Anmeldeverhalten der Zielgruppe (z.B. Anmeldung zu einer Lehrerfortbildung ⇨ Selbstselektion). Die Schwierigkeit für Evaluierende besteht dann darin, eine möglichst äquivalente Vergleichsgruppe zu finden, die sich nicht durch wesentliche Merkmale von der Interventionsgruppe unterscheidet. Handelt es sich um ein Programm mit partieller Erfassung, nimmt also noch nicht die gesamte

4 Für weitere Spezifikationen dieses Grundgedankens vgl. Frey/Frenz (1982).
5 Für eine intensivere Beschäftigung mit der Thematik siehe Bellmann/Müller (2011).

Zielgruppe am Programm teil, so können „Wartegruppen" als Kontrollgruppe herangezogen werden. Bei Programmen, die die gesamte Zielgruppe umfassen, ist die Suche nach Vergleichsgruppen nochmals schwieriger.

Sobald die Zuweisung zu den Versuchsgruppen nicht per Zufall erfolgen kann (d.h., „nur" ein *quasi*-experimentelles Design umgesetzt werden kann), werden Annahmen darüber benötigt, welche Merkmale der Versuchspersonen ggf. die Wirkungen (abhängige Variable) beeinflussen, um diese Merkmale bei den statistischen Analysen entsprechend berücksichtigen zu können. Da diese Annahmen aufgrund fehlender Theorien eher Vermutungen sind, werden häufig hilfsweise gängige soziodemographische Variablen wie z.B. Alter, Schulform etc. in die Berechnungen einbezogen, ohne sicher zu sein, alle entscheidenden Variablen erfasst zu haben. Entsprechendes gilt für das Matching, bei dem für Objekte der Interventionsgruppe statistische Zwillinge gesucht werden (und das zudem das Vorliegen von entsprechenden Datensätzen voraussetzt).

2.2 *Was* wirkt denn eigentlich?

Eine Prämisse für ein experimentelles Design ist, dass die untersuchte Maßnahme klar identifizierbar sein und relativ standardisiert durchgeführt werden muss. Bei Experimenten zur Wirksamkeit von Medikamenten ist die Intervention in hohem Maße standardisiert: Die *Bestandteile* des Medikaments und die *Dosis* werden kontrolliert. Im pädagogischen Bereich ist die Standardisierung ungleich schwieriger (vgl. Hammersley 2015).

Vielzahl potenzieller Wirkfaktoren
Bereits in der Grundlagenforschung ist es eine kaum zu lösende Aufgabe, die theoretischen Vorstellungen in konkrete Variablen eines Feldexperiments umzusetzen (vgl. Frey/Frenz 1982). Im Kontext von Evaluationen mit starker Anwendungsorientierung gilt dies umso mehr: Der Ausgangspunkt von Evaluationen sind in der Regel Programme, die aus der Praxis heraus als Antwort auf spezifische Anforderungen/Probleme entwickelt wurden. Die Überprüfung der Wirksamkeit stand dabei nicht im Fokus. In der Konsequenz bedeutet dies, dass ein Programm häufig aus einer Vielzahl von Interventionen besteht, d.h., eine Vielzahl von möglichen „Wirkfaktoren" beinhaltet, für deren Zusammenwirken selten explizierte Annahmen bestehen.

Ein Beispiel: Es wird das Problem erkannt, dass Schulen mit niedrigem Sozialindex hinsichtlich der Leistungsergebnisse ihrer Schülerinnen und Schüler hinter anderen Schulen zurückbleiben. Es wird ein Programm zur Unterstützung dieser Schulen aufgelegt, um die fachlichen Kompetenzen der Schülerinnen und Schüler zu verbessern. Dieses Programm wird in der Regel nicht aus einer einzelnen, eng umschriebenen Maßnahme bestehen (z.B. zusätzliche einstündige wöchentliche Förderung in

Mathematik durch Mathematikfachkräfte mit genauer Durchführungsanweisung). Eher wird – ausgehend von Überlegungen zu wichtigen Einflussfaktoren auf Leistungsergebnisse und pragmatischen Erfordernissen – ein Maßnahmenbündel „geschnürt", und es werden verschiedene Maßnahmen ergriffen, die z.B. auf Ebene von Schulleitung (Schulleitungshandeln), Lehrpersonen (Unterrichtsentwicklung), Eltern (elterliche Unterstützung) und Schülerinnen und Schülern (z.B. Lernberatung) ansetzen könnten. Andere, bereits für alle Schulen implementierte Maßnahmen, wie z.B. die Lernförderung bei nicht ausreichenden Leistungen, werden beibehalten.

Wird die Fachleistung der Schülerinnen und Schüler der Programmschulen und möglichst ähnlicher Schulen[6] in Zeitreihen erfasst, so könnte sich herausstellen, dass sich sowohl in der Interventionsgruppe als auch in der Vergleichsgruppe gleichermaßen Schulen mit und ohne deutliche Leistungssteigerungen finden lassen. So berichtet Berliner (vgl. 2002, S. 19) von Programmevaluationen, die ergaben, dass die Schülerleistungen innerhalb eines Programms stärker variierten als zwischen den Programmen. Wie lässt sich das erklären?

Macht des Kontexts/Vielzahl von Interaktionen
Selbst wenn die Kernelemente einer Intervention als „Ausführungshinweise" schriftlich fixiert sind (das ist selten der Fall), ist der konkrete Verlauf im pädagogischen Kontext in verschiedenen Umgebungen nicht identisch, *kann* er nicht identisch sein: U.a. Berliner (2002) macht darauf aufmerksam, dass einerseits zahlreiche (lokale) Kontextbedingungen die konkrete Ausgestaltung pädagogischer Aktivitäten mitbestimmen (*Power of Contexts*) und dass andererseits die Interaktionen zwischen den Beteiligten zu völlig unterschiedlichen Verläufen führen (*Ubiquity of Interactions*).

Als Beispiel führt er das Klassenzimmer an, in dem verschiedene Charakteristika der Lehrpersonen (z.B. Ausbildung, Motivation) mit Eigenschaften der Schülerinnen und Schüler (z.B. IQ, sozioökonomischer Status, Lernmotivation) ebenso interagieren wie mit Merkmalen der Umwelt (z.B. Lehrpläne, Jugendarbeitslosigkeit) und die Wirkungsrichtung oft unklar sei: „Moreover, we are not even sure in which directions the influences work, and many surely are reciprocal. Because of the myriad interactions, doing educational science seems very difficult, while science in other fields seems easier." (Berliner 2002, S. 19)

Kelle (2006) macht auf ein Grundproblem der Kausalanalyse sozialen Handelns aufmerksam: Diejenigen, die das Programm planen oder anordnen, können nur die *Handlungsbedingungen* der Akteure vor Ort *direkt* beeinflussen, nicht aber deren Handlungen selber. Da diese individuellen Akteure ggf. konkurrierende Handlungsziele verfolgen, müsse gerade in Interventionsstudien

6 Wobei sich auch hier die Frage stellt, welches „ähnliche" Schulen sind: ähnlich hinsichtlich von Merkmalen der Schülerinnen und Schüler (z.B. Sozialindex) oder der Lehrpersonen (z.B. Alter, unterrichtsrelevante Einstellungen) etc.

> „damit gerechnet werden, dass die Betroffenen in nicht vorhersagbarer Weise auf Interventionsmaßnahmen reagieren und dabei in kreativer Weise jene Handlungsbedingungen verändern, die durch die Intervention sozialtechnologisch geschaffen und beeinflusst werden sollen." (Ebd., S. 133)

Die hier aus Forschersicht als „Störfaktor" kritisierte Variabilität von pädagogischen Maßnahmen ist aus der Sicht von Pädagogen und Pädagoginnen oftmals erwünscht und für den Erfolg einer Intervention unverzichtbar. So muss z.B. eine Lehrperson ihr eigenes Verhalten an die Klasse adaptieren, um Lernfortschritte zu ermöglichen. In welchem Ausmaß eine Intervention im Feld variiert wird – darauf haben die Evaluierenden in der Regel keinen Einfluss. Es bleibt ihnen lediglich, die *Implementationstreue* zu erheben, um Umfang, Genauigkeit und Qualität der Implementation abschätzen zu können. Evaluationsergebnisse können dann eher eingeordnet werden. Die vergleichsweise geringen Lernfortschritte im Bereich Sprache in einer Kindertagesstätte können dann z.B. evtl. mit der langfristigen Erkrankung der Förderfachkraft oder mit deutlichen Abweichungen vom Ursprungsprogramm in Verbindung gebracht werden – und gehen weniger zulasten des Programms.

Dilemma interne und externe Validität (ökologische Validität)
Forschungen auf der Basis von (Quasi-)Experimenten befinden sich in einem Dilemma zwischen interner und externer Validität: Die interne Validität, d.h. die eindeutige kausale Interpretierbarkeit eines Zusammenhangs (z.B. kann die Veränderung der Fachleistung Mathematik *eindeutig* auf die infrage stehende Maßnahme zurückgeführt werden), ist am ehesten zu erreichen, wenn eine Intervention unter hochkontrollierten und deshalb recht künstlichen (Labor-)Bedingungen durchgeführt wird – die unabhängige Variable somit eindeutig identifizierbar ist. Fraglich ist dann jedoch die externe (ökologische) Validität: Gelten die unter künstlichen Bedingungen erlangten Ergebnisse auch in realen Situationen, können sie auf Kontexte verallgemeinert werden, die nicht untersucht wurden?

Wird ein stärker realitätsnahes Forschungsdesign genutzt, z.B. indem eine Intervention von verschiedenen Personen in verschiedenen Klassen durchgeführt wird, ist die Gefahr unkontrollierbarer oder unbeachteter Störeinflüsse groß und die interne Validität, d.h. die eindeutige Rückführbarkeit einer Wirkung auf die Maßnahme, gefährdet.

2.3 Wie lassen sich Wirkungen *messen*?

Als „Messen" wird Kriz und Lisch (1988, S. 175) folgend „die Zuordnung von Zahlen (numerisches Relativ) zu Objekten und deren Eigenschaften (empirisches Relativ) mit dem Ziel einer isomorphen oder homomorphen Abbildung" verstanden. Diese „Zuordnung von Zahlen" ist je nach Zielvariable unterschiedlich schwierig: Soll die

Wirksamkeit eines blutdrucksenkenden Medikaments untersucht werden, so lässt sich ein Blutdruckmessgerät verwenden. Eine wesentlich höhere Anforderung stellt jedoch die Messung z.B. von Leistung, Lernmotivation, Unterrichtsverhalten, Kooperation der Lehrpersonen etc. dar. Diese Herausforderung gilt zwar nicht nur für (quasi-)experimentelle Designs, beschädigt jedoch deren Anspruch, besonders überzeugende Belege für die Wirksamkeit einer Maßnahme zu liefern.

Frey und Franz (1982, S. 231) sehen es zurecht als eine der zentralen Fragen an, wie man ein Messinstrument „finden (bzw. konstruieren) könne, das in der Lage ist, das in der Hypothese ideell fixierte Verhaltensphänomen O (Observation) tatsächlich empirisch zu erfassen." Hierzu ist zunächst ein umfangreiches Wissen über die zu messenden Variablen notwendig – das auch die Frage umfasst, ob diese überhaupt eine quantitative Struktur aufweisen, also eine metrische Messung angemessen ist (vgl. Hammersley 2013, S. 69f.). Welche Annahmen oder gar Theorien gibt es dazu, was „Unterrichtsqualität", „Lehrerkooperation" etc. ausmacht, bzw. was soll mit dem Programm genau erreicht werden?

Im nächsten Schritt müssen Messinstrumente gefunden bzw. konstruiert werden, die in der Lage sind, diese Konzepte abzubilden. In der Evaluationspraxis gibt es hierzu zwei Vorgehensweisen:

(1) Es wird (ressourcensparend) auf bereits in anderen Studien verwendete Messinstrumente zurückgegriffen. Für den Bereich Schule liegen z.B. zahlreiche Fragebögen zu zentralen Konzepten wie Klassenklima, Unterrichtsverhalten etc. vor.[7] Die Gefahr besteht hier, dass mit Messinstrumenten, die nicht auf das Programm zugeschnitten wurden, der Erfolg der Maßnahme nicht hinreichend erfasst werden kann:

> „Ein dann routinemäßig geübter Rückgriff auf vorhandene und vermeintlich bewährte standardisierte Messinstrumente kann erhebliche Probleme mit sich bringen, wenn die hiermit gemessenen Variablen zu unspezifisch sind, um den Erfolg zu erfassen." (Kelle 2006, S. 127)

(2) Es werden neue Messinstrumente entwickelt. Dieses ist voraussetzungsvoll und i.d.R. mit hohem Aufwand verbunden.

Problematisch erscheint zudem, dass vornehmlich solche Messinstrumente gewählt werden, die angesichts von Zeitknappheit und Kostendruck mit möglichst geringem Aufwand angewendet werden können, wie z.B. Selbsteinschätzungsskalen. Sie haben den Vorteil, dass sie die Untersuchenden „quasi in Sekundenschnelle mit quantitativer Information über Parameter versorgen, auch wenn deren empirische Korrelate

7 Viele Skalen samt Vergleichswerten werden leicht zugänglich vom Deutschen Institut für Pädagogische Forschung in der „Datenbank zur Qualität von Schule" (DaQS; URL: http://daqs.fachportal-paedagogik.de/; Zugriffsdatum: 11.04.2016) vorgehalten.

umstritten oder gänzlich unklar sind." (Frey/Frenz 1982, S. 255) So werden beispielsweise zur Erhebung von Unterrichtsverhalten Einschätzungsskalen für Lehrpersonen oder Schülerinnen und Schüler genutzt – das komplexe Unterrichtsgeschehen somit auf leicht erfassbare und methodisch unkompliziert erhebbare Items reduziert.

Evaluierende befinden sich in einem Dilemma zwischen Wissenschaftlichkeit und Praktikabilität: Die Erhebung eines Konstrukts sollte gegenstandsangemessen erfolgen, was ggf. umfangreichere und methodisch anspruchsvollere Erhebungen erforderlich macht. So ist die Messung einer Variablen in der Regel umso aufwendiger, je höher sie in der Hierarchie der Resultatsarten steht: Outputs (wie z.B. Zahl der Teilnahmen an einer Fortbildung, Zufriedenheit) sind einfacher zu messen als Wissen oder Kenntnisse (Outcome I), diese wiederum einfacher als Verhalten (Outcome II) etc.[8] Andererseits müssen neben dem Zeit- und Kostenrahmen auch datenschutzrechtliche Bestimmungen beachtet werden, die ggf. das Gewünschte einschränken (z.B. die Beobachtung/Befragung von Schülerinnen und Schülern). Auch muss berücksichtigt werden, dass Evaluationen nur dann zu nützlichen Ergebnissen führen, wenn die Datengebenden die Datenerhebungsmethoden akzeptieren. Das heißt: Obwohl die Messgenauigkeit i.d.R. mit der Länge des Messinstruments steigt, darf dieses nicht zu umfangreich sein. Auch ist nicht jede Messmethode überall einsetzbar; so ist beispielsweise die Akzeptanz von Wissenstests bei Lehrpersonen eher gering.

2.4 Erkenntnisgewinn von (Quasi-)Experimenten

Bei Durchführung einer (quasi-)experimentaldesigngesteuerten Evaluation sollten sich Auftraggebende und Evaluierende nicht nur über zahlreiche praktische Probleme und die sehr hohen Kosten im Klaren sein, sondern auch über den potenziellen Erkenntnisgewinn, der mit dieser Designvariante verbunden ist.

Im Idealfall ergeben sich in der Interventionsgruppe gegenüber der Kontrollgruppe deutlich höhere Ausprägungen der Zielvariablen. Unklar bleibt jedoch, welche Wirkmechanismen hierfür ausschlaggebend waren und ob sich dementsprechend die Ergebnisse in anderen Kontexten replizieren lassen. Auch lassen die Ergebnisse selbst bei sehr sorgfältig und aufwendig gestalteten Quasi-Experimenten vielfältigen Interpretationsspielraum. Sind z.B. für die geringfügig höheren Leistungsergebnisse der Schülerinnen und Schüler, die mit der „Gruppenrallye im Mathematikunterricht" unterrichtet wurden, Novitätseffekte der Methode verantwortlich (vgl. Wandeler et al. 2015)?

8 Bei der Benennung der Resultatsarten orientiere ich mich an der Einteilung von W. Beywl. URL: http://www.eval-wiki.org/w_glossar/images/f/fb/Variantentafel-Resultatsarten-fuer_Glossar.pdf; Zugriffsdatum: 29.11.2015.

Nicht selten ergeben sich nur sehr geringe Unterschiede zwischen den Gruppen – was angesichts der multifaktoriellen Bedingtheit vieler Zielvariablen nicht verwundert. Es könnte sogar der Fall eintreten, dass die Ergebnisse nicht in die gewünschte Richtung weisen. Erst mithilfe qualitativer Methoden können Erklärungen für solche unerwünschten Effekte gefunden werden (vgl. Kelle 2006). Ein ausschließlich auf quantitativen Daten fußendes Design ist geeignet, unter bestimmten Bedingungen ein überblicksartiges Wissen über den Erfolg oder Misserfolg eines Programms oder einer Intervention zu generieren. Die Wirkmechanismen bleiben jedoch im Dunkeln; es können keine (erwünschten oder unerwünschten) Nebenwirkungen erfasst und keine Hinweise auf die Verbesserung eines Programms generiert werden.

3. Schluss

Der Beitrag sollte deutlich machen, dass es kein generell überlegenes Erhebungsdesign oder keine überlegenen Erhebungsmethoden zum Nachweis der Wirksamkeit gibt – jedes Design, jede Methode hat Stärken und Schwächen. Insofern schlägt Hammersley (2015) vor, eher von einer Matrix als von einer Hierarchie der Evidenz auszugehen.

Auch Berliner warnt davor, Wissenschaft und Methode zu verwechseln. Ein experimentelles Design sei eine Methode oder Technik, aber nicht mit Wissenschaft gleichzusetzen:

> "But to think that this form of research is the only 'scientific' approach to gaining knowledge – the only one that yields trustworthy evidence – reveals a myopic view of science in general and a misunderstanding of educational research in particular." (2002, S. 18)

Kelle (2006, S. 134) zufolge wird die Analyse kausaler Beziehungen „oft zu Unrecht für ein genuines und exklusives Feld quantitativer Datenanalyse" gehalten. Qualitative Forschung halte Lösungen bereit zur Identifikation unterschiedlicher kausaler Pfade und von (erwünschten und unerwünschten) Nebenwirkungen.

Es kann zusammenfassend festgehalten werden: Wirkungen kausal auf ein Programm zurückzuführen, ist in der Regel aufwendig und teuer. Es ist für jedes einzelne Evaluationsprojekt abzuwägen, welches Design (*one-shot*, mehrere Messzeitpunkte, …) und welche Methoden (quantitative, qualitative) oder welche Kombinationen von Methoden in Anbetracht der Kontextbedingungen (z.B. Ressourcen) am besten geeignet sind, die Wirkungen eines Programms nachzuweisen. Während quantitative Methoden geeignet sind, ein überblicksartiges Wissen zu den Wirkungen eines Programms zu generieren, liegt die Stärke qualitativer Methoden in der Erkundung der zugrunde liegenden Wirkmechanismen. Insofern erscheint eine Kombination

quantitativer und qualitativer Methoden beim Führen eines Wirkungsnachweises geboten – aber angesichts oftmals begrenzter Ressourcen schwer realisierbar.

Und manchmal gilt vielleicht auch die Mahnung Hammersleys (2015, S. 9), wonach Evaluationen nicht auf alle drängenden Fragen Antworten geben und die Validität der Ergebnisse garantieren können. Um nicht falsche Erwartungen zu wecken, muss auch dies manchmal ausgesprochen werden:

> "Unfortunately, we must face the fact that research cannot always provide answers to questions that are seen as pressing by policymakers and practitioners. Nor can we guarantee the validity of our findings. And we need to make this clear to lay audiences. So, there is a danger of over-promising or over-claiming."

Anhang: „Checkliste" für eine (quasi-)experimentaldesign-gesteuerte Evaluation

Ohne Anspruch auf Vollständigkeit werden im Folgenden checklistenartig einige – z.T. bereits angesprochene – Aspekte aufgeführt, die geprüft werden sollten, sobald ein (quasi-)experimentelles Design in Betracht gezogen wird.

Wie sollte das zu evaluierende Programm beschaffen sein?

1) Das zu evaluierende Programm lässt angesichts seines zeitlichen Umfangs nennenswerte Wirkungen erwarten.
2) Das Programm ist teuer, so dass eine Wirkungsanalyse bedeutsam ist.
3) Das Programm hat potenziell eine große Reichweite, betrifft also viele Personen und wird voraussichtlich langfristig eingesetzt.
4) Das Programm beinhaltet wenige eng umschriebene Kernelemente, deren angenommene Wirkungen auf das Outcome expliziert sind (Wirkmodell).
5) Die Ziele des Programms sind schriftlich fixiert und evaluierbar formuliert (*smart*).
6) Die Programmstabilität ist im Rahmen des Möglichen gewährleistet: (a) Das Programm ist schriftlich fixiert, um eine ähnliche Durchführung durch verschiedene Personen wahrscheinlich zu machen. (b) Das Programm ist „ausgereift", d.h., seine Konzeptschwächen wurden bereits behoben, so dass anzunehmen ist, dass künftig keine wesentlichen Konzeptänderungen vorgenommen werden.
7) Das Programm sollte möglichst „immun" gegen (wechselnde) äußere Rahmenbedingungen sein, z.B. gegenüber Änderungen politischer Vorgaben.

Was sollte bei der Evaluation berücksichtigt werden?

1) Der Projektzeitraum der Evaluation ist angemessen: (a) Die Evaluation kann rechtzeitig beginnen, so dass Erhebungen zur Ausgangslage erfolgen können (b) Die

Laufzeit der Evaluation wird so gewählt, dass die erwarteten Wirkungen innerhalb dieser Zeit eintreten können.

2) Es stehen ausreichend Ressourcen für die Evaluation zur Verfügung. Gerade (quasi-)experimentelle Designs sind extrem teuer, da Daten zu mehreren Messzeitpunkten erhoben werden müssen.
3) Es liegt ein möglichst zuverlässiges Instrument zur Messung der Wirkungen vor (damit u.a. auch kleine Wirkungen erfasst werden können) – oder es stehen Zeit, Geld und Expertise zur Verfügung, ein solches zu entwickeln.
4) Es ist möglich, eine zur Versuchsgruppe äquivalente Vergleichsgruppe zu bilden (z.B. bei einem Programm mit partieller Erfassung eine Wartegruppe) oder auf Daten zuzugreifen, auf deren Grundlage statistische Zwillinge gebildet werden können (*matching*). Dabei liegen möglichst Annahmen darüber vor, welche Merkmale der Untersuchten die Wirkungen moderieren.
5) Die geplante Evaluation wird durch datenschutzrechtliche Bestimmungen hinsichtlich des zu erwartenden Erkenntnisgewinns nicht übermäßig eingeschränkt.
6) Der (langfristige) Zugriff auf relevante Daten ist gesichert: Es ist zu erwarten, dass die Datengebenden über einen langen Zeitraum für mehrere Datenerhebungen zur Verfügung stehen, zur Teilnahme bereit/verpflichtet und im Projektverlauf „auffindbar" sind (Gefahr des *Dropout*).
7) Der voraussichtlich durch die Evaluation generierte Nutzen steht in einem angemessenen Verhältnis zu dem (i.d.R. sehr großen) Aufwand.
8) Ergebnisse der Evaluation werden nicht sofort benötigt.

Literatur und Internetquellen

Bellmann, J./Müller, T. (Hrsg.) (2011): Wissen, was wirkt. Kritik evidenzbasierter Pädagogik. Wiesbaden: VS.

Berliner, D.C. (2002): Educational Research: The Hardest Science of All. In: Educational Researcher 31, H. 8, S. 18-20.

Beywl, W. (2006): Evaluationsmodelle und qualitative Methoden. In: Flick, U. (Hrsg.): Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek: Rowohlt-Taschenbuch-Verlag, S. 92-116.

Beywl, W./Niestroj, M. (^{2}2009): Das A-B-C der wirkungsorientierten Evaluation. Glossar – Deutsch/Englisch – der wirkungsorientierten Evaluation. Köln: Univation – Institut für Evaluation.

DeGEval – Gesellschaft für Evaluation (18.06.2012): Empfehlungen für Auftraggebende von Evaluationen. Eine Einstiegsbroschüre für den Bereich der Öffentlichen Verwaltung. Mainz: DeGEval – Gesellschaft für Evaluation. URL: http://www.degeval.de/fileadmin/Publi kationen/Publikationen_Homepage/DeGEval_-_Empfehlungen_Auftraggebende.pdf; Zugriffsdatum 19.12.2015.

Deutsches Jugendinstitut (2010): Vergabe und Begleitung externer Evaluationen in der Kinder- und Jugendhilfe. Ein Leitfaden für Auftraggebende. München: DJI (Projekt exe).

Frey, S./Frenz, H.-G. (1982): Experiment und Quasi-Experiment im Feld. In: Patry, J.-L. (Hrsg.): Feldforschung. Methoden und Probleme sozialwissenschaftlicher Forschung unter natürlichen Bedingungen. Bern: Huber, S. 229-258.

Hammersley, M. (2013): The Myth of Research-Based Policy and Practice. London: Sage.

Hammersley, M. (2015): Against 'Gold Standards' in Research: On the Problem of Assessment Criteria. Frühjahrstagung AK Methoden der DeGEval am 29.05.2015, Saarbrücken. URL: http://www.degeval.de/fileadmin/users/Arbeitskreise/AK_Methoden/Hammersley_Saarbrucken.pdf; Zugriffsdatum 08.12.2015.

Kelle, U. (2006): Qualitative Evaluationsforschung und das Kausalitätsparadigma. In: Flick, U. (Hrsg.): Qualitative Evaluationsforschung. Konzepte, Methoden, Umsetzungen. Reinbek: Rowohlt-Taschenbuch-Verlag, S. 117-134.

Kriz, J./Lisch, R. (1988): Methoden-Lexikon für Mediziner, Psychologen, Soziologen. München: Psychologie Verlags Union.

Pant, H.A. (2014): Aufbereitung von Evidenz für bildungspolitische und pädagogische Entscheidungen: Metaanalysen in der Bildungsforschung. In: Zeitschrift für Erziehungswissenschaft 17, Sonderheft 27, S. 79-99.

Wandeler, C./Niggli, A./Villiger, C./Aebischer, M./Leopold, P. (2015): Ein Quasi-Experiment zur Gruppenrallye im Mathematikunterricht: Hält die Methode, was sie verspricht? In: Empirische Pädagogik 29, H. 2, S. 161-188.

Wolff, J. (2015): Evaluation eines komplexen Fortbildungsprogramms für Schulen und ihre Lehrkräfte. Planung – Design – Herausforderungen bei der Umsetzung eines quasi-experimentellen Designs. In: Grimm, A./Schoof-Wetzig, D. (Hrsg.): Was wirklich wirkt!? Effektive Lernprozesse und Strukturen in Lehrerfortbildung und Schulentwicklung. Rehburg-Loccum: Evangelische Akademie Loccum, S. 125-142.

Jutta Wolff, geb. 1964, Wissenschaftliche Referentin für Evaluation im Institut für Bildungsmonitoring und Qualitätsentwicklung (IfBQ), Hamburg.
Anschrift: Institut für Bildungsmonitoring und Qualitätsentwicklung (IfBQ), Beltgens Garten 25, 20537 Hamburg
E-Mail: Jutta.Wolff@ifbq.hamburg.de

DDS – Die Deutsche Schule
108. Jahrgang 2016, Heft 2, S. 149-162

Susanne Giel

Vom Nutzen der Programmtheorie in Evaluationen im Schulkontext[1]

Zusammenfassung
Lehrerinnen, Lehrer und Schulleitungen stehen immer wieder vor neuen Herausforderungen, die gemeistert werden müssen. (Selbst-)Vergewisserung über gelungene, wirkungsvolle Lösungsstrategien ist wichtig; klassische Designs der Bildungsforschung sind jedoch häufig nicht anwendbar, oder ihre Ergebnisse können nicht ohne weiteres in der Schulpraxis genutzt werden. Eine auf Programmtheorien basierende Evaluation versucht, die Lücken zu schließen, wirkungsorientiert und gleichzeitig praxisorientiert vorzugehen. Im Unterschied zu üblichen Schulevaluationen steht dabei nicht die Schule als Organisation, sondern das pädagogische Programm bzw. Handeln auf dem Prüfstand. Der Artikel erläutert die Grundzüge eines theoriebasierten Evaluationskonzepts und illustriert sie am Beispiel einer fiktiven Evaluation von Willkommensklassen. Abschließend reflektiert die Autorin den Nutzen dieses Vorgehens für die Schulpraxis, für das Wissen um Lern- und Bildungsprozesse und auch für methodische Entscheidungen bei Evaluationen.
Schlüsselwörter: Programmevaluation, Programmtheorie, theoriebasierte Evaluation, Innovationen, Wirkungen, Blackbox, Überprüfung von Programmtheorien

The Use of Program Theory in Evaluation in the Context of Schools

Summary
Pedagogic staff in schools consistently faces new challenges. It has to make sure to find succeeding and effective solutions and answers. Classical educational research designs are often either inapplicable, or their results cannot simply be put into practice within the particular context. Theory-based evaluation tries to close the gap by choosing an impact-oriented and simultaneously practice-oriented approach. This article explains the main functions of program theories in evaluations. Unlike typical school evaluation, a theory-based approach puts the educational program, not the school as an organization in its center. The text illustrates the main features of program theories along the notional evaluation of educational efforts to integrate refugee children into the regular school sys-

1 Für die kritischen Kommentare und Rückfragen danke ich herzlich Simone Stroppel.

tem. Concluding, the author reflects the use of this approach for practitioners, for gaining knowledge about learning processes and educational processes, as well as for methodical decisions in the course of evaluations.
Keywords: program evaluation, program theory, theory-based evaluation, innovations, effects, black box, testing program theories

1. Ausgangslage

Bildungsforschung blickt in Deutschland auf eine lange Tradition zurück. Seit dem sogenannten PISA-Schock spielen auch vermehrt Bewertungen eine große Rolle, und somit hat Evaluation an Schulen Einzug gehalten. Im föderalen Bildungssystem wurde dazu eine breite Palette an Verfahren entwickelt (vgl. URL: http://www.bildungsserver.de/Schulentwicklung-Institute-und-Materialien-der-Laender-5079.html; Zugriffsdatum: 11.04.2016). Diese Verfahren reichen von individuellen Selbstevaluationen, die in erster Linie die Zielerreichung konkreter Unterrichtsvorhaben überprüfen, über externe Peer-Evaluationen bzw. Schulbegehungen, die vor allem Schulentwicklungsprozesse in den Blick nehmen, bis hin zu fächer- und klassenstufenspezifischen Lernstandserhebungen, die Vergleiche zwischen z.B. Schulklassen, Schulen, Bundesländern und Ländern ermöglichen. Insofern kann festgestellt werden, dass es in Schulen eine vielfältige Evaluationspraxis gibt (siehe auch Griese 2016).

Teilweise im Rahmen der Qualitätsdebatte angestoßen, teilweise durch allgemeine gesellschaftliche Entwicklungen und Ereignisse ausgelöst, finden an Schulen innovative Prozesse statt, die Antworten auf aktuelle Herausforderungen liefern sollen. Schlaglichtartig können folgende Handlungsfelder genannt werden: Sprachförderung, interkulturelle Öffnung, Demokratieerziehung und kulturelle Bildung als neue fächerübergreifende Aufgabenstellungen, die Einbeziehung externer Kooperationspartner (insbesondere im Rahmen der Ganztagsbeschulung) und verändertes Lernen durch digitale Medien und das Internet. Aktuell stellt die Integration von geflüchteten Kindern besondere Anforderungen an viele Schulen.

Auf alle diese Herausforderungen müssen Schulen Antworten finden, indem sie Innovationen erproben und sich damit oft auf unbekanntem Terrain bewegen. In solchen Situationen sind Wissen und Erfahrung wertvoll. Evaluation als wissenschaftliche Dienstleistung verfügt über Vorgehensweisen und Methoden, die solche Informationen generieren und die praktisches Handeln wirkungsorientiert unterstützen (vgl. Böttcher/Hense in diesem Heft, S. 117-135). Hierbei sind Antworten auf Fragen danach, was funktioniert bzw. wie neue Handlungsansätze funktionieren, von großem Interesse für die beteiligten Akteure. Die traditionellen Evaluationsstrategien stoßen in solchen Settings jedoch an ihre Grenzen (vgl. dazu ausführlich Giel 2013, S. 55ff.).

2. Grenzen klassischer Forschungsdesigns

Traditionell gelten in der empirischen Bildungsforschung Kontrollgruppenvergleiche als Königsweg, insbesondere dann, wenn es darum geht, die Wirkungen pädagogischen Handelns zu erfassen. Das Grundprinzip besteht darin, mögliche Einflussfaktoren zwischen der Ursache (der pädagogischen Intervention) und der daraus resultierenden Wirkung (bspw. auf Wissen und Kompetenzen) zu kontrollieren, um damit eindeutig den Einfluss der Intervention auf das Resultat zu erfassen. Dies geschieht im Idealfall durch Randomisierung, also eine zufällige Verteilung von Teilnehmenden auf Versuchs- und Kontrollgruppen.

Häufig angewendet werden auch solche Designs, bei denen (Test-)Daten zu verschiedenen Messzeitpunkten – in der Regel vor und nach einer Intervention – herangezogen werden, um Aussagen darüber zu treffen, welche Resultate mit einem neuen Bildungsangebot, einer neuen Lehrmethode etc. erzielt werden können. Beide Vorgehensweisen sind voraussetzungsvoll: Sie bedürfen ausreichend hoher Fallzahlen, bei Kontrollgruppenvergleichen eines Mindestmaßes an Einfluss auf die Gruppenbildung und grundsätzlich der Chance, bereits vor dem Beginn der Intervention zu messen.

Im zuvor geschilderten Szenario, bei dem Neuerungen in den Schulalltag eingeführt werden, ergeben sich weitere Voraussetzungen, die nur selten gegeben sind: Es muss gewährleistet sein, dass die Umsetzung plangemäß stattfindet und dass sie stabil (in vergleichbarer Weise) realisiert wird. Außerdem benötigen sowohl Kontrollgruppenvergleiche als auch bloße Vorher-Nachher-Messungen eine solide Wissensbasis darüber, welche Resultate die Neuerung überhaupt auslösen *kann*, die dementsprechend zu messen wären (vgl. Kromrey 2001).

Auch ein Vorgehen, das sich auf die Überprüfung der Zielerreichung konzentriert, weist deutliche Grenzen auf. Gerade dann, wenn sich Akteure in Schulen mit der Einführung innovativer Prozesse auf Neuland begeben, ist die Formulierung von operationalisierbaren Zielen sehr herausfordernd. Es fehlen Erfahrungen, um abschätzen zu können, was realistisch erreichbar ist und wo die Grenze zwischen Gelingen und Nicht-Gelingen verläuft. Vergleichs- bzw. Kontrollgruppendesigns wie letztlich auch zielorientierte Evaluationen müssen sich darüber hinaus mit dem sogenannten Blackbox-Problem auseinandersetzen (vgl. z.B. Pawson/Tilley 2004, S. 11): Solange ausschließlich der Input und die Resultate betrachtet werden, bleibt der Transformationsprozess dazwischen ausgeblendet. Wodurch nun genau Ergebnisse erzielt bzw. weswegen eben keine Erfolge gemessen werden können, dazu liefern diese Designs keine Hinweise. Lehrerinnen und Lehrer in der Praxis haben dazu zwar häufig Vermutungen; diese werden jedoch nicht explizit und systematisch betrachtet.

Auch die Bedingungen, unter denen Erfolge (nicht) erzielt werden, bleiben von diesen Forschungsdesigns ausgespart.

Zusammenfassend lässt sich sagen, dass die klassischen Herangehensweisen bei der Evaluation von Neuerungen zwei Begrenzungen aufweisen: Zum einen fehlt es noch an Wissen über relevante Dimensionen des Gelingens bzw. des Nicht-Gelingens, auf die man zurückgreifen kann, um dementsprechende Daten zu erheben. Zum anderen bleibt die Brücke zwischen Handeln und Resultaten genauso ausgeblendet wie der Kontext, in dem und für den die Lösungen gesucht werden. Genau hier setzt der im Folgenden vorgeschlagene Evaluationsansatz an, der sogenannte Programmtheorien in seinen Mittelpunkt stellt.

3. Auf Programmtheorien basierende Evaluationen

Theoriebasierte Evaluationskonzepte haben ihren Ursprung in der Evaluation von Programmen, also nicht in der Evaluation von Organisationen (wie bspw. Schulen) oder Produkten (wie bspw. Lehrmedien). Programme werden in der Regel mit Hilfe eines intentional aufeinander bezogenen Sets an verschiedenen Einzelprojekten umgesetzt, die wiederum aus einem Bündel einzelner Maßnahmen bestehen. Ob nun komplexe Programme, fest umrissene Maßnahmen oder einzelne Interventionen – sie werden unter spezifischen Bedingungen geplant und umgesetzt und sind auf spezifische Ziele ausgerichtet (siehe URL: http://eval-wiki.org/glossar/Programm; Zugriffsdatum: 11.04.2016). Im Kontext Schule sind aktuelle Programme beispielsweise BiSS („Bildung durch Schrift und Sprache“), das Landesprogramm „Gute gesunde Schule“ in Berlin oder „Kultur und Schule“ in Nordrhein-Westfalen. Programmtheorien – das konzeptionelle Herzstück des im Folgenden dargestellten Evaluationskonzepts – können sich auf komplexe Programme, Projekte und auch einzelne Maßnahmen beziehen.

Im Mittelprunkt der Evaluation stehen das Programm, das pädagogische Handeln, möglicherweise auch einzelne Maßnahmen des Programms (vgl. Balzer/Beywl in diesem Heft, Abschnitt 2, S. 193-197; Böttcher/Hense in diesem Heft, S. 117-135). Explizit nicht evaluiert wird die Organisation Schule. Sie stellt zwar einen Teil der Bedingungen des Programms, wie bspw. die Anzahl und Qualifikationen von Lehrerinnen und Lehrern, vorhandene Räume und deren Ausstattung, die Lehr- und Lernkultur, qualitätssichernde Verfahren etc. Daneben könnten auch der Sozialraum, in dem die Schule liegt, die Eigenschaften von Schülerinnen und Schüler und deren Familien, die eine Schule besuchen, zusätzliche finanzielle Mittel, die zur Umsetzung des Programms zur Verfügung stehen, wichtige Bedingungen des Programms bieten, die sich als hinderlich oder förderlich für die Programmumsetzung und Programmerfolge erweisen.

Es gilt zu klären, was sich hinter dem Begriff „Theorie" verbirgt. Ganz allgemein bezeichnet man einen bereits empirisch bestätigten oder noch vermuteten Zusammenhang zwischen mindestens zwei Sachverhalten als Theorie (vgl. z.B. Kromrey 2009, S. 47). In theoriebasierten Evaluationskonzepten wird der Begriff „Theorie" tatsächlich auch so breit verstanden: Er reicht von Alltagstheorien, die auf persönlichen oder professionellen Erfahrungen beruhen, bis hin zu empirisch überprüften Theorien aus der Forschung oder aus vorangegangenen Evaluationen. Der Begriff Programmtheorie lässt sich in Anlehnung an Bickman (vgl. 1987, S. 5f.) entsprechend definieren als Annahme darüber, in welcher Weise ein Programm Veränderungen herbeiführt. Eine Programmtheorie muss also mindestens geplante Interventionen eines Programms oder einer Maßnahme und die damit auszulösenden Resultate enthalten. Ergänzend sind außerdem – zumindest Pawson/Tilley (2004) betonen das – die relevanten Bedingungen (wie bspw. die Charakteristik der Zielgruppe, eingesetzte Ressourcen etc.) zu explizieren.

Ein auf Programmtheorien basierendes Evaluationskonzept verfolgt das Anliegen, Licht in die Blackbox zu werfen, die traditionelle Evaluationskonzepte oftmals hinterlassen. Die klassischen Fragestellungen von Evaluationen „Wirken Programme?" oder „Erreichen Programme ihre Ziele?" erhalten so einen neuen Akzent. Mit diesem Konzept rücken Fragen danach, *wie* Programme wirken, wie sie ihre Ziele erreichen, ins Zentrum der Aufmerksamkeit. Patricia Rogers (2000), eine der wichtigen Protagonistinnen des theoriebasierten Evaluationsansatzes, überschreibt sehr treffend einen ihrer Artikel mit: „Program Theory. Not Whether Programs Work, but How They Work".

Was bedeutet dies nun für die Evaluation? Zunächst einmal ist festzustellen, dass jede Evaluation über einen theoretischen Kern verfügt. In theoriebasierten Evaluationen wird genau dieser theoretische Kern explizit gemacht und systematisch herausgearbeitet. Dabei werden die Annahmen, die der Evaluation und dem Programm zugrunde liegen, aus dem Programm selbst abgeleitet und nicht – zumindest nicht nur – aus einer wissenschaftlichen Perspektive an das Programm herangetragen. Dadurch wird gewährleistet, dass die Evaluation nicht abstrakt bleibt, sondern die Praxis des Programms miteinbezieht. Die explizierten Programmtheorien bilden die Grundlage für die Konzipierung, das Design und die Durchführung der Evaluation; sie werden in der Datenauswertung und -interpretation genutzt und bilden damit die Basis für die Bewertungen, die jede Evaluation vorzunehmen hat (vgl. Coryn et al. 2011, S. 201; Weiss 1995).

Mögliche Quellen, die Evaluatorinnen und Evaluatoren zum Erschließen von Programmtheorien nutzen können, sind u.a. Leitlinien und schriftlich niedergelegte Konzepte von Programmen, Projekten und Maßnahmen. Da Programmdurchführende auf der Grundlage ihrer persönlichen Erfahrungen und ihrer fachlichen Annahmen darüber handeln, wie ihre Aktivitäten Veränderungen oder

Stabilisierungen herbeiführen, ist es notwendig, auch diese Erfahrungen und Annahmen herauszuarbeiten. Mit Hilfe von Beobachtungen und Befragungen können solche vorwiegend impliziten Programmtheorien rekonstruiert werden (vgl. z.B. Haubrich 2009; Giel 2015).

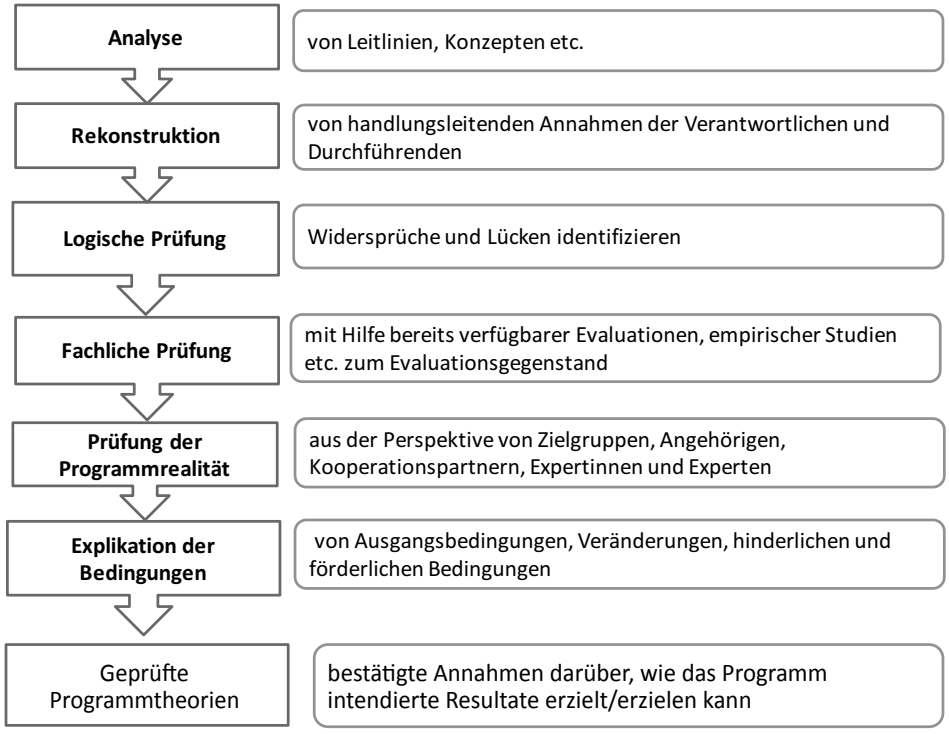
Natürlich besteht die Evaluation nicht nur darin, die das Programm strukturierenden und konturierenden Annahmen sichtbar zu machen, sondern diese auch zu prüfen. Zunächst ist eine logische Prüfung durchzuführen, die Widersprüche und vor allem Lücken identifiziert. Eine weitere Überprüfung aus externer Sicht erfolgt durch einen Abgleich mit zugänglichen, für das Programm relevanten Forschungs- und Evaluationsergebnissen. Schließlich gilt es, die Annahmen aus dem Programm und über das Programm mit der Wirklichkeit des Programms abzugleichen. Dabei wird etwa folgenden Fragen nachgegangen: Wird das Programm umgesetzt wie geplant? Werden mit den umgesetzten Aktivitäten die intendierten Resultate erzielt? Unter welchen Bedingungen lassen sich mit spezifischen Interventionen die intendierten Resultate erreichen, unter welchen Bedingungen ist das nicht der Fall?

Hierbei steht Evaluatorinnen und Evaluatoren das gesamte Repertoire an Methoden der Daten- und Informationsbeschaffung zur Verfügung: Befragungen, Beobachtungen, Tests, Sekundärdatenanalysen. Ob die Daten in eher standardisierter oder offener Weise erhoben werden, ist dabei von den Gegebenheiten im Programm, dem Auftrag an die Evaluation und nicht zuletzt von den Programmtheorien abhängig. Einen Überblick über die Bestandteile einer auf Programmtheorien basierenden Evaluation gibt Abbildung 1 auf S. 155.

Auch wenn dieses Schema eine stringente Abfolge suggeriert, so ist in der Evaluationspraxis, gerade bei Programmen, die sich noch in der Entwicklung befinden, von einer flexiblen Umsetzung auszugehen. Eine strikte Unterteilung in zunächst die Explizierung bzw. Rekonstruktion der Programmtheorie, dann deren Überprüfung ist dem Evaluationsgegenstand nicht angemessen. Wenn Programme noch experimentieren und Suchbewegungen nach passenden Antworten unternehmen, dann empfiehlt es sich, die Rekonstruktion der Programmtheorie mit deren Überprüfung zu verschränken. Damit bleibt Spielraum für Programmverbesserungen, -anpassungen und -weiterentwicklungen – bereits im Verlauf der Evaluation.

Ein Vorgehen, bei dem Erhebungsergebnisse mit den Programmbeteiligten kontinuierlich geteilt werden, ermöglicht nicht nur einen hohen Nutzen für das Programm selbst. Evaluatorinnen und Evaluatoren erhalten in diesem Rahmen immer wieder neue Informationen über die tatsächliche Umsetzung und die damit gemachten Erfahrungen. Diese Daten werden dann systematisch für die Weiterentwicklung und Überprüfung der Programmtheorie genutzt.

Abb. 1: Bestandteile einer Evaluation, die auf Programmtheorien basiert



Quelle: eigene Darstellung

Zusammenfassend lässt sich das hier umrissene Evaluationskonzept dadurch charakterisieren, dass Programmtheorien den Dreh- und Angelpunkt von Evaluation bilden. Sie rahmen den gesamten Prozess und bestimmen, welche Aspekte empirisch zu untersuchen sind. Dabei gibt das Programm den Takt vor und nicht von außen an das Programm herangetragene Theorien oder ein wenig flexibles Untersuchungsdesign, wie der zu Beginn genannte randomisierte Kontrollgruppenvergleich.

4. Programmtheorien bei Evaluationen am Beispiel von „Willkommensklassen“

Die im vorangegangenen Abschnitt aufgezeigten idealtypischen Schritte sollen im Folgenden an einem hypothetischen Beispiel illustriert werden. Im Mittelpunkt stehen hierbei:

- mögliche Fragestellungen von auf Programmtheorien basierenden Evaluationen im Kontext Schule,
- mögliche Quellen für Programmtheorien,

- Möglichkeiten zur Überprüfung der Programmtheorien anhand der Programmwirklichkeit und
- Strategien zur Bewertung von Programmen entlang der Programmtheorie.

Die Wahl für das Beispiel fällt auf die Integration von geflüchteten Kindern in die Regelschulen. In Berlin werden hierzu „Lerngruppen für Neuzugänge ohne Deutschkenntnisse" eingerichtet, die auch „Willkommensklassen" genannt werden; in Bayern spricht man von „Übergangs-" und in Nordrhein-Westfalen von „Seiteneinsteigerklassen". Angenommen wird eine Evaluation in Berlin, die den Auftrag erhält, Aufschluss darüber zu geben, wie das Lernangebot zu gestalten ist, damit die Integration in die Regelschule gut gelingt.

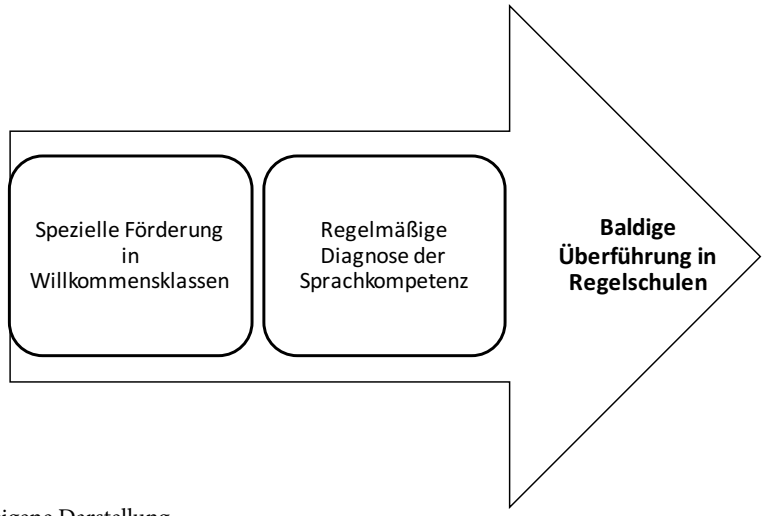
Zu berücksichtigen ist, dass Handlungsdruck besteht und dass Entscheidungen auf der Basis von wenig gesicherten Informationen getroffen werden müssen. In dieser Situation ist eine Evaluation, die eine gute Praxis identifizieren und beschreiben kann, um Entscheiderinnen und Entscheider sowie Lehrkräfte mit nützlichen Informationen und Anregungen für eine gelingende Umsetzung zu versorgen, angemessen. Diesen Bedarf deckt eine Evaluation, die Programmtheorien in den Mittelpunkt stellt, indem sie den Fokus darauf legt, *wie* überhaupt die Vermittlung von sprachlichen Kompetenzen gelingen bzw. *wie* den geflüchteten Kindern die Integration in Klassen der Regelschule ermöglicht werden kann.

Eine *erste Analyse offizieller Dokumente* untersucht die Grundannahmen des Programms. Hinweise liefert der Tagesspiegel vom 20. Juni 2015, der von einem Interview mit dem Berliner Bildungsstaatssekretär berichtet:

> „‚Ziel des Unterrichts ist die schnellstmögliche Integration der Kinder und Jugendlichen in den regulären Unterricht.' Damit das klappt, sollen die Lehrer alle paar Monate die Deutschkenntnisse der Schüler prüfen und sie je nach Wissensstand, spätestens aber nach einem Schuljahr, in die reguläre Klasse verabschieden." (URL: http://www.tagesspiegel.de/berlin/fluechtlings kinder-in-der-schule-willkommensklassen-zwischen-integration-und-abschiebung/11910496.html; Zugriffsdatum: 11.04.2016)

Im Einklang mit dieser öffentlichen Aussage steht der zentrale Handlungsleitfaden für Lehrkräfte „Von der Lerngruppe für Neuzugänge ohne Deutschkenntnisse in die Regelklasse. Ein dokumentierendes Verfahren", das vom Landesinstitut für Schule und Medien Berlin-Brandenburg (LISUM 2014) herausgegeben wird. In diesem Leitfaden liegt der zentrale Akzent auf verschiedenen Diagnoseverfahren, Fragen- und Kriterienkatalogen zur Feststellung der deutschen Sprachkompetenz von Kindern. Nach dieser – hier natürlich nur verkürzt dargestellten – Sekundärdatenanalyse könnte die Evaluation mit einem ersten Entwurf für eine noch sehr grobe Programmtheorie starten (siehe Abbildung 2).

Abb. 2: Beispiel für eine erste grobe Programmtheorie



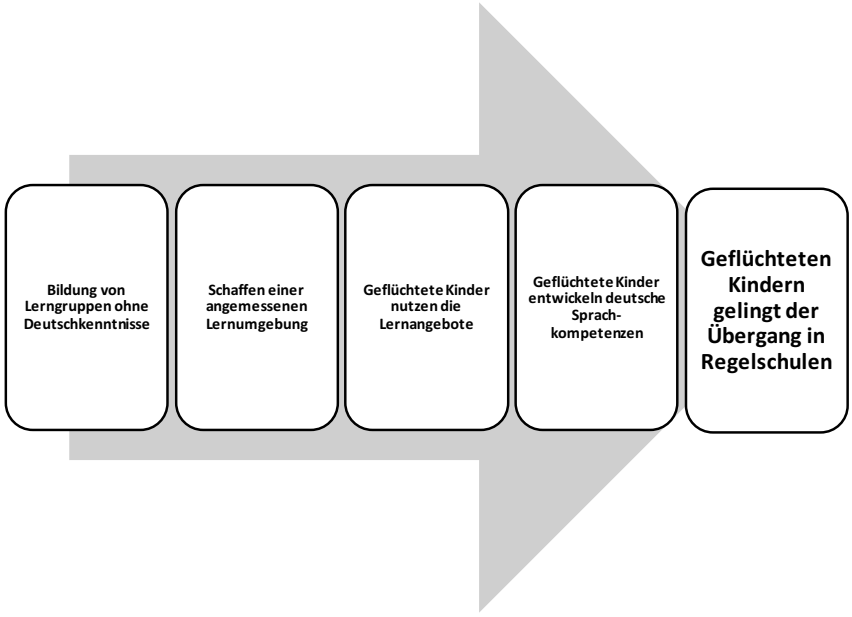
Quelle: eigene Darstellung

Es ist davon auszugehen, dass die den „Willkommensklassen" unterliegenden Programmtheorien wesentlich ausdifferenzierter sind. Bereits eine erste *logische Prüfung* deckt auf, dass der Zeitraum der Separierung der geflüchteten Kinder von Schülerinnen und Schülern der Regelschulen möglichst kurz gehalten werden soll, um möglichst frühzeitig die Integration zu starten. Die zentrale Intervention wird in unterschiedlichen Test- und Diagnoseverfahren gesehen. Damit soll gewährleistet werden, dass kein Kind zu lange in den Willkommensklassen bleibt. Jedoch wird nicht deutlich, wie die eigentliche Vermittlung von deutscher Sprachkompetenz geschehen soll. Hier ist eindeutig eine Lücke in der Programmtheorie feststellbar.

In der Praxis wird es dazu sicherlich Konzepte und Strategien geben. Zur weiteren Differenzierung der Programmtheorie müsste die Evaluation in einem nächsten Schritt *rekonstruieren*, wie in den Willkommensklassen gearbeitet wird, konzeptionell wie auch ganz konkret und praktisch. Neben einer Recherche in den Konzeptpapieren der Träger der Willkommensklassen und in den Fortbildungsangeboten des Berlin-Brandenburger Lehrerfortbildungsinstituts gilt es vor allem herauszufinden, welche Lehrmaterialen und -medien, welche Lehr- und Lernmethoden umgesetzt werden. Hier bieten sich Einzel- und Gruppenbefragungen von Lehrerinnen und Lehrern sowie teilnehmende Beobachtungen von Unterrichtssequenzen an. Es ist davon auszugehen, dass diese Daten weitere Erkenntnisse über die Aktivitäten und Wirkannahmen der Programmumsetzenden liefern, etwa dazu, *wie* die *Bildung der Lerngruppen* organisatorisch verläuft, *welche Unterrichtsmethoden* eingesetzt werden, *welche Lernziele* damit erreicht werden sollen und *wie* der *Übergang in die Regelschulen* gestaltet wird.

Diese Wirkannahmen gilt es in einem nächsten Schritt einer *fachlichen Prüfung* zu unterziehen. Hierbei bietet sich ein Rückgriff auf konstruktivistische Lerntheorien an. Aus diesen weithin akzeptierten theoretischen Grundannahmen über erfolgreiches Lernen lassen sich Prüffragen an die Wirkannahmen der Durchführenden richten: Inwieweit enthalten die eingesetzten Methoden, Materialien und Medien aktivierende Elemente, wie breit sind die Angebote an Lernmöglichkeiten? Wodurch wird der Zugang zu den Lernangeboten ermöglicht? Ausführungen hierzu müssen an dieser Stelle notwendigerweise allgemein bleiben; für eine Konkretisierung und Präzisierung wäre eine tatsächlich durchgeführte Evaluation notwendig. Zur groben Orientierung lassen sich zunächst logisch und fachlich geprüfte Annahmen darüber ableiten, wie durch die Willkommensklassen der Übergang in Regelschulen gelingen soll (vgl. Abbildung 3).

Abb. 3: Programmtheorie nach logischer und fachlicher Prüfung



Quelle: eigene Darstellung

Bislang wurden die Programmtheorien, also die Annahmen darüber, wie mit den Willkommensklassen eine Integration der geflüchteten Kinder gelingen *kann* und *soll*, aus Programmsicht und aus fachlicher, durchführender Perspektive herausgearbeitet und expliziert. Inwieweit dies in der Umsetzung gelingt, das muss nun an der *Programmrealität überprüft* werden. Neben allgemeinen Daten zur Programmumsetzung (z.B. wie viele/welche Lehrkräfte im Einsatz sind, wie viele/welche Kinder an den Willkommensklassen teilnehmen, wie viele/welche Kinder in Regelschulen wechseln) sind die Annahmen unbedingt auch aus Zielgruppensicht zu überprüfen.

Die Überprüfung aus Zielgruppensicht stellt darauf ab, wie die Lernangebote von den geflüchteten Kindern aufgenommen werden, wie zielgruppenadäquat die Lernarrangements gestaltet sind, wie die Lernangebote von den Kindern genutzt werden. Daneben ist die Einschätzung der Kinder dazu wichtig, inwieweit sie sich auf den Übergang vorbereitet fühlen, bzw. das Urteil derjenigen, die bereits in den Regelschulen angekommen sind, wie sie den Übergang erlebt haben. Wahrscheinlich bietet es sich an, auch die Einschätzungen der Familienangehörigen bzw. des sozialen Umfelds wie auch die der Lehrerinnen und Lehrer der aufnehmenden Regelschulklassen zum Lern- und Integrationserfolg einzubeziehen. Zur Feststellung der Lernerfolge können darüber hinaus die dokumentierten Diagnoseverfahren genutzt werden.

Ein Abgleich zwischen den rekonstruierten Programmtheorien und den Daten der Überprüfung der Programmrealität aus den diversen Perspektiven gibt Aufschluss darüber, inwieweit die Annahmen als zutreffend eingeschätzt werden können. Anders ausgedrückt lassen sich bewertende Aussagen dazu ableiten, ob das pädagogische Handeln die gewünschten Resultate tatsächlich erzielt. Unter der Aufgabenstellung, gute Praxis zu identifizieren, ist es sicherlich auch notwendig, die *Bedingungen*, unter denen eine Integration (entlang der Programmtheorie) gelungen bzw. nicht gelungen ist, *auszuleuchten*. Folgende Aspekte sind neben anderen von Bedeutung:

- die Zusammensetzung der Zielgruppe, bspw. hinsichtlich Alter und Altersspanne in der Gruppe, Bildungsstand, besonderer Belastungen, der Fluktuation in der Gruppe etc.;
- zur Verfügung stehende Ressourcen: Qualifikationen der umsetzenden Lehrkräfte, Ausstattung der Räume, Ausstattung mit Lehr- und Lernmaterial;
- Kooperationen zwischen Willkommensklassen, Regelschulbetrieben und weiteren unterstützenden Angeboten.

Am Ende steht die Beschreibung gelingender Praxis, also eine Darstellung, welche konkreten pädagogischen Interventionen unter welchen ausgewiesenen Bedingungen intendierte Resultate erzielen können.

In diesem konstruierten Beispiel steckt so viel Realität, dass absehbar ist, dass in der Programmumsetzung experimentiert wird, dass viele situative Entscheidungen getroffen werden und weniger auf langfristig entwickelte Konzepte zurückgegriffen werden kann. Ebenso ist davon auszugehen, dass sich die Praxis permanent verändert, die eingesetzten Fachkräfte Erfahrungen sammeln, daraus Schlüsse ziehen und sich den veränderlichen Bedingungen anpassen. Evaluatorinnen und Evaluatoren gehen diese Bewegungen mit und nehmen kontinuierlich Bewertungen vor: Sie bewerten, ob die Programmtheorien logisch und fachlich konsistent sind, ob sich die Annahmen auch in der Praxis bewähren, und finden heraus, welche Annahmen der Programmdurchführenden in der praktischen Umsetzung scheitern oder nicht die intendierten Resultate zeigen. Daraus lassen sich jeweils Optimierungsvorschläge ableiten. Die

Evaluierenden stellen durch regelmäßigen Austausch mit Programmverantwortlichen und -umsetzenden sicher, dass die Programmtheorien auf dem aktuellen Stand sind, und die Programmakteure können kontinuierlich ihr Handeln datenbasiert verbessern.

5. Reflexion der Nutzungspotenziale von Evaluationen, die sich auf Programmtheorien stützen

Der vorangestellte Entwurf für die Umsetzung einer auf Programmtheorien basierenden Evaluation muss notwendigerweise – da es sich um ein hypothetisches Beispiel handelt – oberflächlich bleiben. Trotzdem sollte deutlich geworden sein, dass sich mit dem hier präsentierten Konzept Evaluationen deutlich von Lernstandsmessungen und in Schul-Selbstevaluationstools verbreiteten Selbsteinschätzungen unterscheiden. Das wichtigste Kennzeichen theoriebasierter Evaluationen besteht darin, dass sie nicht lediglich Resultate deskriptiv erfassen, sondern konsequent den Zusammenhang zwischen (pädagogischen) Aktivitäten und deren Resultaten untersuchen. Damit steht durchgängig die Wirkfähigkeit von Programmen auf dem Prüfstand. Ein solches Vorgehen schafft Transparenz und trägt erfahrungsgemäß dazu bei, das Wirkungspotenzial von Programmen realistischer einzuschätzen.

Die sukzessive Entwicklung und Überprüfung der Programmtheorien ermöglicht es, mit den Entwicklungen von solchen Programmen mitzugehen, die sich selbst noch auf der Suche nach Lösungen befinden, Neues ausprobieren und Umsetzungsstrategien revidieren. Eine auf Programmtheorien beruhende Evaluation ist in der Lage, diese Suchbewegungen und Entwicklungen nachzuvollziehen und sich flexibel den Bedingungen im Programm anzupassen.

Zusätzlich unterstützt die Evaluation die Entwicklung des Programms. Sie ermöglicht es, genauer zu verstehen, wieso Erfolge eintreten oder ausbleiben. Dadurch, dass die Programmtheorien nicht schlicht von Evaluierenden gesetzt, sondern im kontinuierlichen Austausch mit Programmakteuren, Beteiligten und Betroffenen entwickelt und mit externen theoretischen Wissensbeständen angereichert werden, ist ein fortwährendes Lernen abgesichert. Durch die Explikation förderlicher und womöglich auch hinderlicher Bedingungen ist sichergestellt, dass mögliche Stellschrauben identifiziert werden. Die Berücksichtigung dieser Bedingungen ermöglicht es, transferfähiges Wissen zu generieren und dieses Wissen an andere Standorte und in die Fachöffentlichkeit weiterzugeben. Unter Einbeziehung der Praxis erarbeitete Programmtheorien gewährleisten, dass sich Theorie und Praxis gegenseitig befruchten.

Das hypothetische Beispiel hat deutlich gemacht, dass eine auf Programmtheorien basierende Evaluation durchaus aufwändig und anspruchsvoll ist und somit von Schulen in Eigenverantwortung – im Sinne von individueller, kollegialer Selbstevaluation – nicht zu „stemmen" ist und dass nur wenige Schulen über die erforderlichen Kompetenzen für solche anspruchsvollen Inhouse-Evaluationen verfügen dürften (vgl. Beywl/Balzer in diesem Heft, S. 191-204). Jedoch kann der konsequente Rückbezug auf die Frage nach dem „Wie" eine wertvolle Anregung für die verschiedenen Arten schulinterner Evaluationen liefern. Der Fokus darauf, wie und mit welchen Interventionen, auf der Grundlage welcher theoretischer Wirkannahmen Lehrerinnen und Lehrer zum Lernerfolg beitragen wollen, befördert wirkungsorientiertes Handeln – und auch Evaluieren.

Literatur und Internetquelle

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

Bickman, L. (1987): The Function of Program Theory. In: Ders. (Hrsg.): Using Program Theory in Evaluation. New Directions for Program Evaluation. San Francisco, CA: Jossey-Bass, S. 5-18.

Böttcher, W./Hense, J. (2016): Evaluation im Bildungswesen – eine nicht ganz erfolgreiche Erfolgsgeschichte. In: Die Deutsche Schule 108, H. 2, S. 117-135.

Coryn, C.L./Noakes, L.A./Westine, C.D./Schröter, D.C. (2011): A Systematic Review of Theory-driven Evaluation Practice from 1990 to 2009. In: American Evaluation Association (Hrsg.): American Journal of Evaluation 32, H. 2, S. 199-226.

Giel, S. (2013): Theoriebasierte Evaluation. Konzepte und methodische Umsetzungen. Münster u.a.: Waxmann.

Giel, S. (2015): Wirkungen auf der Spur. In: Giel, S./Klockgether, K./Mäder, S.: Evaluationspraxis. Professionalisierung – Ansätze – Methoden. Münster u.a.: Waxmann. S. 111-130.

Griese, C. (2016): Evaluation in der Schule. In: Griese, C./Marburger, H./Müller, T. (Hrsg.): Bildungs- und Bildungsorganisationsevaluation. Ein Lehrbuch. Berlin: de Gruyter, S. 163-187.

Haubrich, K. (2009): Sozialpolitische Innovation ermöglichen. Die Entwicklung der rekonstruktiven Programmtheorie-Evaluation am Beispiel der Modellförderung in der Kinder- und Jugendhilfe. Evaluation innovativer multizentrischer Programme. Münster u.a.: Waxmann.

Kromrey, H. (2001): Evaluation – ein vielschichtiges Konzept. In: Sozialwissenschaften und Berufspraxis 24, H. 2, S. 105-131.

Kromrey, H. ([12]2009): Empirische Sozialforschung. Überarb. und ergänzte Aufl. Stuttgart: Lucius & Lucius.

LISUM (Hrsg.) (2014): Von der Lerngruppe für Neuzugänge ohne Deutschkenntnisse in die Regelklasse. Ein dokumentierendes Verfahren. URL: http://bildungsserver.berlin-brandenburg.de/fileadmin/bbb/unterricht/unterrichtsentwicklung/Durchgaengige_Sprachbildung/Publikationen_sprachbildung/Lerngruppe_fuer_Neuzugaenge_ges_WEB_2014_12.pdf; Zugriffsdatum: 30.12.2015.

Pawson, R./Tilley, N. ([7]2004): Realistic Evaluation. London/Thousand Oaks, CA/New Delhi: Sage.

Rogers, P.J. ([2]2000): Program Theory. Not Whether Programs Work, but How They Work. In: Stufflebeam, D.L./Madaus, G.F./Kellaghan, T. (Hrsg.): Evaluation Models. Viewpoints on Educational and Human Services Evaluation. Boston, MA/Dordrecht/London: Kluwer Academic Publishers, S. 209-232.

Weiss, C.H. (1995): Nothing As Practical As Good Theory: Exploring Theory-based Evaluation for Comprehensive Community Initiatives. In: Connell, J.P./Kubisch, A.C./Schorr, L.B./Weiss, C.H.: New Approaches to Evaluation Community Initiatives. Concepts, Methods, and Contexts. Washington: Aspen Institute.

Susanne Giel, Dr., geb. 1963, Gesellschafterin und Projektleiterin bei Univation – Institut für Evaluation GmbH.
Anschrift: Univation – Büro Berlin, Gubener Straße 25, 10243 Berlin
E-Mail: susanne.giel@univation.org

DDS – Die Deutsche Schule
108. Jahrgang 2016, Heft 2, S. 163-175

Sylvia Rahn/Sabine Gruehn/Miriam Keune/Christoph Fuhrmann

Aus Schüleraussagen lernen?! – Auf dem Weg zu einer professionellen Feedbackkultur an Schulen

Zusammenfassung
Lehrkräfte sind aufgefordert, die Qualität ihres Unterrichts mithilfe von Schülerrückmeldungen zu überprüfen. In diesem Beitrag wird diskutiert, was unter einer „professionellen Feedbackkultur" in Schulen im Hinblick auf die Verwendung des Schülerfeedbacks zur Messung von Unterrichtsqualität zu verstehen ist und auf welche Verwendungsweisen angesichts des empirischen Forschungsstands zur Validität von Schülerratings zur Unterrichtsqualität möglicherweise auch besser zu verzichten wäre.
Schlüsselwörter: Unterrichtsevaluation, Unterrichtsqualität, Schülerfeedback, Schülerratings, Validität, Biasfaktoren, Verzerrungsfaktoren

Learning from Students' Statements?! – Towards a Professional Feedback Culture at Schools

Summary
Teachers are encouraged to check the quality of their teaching by using student feedback. This paper discusses what a "professional feedback culture" in schools in the context of measuring teaching quality means. The second question is, whether some uses of student feedback are not useful in light of the empirical research literature on validity of student ratings of teaching quality.
Key words: course evaluation, course quality, student feedback, student ratings, validity, bias factors, distortion factors

1. Einleitung

Lehrkräfte sehen sich bereits seit längerem – z.B. in Empfehlungen zur internen schulischen Evaluation – dazu aufgefordert, den Wahrnehmungen und Einschätzungen ihrer Schülerinnen und Schüler zur Qualität des Unterrichts systematisch Aufmerksamkeit zu schenken, um hieraus produktive Konsequenzen für die Unterrichts- und eigene Professionalitätsentwicklung abzuleiten. Erneutes Gewicht

erhält diese seit den 1990er-Jahren immer wieder formulierte Forderung nach dem Auf- und Ausbau einer „professionellen Feedbackkultur" an Schulen etwa durch die Standards der Kultusministerkonferenz (KMK) für die bildungswissenschaftliche Lehrerbildung, in denen die Nutzung von „Rückmeldungen anderer" als Ausbildungsziel genannt ist (KMK 2014, S. 13), oder durch die breite Rezeption der Hattie-Studie (Hattie 2013). Zwar wird Schülerfeedback an einigen Schulen bereits regelmäßig und systematisch genutzt; von einer flächendeckenden und selbstverständlichen Verwendung von Schülerrückmeldungen sind wir jedoch auch nach jahrelanger Diskussion um Evaluation, Qualitätsmanagement und Qualitätsentwicklung an Schulen noch weit entfernt. Hinzu kommt, dass sich die Art und Weise, wie das Schülerfeedback erhoben, ausgewertet und interpretiert wird, im inter- wie nationalen Vergleich der Verfahren erheblich unterscheidet.

Deshalb ist es das Ziel dieses Beitrages zu konkretisieren, was mit einer „professionellen Feedbackkultur" in Schulen im Hinblick auf die Verwendung des Schülerfeedbacks zur Messung von Unterrichtsqualität im Einzelnen gemeint sein könnte und auf welche Verwendungsweisen angesichts des empirischen Forschungsstands zur Validität von Schülerratings zur Unterrichtsqualität möglicherweise auch besser zu verzichten wäre.

Zu diesem Zweck wird im Folgenden zunächst ein knapper Überblick darüber gegeben, wie Schüleraussagen derzeit in deutschen Sekundarschulen zu Evaluationszwecken erhoben und ausgewertet werden. Auf der Grundlage des internationalen Forschungsstandes sowie auf der Basis der Ergebnisse einer abgeschlossenen empirischen Untersuchung zu den „Determinanten des Schülerfeedbacks" (Rahn/Gruehn/Böttcher 2015), in der Schülerratings zur Unterrichtsqualität in verschiedenen Fächern von insgesamt 2.008 Schülerinnen und Schülern der 11. Jahrgangsstufe an 38 beruflichen Gymnasien und 10 Gesamtschulen erhoben und ausgewertet wurden, wird sodann diskutiert, wie die Validität des Schülerfeedbacks einzuschätzen ist. Der Beitrag mündet schließlich in Empfehlungen, worauf bei der praktischen Verwendung von Schülerrückmeldungen zu achten wäre.

2. Der praktische Status quo – Schülerratings zur Unterrichtsqualität als Datenquelle in der Unterrichtsforschung und in der Schul- und Unterrichtsevaluation

Schüleraussagen werden auf vielfältige Weise zur Beschreibung, Diagnose, Evaluation und Entwicklung von Schule und Unterricht verwendet. Im Kontext der Unterrichtsforschung werden Schülerratings zur Unterrichtsqualität in der Regel fachbezogen erhoben und auch so zurückgemeldet. Als ein Beispiel hierfür lässt sich die Untersuchung „Qualitätssicherung in Schule und Unterricht" (kurz: QuaSSU) anfüh-

ren, in der bayerische Hauptschul-, Realschul- und Gymnasialklassen zur Qualität ihres Mathematikunterrichts standardisiert befragt wurden (vgl. Ditton 2002). Wenn die Schülerinnen und Schüler der befragten Klassen dem zustimmten, konnten die Lehrkräfte der beteiligten Schulen Rückmeldungen zu den Ergebnissen der Schülerbefragungen erhalten. In diesen Feedbacks wurden die Mittelwerte für die jeweilige Schulklasse im Vergleich mit dem Mittelwert der Gesamtstichprobe der jeweils gleichen Schulart dargestellt.

Etwas andere Rückmeldungen erhielten die Lehrkräfte in der Studie „Deutsch Englisch Schülerleistungen International" (kurz: DESI) für die Fächer Deutsch und Englisch (vgl. Klieme 2008). In diesem Zusammenhang erhielten Lehrkräfte, falls gewünscht, von den befragten Neuntklässlerinnen und Neuntklässlern Rückmeldungen zur Klassenführung sowie zur Motivation der Jugendlichen für fachbezogene Inhalte. Die Ergebnisse wurden lediglich den Lehrkräften der jeweils untersuchten Klasse mitgeteilt, ohne einen Referenzwert einer Vergleichsgruppe anzugeben.

In der Praxis der Schul- und Unterrichtsevaluation wird die wahrgenommene Unterrichtsqualität zwar auch in der Regel fachbezogen erhoben, aber – anders als in der Unterrichtsforschung – nicht immer fachbezogen ausgewertet. In Bezug auf die Rückmeldungen des Schülerfeedbacks an die Lehrkräfte finden sich verschiedene Ansätze von der individuellen Rückmeldung der Daten der eigenen Klasse(n) an einzelne Lehrkräfte über mehrperspektivische Vergleiche innerhalb einer Schule (beides fachbezogene Auswertungen) bis hin zu Vergleichen zwischen Schulen (fachübergreifende Auswertung).

Ein Beispiel für individuelle Rückmeldungen an Lehrkräfte ist das in NRW, Sachsen und Thüringen implementierte Konzept „Schüler als Experten für Unterricht" (kurz: SEfU). Mit Hilfe von 35 Kernaussagen – basierend auf einem einschlägigen Qualitätsmodell der Schulforschung – wird die Schülersicht auf Unterricht online als „Expertenurteil" für die Unterrichtsqualität erfasst und mit der Unterrichtswahrnehmung der jeweiligen Lehrperson über ein ebenfalls online verfügbares Auswertungstool verglichen (vgl. Kämpfe 2009; ten Venne/Nachtigall/Kämpfe-Hargrave 2010). Durch die wiederholte Nutzung können Lehrende das Schülerfeedback mehrerer Lerngruppen, die sie unterrichten, im Querschnitt oder Längsschnitt vergleichen.

Ein anderes, von der Kultusministerkonferenz gefördertes Angebot ist das Programm „Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung" (kurz: EMU) (vgl. Helmke/Helmke 2014; Helmke et al. 2015). Das Programm bietet über eine Internetplattform frei zugängliche Instrumente an, mit denen Lehrkräfte die Qualität des eigenen Unterrichts mithilfe der Sichtweise eines oder mehrerer Kollegen und/oder der Schülersicht analysieren und mehrperspektivisch vergleichen können. Für die Auswertung steht ein Auswertungsprogramm zur Verfügung, das nach der

Dateneingabe die Verteilung der verschiedenen Perspektiven in Form von Profillinien anzeigt.

Während SefU und EMU stark auf den Vergleich verschiedener Perspektiven auf denselben Unterricht setzen, werden im „Netzwerk Schülerbefragung", das als drittes Beispiel genannt werden soll, der fächerübergreifende Vergleich und der Vergleich zwischen verschiedenen Schulen betont. Bei diesem Zusammenschluss aus Schulen in Berlin und Brandenburg geben Lernende ein Feedback in allen unterrichteten Fächern und in sämtlichen Bildungsgängen ab (vgl. Wagner 2010, S. 292). Im „Netzwerk Schülerbefragung" wird das Feedback der Klasse den Lehrenden im Vergleich mit dem „aggregierten Bewertungsprofil" sowohl ihrer eigenen Schule als auch aller Netzwerkschulen mitgeteilt, und das unabhängig vom jeweils unterrichteten Fach (vgl. ebd.).[1]

Neben den vorgestellten Programmen, die ausschließlich unterrichtsbezogene Schülerfeedbacks generieren, werden Schülerwahrnehmungen auch in nahezu allen auf die Schule als Organisation bezogenen Evaluationsverfahren und Qualitätsmanagementkonzepten genutzt. Dies gilt u.a. für das Konzept von „Selbstevaluation in Schulen" (kurz: SEIS) der Bertelsmann Stiftung, das Modell der „European Foundation for Quality Management" (kurz: EFQM) und die Konzepte zur Schul- und Unterrichtsentwicklung, die vom Institut für Schulentwicklungsforschung in Dortmund entwickelt wurden. Üblich ist im Zuge solcher Konzepte, dass die Schülerinnen und Schüler schriftlich zur Unterrichtsqualität an ihrer Schule befragt werden, ohne dass zwischen verschiedenen Fächern und Lehrpersonen unterschieden würde. Auf dieser Basis werden dann Schulmittelwerte berechnet, die zu den Durchschnittswerten der Unterrichtsqualität an allen Schulen derselben Schulform ins Verhältnis gesetzt werden. Die einzelne Lehrkraft bekommt keine individuelle Rückmeldung. Diese Vorgehensweise unterscheidet sich damit erheblich von den auf den Unterricht und das Feedback identifizierbarer Lerngruppen bezogenen Projekten der Unterrichtsforschung und von den Evaluationsverfahren, die das Schülerfeedback primär in den Dienst der individuellen Unterrichtsentwicklung der einzelnen Lehrkraft stellen (vgl. Thiel/Ulber 2007).

Per Saldo bietet der knappe Überblick über den Status quo zur Nutzung von Schüleraussagen in der Schul-, Unterrichts- und Lehrevaluation ein „buntes" Bild, das die Frage nach den Kriterien und Ansprüchen aufwirft, denen die Evaluationsverfahren selbst genügen sollten.

1 Zu den Aktivitäten des „Netzwerks Schülerbefragung" siehe neben den Veröffentlichungen von Wagner (2009, 2010) auch exemplarisch den Bericht eines Oberstufenzentrums. URL: http://www.peter-lenne-schule.de/images/pdfs/schuelerbefragung_2014/2014-03-26_ergebnisse_schuelerbefragung_2014.pdf; Zugriffsdatum: 28.01.2016.

3. Validität von Schülerratings zur Unterrichtsqualität im Spiegel empirischer Forschung

Was wissen wir nun aber darüber, wie valide Schülerratings im Zuge standardisierter Fragebogenerhebungen die Unterrichtsqualität messen? Zur Beantwortung dieser Frage werden im Folgenden empirische Forschungsbefunde zur – perspektivenspezifischen – Validität sowie zu den Bias- und Verzerrungsfaktoren der Schülerratings berichtet. Auf eine Darstellung der Ergebnisse zur prognostischen Validität, d.h. zur Vorhersagekraft der Schülerfeedbacks für zukünftige Schülerleistungen, wird hier – angesichts des begrenzten Raums – verzichtet, da diese Fragestellung zwar die Bildungsforschung durchaus beschäftigt hat (vgl. u.a. Clausen 2002; Gruehn 2000, S. 199; Kunter 2005; Lüdtke et al. 2006), in der schulischen Evaluationspraxis aber kaum eine Rolle spielt.

3.1 Validität des Schülerfeedbacks

Um einschätzen zu können, wie zuverlässig und valide die Prozessqualität des Unterrichts mithilfe von Schülerbefragungen erfasst werden kann, wird untersucht, wie gut das Schülerfeedback mit der Unterrichtswahrnehmung und -beurteilung von Lehrkräften einerseits und der von externen Beobachtern und Beobachterinnen andererseits übereinstimmt bzw. worin sich die drei Perspektiven jeweils unterscheiden. Der Perspektivenvergleich zeigt, dass die Ratings zwischen Schülerinnen und Schülern sowie externen Beobachterinnen und Beobachtern erheblich korrelieren, dass sie aber kein perspektivenübergreifendes Messmodell zur Unterrichtsqualität darstellen (vgl. Clausen 2002; Greimel 2002). Zwischen den Schüler- und Lehrkräfteeinschätzungen bestehen hingegen eher geringe Übereinstimmungen (vgl. Clausen 2002; Baumert/Kunter 2006), was u.a. daran liegt, dass die Angaben von Lehrkräften vielfach Selbstbeurteilungen darstellen und deshalb selbstwertdienlichen Verzerrungen unterliegen. Parallelen bestehen daher nur bei Unterrichtsmerkmalen, die eine längere gemeinsame Erfahrung unterrichtlicher Interaktion und Kommunikation voraussetzen wie etwa bei der Klassenführung (vgl. Clausen 2002; Baumert/Kunter 2006). Dies wurde unlängst auch in unserer einleitend erwähnten fächervergleichenden Untersuchung in der gymnasialen Oberstufe bestätigt: Aspekte der Klassenführung werden sowohl perspektiven- als auch fachübergreifend einheitlich wahrgenommen.

Für die meisten Unterrichtsmerkmale gilt, dass ihre Erfassung und Beurteilung nur perspektivenspezifisch valide möglich ist, wobei keine der drei Perspektiven grundsätzlich näher an der „wahren" Unterrichtsqualität ist als die anderen. Für die perspektivenspezifische Validität des Schülerfeedbacks lässt sich der Forschungsstand dahingehend zusammenfassen, dass Unterrichtsbeschreibungen aus Schülersicht undifferenzierter ausfallen als dies wünschenswert wäre (vgl. Gruehn 2000, S. 142ff.;

Wagner 2010, S. 292). Überdies unterliegt das Schülerfeedback einem so genannten „affektiven Haloeffekt", d.h., der Gesamteindruck, den die Befragten vom Unterricht haben, wirkt sich auch auf die Einschätzung seiner einzelnen Prozessmerkmale aus (vgl. Clausen 2002; Wagner 2008; Wagner 2010). Insgesamt führt all dies jedoch nicht dazu, dass das Schülerfeedback es nicht erlauben würde, bei der Beschreibung und Bewertung der Unterrichtsqualität deutlich zwischen Unterrichtsmerkmalen zu unterscheiden. Im Gegenteil: Stärken- und Schwächenprofile sind im Mittelwertvergleich und in den sie visualisierenden Profillinien deutlich zu erkennen.

3.2 Der Einfluss von Bias- und Verzerrungsfaktoren

Zur Analyse der Validität des Schülerfeedbacks wurde zudem intensiv der Frage nachgegangen, inwieweit die Schülerratings durch Merkmale beeinflusst und verzerrt sind, die selbst nicht Ausdruck der Unterrichtsqualität sind (echte Biasfaktoren) oder aber sich der direkten Beeinflussung durch die Lehrperson entziehen (Fairnessvariablen) (vgl. Rindermann 2001). Bei der Beantwortung dieser Frage ist die Unterrichtsforschung in weiten Teilen der Forschung zur Evaluation der Hochschullehre gefolgt. Bislang wurde empirisch vor allem geprüft, inwieweit das Geschlecht der Lernenden, das Geschlecht der Lehrperson oder die Interaktion zwischen beiden, die Leistungsstände der Schülerinnen und Schüler und die Benotungspraxis der Lehrkräfte, die Sympathie mit der Lehrkraft sowie das Fachinteresse der Lernenden mit dem Schülerfeedback korrelieren.

Hinsichtlich der Bedeutung des Geschlechts der Lehrenden bzw. der Lernenden und der Interaktion zwischen beiden bleibt der Forschungsstand sowohl für den schulischen als auch für den hochschulischen Lehr-Lern-Kontext uneindeutig (vgl. Greimel-Fuhrmann/Geyer 2005, S. 115; Gruehn 2000, S. 78). Der Gedanke liegt nahe, dass die Geschlechtereffekte durch das mit dem Geschlecht der Lernenden kovariierende Interesse am Unterrichtsfach zustande kommen und nicht als Ausdruck allgemeingültiger geschlechtsspezifischer Beurteilungstendenzen zu verstehen sind. Genau dies war in der Untersuchung zu den Determinanten des Schülerfeedbacks in vier der fünf einbezogenen Unterrichtsfächer der Fall. In Deutsch, BWL/Rechnungswesen, Erziehungswissenschaft und Elektrotechnik ließ sich bei statistischer Kontrolle weiterer Schülermerkmale kein Geschlechtereffekt bei der Unterrichtsbeurteilung feststellen. Nur im Fach Mathematik zeigte sich trotz Berücksichtigung der Lernmotivation und des Fachinteresses der Schülerinnen und Schüler noch ein schwacher signifikanter Geschlechtereffekt.

Auch die Befürchtung, dass die Schüler- und Studierendenwahrnehmungen durch die erwarteten und faktisch erhaltenen Noten verzerrt werden, kann weder klar entkräftet noch zweifelsfrei empirisch gestützt werden (vgl. Aleamoni 1999; Böttcher/Grewe 2010; Rindermann 2001). Insbesondere die Geltung der so genannten „Grading-

Leniency-Hypothese", wonach sich eine relativ zu den faktischen Leistungen milde Beurteilung der Schülerleistungen in einem ebenfalls günstigen Schülerfeedback niederschlage (vgl. Marsh/Roche 1997; Wagner 2008, S. 23 und S. 124), bedarf noch eindeutiger empirischer Belege.

Klarer ist der Forschungsstand zum Zusammenhang zwischen Sympathie und Schülerfeedback. Er wurde sowohl für die Beurteilung der Lehrqualität im Fach Rechnungswesen an österreichischen Handelsakademien (vgl. Greimel-Fuhrmann/Geyer 2005) als auch am Beispiel der DESI-Daten für die Fächer Deutsch und Englisch belegt. In unserer Studie zum Schülerfeedback in der gymnasialen Oberstufe war der Zusammenhang zwischen der Sympathie mit der Lehrkraft und der wahrgenommenen Unterrichtsqualität in allen fünf Fächern signifikant und mit ähnlicher Effektstärke nachweisbar.

Besonders wichtig ist für das Schülerfeedback zur Unterrichtsqualität, welches Interesse die Schülerinnen und Schüler dem Unterrichtsfach entgegenbringen. In der Forschung zur Evaluation der Hochschullehre ist der Zusammenhang zwischen dem Interesse am Lehrinhalt und der wahrgenommenen Lehrqualität gut belegt (vgl. Spiel/Gössler 2000; Rindermann 2001, S. 179ff.). Dabei variiert der Stellenwert, den das Interesse für die studentischen Einschätzungen hat, nach Veranstaltungsart und Veranstaltungskontext (vgl. El-Hage 1996).

Damit ließe sich auch für schulische Kontexte erwarten, dass sich das Fachinteresse der Lernenden nicht nur auf das Schülerfeedback auswirkt, sondern dass es dies in Abhängigkeit vom Unterrichtsfach mit unterschiedlicher Stärke tut. Zwar konnte Ditton (vgl. 2002, S. 280) für den Mathematikunterricht an bayerischen Schulen der Klasse 9 mit Hilfe von Mehrebenenanalysen zeigen, dass sich die Interessantheit und Wichtigkeit des Faches nicht mehr verzerrend auf die wahrgenommene Unterrichtsqualität auswirkt, wenn statt der individuellen Werte die Klassenmittelwerte verwendet werden – sich durch die Aggregation der Daten also der Verzerrungseinfluss ausmittelt. Bei fächerübergreifenden Stichproben kann dieser Effekt allerdings nicht ohne weiteres vorausgesetzt werden. Denn nur wenn die Verzerrungsvariablen innerhalb der Klassen normalverteilt sind, sind Individualdaten verzerrt, Klassenmittelwerte aber nicht.

Dies ist im Falle von fächer- und bildungsgangübergreifenden Stichproben, wie sie in der Unterrichtsforschung selten, aber in Qualitätsmanagementverfahren häufig vorkommen, für den Einfluss des Fachinteresses aber nicht zu erwarten. Mit andauerndem Schulbesuch entwickeln Jugendliche ein sich zunehmend ausdifferenzierendes Interessenprofil. Erwartungsgemäß konnten deshalb in dem Projekt zu den Determinanten des Schülerfeedbacks in der Oberstufe im berufsbildenden Zweig ein signifikanter Effekt des Unterrichtsfachs auf das Fachinteresse und ein Interaktionseffekt mit dem besuchten Gymnasialzweig nachgewiesen werden. Im

Einzelnen bedeutet das, dass das Fachinteresse der Lernenden in den profilbildenden Fächern der beruflichen Gymnasien höher ist als in Mathematik oder Deutsch und dass sich Lehrende, die Mathematik im pädagogischen Gymnasialzweig oder Deutsch im technischen Gymnasium unterrichten, mit einem geringen durchschnittlichen Fachinteresse konfrontiert sehen. In der Folge konnte in der Oberstufenstudie neben dem Einfluss des individuellen Schülerinteresses auch ein signifikanter Effekt des Unterrichtsfachs als Klassenmerkmal auf die Gesamtbeurteilung des Unterrichts durch die Lernenden belegt werden.

Überdies wirkten sich nicht nur einzelne potenzielle Verzerrungsmerkmale, sondern auch die Unterrichtsmerkmale im Fächervergleich verschieden auf die wahrgenommene allgemeine Unterrichtsqualität aus, so dass die Frage, ob die Determinanten des Schülerfeedbacks über die Fächer hinweg vergleichbar sind (vgl. Ditton 2002, S. 282; Greimel-Fuhrmann 2003, S. 261), für die gymnasiale Oberstufe zu verneinen ist. Auch in DESI erwies sich das Schülerfeedback nicht als fachübergreifend valide (vgl. Wagner 2008, S. 131).

Somit spricht der Forschungsstand zwar insgesamt dafür, dass Schülerinnen und Schüler „durchaus" als „Auskunftsgeber über unterrichtliche Bedingungen und Lernarrangements fungieren" können (Gruehn 2000, S. 211; vgl. hierzu u.a. Baumert et al. 2004; Greimel-Fuhrmann 2003; Wittwer 2008), dass bei der Rezeption und Verwendung der Daten aber auch die skizzierten Einschränkungen bedacht sein sollten.

4. Kriterien und Empfehlungen zur Verwendung von Schülerratings an Schulen

Eine schulische Feedbackkultur, deren Ziel es ist, dass Lehrpersonen aus den Wahrnehmungen und Urteilen ihrer Schülerinnen und Schüler etwas für die eigene Professionalitätsentwicklung lernen können, setzt zweierlei voraus: Erstens sollten die Daten das gesamte Spektrum der Wahrnehmungen und Einschätzungen der Lerngruppe zeigen und zweitens ein sachlich hinreichend differenziertes Meinungsbild der Lernenden darstellen. Damit man ein Meinungsbild erhält, das die gesamte Verteilung der Schülerwahrnehmungen und Akzeptanzurteile in einer Klasse abbildet, sollten die Schülerinnen und Schüler ihre Wahrnehmungen niedrigschwellig äußern können, was sich durch leichte Zugänglichkeit und einfache Handhabbarkeit des Erhebungsinstruments sowie Gewährleistung der Anonymität der Lernenden erreichen lässt. Die eingangs beschriebenen Feedback-Instrumente zur individuellen Rückmeldung an Lehrkräfte erfüllen diese vier Anforderungen (vollständig hinsichtlich der Verteilung innerhalb der Lerngruppe, hinreichend differenziert, niedrig-

schwellig und anonym); allerdings kommt eine solche Feedbackkultur nicht ohne periodisch wiederkehrende standardisierte Fragebogenerhebungen aus.

Während Schülerinnen und Schüler die Möglichkeit haben müssen, ihre Unterrichtseinschätzungen „ohne Ansehen ihrer Person" abgeben zu können, kann dies für Lehrkräfte nicht gelten. Die Feedbackkultur an Schulen sollte sich zwar auch, aber nicht nur an das gesamte Kollegium als so genannte „Professionellengemeinschaft" richten. Auch die einzelne Lehrkraft als verantwortliche Akteurin/verantwortlicher Akteur ist anzusprechen. Wenn Schülerurteile zu Evaluationszwecken erhoben und ausgewertet werden, sollte mit anderen Worten eindeutig identifizierbar sein, welches Unterrichtsangebot und welche einzelne Lehrkraft konkret Gegenstand der Wahrnehmung und Beurteilung ist. Denn dies bildet die Grundlage dafür, dass Lehrende ein individuelles Stärken- und Schwächenprofil ihrer eigenen Instruktionsqualität erhalten.

Dies schließt die spätere Aggregation der Daten auf Abteilungs- und/oder Schulebene zwar nicht aus, aber ein rein organisationsbezogenes Qualitätsmanagement, das ausschließlich mit solchen aggregierten Daten arbeitet, erfüllt das Kriterium nicht. Eine solche Praxis droht stattdessen, den individuellen Unterricht als entscheidende Ebene der Leistungserbringung in den Schulen zu verfehlen. Sie bringt – ähnlich wie Ansätze der Klimaforschung – die einzelne Lehrkraft als Adressaten des Schülerfeedbacks gewissermaßen zum Verschwinden und erschwert oder verhindert sogar, dass die einzelne Lehrperson Konsequenzen für die eigene Professionalitätsentwicklung ziehen kann. Qualitätsmanagementverfahren, die stets die gesamte Schule in den Blick nehmen, können so zwar etwaigen Widerständen entgehen, die aus der Angst vor dem personenbezogenen Vergleich resultieren, verschenken aber auch die Chancen eines solchen Vergleichs und drohen damit notorisch ihren eigenen Anspruch auf Unterrichtsentwicklung zu verfehlen (vgl. Rahn 2008).

Der Entwicklung des Unterrichts dient es hingegen, wenn Lehrende die Chance haben, ein Schülerfeedback-basiertes Stärken-Schwächen-Profil und damit eine Rückmeldung mit individueller Bezugsnorm zu erhalten. Dass Leistungsrückmeldungen, in denen die Leistungen einer Person aufgrund der eigenen Leistungsentwicklung beurteilt werden, günstig für die Lern- und Leistungsmotivation von Lernenden sind, ist in der pädagogischen Psychologie gut belegt (vgl. Rheinberg 2001). Somit liegt die Hypothese nahe, dass auch die Lern- und Professionalitätsentwicklungsprozesse von Lehrpersonen bei der Rückmeldung von Schülerurteilen mit einer individuellen Bezugsnorm unterstützt werden können. Es wäre zu erwarten, dass sich die Selbstwirksamkeitserwartungen, d.h. die Überzeugung, aufgrund eigener Lehrkompetenz gute Lehrangebote erzeugen zu können, von Lehrkräften, die zunehmend besseres Schülerfeedback erhalten, positiv entwickeln.

Schülerfeedback birgt für die Lehrenden aber auch die Chance, das eigene Fähigkeitsselbstbild mit Hilfe der Schüleraussagen zu validieren und durch den sozialen Vergleich mit anderen Lehrkräften auf seinen Realitätsgehalt zu prüfen. Dies setzt allerdings voraus, dass in den Evaluationsverfahren ein Maßstab für solche sozialen Vergleiche, d.h., eine soziale Bezugsnorm zur Verfügung steht, was – wie oben gezeigt – nicht immer üblich ist. Wenn z.B. die Schülerinnen und Schüler einer Klasse das Lehrangebot einer Lehrperson hinsichtlich der Klarheit der Unterrichtssprache auf einer sechsstufigen, an der Notenfolge angelehnten Skala mit dem Mittelwert 2,8 bewerten – wie hat man diesen Wert dann zu interpretieren? Erst der soziale Vergleich zwischen Rückmeldungen verschiedener Lerngruppen und verschiedenen Lehrpersonen erlaubt es, einen konkreten Wert als relativ gut oder schlecht zu verorten. Insofern ist fraglich, ob eine professionelle Evaluationskultur an Schulen tatsächlich auf den sozialen Vergleich verzichten kann (siehe hierzu Thiel/Ulber 2007). Bekanntlich sind solche sozialen Vergleiche in der Leistungsbeurteilung einerseits wichtig, um eine realistische Rückmeldung über den eigenen Leistungsstand zu erhalten und einen angemessenen sachlichen Leistungsstandard zu justieren. Andererseits stoßen sie aber auch auf erheblichen Widerstand bei den Lehrenden und bergen Risiken für das Fähigkeitsselbstbild, d.h., auch für die Lern- und Leistungsmotivation des pädagogischen Personals.

Wenn aber verglichen wird – das gilt nicht nur für den Vergleich von Schülerleistungen –, dann muss der Vergleich fair sein. Und dies bedeutet, dass er – angesichts der empirischen Belege – nur innerhalb der einzelnen Unterrichtsfächer vorgenommen werden sollte. Andernfalls besteht die Gefahr, dass man die Leistungen von Lehrenden, die ihre Arbeit unter schwierigen Bedingungen erbringen müssen, systematisch unterschätzt und diejenigen zusätzlich belobigt, die bereits von guten Arbeitsbedingungen wie etwa einem hohen Fachinteresse der Lerngruppe profitieren.

5. Schlussbetrachtung und Ausblick

Schülerfeedback – so ist zu resümieren – sollte für die Lernenden anonym, hinreichend differenziert und mit Hilfe sorgfältig entwickelter Instrumente erhoben werden. Geeignete Verfahren stehen hierzu zur Verfügung. Individuelle Vergleiche sollten durch – faire – soziale Vergleiche, die aufgrund der oben skizzierten Befunde *innerhalb* von Unterrichtsfächern, aber nicht zwischen Unterrichtsfächern möglich sind, ergänzt werden können. Dies legt allerdings Modifikationen in den Rückmeldeverfahren einiger der bis dato implementierten Instrumente nahe.

Mehrheitlich – darauf ist abschließend ausdrücklich hinzuweisen – birgt das Schülerfeedback für die einzelne Lehrkraft, zumindest was die allgemeine Unterrichtseinschätzung betrifft, keine Überraschungen. Wen sollte das auch wundern:

Lehrpersonen kennen ihre Klassen, so dass der Vergleich zwischen dem tatsächlichen und dem von der Lehrkraft erwarteten Feedback etwa in der Studie zu den Determinanten des Schülerfeedbacks in der Oberstufe in 70 Prozent der Fälle nur geringe Abweichungen zeigte. Rund 30 Prozent der Lehrenden erlebten allerdings eine Überraschung, darunter ca. 12 Prozent eine positive, d.h., die Klasse beurteilte die Unterrichtsqualität besser als erwartet. Bedenkt man zudem, und dies sei an dieser Stelle ebenfalls betont, dass die Einschätzungen der Schülerinnen und Schüler zur Qualität der Lehrerarbeit und des Unterrichts ohnehin überwiegend positiv, die Lernenden mit der Unterrichtsqualität also zufrieden sind, dann stellt die Erhebung von Schülerfeedback eine der Gelegenheiten dar, bei denen sich Lehrkräfte in ihrem Beruf als erfolgreich erleben können. In einem Beruf, der im Ruf steht, durch ein notorisches Defizit an sozialer Anerkennung gekennzeichnet zu sein, könnte das ein nicht zu unterschätzender Gewinn sein.

Wenn das Schülerfeedback aber, wie in der Oberstufenstudie bei rund 17 Prozent der Klassen, unerwartet negativ ausfällt, die Klassen der Lehrperson also einen Änderungsbedarf signalisieren, den sie selbst nicht wahrgenommen hat, kann dies im besten Fall den Impuls für einen Entwicklungsprozess geben – nicht mehr, aber auch nicht weniger.

Literatur und Internetquellen

Aleamoni, L.M. (1999): Student Rating Myths versus Research Facts from 1924 to 1998. In: Journal of Personnel Evaluation in Education 13, H. 2, S. 153-166.

Baumert, J./Kunter, M. (2006): Stichwort: Professionelle Kompetenz von Lehrkräften. In: Zeitschrift für Erziehungswissenschaft 9, H. 4, S. 469-520.

Baumert, J./Kunter, M./Brunner, M./Krauss, S./Blum, W./Neubrand, M. (2004): Mathematikunterricht aus Sicht der PISA-Schülerinnen und -Schüler und ihrer Lehrkräfte. In: Prenzel, M./Baumert, J./Blum, W./Lehmann, R./Leutner, D./Neubrand, M./Pekrun, R./Rolff, H.-G./Rost, J./Schiefele, U. (Hrsg.): PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs. Münster u.a.: Waxmann, S. 314-354.

Böttcher, W./Grewe, C.M. (2010): Eine Untersuchung zur Wirksamkeit der studentischen Lehrveranstaltungskritik am Beispiel der Westfälischen Wilhelms-Universität Münster. In: Pohlenz, P./Oppermann, A. (Hrsg.): Lehre und Studium professionell evaluieren: Wie viel Wissenschaft braucht die Evaluation? Bielefeld: Universitäts Verlag Webler, S. 73-82.

Clausen, M. (2002): Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität. Münster u.a.: Waxmann.

Ditton, H. (2002): Lehrkräfte und Unterricht aus Schülersicht. Ergebnisse einer Untersuchung im Fach Mathematik. In: Zeitschrift für Pädagogik 48, H. 2, S. 262-286.

El-Hage, N. (1996): Lehrevaluation und studentische Veranstaltungskritik. Projekte, Instrumente und Grundlagen. Bonn: Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie.

Greimel, B. (2002): Lehrerevaluation durch Beurteilung der Lernenden – eine Analyse des Standes der Evaluationsforschung. In: Zeitschrift für Berufs- und Wirtschaftspädagogik 98, H. 2, S. 197-224.

Greimel-Fuhrmann, B. (2003): Evaluation von Lehrerinnen und Lehrern. Einflussgrößen auf das Gesamturteil von Lernenden. Innsbruck/Wien/München: Studienverlag.

Greimel-Fuhrmann, B./Geyer, A. (2005): Die Wirkung von Interesse und Sympathie auf die Gesamtbeurteilung in der Lehrevaluation. Direkte und indirekte Effekte unter Berücksichtigung des Lehrverhaltens. In: Empirische Pädagogik 19, H. 2, S. 103-120.

Gruehn, S. (2000): Unterricht und schulisches Lernen. Schüler als Quellen der Unterrichtsbeschreibung. Münster u.a.: Waxmann.

Hattie, J. (2013): Lernen sichtbar machen. Überarb. deutschsprachige Ausgabe von „Visible Learning", besorgt von Wolfgang Beywl und Klaus Zierer. Baltmannsweiler: Schneider Verlag Hohengehren.

Helmke, A./Helmke, T. (2014): Unterrichtsanalyse mit EMU (Evidenzbasierte Methoden der Unterrichtsentwicklung). In: Journal für Schulentwicklung 18, H. 1, S. 55-57.

Helmke, A./Helmke, T./Lenske, G./Pham, G./Praetorius, A.-K./Schrader, F.-W./Ade-Thurow, M. (2015): Evidenzbasierte Methoden der Unterrichtsdiagnostik und -entwicklung. Version 5.0 (15.01.2015). URL: http://unterrichtsdiagnostik.info/; Zugriffsdatum: 28.01.2016.

Kämpfe, N. (2009): Schülerinnen und Schüler als Experten für Unterricht. In: Die Deutsche Schule 101, H. 2, S. 149-163.

Klieme, E. (2008) (Hrsg.): Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie. Weinheim u.a.: Beltz.

KMK (Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland) (2014): Standards für die Lehrerbildung: Bildungswissenschaften (Beschluss der Kultusministerkonferenz vom 16.12.2004 i.d.F. vom 12.06.2014). URL: http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung-Bildungswissenschaften.pdf; Zugriffsdatum: 28.01.2016.

Kunter, M. (2005): Multiple Ziele im Mathematikunterricht. Münster u.a.: Waxmann.

Lüdtke, O./Trautwein, U./Kunter, M./Baumert, J. (2006): Analyse von Lernumwelten. Ansätze zur Bestimmung der Reliabilität und Übereinstimmung von Schülerwahrnehmungen. In: Zeitschrift für pädagogische Psychologie 20, H. 1, S. 85-96.

Marsh, H.W./Roche, L.A. (1997): Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility. In: American Psychologist 52, H. 11, S. 1187-1197.

Rahn, S. (2008): Qualität, Qualitätssicherung und Qualitätsentwicklung im beruflichen Schulwesen – Charakteristika und Reichweite aktueller Konzepte. In: bwp Spezial Hochschultage Berufliche Bildung. URL: http://www.bwpat.de/ht2008/ft06/rahn_ft06-ht2008_spezial4.pdf; Zugriffsdatum: 28.01.2016.

Rahn, S./Gruehn, S./Böttcher, W. (2015): Determinanten des Schülerfeedbacks: eine fächervergleichende Analyse von Schüleraussagen zur Unterrichtsqualität an allgemeinbildenden und beruflichen Gymnasien. Unveröffentlichter Abschlussbericht des gleichnamigen DFG-Projekts (RA 2380/1-1|GR 1951/2-1).

Rheinberg, F. (2001): Bezugsnormen und schulische Leistungsbeurteilung. In: Weinert, F.E. (Hrsg.): Leistungsmessungen in Schulen. Weinheim: Beltz, S. 59-71.

Rindermann, H. (2001): Lehrevaluation. Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts. Landau: Empirische Pädagogik.

Spiel, C./Gössler, M. (2000): Zum Einfluss von Biasvariablen auf die Bewertung universitärer Lehre durch Studierende. In: Zeitschrift für Pädagogische Psychologie 14, H. 1, S. 38-47.

ten Venne, M./Nachtigall, C./Kämpfe-Hargrave, N. (2010): SEfU – Schüler als Experten für Unterricht. In: Informationsschrift Recht und Bildung des Instituts für Bildungsforschung und Bildungsrecht 7, H. 1, S. 17-21.

Thiel, F./Ulber, D. (2007): Unterrichtsentwicklung durch Evaluation. In: Bauer, K.-O. (Hrsg.): Evaluation an Schulen. Theoretischer Rahmen und Beispiele guter Evaluationspraxis. Weinheim u.a.: Juventa, S. 163-186.

Wagner, C. (2009): Wissenschaftlich und dennoch praktikabel? Nutzbarkeit von Daten aus Fragebogenerhebungen im Rahmen der internen Evaluation an Schulen. In: Münk, D./Deißinger, T./Tenberg, R. (Hrsg.): Forschungserträge aus der Berufs- und Wirtschaftspädagogik. Probleme, Perspektiven, Handlungsfelder und Desiderata der beruflichen Bildung in der Bundesrepublik Deutschland, in Europa und im internationalen Raum. Opladen: Budrich, S. 40-49.

Wagner, C. (2010): Unterrichtsentwicklung durch Evaluation am Beispiel des „Netzwerk Schülerbefragung". In: Die berufsbildende Schule 62, H. 10, S. 291-297.

Wagner, W. (2008): Methodenprobleme bei der Analyse der Unterrichtswahrnehmung aus Schülersicht. Am Beispiel der Studie DESI (Deutsch Englisch Schülerleistungen International) der Kultusministerkonferenz. Koblenz-Landau: Universität, FB Psychologie.

Wittwer, J. (2008): What Influences the Agreement among Student Ratings of Science Instruction? In: Prenzel, M./Baumert, J. (Hrsg.): Vertiefende Analysen zu PISA 2006. Wiesbaden: VS, S. 205-220.

Christoph Fuhrmann, M.A., geb. 1963, wissenschaftlicher Mitarbeiter am Institut für Bildungsforschung in der School of Education der Bergischen Universität Wuppertal.
E-Mail: christoph.fuhrmann@uni-wuppertal.de

Miriam Sharon Keune, Dr., geb. 1980, wissenschaftliche Mitarbeiterin am Institut für Bildungsforschung in der School of Education der Bergischen Universität Wuppertal.
E-Mail: miriam.keune@uni-wuppertal.de

Sylvia Rahn, Prof. Dr., geb. 1967, Professorin für Berufsbildungsforschung am Institut für Bildungsforschung in der School of Education der Bergischen Universität Wuppertal.
E-Mail: sylvia.rahn@uni-wuppertal.de

Anschrift: Bergische Universität Wuppertal, Gaußstraße 20, 42119 Wuppertal

Sabine Gruehn, Prof. Dr., geb. 1966, Professorin für Erziehungswissenschaft mit dem Schwerpunkt Schultheorie/Schulforschung an der Westfälischen Wilhelms-Universität Münster.
E-Mail: sabine.gruehn@uni-muenster.de

Anschrift: Westfälische Wilhelms-Universität Münster, Bispinghof 5/6, 48143 Münster

BERICHTE ZUM SCHWERPUNKTTHEMA

DDS – Die Deutsche Schule
108. Jahrgang 2016, Heft 2, S. 176-190

Klaudia Schulte/Detlef Fickermann/Markus Lücken

Das Hamburger Prozessmodell datengestützter Schulentwicklung

Zusammenfassung
Aufgrund der seit 1997 eingeführten diversen datengestützten Maßnahmen der Schul- und Qualitätsentwicklung stehen Akteure auf allen Ebenen des Bildungssystems vor der Aufgabe, aus einer Vielzahl von Daten informierte, sinnvolle und nachvollziehbare Entscheidungen zur Optimierung von Schule und Unterricht zu treffen und umzusetzen. Bisher fehlt es jedoch flächendeckend an systematischen Überlegungen zur Orchestrierung der einzelnen Maßnahmen; vorliegende Wirkmodelle beschreiben jeweils nur Teilaspekte. Auf der Basis theoretischer Überlegungen und wissenschaftlicher Befunde wird ein in dem Hamburger Projekt „Integrierte Datennutzung in allgemeinbildenden Schulen" (IDA) entwickeltes Prozessmodell datengestützter Schulrückmeldungen vorgestellt, welches das vielfach verwendete Kontext-Input-Prozess-Output-Modell mit zwei Qualitätszyklen mehrebenenanalytisch verbindet und verschiedene datengestützte Rückmeldungen sowie deren Verhältnis zueinander abbildet. Damit wird versucht, einen ersten Beitrag zu einer besseren Orchestrierung der verschiedenen Maßnahmen auf den verschiedenen Ebenen des Bildungssystems zu leisten.
Schlüsselwörter: datengestützte Schulentwicklung, Bildungsmonitoring, Governance

The Hamburg Process Model of Data-based School Development

Summary
Due to the implementation of diverse data-based measures for school and quality development, actors on all levels of the educational system face the challenge to base their decisions for optimizing school and education on a variety of data. Until now, systematic reflections on how to orchestrate the different measures and data sources are missing; effect models only illustrate partial aspects. On the base of theoretical reflections and scientific findings, the article introduces a process model of data-based school evaluation feedback, which connects the well-known context-input-process-output-model with

two quality cycles in a multilevel way and illustrates different data types as well as their connection. The model was developed in the project "Integrated Use of Data in General Education" in Hamburg. By this, we try to make a contribution to an orchestration of the different measures on the diverse levels of the educational system.
Keywords: data-based school development, educational monitoring, governance

1. Einleitung

In den letzten 15 Jahren hat sich die Perspektive, mit der Bildungssysteme sowie die Schul- und Unterrichtsqualität betrachtet werden, grundlegend verändert. Die Konstanzer Beschlüsse der Kultusministerkonferenz (KMK) aus dem Jahr 1997 (vgl. KMK 1997) und die im Jahr 2001 veröffentlichten schlechten PISA-Ergebnisse Deutschlands (vgl. Baumert 2001) haben in der Bundesrepublik Deutschland zu einem neuen Steuerungsmodell des deutschen Schulwesens auf der Grundlage evidenzbasierter Qualitätssicherung und -entwicklung geführt. Ihre diesbezüglichen Überlegungen hat die KMK im Jahre 2006 in einer Gesamtstrategie zum Bildungsmonitoring zusammengefasst, die im Jahre 2015 aktualisiert worden ist (vgl. KMK 2006; 2015). Im Rahmen dieser Strategie wurden u.a. die Teilnahme an den internationalen und nationalen Leistungsvergleichsstudien, die Einführung von Vergleichsarbeiten in allen Bundesländern sowie die Implementation externer Evaluationen durch Schulinspektionsverfahren beschlossen.

Aufgrund der Einführung dieser und weiterer Maßnahmen stehen nun Akteure auf allen Ebenen des Bildungssystems vor der Aufgabe, aus dieser Datenfülle informierte, sinnvolle und nachvollziehbare Entscheidungen zur Optimierung von Schule und Unterricht zu treffen und umzusetzen. So werden beispielsweise auf der Ebene der Bildungsadministration anhand von Monitoring-Daten Ressourcen umverteilt, oder Schulaufsichten nutzen die Schulergebnisse der Vergleichsarbeiten, um mit der Schulleitung gemeinsam Schulentwicklungsziele zu beschließen. Auf Ebene der Schule wird wiederum verlangt, dass die Schulleitung und das Kollegium die Ergebnisse interner und externer Evaluationen nutzen, um daraus gezielt Maßnahmen der Schul- und Unterrichtsentwicklung abzuleiten. Während nach und nach die verschiedenen Akteure auf allen Ebenen mit der Interpretation einzelner Datenquellen zunehmend vertrauter werden, fehlt es dennoch an systematischen Überlegungen, wie die verschiedenen Maßnahmen auf den verschiedenen Ebenen des Bildungssystems orchestriert werden können (vgl. Böttcher 2013).

Datengestützte Schul- und Unterrichtsentwicklung sowie Perspektiven in Hamburg

Die meisten Bundesländer haben in den letzten 15 Jahren eigene, die Gesamtstrategie der KMK umsetzende und z.T. ergänzende Instrumente und Verfahren der externen Evaluation eingeführt, die ihnen – so die erklärte Absicht – systematisch und datengestützt Aufschluss über Leistungen und Herausforderungen auf den verschiedenen Ebenen des Bildungssystems geben sollen. Hamburg ist in dieser Hinsicht besonders weit fortgeschritten. Neben auf sechs Jahrgänge erweiterten jährlichen Vergleichsarbeiten (Kompetenzen ermitteln [KERMIT]; vgl. Lücken et al. 2014 und Schulte/Lücken 2015) in Deutsch, Mathematik, Naturwissenschaften und Englisch mit Angaben zur Kompetenzentwicklung der Schülerinnen und Schüler und Vergleichswerten von Schulen mit ähnlicher sozialer Zusammensetzung der Schülerschaft („faire Vergleiche"; vgl. Schulte/Hartig/Pietsch 2014) erhalten die Schulen in regelmäßigen Abständen eine Rückmeldung der Hamburger Schulinspektion (vgl. Pietsch/Scholand/Schulte 2015), den jährlichen Datenreport „Schule im Überblick" (SchÜb; vgl. IfBQ 2015, S. 19), der das sozialräumliche Umfeld der Schule sowie die soziale und ethnische Zusammensetzung der Schülerschaft abbildet, eine jährliche Rückmeldung zur Umsetzung der Maßnahmen des Sprachförderkonzepts (vgl. May/Berger 2014) sowie seit 2015 eine jährliche Rückmeldung zu den Ergebnissen der Zentralen Abschlussprüfungen mit den Vergleichswerten des Vorjahres und Ergebnissen relevanter Referenzgruppen (Schulen mit vergleichbarer Schülerschaft, Schulen derselben Schulform, alle Schulen). Darüber hinaus arbeiten alle Hamburger Schulen mit weiteren Daten, z.B. Noten, Daten aus internen Evaluationen, Schülerfeedbacks etc.

Um die zahlreichen Rückmeldungen und Auswertungen sowohl für die einzelnen Schulen als auch für die „Steuerleute" auf den verschiedenen Ebenen (Schulaufsichten, Leitung der Behörde für Schule und Berufsbildung) und für das Parlament (Bürgerschaft) zu systematisieren und deutlich stärker methodisch und inhaltlich aufeinander beziehen zu können, wurde im Institut für Bildungsmonitoring und Qualitätsentwicklung (IfBQ) das Projekt „Integrierte Datennutzung an allgemeinbildenden Schulen" (IDA) installiert (1. Phase: Juli 2014 bis August 2015; eine 2. Phase wird derzeit [Januar 2016] vorbereitet). Auftrag des Projekts in der ersten Phase war es, eine empirische Bestandsaufnahme der aktuellen Nutzungssituation datengestützter Rückmeldungen durch Schulen und andere Akteure in Hamburg vorzunehmen sowie Ideen zu entwickeln, wie eine systematische Datennutzung in Hamburg weiter unterstützt und gefördert werden kann. Im Rahmen des Projekts wurden insgesamt über 300 unterschiedliche Akteure befragt: Schulleitungen, Datenbeauftragte an Schulen, Lehrkräfte, Schulaufsichtsbeamte und -beamtinnen, externe Wissenschaftlerinnen und Wissenschaftler sowie Kolleginnen und Kollegen aus anderen Bundesländern.

Der vorliegende Beitrag führt zunächst kurz theoretisch in das Feld der „Neuen Steuerung" und in das entsprechende, durch unterschiedliche Ansätze und Perspektiven geprägte Forschungsfeld ein und stellt einige ausgewählte Modelle zur evidenzgestützten Schul- und Unterrichtsentwicklung vor. Anschließend werden relevante Wirkfaktoren für die Nutzung datengestützter Rückmeldungen übersichtsartig dargestellt.

Auf der Basis theoretischer Überlegungen und wissenschaftlicher Befunde sowie der Befragungsergebnisse wird ein in dem Projekt IDA entwickeltes, integriertes Prozessmodell datengestützter Schulrückmeldungen vorgestellt, welches das vielfach verwendete Kontext-Input-Prozess-Output-Modell (KIPO-Modell) mit zwei Qualitätszyklen mehrebenenanalytisch verbindet und verschiedene datengestützte Rückmeldungen und deren Verhältnis zueinander abbildet.

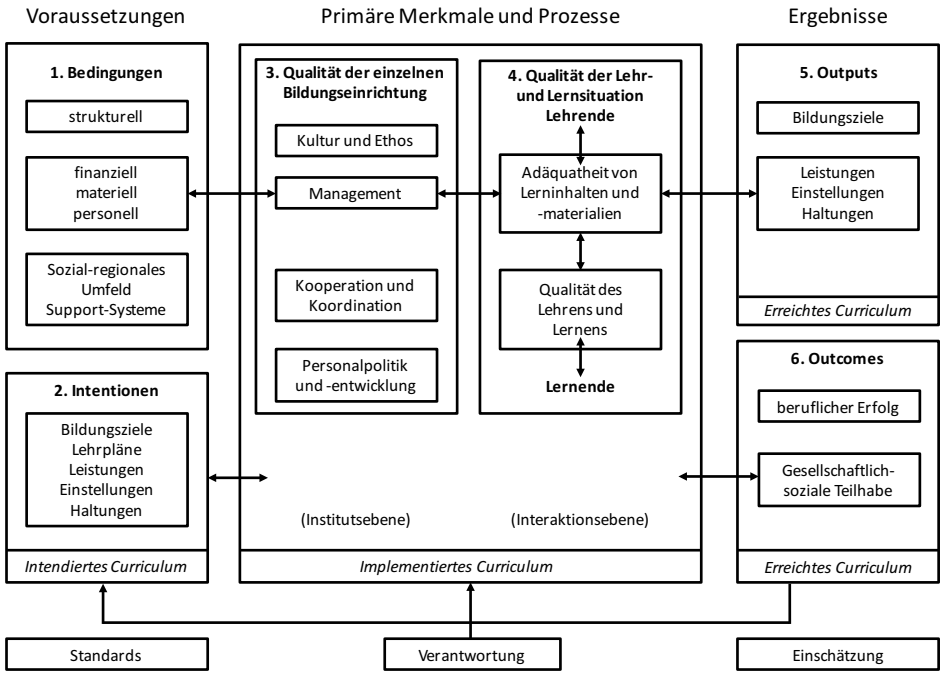
Mit dem vorliegenden Artikel und dem Prozessmodell wird versucht, einen ersten Beitrag zu einer besseren Orchestrierung der verschiedenen Maßnahmen auf den verschiedenen Ebenen des Bildungssystems (s.o.) zu leisten.

2. Datennutzung aus wissenschaftlicher Sicht

2.1 Das KIPO-Modell

Die KIPO-Modelle der Schuleffektivitätsforschung (vgl. Abb. 1; Ditton 2007; Scheerens 2000) postulieren, dass Schüler-Outcomes durch die Kombination und Interaktion von Kontextfaktoren mit Input- und Prozessfaktoren determiniert werden. Dabei wird auf Seiten der Input-Faktoren zwischen finanziellen, materiellen und personellen Aspekten unterschieden, wie z.B. der Schülerzahl bzw. Merkmalen der Schülerinnen und Schüler (Alter, Geschlecht, Migrationshintergrund). Prozessmerkmale beschreiben die schulischen Prozesse innerhalb der Klasse und innerhalb der Schule. Output-Aspekte beschreiben erreichte Leistungen bzw. Einstellungen. Als Kontextfaktoren zählen typische „Kovariablen" wie Schulgröße oder die Lage der Schule. Das neue Steuerungsmodell impliziert im Sinne eines klassischen KIPO-Modells der Schulqualität also eine Konzentration auf erreichte Leistungen „sowie gesellschaftlich erwünschte Werthaltungen und Einstellungen der Schüler/innen" (Peek 2006, S. 1345; vgl. Altrichter/Kanape-Willingshofer 2012; siehe auch Fends Erweiterung um die Komponente eines mehrebenenanalytisch zu verstehenden Angebots-Nutzungs-Modells: Fend 2008).

Abb. 1: Modell zur Qualität und Qualitätsentwicklung im Bildungsbereich



Quelle: Eigene Darstellung nach Ditton 2007

2.2 Multifunktionalität datengestützter Schulentwicklung

International unterscheidet sich die Umsetzung einer outputorientierten Steuerung in Bezug auf die Konsequenzen, die an empirisch generierte Evidenzen geknüpft werden. In „High-Stakes"-Systemen können ungünstige Ergebnisse mit weitreichenden Konsequenzen wie Ressourceneinschränkungen verbunden werden (z.B. in den USA oder Großbritannien). In „Low-Stakes"-Systemen gehen solche Ergebnisse mit weniger weitreichenden Konsequenzen einher (z.B. in Deutschland oder den Niederlanden). Demski et al. (2012) sowie van Ackeren, Binnewies und Clausen (2013) unterscheiden verschiedene weitere Aspekte der evidenzbasierten Schulentwicklung: Evidenzquellen im engeren Sinne mit einer ausgeprägten Wissenschaftsorientierung (z.B. Vergleichsarbeiten) versus Evidenzquellen im weiteren Sinne, die eher interne Impulse setzen (z.B. Schülerfeedback), schulintern versus -extern erhobene Daten, die Art der Datenerhebung (inwieweit wurden (sozial-)wissenschaftliche Methoden angewendet), die Aussageebene (Bildungssystem, Schule, Klasse oder Schüler/Schülerin), der spezifische Steuerungsanspruch (formalisiert, explizit vs. informell, implizit) sowie die Konkretheit der Handlungsempfehlungen bzw. die tatsächliche oder intendierte Handlungsnähe. Van Ackeren et al. (2011) grenzen vom

evidenzbasierten Wissen das „Ersatzwissen" ab, das sie als „Sammlung von Leitideen, die ohne exakten Nachweis ihrer Wirksamkeit mittels gezielter und systematischer Analyse angewendet werden" (S. 56), begreifen.

Vor allem den Evidenzquellen im engeren Sinne werden teils spannungsreiche Doppelfunktionen (vgl. Altrichter 2010) bzw. Multifunktionen, wenn man die Funktionen der Normendurchsetzung und Wissensgewinnung berücksichtigt (vgl. Landwehr 2011), zugeschrieben: Sie sollen auf der einen Seite der Rechenschaftslegung gegenüber übergeordneten Steuerungsebenen dienen und haben damit eine Kontrollfunktion inne. Auf der anderen Seite sollen sie Impulse für die Schulentwicklung setzen. Dieses grundsätzliche Spannungsverhältnis spiegelt sich auch in der Gesamtstrategie der KMK (2006) wider (vgl. Bohl/Kleinknecht/Maier 2008). Im Zuge der zunehmenden Veröffentlichung schulischer Leistungsindikatoren sowie der Etablierung von Quasimärkten durch Dezentralisierung, Schulautonomie und freie Schulwahl entstehen darüber hinaus Wettbewerbssituationen, denen die Schulen ausgesetzt sind (vgl. Bellmann/Weiß 2009).

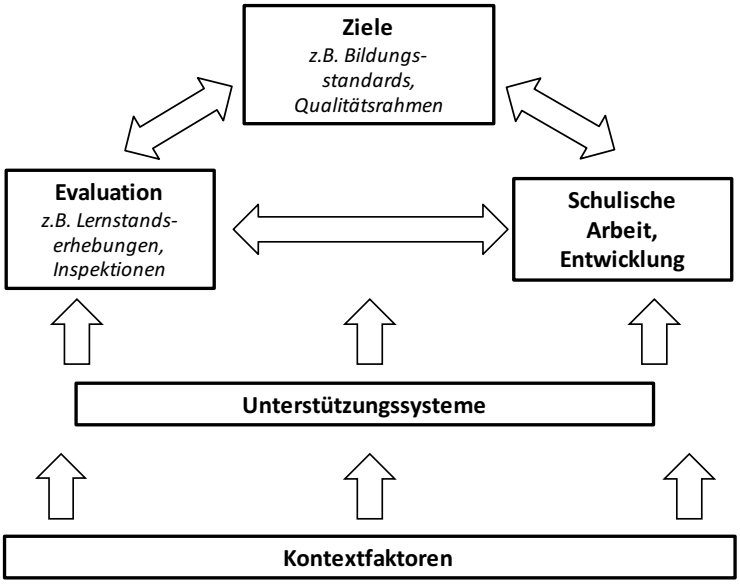
2.3 Wirkmodelle outputorientierter Steuerung

Zu den Wirkungen outputorientierter Steuerung liegen Modelle verschiedener Autorinnen und Autoren vor:

Altrichter (2010) beschreibt einen einfachen Regelkreis outputorientierter Steuerung, d.h. der evidenzbasierten Steuerung mit Fokus auf Leistungen, in einem „Idealmodell" (siehe Abb. 2): In diesem Modell werden durch Standardsetzungen und Qualitätsrahmen Ziele vorgegeben, deren Erreichung in der schulischen Arbeit durch Evaluationen überprüft wird. Der Regelkreis von Altrichter gibt einen guten ersten Überblick, ist jedoch für eine differenzierte Betrachtung zu unspezifisch und leistet keinen Beitrag zur Integration der einzelnen Maßnahmen.

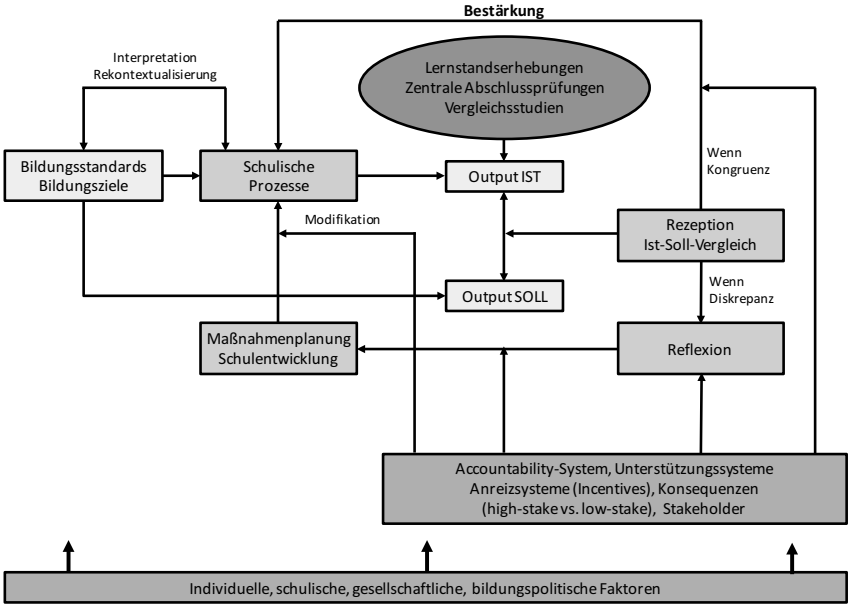
Maag Merki (2010) expliziert ein komplexeres Wirkungsmodell der outputorientierten neuen Steuerung (siehe Abb. 3). In dem Modell wird spezifiziert, dass über Ist-Soll-Diskrepanzen Impulse für Handlungen im Schulkontext gesetzt werden, die zu schulischen Entwicklungsmaßnahmen mit dem Ziel einer weniger starken Diskrepanz führen sollen. Dabei treten verschiedene Einflussfaktoren auf, z.B. das jeweilige Accountability- oder Unterstützungs-System, aber auch individuelle und schulische Faktoren. Obwohl das Modell aktuelle Reformbestrebungen beschreibt, bleibt offen, inwiefern weitere Instrumente der Outputorientierung, z.B. Verfahren der externen Evaluation, in das Wirkungsmodell integriert werden können.

Abb. 2: Regelkreis der outputorientierten Steuerung



Quelle: Eigene Darstellung nach Altrichter 2010

Abb. 3: Wirkungsmodell von Bildungsstandards und outputübergreifenden Verfahren



Quelle: Eigene Darstellung nach Maag Merki 2010, S. 154

Das Modell von Maag Merki beschränkt sich darüber hinaus zu stark auf eine Innensicht der Schule bei der Umsetzung von Vorgaben mit Hilfe von externen Evaluationsdaten. Die Außenperspektive und die Mehrebenenstruktur des Schulsystems werden in ihrem Modell zwar unterstellt, aber nicht ausdifferenziert.

Grundsätzlich ist es sinnvoll, bei Wirkmodellen Ansätze aus der Governance-Forschung zu integrieren, indem berücksichtigt wird, wie die Akteure auf den verschiedenen Ebenen des Bildungssystems die Vorgaben und Entwicklungen im Sinne von Fend (2008) rekontextualisieren. Gleichzeitig sollten auch Erkenntnisse der Implementationsforschung mit in die Beschreibung eines Wirkmodells outputorientierter Steuerung einfließen, in dem der Top-Down-Fluss von Vorgaben der Bildungsadministration und deren Umsetzung in der Schule explizit verfolgt (vgl. van Ackeren et al. 2011) und mögliche Implementationsbrüche (vgl. Zlatkin-Troitschanskaia 2006) mitgedacht werden.

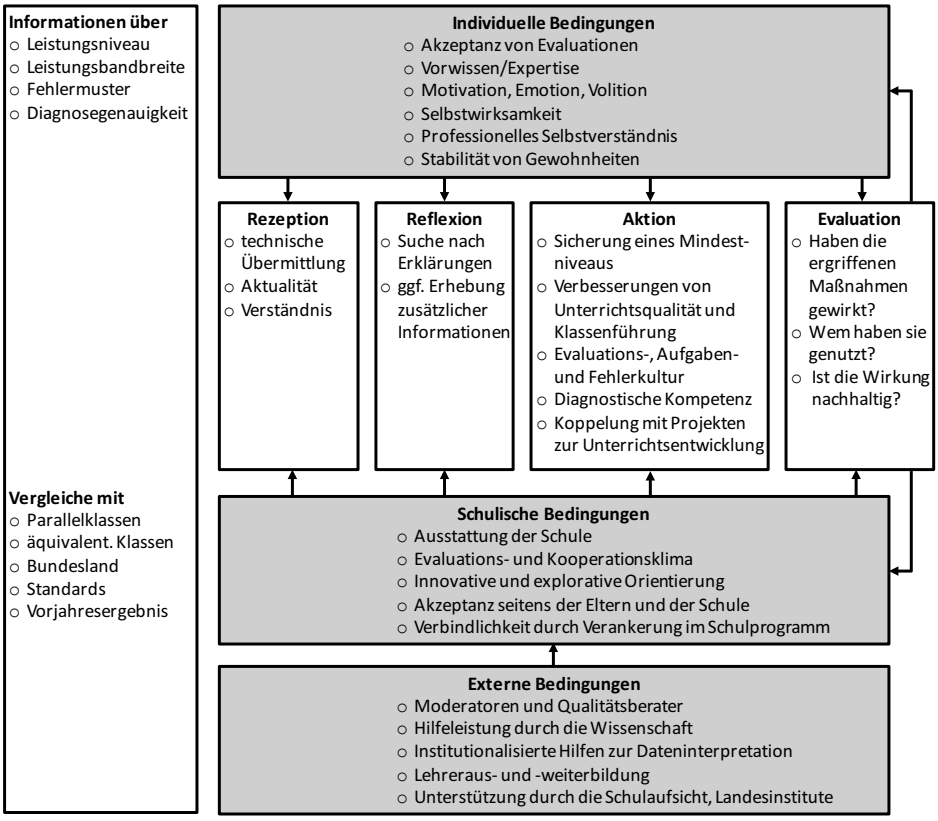
3. Wirkfaktoren für die Nutzung von Rückmeldungen

Die Nutzung von Rückmeldungen ist abhängig von der Bereitschaft und Kompetenz, die Rückmeldungen zu lesen, zu verstehen und daraus Maßnahmen abzuleiten. Das gilt nicht nur für eine einzige Rückmeldung, sondern im Sinne der Orchestrierung für eine Ansammlung von verschiedenen Datenquellen, deren Informationen sich überschneiden, aber auch gegenseitig widersprechen können.

Altrichter (2010) schätzt die Qualität der Rückmeldungen in Deutschland insgesamt als relativ hoch ein; in den letzten Jahren habe es umfangreiche Bemühungen in Richtung einer höheren Qualität gegeben. Nach Landwehr (2014) sollten Leistungsergebnisse jedoch durch die Lehrpersonen mit Prozessinformationen angereichert werden, da diese die Lernergebnisse verursachen; er beschreibt dies allerdings als „außerordentlich anspruchsvoll" (S. 2) und als „Paradigmenwechsel im Unterrichtsverständnis" (ebd.). Dies sei zusätzlich dadurch erschwert, dass die Tests nicht durch die Lehrpersonen selbst, sondern durch Externe konstruiert würden. Darüber hinaus seien viele Leistungsdaten für Lehrkräfte angstbesetzt. In der Auseinandersetzung mit Daten sei die Schulleitung gefordert, die dies als Teil des schulinternen Qualitäts- und Personalentwicklungskonzepts verstehen müsse.

Das Rahmenmodell von Helmke (2004, siehe Abb. 4; vgl. auch Groß Ophoff/Hosenfeld/Koch 2007) zu Faktoren der Nutzung von Vergleichsarbeiten integriert verschiedenste Befunde der Schulforschung und unterscheidet die vier prozesshaften Schritte Rezeption, Reflexion, Aktion und Evaluation. Helmke beschreibt die vielfältigen externen, schulischen und individuellen Bedingungen, die die Nutzung leistungsbezogener Daten in der Schule beeinflussen können. Das Modell unterscheidet Angebote

Abb. 4: Rahmenmodell zur pädagogischen Nutzung von Vergleichsarbeiten

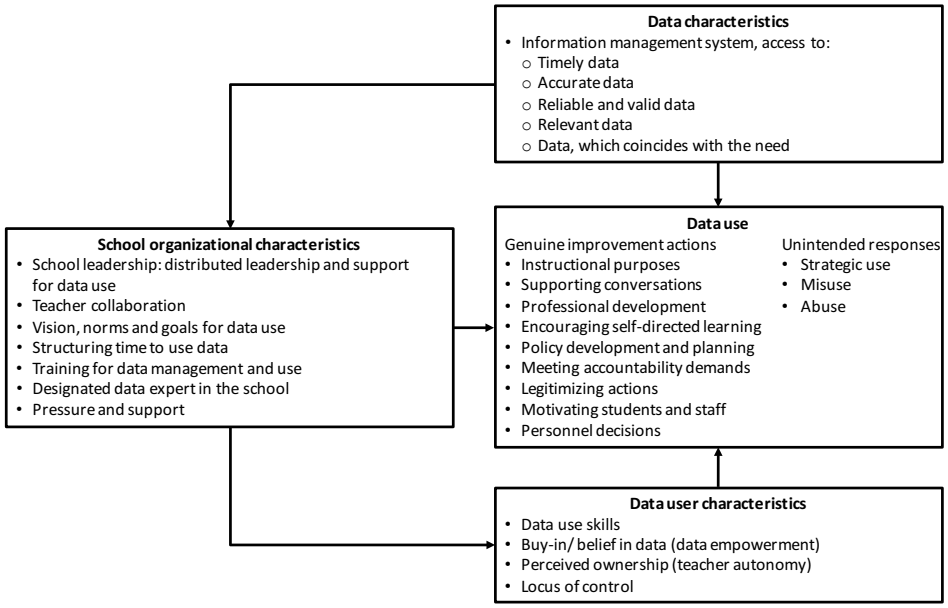


Quelle: Eigene Darstellung nach Groß Ophoff/Hosenfeld/Koch 2007, nach Helmke 2004

von eben diesen komplexen Nutzungsaspekten der einzelnen Lehrperson. In Helmkes Modell werden die Mehrebenenstruktur des Schulsystems sowie eine Beschreibung der Zusammenhänge der Verknüpfung von Handlungs- und Lernvorgängen allerdings nicht berücksichtigt (vgl. Altrichter 2010).

Schildkamp und Kuiper (2010) fassen aus einem Literaturüberblick der internationalen Forschung (z.B. Datnow/Park/Wohlstetter 2007) extrahierte Faktoren zusammen, die Datennutzung an Schulen allgemein fördern und behindern. Sie unterscheiden dabei zwischen den Charakteristika der Daten, der Schulorganisation und der Nutzer (siehe Abb. 5).

Abb. 5: Einflussfaktoren auf Datennutzung



Quelle: Eigene Darstellung nach Schildkamp/Kuiper 2010, S. 485

4. Das Hamburger Prozessmodell datengestützter Schulentwicklung

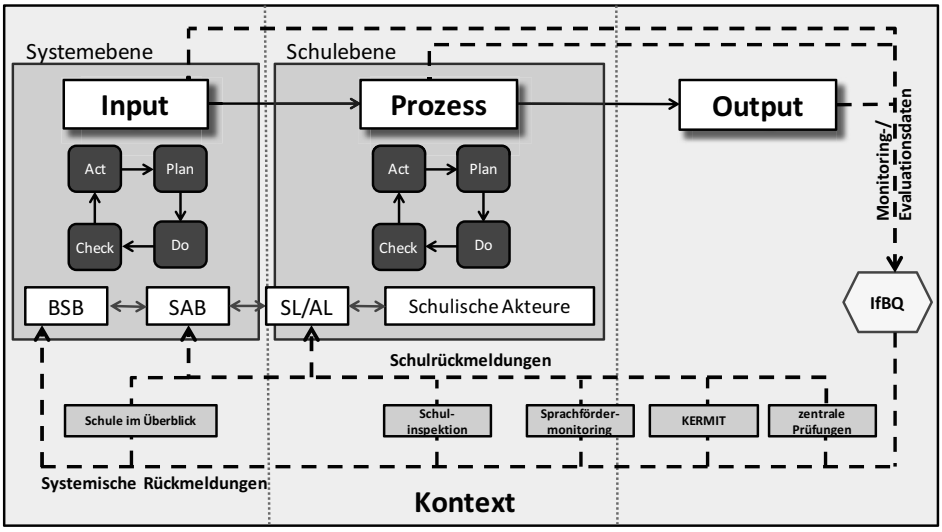
Die bislang vorliegenden Wirk- und Prozessmodelle evidenzbasierter Schulentwicklung beziehen sich zumeist auf Vergleichsarbeiten bzw. outputbasierte Verfahren oder sind sehr allgemein gehalten. In keinem Modell wird die komplexe Mehrebenenstruktur des Schulsystems abgebildet, und es finden auch keine systematischen Verschränkungen mit Qualitätszyklen statt.

Im Projekt „IDA" wurde deshalb zunächst im Sinne einer Heuristik ein neues Prozessmodell zur Beschreibung datengestützter Schulentwicklung in Hamburg entwickelt (siehe Abb. 6). In dem Modell verbinden sich zwei Qualitätsentwicklungszyklen mehrebenenanalytisch mit dem klassischen KIPO-Modell; es wird systematisch zwischen der Schulinnen- und der Schulaußenwelt unterschieden.

Das IfBQ in seiner Funktion als Monitoring- und Qualitätsentwicklungsinstanz misst verschiedene Daten des Schulsystems: Kontext-, Input-, Prozess- und Outputinformationen (gestrichelte Linien). Diese Daten werden analysiert und in verschiedenen Formaten sowohl den Schulleitungen (SL/AL) als auch der Systemebene (Behörde für Schule und Berufsbildung: BSB) sowie vermittelnden Ebenen, wie

den Schulaufsichten (SAB), zur evidenzbasierten Steuerung zur Verfügung gestellt. Die Verarbeitung der Rückmeldungen geschieht idealerweise durch die Rekontextualisierung in zwei Qualitätszyklen: einem schulischen Qualitätszyklus unter Einbindung der schulischen Akteure sowie einem Qualitätszyklus auf Systemebene.

Abb. 6: Hamburger Prozessmodell datengestützter Schulentwicklung



Quelle: eigene Darstellung

Typischerweise werden in schulischen Qualitätsentwicklungszyklen die Schritte „Plan – Do – Check – Act" unterschieden. Der Schritt „Plan" in einem Qualitätsentwicklungszyklus meint dabei die Zielsetzungen sowie darauf aufbauende Maßnahmenplanungen zur Zielerreichung. Als zweites folgt der Schritt „Do", d.h. die Umsetzung der beschlossenen Maßnahmen. Der Schritt „Check" umfasst die Prüfung der Maßnahmen- und Zielerreichung, z.B. anhand von internen und externen Evaluationsdaten. Der letzte Schritt „Act" beschreibt schließlich auf dem Hintergrund der Überprüfung des aktuellen Standes die Re-Formulierung von Zielen bzw. die Schärfung der eingeleiteten Maßnahmen.

Im Fall des zweiten Qualitätszyklus auf der Systemebene erfolgt die Zielsetzung („Plan") durch die Bürgerschaft oder durch die Behördenleitung. „Do" umfasst die Implementation der beschlossenen Maßnahmen in der Regel durch die Administration. Das Maß der Zielerreichung („Check") wird vom IfBQ durch Auswertungen der erhobenen Daten auf der Systemebene überprüft und zurückgemeldet. In idealtypischer Weise der Umsetzung einer evidenzbasierten Steuerung dienen diese Systemanalysen der Reformulierung der politischen Ziele bzw. der Schärfung der eingeleiteten Maßnahmen („Act").

Die Rückmeldungen werden in den Schulen und auf der Systemebene unterschiedlich rezipiert und rekontextualisiert; es werden Ist-Soll-Vergleiche vorgenommen und bei Diskrepanzen gegebenenfalls Maßnahmen geplant und umgesetzt, so dass sowohl Input-Faktoren als auch schulische Prozesse beeinflusst werden. Der schulische ist mit dem systemischen Qualitätszyklus über das KIPO-Modell verbunden: Die politischen oder von der Behördenleitung beschlossenen Maßnahmen, seien es Ressourcenausstattungen, Rahmenlehrpläne oder andere Vorgaben, gehören im KIPO-Modell zum Input einer Schule. Sie beeinflussen die Schulinnenwelt und führen über eine geänderte Prozessgestaltung zu einem anderen Output, der wiederum vom IfBQ erhoben und sowohl der Schule als Rückmeldung zu ihrer Prozessgestaltung als auch auf Systemebene der Behördenspitze zur politischen Bewertung der eingeleiteten Maßnahmen zur Verfügung gestellt wird.

Denkbar ist es, das Modell um Unterstützungssysteme, die ihren Fokus auf die Schul- und Unterrichtsentwicklung richten, zu erweitern. Dies sind einerseits das Landesinstitut und andererseits das IfBQ, das schulische Akteure bei der Interpretation von Daten unterstützen und Vorschläge zu ihrer Nutzung unterbreiten kann.

In dem Modell ist gekennzeichnet, welche Rückmeldungen in Hamburg bereitgestellt werden; z.T. sind davon unterschiedliche weitere schulische Akteure betroffen. So erhält z.B. die Schulleitung die KERMIT-Rückmeldungen und gibt die jeweils relevanten Rückmeldungsteile an Lehrkräfte, an die Fachkonferenzen und – falls in der Schule vorhanden – auch an die Beauftragte oder den Beauftragten für Evaluationen, an die didaktische Leitung oder an das sonderpädagogische Fachpersonal weiter. Darüber hinaus besprechen in bestimmten Jahrgängen die Lehrkräfte die individuellen Rückmeldungen der Schülerinnen und Schüler mit diesen und deren Eltern in Lernentwicklungsgesprächen.

Anhand der graphischen Verortung der Rückmeldungen im Bereich des Inputs, Prozesses oder Outputs ist im Modell direkt ersichtlich, auf welche Aspekte sich die Rückmeldungen beziehen. Dabei erfassen zwei Rückmeldungen mehrere Aspekte; der jeweils wichtigere Aspekt ist im Modell visualisiert worden: Die Rückmeldung „Schule im Überblick“ stellt neben Input-Aspekten wie z.B. der Anzahl von Schülerinnen und Schülern auch Kontextdaten wie Vergleichsdaten von Schulen mit ähnlicher Schülerzusammensetzung, Daten zum Stadtteil und zum Bezirk sowie einen Output-Aspekt dar: die Abschlussquoten. Darüber hinaus werden in der Rückmeldung zur Umsetzung der Sprachfördermaßnahmen nicht nur Output-Aspekte (Fördererfolge), sondern auch prozessuale Aspekte (Qualität der Umsetzung der Maßnahmen) sowie Input-Aspekte (Einsatz der Förderressourcen) thematisiert.

In das Modell können weitere Daten integriert werden, die bspw. in Schulen selbst erhoben oder verwendet werden, wie beispielsweise interne Evaluationsdaten, Leistungstests zur Diagnose von Förderbedarfen oder Daten zum Unterrichtsausfall.

Je nach Ziel der Messung lassen sich auch diese Daten den Input-, Prozess- und Output-Bereichen zuordnen.

6. Fazit

Bei dem im Projekt IDA entwickelten Prozessmodell handelt es sich zunächst um eine Heuristik, in der das KIPO-Modell mit zwei Qualitätsentwicklungszyklen unter Beachtung der Mehrebenenstruktur des Schulsystems und der doppelten Nutzung von Instrumenten zur Qualitätsmessung verbunden worden ist. Der weitere Verlauf der Diskussion im geplanten Folgeprojekt IDA 2.0 wird zeigen, wie tragfähig das Modell für die Beschreibung funktionaler Zusammenhänge und der Handlungsräume der verschiedenen Akteure ist. Übergeordnetes Ziel ist dabei, Schulen in ihren jeweils eigenen Schulentwicklungsprozessen durch optimierte und aufeinander bezogene Rückmeldungen und durch die Bereitstellung von Daten optimal zu unterstützen.

Literatur und Internetquellen

Altrichter, H. (2010): Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In: Altrichter, H./Maag Merki, K. (Hrsg.): Handbuch Neue Steuerung im Schulsystem. Wiesbaden: VS, S. 219-254.

Altrichter, H./Kanape-Willingshofer, A. (2012): Bildungsstandards und externe Überprüfung von Schülerkompetenzen: Mögliche Beiträge externer Messungen zur Erreichung der Qualitätsziele der Schule. In: Nationaler Bildungsbericht Österreich. Hrsg. vom Bundesministerium für Unterricht, Kunst und Kultur und vom Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des Österreichischen Schulwesens. Graz: Leykam, S. 355-394.

Baumert, J. (2001): PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske + Budrich.

Bellmann, J./Weiß, M. (2009): Risiken und Nebenwirkungen Neuer Steuerung im Schulsystem: Theoretische Konzeptualisierung und Erklärungsmodelle. In: Zeitschrift für Pädagogik 55, H. 2, S. 286-308.

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

Bohl, T./Kleinknecht, M./Maier, U. (2008): Datenbasierte Selbst- und Fremdevaluation: Eine exemplarische Analyse des Steuerungskonzeptes in Baden-Württemberg. In: Die Deutsche Schule 100, H. 4, S. 459-466.

Datnow, A./Park, V./Wohlstetter, P. (2007): Achieving with Data: How High-performing School Systems Use Data to Improve Instruction for Elementary Students. Center on Educational Governance, Rossier School of Education, University of Southern California. URL: http://www.newschools.org/viewpoints/AchievingWithData.pdf; Zugriffsdatum: 16.02.2007.

Demski, D./Rosenbusch, C./van Ackeren, I./Clausen, M./Schmidt, U. (2012): Steuerung von Schule durch evidenzbasierte Einsicht? Konzeption und erste Befunde des Forschungsverbundes EviS. In: Hornberg, S. (Hrsg.): Deregulierung im Bildungswesen. Münster u.a.: Waxmann, S. 131-150.

Ditton, H. (2007): Kompetenzaufbau und Laufbahnen im Schulsystem: Ergebnisse einer Längsschnittuntersuchung an Grundschulen. Münster u.a.: Waxmann.

Fend, H. (2008): Schule gestalten: Systemsteuerung, Schulentwicklung und Unterrichtsqualität. Lehrbuch. Wiesbaden: VS.

Groß Ophoff, J./Hosenfeld, I./Koch, U. (2007): Formen der Ergebnisrezeption und damit verbundene Schul- und Unterrichtsentwicklung. In: Empirische Pädagogik 21, H. 4, S. 411-427.

Helmke, A. (2004): Von der Evaluation zur Innovation: Pädagogische Nutzbarmachung von Vergleichsarbeiten in der Grundschule. In: Seminar, H. 2, S. 90-112.

IfBQ (Institut für Bildungsmonitoring und Qualitätsentwicklung) (2015): Tätigkeitsbericht 2012-2014. Hamburg: IfBQ. URL: http://www.hamburg.de/contentblob/4468054/data/pdf-ifbq-taetigkeitsbericht-2015.pdf; Zugriffsdatum: 04.04.2016.

KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland) (1997): Grundsätzliche Überlegungen zu Leistungsvergleichen innerhalb der Bundesrepublik Deutschland: Konstanzer Beschluss. Beschluss der Kultusministerkonferenz vom 24.10.1997. Bonn/Berlin: KMK.

KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland) (2006): Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring. München: Wolters Kluwer.

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

Landwehr, N. (2011): Thesen zur Wirkung und Wirksamkeit der externen Schulevaluation. In: Quesel, C./Husfeldt, V./Landwehr, N./Steiner, P. (Hrsg.): Wirkungen und Wirksamkeit der externen Schulevaluation. Bern: hep, S. 35-69.

Landwehr, N. (2014): Von den Check-Daten zu den UE-Taten: Wie gelangt man von den Testdaten zur Konzipierung von Unterrichts- und Schulentwicklungsmaßnahmen? Unveröffentlichtes Manuskript. Fachhochschule Nordwestschweiz, Pädagogische Hochschule, Windisch.

Lücken, M. et al. (2014): KERMIT – Kompetenzen ermitteln. In: Fickermann, D./Maritzen, N. (Hrsg.): Grundlagen für eine daten- und theoriegestützte Schulentwicklung. Konzeption und Anspruch des Hamburger Instituts für Bildungsmonitoring und Qualitätsentwicklung (IfBQ). Münster u.a.: Waxmann, S. 127-154.

Maag Merki, K. (2010): Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In: Altrichter, H./Maag Merki, K. (Hrsg.): Handbuch Neue Steuerung im Schulsystem. Wiesbaden: VS, S. 145-169.

May, P./Berger, C. (2014): Diagnostik als Grundlage des Hamburger Sprachförderkonzepts. In: Fickermann, D./Maritzen, N. (Hrsg.): Grundlagen für eine daten- und theoriegestützte Schulentwicklung. Konzeption und Anspruch des Hamburger Instituts für Bildungsmonitoring und Qualitätsentwicklung (IfBQ). Münster u.a.: Waxmann, S. 155-177.

Peek, R. (2006): Dateninduzierte Schulentwicklung. In: Buchen, H. (Hrsg.): Professionswissen Schulleitung. Weinheim: Beltz, S. 1343-1366.

Pietsch, M./Scholand, B./Schulte, K. (Hrsg.) (2015): Schulinspektion in Hamburg: Der erste Zyklus 2007-2013: Grundlagen, Befunde und Perspektiven. Münster u.a.: Waxmann.

Scheerens, J. (2000): Improving School Effectiveness (Fundamentals of Educational Planning, No. 68). Paris: UNESCO, International Institute for Educational Planning.

Schildkamp, K./Kuiper, W. (2010): Data-informed Curriculum Reform: Which Data, What Purposes, and Promoting and Hindering Factors. In: Teaching and Teacher Education 26, H. 3, S. 482-496.

Schulte, K./Hartig, J./Pietsch, M. (2014): Der Sozialindex für Hamburger Schulen. In: Fickermann, D./Maritzen, N. (Hrsg.): Grundlagen für eine daten- und theoriegestützte Schulentwicklung. Konzeption und Anspruch des Hamburger Instituts für Bildungsmonitoring und Qualitätsentwicklung (IfBQ). Münster u.a.: Waxmann, S. 67-80.

Schulte, K./Lücken, M. (2015): Der Einfluss schulischer Prozesse auf die Lernentwicklung der Schülerinnen und Schüler an weiterführenden Schulen in Hamburg. In: Pietsch, M./Scholand, B./Schulte, K. (Hrsg.): Schulinspektion in Hamburg. Der erste Zyklus 2007-2013: Grundlagen, Befunde und Perspektiven. Münster u.a.: Waxmann, S. 317-339.

van Ackeren, I./Binnewies, C./Clausen, M. (2013): Welche Wissensbestände nutzen Schulen im Kontext von Schulentwicklung? Theoretische Konzepte und erste Befunde des EviS-Verbundprojektes im Überblick. In: Die Deutsche Schule, 12. Beiheft, S. 51-71.

van Ackeren, I./Zlatkin-Troitschanskaia, O./Binnewies, C./Clausen, M./Dormann, C./Preisendörfer, P. et al. (2011): Evidenzbasierte Schulentwicklung: Ein Forschungsüberblick aus interdisziplinärer Perspektive. In: Die Deutsche Schule 103, H. 2, S. 170-184.

Zlatkin-Troitschanskaia, O. (2006): Steuerbarkeit von Bildungssystemen mittels politischer Reformstrategien: Interdisziplinäre theoretische Analyse und empirische Studie zur Erweiterung der Autonomie im öffentlichen Schulwesen. Frankfurt a.M.: Peter Lang.

Dr. Klaudia Schulte, geb. 1980, wiss. Referentin im Institut für Bildungsmonitoring (IfBQ), Hamburg.
E-Mail: Klaudia.Schulte@ifbq.hamburg.de

Detlef Fickermann, geb. 1952, Leiter der Stabsstelle Forschungskoordination und Datengewinnungsstrategie im Institut für Bildungsmonitoring (IfBQ), Hamburg.
E-Mail: Detlef.Fickermann@ifbq.hamburg.de

Dr. Markus Lücken, geb. 1971, wiss. Referent im Institut für Bildungsmonitoring (IfBQ), Hamburg.
E-Mail: Markus.Luecken@ifbq.hamburg.de

Anschrift: Institut für Bildungsmonitoring und Qualitätsentwicklung (IfBQ), Beltgens Garten 25, 20537 Hamburg

DDS – Die Deutsche Schule
108. Jahrgang 2016, Heft 2, S. 191-204

Wolfgang Beywl/Lars Balzer

Aufbau von Evaluationskompetenzen für interne Schulevaluation durch projektbezogene Fortbildung

Zusammenfassung

Interne Evaluationen werden aktuell wieder stärker diskutiert. Der Beitrag definiert zunächst schulinterne Evaluation und identifiziert Varianten nach Größe des Evaluationsgegenstandes und nach der Bearbeitungsweise. Für drei Typen von Varianten werden erforderliche Kompetenzen skizziert. Das Kompetenzportfolio für schulinterne Evaluationen lässt sich inhaltlich in zehn Schritte gliedern, nach denen Evaluationen geplant und durchgeführt werden können. Es wird abschließend beschrieben, wie diese Kompetenzen in einer Evaluationsfortbildung entlang der zehn Schritte projektbezogen erarbeitet und angewendet werden können.

Schlüsselwörter: Selbstevaluation, interne Evaluation, Schulevaluation, Unterrichtsevaluation, Unterrichtsentwicklung, Schulentwicklung

Evaluation Capacity Building for Internal School Evaluation by Project-centered Training

Summary

Internal evaluations are currently being discussed more intensively. The paper initially defines internal school evaluation and identifies variants, depending on the size of the evaluand and the treatment mode. Required competencies are outlined for three types of variants. The competence portfolio for internal school evaluations can be broken down to ten steps, which guide evaluation planning. The paper outlines how the skills can be acquired in an evaluation training alongside the ten steps, which includes a real evaluation project.

Keywords: self-evaluation, internal evaluation, school evaluation, evaluation of teaching, improvement of teaching, school development

1. Wichtigkeit, Seltenheit und neue Chancen

Interne Evaluation gilt vielfach als zentral für die schulische Qualitätsentwicklung, als wirkmächtig zur Entwicklung von Unterricht und Schule. Nevo (2009) erachtet die interne Evaluation als nachhaltige Investition, um das Wissensmanagement der Bildungsorganisation zu stärken, und auch als Chance, Steuerungsentscheidungen vom Zentrum an die Peripherie, von der zentralen Bildungsverwaltung an die lokale Schule zu verlagern. Die deutsche Kultusministerkonferenz fordert im Zusammenhang mit der Einführung von Bildungsstandards, „verstärkt Unterrichtsentwicklung zu betreiben, […] auch, sich regelmäßig des Erfolgs der Arbeit zu vergewissern (interne Evaluation)" (KMK 2005, S. 10). Pietsch (2011) weist aus, dass interne Evaluation in den Schulgesetzen der meisten Bundesländer eine prominente Rolle einnimmt. Für die Deutschschweizer Kantone gilt dies in ähnlicher Weise (vgl. IDES 2007). Lehrbuchautoren und -autorinnen betonen, wie wichtig interne Evaluation für die Qualität der Schule sei (vgl. den kritischen Überblick von Zenkel 2015).

Empirische Studien zeichnen ein enttäuschendes Bild in Bezug auf Akzeptanz und Nachhaltigkeit: Bei Schulleitenden der D-A-CH-Länder gehören Evaluationen und andere datenbezogene Aufgaben zu den bei weitem unbeliebtesten und belastendsten (vgl. Huber/Wolfgramm/Kilic 2013). Pietsch (2011, S. 13) konstatiert auf Basis mehrerer Quellen (PISA 2003, Syntheseberichte von Schulinspektionen/externer Schulevaluation, GEW-TALIS), dass schulinterne Evaluation „kaum entwickelt wurde und derzeit nicht zu den Stärken schulischer Arbeit gehört". Dieser Befund gilt auch für die USA, das „Mutterland" der Schulevaluation. Hier bleibe eine entwicklungsorientierte schulische Programmevaluation stets sekundär gegenüber den stark verbreiteten standardisierten landesweiten Leistungsmessungen (vgl. King/Rohmer-Hirt 2011, S. 76). Gärtner (2015, S. 34) spricht auf Basis eigener empirischer Untersuchungen zu Brandenburg sowie einer Auswertung internationaler Studien von einem „geringen Anteil nutzbringender interner Evaluation".

Während also *auf der konzeptionellen Ebene* gesetzliche Grundlagen, bildungstheoretische Annahmen und Befunde der Lehr-/Lernforschung dezidiert für interne Evaluation sprechen, so wird sie *in der Praxis* – wenn überhaupt – allzu oft als lästige Pflichtaufgabe abgearbeitet, da ihr kaum konkreter Nutzen für Schule und Unterricht zugetraut wird.

Eine zweite Chance für die interne Evaluation ergibt sich aus der – gerade im deutschsprachigen Raum – breit rezipierten Meta-Meta-Analyse von John Hattie (2014) zu den Einflussfaktoren auf schulische Lernleistungen. Eine seiner zentralen Schlussfolgerungen lautet, dass Lehrpersonen und Schulleitende zu Evaluierenden ihrer eigenen professionellen Tätigkeit werden sollen. Dies betrifft besonders den

starken Faktor „Formative Evaluation des Unterrichts" (Hattie 2015, S. 215). Diese Forderung untermauert das von Unterrichts-, Schul- und Bildungsforschern (z.B. Andreas Helmke, Eckhard Klieme, Hilbert Meyer, Michael Schratz) vertretene Plädoyer für interne Evaluation. Es gilt, interne Evaluation konsequent auf die Verbesserung des Unterrichts und die Stärkung des Lernens (der Schülerinnen und Schüler wie der Lehrkräfte) auszurichten, also auf die Mikro-Ebene des Unterrichts und die Meso-Ebene der unterrichtsübergreifenden Kooperation. Derart lernorientiert Evaluierende sollten sich vor allem beim Aufbau von Evaluationsvermögen und Evaluationskultur in der Schule engagieren (vgl. Dahler-Larson 2009). Es geht um schulinternes Evaluationsvermögen – verstanden als Fähigkeit einer Bildungsorganisation und ihrer Mitglieder, Evaluationen zu planen, durchzuführen und zu nutzen, um daraus individuell und organisational zu lernen. Hierfür gibt es nach Cousins et al. (2014) zwei unabdingbare Voraussetzungen:

- ermöglichende und stützende „Strukturen", d.h. Evaluationsaufgaben in Stellenbeschreibungen der Lehrenden, entsprechende Arbeitszeitkontingente, fixierte Arbeitsgruppen-/Unterrichtsteam-Termine zur Erarbeitung sowie schulweite Kommunikations- und Präsentationskanäle zur Vermittlung von Ergebnissen und zur Einleitung ihrer schulweiten Nutzung;
- Kompetenzen zur Planung und Umsetzung von Evaluationen, die sich Schulinterne aneignen.

Welches Verständnis von interner Evaluation dafür zugrunde zu legen ist und welche Kompetenzprofile je nach Gegenstand der Evaluation und dessen Größe erforderlich sind, behandelt der nachfolgende Abschnitt.

2. Verortung und Varianten interner Schulevaluation

Wie im einleitenden Beitrag von Böttcher und Hense (2016) in diesem Heft wird Evaluation hier als systematisch geplantes und umgesetztes, auf den jeweiligen Gegenstand maßgeschneidertes Beschreibungs- und Bewertungs-Verfahren verstanden, das

- auf spezifische Handlungsfolgen ausgerichtet ist,
- dabei Methoden der empirischen Sozialforschung einsetzt und
- Bewertungen nachvollziehbar auf Kriterien abstützt.

In Bezug auf den von ihr zu verfolgenden Zweck kann eine Evaluation unterschiedliche[1] Prioritäten setzen:

- Formative Evaluation soll zur optimalen „Ausformung" des Evaluationsgegenstandes, z.B. eines bestimmten Lehr-Lernarrangements, beitragen. Sie soll die-

1 Diese Unterscheidung geht auf Michael Scriven (1972/1966) zurück.

sen stabilisieren oder verändern helfen und verfolgt primär den Zweck der Verbesserung.
- Summative Evaluation will einen „Summenstrich", eine Bilanz zum Evaluationsgegenstand ziehen. Typische Zwecke sind die Rechenschaftslegung (z.B. gegenüber den Bildungsbehörden) und die Vorbereitung von Richtungsentscheidungen.

Im Bildungsbereich richten sich Evaluationen vielfach auf *Programme* („Unterrichtssequenzen", „Kurse", „Angebote", „Projekte" usw.; vgl. Abschnitt 4 in Böttcher/Hense 2016). Entsprechend tragen viele englischsprachige Standardwerke *Program Evaluation* im Titel. Programme basieren auf einem (möglichst schriftlichen) Konzept. In der Schule sind solche *„sichtbaren" Programme* (wie z.B. eine Projektwoche oder die Berufswahlorientierung) die Ausnahme gegenüber dem Regelfall des *alltäglichen Unterrichts*. Dabei weisen beide ähnliche Grundelemente auf: Lernausgangslagen, die auf Ziele hin verändert werden sollen, sowie darauf zugeschnittene Methoden/Interventionen, die das Lernvermögen der Schülerinnen und Schüler mobilisieren und schließlich zu Lernresultaten führen sollen. Konzepte für Unterricht und damit das Kernelement der Programmförmigkeit sind zwar gegeben, doch – wie auch der Unterricht selbst – *kaum* für Dritte *sichtbar*. Der „normale" Unterricht einerseits, spezifische schulische Angebote andererseits erfordern daher unterschiedliche Ansätze der Evaluation (siehe Abbildung unten). Dieser Beitrag konzentriert sich auf programm-/projektförmig abgegrenzte bzw. abgrenzbare mittelgroße Evaluationsgegenstände.

Programme im Bildungsbereich werden von *Organisationen* durchgeführt und (mit-)verantwortet. Aus dem für Organisationsförmigkeit konstitutiven Element der Mitgliedschaft (vgl. Kühl 2011) ergibt sich als weitere Unterscheidung die nach Evaluationsarten: zum einen der Fall, dass die Evaluation wesentlich durch Mitglieder[2] der Schule geplant und umgesetzt wird (interne Evaluation), und zum anderen der, dass dies Nicht-Mitglieder der Organisation tun (externe Evaluation).

Für eine bestimmte Evaluation sollte ein Schwerpunkt entweder auf formativ oder summativ gelegt werden. Hieraus ergeben sich Präferenzen für die eine oder andere Art der Durchführung: Die summative Rolle kann durch externe Evaluationsfachleute, die keine Interessenkonflikte haben, glaubwürdiger wahrgenommen werden. Hingegen wird Organisationsinternen ob ihrer genauen Kenntnisse der Schülerschaft, der sozialen Beziehungen und des sozialen Kontextes der Schule eher zugetraut, dass sie mit unmittelbarem Nutzen für die pädagogische Praxis formativ evaluieren können. Oft ist es sinnvoll, intern Evaluierende extern zu beraten, damit

2 Dies sind namentlich Lehrkräfte, Schulleitende, evtl. weitere Angestellte. Sie treten der Schule aus freiem Willen per Arbeitsvertrag bei und können ebenso austreten. Freiwilligkeit ist für Schülerinnen und Schüler zumindest während der Pflichtschuljahre nicht gegeben. Organisationstheoretisch gesehen sind sie keine Mitglieder; evaluationstheoretisch zählen sie zu den Stakeholdern (s.u. Abschnitt 4, 2. Schritt).

sie blinde Flecken aufdecken oder sich aus ihrer Organisationsmitgliedschaft ergebende Herausforderungen besser bewältigen (tatsächliche oder vermeintliche pädagogische Wertkonflikte, Widerstände von Kolleginnen und Kollegen, die Autonomie- bzw. Kontrollverlust befürchten, u.v.m.).

Auf Makroebene, z.B. bei der Kooperation von Grund- mit weiterführenden Schulen, wird besser extern evaluiert, denn die jeweils Internen geraten allzu schnell in Interessenkonflikte. Die Mikro-Ebene des Unterrichts gehört hingegen in die Domäne der internen Evaluation. Dazwischen, auf der Meso-Ebene, sind beide Hauptarten geeignet, die interne allerdings eher, wenn es formativ um Verbesserung geht.

Interne Evaluation wird im Bildungsbereich oft als „Selbstevaluation" bezeichnet (vgl. Buhren 2007; Burkard/Eikenbusch 2000; von Saldern 2010), deren Pendant als „Fremdevaluation".[3] Im Unterschied zur durch (Nicht-)Mitgliedschaft leicht operationalisierbaren Gegenüberstellung extern/intern ist die Unterscheidung fremd/selbst komplex und basiert auf dem Konzept der *Kultur*. Diese Differenz kann sich aus (Professions-)Kulturen und damit verbundenen Werten von Evaluierenden einerseits, von für das pädagogische Programm verantwortlichen Stakeholdern andererseits herleiten.[4]

Wie Abbildung 1 zu entnehmen, gibt es viele Varianten schulinterner Evaluation, je nach der *Größe der zu bearbeitenden Evaluationsgegenstände* und je nach gewählter *Bearbeitungsweise*:

Bezüglich *Größe* des Gegenstands gibt es folgende Pole:
- fokussierte Evaluationen, bezogen auf einen pädagogischen Gegenstand z.B. im Umfang weniger Unterrichtsstunden;
- ausgedehnte Evaluationen zu schulweit oder gar übergreifend relevanten Fragestellungen, die z.B. aus dem Schulprogramm abgeleitet oder durch die Ergebnisse externer Evaluationen/Inspektionen ausgelöst sind.

Besonders letztgenannte richten sich oft auf die Schule als Organisation. Zwischen diesen Polen liegen zahlreiche Evaluationsgegenstände unterschiedlichen Umfangs.

3 Wissenschaftstheoretisch präziser (vgl. Eckensberger 2015) werden „etische" (fremd/generalisierend/standardisiert) von „emischen" (selbst/ideographisch/responsiv) Zugängen unterschieden. Etisch wird der Gegenstand von außerhalb des Systems evaluiert, mit feststehenden Fragestellungen und allgemein gültigen Bewertungskriterien, auch um ihn mit anderen zu vergleichen. Emisch wird von innen her evaluiert, basierend auf systemspezifischen Fragestellungen und Kriterien. Angesichts solch spannungsgeladener Themen wie Koedukation, Inklusion oder leistungsdifferenzierter Klassenbildung ist offensichtlich, welchen Unterschied ein etischer oder emischer Evaluationszugang macht. Dies ist besonders in der deutschsprachigen Forschung zur Schulevaluation kaum theoriebasiert thematisiert.

4 Als „fremd" wird z.B. von pädagogischen Fachpersonen wahrgenommen, wenn ihre Angebote durch eine von Juristen oder Betriebswirtinnen geführte Stelle wie etwa ein Rechnungsprüfungsamt evaluiert werden.

Einige von diesen sind in Spalte 3 der Abb. 1 aufgeführt. Die Größe eines Evaluationsgegenstandes ergibt sich aus seiner zeitlichen Dauer, der Anzahl der als Lehrende oder Lernende oder anderweitig Mitwirkenden, der Menge erforderlicher Abstimmungs- und Übergabepunkte u.a.

Die geeignete *Bearbeitungsweise* hängt mit der Größe des Evaluationsgegenstandes zusammen:

- Ein wenige Unterrichtsstunden umfassender, sehr fokussierter Evaluationsgegenstand kann von einer Lehrkraft (oder einem Tandem) selbständig und beiläufig im Unterricht evaluiert werden. Unterrichten und Untersuchen können *integriert,* in einem Zug durchgeführt werden. In Fortführung der von Schön (1983) für professionelles Lehren geprägten *reflection in action* könnte man von *evaluation in action* sprechen. Es gibt Nähen zur Aktionsforschung (vgl. Altrichter/Posch 2007) oder zur Selbstevaluation als Bestandteil methodischen Handelns (vgl. Bestvater/Beywl 2015). Diese Bearbeitungsweise wird „individuell/Tandem" genannt.
- Größere interne Evaluationen müssen hingegen von speziell dafür zuständigen Lehrpersonen mit reduzierter Unterrichtsverpflichtung oder nicht unterrichtenden Evaluationsfachleuten *arbeitsteilig* bearbeitet werden. Diese tragen keine (Haupt-)Verantwortung für das zu evaluierende Bildungsangebot. In sehr großen Schulen gibt es evtl. eine eigene Stabsstelle „Evaluation". Sie ist für die Planung und Durchführung der Evaluation zuständig, während (andere) Lehrpersonen den Unterricht/die pädagogischen Maßnahmen steuern. Als Interne sind sie dem Wert- und Normensystem der Schule verpflichtet, und das Kollegium verlangt von ihnen Loyalität. Sie evaluieren im Auftrag der Schulleitung und in Abstimmung mit weiteren Beteiligten. Diese Variante der internen Evaluation wird hier zur besseren Unterscheidbarkeit „Inhouse-Evaluation" genannt. Oft werden diese Ansätze (missverständlich) als „Schulevaluation" oder „institutionelle Selbstevaluation" bezeichnet. Diese die pädagogische Arbeit „unterstützende" Bearbeitungsweise wird im Anschluss an Zenkel (2015) als „auxiliar" bezeichnet.
- Für Evaluationsgegenstände in dazwischen liegenden Größenordnungen eignet sich die kollegiale Bearbeitungsweise, angesiedelt zwischen individuell/Tandem und auxiliar. Hier arbeiten drei oder mehr Lehrkräfte zusammen, im Rahmen einer zumindest minimalen Projektorganisation.

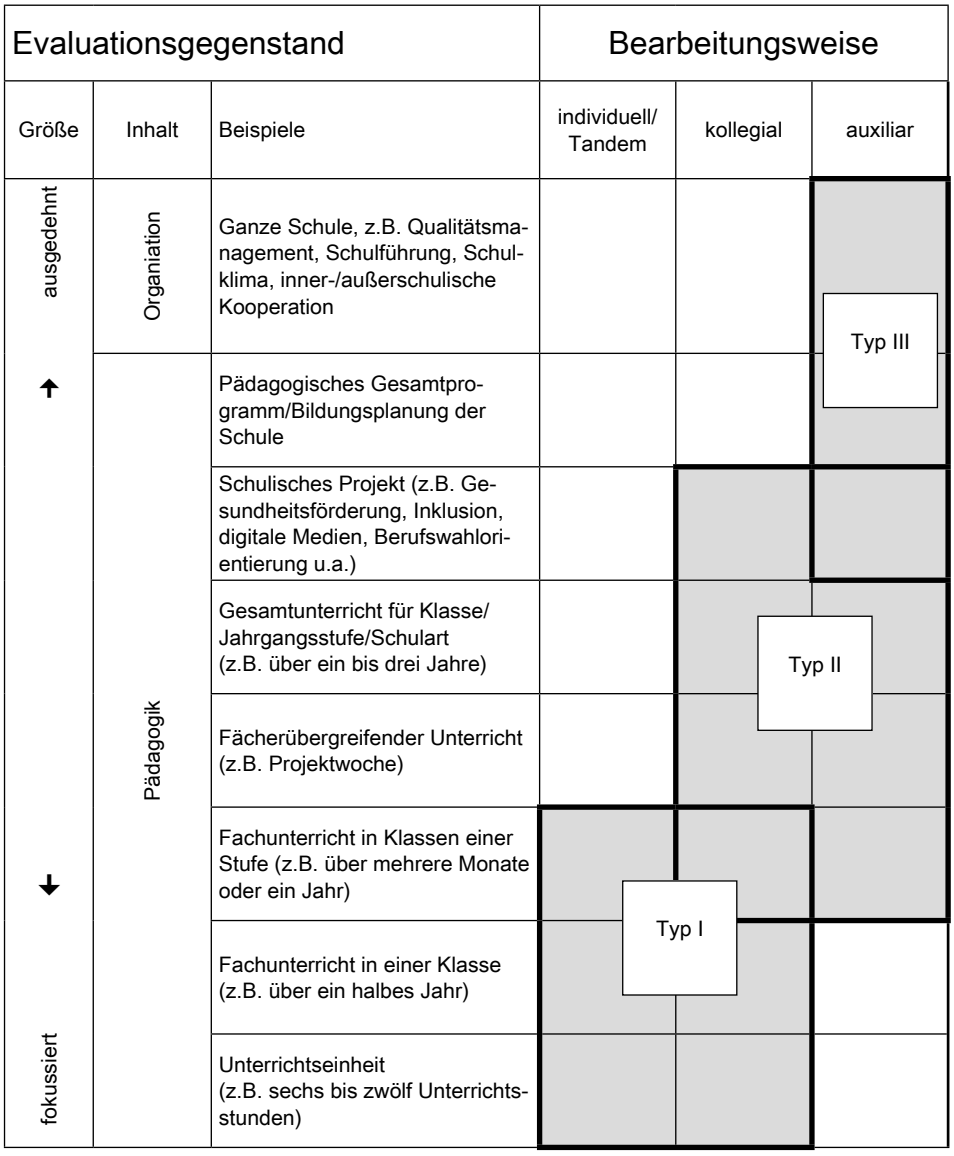
Soweit an Schulen Unterrichtsteams, Fachschaften oder Qualitätsgruppen etabliert sind und es für diese verbindliche und anerkannte Ressourcen, Zeiten und Abläufe gibt, ist es möglich, kollegial zu evaluieren. Wenn jedoch diese strukturellen Voraussetzungen (vgl. Cousins et al. 2014) schwach sind, sind individuelle/Tandem-Ansätze und auxiliar angelegte Evaluationen leichter realisierbar.

Wie Abbildung 1 zu entnehmen ist, ergeben sich durch die Kombination der Merkmale *Bearbeitungsweise* (Unterarten der internen Schulevaluation) und *Größe des Evaluationsgegenstandes* zahlreiche Varianten. Diese können zu den Typen I bis

III zusammengefasst werden, wobei es Überschneidungsbereiche gibt. Die folgende Abbildung kann auch als Prüfschema verwendet werden:

- Welche Bearbeitungsweise passt zum Evaluationsgegenstand?
- Welche Evaluationsgegenstände sind mit gegebenen Ressourcen und Kompetenzen bearbeitbar?

Abb. 1: Varianten interner Schulevaluation



Quelle: eigene Darstellung

3. Kompetenzerwerb für schulinterne Evaluation

Die Annahme, dass für interne Schulevaluation keine spezifischen Kompetenzen erforderlich seien, ist weit verbreitet. So wird z.B. unterstellt, diese würden durch die seit über zehn Jahren aufgebauten Online-Selbstevaluations-Portale[5] überflüssig. Die teils kostenlos verwendbaren Fragebogen und andere „qualitativ hochwertige Instrumente" (vgl. Gärtner 2015) könnten durch Evaluationslaien sicher angewendet werden (vgl. von Saldern 2010). Allerdings ist – wie auch Gärtner festhält – der erhoffte durchschlagende Erfolg in Bezug auf den Aufbau einer lebendigen internen Evaluationskultur in den allermeisten Schulen ausgeblieben.

Hingegen erfordert eine nachhaltige, für Unterricht und Schule nutzenstiftende Evaluation evaluationsspezifische Kompetenzen (vgl. Abschnitt 3 in Böttcher und Hense 2016 in diesem Heft sowie Cousins et al. 2014). In einer Expertenorganisation wie der Schule werden Evaluationsvermögen und -kultur dann entstehen können, wenn möglichst viele der dort tätigen Lehrexperten und -expertinnen Evaluationskompetenz als gelebten Bestandteil ihrer Professionalität betrachten. Evaluationen vom Typ I sind insofern grundlegend. Die gesamte Bandbreite schulinterner Evaluation zu aktivieren, ist Aufgabe der Schulleitung. Ihr Engagement ist ausschlaggebend für den Erfolg (vgl. Hattie/Masters/Birch 2016). Hattie (2014, S. 183) betont als dafür wichtige Haltung: „Lehrpersonen/Schulleitende sind überzeugt, dass ihre fundamentale Aufgabe darin besteht, ihr Lehren und das Lernen und die Lernleistung der Schülerinnen und Schüler wirkungsorientiert zu evaluieren". Es dürfte sehr wenige Fachkräfte geben, die vertiefte Expertise für alle drei der nachfolgend skizzierten Typen interner Evaluation aufweisen. Daher ist in der Regel eine arbeitsteilige Zuständigkeit angeraten.

Typ I Individuell-kollegiale unterrichtsintegrierte Evaluation
Evaluationskompetenz für Typ I ist eine Erweiterung vorhandener didaktischer Experten-Kompetenz, die in mehrjähriger praktischer Unterrichtstätigkeit ausgebaut worden ist. Datenerhebung und -auswertung werden möglichst ohne Zusatzaufwand in den Unterricht integriert. Die Untersuchungen sind so angelegt, dass sie für das Erreichen von Lernzielen, das eigene pädagogische Lernen und den Dialog mit den Lernenden über den Unterricht möglichst unmittelbar Nutzen stiften können. Dies meint Hattie (2014, S. 164) mit „formativer Evaluation des Unterrichts". Doch auch diese „Nebentätigkeit" (vgl. Böttcher/Hense 2016 in diesem Heft) erfordert Kompetenzaufbau und Einüben in die Evaluationspraxis. Der von über 300

5 Um nur einige zu nennen: Der Vorreiter, SEIS, ursprünglich von der Bertelsmann Stiftung gefördert, hat seinen Betrieb nach zwölf Jahren Ende 2015 eingestellt. Das Portal IQUES-Online ist in Deutschland und der Schweiz seit ca. zehn Jahren verbreitet und zeichnet sich u.a. dadurch aus, dass es zahlreiche Materialien zur Unterrichtsentwicklung bereitstellt. Mehrere Bundesländer finanzieren bzw. betreiben seit einigen Jahren eigene Online-Instrumentensammlungen.

Lehrkräften erprobte Ansatz LUUISE (Lehrpersonen unterrichten und untersuchen integriert, sichtbar und effektiv) ermöglicht dies in kollegial gestützter, projektförmiger schulinterner Weiterbildung (siehe Härri 2015; URL: http://tinyurl.com/Luuise; Zugriffsdatum: 11.04.2016). Für die nachhaltige Optimierung des Lehrerhandelns sind derart auf Sichtbarmachen der eigenen Wirksamkeit angelegte Fortbildungen besonders effektiv (vgl. Lipowsky/Rzejak 2015).

Typ II Kollegial-auxiliare unterrichtsübergreifende Evaluation
Zusätzlich zu den Kompetenzen von Typ I sind für Typ II Evaluationskompetenzen erforderlich, welche die Befähigung zur Planung und Umsetzung fokussierter Evaluationen bis hin zu mittelgroßen Projekten umfassen. Sie können in einschlägigen Weiterbildungen erworben und in gecoachten Evaluationsvorhaben erprobt werden. Dabei bietet es sich an, die Weiterbildung inhaltlich-thematisch nach einer für die Evaluationspraxis orientierenden Schrittfolge zu gliedern (vgl. Textkasten), so dass der Transfer in die Praxis unmittelbar anschließen kann.

Typ III Auxiliare Projekt-/Programmevaluation – Ebene Schule
Diese verlangt eine umfassende Evaluationskompetenz. Es geht um die Befähigung, adaptiv für den jeweiligen Evaluationsgegenstand und seinen Kontext die leitenden Evaluationsfragestellungen sowie geeignete Evaluationsmodelle (vgl. Beywl 2011) auszuwählen und diese konzeptionell wie zeitlich in die Schulentwicklung einzupassen. Neben grundständigem didaktischen sind auch ausgeprägtes empirisches Untersuchungswissen und -können sowie Beratungskompetenz erforderlich. Dieses Kompetenzprofil (vgl. Russ-Eft et al. 2008) kann in einem postgradualen wissenschaftlichen (Weiterbildungs-)Studium erworben und in mehrjähriger Evaluationspraxis vertieft werden.

Nachfolgend wird geschildert, wie Kompetenzen für interne Schulevaluationen des Typs II berufsbegleitend erworben werden können. Um dies zu erreichen, wurde die nachfolgend skizzierte zweiteilige Evaluationsweiterbildung konzipiert und erprobt.

Im ersten Teil der Fortbildung wird das evaluative Grundlagenwissen und -können erarbeitet. Dabei sollen die Teilnehmenden die systematische 10er-Schrittfolge für Planung und Durchführung von Evaluationen (vgl. Textkasten) erläutern, begründen und auf ein eigenes Evaluationsprojekt planend anwenden können.

Im *ersten Schritt* wird der *Evaluationsgegenstand* bestimmt. Dafür wird eine kurze, für Außenstehende nachvollziehbare Erstbeschreibung mit allen zentralen Elementen erstellt. Oft gilt es, die durch den Evaluationsgegenstand verfolgten Ziele zu präzisieren und ihnen Interventionen/Methoden/Lehr-Lernarrangements, Sozialformen, Medien/Materialien u.a. zuzuordnen (vgl. den Beitrag von Giel 2016 in diesem Heft, S. 149-162).

Im *zweiten Schritt* werden die am Evaluationsgegenstand interessierten *Akteure* bestimmt. Deren Werte und Interessen bezüglich des Evaluationsgegenstands und ihre Informationsbedarfe an die Evaluation sind zu identifizieren.

Im *dritten Schritt* geht es um die Festlegung der *Evaluationszwecke*: Unterscheidbar sind entscheidungsorientierte, rechenschaftslegungsorientierte, verbesserungsorientierte und wissensgenerierende Evaluation. Die auf die Zwecke abgestimmten *Evaluationsfragestellungen* fassen in Worte, welche Informationen die vorgesehenen Nutzenden benötigen.

Bewertungskriterien als im Vorhinein ausgewiesene Referenzmaßstäbe sind Voraussetzung für jede systematische und faire Bewertung. Sie werden im *vierten Schritt* festgelegt und bieten eine Bezugsbasis, auf die sich Werturteile zum Evaluationsgegenstand stützen.

Im *fünften Schritt* planen die Evaluierenden das methodische Vorgehen. Es gilt, ein passendes *Erhebungsdesign* sowie angemessene *Erhebungsmethoden* auszuwählen und einen Datenerhebungsplan zu entwerfen.

Die empirische Phase der Evaluation beginnt im *sechsten Schritt* mit der *Durchführung der Erhebungen*. Dabei sind die logistische und technische Seite ebenso zu beachten wie der Umgang mit Störungen und Konflikten.

Auf die Erhebungen folgen im *siebten Schritt* die *Auswertung* und *Interpretation* der Daten. Diese werden komprimiert, verbal zusammengefasst sowie in Bezug zu den Fragestellungen in deutende, erklärende Sinnzusammenhänge gestellt und zu Schlussfolgerungen verdichtet. Bewertungen und *Bewertungssynthese* erfolgen in Referenz auf die Bewertungskriterien.

Damit eine Evaluation Nutzen für Schule und Unterricht entfalten kann, muss im *achten Schritt* über ihre Ergebnisse *berichtet* werden: mündlich oder schriftlich, in Workshops, auf Papier oder mittels elektronischer Medien.

Evaluationen lohnen sich dann, wenn sie für und durch die schulische Praxis *genutzt* werden, was im *neunten Schritt* eingeleitet wird. Es geht nicht allein um die Verwendung der *Ergebnisse*, sondern auch darum zu klären, was der Evaluations*prozess* bereits an Nützlichem für die Bildungspraxis erbracht hat.

Im abschließenden *zehnten Schritt* kann die Evaluation selbst bewertet werden. Mit der *Evaluation der Evaluation* ist diese abgeschlossen, und der Evaluationszyklus schließt sich, um bei Bedarf erneut zu beginnen.

Zwei didaktische Prinzipien leiten diesen ersten Teil der Weiterbildung: Mit dem „umgedrehten Unterricht", dem *flipped classroom* (vgl. Bergmann/Sams 2012), wird die übliche Aufteilung dessen, was innerhalb (Wissensvermittlung) und außerhalb (Transfer/Anwendung) des Schulungsraumes stattfindet, „umgedreht". Die Teilnehmenden erarbeiten sich das Evaluationswissen im Vorfeld und außerhalb der Präsenzphasen – selbstgesteuert, asynchron, im eigenen Lerntempo und ortsunabhängig. Dies geschieht durch die Auseinandersetzung mit dem an der 10er-Schrittfolge orientierten Lehrtext mit kapitelweisen Lernzielen, Fallbeispielen sowie Übungsaufgaben mit Beispiellösungen (vgl. Balzer/Beywl 2015). Als zweites didaktisches Prinzip wird der „Projektmethode" folgend (vgl. Frey 2012, S. 218-234; Pfäffli 2015) ein – wenn möglich echtes – selbst oder z.B. von der Schulleitung in Auftrag gegebenes Evaluationsvorhaben geplant.

Während der zwei bis drei Präsenzphasen à 1 bis 1,5 Tage wird kaum noch Stoff vermittelt. Hingegen vertiefen und übertragen hier die Teilnehmenden das theoretisch Erarbeitete auf ihre eigene Bildungspraxis. Sie erstellen einen Evaluationsplan für ein konkretes Vorhaben in ihrem schulischen Kontext. Dabei arbeiten sie alleine, in Tandems oder in Kleinstgruppen und werden von den Kursleitenden beraten. Diese lenken den Lern- und Planungsprozess durch Arbeitsaufträge, die auf die Schritte des Evaluationsprozesses abgestimmt sind. Dabei auftauchende Fragen und Probleme können an den realen Beispielen der Teilnehmenden diskutiert werden. In selbstorganisierten Lerngruppen – hier agieren die Teilnehmenden präsent, per Telefon oder online – können darüber hinaus Fragen geklärt werden.

Die Präsenzphasen haben je ca. vierwöchigen Abstand zueinander. Damit bleibt dazwischen genügend Zeit für die Auseinandersetzung mit den neuen Lerninhalten bzw. das Lesen der entsprechenden Buchkapitel sowie für Vertiefungen und Anwendungen. Parallel beantworten die Kursleitenden Rückfragen via E-Mail oder Lernplattform.

Mit diesem Vorgehen wird die Kompetenzvermittlung direkt mit der praktischen Umsetzung verbunden. Es werden für die Schul- bzw. Unterrichtspraxis nützliche Evaluationen vorbereitet. Am Schluss des letzten Präsenztages verfügen alle Teilnehmenden (evtl. als Tandem) über einen umsetzbaren Evaluationsplan, eventuell bereits ergänzt um ein vorgetestetes Erhebungsinstrument. Dieser ist Grundlage für den zweiten Teil der Evaluationsweiterbildung, mit dem der Transfer in die Praxis sichergestellt wird.

Soll das empirische Methodenwissen ebenfalls erworben werden, sind zusätzliche, deutlich zeitintensivere Zusatzweiterbildungen einzuplanen. Oft wird es angeraten sein, dass sich Schulinterne etwa zur Konzeption eines Interviewleitfadens oder zur Konstruktion eines Fragebogens an eine evaluatorische (Methoden-)Beratung wenden, die über Kompetenzen des Typs III verfügt.

Im zweiten Teil folgt die Durchführung des geplanten Evaluationsprojektes in der Schule (teils beginnt diese schon im Verlauf des ersten Teils). Dabei ist auch zu klären, wieviel Wissen und Können insbesondere im empirisch-methodischen Bereich (Fragebogenerstellung/Interviewdurchführung) schulintern vorhanden ist, um die erforderlichen Datenerhebungsinstrumente zu erarbeiten und einzusetzen und um die resultierenden Daten auszuwerten. Je nach schulintern vorhandenen Kompetenzen und Ressourcen kann das Evaluationsprojekt mit mehr oder weniger intensiver externer Unterstützung durchgeführt werden.

An die (erstmalige) Durchführung kann sich eine Folge-Weiterbildung anschließen, in der die eigene Evaluationspraxis kritisch reflektiert und daran anschließend weitere oder vertiefte Evaluationskompetenzen aufgebaut werden.

Resümee

Dieser Beitrag skizziert einen Weg, wie das bislang enttäuschend wenig genutzte Potenzial der internen Schulevaluation für Bildung und Lernen entfaltet werden kann. Als zentrale Gelingensbedingung wird die Erweiterung der Evaluationskompetenzen von Lehrpersonen angesprochen. Über das Evaluations-*Wissen* und das handwerkliche Untersuchen-*Können* hinaus geht es um die Geistes-*Haltung*, dass systematisches Evaluieren zur pädagogischen Arbeit dazugehört. In praxisintegrierten Fortbildungen kann es gelingen, ein solches *Mindset* von Bildungsprofessionellen zu stärken, wenn der Nutzen des Evaluierens für die pädagogische Praxis nach kurzer Zeit eintritt und wenn er sich als nachhaltig erweist. Wie jede Bildungsanstrengung verlangt dies Investitionen, d.h., insbesondere Zeit für die eigene Fortbildung sowie Anreize im Berufsauftrag der Lehrkräfte.

Literatur und Internetquellen

Altrichter, H./Posch, P. ([4]2007): Lehrerinnen und Lehrer erforschen ihren Unterricht. Unterrichtsentwicklung und Unterrichtsevaluation durch Aktionsforschung. Bad Heilbrunn: Klinkhardt.

Balzer, L./Beywl, W. (2015): evaluiert. Planungsbuch für Evaluationen im Bildungsbereich. Bern: hep.

Bergmann, J./Sams, A. (2012): Flip Your Classroom: Reach Every Student in Every Class Every Day. Washington: International Society for Technology in Education.

Bestvater, H./Beywl, W. (2015): Gelingensbedingungen der Selbstevaluation. In: Bolay, E./Iser, A./Weinhardt, M. (Hrsg.): Methodisch Handeln – Beiträge zu Maja Heiners Impulsen zur Professionalisierung der Sozialen Arbeit. Wiesbaden: VS, S. 133-144.

Beywl, W. (2011): Modelle der Evaluation personenbezogener Dienstleistungen. In: Beck, I./Greving, H. (Hrsg.): Gemeindeorientierte Dienstleistungen. Behinderung, Bildung, Partizipation. Stuttgart: Kohlhammer, S. 169-173.

Böttcher, W./Hense, J. (2016): Evaluation im Bildungswesen – eine nicht ganz erfolgreiche Erfolgsgeschichte. In: Die Deutsche Schule 108, H. 2, S. 117-135.

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

Burkard, C./Eikenbusch, G. (2000): Praxishandbuch Evaluation in der Schule. Berlin: Cornelsen Scriptor.

Cousins, J.B./Goh, S.C./Elliott, C.J./Bourgeois, I. (2014): Framing the Capacity to Do and Use Evaluation. In: New Directions for Evaluation, H. 141, S. 7-23.

Dahler-Larson, P. (2009): Learning-oriented Educational Evaluation in Contemporary Society. In: Ryan, K.E.C./Bradley, J. (Hrsg.): The SAGE International Handbook of Educational Evaluation. Thousand Oaks, CA: Sage, S. 307-322.

Eckensberger, L.H. (2015): Integrating the Emic (Indigenous) with the Etic (Universal) – A Case of Squaring the Circle or for Adopting a Culture Inclusive Action Theory Perspective. In: Journal for the Theory of Social Behaviour 45, H. 1, S. 108-140.

Frey, K. ([12]2012): Die Projektmethode. Der Weg zum bildenden Tun. Weinheim: Beltz.

Gärtner, H. (2015): Feedbackkultur auf mehreren Ebenen. In: SchulVerwaltung spezial 17, H. 1, S. 34-35.

Giel, S. (2016): Vom Nutzen der Programmtheorie in Evaluationen im Schulkontext. In: Die Deutsche Schule 108, H. 2. S. 149-162.

Härri, R. (2015): Evidenzbasierte Unterrichts- und Schulentwicklung – Wenn Lehrpersonen zu Selbstevaluatoren werden und Unterrichtserfolge sichtbar machen. In: Erziehung und Unterricht 166, H. 1-2, S. 156-163.

Hattie, J.A.C. (2014): Lernen sichtbar machen für Lehrpersonen. Überarbeitete deutschsprachige Ausgabe von „Visible Learning for Teachers". Besorgt von Wolfgang Beywl und Klaus Zierer. Baltmannsweiler: Schneider Verlag Hohengehren.

Hattie, J.A.C. ([3]2015): Lernen sichtbar machen. Überarbeitete und erweiterte deutschsprachige Ausgabe von „Visible Learning". Besorgt von Wolfgang Beywl und Klaus Zierer. Baltmannsweiler: Schneider Verlag Hohengehren.

Hattie, J.A.C./Masters, D./Birch, K. (2016): Visible Learning into Action. International Case Studies of Impact. London: Routledge.

Huber, S.G./Wolfgramm, C./Kilic, S. (2013): Vorlieben und Belastungen im Schulleitungshandeln: Ausgewählte Ergebnisse aus der Schulleitungsstudie 2011/2012 in Deutschland, Österreich, Liechtenstein und der Schweiz. In: Jahrbuch Schulleitung, S. 259-271.

IDES (Informationen Dokumentation Erziehung Schweiz) (2007): Evaluation der Bildungseinrichtungen und des Bildungssystems. Kapitel 9 des Schweizer Beitrags für die Datenbank „Eurybase – The Database on Education Systems in Europe". URL: http://www.ides.ch/dyn/bin/12961-13438-1-eurydice_09d.pdf; Zugriffsdatum: 08.03.2016.

King, J.A./Rohmer-Hirt, J.A. (2011): Internal Evaluation in American Public School Districts: The Importance of Externally Driven Accountability Mandates. In: New Directions for Evaluation, H. 132, S. 73-86.

KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland) (2005): Bildungsstandards der Kultusministerkonferenz. München: Wolters Kluwer.

Kühl, S. (2011): Organisationen – eine sehr kurze Einführung. Wiesbaden: VS.

Lipowsky, F./Rzejak, D. (2015): Wenn Lehrer zu Lernern werden – Merkmale wirksamer Lehrerfortbildung. In: Lin-Klitzing, S./Di Fuccia, D./Stengl-Jörns, R. (Hrsg.): Auf die Lehrperson kommt es an? Beiträge zur Lehrerbildung nach John Hatties „Visible Learning". Bad Heilbrunn: Klinkhardt, S. 144-160.

Nevo, D. (2009): Accountability and Capacity Building: Can they Live Together? In: Ryan, K.E./Cousins, J.B. (Hrsg.): The SAGE International Handbook of Educational Evaluation. Los Angeles, CA: Sage, S. 291-303.

Pfäffli, B.K. ([2]2015): Lehren an Hochschulen. Eine Hochschuldidaktik für den Aufbau von Wissen und Kompetenzen. Bern: Haupt.

Pietsch, M. (2011): Die Evaluationspraxis an deutschen Schulen. Ein Überblick aus empirischer Sicht. In: Schulmanagement 42, H. 4, S. 12-14.

Russ-Eft, D.F./Bober, M.J./de la Teja, I./Foxon, M./Koszalka, T.A. (2008): Evaluator Competencies: Standards for the Practice of Evaluation in Organizations. San Francisco, CA: Jossey-Bass.

Schön, D.A. (1983): The Reflective Practitioner. How Professionals Think in Action. New York: Basic Books.

Scriven, M. (1972): Die Methodologie der Evaluation. In: Wulf, C. (Hrsg.): Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München: Piper, S. 60-91 (engl. Originalausgabe 1966: The Methodology of Evaluation. Lafayette, IN: Purdue University).

von Saldern, M. (Hrsg.) (2010): Selbstevaluation von Schule: Hintergrund – Durchführung – Kritik. Nordersted: BoD.

Zenkel, S. (2015): Selbstevaluation und neue Autonomie der Schule. Kritische Anmerkungen zu vermeintlichen Selbstverständlichkeiten. Berlin: LIT.

Wolfgang Beywl, Prof. Dr., geb. 1954, Leiter der Professur für Bildungsmanagement sowie Schul- und Personalentwicklung an der Pädagogischen Hochschule Nordwestschweiz; wissenschaftlicher Leiter von Univation – Institut für Evaluation, Köln.
Anschrift: PH FHNW, Bahnhofstrasse 6, 5210 Windisch, Schweiz
E-Mail: wolfgang.beywl@fhnw.ch

Lars Balzer, Dr., geb. 1972, Leiter der Fachstelle Evaluation am Eidgenössischen Hochschulinstitut für Berufsbildung.
Anschrift: EHB IFFP IUFFP, Kirchlindachstrasse 79, 3052 Zollikofen, Schweiz
E-Mail: lars.balzer@ehb-schweiz.ch

DDS – Die Deutsche Schule
108. Jahrgang 2016, Heft 2, S. 205-215

Hans Merkens

Evaluation in Schulen – Notwendigkeit und Grenzen

Zusammenfassung
Schulen werden gegenwärtig in vielen Bundesländern extern evaluiert und sind zusätzlich gefordert, sich intern zu evaluieren. Dazu wird ihnen Unterstützung von Dritten angeboten. Im Beitrag werden Formen der Evaluation und eine Abgrenzung zum Monitoring sowie das Verhältnis zur Schulinspektion dargestellt. Die mögliche Funktion von Evaluation im Rahmen der Neuen Steuerung wird erläutert. Nachdem Standards für die Evaluation aus zwei verschiedenen Perspektiven benannt worden sind, werden abschließend einige Beispiele zur Praxis der Evaluation angeführt.
Schlüsselwörter: externe Evaluation, interne Evaluation, Schulinspektion, Monitoring, Neue Steuerung, Standards für Evaluation

Evaluation in Schools – Necessity and Limits

Summary
Presently, schools are evaluated in many States of the Federal Republic of Germany externally and are in addition demanded to evaluate themselves internally. Support of third parties is offered to them. In this contribution forms of evaluation and a demarcation to monitoring as well as the relation to school inspection are shown. The possible function of evaluation within the scope of governance is explained. After evaluation standards have been named from two different perspectives, finally some examples are given for the practice of evaluation.
Keywords: external evaluation, internal evaluation, school inspection, monitoring, governance, evaluation standards

1. Einleitung

Erträge von Bildungseinrichtungen werden in Deutschland seit den letzten Jahren verstärkt kontrolliert (vgl. Döbert/Dedering 2008). Das war international bereits lange der Fall. Untersuchungen zu *School Effectiveness* und *School Improvement* sowie ähnlichen Themenstellungen haben eine Tradition (vgl. Scheerens/Creemers 1989;

Sammons 1999; Mortimore 1998; Reynolds et al. 2014). Im Bildungssystem wurde in Deutschland im 21. Jahrhundert von der Input- zur Output-Steuerung gewechselt. Aus diesem Grund muss man sich vermehrt der Ergebnisse des Arbeitens im Bildungssystem und der einzelnen Schulen versichern. In einem allgemeinen Verständnis dient dem Evaluation (vgl. Abs/Klieme 2005, S. 45).

2. Begriffsbestimmung

Eine klassische Beschreibung findet sich bei Gronlund (1968, S. 1f.), der Evaluation über das methodische Vorgehen in vier Schritten erfasst hat. Danach muss erstens die Bestimmung der Ziele erfolgen, zweitens müssen diese anschließend operationalisiert werden, es schließt sich drittens die Auswahl bzw. Konstruktion der Messinstrumente an; diese werden eingesetzt und ausgewertet. Abschließend kann viertens festgestellt werden, wieweit die Schüler und Schülerinnen die intendierten Ziele erreicht haben. Heute sind systematische Definitionen üblich. Pragmatisch hat die Gesellschaft für Evaluation (DeGEval 2004, S. 14) bestimmt: „Evaluation ist die systematische Untersuchung des Nutzens oder Wertes eines Gegenstandes." Ähnlich hat Köller (2015, S. 330) Evaluation „als jegliche Art der zielgerichteten und zweckorientierten Festsetzung des Wertes einer Sache" definiert. Kuper (2005, S. 7) hat dem noch den Anwendungsaspekt hinzugefügt und drei Dimensionen unterschieden: „Jede Evaluation beinhaltet an Erfahrung gebundene Aussagen über Tatsachen, normative Aussagen als Grundlage für Bewertungen und prognostische Aussagen in der Form von Entscheidungen über zukünftiges Handeln." Während Köller (vgl. 2015, S. 333) den Graben zwischen Aussagen über Tatsachen und normativen Aussagen zu schließen versucht, indem er einen theoretischen bzw. konzeptionellen Überbau für jede Evaluation als notwendig erachtet, hält Kuper (vgl. 2005) das Verhältnis zwischen den verschiedenen Aussagetypen für ungelöst. Beide Varianten lassen erkennen, dass mit Evaluation nicht der Anspruch verbunden wird, eine Tatsache als solche abzubilden, sondern dass die Tatsache aus einem bestimmten Blickwinkel erfasst werden soll. Dessen Auswahl bedarf der Begründung. Anschließend kann dann festgestellt werden, ob man mit dem Ergebnis zufrieden ist oder eine Verbesserung für erforderlich hält. Wie diese zu erreichen ist, kann in der Regel der Evaluation nicht entnommen werden.

3. Zur Historie von Evaluationen

Evaluationen haben in den USA eine lange Tradition. Dabei waren zunächst Messungen von Schülerleistungen üblich. Guba und Lincoln (1989) haben ein Generationenmodell mit vier Phasen dargestellt (vgl. Kuper 2005, S. 28ff.; von Kardorff 2010, S. 241f.). Danach wurde in der ersten Phase gemessen; es wurden beispiels-

weise Schülerleistungen erhoben. In der zweiten Phase, die sie mit Beschreibung gekennzeichnet haben, ging es im Wesentlichen um die Erfassung und Erfüllung pädagogischer Programme. Die dritte Phase war durch eine Verbindung mit sozial- und bildungspolitischen Reformprogrammen gekennzeichnet; das Ziel war Politikberatung. Die letzte Phase, der sie sich auch verbunden fühlen, wird von ihnen als responsiv gekennzeichnet; hier steht der Einbezug der Akteure im Feld, das evaluiert wird, im Mittelpunkt. Während anfangs bei der Datenerhebung ausschließlich quantitative Methoden eingesetzt worden sind, hat vor allem beim vierten Schritt der Einbezug qualitativer Methoden zugenommen. Diese Phasen verlaufen auch parallel zueinander weiter; es geht in erster Linie darum, wie sich die Perspektiven im Laufe der Zeit von Phase zu Phase erweitert haben. Deutlich wird allerdings, dass nicht von einem einheitlichen Verständnis von Evaluation ausgegangen werden kann.

3.1 Formative und summative Evaluation

Die Unterscheidung von formativer und summativer Evaluation geht auf Scriven (1967) zurück, der sie im Kontext der Curriculumentwicklung eingeführt hat. Es handelt sich nach der Einteilung von Guba und Lincoln (1989) um Phasen der Beschreibung: Formative Evaluation bezeichnete Evaluationen im Prozess der Curriculumentwicklung; summative Evaluation bezog sich auf die Evaluation des Curriculums als Produkt. Köller (2015) hat die formative Evaluation – auf den Unterricht bezogen – als Prozessevaluation bezeichnet. Mit diesem Hinweis wird ein Bezug zum Unterricht sichtbar, bei dem klassisch zwischen Unterrichtsplanung, -durchführung und -nachbereitung unterschieden werden kann (vgl. Merkens 2010, S. 69ff.). Alle drei Phasen können evaluiert werden. Die summative Evaluation zählt zur dritten Phase. Formativen Evaluationen kommt darüber hinaus im Prozess der Schulentwicklung eine große Bedeutung zu.

3.2 Interne versus externe Evaluation

Der im Zusammenhang mit der Evaluation von Schulen gerne gebrauchte Begriff Selbstevaluation wird von einigen Autoren abgelehnt, weil nicht eindeutig zu definieren sei, wer mit „selbst“ gemeint sein könnte (vgl. Abs/Klieme 2005, S. 48; Berkemeyer/Müller/van Holt 2016, S. 212). Im englischen Sprachraum ist der Terminus „self-evaluation“ geläufig (vgl. Chapman/Sammons 2013). Wenig umstritten ist die Unterscheidung zwischen interner und externer Evaluation. Bei einer Befragung von Schulleitungen hat sich in Österreich herausgestellt, dass die interne Evaluation als wichtiger betrachtet wird, als dies für Ergebnisse externer Evaluationen zutrifft (vgl. Altrichter/Kemethofer 2015, S. 297). Internen Evaluationen kommt z.B. im Rahmen der Schulentwicklung eine große Bedeutung zu. Das gilt sowohl für die Evaluation des IST-Standes am Beginn als auch für die Begleitung des Prozesses und auch für

die Auswertung am Ende des Prozesses. Sie weisen in diesem Format eine hohe Ähnlichkeit zur Survey-Feedback-Methode auf, die in der Organisationsentwicklung eine Tradition hat (vgl. French/Bell 1990). Externe Evaluation wird von wissenschaftlichen Instituten, privaten Unternehmen oder Experten und Expertinnen in Schulen (vgl. Stockmann 2010, S. 22) oder von einer Bildungsverwaltung durchgeführt. Durch sie soll geprüft werden, wieweit vorgegebene Ziele sowohl bei der Organisation als auch bei den Ergebnissen einer Schule erreicht werden. Sie kann ein Auslöser für Schulentwicklung sein (vgl. Husfeldt 2011; Dedering 2012; van Ackeren/Klemm/Kühn 2015). Ziel von interner und externer Evaluation ist das Gewinnen von Steuerungswissen. Allerdings ist die Unterscheidung zwischen interner und externer Evaluation nicht immer trennscharf, weil auch bei internen Evaluationen externe Berater und Beraterinnen hinzugezogen und Instrumente eingesetzt werden können, die extern entwickelt worden sind.

4. Abgrenzungen

Evaluation wird gegenwärtig in vielen Fällen synonym mit Monitoring bzw. Schulinspektion verwendet (vgl. Pietsch 2010). Mit Monitoring wird allgemein die systematische Erfassung, Beobachtung, Messung bzw. Überwachung eines Vorgangs oder eines Prozesses bezeichnet. Systemmonitoring hat gegenwärtig im Bildungssystem Konjunktur. Im Unterschied zur Evaluation entfällt der Anspruch zu bewerten; dieser ist nur implizit insoweit enthalten, als die Auswahl der zu beobachtenden Tatsachen eine Wertung nach Wichtigkeit enthält, wenn z.B. Schülerleistungen im Bildungssystem stichprobenartig erhoben werden, wie das bei den internationalen Vergleichsstudien PISA (vgl. Deutsches PISA-Konsortium 2001) oder IGLU (vgl. Bos et al. 2003) der Fall ist, oder wenn über das Bildungssystem berichtet wird (vgl. Autorengruppe Bildungsberichterstattung 2014). Schulinspektionen sind für die einzelne Schule eine Variante der externen Evaluation (vgl. Holtappels 2008). Sie werden auch von den Initiatoren so verstanden (vgl. z.B. Schulinspektion Berlin 2016).

5. Evaluation und Neue Steuerung

Evaluation kann im Rahmen der Neuen Steuerung eine zentrale Funktion zukommen. Das trifft insbesondere für den Fall der evidenzbasierten Steuerung zu (vgl. Altrichter/Kemethofer 2015) und hängt mit der Hinwendung zur Output-Steuerung (vgl. Fend 2011, S. 4) sowie der zunehmenden Gewährung von Autonomie für die einzelne Schule zusammen (vgl. Dubs 2011; Abs/Klieme 2005, S. 46). Mit Evaluationen soll dann auf der Makroebene des Bildungssystems, aber auch auf der Mikroebene der einzelnen Schule die Wirksamkeit von Maßnahmen nachgewiesen werden, indem die Steigerung der Lernergebnisse als Erfolgskriterium dient (vgl.

Fend 2011, S. 9). Der Schwerpunkt der bisherigen empirischen Forschung liegt dabei auf der Mikroebene der einzelnen Schule. Hier lassen sich Quantität und Qualität der Bildungsangebote am einfachsten überprüfen. Kyriakides und Creemers (2008) haben mit einem multidimensionalen Modell den Einfluss von Faktoren auf der Klassenebene auf Schülerleistungen in Mathematik, Griechisch sowie Religion bestätigt. Viel schwieriger ist es allerdings, die Wirksamkeit von Maßnahmen der Steuerung im Mehrebenensystem des Schulwesens nachzuweisen (vgl. Maag Merki 2016; Fend 2011). Neben Schwierigkeiten der praktischen Realisierung hat Fend (ebd., S. 12) vor allem auf die theoretische Frage hingewiesen, wie Mechanismen, die erwünschte Steigerungen der Lernergebnisse bewirken sollen, identifiziert werden können. Ein Blick auf den Stand der Governance-Forschung belegt, dass die Annahme einer einfachen Wirkungskette nicht tragfähig ist (vgl. Altrichter/Maag Merki 2016; Berkemeyer 2010). Davon zu unterscheiden ist die Steuerung in der einzelnen Schule (vgl. Berkemeyer/Müller/van Holt 2016). Hier zeichnet sich ab, dass es vermehrt über die Nutzung von Feedback-Systemen zu Annäherungen an Evaluation kommen wird (vgl. Altrichter/Moosbrugger/Zuber 2016, S. 237).

6. Standards für die Evaluation

Dem Aspekt der Standards ist sowohl national als auch international Aufmerksamkeit geschenkt worden (vgl. z.B. Abs/Klieme 2005). Bei den Standards der Evaluation wird immer wieder auf die Arbeiten des Joint Committee (1994) verwiesen. Für Deutschland hat die DeGEval diese Standards übersetzt und leicht an die hiesigen Bedürfnisse angepasst (vgl. DeGEval 2004). Es handelt sich um ein additives Konzept, in dem aufgelistet wird, welche Standards einzuhalten sind und welche Kategorien in den einzelnen Dimensionen beachtet werden sollen. Für jede dieser Dimensionen werden Standards benannt. Die Reihenfolge der vier Dimensionen Nützlichkeit, Durchführbarkeit, Fairness und Genauigkeit lässt erkennen, dass der Anwendungsbezug bei diesen Standards dominiert. In der Tradition von Gronlund (1968) wird das methodische Vorgehen spezifiziert. Das steht in einem gewissen Spannungsverhältnis zu der von Sozialwissenschaftlern und Sozialwissenschaftlerinnen verfassten Literatur zur Evaluation bzw. deren Durchführung, bei der oft methodische und methodologische Aspekte obsiegen (vgl. Köller 2015). Vor diesem Hintergrund überrascht nicht, dass auch das Verständnis für Standards bei Evaluationen variiert. Abs und Klieme (vgl. 2005, S. 58) haben z.B. das Einhalten von Standards beim Einsatz von Leistungstests noch am ehesten als erfüllbar angesehen. Ihrer kritischen Bilanz ist aber zu entnehmen, dass allgemein das Einhalten von Standards bei der Evaluation von Schulen nicht erfüllbar zu sein scheint. Dabei interpretieren sie Standards aus der Sicht der Sozialwissenschaften. Das muss in einem gewissen Widerspruch zu den Erwartungen der Auftraggeber von Evaluationen einerseits und denen der Nutzer von Ergebnissen andererseits stehen.

Mit den Standards hängt die Auswahl der Methoden eng zusammen, die zur Gewinnung der Daten eingesetzt werden. Ursprünglich wurde die abhängige Variable als Schülerleistung mit quantitativen Methoden erhoben. Es wurden in der Regel Schulleistungstests eingesetzt. Diese Form der Messung ist auch heute noch üblich; allerdings hat sich eine wesentliche Änderung bei der Bezugsnorm ergeben. An die Stelle der normorientierten Messung ist eine kriteriale getreten. Es geht nicht mehr darum, eine relative Position in einer Normalverteilung zu bestimmen, sondern es wird angestrebt, das Erreichen bzw. den Grad des Erreichens von Lernzielen zu messen. Die angewendeten statistischen Methoden sind im Zuge dieser Veränderung immer elaborierter geworden (vgl. Köller 2015). Das weist den Nachteil auf, dass für Nichtsozialwissenschaftler und -sozialwissenschaftlerinnen die Nutzung der Daten erschwert wird. Generell kann davon ausgegangen werden, dass bei externen Evaluationen quantitative Methoden bevorzugt werden, weil auf diese Weise die Vergleichbarkeit der Ergebnisse von verschiedenen Untersuchungseinheiten leichter zu sichern ist.

Bei internen Evaluationen einzelner Schulen werden demgegenüber häufig qualitative Methoden eingesetzt. Das Spektrum reicht dabei von teilnehmender Beobachtung im Unterricht bzw. beim Schulleben über Interviews mit Personen – Schulleitung, Lehrkräften, Schülern und Schülerinnen, Eltern – und Gruppendiskussionen bis hin zu Dokumentenanalysen (Schulprogramm). Allerdings wird auch für diesen Typ von Evaluationen zunehmend ein Instrumentarium zur Verfügung gestellt, das eine Auswertung mit quantitativen Methoden ermöglicht. Das reicht von Fragebögen mit geschlossenen Antwortformaten bzw. wenigen offenen Fragen bis hin zu Anleitungen für die Kategorisierung von Beobachtungen im Schulleben bzw. Unterricht (vgl. Pietsch 2010). Während bei den quantitativen Methoden die Sicherung methodischer Standards im Zentrum steht, wird bei den qualitativen Methoden die Annäherung an die Besonderheiten des Untersuchungsobjektes angestrebt. In beiden Fällen ist die Nachvollziehbarkeit der Interpretationen eine wesentliche Bedingung. Dazu gibt es bei den qualitativen Methoden entsprechende Anleitungen, z.B. in Form der Auswertungsstrategie der qualitativen Inhaltsanalyse (vgl. Mayring 2010).

7. Zur Praxis der Evaluation

Abs und Klieme (vgl. 2005, S. 48ff.) unterscheiden bei der Evaluation zwischen einem Forschungs-, einem Entwicklungs- und einem Legitimations- sowie Kontrollparadigma. Während bei dem Forschungsparadigma die Idealform das Experiment darstellt (vgl. auch Köller 2015), wird bei dem Entwicklungsparadigma die Verbesserung der Institution angestrebt, und bei dem dritten Paradigma geht es im Kern um Rechtfertigung. Im Folgenden wird dem ersten dieser Paradigmen keine große Aufmerksamkeit geschenkt.

Gegenwärtig gibt es eine Flut von Evaluationen. Evaluation zählt in Schulen bereits zum Alltag: Sie ist in vielen Bundesländern in den Schulgesetzen vorgeschrieben und wird intern oder extern durchgeführt. Dabei sind häufig die Schule sowie die Organisation des Unterrichts von Interesse. Eine praxisorientierte Anleitung haben Chapman und Sammons (2013) verfasst. Sie enthält viele Anregungen, was warum zu veranlassen ist.

Zur Unterstützung der Schulen bei ihren internen Evaluationen ist in Deutschland ein Markt entstanden, in dem Externe ihre Unterstützung für die Durchführung von Evaluationen anbieten. Am sichtbarsten hatte sich die Bertelsmann Stiftung mit ihrem Instrument SEIS positioniert (vgl. Holland 2009). Dieses Instrument steht seit 2016 nicht mehr zur Verfügung. In vielen Fällen ziehen Schulen externe Berater und Beraterinnen hinzu.

Auf der Basis gesetzlicher Vorschriften müssen Schulen heute Schul- und Unterrichtsqualität kontrollieren. Maritz et al. (2006) haben ein Bewertungsbuch für Schulen auf der Basis der *European Foundation for Quality Management* (EFQM) vorgelegt, das den Vorteil bietet, verschiedene Facetten von Qualität zu berücksichtigen. Bei der praxisbezogenen Anleitung zu diesem Qualitätsmanagementsystem wird der Bezug zur Evaluation deutlich. Zur Evaluation werden z.B. in Berlin sowohl interne Evaluationen durchgeführt als auch Ergebnisse der externen Evaluation in der Form des Abschneidens bei Vergleichsarbeiten sowie der Auswertung der Berichte von Schulinspektionen herangezogen (vgl. Senatsverwaltung für Bildung, Jugend und Wissenschaft 2013).

Entgegen den bisherigen Erläuterungen zur Evaluation ist die Steuerung des Lernerfolgs bei den Schülern und Schülerinnen in vielen Fällen kein zentrales Thema. Oft werden vielmehr andere Aspekte des Schullebens evaluiert. Viele Evaluationen dienen der Zielsetzung, eventuell als notwendig angesehene Verbesserungen auf der Schulebene anzustoßen. Ko, Hallinger und Walker (2015) haben in Hongkong diese Zielsetzung als zu anspruchsvoll empfunden: Sie konnten erstens Effekte eher auf der Fachbereichs- als auf der Schulebene nachweisen und haben zweitens Daten aus Längsschnittuntersuchungen für notwendig erachtet, wenn Effekte bei den Schülerleistungen überprüft werden sollen. In der Regel stehen allerdings Daten nur aus Querschnittsuntersuchungen zur Verfügung.

Eine interessante Studie haben Scribner et al. (1999) vorgelegt, die am Beispiel der Einführung professioneller Lerngemeinschaften eine Evaluation dieses Prozesses beschrieben haben. Nach Park und Lee (2015) ist es allerdings im internationalen Vergleich schwierig, auf der Schulebene Bedingungen zu identifizieren, die das Entstehen professioneller Lerngemeinschaften unterstützen. In einer anderen Studie wird die Wirksamkeit eines Präventions-Curriculums zum Abbau aggressiven Verhaltens bei Schülern und Schülerinnen evaluiert (vgl. Schick/Cierpka 2013).

International ist ein häufiges Thema, welchen Beitrag Lehrkräfte am Erfolg von Schülern und Schülerinnen haben (vgl. Reynolds/Mujis/Treharne 2003). Das ist bereits in vielen Studien untersucht worden (vgl. auch Hallinger/Heck/Murphy 2014). Hier war vor allem von Interesse, welchen Einfluss die Evaluation der Lehrkräfte auf den Erfolg der Schülerinnen und Schüler hat. Die Ergebnisse lassen Vorsicht angeraten sein. Das hängt sicher auch damit zusammen, dass die Modi der Evaluation kein einheitliches Bild abgeben (vgl. Murphy/Hallinger/Heck 2013). Diese Fragestellung ist in Deutschland bisher weniger untersucht worden. Ihr käme bei Fragen zur Personalentwicklung bzw. Weiterbildung Bedeutung zu.

Innerhalb der einzelnen Schule können somit bei der Evaluation die Schule insgesamt (Schulqualität, Schulinspektion), der Unterricht (Unterrichtsqualität) bzw. die Tätigkeit der Lehrpersonen im Fokus stehen. Falls Mängel entdeckt werden, sollen diese im Anschluss behoben werden. Eine Übersicht zum Forschungsstand bei internen Evaluationen findet sich bei Berkemeyer, Müller und van Holt (2016).

8. Zusammenfassung

Evaluation ist kein klar definierter Begriff. Es gibt vielmehr unterschiedliche Varianten und auch wissenschaftliche Traditionen, aus denen heraus ein Verständnis von Evaluation entwickelt worden ist. Deshalb gibt es bei der Praxis der Evaluation erhebliche Differenzen. Trotz dieser Einschränkung kommt Evaluationen innerhalb des Bildungssystems und bei der einzelnen Schule zunehmende Bedeutung zu. Weil inzwischen von der Kontrolle des Outputs bei Maßnahmen im Bildungssystem Hinweise auf mögliche Defizite erwartet werden, muss konsequenterweise evaluiert werden. Es wird allerdings darauf ankommen, das Verständnis in diesem Bereich weiterzuentwickeln. Die Selbstverständlichkeit, mit der der Begriff verwendet wird, trügt.

Literatur und Internetquellen

Abs, H.J./Klieme, E. (2005): Standards für die schulbezogene Evaluation. In: Zeitschrift für Erziehungswissenschaft, 4. Beiheft, S. 45-62.

Altrichter, H./Kemethofer, D. (2015): Neue Ansätze der Steuerung des Schulsystems und die Einstellung von Schulleitungen. In: Bildung und Erziehung 68, H. 3, S. 291-310.

Altrichter, H./Maag Merki, K. (Hrsg.) ([2]2016): Handbuch Neue Steuerung im Schulsystem. Wiesbaden: VS.

Altrichter, H./Moosbrugger, R./Zuber, J. ([2]2016): Schul- und Unterrichtsentwicklung durch Datenrückmeldung. In: Altrichter, H./Maag Merki, K. (Hrsg.): Handbuch Neue Steuerung im Schulsystem. Wiesbaden: VS.

Autorengruppe Bildungsberichterstattung (Hrsg.) (2014): Bildung in Deutschland 2014. Ein indikatorengestützter Bericht mit einer Analyse zur Bildung von Menschen mit Behinderungen. Bielefeld: Bertelsmann.

Berkemeyer, N. (2010): Die Steuerung des Schulsystems. Wiesbaden: VS.

Berkemeyer, N./Müller, S./van Holt, N. (2016): Schulinterne Evaluation – nur ein Instrument zur Selbststeuerung von Schulen? In: Altrichter, H./Maag Merki, K. (Hrsg.) ([2]2016): Handbuch Neue Steuerung im Schulsystem. Wiesbaden: VS, S. 209-234.

Bos, W. et al. (Hrsg.) (2003): Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich. Münster u.a.: Waxmann.

Chapman, C./Sammons, P. (2013): School Self-Evaluation for School Improvement. What Works and Why. CfBT Education Trust. Reading: CfBT.

Dedering, K. (2012): Schulinspektion als wirksamer Weg der Systemsteuerung? In: Zeitschrift für Pädagogik 58, H. 1, S. 69-88.

DeGEval – Gesellschaft für Evaluation ([4]2004): Standards für Evaluation. Köln: DeGEval – Gesellschaft für Evaluation e.V.

Deutsches PISA-Konsortium (Hrsg.) (2001): PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Opladen: Leske + Budrich.

Döbert, H./Dedering, K. (2008): Externe Evaluation von Schulen in vergleichender Perspektive – eine Einführung. In: Dies. (Hrsg.): Externe Evaluation von Schulen. Münster u.a.: Waxmann, S. 11-22.

Dubs, R. (2011): Die teilautonome Schule. Ein Beitrag zur Ausgestaltung aus politischer, rechtlicher und schulischer Sicht. Berlin: Sigma.

Fend, H. (2011): Die Wirksamkeit der Neuen Steuerung – theoretische und methodische Probleme ihrer Evaluation. In: Zeitschrift für Bildungsforschung 1, H. 1, S. 5-24.

French, W.L./Bell, C.H. Jr. ([3]1990): Organisationsentwicklung. Bern: Haupt.

Gronlund, N.E. (1968): Readings in Measurement and Evaluation. New York: The Macmillan Company.

Guba, E.G./Lincoln, Y.S. (1989): Fourth Generation Evaluation. Thousand Oaks, CA: Sage.

Hallinger, P./Heck, R.H./Murphy, J. (2014): Teacher Evaluation and School Improvement: Analysis of the Evidence. In: Educational Assessment, Evaluation and Accountability 25, H. 1, S. 5-28.

Holland, D. (2009): Interne Evaluation der Georg-Forster-Gesamtschule Wörrstadt mit Hilfe von SEIS. In: Appel, S./Ludwig, H./Rother, U./Rutz, G. (Hrsg.): Jahrbuch Ganztagsschule 2009. Leben, Leisten, Lernen. Bad Schwalbach: Wochenschau Verlag, S. 162-175.

Holtappels, H.-G. (2008): Externe Evaluation durch Schulinspektion und zentrale Prüfungen – Eine Einführung. In: Böttcher, W./Bos, W./Döbert, H./Holtappels, H.-G. (Hrsg.): Bildungsmonitoring und Bildungscontrolling in nationaler und internationaler Perspektive. Münster u.a.: Waxmann, S. 219-222.

Husfeldt, V. (2011): Wirkungen und Wirksamkeit der externen Schulevaluation. In: Zeitschrift für Erziehungswissenschaft 14, H. 2, S. 259-282.

Joint Committee on Standards for Educational Evaluation ([2]1994): The Program Evaluation Standards: How to Assess Evaluations of Educational Programs. Thousand Oaks, CA: Sage.

Ko, J./Hallinger, P./Walker, A. (2015): Exploring Whole School versus Subject Department Improvement in Hong Kong Secondary Schools. In: School Effectiveness and School Improvement 26, H. 2, S. 215-239.

Köller, O. ([2]2015): Evaluation pädagogisch-psychologischer Maßnahmen. In: Wild, E./Möller, J. (Hrsg.): Pädagogische Psychologie. Berlin/Heidelberg: Springer, S. 329-341.

Kuper, H. (2005): Evaluation im Bildungssystem. Stuttgart: Kohlhammer.

Kyriakides, L./Creemers, B.P.M. (2008): Using a Multidimensional Approach to Measure the Impact of Classroom-Level-Factors upon Student Achievement: A Study Testing the Validity of the Dynamic Model. In: School Effectiveness and School Improvement 19, H. 2, S. 183-205.

Maag Merki, K. ([2]2016): Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In: Altrichter, H./Maag Merki, K. (Hrsg.): Handbuch Neue Steuerung im Schulsystem. Wiesbaden: VS, S. 145-170.

Maritz, B. et al. (2006): Bewertungsbuch für Schulen. Eine Anleitung zur Bewertung der Schulqualität auf der Grundlage des Modells der European Foundation for Quality Management (EFQM). Bern: hep.

Mayring, P. ([9]2010): Qualitative Inhaltsanalyse. In: Flick, U./von Kardorff, E./Steinke, I. (Hrsg.) (2010): Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, S. 468-475.

Merkens, H. (2010): Unterricht. Eine Einführung. Wiesbaden: VS.

Mortimore, P. (1998): The Road to Improvement. Reflections on School Effectiveness. Lisse: Swets & Zeitlinger.

Murphy, J./Hallinger, P./Heck, R.H. (2013): Leading via Teacher Evaluation: The Case of the Missing Clothes? In: Educational Researcher 42, H. 6, S. 349-354.

Park, J.-H./Lee, J.Y. (2015): School-Level Determinants of Teacher Collegial Interaction: Evidence from Lower Secondary Schools in England, Finland, South Korea, and the USA. In: Teaching and Teacher Education 50, H. 1, S. 24-35.

Pietsch, M. (2010): Evaluation von Unterrichtsstandards. In: Zeitschrift für Erziehungswissenschaft 13, H. 1, S. 121-148.

Reynolds, D./Mujis, D./Treharne, D. (2003): Teacher Evaluation and Teacher Effectiveness in the United Kingdom. In: Journal of Personnel Evaluation in Education 17, H. 1, S. 83-100.

Reynolds, D./Sammons, P./Fraine, B.D./Damme, J. van/Townsend, T./Teddlie, C./Stringfield, S. (2014): Educational Effectiveness Research (EER): A State-of-the-Art Review. In: School Effectiveness and School Improvement 25, H. 2, S. 197-230.

Sammons, P. (1999): School Effectiveness. Coming of Age in the Twenty-First Century. Lisse: Swets & Zeitlinger.

Scheerens, J./Creemers, B.P.M. (1989): Conceptualizing School Effectiveness. In: Dies. (Hrsg.): Developments in School Effectiveness Research. Special Issue of the International Journal of Educational Research 13, S. 691-706.

Schick, A./Cierpka, M. (2013): International Evaluation Studies of Second Step, a Primary Prevention Programme: A Review. In: Emotional and Behavioral Difficulties 18, H. 3, S. 241-247.

Schulinspektion Berlin (2016). URL: https://www.berlin.de/.../schulinspektion/handbuch_schulinspektion.pdf; Zugriffsdatum: 18.03.2016.

Scribner, J.P./Cockrell, K.S./Cockrell, D.H./Valentine, J.V. (1999): Creating Professional Communities in Schools through Organizational Learning: An Evaluation of a School Improvement Process. In: Educational Administrative Quarterly 25, H. 1, S. 130-160.

Scriven, M. (1967): The Methodology of Evaluation. In: Tyler, R./Gagne, R./Scriven, M. (Hrsg.): Perspectives of Curriculum Evaluation. Chicago, IL: Rand McNally & Co., S. 39-83.

Senatsverwaltung für Bildung, Jugend und Wissenschaft (Hrsg.) (2013): Handlungsrahmen Schulqualität in Berlin. Qualitätsbereiche und Qualitätsmerkmale. URL: www.berlin.de/sen/bildung; Zugriffsdatum: 18.03.2016.

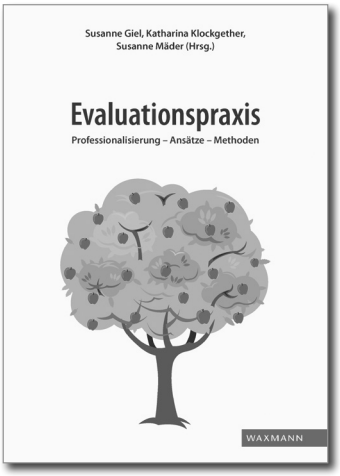
Stockmann, R. (2010): Rolle der Evaluation in der Gesellschaft. In: Stockmann, R./Meyer, W. (Hrsg.): Evaluation. Eine Einführung. Opladen: Budrich, S. 25-53.

van Ackeren, I./Klemm, K./Kühn, S.M. (2015): Entstehung, Struktur und Steuerung des deutschen Schulsystems. Eine Einführung. Wiesbaden: Springer VS.

von Kardorff, E. ([9]2010): Qualitative Evaluationsforschung. In: Flick, U./von Kardorff, E./Steinke, I. (Hrsg.) (2010): Qualitative Forschung. Ein Handbuch. Reinbek: Rowohlt, S. 238-250.

Hans Merkens, Prof. Dr. Dr. h.c., geb. 1937, emeritierter Universitätsprofessor für Erziehungswissenschaft (Arbeitsbereich Empirische Erziehungswissenschaft) an der Freien Universität Berlin.
Anschrift: FU Berlin, Fachbereich Erziehungswissenschaft und Psychologie, Fabeckstr. 69, 14195 Berlin
E-Mail: merken@zedat.fu-berlin.de

UNSERE BUCHEMPFEHLUNG



Susanne Giel,
Katharina Klockgether,
Susanne Mäder
(Hrsg.)

Evaluationspraxis

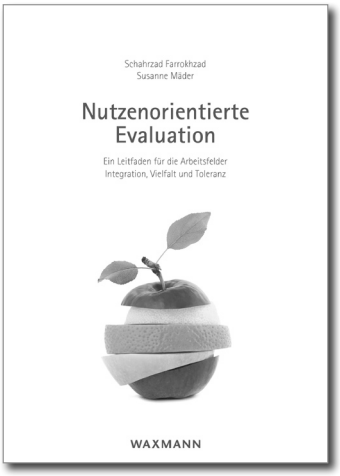
Professionalisierung –
Ansätze – Methoden

2015, 300 Seiten, br., 34,90 €,
ISBN 978-3-8309-3345-8
E-Book: 30,99 €,
ISBN 978-3-8309-8345-3

In der Evaluationspraxis stellen sich vielfach Herausforderungen, zu denen situativ passende und kreative Lösungen gefunden werden müssen. Dieser Band präsentiert Aufsätze aus drei inhaltlichen Bereichen: Professionalisierungsinstrumente, Evaluationsansätze und Evaluationsmethoden. Im ersten Teil werden Quellen für eine Verbesserung der Steuerung vorgestellt. Der Schwerpunkt im zweiten Teil liegt auf nutzungsorientierten Evaluationsansätzen, insbesondere für experimentierende und unausgereifte Programme. Neben Datenerhebungsmethoden werden im dritten Teil auch Methoden für die Gegenstandsklärung sowie die Ergebnisvermittlung inklusive des Bewertungsvorgangs präsentiert.

Das Buch bietet Erfahrenen, Novizinnen und Novizen, Praktikerinnen und Praktikern sowie Studierenden einen anregenden und anschaulichen Einblick in die Evaluationspraxis.

UNSERE BUCHEMPFEHLUNG



Schahrzad Farrokhzad,
Susanne Mäder

Nutzenorientierte Evaluation

Ein Leitfaden für die Arbeitsfelder Integration, Vielfalt und Toleranz

2014, 144 Seiten, br., 24,90 €,
ISBN 978-3-8309-3065-5
E-Book: 21,99 €,
ISBN 978-3-8309-8065-0

Eine Evaluation durchgeführter Programme und Projekte wird von den Auftraggebern häufig zur Pflicht gemacht. Wenn aber die Mittel für diese Projekte reduziert werden, muss die Evaluation besonders ressourcenschonend und mit einem hohen Nutzen für alle Beteiligten durchgeführt werden. In diesem Buch werden nutzenorientierte Strategien und Methoden für die Durchführung von Evaluationen vorgestellt, insbesondere für Evaluationen in den Themenfeldern Integration, Vielfalt und Toleranz. Die evaluationstheoretischen und -praktischen Ausführungen werden am Beispiel konkreter Evaluationsinstrumente veranschaulicht. Das Buch richtet sich an Projektverantwortliche, die Evaluationen selbst durchführen oder externe Evaluationen beauftragen wollen. Auch für Studierende und Evaluierende mit Interesse an einem nutzenorientierten Evaluationsansatz hält der Band praxisnahe Informationen bereit.

ZEITSCHRIFT FÜR EVALUATION

mehr zum Inhalt der ZfEv und zu den Bestellmöglichkeiten ▶

Jahresabo
print: **32,–€**
online: **29,–€**

Die Zeitschrift für Evaluation (ZfEv) publiziert sowohl wissenschaftliche Beiträge als auch praxisorientierte Erfahrungsberichte aus verschiedenen Themenfeldern zur Evaluation.

Sie bietet eine Plattform für

- den fachlichen Dialog zwischen Wissenschaft und Praxis
- die interdisziplinäre Bündelung sektoralen Fachwissens
- den internationalen Austausch im deutschsprachigen Raum
- die Vermittlung und Diskussion neuer Entwicklungen
- die Verbreitung von Standards in der Evaluation
- Ausschreibungen, Literatur, Veranstaltungen etc.

www.waxmann.com