

Zusammenfassungen der Kapitel

Teil I „Was“ und „wozu“? Evaluationsforschung im Überblick

1 Was ist Evaluation?

Evaluation und Evaluationsforschung. In diesem Kapitel ging es um die Frage, was Gegenstand der Evaluation ist. Hierbei wurde zwischen den beiden Begriffen Evaluation und Evaluationsforschung unterschieden:

- Evaluation umfasst alle Arten der Bewertung von Gegenständen, inklusive einer Bewertung des Prozesses der Bewertung.
- Evaluationsforschung ist eine Form der Bewertung, die sich auf Anwendung sozialwissenschaftlicher Methoden und damit eine regelgeleitete Bewertung stützt.

Evaluationskriterien. Evaluation und Evaluationsforschung befassen sich mit der Bewertung bestimmter Gegenstände (meist handelt es sich um unterschiedliche Formen von Maßnahmen oder Interventionen) anhand bestimmter Kriterien. Zu diesen Kriterien gehören

- der Bedarf an einer Intervention,
- die Qualität des Interventionskonzepts,
- die Qualität der Einführung und Implementation der Interventionsmaßnahme,
- ihre Wirksamkeit (Effektivität) und
- ihre Kosten-Nutzen-Bilanz (Effizienz).

Historische Entwicklung und theoretische Positionen. Epochale historische Entwicklungen, allgemein-politisch begründete Phasen sowie fachpolitische Entscheidungen der Disziplin Evaluation wurden dargestellt – bis zur Gründung von Evaluationsgesellschaften. Diese chronologische Darstellung wurde durch theoretische Positionen innerhalb der Evaluationsforschung ergänzt. Hier konnte unterschieden werden zwischen methodischen und methodologischen Orientierungen, bewertungsorientierten und nutzungsorientierten Ansätzen sowie dem normativen Ansatz.

Evaluationsformen. Die verschiedenen Arten der Evaluationsforschung wurden auf der Basis von Kriterien wie Rahmenbedingungen, Evaluationsmodell, Ausrichtung der Evaluation, Zeitpunkt der Evaluation, Durchführungsmodi sowie Metaevaluation eingeteilt.

Anwendungsfelder. Am Ende wurden Evaluation und Evaluationsforschung von den Bereichen Diagnostik, Methodologie und Methodik sowie gegenüber psychologischen und nicht-psychologischen Anwendungsfeldern abgegrenzt.

2 Aufgaben, Standards und Modelle der Evaluation

Worin besteht der Unterschied zwischen einer (Alltags-)Evaluation und der Evaluationsforschung? Zielsetzungen der Evaluationsforschung umfassen im Wesentlichen Zustandsevaluationen, Veränderungsevaluationen und Wirksamkeitsevaluationen.

Gütekriterien. Welche Gütekriterien können herangezogen werden, um die Evaluationsforschung selbst unter die Lupe zu nehmen? Da sich Evaluationsforschung unter bestimmten Randbedingungen (Disziplinen, Bereichen und Fragestellungen) vollzieht, gilt die Präzisierung der Fragestellung bereits als ein erstes Gütekriterium. Hinzu kommen allgemeine Gütekriterien, die als allgemeine Rahmenbedingungen verstanden werden, und klassische Gütekriterien, wie sie aus der Psychometrie bekannt sind. Darüber hinaus spielen die beiden Kriterien interne Validität (Kausalitätsannahmen) und externe Validität (Verallgemeinerbarkeit) eine besondere Rolle.

Standards. In diesem Kontext sind auch Überlegungen zu Standards angesiedelt. Sie gelten als die jeweiligen Regeln der Kunst und sind auf der Basis von Konventionen entstanden.

Theoretische Positionen. War bereits in Kapitel 1 in einem historischen Kontext die Frage gestellt worden, welche Vorstellungen und Modelle die Entwicklung der Evaluationsforschung beeinflusst haben, so stand nunmehr die Frage an, wie Vertreter bestimmter Richtungen ihre Position präzisiert haben. Hierzu wurde je eine Position aus den methodischen und methodologischen Ansätzen, bewertungsorientierten Vorgehensweisen, nutzungsorientierten Zugängen und normativen Modellen dargestellt.

Teil II „Wann“ und „warum“? Inhaltliche Fragestellungen

3 Beschreibung und Bewertung von Zuständen

Ist-Zustände. Ist-Zustände können auf der Basis qualitativer und/oder quantitativer Informationen beschrieben werden. Anhand solcher Zustandsbeschreibungen kann beurteilt werden, ob es angezeigt ist, Maßnahmen zu ihrer Veränderung (Interventionen) einzuleiten. Ferner sollen Zustandsbeschreibungen helfen, die angemessene Art von Intervention zu ermitteln und zu planen (Bedarfsanalyse, s. Abschn. 7.1).

Erhält ein Evaluator den Auftrag, einen Ist-Zustand zu analysieren, sollte er sich zunächst mit folgenden Fragen befassen:

- Was ist das genaue Ziel, die genaue Funktion des Evaluationsvorhabens?
- Gibt es neben dem explizierten Ziel noch weitere, »verborgene« implizite Ziele bzw. Funktionen?
- Sind diese impliziten Funktionen ethisch oder fachlich gerechtfertigt?
- Bestehen Konfliktpotenziale zwischen den Beteiligtegruppen bzw. den von den Evaluationsergebnissen direkt oder indirekt Betroffenen (einschließlich der zu evaluierenden Personen)?
- Lassen sich solche Konflikte lösen?
- Welche Beteiligten haben welche Rollen und Interessen im Evaluationsprozess? Widersprechen sich diese Interessen? Ist mit Rollenkonflikten zu rechnen? Sind diese lösbar?

- Sind die Bestimmungsstücke des Evaluationsvorhabens (Ziel, Anwendungsbereich, Evaluationsobjekte, Evaluationsgegenstand bzw. -kriterien, Informationsquellen, Kontextbedingungen, Methodik, Ergebnisnutzung) hinreichend genau definiert?
- Liegen bereits verfügbare Instrumente bzw. Ansätze zur Datengewinnung (Operationalisierung des Evaluationsgegenstands bzw. -kriteriums) vor?

Stichproben. Will man eine Population über quantitative Merkmale beschreiben, die über die Personen in der Population hinweg variieren, so zieht man – falls eine Totalerhebung unmöglich oder unökonomisch ist – eine Stichprobe heran. Ist das zu erfassende Merkmal in der Population normalverteilt, kann die entsprechende Populationsverteilung vollständig über die Stichprobenkennwerte Mittelwert und Standardabweichung beschrieben werden. Voraussetzung ist, dass die Stichprobe repräsentativ für die Population ist, d. h. dass es sich um eine probabilistische Stichprobe handelt. Gelegenheitsstichproben sind keine probabilistischen Stichproben.

Prävalenz und Inzidenz. Im Falle kategorialer Variablen können Kennwerte aus der Epidemiologie zur Beschreibung von Ist-Zuständen in der Population nützlich sein:

- die Prävalenz (Punktprävalenz, Periodenprävalenz),
- die Lebenszeitprävalenz und
- die Inzidenzrate (Inzidenz).

Vergleiche. Um Ist-Zustände bewerten zu können, müssen in den meisten Fällen Vergleiche angestellt werden. Vergleichskriterien können

- normativer Natur (Idealnormen, Realnormen; Norm) oder
- empirischer Natur (sozialer und temporaler Vergleich) sein.

Solche Vergleiche können statistisch abgesichert werden: Hierzu werden inferenzstatistische Verfahren verwendet. Da die Ergebnisse inferenzstatistischer Tests jedoch sehr stark von der Größe der Stichprobe abhängen, aus der die Beobachtungen gezogen wurden, wurde zusätzlich vorgeschlagen, die Bewertung empirisch vorgefundener Ist-Zustände auf der Basis ihrer praktischen Bedeutsamkeit vorzunehmen. In diesem Zusammenhang sind Effektstärkenmaße und die von Cohen (1988) vorgeschlagene Heuristik zur Beurteilung der Größe dieser Maße wichtige Hilfsmittel. Letzten Endes sollte man vor allem anhand inhaltlicher Erwägungen beurteilen, ob ein vorgefundenes empirisches Ergebnis zur Beschreibung eines Ist-Zustandes praktisch bedeutsam ist oder nicht.

4 Beschreibung und Bewertung von Veränderungen

Veränderungsanalysen kann man dahingehend kategorisieren, ob sie

- quantitativ (die Veränderung kann in Zahlen ausgedrückt werden) oder
- qualitativ (die Veränderung ist mehrdimensional oder nicht adäquat zu quantifizieren)

orientiert sind. Die Veränderung kann auf der Ebene

- einzelner Individuen bzw. Objekte oder
- Gruppen von Individuen (inkl. ganzer Populationen) bzw. Objekten

beschrieben werden.

Intra- und interindividuelle Veränderungen. Dabei können

- echte intraindividuelle Veränderungen (Daten von den gleichen Personen zu verschiedenen Messzeitpunkten; Messwiederholung) oder
- interindividuelle Veränderungen (zu den unterschiedlichen Messzeitpunkten werden jeweils unterschiedliche Personen getestet)

ausgewertet werden. Auf Einzelfallebene kann die Veränderung zwischen zwei Messzeitpunkten entweder als einfache Differenz (d), als standardisierte Differenz (dst) oder aber in Form einer klinischen Signifikanz (dk) angegeben werden. Letztere ermöglicht es auch, das Ausmaß der Veränderung inferenzstatistisch abzusichern.

Für die inferenzstatistische Absicherung von Veränderungen auf Gruppenebene hängt die Wahl des geeigneten Verfahrens von der Art der Analyseeinheit und der Art der Veränderungsmessung ab. Dabei spielt ferner eine Rolle, ob

- eine Analyse von Art und Ausmaß des zeitlichen Unterschieds (Differenz),
- eine Analyse des zeitlichen Trends oder
- eine Analyse der Variabilität über die Zeit hinweg

im Mittelpunkt des Interesses steht.

Der Unterschied zwischen zwei Messzeitpunkten kann

- indirekt (mit Hilfe einer Vorher-Nachher-Messung),
- direkt oder
- quasi-indirekt (mit Hilfe einer retrospektiven Vorher-Messung)

ermittelt werden.

Regressionseffekt. Mit jedem dieser Ansätze sind Vor- und Nachteile verbunden. Ein Nachteil der indirekten Veränderungsmessung ist die Gefahr artifizieller Regressionseffekte: Werte, die zu einem Messzeitpunkt weit vom Mittelwert abweichen, weichen zum anderen Messzeitpunkt weniger weit vom Mittelwert ab. Der Regressionseffekt kann Veränderung vorgaukeln, die in Wirklichkeit gar nicht (oder sogar in umgekehrter Richtung!) stattgefunden hat.

Moderatorvariablen. Eine Suche nach Moderatoren der Veränderung ist nicht nur statistisch sinnvoll (denn Interaktionseffekte können Haupteffekte der Zeit überlagern!), sondern auch inhaltlich angezeigt. Für die Wahl der Moderatorvariablen gibt es – im Idealfall – konzeptuelle oder empirische Argumente.

5 Evaluation der Wirkung und Wirksamkeit einer Maßnahme

Wirksamkeit. In Bezug auf das Evaluationskriterium Wirksamkeit wurde festgestellt, dass im Einzelfall zunächst meist eine Präzisierung nötig ist. Es ist Aufgabe des Evaluators, gegebenenfalls gemeinsam mit den Auftraggebern der Untersuchung zu ermitteln,

- in welchen Phänomenbereichen sich die Wirksamkeit manifestieren sollte, d. h. welche Konstrukte gemessen werden sollen,
- in welchen Modalitäten das zu messende Phänomen vorliegt,
- welche Facetten des Konstrukts zur Messung herangezogen werden sollen,
- welche Datenquellen für das Konstrukt am ehesten informativ und geeignet sein dürften und
- welche Methoden zur Datengewinnung am ehesten angemessen erscheinen.

Gütekriterien. Wirksamkeitsindikatoren müssen bestimmte Gütekriterien erfüllen. Hierbei ist es nicht nur notwendig, dass sie ausreichend reliabel (messgenau bzw. messfehlerfrei) und valide sind (d. h. tatsächlich das messen, was gemessen werden soll), sondern zusätzlich

- eine inhaltliche Symmetrie zum Interventionsgegenstand aufweisen,
- neutral gegenüber der Art der Intervention sind,
- eine angemessene Schwierigkeit für die Zielgruppe aufweisen und
- (im Falle von indirekten Veränderungsmessungen) optimal änderungssensitiv sind.

Wirkmodelle. Oft beschränken sich Evaluationsvorhaben im Kontext von Interventionsmaßnahmen auf eine reine Analyse der Wirksamkeit der Maßnahme, d. h. darauf, ob die erwarteten Effekte tatsächlich eingetreten sind. Die – theoretisch, aber auch praktisch – weitaus interessantere Frage ist allerdings, welche Faktoren zu einer beobachteten Wirksamkeit beigetragen haben, also welche spezifischen Wirkungen mit der Maßnahme verknüpft waren.

Zu den möglichen Wirkungen von Maßnahmen gehören spezifische intendierte Wirkungen, aber auch nicht-intendierte Neben- oder Folgewirkungen. Diese können mehr oder weniger vorhersehbar sein. Vorhersehbare Neben- oder Folgewirkungen sollten im Wirkmodell spezifiziert werden, unvorhersehbare Neben- und Folgewirkungen müssen hingegen explorativ ermittelt werden.

Im Wirkmodell einer Maßnahme sollte darüber hinaus spezifiziert sein,

- warum mit einer Wirksamkeit gerechnet wird (d. h. wie die Wirksamkeit theoretisch begründet wird),
- von welchen Randbedingungen sie abhängt (Moderatorvariablen),
- wie die Wirksamkeit empirisch getestet werden kann,
- wann mit einer Wirksamkeit zu rechnen ist,
- wie lange sie anhält (Persistenz) und
- inwiefern mit Anforderungs- und Situationstransfer zu rechnen ist und wieso.

Brutto- und Nettowirkungen. Spezifische intendierte Wirkungen und Neben- bzw. Folgewirkungen werden als Nettowirkungen einer Maßnahme bezeichnet. Zusätzlich zu diesen Nettowirkungen kann die Wirksamkeit einer Intervention auch auf

- maßnahmenunspezifische Wirkungen (d. h. die Effekte sind gar nicht auf die spezifischen Wirkungen dieser Maßnahme zurückzuführen) und/oder
- externe Wirkungen (z. B. einen Placebo-Effekt)

zurückgeführt werden. Ein Evaluationsvorhaben, das diese Wirkungen nicht empirisch auseinanderhalten kann, erfasst lediglich die Bruttowirkungen einer Maßnahme. In diesem Fall ist

die interne Validität der Evaluationsuntersuchung gering. Es ist nicht mehr möglich, einen empirisch gefundenen Effekt tatsächlich kausal auf eine Wirkung der Intervention zu attribuieren. Neben der internen spielt die externe Validität des Untersuchungsdesigns eine wichtige Rolle: Ist anzunehmen, dass sich der Wirksamkeitsbefund auch in anderen Durchführungskontexten replizieren lässt? Die Frage kann umso eher mit ja beantwortet werden, je mehr empirische Replikationen in unterschiedlichen Kontexten es bereits gibt.

Rolle der Grundlagenforschung. Die Hypothesen, die sich aus einem Wirkmodell ergeben, sollten theoretisch begründet werden. Solche Begründungen stammen meistens aus der Grundlagenforschung; hier wird also die enge Verzahnung zwischen Theorie und Praxis offenbar. Dabei ist eine Wirksamkeitsuntersuchung allerdings nicht ausschließlich durch die Qualität der zugrunde liegenden Grundlagentheorie beeinflusst, auch

- die Qualität der technologischen Überführung einer Grundlagentheorie in eine Interventionshypothese,
- die Qualität der Implementation einer Intervention und
- die methodische Qualität der Evaluationsstudie beeinflussen die Wirksamkeitsbefunde.

6 Evaluation des Nutzens und der Kosten einer Maßnahme

Effizienzanalysen. Während die Effektivitätsanalyse nach der Wirksamkeit einer Maßnahme fragt, steht bei der Effizienzanalyse die Frage nach dem Nutzen und der Kosten-Nutzen-Bilanz im Vordergrund. Effizienzanalysen können zum einen durchgeführt werden, um einer Entscheidung für eine konkrete Maßnahme (und gegen eine andere Maßnahme) eine rationale Basis zu verleihen (a priori Effizienzanalysen), zum anderen dienen Effizienzanalysen dem Zweck, im Nachhinein zu beurteilen, ob sich die Durchführung der Maßnahme tatsächlich gelohnt hat (a posteriori Effizienzanalysen). Kosten, Wirkungen und Werte werden quantifiziert und auf einer gemeinsamen Einheit (z. B. Geld) abgetragen. Dieses Vorgehen hilft, begründete Entscheidungen für oder gegen Maßnahmen zu treffen und damit nicht nur volkswirtschaftlichen Schaden zu verhindern, sondern auch Versorgungssysteme zu verbessern und öffentliche Mittel zielgerichteter und Erfolg versprechender zu investieren. Auf der anderen Seite ist es ebenso falsch zu glauben, dass jede erhoffte oder tatsächlich eingetretene Wirkung einer Maßnahme in geldwerten Einheiten ausgedrückt werden kann.

Kostenarten. Bei der Kostenplanung sind unterschiedliche Kostenarten (fixe vs. variable, einmalige vs. wiederkehrende, direkte/manifeste vs. indirekte/latente) einzukalkulieren. Besonders um Unwägbarkeiten und ihre ökonomischen Konsequenzen zu antizipieren, bieten sich Expertenpanels und Planspieltechniken an.

Nutzen. Der Nutzen einer Maßnahme ist definiert als Wirksamkeit mal Wert; im Falle mehrerer unabhängiger Effekte auf unterschiedlichen Wirksamkeitsindikatoren können die mit ihrem jeweiligen Wert gewichteten Effekte aufaddiert werden (MAUT-Technik) – diese einfache Formel erinnert an Erwartungs-mal-Wert-Ansätze in der Psychologie und der ökonomischen Entscheidungstheorie.

Der Wert einer Wirkung kann dabei entweder

- in Geldeinheiten ausgedrückt, insbesondere wenn es sich um einen objektiven Wert handelt,
- oder von den Beteiligten subjektiv angegeben werden, z. B. in Form einer Ratingskala von 0 bis 100.

Der Zusammenhang zwischen Nutzen und Wirksamkeit hängt davon ab, wie viel Wert einer Wirksamkeit in unterschiedlichen Bereichen des Wirksamkeitskontinuums beigemessen wird. Evaluatoren sollten großen Wert auf Perspektivenunterschiede bei der Wertigkeitszu-messung zwischen unterschiedlichen Beteiligten(-gruppen) legen: Ist zu erwarten, dass sich der Wert einer Wirkung zwischen den Beteiligtegruppen stark unterscheidet, so ist für jede Gruppe eine eigene Nutzenabschätzung vorzunehmen.

Modelle der Effizienzschtzung. Je nachdem, ob die Wirksamkeit (bzw. der Nutzen) einer MaBnahme in der gleichen Einheit dargestellt werden kann wie ihre Kosten, werden zwei Modelle der Effizienzanalyse unterschieden:

- Kosten-Nutzen-Analysen erlauben eine direkte Verrechnung von Kosten und Nut-zen, entweder in Form eines Nettonutzens (NN), eines Nutzenquotienten (NQ) oder einer Profitrate (PR).
- Kosten-Effektivitäts-Analysen erlauben es, die Kosten einer MaBnahme bezogen auf eine sinnvolle Wirksamkeitseinheit anzugeben, z. B. Anzahl »geheilte« Patienten o-der Standardabweichungseinheiten auf einem metrischen Wirksamkeitskriterium. Die Analyse der Nutzenschwelle hilft dabei, den für die Nützlichkeit erforderlichen Mindesteffekt einer MaBnahme zu quantifizieren.

Nutzenschwelle (»Break-even-Point«). Dieser Punkt beschreibt den Fall, dass der Nutzen einer MaBnahme exakt identisch mit den Kosten ist. Die Bestimmung der Nutzenschwelle erlaubt eine Aussage darüber, wie effektiv ein Programm oder eine MaBnahme mindestens sein muss, damit sich die Einführung lohnt (damit also der Nutzen größer ist als die Kosten).

7 Prospektive Evaluation und MaBnahmenplanung

Unter prospektiver (nach vorne schauender) Evaluation verstehen wir Aktivitäten, die im Zusammenhang mit der Feststellung des Interventionsbedarfs sowie der Bewertung der Kon-zeption der geplanten MaBnahme relevant sind.

Problembestimmung. Die Feststellung des Interventionsbedarfs umfasst objektivierbare und subjektive Komponenten. Objektivierbar – und empirisch überprüfbar – sind bspw. Annah-men über Art, AusmaB und Verbreitung eines vermuteten Problems. Je spezifischer die In-formationen sind, die man in Bezug auf diese Parameter gewinnen kann, desto präziser kann die MaBnahme geplant werden, und desto größer sind sowohl ihre Erfolgchancen als auch ihre Effizienz.

Zielgruppe. Bei der Bestimmung der Zielgruppe ist zu fragen,

- aus welchen »Einheiten« die Zielgruppe besteht,

- ob die Maßnahme eher universell oder zielgerichtet (d. h. indiziert oder selektiv) geartet ist und
- welches die Kriterien für die Inklusion von Personen bzw. »Einheiten« in die Zielgruppe sind.

Bei einer Primärprävention (selektive Prävention) ist zu beachten, dass die Inklusion nicht anhand des eigentlichen Problemverhaltens, sondern anhand theoretisch oder empirisch begründeter Risikobedingungen vorgenommen wird.

Over- und Underinclusion. Bei der Zuordnung von Personen aus der Population zur Zielgruppe sind zwei Fehler möglich, die als Overinclusion und Underinclusion bezeichnet werden. Die Fehlerwahrscheinlichkeit richtet sich danach,

- wie eng oder weit das Inklusionskriterium definiert ist und
- wie sensitiv bzw. »schwierig« (im psychometrischen Sinne) das Messinstrument ist, anhand dessen die Zuordnung vorgenommen wird.

Zielhierarchien. In vielen Fällen bietet es sich an, die Ziele einer Maßnahme in Form von Zielhierarchien darzustellen und auf diese Weise zu veranschaulichen,

- welche Arbeitsschritte in welcher Abfolge zur Zielerreichung nötig sind,
- welche Bedingungen für die Zielerreichung notwendig oder hinreichend sind und
- wie ggf. abstrakt gefasste Oberziele konkretisiert (und damit empirisch messbar gemacht) werden können.

Konzeptionsanalyse. Der zweite große Baustein einer prospektiven Evaluation – neben der Bedarfsprüfung – besteht in der Analyse und Bewertung der Maßnahmenkonzeption. Dabei ist zunächst nach strukturellen Eigenschaften der Maßnahme zu fragen: Wie sollte eine Maßnahme gestaltet sein, um dem Problemzustand, dem angestrebten Soll-Zustand bzw. der intendierten Art der Veränderung möglichst gut gerecht zu werden?

Experten. Evaluatoren dürfen sich nicht davor scheuen, das Wirkmodell einer geplanten Maßnahme unter die Lupe zu nehmen und die darin implizit enthaltenen Annahmen zum Gegenstand einer Diskussion unter den Beteiligten zu machen. Strukturierte Expertengremien können bei dieser schwierigen Aufgabe helfen und den Evaluator entlasten.

Rahmenbedingungen. Schließlich ist zu fragen, ob für die Durchführung einer bestimmten Maßnahme überhaupt die Rahmenbedingungen gegeben sind. Solche Rahmenbedingungen schließen ein:

- notwendige Voraussetzungen auf Seiten der Maßnahmeempfänger,
- notwendige Voraussetzungen auf Seiten der Durchführenden,
- notwendige Voraussetzungen in Bezug auf den Kontext (Personal, Räume, Zeit), die rechtlichen und die finanziellen Bedingungen, aber auch
- Akzeptanz aller Beteiligengruppen und die Motivation zur Umsetzung.

8 Formative Evaluation: Programmoptimierung und Implementationskontrolle

Während eine summative Evaluation darauf angelegt ist, die Wirksamkeit einer Maßnahme abschließend zu bewerten bzw. im Sinne einer prospektiven Wirksamkeitsabschätzung zu

antizipieren, geht es bei der formativen Evaluation um eine Optimierung der Programmdurchführung bzw. eine Anpassung an die gegebenen Durchführungsbedingungen.

Probedurchlauf. Um zu überprüfen, ob eine erfolgreiche Umsetzung der Maßnahme unter den gegebenen Praxisbedingungen überhaupt möglich ist, können Probelaufe mit kleinen Stichproben bzw. mit ausgewählten Programmeinheiten durchgeführt werden. Stellen sich bei diesen Durchläufen Probleme mit den Implementationsbedingungen (z. B. fehlende finanzielle Ressourcen, geringe Akzeptanz bei den Teilnehmern, suboptimales Material, nicht ausreichend geschulte Trainer) heraus, so müssen diese verändert werden. Stellen sich hingegen Probleme mit dem Programm selbst heraus – etwa weil bestimmte Annahmen des Wirkmodells nicht erfüllt sind (z. B. ein Spinning-Kurs im Fitnessclub, der für Anfänger viel zu schwer konzipiert ist) – muss die Programmkonzeption gegebenenfalls verändert werden. Diese Form der formativen Evaluation ist innerhalb der Phase der prospektiven Evaluation angesiedelt.

Prozess- bzw. Zwischenevaluation. Eine zweite Möglichkeit, formative Evaluation zu betreiben, besteht darin, regelmäßige stichprobenartige Zwischenevaluationen durchzuführen, bei denen Parameter der Wirksamkeit erfasst werden. Anhand dieser Zwischenevaluationen können Durchführungsprobleme erkannt und beseitigt und Programmstrukturen gegebenenfalls verändert werden. Man spricht hier auch von Prozessevaluation.

Programm-Monitoring. Mit der formativen Evaluation verwandt, aber nicht identisch, ist die Idee eines fortlaufenden Programm-Monitoring. Dabei wird überprüft, ob die Struktur einer Maßnahme, die ihr zugrunde liegenden Interventionsprinzipien, bestimmte notwendige Wirksamkeitsvoraussetzungen (z. B. Akzeptanz) usw. auch tatsächlich gegeben und nachhaltig gewährleistet sind. Im Rahmen sog. Management Information Systeme (MIS) können solche objektiven Daten zur Implementationsqualität dazu verwendet werden, den Programmverantwortlichen kontinuierlich Informationen über den Fortgang der Maßnahme zu liefern.

Prüfung der Ausführungsintegrität. Bei hoch-standardisierten Programmen kann es essenziell sein, ein Programm genau so durchzuführen wie es z. B. in einem Manual vorgesehen ist. Die Ausführungsintegrität ist dabei nicht nur eine Maßnahme zur Sicherung der internen Validität eines Evaluationsdesigns; sie ist eventuell sogar eine zentrale Bedingung für die Wirksamkeit einer Maßnahme.

Methoden. Zu den Methoden der formativen Evaluation und der Implementationskontrolle gehören weniger die quantitativen als eher die qualitativen Formen der Datengewinnung, also bspw. Gespräche, Interviews, Beobachtungen, Aktenauswertungen, Checklisten. Entsprechend muss auch die Datenauswertung vollzogen werden. Insbesondere die kommunikativen Methoden formativer Evaluation (z. B. Expertenrunden) sollten in Form eines Kommunikationsplans vorab eingeplant werden.

Teil III „Wie“ und „womit“? Methodische Aspekte der Evaluationsforschung

9 Designfragen: Planung von Evaluationsuntersuchungen

Evaluationsuntersuchungen zur Prüfung der Wirksamkeit einer Maßnahme sollen zum einen die Frage beantworten helfen, ob eine Veränderung auf der AV tatsächlich eingetreten ist. Zum anderen sollen sie nachweisen, dass ein beobachtbarer Effekt in der Tat kausal auf die Maßnahme zurückzuführen ist. Voraussetzung für die kausale Interpretierbarkeit eines Effekts ist, dass es keine plausiblen Alternativerklärungen für das Zustandekommen des Effekts gibt. Ein Untersuchungsdesign, das in der Lage ist, unterschiedliche Alternativerklärungen gezielt zu prüfen oder von vornherein auszuschließen, wird als intern valide bezeichnet. Interne Validität ist von daher eines der wichtigsten Gütekriterien für ein Untersuchungsdesign, nicht nur in der experimentellen Grundlagenforschung, sondern auch in der angewandten Evaluationsforschung.

Konfundierung. Bei der Prüfung oder Ausschaltung von Alternativerklärungen geht es vor allem um den potenziell systematischen Einfluss von Störvariablen. Störvariablen sind mit der AV korreliert; sind sie zusätzlich auch mit der UV korreliert, spricht man von Konfundierung. Eine gleichsinnige Konfundierung führt zu artifiziellen Bestätigungen der empirischen Hypothese. Eine gegensinnige Konfundierung führt zu artifiziellen Widerlegungen der Hypothese. Beide machen kausale Interpretationen über die Wirkung der UV unmöglich.

Kontrolle von Messwiederholungsartefakten. Wird der empirische Nachweis der Wirksamkeit einer Maßnahme über die durchschnittliche Veränderung auf den Wirksamkeitskriterien nach der Intervention im Vergleich zu vorher ermittelt (Messwiederholung), so lautet eine typische Alternativerklärung: »Die Veränderung wäre auch ohne die Intervention aufgetreten«. Dann hätte die Veränderung gar nichts mit der Maßnahme als solcher zu tun, sondern ist vielmehr auf den systematischen Einfluss von Störvariablen zurückzuführen. Zu diesen können die in Abschnitt 4.2 angesprochenen Messwiederholungsartefakte (Erinnerung, Testübung, Sensibilisierung für die Fragestellung, Response Shifts, Reifung, Regressionseffekt) gehören, aber auch spezielle Störeinflüsse wie selektiver Drop-out/experimentelle Mortalität. Solche Messwiederholungsartefakte können durch die Realisierung einer Kontrollbedingung kontrolliert werden, die der Interventionsbedingung hinsichtlich aller relevanten Variablen ähnlich ist und sich von dieser nur insofern unterscheidet, als hier die Intervention nicht (oder erst später) durchgeführt wird (unbehandelte Kontrollgruppe).

Kontrolle von maßnahmenunspezifischen Effekten. Eine zweite Klasse von Alternativerklärungen besagt, dass der beobachtete Effekt gar nicht auf die Spezifika der durchgeführten Maßnahme zurückzuführen ist, sondern vielmehr auf unspezifische Wirkungen, z. B. auf die Tatsache, dass überhaupt irgendeine Intervention stattgefunden hat. Solche Alternativerklärungen können nur mit Hilfe einer unspezifisch behandelten Kontrollgruppe getestet werden; nur solche Kontrollgruppen sind geeignet, die Nettowirkungen einer Maßnahme empirisch zu isolieren.

Kontrolle von Konfundierungen. Prinzipiell ist die Menge aller Störvariablen unbegrenzt; daher wird es in der Praxis auch nicht möglich sein, sie alle auszuschalten (wie es bspw. in der strengen Kausalitätsbedingung gefordert ist). Perfekte interne Validität wird demnach bei keinem einzigen (sozialwissenschaftlichen) Untersuchungsdesign gegeben sein. Dennoch ist es wichtig, die interne Validität des Designs so gut wie möglich zu sichern. Typische Störvariablen, die im Kontext von Wirksamkeitsevaluationen unter Umständen mit der UV (Bedingung) korreliert sind, sind

- (1) Ausgangswerte,
- (2) persongebundene Störvariablen, die die Wirksamkeit beeinflussen,
- (3) Reifung bzw. Spontanerholung,
- (4) maßnahmenunspezifische Wirkungen,
- (5) zwischenzeitliches Geschehen,
- (6) selektiver Drop-out,
- (7) alternative Interventionen und Spillover-Effekte,
- (8) mangelnde Konstruktvalidität und Eignung des Messinstruments,
- (9) Unreliabilität des Messinstruments,
- (10) externe Wirkungen,
- (11) Sequenzeffekte bei messwiederholten Designs.

Simpson-Paradox. Eine besondere Form der Konfundierung stellt das sog. Simpson-Paradox dar: Hier fällt ein Effekt der UV auf die AV zwar in allen Teilstichproben (z. B. Männer und Frauen) homogen aus, aber sobald man über die Teilstichproben hinweg aggregiert, kehrt sich dieser Effekt um.

Designs ohne Kontrollgruppe. Anders als in kontrollierten experimentellen Kontexten ist es im Rahmen von Evaluationsuntersuchungen oft nicht möglich, die Zuweisung von Untersuchungseinheiten zu Bedingungen vollständig dem Zufall zu überlassen – in diesem Fall ist das Design lediglich quasi-experimentell. Manchmal ist es sogar nicht einmal möglich (oder ethisch vertretbar), überhaupt eine Kontrollgruppe zu realisieren. In solchen Fällen kann man auf Untersuchungsdesigns ohne Kontrollgruppen zurückgreifen, z. B.

- Designs mit mehreren Messzeitpunkten,
- Designs mit mehreren nicht-äquivalenten abhängigen Variablen," Designs mit Kohorten-Kontrollgruppen sowie
- Kohorten-Designs mit Parallelisierung.

10 Auswertungsfragen: Deskriptiv- und inferenzstatistische Methoden

Ausreißer und Extremwerte. Eine Möglichkeit der Ausreißerdiagnose stellen so genannte grafische Tests dar, bspw. das Box-Whisker-Diagramm. Ist das Merkmal annähernd normalverteilt, kann man zur Diagnose von Ausreißer- und Extremwerten Standardwerte oder Distanzwerte (wie etwa die Mahalanobis-Distanz) verwenden. Generell gilt: Ausreißer sollten nur dann aus der Analyse eliminiert werden, wenn es sich nachweislich um »falsche« Werte han-

delt, die etwa durch falsch verstandene Instruktionen, fehlerhaftes Antworten oder Fehler bei der Dateneingabe zustande gekommen sind.

Verteilungsform. Inferenzstatistische Tests, die davon ausgehen, dass das gemessene Merkmal in der Population normalverteilt ist (sog. parametrische Tests), können – zumindest bei kleinen Stichproben – zu fehlerhaften Ergebnissen führen, wenn die Rohwerte stark von einer Normalverteilung abweichen. Mit dem Kolmogorov-Smirnov- oder dem Shapiro-Wilk-Test kann die Normalverteiltheit der Rohdaten inferenzstatistisch geprüft werden. Ist die Verteilung eingipflig, aber schief (bzw. steil), so helfen gegebenenfalls Transformationen der Rohwerte, um eine Normalverteilung zu approximieren.

Fehlende Werte. Unproblematisch sind fehlende Werte nur dann, wenn die Stichprobe groß ist und die fehlenden Werte zufällig verteilt sind. Wie mit fehlenden Personen bzw. mit fehlenden Datenpunkten auf einzelnen Variablen umzugehen ist, sollte im Einzelfall entschieden werden. Ggf. sollte man auf Imputationsmethoden zurückgreifen. Von einfachen Ersetzungsmethoden (Ersetzung durch den Mittelwert, »last observation carried forward«) ist abzuraten.

Inferenzstatistische Verfahren. Die Wahl eines geeigneten inferenzstatistischen Verfahrens hängt ab von

- dem Populationsparameter, über den eine Annahme gemacht wird,
- der Hypothese in Bezug auf diesen Parameter (Abweichung von einem fixen Wert, Unterschied zwischen Bedingungen/Gruppen, intraindividuelle Veränderung),
- den Anforderungen an die Verteilung des in Frage stehenden Merkmals in der Population (parametrische oder nicht-parametrische Tests),
- dem Skalenniveau des Merkmals,
- der Anzahl der unabhängigen Variablen (»Faktoren«) und der Anzahl der Ausprägungen auf diesen Faktoren (»Faktorstufen«),
- der Gültigkeit von Annahmen und Voraussetzungen (z. B. Varianzhomogenität).

Insbesondere die Voraussetzung, dass Messwerte (bzw. etwaige Fehler und Residualeinflüsse) zwischen den getesteten Untersuchungseinheiten voneinander unabhängig sind, ist zentral: Eine Verletzung dieser Bedingung führt unweigerlich zu einer Erhöhung der statistischen Irrtumswahrscheinlichkeiten!

Im Falle messwiederholter Analysen ist eine weitere Voraussetzung, dass die Matrix der Varianzen und Kovarianzen zwischen den Messzeitpunktpaaren zirkulär ist (Sphärizitätsannahme). Ist diese Bedingung nicht erfüllt, können die Freiheitsgrade des entsprechenden Tests mit einem Korrekturfaktor (z. B. dem Greenhouse-Geisser-Epsilon) gewichtet werden. Alternativ sollte auf andere Verfahren (z. B. Mehrebenenmodelle) ausgewichen werden.

Teststärke. Die Teststärke ist definiert als die Wahrscheinlichkeit, mit der ein Effekt (bei gegebenem α -Niveau und gegebener Stichprobengröße) signifikant wird, wenn er in der zuvor spezifizierten Größe tatsächlich existiert. Die Teststärke kann über

- eine Erhöhung des α -Fehlerniveaus,
- eine Erhöhung der Stichprobengröße (n) sowie

- Maßnahmen, die zur Steigerung der Reliabilität der Messung beitragen, erhöht werden.

Fehlerkumulierung. Ein Problem bei der multiplen Testung der gleichen inhaltlichen Hypothese über mehrere Tests ist die Kumulierung der Fehlerwahrscheinlichkeiten α und β (Fehlerkumulierung). Dieser kann mit

- Kontrastanalysen,
- multivariaten Analyseverfahren oder
- einer Reduzierung der Fehlerwahrscheinlichkeiten für die einzelnen Tests

begegnet werden.

Auspartialisierung. Zur Erhöhung der internen Validität eines Designs, aber auch zur Erhöhung der Reliabilität der AV, können Störvariablen auspartialisiert werden. Störvariablen können dabei entweder vorab erfasste weitere Variablen sein (Kovarianzanalysen); bei messwiederholten Analysen kann auch der Pre-Test wie eine Störvariable behandelt und auspartialisiert werden.

Dadurch wird die Messung um all jene Effekte bereinigt, die mit Unterschieden im Pre-Test in Verbindung stehen. Man spricht dann von einem Autoregressor-Modell. Werte späterer Messzeitpunkte, welche um ihre jeweiligen Pre-Testwerte bereinigt wurden, nennt man auto-residualisierte Werte. Hierzu muss sichergestellt sein, dass (a) es keine Interaktion zwischen den Pre-Testwerten und der UV gibt und dass (b) die Pre-Testwerte möglichst messfehlerfrei gemessen wurden, da es ansonsten zu artifiziellen Effekten (Lords Paradox) kommen kann.