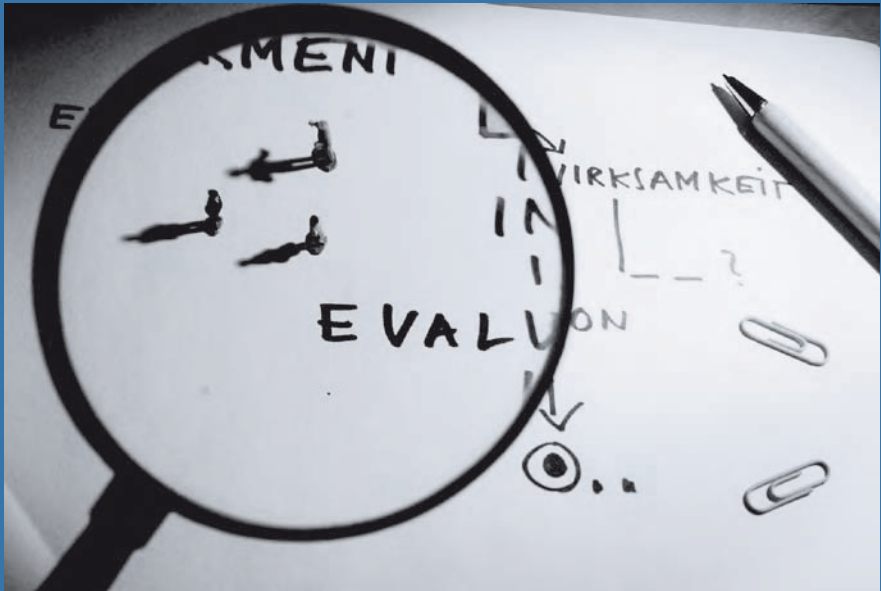


Jan Hense
Stefan Rädiker
Wolfgang Böttcher
Thomas Widmer
(Hrsg.)

FORSCHUNG ÜBER EVALUATION

BEDINGUNGEN, PROZESSE UND WIRKUNGEN



WAXMANN

Forschung über Evaluation

Jan Hense, Stefan Rädiker,
Wolfgang Böttcher, Thomas Widmer
(Hrsg.)

Forschung über Evaluation

Bedingungen, Prozesse, Wirkungen



Waxmann 2013
Münster / New York / München / Berlin

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar

ISBN 978-3-8309-2829-4

© Waxmann Verlag GmbH, 2013

Postfach 8603, 48046 Münster

www.waxmann.com

info@waxmann.com

Umschlaggestaltung: Anne Breitenbach, Tübingen

Umschlagabbildung: Conny Türk, Willich

Druck: Hubert & Co., Göttingen

Gedruckt auf alterungsbeständigem Papier, säurefrei gemäß ISO 9706



Mix

Produktgruppe aus vorbildlich bewirtschafteten
Wäldern und anderen kontrollierten Herkünften

www.fsc.org Zert.-Nr. SGS-COC-005773
© 1996 Forest Stewardship Council

Printed in Germany

Alle Rechte vorbehalten. Nachdruck, auch auszugsweise, verboten.
Kein Teil dieses Werkes darf ohne schriftliche Genehmigung des
Verlages in irgendeiner Form reproduziert oder unter Verwendung
elektronischer Systeme verarbeitet, vervielfältigt oder verbreitet werden.

Inhalt

Jan Hense, Stefan Rädiker, Thomas Widmer, Wolfgang Böttcher

Notwendigkeit und Potenzial einer Forschung über Evaluation..... 7

Teil 1 Ausgangsbedingungen von Evaluation

Ursula Koch

Datenauswertungskompetenzen von Lehrkräften 21

Markus Seyfried, Philipp Pohlenz

Professionalisierung von Qualitätsentwicklung und Evaluation
der Hochschullehre – zwischen Kontrolle und Selbstreflexion 43

Teil 2 Evaluationsprozesse

Stefan Rädiker

Evaluation von Bildungsprozessen in Weiterbildungsorganisationen: Eine
Studie zu Status quo, Umsetzungsproblemen und Kompetenzanforderungen.... 65

Matthias Müller

Evaluation in der Sozialen Arbeit: Eine Studie zum Verhältnis
von Autonomie und Heteronomie..... 87

Teil 3 Wirkungen und Folgen von Evaluation

Holger Gärtner

Wirksamkeit von Schülerfeedback als Instrument der Selbstevaluation
von Unterricht..... 107

André Nowakowski

Evaluationsrückmeldungen als Untersuchungsgegenstand: Studien zum
Zusammenhang mit Lehrerorientierungen 125

Mary Sandermann

Wie bedeutsam sind soziale Kompetenzen für Evaluationsnutzungen? 143

Kathrin Frey, Thomas Widmer

Vom Eigenen oder Fremden lernen bei der evidenzbasierten
Politikgestaltung? 165

Teil 4 Übergreifende Fragestellungen

Wolfgang Meyer, Jörg Rech

Determinanten der Evaluationskultur –
Das Beispiel Entwicklungszusammenarbeit 189

Wolfgang Böttcher, Nina Hoglebe, Sabrina Schulz

Evidenzgewinn und Generalisierbarkeit durch systematische Auswertungen
von Evaluationen? Das Beispiel der Erziehungswissenschaft 211

Wolfgang Böttcher, Jan Hense, Miriam Keune

Schulinspektion als eine Form externer Evaluation –
ein Forschungsüberblick 231

Jan Hense, Thomas Widmer

Ein Überblick zum internationalen Stand der Forschung über Evaluation 251

Verzeichnis der Autorinnen und Autoren 279

Jan Hense, Stefan Rädiker, Thomas Widmer, Wolfgang Böttcher
**Notwendigkeit und Potenzial einer Forschung
über Evaluation**

Die Evaluation hat sich in den vergangenen Jahren in unterschiedlichen gesellschaftlichen Handlungsfeldern als wissenschaftsgestützte Methode zur Untersuchung von Prozessen, Wirkung und Nutzen sozialer Praxis etabliert (z. B. Widmer, Beywl & Fabian, 2009). Gewöhnlich wird von Evaluationen erwartet, dass sie einen Beitrag zur Optimierung von Steuerungsprozessen im jeweiligen Handlungsfeld leisten, indem sie Verbesserungsmöglichkeiten aufzeigen, über Entscheidungsprozesse informieren oder zur Legitimation und Erfolgskontrolle beitragen.

Im Laufe der immer noch andauernden Expansionsphase der Evaluation wurden wiederholt Sinn- und Nutzenfragen gestellt (z. B. Brüsemeister & Eubel, 2008; Matthies & Simon, 2008): Lohnt sich der Aufwand, der für einzelne Evaluationsstudien und für kontinuierlich arbeitende Evaluationssysteme betrieben wird? Tragen Evaluationsverfahren tatsächlich zur verbesserten Steuerung sozialer Handlungssysteme bei und wenn ja, unter welchen Bedingungen? Wie lässt sich Evaluation so gestalten, dass sie die an sie gerichteten Erwartungen erfüllen kann? Welche erwünschten und nicht erwünschten Wirkungen hat Evaluation als dauerhaft institutionalisierte oder punktuelle Maßnahme?

Diese und ähnliche Fragen sind in den vergangenen Jahren auch im deutschsprachigen Raum zunehmend Gegenstand empirischer Forschungsarbeiten unterschiedlicher Fachdisziplinen geworden. Stellvertretend nennen lassen sich hier Studien zur Ergebnisrezeption im Kontext von Schulleistungsstudien (z. B. Schrader & Helmke, 2004; Schneewind, 2007), zu Umsetzung und Folgen evaluativer Verfahren wie Akkreditierung und Zertifizierung (z. B. Kaufmann, 2012; Serrano-Velarde, 2008) oder zum Nutzen von Programmevaluation im Präventions- und Gesundheitsbereich (z. B. Keller-Ebert, 2006).

In diesem einleitenden Kapitel wollen wir zunächst argumentieren, warum wir es für notwendig halten, entsprechende Forschungsarbeiten in den Kontext einer übergeordneten Forschungsperspektive einzubetten. Anschließend möchten wir den Begriff der „Forschung über Evaluation“ konkretisieren und vier inhaltliche Teilbereiche möglicher Studien unterscheiden. Diese vier Schwerpunktbereiche der Forschung über Evaluation dienen dann auch zur Gliederung der Einzelbeiträge des Sammelbands, die abschließend vorgestellt werden.

1 Wozu Forschung über Evaluation?

Trotz vielfältiger Forschungsaktivitäten hat sich im Gegensatz zur nordamerikanischen *research on evaluation* (siehe Hense & Widmer in diesem Band) hierzulande noch kein gemeinsamer transdisziplinärer wissenschaftlicher Diskurs herausgebildet. Vielmehr verlaufen die entsprechenden Diskurse, trotz inhaltlich eng verwandter Fragestellungen, derzeit weitgehend isoliert voneinander in verschiedenen Fachdisziplinen wie etwa Erziehungswissenschaft, Politikwissenschaft oder Soziologie respektive in den verschiedenen Handlungsfeldern wie Bildung, Entwicklungszusammenarbeit oder Gesundheit. Teils gibt es sogar innerhalb der Disziplinen eine weitere Zersplitterung des Diskurses in verschiedene Anwendungsfelder. So werden etwa in der Pädagogik die Diskurse zur Evaluation in Schule, Hochschule oder Weiterbildung zu großen Teilen unabhängig voneinander geführt. Aus verschiedenen Gründen halten wir es aber für lohnend, den gegenwärtigen Zustand der Dissoziation zu überwinden.

- In den verschiedenen Disziplinen und Anwendungsfeldern werden teils sehr ähnliche Probleme und Fragestellungen bearbeitet. Trotz nicht zu bestreitender disziplinärer und feldspezifischer Unterschiedlichkeiten, lassen sich die untersuchten Evaluationsverfahren aber zumindest in Teilbereichen (z. B. Ziele von Evaluationen, methodische Herangehensweisen, institutionelle Einbettung, Nutzungsproblematik) sinnvoll zueinander in Beziehung setzen. Daher steht zu erwarten, dass durch eine transdisziplinär und bereichsübergreifend orientierte Perspektive, die den Austausch über theoretische und methodische Herangehensweisen sowie den Vergleich von Befunden impliziert, ein weitergehender Erkenntnisgewinn und ein tieferes Verständnis der jeweils fokussierten lokalen Problem- und Fragestellungen möglich wird.
- Akzeptiert man die Position, dass Evaluation sich trotz aller Unterschiedlichkeit in ihren konkreten Realisierungen auf einen gemeinsamen ideengeschichtlichen, theoretischen und formal-logischen Kern zurückführen lässt (z. B. Mark, Henry & Julnes, 1999; Scriven, 1980; Shadish, Cook & Leviton, 1991), wird eine transdisziplinär bewusste Forschung sich zunehmend auf einen gemeinsamen empirischen Korpus und ein gemeinsames Theoriegebäude stützen können. Dies sollte dazu führen, dass das „Rad“ nicht in jedem Anwendungsbereich der Evaluation jeweils „neu erfunden“ werden muss, sondern zunehmend Erkenntnisse bereit stehen, die zumindest tentativ auf neue Anwendungsfelder übertragbar sind. Eine wichtige Bedeutung hat dabei auch der international theoretische und empirische Forschungsstand zur Evaluation, der in vielen Arbeiten mit fachspezifischer Perspektive aber kaum zur

Kenntnis genommen wird, obwohl umfangreiche theoretische Überlegungen und empirische Erfahrungen seit den 1970er Jahren verfügbar wären (siehe Hense & Widmer in diesem Band).

- Relevant sind die Ergebnisse eines transdisziplinären Diskurses zur Forschung über Evaluation schließlich nicht alleine aus akademischem Interesse, sondern auch im Hinblick auf eine zunehmende Professionalisierung von Evaluation. Denn während Evaluation oft mit dem Ruf nach einer evidenz-basierten Praxis begründet wird, stellt sich die berechnigte Frage, mit welcher systematischen Evidenz sich die Praxis der Evaluation begründen lässt. Zwar liegen nicht wenige präskriptive Modelle und Ansätze der Evaluation vor, die mit unterschiedlicher normativer Fundierung Vorgaben dazu machen, wie Evaluationen konkret zu planen und umzusetzen sind (Stufflebeam, 2001a). Ob die vielen impliziten Annahmen darüber, wie eine „gute“ Evaluation „richtig“ zu gestalten ist, sich aber jenseits von Erfahrungswissen und individueller Expertise begründen lassen, bleibt dabei offen. Ein theoretisch und empirisch fundierter Erkenntnisstand, aus dem heraus sich professionelles Handeln und die Ausbildung von Evaluierenden begründen lässt, ist aber ein unverzichtbares Element von Professionalisierung (Mieg, 2010).

2 Was ist Forschung über Evaluation?

Forschung über Evaluation verstehen wir, wie oben eingeführt, als Forschung, die Evaluation selbst zum Untersuchungsgegenstand hat und mit dem Ziel betrieben wird, die theoretische und empirische Befundlage zur Evaluation zu verbessern. Obwohl diese Definition explizit auch die Theorieentwicklung umfasst, konzentrieren wir uns in diesem Band vorwiegend auf empirisch orientierte Arbeiten.

Um begriffliche Unklarheiten zu vermeiden, wollen wir „Forschung über Evaluation“ von drei anderen Begriffen abgrenzen, mit denen Verwechslungsgefahr bestehen könnte:

- In Analogie zu etablierten Komposita wie „Genforschung“, „Meinungsforschung“ oder „Lehr-Lern-Forschung“ könnte *Evaluationsforschung* als Synonym von „Forschung über Evaluation“ gelesen werden. Obwohl eine solche Verwendung inhaltlich plausibel und nahe liegend wäre, erscheint es uns angesichts der teilweise etablierten abweichenden Bedeutung von „Evaluationsforschung“ nicht sinnvoll, hier eine Umprägung anzustreben. In einem Teil der Literatur wird der Begriff der Evaluationsforschung als Synonym

oder Alternative zu „Evaluation“ verwendet¹, oft auch mit der davon abgrenzenden Betonung eines wissenschaftlichen Anspruchs gegenüber alltäglichen, nicht-wissenschaftlichen Bewertungsprozessen (z. B. Wottawa & Thierau, 1998; Stockmann, 2006). Daher bevorzugen wir den etwas sperrigeren, inhaltlich aber eindeutigen Begriff der „Forschung über Evaluation“.

Abzugrenzen ist der Begriff inhaltlich darüber hinaus von zwei weiteren Untersuchungsarten dritter Ordnung, die Evaluationen zum Gegenstand haben können, aber andere Ziele verfolgen als die Forschung über Evaluation:

- Das erste dieser Konzepte, *Meta-Evaluation*, untersucht ebenfalls Evaluationen, hat dabei allerdings das Ziel der Bewertung ihrer Qualität (Stufflebeam, 2001b), etwa durch Überprüfung, inwiefern sie gängige Standards der Evaluation (z. B. Gesellschaft für Evaluation [DeGEval], 2008; Widmer, Landert & Bachmann 2000; Yarbrough, Shulha, Hopson & Caruthers, 2011) beachten. Übergänge zur Forschung über Evaluation bestehen hier dann, wenn Meta-Evaluationen nicht nur praxisorientierte Einzelfallstudien sind, sondern wenn Evaluationen, etwa in einem bestimmten Politikfeld oder Administrationsbereich, vergleichend im Hinblick auf Qualitätsaspekte untersucht werden (z. B. Widmer, 1996).
- Ebenfalls Gegenstand von Untersuchungen können Evaluationen im Kontext von *Meta-Analysen* oder *Evaluationssynthesen* sein. Meta-Analysen verdichten die Befundlage zu einer bestimmten Maßnahme (z. B. einer bestimmten Lernmethode, einer Therapieform oder einem Präventionsprogramm), um den Erkenntnisstand zur Wirksamkeit dieser Maßnahme z. B. in Form eines einheitlichen Effektstärkemaßes zusammenzufassen (siehe den Beitrag von Böttcher, Hogrebe und Schulz in diesem Band). Evaluationssynthesen führen die Befunde von vorbestehenden Evaluationsstudien zu einer oder mehreren Maßnahmen oder zu einer Problematik mit einem qualitativen Zugang zusammen (General Accounting Office [GAO], 1992; Balthasar, 2000). Das Ziel der beiden Zugänge ist also der Erkenntnisgewinn auf der Ebene des Evaluations- oder Untersuchungsgegenstands und nicht, wie bei der Forschung über Evaluation, auf der Ebene der Evaluation. Auch hier sind natürlich Übergänge denkbar. So beinhalten viele Meta-Analysen eine meta-evaluative Komponente, indem sie eine Bewertung der methodischen Qualität

1 Diese Begriffsverwendung wird oft mit einem Verweis auf Suchman (1967) begründet. Da er allerdings von „evaluative research“ sprach, wäre eigentlich „evaluative Forschung“ die korrekte Übersetzung.

der jeweiligen Studie einbeziehen (Ahn, Ames & Myers, 2012). Weitere Übergänge zur Forschung über Evaluation bestehen etwa darin, dass Meta-Analysen eine wichtige Informationsquelle über die Design- und Methodenwahl in Evaluationsstudien darstellen.

In Anlehnung an die verbreitete Logik allgemeiner Produktionsmodelle lassen sich Studien zur Forschung über Evaluation danach strukturieren, welchen Bereich des Evaluationsgeschehens sie fokussieren (Mark, 2008; Hense & Widmer in diesem Band):

1. Studien zu *Ausgangsbedingungen* (Inputs) von Evaluation fokussieren Aspekte wie die Qualifikation und Qualifizierung von Evaluierenden, den sozialen Kontext von Evaluation und weitere Voraussetzungen und Rahmenbedingungen von Evaluation.
2. Die Umsetzung bzw. der *Prozess* der Evaluation ist ein weiterer möglicher Untersuchungsgegenstand. Entsprechende Studien untersuchen etwa die Interaktion mit Beteiligten und Betroffenen einer Evaluation, Schwierigkeiten bei der praktischen Umsetzung von Evaluationen, ethische Probleme und Konflikte in der Evaluation oder Aspekte der Methodenwahl.
3. Ein dritter Gegenstandsbereich sind *Wirkungen und Folgen* (Outcomes) von Evaluation. Dazu gehören etwa Fragen nach der Evaluationsnutzung, nach allgemeineren Einflüssen von Evaluation auf ihren Gegenstand während und nach der Evaluation sowie intendierte und nicht intendierte Wirkungen von Evaluation.
4. Schließlich lassen sich noch *übergeordnete Fragestellungen* nennen, die sich nicht trennscharf einem der drei obigen Bereiche zuordnen lassen. Beispiele sind Untersuchungen, die verschiedene präskriptive Evaluationsmodelle oder Evaluationen als Ganzes zum Gegenstand haben.

3 Zu diesem Sammelband

Vor dem Hintergrund der oben vorgenommenen begrifflichen Klärung zur Forschung über Evaluation will dieser Sammelband eine transdisziplinäre Plattform für die Präsentation und Diskussion von Forschungsprojekten bereitstellen, die Evaluationen und Evaluationssysteme als Ganzes oder in Teilaspekten zum Gegenstand haben. Die Beiträge beziehen sich jeweils auf eigene empirische Arbeiten der jeweiligen Autorinnen und Autoren oder fassen den Forschungsstand in einem bestimmten Anwendungsbereich zusammen.

Da wir den Sammelband als Beitrag zu einem ersten Schritt auf dem Weg zu einer Etablierung eines transdisziplinären Diskurses zur Forschung über Evaluation verstehen, verfolgen wir nicht den Anspruch, einen wie oben skizzierten gemeinsamen Bezugsrahmen anzulegen respektive den Beitragenden einen solchen aufzuerlegen. Ziel ist zunächst, die Vielfältigkeit bisher vorliegender Arbeiten aus verschiedenen Disziplinen, Fachtraditionen und Anwendungsfeldern sichtbar zu machen, um sie und ähnliche weitere Arbeiten in einem zweiten Schritt für eine Diskussion von Gemeinsamkeiten im oben angedeuteten Sinne anschlussfähig zu machen.

Der Sammelband geht zurück auf eine Fachtagung gleichen Namens, die der Arbeitskreis „Aus- und Weiterbildung in der Evaluation“ in der Gesellschaft für Evaluation (DeGEval) im April 2012 in Berlin veranstaltet hat.² Er umfasst einen Großteil der dort präsentierten Forschungsarbeiten in ausgearbeiteter Form und wurde um einige weitere Arbeiten ergänzt, um das Gesamtbild inhaltlich abzurunden.

Die Einzelbeiträge sind analog zur oben dargestellten Strukturierung nach den fokussierten Bereichen des Evaluationsgeschehens in vier Teile gegliedert. Dabei soll die Einteilung jeweils vor allem die Schwerpunktsetzung der jeweiligen Studie betonen, ohne auszuschließen, dass auch Aspekte berührt werden, die in einen der anderen Bereiche fallen.

Teil 1 beinhaltet Beiträge, die förderliche und hinderliche Ausgangsbedingungen der Evaluation fokussieren. Untersucht werden also jeweils verschiedene Aspekte des Evaluationsgeschehens, die Einfluss auf Umsetzung und Wirkungen von Evaluationen haben.

- *Ursula Koch* beschäftigt sich in ihrem Beitrag mit der Datenauswertungskompetenz, die Lehrkräfte benötigen, um Evaluationsdaten sinnvoll nutzen zu können, etwa zur Unterrichtsverbesserung. Hintergrund ihrer Arbeit sind die sogenannten Vergleichsarbeiten, die seit einigen Jahren an deutschen Schulen geschrieben werden, um Informationen zur Leistungsfähigkeit einzelner Schulen und ihrer Schüler/innen zu erhalten und den Schulen Daten für die (Selbst-)Evaluation zu liefern. Die Autorin berichtet in ihrem Beitrag über die Ergebnisse eines Quasi-Experiments, bei dem die Lehrer/innen im Rahmen einer Fortbildung einmal mit und einmal ohne Softwareunterstützung die Ergebnisse analysiert haben. Zu mehreren Zeitpunkten während und

2 Die einzelnen Präsentationen der Fachtagung sind im Internetangebot der DeGEval auf den Seiten des Arbeitskreises unter <http://www.degeval.de/arbeitskreise/aus-und-weiterbildung-in-der-evaluation/bisherige-aktivitaeten> verfügbar.

im direkten Anschluss an die Fortbildung wurde die Datenkompetenz der Lehrenden erhoben.

- *Markus Seyfried* und *Philipp Pohlenz* werfen für den Kontext Hochschule die Frage auf, wie und woran sich strukturelle Professionalisierung im Bereich der Evaluation von Lehre und Studium erkennen lässt. Im Anschluss an eine umfassende Klärung dieser Frage auf konzeptioneller Ebene werden erste Ergebnisse einer empirischen Erhebung vorgestellt. Unter Rückgriff auf die Daten der Hochschulrektorenkonferenz von 105 deutschen Universitäten haben die Autoren die organisatorische Anbindung von Qualitätsentwicklungs- bzw. Evaluationseinheiten untersucht, wobei zwischen der „klassischen Verwaltung“, den „Stabsstellen“ und „anderen Organisationseinheiten“ unterschieden wurde.

Teil 2 des Sammelbands präsentiert Studien, die sich auf den Evaluationsprozess, also auf Aspekte der Planung und Umsetzung von Evaluationen konzentrieren.

- *Stefan Rädiker* widmet sich in seinem Beitrag der Evaluation von Bildungsprozessen in Weiterbildungsorganisationen. Der Autor stellt die Ergebnisse einer qualitativen Studie unter Organisationen vor, welche das Qualitätsmanagementmodell der „Lernerorientierten Qualitätstestierung in der Weiterbildung (LQW)“ anwenden und in diesem Kontext verpflichtet sind, ihre Bildungsprozesse zu evaluieren. Anhand von Dokumentenanalysen von Selbstberichten sowie Leitfadeninterviews mit Evaluationsverantwortlichen wurden sowohl Status quo der Evaluationspraxis als auch die im Evaluationsprozess auftretenden Probleme untersucht. Die empirischen Ergebnisse flossen in ein Anforderungsprofil für Evaluierende in Weiterbildungsorganisationen ein, das am Ende des Beitrags kurz vorgestellt wird.
- *Matthias Müller* hat sich mit Evaluation in der Sozialen Arbeit auseinandergesetzt. Mithilfe eines wissenssoziologischen Zugangs rekonstruiert der Autor die Handlungsorientierungen und Deutungsmuster von Fach- und Führungskräften eines Wohlfahrtsverbandes zu Evaluationen und legt dabei besonderen Wert auf die Perspektive der Macht. Als Datenbasis dienen ihm zwölf Einzelinterviews sowie vor allem eine Gruppendiskussion mit den Fach- und Führungskräften des Verbandes. Beide Datenquellen wurden durch eine Verknüpfung der Grounded-Theory-Methodologie mit dem Verfahren der Sequenzanalyse ausgewertet.

Teil 3 fokussiert Wirkungen und Folgen von Evaluationen, wobei sich die Beiträge vor allem der Frage nach der Nutzung und dem Nutzen von Evaluationen widmen. Eng damit verbunden ist die Frage, welche Faktoren die Nutzung respektive den Nutzen jeweils beeinflussen.

- *Holger Gärtner* fokussiert in seinem Beitrag den Bereich des Schulwesens, nämlich die potentielle Wirkung von Schülerfeedback als einem Instrument zur Selbstevaluation des beruflichen Handelns: Nutzen Lehrkräfte die gewonnenen Informationen aus der Selbstevaluation für die Weiterentwicklung des Unterrichts? Und wenn ja: wie? Der Autor beantwortet diese Fragen am Beispiel eines Selbstevaluationsportals, das allen Lehrkräften in Berlin sowie Brandenburg zur freiwilligen Nutzung zur Verfügung steht. Anhand einer qualitativen Vorstudie mit Leitfadeninterviews wurden erste Hinweise auf die Wirkmechanismen eines Schülerfeedbacks in Erfahrung gebracht. Im Anschluss wurde eine auf den Ergebnissen aufbauende Online-Fragebogenerhebung durchgeführt, an der sich 305 Nutzerinnen und Nutzer des Selbstevaluationsportals beteiligten.
- Können unterschiedliche Inhalte von Evaluationsrückmeldungen die Lehrgestaltung beeinflussen? *André Nowakowski* stellt in seinem Beitrag drei Studien vor, welche diese Frage im Kontext der Hochschulevaluation mit Blick auf die zunehmende Kompetenzorientierung im Bildungsbereich untersuchen. In einem qualitativ ausgewerteten Experiment erhielten Lehrende ein ergebnisorientiertes Feedback, während einer Kontrollgruppe Rückmeldungen aus einem prozessorientierten Fragebogen vorgelegt wurden. In zwei weiteren quantitativ ausgewerteten Experimenten wurden den Lehrenden ebenfalls ergebnisorientierte und prozessorientierte Rückmeldungen präsentiert. Als abhängige Variable wurde in allen Teilstudien untersucht, inwieweit die unterschiedlichen Rückmeldungen unterschiedliche Lehrorientierungen (Lehrenden- vs. Studierendenfokussierung) anstoßen.
- *Mary Sandermann* geht in ihrem Beitrag der Frage nach, welche Bedeutung soziale Kompetenzen der Evaluierenden und der Nutzenden sowie die Interaktionsgestaltung auf Evaluationsnutzungen bei Programmevaluationen haben. Als Untersuchungsgegenstand dienten sieben externe Programmevaluationen von universitären Mentoringprogrammen, wobei jeweils die externen Evaluierenden und die jeweiligen Hauptinteraktionspartner/innen befragt wurden. Der Studie lag ein Untersuchungsmodell zugrunde, das sechs verschiedene soziale Kompetenzen sowie Umfeld- und Prozessmerkmale der Evaluation berücksichtigt. Zur Erfassung der sozialen Kompetenzen und des Interaktionsgeschehens hat die Autorin Konstrukt-Interviews mit einem stan-

dardisierten Fragebogen kombiniert. Die Ergebnisse der Interviews und der Fragebögen wurden mithilfe einer Fuzzy-Set Qualitative Comparative Analysis ausgewertet.

- *Kathrin Frey* und *Thomas Widmer* haben die Frage untersucht, ob und inwieweit politische Entscheidungsträger/innen in der Schweiz evidenzbasierte Erkenntnisse bei Gesetzes-Novellierungen nutzen. Sie unterscheiden dabei das „Eigenlernen“ (Rückgriff auf Evaluationsergebnisse aus der Schweiz) und das „Fremdlernen“ (Rückgriff auf Evaluationsergebnisse außerhalb der Schweiz) und gehen unter anderem der Hypothese nach, dass Fremdlernen nur dann zum Tragen kommt, wenn die eigene Evidenz als unzureichend beurteilt wird. In qualitativen Fallstudien, die sich auf Dokumentenanalysen und Leitfadeninterviews stützen, werden zwölf Gesetzes-Novellierungen aus unterschiedlichen Politikbereichen analysiert, wie z. B. der Verkehrspolitik, der Migrationspolitik und der Gesundheitspolitik.

Teil 4 versammelt Beiträge, die sich mit übergreifenden Fragestellungen auseinandersetzen oder mehrere Aspekte der unser Gliederung zugrunde liegenden Input-Process-Outcome-Logik betreffen.

- *Jörg Rech* und *Wolfgang Meyer* beschäftigen sich in ihrem Beitrag mit dem Thema Evaluationskultur in der Evaluationszusammenarbeit. Zur Analyse der Evaluationskultur führten die Autoren eine Sekundäranalyse der Daten einer sogenannten Systemprüfung durch. Diese Prüfung hat im Jahr 2007 mithilfe von Dokumentenanalysen, Auswertungen von leitfadengestützten Self-Assessments und Interviews in 19 Organisationen der Entwicklungszusammenarbeit den damaligen Stand der eingesetzten Evaluationssysteme in diesem Sektor erhoben und bewertet. Zur Analyse der vorliegenden Daten setzen die Autoren auf eine Operationalisierung der Einflussfaktoren auf die „Evaluationskultur“ anhand von vier Dimensionen: nationale Vorgaben, sektorale Vorgaben, institutionelle Vorgaben und operationale Vorgaben.
- Systematische Reviews, die Forschungsergebnisse zu einem Thema systematisch zusammenstellen, gewinnen seit mehreren Jahren besonders im Kontext der Evidenzbasierung an Bedeutung. *Wolfgang Böttcher*, *Nina Hogrebe* und *Sabrina Schulz* widmen sich vor diesem Hintergrund der Frage, inwieweit systematische Reviews im Feld der Evaluation brauchbar sein können. Am Beispiel der Erziehungswissenschaft gehen sie der Frage nach, inwieweit sich aus empirisch gewonnenem Wissen aus Evaluationen generalisierbares Wissen für die Gestaltung von politischem und praktischem Wissen erzeugen lässt. Unter Beachtung von Gemeinsamkeiten und Unterschieden von (grund-

legender) Forschung und Evaluation plädiert der Beitrag für die institutionelle Verankerung von „Systematischen Evaluationen“ im deutschsprachigen Raum.

- Seit einigen Jahren hat sich in Deutschland die Schulinspektion als externe Evaluation von Schulen umgesetzt. Explizit werden hier mit der Datengewinnung die Bewertung und Qualitätsentwicklung der Schulen verbunden. *Wolfgang Böttcher, Jan Hense und Miriam Keune* geben einen Überblick über den empirischen Forschungsstand zu Nutzung und Nützlichkeit dieses Evaluationsverfahrens. Die Befunde dieser spezifischen Studien werden vor dem Hintergrund der allgemeinen Nutzungsforschung im Feld der Evaluation interpretiert.
- Einen Blick auf die englischsprachige Forschung über Evaluation nordamerikanischer Provenienz werfen abschließend *Jan Hense und Thomas Widmer*. Sie beschreiben die Entwicklung des Forschungsfelds von den Anfängen der modernen Expansionsphase der Evaluation in den 1960er Jahren bis heute. Darüber hinaus geben sie einen Überblick über Gegenstände und Fragestellungen, die in diesem Zeitraum empirisch bearbeitet wurden. Abschließend werfen sie vor diesem Hintergrund sowie den Beiträgen des vorliegenden Sammelbands einen Blick auf Desiderata und Aussichten der zukünftigen Forschung über Evaluation.

Bei der Erstellung dieses Bandes konnten wir auf zahlreiche Unterstützung bauen – vom Recherchieren über das Korrekturlesen bis hin zum Formatieren und Layouten. Unser herzlicher Dank geht hier vor allem an Mailin Gunkel, Patrick Plettenberg und Tamara Reisdorf. Wir bedanken uns zudem bei allen Autorinnen und Autoren, die zur Entstehung dieses Bandes beigetragen haben und durch ihr Engagement und ihren Einsatz die Forschung über Evaluation im deutschsprachigen Raum vorantreiben. Wir glauben und hoffen, dass dieser Band einen Beitrag dazu leistet, Evaluationen und ihre Wirkung nicht für selbstverständlich hinzunehmen, sondern Evaluationen, ihren Wirkmechanismen sowie ihrem Potenzial und ihren Grenzen mit kritisch positivem Blick zu begegnen. Wir wünschen uns, dass die Beiträge in diesem Band andere dazu inspirieren werden, die dargestellten Forschungsergebnisse anzuwenden, aber auch darauf aufbauend weitere Forschung über Evaluation zu betreiben, um Evaluationen (noch) nützlicher, durchführbarer, fairer und genauer zu gestalten, wie dies von einschlägigen Evaluationsstandards eingefordert wird.

Die Herausgeber, im Juni 2013

Literatur

- Ahn, S., Ames, A. J. & Myers, N. D. (2012). A Review of Meta-Analyses in Education: Methodological Strengths and Weaknesses. *Review of Educational Research*, 82(4), 436–476.
- Balthasar, A. (2000). Evaluationssynthesen: Bedingungen, Möglichkeiten und Grenzen. *LeGes – Gesetzgebung & Evaluation*, 11(1), 13–25.
- Brüsemeister, T. & Eubel, K.-D. (Hrsg.) (2008). *Evaluation, Wissen und Nichtwissen*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- DeGEval – Gesellschaft für Evaluation e.V. (Hrsg.) (2008). *Standards für Evaluation* (4. unveränderte Aufl.). Mainz.
- GAO – General Accounting Office (Hrsg.) (1992). *The Evaluation Synthesis*. Washington, D.C.: GAO.
- Kaufmann, B. (2012). *Akkreditierung als Mikropolitik. Zur Wirkung neuer Steuerungsinstrumente an deutschen Hochschulen*. Wiesbaden: VS Verlag.
- Keller-Ebert, C. (2006). Ziehen Auftraggebende Nutzen aus Evaluationen? Ergebnisse einer Follow-up-Studie. *Sozialwissenschaften und Berufspraxis*, 1, 101–107.
- Mark, M. M. (2008). Building a better evidence base for evaluation theory. Beyond general calls to a framework of types of research on evaluation. In N. L. Smith & P. R. Brandon (Hrsg.), *Fundamental issues in evaluation* (S. 111–134). New York: Guilford Press.
- Mark, M. M., Henry, G. T. & Julnes, G. (1999). Toward an integrative framework for evaluation practice. *American Journal of Evaluation*, 20(2), 177–102.
- Matthies, H. & Simon, D. (Hrsg.) (2008). *Wissenschaft unter Beobachtung. Effekte und Defekte von Evaluationen*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Mieg, H. A. (2010). Professionalisierung: eine dreifache Autonomieregulation. In K. von Schlieffen (Hrsg.), *Professionalisierung und Mediation* (S. 15–26). München: Beck.
- Schneewind, J. (2007). *Wie Lehrkräfte mit Ergebnisrückmeldungen aus Schulleistungsstudien umgehen*, Freie Universität Berlin. Verfügbar unter: http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000002819 [13.06.08].
- Schrader, F.-W. & Helmke, A. (2004). Von der Evaluation zur Innovation? Die Rezeptionsstudie WALZER. Ergebnisse der Lehrerbefragung. *Empirische Pädagogik*, 18, 140–161.
- Scriven, M. (1980). *The Logic of Evaluation*. Inverness, CA: Edgepress.
- Serrano-Velarde, K. (2008). *Evaluation, Akkreditierung und Politik. Zur Organisation von Qualitätssicherung im Zuge des Bolognaprozesses*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Shadish, W. R., Cook, T. D. & Leviton, L. C. (1991). *Foundations of program evaluation. Theories of practice*. Newbury Park, CA: Sage.
- Stockmann, R. (2006). *Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder*. (3. Aufl.). Münster: Waxmann.
- Stufflebeam, D. L. (2001a). Evaluation models. *New Directions for Evaluation*, 89, 7–98.
- Stufflebeam, D. L. (2001b). The metaevaluation imperative. *American Journal of Evaluation*, 22(2), 183–209.

- Suchman, E. A. (1967). *Evaluative research. Principles and practices in public service and social action programs*. New York: Russell Sage Foundation.
- Widmer, T. (1996). *Meta-Evaluation. Kriterien zur Bewertung von Evaluationen*. Bern: Haupt.
- Widmer, T., Beywl, W. & Fabian, C. (Hrsg.) (2009). *Evaluation. Ein systematisches Handbuch*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Widmer, T., Landert, C. & Bachmann, N. (2000). *Evaluations-Standards der Schweizerischen Evaluationsgesellschaft (SEVAL-Standards)*. Bern/Genève: SEVAL.
- Wottawa, H. & Thierau, H. (1998). *Lehrbuch Evaluation*. Bern: Huber.
- Yarborough, D. B., Shulha, L. M., Hopson, R. K. & Caruthers, F. A. (2011). *The program evaluation standards. A guide for evaluators and evaluation users* (3rd ed.). Los Angeles: Sage.

Teil 1

Ausgangsbedingungen von Evaluation

1 Hintergrund

Seit 1995 beteiligt sich die Bundesrepublik Deutschland wieder an internationalen Schulleistungsuntersuchungen (TIMSS im Jahr 1995, Baumert et al., 1997; PISA im Jahr 2000, Deutsches PISA-Konsortium, 2010). Zudem führen und führten einige Bundesländer eigene Vollerhebungen durch (z. B. QuaSUM, Peek, 2004). Heute sind diese Maßnahmen der Qualitätsbemühungen in der Gesamtstrategie zum Bildungsmonitoring der Kultusministerkonferenz (KMK) aus dem Jahr 2006 zusammengefasst. Einen Baustein stellen die „Vergleichsarbeiten zur landesweiten Überprüfung der Leistungsfähigkeit einzelner Schulen“ (KMK, S. 1) dar.

Als eine Konsequenz der verschiedenen Datenerhebungen fordern die Bildungsstandards für die Lehrerbildung (KMK, 2004) von den Lehrkräften die Kompetenz der Rezeption und Nutzung von Ergebnissen der Bildungsforschung. In diesem Zusammenhang wird als wichtiger Aspekt das empirische Verständnis der Lehrkräfte thematisiert (Peek & Dobbelsstein, 2006). In der hier vorgestellten Untersuchung wurde der Frage nachgegangen, inwieweit die für das Verständnis notwendige (Daten-)Kompetenz bei Lehrkräften vorhanden ist und ob diese durch entsprechende Fortbildungen verbessert werden kann.

Das Projekt Vergleichsarbeiten in der Grundschule (VERA)¹, in dessen Rahmen die vorliegende Untersuchung durchgeführt wurde, wird im Folgenden (Abschnitt 2) kurz vorgestellt. Da die Interpretation der dort gegebenen Rückmeldungen statistisches und forschungsmethodisches Wissen bei Lehrkräften voraussetzt, können Erkenntnisse aus der Statistikdidaktik (Abschnitt 3) zur Bestimmung von dabei auftretenden Schwierigkeiten genutzt werden. Zur Messung und Verbesserung der Datenkompetenz von Lehrkräften wurden Fortbildungen durchgeführt (Abschnitt 4). Die wichtigsten dabei gewonnenen Ergebnisse werden zum Abschluss zusammengefasst und in Hinblick auf weitere Entwicklungsschritte diskutiert (Abschnitt 5).

1 Zu den Vergleichsarbeiten in der 8. Jahrgangsstufe siehe Institut zur Qualitätssicherung im Bildungswesen [IQB] (2012).

2 Vergleichsarbeiten

Vergleichsarbeiten in der Grundschule (VERA), wie sie von Helmke und Hosenfeld (2003) zusammen mit dem rheinland-pfälzischen Ministerium für Bildung, Wissenschaft, Jugend und Kultur konzipiert wurden, werden seit 2003 jährlich durchgeführt. Seit 2007 beteiligen sich alle 16 Bundesländer an VERA.² Die flächendeckend durchgeführten Tests, die auf den Bildungsstandards der KMK basieren, werden in jährlich wechselnden Inhaltsbereichen der Fächer Deutsch (z. B. Leseverständnis) und Mathematik (z. B. Raum und Form) geschrieben.

Hauptziel der Vergleichsarbeiten ist das Anstoßen von Unterrichts- und Schulentwicklung.³ Durch die gebotene zusätzliche Außenperspektive der Klassen- bzw. Schulleistung soll die Diskussion in den Fachkollegien angeregt und unterstützt werden. Lehrkräfte werden hierfür aktiv an der Durchführung der Vergleichsarbeiten und der Korrektur der Tests beteiligt, und mit den gelieferten Materialien werden fachliche, fachdidaktische und pädagogisch-psychologische Impulse zur Analyse und Weiterarbeit gegeben.

Vergleichsarbeiten sind als Bausteine zur (Selbst-)Evaluation von Schulen gedacht. Da nur wenige Angaben zum Kontext (z. B. Schule liegt im sozialen Brennpunkt) erhoben werden, müssen die Leistungsrückmeldungen von den Lehrkräften mit zusätzlichen Daten (z. B. über den Unterricht) kombiniert werden. Die Nutzung der Daten liegt somit in der Hand der Lehrkräfte.

2.1 Rückmeldungen

Die Bemühungen zu der im Jahr 2004 von Kohler und Schrader aufgeworfenen „Frage der Rückmeldegestaltung“ (S. 6) haben laut Posch (2009) dazu geführt, dass die Rückmeldungen, welche die auswertenden Institute den Lehrkräften zur Verfügung stellen, ein hohes Niveau erreicht haben. Sie können nach ihrer Aggregatsebene, ihrem Abstraktionsgrad und den angebotenen Vergleichsmaßstäben charakterisiert werden.

2 Die Aufgabenentwicklung wurde bis 2009 durch eine Arbeitsgruppe an der Universität Koblenz-Landau, Campus Landau verantwortet, ab dem Jahr 2010 vom IQB. Die Durchführung der Vergleichsarbeiten liegt in der Verantwortung der einzelnen Bundesländer, die dafür ihre Landesinstitute oder universitäre Institute beauftragen.

3 Die Unterrichts- und Schulentwicklung wird im Rahmen der Vergleichsarbeiten als ein Ziel genannt. Dabei wird davon ausgegangen, dass z. B. die Diskussionen in den Fachkonferenzen sowohl auf die Unterrichts- als auch auf die Schulebene wirken.

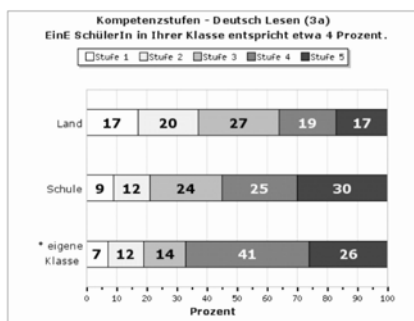
Aggregatsebene der Rückmeldung

Vergleichsarbeiten (z. B. Lernstand 9, Peek & Dobbelstein, 2006; Thüringer Kompetenztests, Nachtigall & Jantowski, 2007; VERA 3, Zimmer-Müller, Hosenfeld & Koch, in Druck) stellen entsprechend der Forderung von Brunner und Light (2008) nach differenzierten Informationen und den Wünschen der Lehrkräfte (Rolff & Gathen, 2008) Daten auf Schülerebene bereit. Um darüber hinaus das vorrangige Ziel der Schul- und Unterrichtsentwicklung zu ermöglichen, sind Analysen auf übergeordneter Ebene notwendig. Vergleichsarbeiten bieten hierfür Daten zur Leistung der Klasse, der Schule sowie verteilungsbezogene Vergleichswerte auf Landesebene. Zusätzlich werden adjustierte Werte berichtet, bei denen die Verteilungen um Einflussfaktoren (z. B. Schule liegt im sozialen Brennpunkt) bereinigt sind, die nicht von den Lehrkräften verändert werden können (sog. fairer Vergleich, Isaac & Hosenfeld, 2008).

Lösungshäufigkeit und kriteriumsorientierte Rückmeldung

Unmittelbar nach Abschluss der Dateneingabe erhalten die Lehrkräfte Lösungshäufigkeiten rückgemeldet. Zusätzlich wird eine, seit 2010 an die Bildungsstandards gebundene, kriteriumsorientierte Rückmeldung angeboten. Entsprechend der empirisch ermittelten Aufgabenschwierigkeiten werden für die verschiedenen Stufen die jeweils beherrschten Kompetenzen beschrieben (Helmke & Hosenfeld, 2004). Mithilfe dieser kriterialen Beschreibung soll eine inhaltliche Interpretation der Rückmeldungen für die Lehrkräfte vereinfacht werden. Gleichwohl kann durch den weiteren Abstraktionsschritt der Informationen die Rezeption erschwert werden. Eine beispielhafte Kompetenzstufenverteilung auf Klassen-, Schul- und Landesebene, wie sie den Lehrkräften vom Zentrum für Empirische Pädagogische Forschung (zepf) zurückgemeldet wird, zeigt die Abbildung 1 (für eine vertiefte Darstellung siehe Zimmer-Müller et al., in Druck).

Abb. 1: Kompetenzstufenverteilungen der Klasse, der Schule und des Landes im Vergleich



Rückmeldungen auf Aufgabenebene

Neben den personenzentrierten Rückmeldungen werden Informationen zu den Aufgaben angeboten, um die fachlich-inhaltliche Auseinandersetzung damit anzuregen. Die Angaben, wie viele Schüler und Schülerinnen die einzelnen Aufgaben lösen konnten und die korrespondierenden Informationen auf Landesebene verdeutlichen besondere Schwierigkeiten bzw. Kompetenzen innerhalb der Klasse. Im Zusammenspiel mit den angebotenen Materialien, in denen die einzelnen Aufgaben fachdidaktisch erläutert und Hinweise zur Weiterarbeit im Unterricht gegeben werden, sollen sie die Implementation einer neuen Aufgabenkultur fördern.

2.2 Nutzung und ihre Bedingungen

In den letzten Jahren wurden mehrere Studien zu verschiedenen externen Rückmeldesystemen durchgeführt. In den meisten Fällen handelt es sich um quantitativ ausgerichtete Befragungen, bei denen die an den Vergleichsarbeiten beteiligten Lehrkräfte um Selbstauskünfte zur Rezeption/Reflexion und daraus abgeleiteten Aktionen gebeten wurden (Dedering, 2011).

Die Verständlichkeit der Rückmeldungen von Vergleichsarbeiten wird von den Lehrkräften dabei meist positiv bewertet (Bonsen, Büchter & Peek, 2006; Koch, Groß Ophoff, Hosenfeld & Helmke, 2006; Nachtigall & Jantowski, 2007). Allerdings wird festgestellt, dass Lehrkräften das Lesen und Interpretieren bzw. Ableiten von Unterrichtsmaßnahmen aus externen Rückmeldungen schwer fällt (Koch et al., 2006; Kohler & Schrader, 2004), auch wird das fehlende empirische Verständnis der Lehrkräfte beklagt (Peek & Dobbelstein, 2006). Da das Verständnis der rückgemeldeten Daten im Modell der Unterrichtsentwicklung (Helmke, 2004) als Voraussetzung für das angemessene Ableiten von Unterrichtsentwicklungsmaßnahmen angesehen wird, kann diese Befundlage als kritische Einschränkung für eine zielgerichtete Nutzung der Rückmeldungen angesehen werden.

Die im Projekt „DATAUSE – Nutzung von Daten für die Verbesserung von Schul- und Schülerleistung“ befragten Lehrkräfte machen u. a. eigene fehlende Fähigkeiten und mangelnde Unterstützung für die geringe Nutzung von Daten im Allgemeinen zur Schul- und Unterrichtsentwicklung verantwortlich (Schildkamp, 2011).

3 Datenkompetenz

Zur Bestimmung und Verbesserung der Datenkompetenz der Lehrkräfte kann auf das Wissen der Statistikdidaktik und die Verfahren der explorativen Datenanalyse zurückgegriffen werden. Für Pfannkuch und Wild (2004) ist die Einsicht notwendig, dass viele Gegebenheiten des Lebens ohne geeignete Daten und deren Analyse nicht angemessen beurteilt werden können. Das Prinzip der Varianz stellt dabei ein wichtiges Konzept der Statistik und der Weltsicht dar. Gal (2004) betont neben den kognitiven Kompetenzen vor allem die Fähigkeit, statistische, datenbezogene Argumente kritisch bewerten und die daraus abgeleiteten Schlussfolgerungen erörtern zu können.

3.1 Explorative Datenanalyse (EDA)

Von den Lehrkräften wird bei der Interpretation der Rückmeldungen eine Vorgehensweise gefordert, die viele Gemeinsamkeiten mit der von Tukey (1977) entworfenen explorativen Datenanalyse hat. Die von ihm genutzte Metapher des „Datendetektivs“ (S. 1) beschreibt das Wechselspiel zwischen Fragen stellen, Hypothesen bilden und der Suche nach belegenden und widersprechenden Indizien. Dies geschieht in einer engen Verknüpfung mit dem Kontext, aus dem die Daten stammen. Entsprechend sollen die Lehrkräfte anhand der erhaltenen Rückmeldungen Hypothesen über Ursachen formulieren und deren Gültigkeit schrittweise über die Integration weiterer Informationen prüfen. Dabei ist die Berücksichtigung von Kontextwissen, das nur ihnen vorliegt, für das Verständnis und die Interpretation der Rückmeldungen Grundvoraussetzung. Bei der explorativen Datenanalyse dürfen Daten nicht ohne gründliche Analyse durch die Angabe von Mittelwert und Streuungsmaß zusammengefasst werden (Biehler, 1999). So sollte auch bei der Analyse von Rückmeldungen aus Vergleichsarbeiten die gesamte Leistungsverteilung der Klasse betrachtet und nicht auf ein „besser oder schlechter als“ reduziert werden. Ebenso wird im Zusammenhang der explorativen Datenanalyse betont, dass auch Daten, die wie die Rückmeldungen aus Vergleichsarbeiten auf kleinen Stichproben (z. B. nur wenige Items pro Schüler/in, nur wenige Schüler/innen in der Klasse) beruhen, mit ihrer Hilfe untersucht werden können (Biehler, 1995).

Eine besondere Bedeutung kommt der Nutzung verschiedener Grafikarten zu. Sie werden nicht nur zur Ergebnispräsentation, sondern als Werkzeug zur Analyse selbst genutzt. Dabei erleichtert das Spielen mit den Daten anhand von Grafiken (Brunner & Light, 2008) – beispielsweise bei der Betrachtung von Werten einzelner Schüler/innen in verschiedenen Verteilungen – das Verständnis vor allem für Analysierende mit wenig Erfahrung (Mandinach, Honey & Light, 2006).

Ebenso können Darstellungen von Subgruppen (über Schüler/innen oder Aufgaben) zur adaptiven Unterrichtsplanung genutzt werden. Curcio (2001) beschreibt drei Fähigkeitsstufen beim Umgang mit grafischen Darstellungen. Auf der untersten Stufe (*reading the data*) gelingt das explizite Lesen von Informationen aus Tabellen und Grafiken. Personen, welche die Ebene *reading between the data* erreicht haben, können Beziehungen zwischen Daten innerhalb einer Grafik herstellen und interpretieren. Auf der obersten Stufe (*reading beyond the data*) können darüber hinaus Verknüpfungen zu Hintergrundwissen gebildet werden.

Da alternative Sichten des untersuchten Sachproblems im Mittelpunkt stehen sollen, stellt die Kommunikation und Diskussion mit anderen über die gefundenen Zusammenhänge einen weiteren wichtigen Schritt der Analyse dar (Biehler, 1999).

3.2 Schwierigkeiten bei der Analyse von Daten

Die Analyse von Rückmeldungen aus Schulleistungstests durch Lehrkräfte wurde bisher nur selten untersucht. Makar und Confrey (2004) führten entsprechende Studien mit Lehrkräften und Lehramtsstudierenden durch, wobei das Hauptziel in der Verbesserung des Mathematikunterrichts im Bereich der Statistik lag. Sie unterscheiden bei der Aufgabe eines Vergleichs von Subgruppen fünf Entwicklungsstufen, von der Gegenüberstellung von individuellen Datenpunkten, über die Nutzung von zusammenfassenden Statistiken bis zum Verständnis der Variabilität. In der von ihnen durchgeführten Fortbildung wurde die Analysefähigkeit der Lehrkräfte verbessert und ein kritisches Hinterfragen z. B. von persönlichen Überzeugungen geschärft. Mandinach et al. (2006) stellen bei Lehrkräften die Bevorzugung einer Betrachtung von einzelnen Schüler/innen fest, die eine Verallgemeinerung auf andere Schüler/innen, andere Klassen, das folgende Schuljahr einschränken. In ähnlicher Weise beschreiben Konold, Higgins, Russell und Khalil (2004) vier Perspektiven vom Blick auf den Wert eines einzelnen Falls bis zur Betrachtung der gesamten Verteilung. Erfahrene Statistiker/innen, welche die oberste Stufe erreicht haben, werden den Einzelfall in Beziehung zur Gesamtverteilung – ihres Mittelwerts und ihrer Varianz – betrachten. Um – basierend auf Rückmeldungen – Maßnahmen für den Unterricht ableiten zu können, ist die Untersuchung der Gesamtverteilung, von Subgruppen und der Wechsel zwischen Individual- und Aggregatsebene notwendig. Konold et al. (2004) heben hervor, dass nicht nur die Unfähigkeit zur Analyse der Aggregatsebene, sondern auch die Vorliebe für den Einzelfall mögliche Analysen und Schlussfolgerungen beeinflusst bzw. behindert. Darüber hinaus gilt es zu bedenken, dass eigene Überzeugungen und Alltagswissen die Datensammlung und Analyse problematisch beeinflussen können (Hancock, Kaput & Goldsmith,

1992). Einerseits erleichtert beispielsweise das Wissen, das Lehrkräfte über den Klassenkontext besitzen, die Analyse, andererseits birgt dieses Kontextwissen jedoch die Gefahr, dass vorschnell Ursachen und angebliche Lösungen ausgemacht werden.

4 Untersuchung zur Datenkompetenz bei Lehrkräften im Rahmen von Vergleichsarbeiten

Als ein Aspekt des angemahnten empirischen Verständnisses von Lehrkräften (Peek & Döbelstein, 2006) wird in der hier vorgestellten Untersuchung die Datenkompetenz der Lehrkräfte betrachtet. Über die Erhebung des Ist-Zustands hinaus interessieren dabei Verbesserungsmöglichkeiten. In den Fortbildungen, im Rahmen der Studie, ging es jedoch nicht isoliert um Datenkompetenz, sondern um die Qualität der Ergebnisanalyse als Grundlage für eine angemessene Ableitung von Maßnahmen (ausführlich hierzu und zu motivationalen Aspekten Koch, 2011).

4.1 Untersuchungsdesign

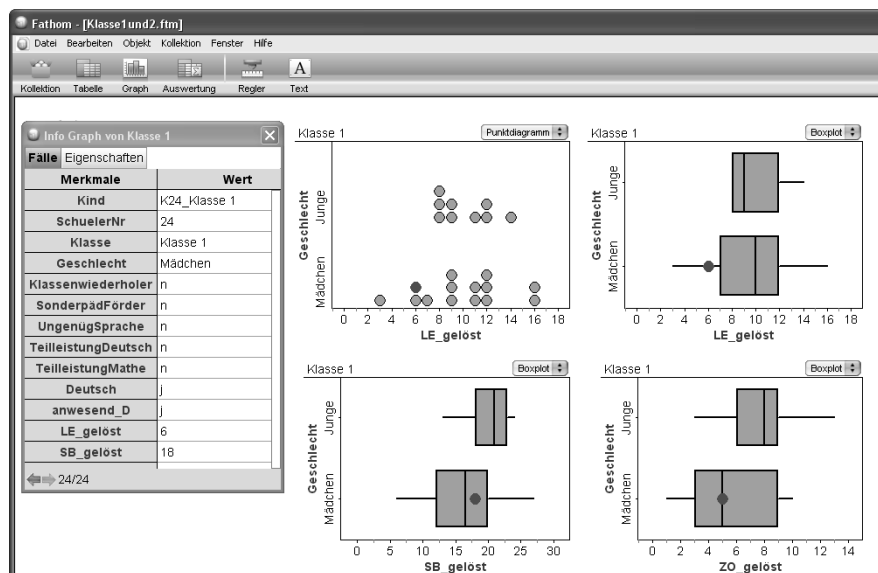
Zur Untersuchung der Fragestellungen – Datenkompetenz der Lehrkräfte, Möglichkeiten zu deren Steigerung und Differenziertheit der Ergebnisanalyse durch die Lehrkräfte – wurde ein quasi-experimentelles Design mit vier Gruppen gewählt. Hierfür wurde eine eintägige Fortbildung entwickelt, die über das rheinland-pfälzische Institut für schulische Fortbildung und schulpädagogische Beratung (IFB) den Lehrkräften zur freiwilligen Teilnahme angeboten wurde.

Der in der explorativen Datenanalyse geforderte dynamische und spielerische Umgang mit Daten wird bei den bisherigen Rückmeldungen aus Vergleichsarbeiten noch kaum geboten. Bei den Fortbildungen wurde deshalb die Analysesoftware Fathom⁴ eingesetzt (Biehler, Hofmann, Maxara & Prömmel, 2006; Key Curriculum Press, 2006). Das Programm (Abb. 2) erleichtert das Verständnis der Daten durch den einfachen Wechsel zwischen verschiedenen Grafikarten. Darüber hinaus werden angewählte Datenobjekte in allen Darstellungen hervorgehoben (multiple verbundene Repräsentationen). In der Abbildung ist z. B. die Anzahl gelöster Leseaufgaben (LE_gelöst) gruppiert nach Geschlecht als Punktediagramm und als Boxplot dargestellt, zusätzlich die Anzahl gelöster Aufgaben im

4 Fathom wurde zur Vermittlung der Datenanalyse und Stochastik in den USA entwickelt. Unter der Leitung von Prof. Biehler entstand eine deutsche Oberfläche inkl. Hilfesystem.

Inhaltsbereich „Sprache und Sprachgebrauch untersuchen“ (SB_gelöst) und im mathematischen Inhaltsbereich „Zahlen und Operationen“ (ZO_gelöst). Im Beispiel ist Schülerin K24_Klasse1 ausgewählt.

Abb. 2: Die Analysesoftware „Fathom“



Um allen Lehrkräften die Untersuchung der Daten mithilfe der Analysesoftware zu ermöglichen, wurde ein verzögertes Treatment für den Einsatz der Software realisiert. In zwei Gruppen wurde nur mit der Software gearbeitet. Die beiden Gruppen unterscheiden sich in der Analysenreihenfolge der Daten der eigenen bzw. fremden Klassen. In den beiden anderen Gruppen wurde zuerst eine Rückmeldung im VERA-Format und im Anschluss mithilfe von Fathom analysiert. Diese beiden Gruppen untersuchten im ersten Teil nur Daten fremder Klassen im VERA-Format, da davon ausgegangen wurde, dass die Lehrkräfte die Rückmeldungen zur eigenen Klasse vorab schon in diesem Format analysiert hatten. Im zweiten Block unterschieden sich die beiden Gruppen darin, ob sie eigene Daten oder die fremder Klassen mithilfe der dynamischen Analysesoftware untersuchten. Vor dem Treatment wurden die bisherigen Erkenntnisse auf Basis der VERA-Rückmeldungen zur eigenen Klasse und die Datenkompetenz der Lehrkräfte erhoben. Während der Analyse der Rückmeldungen notierten die Lehrkräfte die Ergebnisse der Untersuchung. Am Ende des Tages wurden die Lehr-