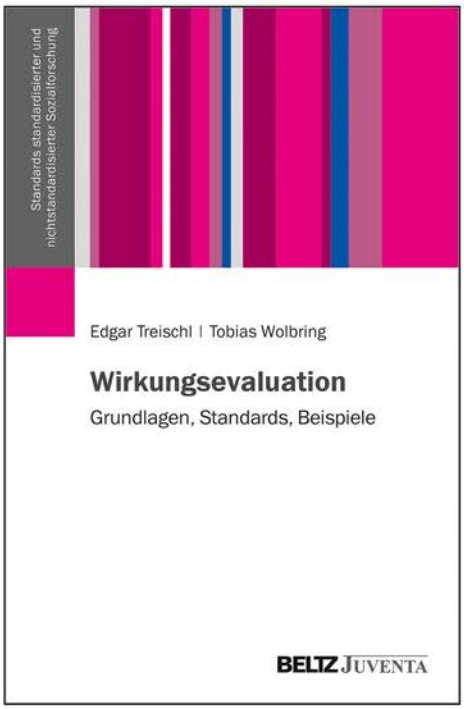

Leseprobe aus Treischl und Wolbring, Wirkungsevaluation, ISBN 978-3-7799-3924-5

http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-3924-5

1 Einleitung

Evaluationen, in einem breiten Sinne verstanden also jedwede Formen der mehr oder weniger systematischen Bewertung des Werts oder der Qualität einer Sache, Person oder Maßnahme (Scriven 1991), sind in vielen gesellschaftlichen Teilbereichen fest etabliert. Man denke nur an die Evaluation des Mindestlohns, die Bewertung von Schulleistungen im Zuge von PISA, Ratings zur Bonität von Unternehmen oder Untersuchungen zur Wirksamkeit von Integrationsmaßnahmen gesellschaftlicher Randgruppen. Seit geraumer Zeit ist im deutschsprachigen Raum vermehrt zu beobachten, dass Evaluationskommissionen einberufen und separate Qualitätssicherungsabteilungen in Organisationen geschaffen werden. Die Vergabe und der Einsatz öffentlicher Mittel werden meist an die Durchführung eines Wirkungsnachweises geknüpft. Während entsprechende Veränderungen in den USA und anderen angloamerikanischen Ländern bereits in den 1970er Jahren eingesetzt haben (Rossi/Lipsey/Freeman 2004), ist für Deutschland und eine Reihe weiterer europäischer Länder insbesondere seit den 2000er Jahren eine Verschiebung in Richtung einer Bewertungsgesellschaft zu diagnostizieren (Stockmann 2010a).

Die Ausbreitung des Evaluationsparadigmas ist dabei eng verknüpft mit gesellschaftlichen Prozessen einer zunehmenden Rationalisierung und Ökonomisierung zuerst der Arbeits- und dann auch der Lebenswelt (Braun/Wolbring 2012; Stockmann 2010a). Die Attraktivität von Evaluationen speist sich dabei wesentlich aus der Erwartung, dass sie eine objektive, rationale und faire Entscheidungsgrundlage, insbesondere für die Allokation knapper Mittel, liefern. Denn die vermeintliche Faktizität und Eindeutigkeit von Evaluationsergebnissen scheint vielfach bereits konkrete Anschlussmaßnahmen zur Verbesserung oder Optimierung nahezulegen und entlastet somit bei Entscheidungen. Im Sinne Poppers (1944/2000) Idee einer „Stückwerk-Sozialtechnik" sollen Evaluationen somit Maßnahmen auf ihre instrumentelle Tauglichkeit prüfen und helfen, diese auf Basis empirischer Evidenz schrittweise weiterzuentwickeln, anstatt ungeprüfte gesellschaftliche Umwälzungen im Ganzen anzustreben (siehe auch Campbell 1969). Eine willkommene Nebenfolge dieser scheinbaren Versachlichung und Objektivierung der Entscheidungsfindung ist, dass Evaluationen den so getroffenen Entscheidungen erhöhte Legitimität verleihen und, im Falle von externer oder interner Kritik oder gar schädlicher Nebenfolgen von Maßnahmen oder Reformen, zur Rechtfertigung herangezogen werden können.

Vor diesem praktischen Verwertungshintergrund erscheint es essenziell, dass Evaluationen tatsächlich das messen, was sie messen sollen, damit die daraus gezogenen Schlussfolgerungen und Konsequenzen korrekt sind. Die Validität der Messung ist damit eine Schlüsselfrage der Evaluationsforschung; die falsche

Interpretation von Evaluationsresultaten einer ihrer wesentlichen Fallstricke. Kritische Stimmen gegenüber der Evaluation weisen in diesem Zusammenhang regelmäßig darauf hin, dass Messungen immer unvollständig sind, also zur Komplexitätsreduktion bestimmte Aspekte und Dimensionen systematisch ausgeblendet werden müssen. Festlegungen, was bei der Qualitätsmessung als wichtig und unwichtig erachtet wird, beeinflussen in Folge den Ausgang des Bewertungsverfahrens. In diesem Sinne dienen Evaluationen nicht nur der Messung, sondern auch der Konstruktion von Qualität (Lamont 2012). Dies erscheint insbesondere dann problematisch, wenn durch die Ausblendung bestimmter Leistungen und (Neben)Wirkungen einzelne Personengruppen oder Bereiche systematisch benachteiligt werden.

Aus diesen und weiteren Gründen könnte man sich nun auf den Standpunkt stellen, wie es manche Autorinnen und Autoren tun, dass Evaluationen per se abzulehnen sind. So wird mitunter angeführt, dass Evaluationen zwar meist nicht mehr als ein modernes Ritual seien (Schwarz 2006), aber oft erhebliche monetäre, zeitliche und mitunter emotionale Ressourcen beanspruchen. Zudem wird kritisiert, dass der potentielle Schaden, den entsprechende Bewertungsverfahren durch Schürung von Konflikten und Verbreitung von Misstrauen verursachen, oftmals den erwarteten Nutzen überwiege. Dies trifft insbesondere dann zu, wenn sich Widerstände formieren oder sich – aufgrund der häufig starken Anreize, in einer Evaluation „gut" abzuschneiden – nicht-intendierte und unerwünschte Effekte einstellen.

Der Umstand, dass Ratings und Rankings nicht nur steuerungswirksam sein können, sondern oft auch Reaktivität bewirken, gehört fraglos zu den Basisweisheiten der Forschung über Evaluation. Menschen reagieren schließlich darauf, dass sie vermessen werden und passen ihr Verhalten an. Evaluationen werden daher von skeptischen Stimmen nicht nur mit Krankheiten („Evaluitis") verglichen (Frey 2007), sondern auch als moderne Herrschaftsinstrumente verstanden, die vermeintlich keinen echten Mehrwert produzieren, aber dazu beitragen, eine bestimmte neoliberale Management-Ideologie durchzusetzen (Münch 2009).

Eine entsprechende grundsätzlich ablehnende Haltung gegenüber Evaluationen löst freilich nicht das Bewertungs- und Entscheidungsproblem, vor welchem individuelle und kollektive Akteure in verschiedenen Kontexten regelmäßig stehen und aufgrund dessen sich Evaluationen in modernen Gesellschaften so großer Verbreitung erfreuen. Eine grundsätzlich ablehnende Haltung, ohne konkrete bessere Alternativen zur Lösung des virulenten Entscheidungsproblems anzubieten, erscheint daher mehr als unbefriedigend.

Dies sei an einem Beispiel der Personalauswahl in Unternehmen verdeutlicht: Fünf Personen bewerben sich auf dieselbe Stelle. Die Kandidatinnen und Kandidaten unterscheiden sich hinsichtlich ihrer Qualifikationen und haben individuelle Stärken und Schwächen. Wie entscheidet eine Personalabteilung, welche Person den Zuschlag für die Stelle erhält? Wie kürzlich für Besetzungsverfahren

für die Wissenschaft vorgeschlagen wurde (Osterloh/Frey 2016), könnte man schlicht würfeln und den Zufall walten lassen. Ein solcher Auswahlmechanismus widerspricht jedoch offensichtlich dem in der westlichen Welt weithin geteilten meritokratischen Prinzip, wonach der bzw. die „beste" Aspirant/in die Stelle erhalten sollte. Eine Verletzung dieses Prinzips dürfte daher das Risiko einer Klage durch Mitbewerber/innen bergen. Die Personalabteilung wird also die „Qualität" der Bewerberinnen und Bewerber hinsichtlich verschiedener Kriterien wie Qualifikationen, Passung und Motivation bewerten wollen und versuchen, diese Bewertungen hinsichtlich verschiedener Einzeldimensionen zu einem Gesamturteil zu aggregieren. Es findet also eine Form der Evaluation statt, selbst wenn hierzu kein systematisches Evaluationsverfahren genutzt wird, also beispielsweise schlicht nach Sympathie oder dem Zeitpunkt des Eingangs der Bewerbungsunterlagen ausgewählt werden würde.

Erkennt man an, dass Bewertungen in diesem und in zahllosen anderen gesellschaftlichen Feldern – z. B. Festlegung eines Mindestlohns, Markteinführung eines neuen Produkts, Bewertung eines Deutschkurses zur Integration von Personen mit Migrationshintergrund – unumgänglich sind, so lässt sich gleichzeitig nicht ernsthaft fordern, auf Evaluationen im weiteren Sinne gänzlich zu verzichten. Entscheidungen werden schließlich bei jeglicher Form der Auswahl getroffen, also selbst dann, wenn nicht entschieden wird (Luhmann 1984). Entsprechend wird auch immer eine (zumindest ordinale) Bewertung von Alternativen vorgenommen. Der damit einhergehende unvermeidliche Entscheidungsdruck hat dabei nach Ansicht mancher Sozialtheoretikerinnen und Sozialtheoretiker in modernen Gesellschaften substantiell zugenommen (Schimank 2005), was mit bekannten Diagnosen einer gesellschaftlichen Rationalisierung (Weber 1920/1981) und Individualisierung (Beck 1986) korrespondiert.

Statt also gänzlich auf Evaluationen zu verzichten, muss nach unserem Erachten vielmehr das Ziel sein, Evaluationsverfahren so zu gestalten, dass sie möglichst informative und belastbare Ergebnisse liefern. Denn Evaluationen können inhaltlich und methodisch besser oder schlechter durchdacht sein. Dies bedeutet freilich auch, die oben genannten Fallstricke, Kritikpunkte und Bedenken ernst zu nehmen und Bewertungsverfahren entweder so zu gestalten, dass diese Probleme nicht auftreten, oder aber die entsprechenden Begrenzungen anzuerkennen und vor allem das Instrument der Evaluation nicht überzustrapazieren.

Um entsprechende Möglichkeiten und Grenzen von Evaluation zu erkennen, bedarf es eines geschulten Auges, was neben inhaltlichen und sozialen Aspekten auch methodische Expertise erfordert. Das vorliegende Buch hat die Vermittlung dieser zentralen Schlüsselkompetenzen zum Ziel: die Formulierung von Mindeststandards und allgemeinen Prinzipien für gelingende Evaluationen, aber auch die Anerkennung von Fallstricken und Begrenzungen von Bewertungsverfahren. Was dabei nicht geleistet werden kann, ist die Vorlage eines einfachen, universell anwendbaren „Kochrezepts". Evaluationen sollten maßgeschneidert sein; ihre

konkrete Gestaltung sollte stets vom spezifischen Evaluationsgegenstand, dem Erkenntnisinteresse und vom Evaluationskontext geleitet sein. Die allgemeinen Leitlinien und Empfehlungen werden daher im Folgenden anhand konkreter Anwendungsbeispiele aus verschiedenen Bereichen (u. a. Arbeitsmarkt, Bildung, Entwicklungszusammenarbeit, Familie, Gesundheit, Kriminalität) illustriert, die spezifische Problemstellungen ausleuchten, aber auch in verschiedenen Feldern wiederkehrende Gefährdungen verdeutlichen. Beispiele sowohl erfolgreicher als auch gescheiterter Evaluationsvorhaben veranschaulichen daher nicht nur Standards und Fallstricke, sondern sollen darüber hinaus den flexiblen, situationsspezifischen Umgang mit entsprechenden Herausforderungen vermitteln.

Das Buch gliedert sich im Folgenden in drei Teile mit insgesamt acht inhaltlichen Kapiteln. In Teil I werden in zwei Kapiteln allgemeine Grundlagen der Evaluation vermittelt. *Kapitel 2* führt dabei die in der Literatur mittlerweile fest etablierte Unterscheidung zwischen alltäglicher und wissenschaftlicher Evaluation ein. Nach Behandlung einiger zentraler, allgemeiner Evaluationsstandards, wie sie etwa von amerikanischen und deutschen Fachverbänden formuliert wurden, wird argumentiert, dass nur die Messung und Bewertung kausaler Wirkungen einer Maßnahme – also nicht jedwede Form der mehr oder weniger systematischen Bewertung – als eine Evaluation im engeren Sinne verstanden werden soll. Diese definitorische Eingrenzung von Evaluation auf die Wirkungsanalyse ist für das vorliegende Buch fundamental, da sie die Fokussierung der weiteren Ausführungen auf den Aspekt kausaler Effekte begründet.

Kapitel 3 behandelt einige grundlegende Entscheidungen und Festlegungen, die vor einer Evaluation zu treffen sind. Es ist in diesem Zusammenhang notwendig, zunächst einige zentrale Unterscheidungen, etwa die Begriffspaare formativ/summativ, extern/intern und qualitativ/quantitativ, einzuführen, die Argumente für und wider die einzelnen Ansätze abzuwägen sowie deren Komplementaritäten und wechselseitige Ergänzung aufzuzeigen. Anschließend werden ausgewählte Evaluationsmodelle präsentiert. Wesentliches Ziel dieses Abschnitts ist es, einen Eindruck von dem praktischen Ablauf eines Evaluationsvorhabens – von der Planung über die Durchführung der Datenerhebung bis zur Berichtslegung und Entscheidung über Anschlussmaßnahmen – zu vermitteln. Insbesondere in der Planungsphase ist eine klare Explikation einer Programmtheorie für den Erfolg eines Evaluationsvorhabens essenziell. Hierbei sollten die Ziele einer Maßnahme und die zugrunde gelegten Kausalhypothesen zum Zusammenhang von Inputs, Outputs und längerfristigen Outcomes herausgearbeitet werden. Diesem Gesichtspunkt wird im dritten und letzten Abschnitt dieses Kapitels gesonderte Aufmerksamkeit gewidmet.

Für die Wirkungs- und Kausalanalyse *(Teil II)* ist neben einer umfassenden und präzisen Programmtheorie auch die Wahl eines geeigneten Forschungsdesigns entscheidend. In *Kapitel 4* wird daher zunächst das fundamentale Problem der kausalen Inferenz erläutert. Verkürzt gesprochen ergibt es sich aus der

Tatsache, dass Menschen zu ein und demselben Zeitpunkt nicht unter zwei Versuchsbedingungen beobachtet werden können. Einer der beiden Zustände bleibt daher stets kontrafaktisch. Auf Grundlage dieses kontrafaktischen Verständnisses von Kausalität wird anschließend aufgezeigt, weshalb Evaluationen auf Basis einfacher quasi-experimenteller Querschnittsdesigns oder nicht-experimenteller Vorher-Nachher-Messungen ohne Drittvariablenkontrolle auf teils starken Annahmen beruhen, die in der Praxis häufig nicht erfüllt sind. In der Regel wird man daher weitreichende Entscheidungen ungerne auf Grundlage einer solchen unzuverlässigen Datenbasis treffen wollen.

Daran anschließend greift *Kapitel 5* drei mögliche Lösungsansätze auf, die für Wirkungsanalysen oft deutlich besser geeignet sind und zentrale Anforderungen kausaler Inferenz Rechnung tragen. Erstens kann man versuchen, dem durch Störgrößen verursachten Problem der kausalen Inferenz mittels statistischer Verfahren, wie Regressionsanalyse und Matching, zu begegnen. Zweitens besteht ein möglicher Ausweg in der Nutzung von Längsschnittdaten und damit einhergehender Veränderungsmessung. Drittens, so wird gezeigt, bietet insbesondere ein (feld)experimenteller Ansatz klare Vorzüge bei der Bearbeitung kausalanalytischer Fragestellungen und wird daher, in seiner Reinform, oftmals als Referenzdesign für Evaluationsvorhaben betrachtet (Kromrey 2001). Alle drei Ansätze sind in der Praxis freilich nicht ohne Probleme, was in Kapitel 5 angedeutet und im folgenden Kapitel vertieft wird.

Kapitel 6 behandelt praktische Fallstricke der Wirkungsevaluation. Nach Einführung einer weit verbreiteten Validitätstypologie wird eine Vielzahl von Gefährdungen für die verschiedenen Formen von Validität (interne, externe, Konstrukt-, statistische Validität) behandelt. Fallstricke können sich zum Beispiel durch systematische Messfehler, fehlende oder ungeeignete Kontrollgruppen, Abweichungen vom Untersuchungsplan sowie intervenierende Ereignisse ergeben. Aber auch Widerstände durch von der Evaluation betroffene Akteure oder Manipulationsversuche durch Personen, die an einem bestimmten Resultat interessiert sind, können Evaluationsvorhaben beschädigen und unterminieren. Dieser Teil des Buches ist stark geprägt durch das Standardwerk „Experimental and Quasi-experimental Designs for Generalized Causal Inference" von Shadish, Cook und Campbell (2002).

Daran anknüpfend werden in *Teil III* drei unterschiedliche Anwendungsbeispiele ausführlicher dargestellt und Möglichkeiten und Grenzen ausgewählter Untersuchungsdesigns für die Wirkungsevaluation aufgezeigt. *Kapitel 7* illustriert mögliche Fallstricke von querschnittlichen Evaluationen am Beispiel der kausalen Wirkung der Klassengröße auf die individuelle Leistung der Schülerinnen und Schüler. *Kapitel 8* stellt längsschnittliche Evaluationen in den Vordergrund und verdeutlicht mögliche Fallstricke anhand der Evaluation der Einführung eines Mindestlohns. Schließlich werden in *Kapitel 9* Potenziale und Grenzen feldexperimenteller Evaluationen diskutiert, basierend auf der Frage nach der

Inhalt

Teil III
Anwendungsbeispiele