

Standards standardisierter und
nichtstandardisierter Sozialforschung

Edgar Treischl | Tobias Wolbring

Wirkungsevaluation

Grundlagen, Standards, Beispiele

BELTZ JUVENTA

Edgar Treischl | Tobias Wolbring
Wirkungsevaluation

Standards standardisierter und nichtstandardisierter Sozialforschung

Herausgegeben von Nicole Burzan |
Paul Eisewicht | Ronald Hitzler

Edgar Treischl | Tobias Wolbring

Wirkungsevaluation

Grundlagen, Standards, Beispiele

BELTZ JUVENTA

Die Autoren

Dr. Edgar Treischl ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Empirische Wirtschaftssoziologie an der Friedrich-Alexander-Universität Erlangen-Nürnberg. Seine Arbeitsschwerpunkte sind Bildungs- und Hochschulforschung, Methoden der empirischen Sozialforschung und Data Science.

Prof. Dr. Tobias Wolbring ist Inhaber des Lehrstuhls für Empirische Wirtschaftssoziologie an der Friedrich-Alexander-Universität Erlangen-Nürnberg. Seine Arbeitsschwerpunkte sind Bildungs- und Hochschulforschung, Methoden der empirischen Sozialforschung und Wirtschaftssoziologie.

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung ist ohne Zustimmung des Verlags unzulässig. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronische Systeme.



Dieses Buch ist erhältlich als:

ISBN 978-3-7799-3924-5 Print

ISBN 978-3-7799-5182-7 E-Book (PDF)

1. Auflage 2020

© 2020 Beltz Juventa

in der Verlagsgruppe Beltz · Weinheim Basel

Werderstraße 10, 69469 Weinheim

Alle Rechte vorbehalten

Herstellung: Ulrike Poppel

Satz: text plus form, Dresden

Druck und Bindung: Beltz Grafische Betriebe, Bad Langensalza

Printed in Germany

Weitere Informationen zu unseren Autor_innen und Titeln finden Sie unter: www.beltz.de

Inhalt

Vorwort	7
1 Einleitung	11
Teil I	
Allgemeine Grundlagen der Evaluation	
2 Grundlagen, Definitionen, Mindestanforderungen	18
2.1 Evaluation und Evaluationsforschung	18
2.2 Standards für Evaluierende und Evaluationen	24
2.3 Evaluation als Wirkungsmessung	28
3 Planung und Ablauf einer Evaluation	35
3.1 Formen der Evaluation	35
3.2 Ablauf von Evaluationen	38
3.3 Programmtheorie: Maßnahmen, Mechanismen, Effekte	55
Teil II	
Wirkungs- und Kausalanalyse	
4 Wirkungsanalyse und das Problem kausaler Inferenz	66
4.1 Das kontrafaktische Modell der Kausalität	67
4.2 Kausale Inferenz mittels Vorher-Nachher-Vergleich	72
4.3 Kausale Inferenz mittels Kontrollgruppenansatz	75
5 Drei Lösungsansätze für das Problem kausaler Inferenz	84
5.1 Spielarten der Subgruppenanalyse mit Querschnittsdaten	84
5.2 Längsschnittdatenanalyse mit Kontrollgruppe	90
5.3 Feldexperimente: Randomisierung als Ausweg?	94
6 Gefährdungen der internen und externen Validität	99
6.1 Interne Validität	99
6.2 Externe Validität	101
6.3 Ausblick: Anwendungsbeispiele	103

Teil III

Anwendungsbeispiele

7	Querschnittliche Evaluationen von Bildungsreformen	106
7.1	Die Klassengröße als bildungspolitische „Stellschraube“	107
7.2	Untersuchungsdesign und Befunde	109
7.3	Praktische Fallstricke und Limitationen	111
7.4	Weitere Forschung zu den Effekten der Klassengröße	119
7.5	Zusammenfassung	122
8	Längsschnittliche Evaluationen von Arbeitsmarktrefor-	123
8.1	Der Mindestlohn zur Reduktion von Einkommensarmut	124
8.2	Untersuchungsdesign und Befunde	127
8.3	Praktische Fallstricke und Limitationen	128
8.4	Weitere Forschung zu den Effekten des Mindestlohns	137
8.5	Zusammenfassung	138
9	Feldexperimentelle Evaluationen	
	von Gesundheitsinterventionen	140
9.1	Lebensmittelampeln zur Steuerung des Ernährungsverhaltens	143
9.2	Untersuchungsdesign und Befunde	145
9.3	Praktische Fallstricke und Limitationen	146
9.4	Weitere Forschung zu den Effekten der Lebensmittelampel	152
9.5	Zusammenfassung	154
10	Fazit und Ausblick	155
	Literaturverzeichnis	159

Vorwort

Der weltweite Erfolg des Evaluationsparadigmas in Wissenschaft und Gesellschaft spiegelt sich in einer Vielzahl an Lehr- und Handbüchern zum Thema.¹ Ziel des vorliegenden Buches kann und soll es daher nicht sein, der langen Liste an Monographien und Sammelbänden zur Evaluation einfach noch ein weiteres Buch hinzuzufügen, das alles bereits Geschriebene noch einmal in eigenen Worten wiederholt. Insbesondere die Geschichte der Evaluation, v. a. in den USA und Europa, wurde ebenso wie die zentralen Grundlagen und grundsätzlichen Herausforderungen von Evaluationen in lesenswerten Beiträgen ausführlich dargestellt und diskutiert. Einige dieser Aspekte (z. B. Leitlinien und Fallstricke von Evaluationen im Allgemeinen) sind so fundamental, dass sie sich auch im vorliegenden Buch nicht vollkommen ausblenden lassen, und daher im ersten Teil dieses Buches nochmals kompakt erläutert werden.

Der Fokus des vorliegenden Buches liegt aber auf der Wirkungsevaluation, also auf einer spezifischen Form der Evaluation, bei der sich ganz spezielle Herausforderungen hinsichtlich der *Identifikation kausaler Effekte* ergeben. Der Bedarf für ein solches, auf die Wirkungsevaluation zugeschnittenes Lehrbuch ergibt sich aus mindestens zwei Gründen. Zum einen behandeln viele der oben genannten Beiträge Fragen der Wirkungsevaluation erstaunlich sparsam. Dies überrascht, da bei Evaluationen doch sehr häufig die Frage im Mittelpunkt steht, welche Wirkungen eine bestimmte Maßnahme entfaltet hat. Zum anderen hat sich in den Sozial- und Verhaltenswissenschaften in den letzten Jahrzehnten ein interdisziplinäres Feld herausgebildet, das sich mit kausalanalytischen Fragen zu Möglichkeit und Grenzen des Nachweises von Ursache-Wirkungszusammenhängen vertieft befasst, ohne jedoch immer expliziten Bezug zu Fragen der Evaluation herzustellen.² Die Literatur ist dabei mitunter sehr technisch, sodass sie nicht allen Leserinnen und Lesern, die an Evaluationsfragestellungen interessiert sind, leicht zugänglich ist.

Ziel dieses Buches ist es daher, diese beiden Literaturstränge zusammenzu-

-
- 1 Ohne Anspruch auf Vollständigkeit seien folgende Lehr- und Handbücher genannt: Fitzpatrick/Sanders/Worthen (2011); Flick (2006); Gollwitzer/Jäger (2009); Birbaumer et al. (2009); Patton (2008); Rossi/Freeman/Hofmann (1988); Rossi/Lipsey/Freeman (2004); Scriven (1991); Stockmann (2000); Stockmann/Meyer (2010); Stufflebeam/Shinkfield (2007); Weiss (1998); Widmer/Beywl/Fabian (2009); Widmer/De Rocchi (2012); Wottawa/Thierau (2003).
 - 2 Neuere einschlägige Texte zur Kausalanalyse sind etwa: Angrist/Pischke (2009; 2015); Best/Wolf (2015); Gangl (2010); Heckman (2005); Imbens/Rubin (2015); Morgan (2013); Morgan/Winship (2015); Murnance/Willett (2010); Pearl (2009); Pearl/Glymour/Jewell (2016); Pearl/Mackenzie (2018); Rosenbaum (2010; 2017); Vanderweele (2015).

führen, Grundlagen und neuere Entwicklungen im Bereich der Kausalanalyse gut verständlich darzustellen und entsprechende Methoden für Evaluationsfragestellungen gewinnbringend nutzbar zu machen. Während die entsprechenden kausalanalytischen Grundlagen im zweiten Teil des Buches eingeführt werden, erfolgt eine ausführliche anwendungsorientierte Illustration im dritten Teil des Buches. Wir hoffen, damit einerseits hilfreiche Hinweise und Impulse für eine evidenzbasierte und kausalanalytisch orientierte Wirkungsevaluation zu liefern. Andererseits soll anhand der Darstellung deutlich werden, dass Kausalanalysen anspruchsvolle und voraussetzungsreiche Vorhaben sind, die einer klaren theoretischen Anleitung, einer gewissenhaften Planung des Untersuchungsdesigns und einer sorgfältigen Umsetzung mit Blick auf mögliche Abweichungen vom ursprünglichen Plan bedürfen.

Zu diesem Zwecke formulieren wir eine – in der Theorie – ideale Konstellation zur Identifikation kausaler Effekte und zeigen dann auf, wie konkrete Evaluationsdesigns in der Praxis von diesem Ideal abweichen können und welche Folgen sich daraus für die Belastbarkeit von Wirkungsevaluationen ergeben. Kritische Stimmen mögen nun einwenden, dass ein solches Vorgehen weltfremd sei, da diese wissenschaftlichen Idealvorstellungen in der Praxis aufgrund von Sachzwängen und unvorhergesehenen Ereignissen niemals realisierbar sind. Wir stimmen vollkommen zu, dass sich Evaluationen stets im Spannungsfeld von Anspruch und Wirklichkeit bewegen und ein gewisses Maß an Pragmatismus bei der praktischen Durchführung unabdingbar ist. Praktische Beschränkungen, insbesondere die Verfügbarkeit von Ressourcen, die politische Durchsetzbarkeit von Untersuchungsvorhaben sowie ethische und datenschutzrechtliche Grenzen, verhindern häufig die Umsetzung idealtypischer Evaluationsdesigns.

Dies sollte jedoch aus unserer Sicht nicht daran hindern, ein kausalanalytisches Ideal zu skizzieren und der vorherrschenden Praxis kontrastierend gegenüberzustellen. Denn erstens sind viele Limitationen von Evaluationsvorhaben nicht Sachzwängen geschuldet, sondern der mangelnden methodischen Schulung von Entscheidungsträgerinnen und -trägern. Die Chance des Erkenntniszuwachses, welche sich im Zuge von Reformen ergibt, verstreicht so oft unnötigerweise mangels besseren Wissens. Zweitens treten methodische Probleme durch den Vergleich mit einem heuristischen Ideal viel deutlicher zu Tage. Möglichst strenge Kritik ist bekanntlich ganz im Sinne des Wissenschaftsverständnisses des Kritischen Rationalismus, denn (praktische) Lösungen können dann am besten entwickelt werden, wenn Probleme präzise benannt sind. Drittens ist die Herausarbeitung der Grenzen und Fallstricke vorherrschender Forschungsdesigns auch deshalb erstrebenswert, da dadurch ein sachgemäßer Umgang mit Evaluationsergebnissen gefördert wird. Wissen darüber, wie belastbar ein empirischer Befund ist und welche Schlüsse man daraus ziehen kann, ist daher nicht nur von rein akademischem, sondern auch von erheblichem praktischen Interesse. Im Sinne einer Meta-Evaluation sollte jede Evaluation daher möglichst auch

selbst einer kritischen Bewertung unterzogen werden, um Grenzen und Verbesserungspotenzial abzuleiten.

Auch dieses Buch wurde im Entstehungsprozess vielen kritischen Evaluationen durch Dritte unterzogen und hat von deren konstruktiven Beiträgen profitiert. Da wir uns seit über 15 Jahren mit dem Thema Evaluation beschäftigen, ist es nicht möglich, allen Personen an dieser Stelle zu danken, die unsere Sicht auf die Evaluationsthematik über die Jahre hinweg geprägt haben. Dennoch seien ausgewählte Personen genannt, ohne deren Unterstützung dieses Buch nicht in seiner jetzigen Form vorliegen würde. Zuvorderst sei Norman Braun posthum gedankt, der uns nicht nur ein Mentor war, sondern auch die Auseinandersetzung mit der Evaluationsthematik angestoßen und stets einen kausalanalytischen Blick auf das Thema eingefordert hat.

Viele Gespräche mit Werner Fröhlich, der als Leiter eines sozialwissenschaftlichen Forschungsinstituts über langjährige, unschätzbare Praxiserfahrung verfügt, haben unseren Blick auf Evaluationen maßgeblich geschärft. Ebenso war die Mitwirkung an über fünfzig internen Evaluationsvorhaben an Schulen der Stadt München für uns prägend wie es zahlreiche Fortbildungen und Workshops zur Evaluation waren, die wir für Lehrkräfte am Pädagogischen Institut der Stadt München sowie für Forschende, Evaluierende und Auftraggebende im Rahmen des Fortbildungsprogramms von GESIS – Leibniz-Institut für Sozialwissenschaften in Mannheim angeboten haben.

Gerhard Krug und Christof Wolf haben Input zu den ersten Materialien des GESIS-Workshops gegeben und uns damit nicht nur vor Fehlern und Inkonsistenzen in der ersten Fassung bewahrt, sondern auch viele hilfreiche Ideen und Anregungen zu diesem Buch geliefert. Die erste Fassung des Buches haben Andreas Eberl, Daniel Großmann, Ulrike Marx und Tilman Wörz gelesen und damit zu einer sowohl in sprachlicher als auch inhaltlicher Hinsicht signifikanten Qualitätssteigerung beigetragen. Für ihre tatkräftige Unterstützung bei der Erstellung des Literaturverzeichnisses ist schließlich Van Anh Nguyen zu danken.

Nürnberg, im Januar 2020

Edgar Treischl und Tobias Wolbring

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 nahelegen 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

kausalen Wirkung einer Lebensmittelkennzeichnung (sogenannte „Lebensmittelampel“).

Das Buch schließt in *Kapitel 10* mit einer Zusammenschau zentraler Einsichten und dem Hinweis, dass nicht nur soziale Interventionen, sondern auch Evaluationen selbst einer strengen Meta-Evaluation unterzogen und im Hinblick auf ihre Kosten und ihren Nutzen (selbst)kritisch hinterfragt werden sollten. Die Berücksichtigung der in diesem Buch formulierten Mindestanforderungen, methodischen Leitlinien und praktischen Empfehlungen gewährleistet hoffentlich, dass dem eigenen Evaluationsvorhaben im Zuge einer entsprechenden Bewertung ein gutes Zeugnis ausgestellt wird.