

Methoden der Evaluationsforschung

Heinrich Wottawa

Der Begriff *Evaluation* wird in der Literatur in sehr vielfältiger Weise verwendet (vgl. die Übersichten in Rossi & Freeman, 1985; Rutman, 1977, und Wittmann, 1985). Für die methodischen Aspekte von Evaluationsstudien ist entscheidend, daß ein typisches Evaluationsprojekt durch folgende Aspekte charakterisiert wird:

- Das Ziel solcher Arbeiten ist nicht allgemeiner Erkenntnisgewinn, sondern eine konkrete Aussage über einen zu bewertenden Sachverhalt.
- Das Evaluationsvorhaben ist in Entscheidungsprozesse, deren Ausgang von den Ergebnissen abhängt, eingebettet.
- Da rationale Entscheidungen immer der Nutzenoptimierung dienen, müssen die in den Evaluationsstudien erfaßten abhängigen Variablen einen entscheidungsrelevanten Nutzenbezug haben.

Gemäß diesen Aspekten liegt von seiten der Zielsetzung und der Rahmenbedingungen ein erheblicher Unterschied zur Grundlagenforschung vor.

1 Methodische Besonderheiten von Evaluationsstudien

Gerade der Praxisbezug von Evaluationsstudien und die damit verbundene unmittelbare Verantwortung würde eine besonders aussagekräftige Methodik nahelegen. Leider zwingen die faktischen Bedingungen vieler Evaluationsvorhaben zu erheblichen Abstrichen gegenüber idealen oder auch nur zweckmäßigen Vorgehensweisen.

1.1 *Interessensbedingte Verzerrungstendenzen*

Die Einbettung in den unmittelbaren Entscheidungsbezug hat zur Folge, daß mit einem erheblichen Druck im Hinblick auf ein bestimmtes Ergebnis durch „Mächtige“ zu rechnen ist; natürlich möchte der für den Personalbereich verantwortliche Manager einen positiven Nachweis des Effektes der von ihm zu verantwortenden Personalentwicklungsarbeit, der Therapeut möchte die Überlegenheit „seiner“ Orientierung bestätigt haben, und von ihrer Position überzeugte Schulpolitiker wollen „ihre“ Vorschläge für das Bildungswesen als die besten bestätigt sehen.

Dieser „Ergebnisdruck“ kann direkt nur gegenüber abhängigen Evaluatoren ausgeübt werden. Bei unabhängigen Untersuchern (siehe dazu auch die Forderungen nach objektiver Qualitätskontrolle durch bereichsunabhängiges Personal entsprechend der DIN-ISO-Norm 9004 zur Qualitätssicherung; vgl. Internationale Organisation für Normung, 1992) wird eher indirekter Einfluß ausgeübt: Einschränkung

der Möglichkeiten zur Herstellung technischer oder organisatorischer Alternativen, begrenzte Zugänglichkeit zu bestimmten Teilgebieten des Untersuchungsfeldes, eine „sorgfältige“ Auswahl der im jeweiligen Praxisfeld verfügbaren Gesprächspartner – all dies (und vieles mehr) kann gezielt zu einer Verzerrung der gefundenen Ergebnisse eingesetzt werden.

1.2 Singularität von Evaluationsstudien

Die angesprochenen Verzerrungen sind besonders unangenehm, weil das in der Grundlagenforschung übliche Korrektiv gegenüber dem Einfluß bestimmter Schulmeinungen oder einzelner „Autoritäten“, nämlich die durch andere Untersucher durchgeführten Wiederholungs- und Kontrollstudien, für die meisten Evaluationsfragestellungen nicht besteht; dies gilt sowohl für die gesellschaftlich relevanten, „großen“ Studien (z.B. Schulsystemvergleiche, Auswirkung der 0,8 Promille-Grenze auf das Trink- und Fahrverhalten in den neuen Bundesländern) als auch für „kleine“ Arbeiten (so läßt sich z.B. die Bewertung des Verhaltens eines bestimmten Trainers während einer Bildungsmaßnahme für ein bestimmtes Unternehmen nur einmal untersuchen, evtl. andere Ergebnisse in Folgestudien können durch Lernen etc. erklärt werden).

Diese *Singularität* von vielen Evaluationsprojekten ließe sich zum Teil einschränken, wenn man die gleiche Studie durch mehrere Untersucher mit unabhängigen Projektplänen gleichzeitig vornehmen ließe. Dem stehen aber zum einen die damit verbundenen Kosten entgegen, zum anderen würden sich möglicherweise die gleichen „allgemeinen Zwänge“ auf alle parallel arbeitenden Untersucher auswirken. Hinzu käme noch die Erschwernis der Praxissituation durch eine zusätzliche Beeinträchtigung der Arbeit als Folge mehrerer parallel durchgeführter Studien und den damit verbundenen Datenerhebungen.

1.3 Grenzen „sauberer“ Methodik

Ein aus der Perspektive der Grundlagenforschung naheliegender Ausweg wäre, in Evaluationsstudien methodisch ganz besonders sorgfältig vorzugehen. An sich wäre es hier wegen der Nicht-Wiederholbarkeit und der zielgerichteten äußeren Einflüsse notwendig, die Standards des Designs, der Stichprobenerhebung und der Messung im Bereich der Grundlagenforschung möglichst zu übertreffen. Diesem berechtigten Wunsch stehen aber folgende Fakten gegenüber:

- Untersuchungsdesigns in der „realen“ Welt sind meist wesentlich „schwächer“ als die Ansätze in der experimentellen Labor-Forschung (vgl. Abschnitt 3.1).
- Das Bilden von echten Zufallsstichproben oder äquivalenten Vergleichsgruppen ist nur selten möglich (Wer schickt sein Kind aus „experimentellen“ Gründen auf eine bestimmte Schulform? Welcher Mitarbeiter bleibt in einer für ihn nicht angenehmen Arbeitsgruppe, nur weil es das Untersuchungsdesign vorsieht?).
- Die intentionalen Verfälschungstendenzen sind unvergleichlich stärker als im Grundlagenbereich; in die Untersuchung aufgenommene Menschen (Schüler, Lehrer, Führungskräfte, Politiker ...) haben eine klare Vorstellung davon, mit

welchen Aussagen sie das Zustandekommen des von ihnen für wünschenswert gehaltenen Evaluationsergebnisses wahrscheinlicher machen können.

Aus diesem Grund ist die Methodik der Evaluationsstudien (vielleicht abgesehen von der Selbstevaluation, wobei man nur die Informationen zur Verbesserung des eigenen beruflichen Verhaltens erhebt bzw. erheben läßt, ohne daß andere die Ergebnisse erfahren) stets eine Gratwanderung zwischen dem wünschenswerten, dem prinzipiell machbaren, dem aufgrund von Ressourcenbegrenzungen tatsächlich möglichen und dem in Anbetracht der verschiedenen „Interessenseinflüsse“ gewünschten Vorgehen. Man kann daher jede unter normalen Feldbedingungen durchgeführte Evaluation im nachhinein methodisch kritisieren, ein Faktum, von dem in der Praxis auch gerne „zielgerichtet“ Gebrauch gemacht wird.

2 Evaluation durch Zusammenfassung von (Experten-) Meinungen

Die erfahrungsgestützte Bewertung von Maßnahmen im Zusammenhang mit Entscheidungsprozessen kann in vielen Fällen ohne neue empirische Studien durch Aufbereitung des vorhandenen Materials erfolgen. Grundlage dafür können allgemeine wissenschaftliche Ergebnisse, metaanalytische Aufbereitungen (vgl. dazu Abschnitt 5) oder auch das systematische Sammeln der Wertungen von „Betroffenen“ sein. Hierzu wurden spezielle Techniken entwickelt.

2.1 Die Delphi-Methode

Die offene mündliche Diskussion unter Experten ist immer schwierig. Es gibt „Meinungsführer“, „paarweise Antagonismen“ und „Beharrungstendenzen“ in der Art, eine einmal geäußerte Meinung nicht mehr vor der Gruppe zurücknehmen zu wollen. Aus diesem Grund wurde von Linstone und Turoff (1975) die Delphi-Methode entwickelt. Ein ausgewählter Kreis von Experten wird zu dem jeweiligen Problembe- reich schriftlich befragt, die Ergebnisse durch ein Leitungsgremium zusammengefaßt und ausgewertet. Die vorhandenen Ergebnisse werden an die einzelnen Experten rückgemeldet, ein neuer Fragenkatalog entworfen und damit wieder ihre Meinung erhoben. Dieser Kreislauf von Informationssammlung, Auswertung, Rückmeldung und neuer Informationssammlung kann so lange fortgeführt werden (üblich sind drei bis fünf Wiederholungen), bis sich ein (annähernder) Konsens abzeichnet. Das Verfahren kann wesentlich beschleunigt werden, wenn für die Arbeit des Leitungsteams entsprechende EDV-Unterstützung in Form geeigneter Auswertungsprogramme vorliegt, auch eine direkte Vernetzung der Experten mit dem zentralen Rechner kann in speziellen Situationen sinnvoll sein. Die derzeit noch relativ seltene Nutzung dieser Methode für Evaluationsvorhaben könnte darauf zurückzuführen sein, daß sie von seiten der erforderlichen Ressourcen her relativ aufwendig ist.

2.2 Bewertung durch Betroffene: Die Planungszelle

Für eine möglichst fundierte *Wissenssammlung* sind naturgemäß Experten die ideale Informationsquelle; geht es aber um die nutzenbezogene *Bewertung* von Maßnahmen,

wäre bei theoretischer Betrachtung die ideale Informationsbasis die „informierte Meinung“ der unmittelbar Betroffenen. Darauf basiert die Idee der „Planungszelle“ (Dienel, 1978), die auch berücksichtigt, daß es keineswegs eine homogene Meinung unter den Betroffenen geben muß. Die Kernelemente sind:

- echte Zufallsauswahl der Teilnehmer aus einer wohldefinierten Menge potentieller Betroffener, die für die Dauer des Entscheidungsfindungsprozesses (bis zu drei Wochen) voll einsatzfähig sind;
- Zusammenfassung der Teilnehmer in mehrere Kleingruppen für die Diskussion/Entscheidungsfindung;
- ausführliche, möglichst alle Perspektiven beleuchtende sachliche Information der Teilnehmer an der Planungszelle durch Experten;
- Moderation und Prozeßbegleitung durch geschultes Personal.

Gelingt es, die hier genannten Bedingungen tatsächlich zu realisieren, ist das methodisch schwierige Problem einer zusammenfassenden Nutzenverrechnung bei komplexen Entscheidungsalternativen (vgl. dazu auch Abschnitt 3.3) in der Form gelöst, daß explizite Verrechnungsmodelle durch die moderierte Diskussion der Teilnehmer ersetzt werden. Es gibt inzwischen eine größere Anzahl von erfolgreichen Anwendungsfällen der Planungszelle, allerdings stößt auch diese Technik auf ähnliche Schwierigkeiten wie das Delphi-Verfahren. Der Aufwand ist sehr groß, vor allem dann, wenn einer größeren Zahl von Betroffenen (etwa 200) die Teilnahme ermöglicht werden soll, und die verantwortlichen Entscheidungsträger stehen natürlich vor dem Problem, faktisch ihre Entscheidungsmacht an ein anderes Gremium abzutreten.

2.3 Die Szenario-Technik

Der relative Nutzen verschiedener Entscheidungsalternativen hängt meist von den sich wandelnden Rahmenbedingungen ab. Gestaltet man z.B. eine Umschulungsmaßnahme „Technischer Zeichner“, so war bis vor einigen Jahren die „beste“ Variante ein ausführliches Trainieren im Umgang mit Zeichengeräten, währenddem heute – als Folge der CAD-Programme – die Vermittlung solcher Kenntnisse nur einen geringen, die von EDV-Benutzung einen sehr hohen Stellenwert haben. Die entscheidungsbezogene Evaluation von Alternativen erfordert daher eigentlich eine fundierte Kenntnis der Bedingungen zu der Zeit, für die die ausgewählte Variante wirksam wird.

Eine methodisch elaborierte Möglichkeit, die Vorstellung über die Zukunft zu fundieren, ist die sogenannte Szenario-Technik. Die wichtigsten Grundaspekte dieses Ansatzes sind:

- eine sorgfältige Analyse der gegenwärtigen Situation unter Einbeziehung der (von Experten) als wichtig angesehenen Aspekte;
- möglichst empiriegestützte Analyse der Zusammenhänge zwischen den einzelnen Aspekten;
- Plausibilitätsabschätzungen für die Entwicklung der wichtigsten Einflußfaktoren;

- Reduktion der komplexen Zukunftsprognosen auf einige wenige, in sich jeweils konsistente Zukunftsbilder („Szenarien“).

Die Bewertung der einzelnen Alternativen erfolgt in Abhängigkeit von den für besonders plausibel gehaltenen Szenarien. Geht man etwa von der folgenden möglichen Zukunft aus:

„Zusammenbruch der demokratischen Gesellschaft als Folge der hohen Arbeitslosigkeit, massiver wirtschaftlicher Druck auf das Überleben des Unternehmens, dadurch bedingt die Herstellung einer Führungskultur in Anlehnung an die Leitbilder der Gründerzeit“,

dann würde sich die sachgerechte Potentialentwicklung von Führungskräften vielleicht auf die Förderung von autoritären, die kritiklose Einpassung in hierarchische Befehlsstrukturen erleichternde Persönlichkeitseigenschaften konzentrieren. Würde man dagegen ein Szenario

„Erheblicher wirtschaftlicher Aufschwung, dadurch bedingt massiver Arbeitskräftemangel, Zunahme der freizeitorientierten Schonhaltung der Mitarbeiter“

für plausibel halten, wären für die Zukunft Führungskräfte gefragt, die in besonderem Maße eine an den Interessen der Mitarbeiter orientierte Arbeitsgestaltung sowie die Fähigkeit zur intrinsischen Motivierung ihrer Mitarbeiter aufweisen.

Eine fundierte Szenario-Analyse stellt einen erheblichen Aufwand dar und erfordert die Unterstützung durch spezielle EDV-Programme. Das wichtigste Anwendungsgebiet dürfte die zukunftsorientierte Maßnahmenplanung in der Wirtschaft sein. Eine nähere Darstellung gibt etwa von Reibnitz (1983).

3 Empirische Evaluationsstudien mit dem Ziel einer retrospektiven Bewertung

Die meisten Evaluationsstudien haben das Ziel, prospektiv die Auswirkung von Maßnahmen für die Zukunft abzuschätzen. Es gibt aber auch Fälle, in denen rückblickend eine Bewertung durchgeführter Maßnahmen erfolgen soll. Die für solche Studien besonders relevanten Methodenfragen, insbesondere die Nutzenmessung und Nutzenverrechnung, werden als Grundlage in gleicher Weise für prospektive Studien benötigt.

3.1 Designerfordernis retrospektiver Evaluationen

Die technischen Aspekte solcher Studien sind wesentlich leichter zu lösen als bei einem zukunftsorientierten Vorgehen. Da keine Verallgemeinerung der Ergebnisse über verschiedene Formen der zu evaluierenden Maßnahmen angestrebt wird, können die

einzelnen Varianten als feste Größen in das Design eingehen. Sie müssen nicht ihrerseits als eine Stichprobe aus einem Universum möglicher Realisierungen aufgefaßt werden. In gleicher Weise ist die Anforderung an die Stichprobenziehung eingeschränkt: Die „Population“ ist durch die Menge der Personen definiert, die der jeweiligen Maßnahme ausgesetzt wurden; nur aus dieser Population muß, soweit keine Totalerhebung möglich ist, eine Stichprobe mit allen damit verbundenen Problemen der Repräsentativität (vgl. dazu Schäffer, in diesem Band) gezogen werden.

Für das Versuchsdesign stellt sich das Problem, daß in jedem Fall ein *Vergleich* möglich sein muß; dieser kann in folgender Weise erfolgen:

- Der Vergleich zwischen einer (oder mehreren) *Versuchsgruppe(n)* und *Kontrollgruppe(n)*; in diesem Fall sollten die Kriterien von zumindest quasi-experimentellen Designs erreicht werden (vgl. dazu Bierhoff und Rudinger, in diesem Band). Ein typisches Beispiel wäre etwa die Untersuchung der Frage, welcher der für vergleichbare Trainingsmaßnahmen eingesetzte Trainer die besten Ergebnisse erbracht hat, um z.B. eine individuelle Prämie davon abhängig zu machen.
- Vergleich mit einem von *außen gesetzten Standard*; ein Beispiel dafür ist die Bewertung einer EDV-Ausbildung danach, ob das Ziel „sechs Wochen nach dem Training verwenden mindestens 90% der Teilnehmer mindestens 80% der in diesem Programm möglichen Optionen“ tatsächlich erreicht wurde und damit eine vertraglich zugesicherte Qualitätseigenschaft des Trainings erfüllt ist.
- Vergleich mit den *subjektiven Erwartungen* der Teilnehmer; dies kann z.B. durch eine explizite Gegenüberstellung der ursprünglichen Forderungen an ein Training mit den Ergebnissen erfolgen (in diesem Fall ersetzen die Teilnehmer-Erwartungen Idealstandard) oder durch direkte Befragung (z.B. „Das Training hat meine Erwartungen voll erfüllt / gerade eben erfüllt / blieb weit hinter meinen Erwartungen zurück“).

Die Designanforderungen sind im ersten Fall am schärfsten, bei den beiden anderen Versionen deutlich schwächer.

3.2 Nutzenmessung

Das Feststellen objektiver Sachverhalte durch Meßverfahren ist in der Psychologie gut etabliert; man kann mit bewährten Techniken nicht nur die Ausprägung von Variablen im Sinne von Test-Dimensionen erfassen, sondern auch qualitative Aspekte wie z.B.

- subjektive Strukturen über Zusammenhangsbeziehungen (z.B. die Heidelberger Strukturlegetechnik, Groeben & Scheele, 1984),
- subjektive „Räume“ auf der Basis multidimensionaler Skalierung (etwa bei Borg, 1981),
- kognitive Strukturen (Erlbaum, Vosmadon & Otony, 1989; Gentner & Stevens, 1983; Kintsch & van Dijk, 1978; Spada & Reimanns, 1988),

- handlungsleitende Entscheidungsstrukturen (für ein Evaluationsbeispiel vgl. etwa Wottawa & Hof, 1987).

Für die Evaluation muß zusätzlich eine Nutzenbewertung erfolgen. Die Ausprägung der objektiv meßbaren Variablen, z.B. die Reduzierung von Absentismus als Folge einer die Arbeitszufriedenheit steigernden Intervention im Betrieb, ist ja in keiner Weise identisch mit dem Grad des dadurch ausgelösten „Nutzens“, der überdies für den gleichen Sachverhalt für verschiedene Betroffene (Unternehmen, unmittelbare Vorgesetzte, Mitarbeiter selbst) sehr unterschiedlich sein kann.

Die *Nutzenmessung*, also die Zuordnung von subjektiv bewerteten Größen zu Variablenausprägungen bzw. evaluierten Objekten/Alternativen kann in vielfältiger Weise erfolgen, insbesondere:

- *direkte Einschätzung*; den Personen werden verschiedene Ausprägungsgrade des objektiven Meßwertes vorgelegt, und sie werden gebeten, innerhalb einer vorgegebenen Bandbreite (z.B. von 0 bis 100) den von ihnen eingeschätzten „Nutzenwert“ anzugeben; dieses Verfahren ist vor allem dann angezeigt, wenn der Nutzen von nur einer einzelnen Variablen einzuschätzen ist, z.B. an dem Maß des Lernfortschritts oder der „Mitarbeiterorientierung“ von Vorgesetzten.
- *Rangreihenbildung*; vorgelegt werden verschiedene Ausprägungsgrade der Meßwerte bzw. Beschreibungen von Evaluationsobjekten mit bestimmten Meßwertergebnissen, diese sind von den Probanden in eine Reihenfolge zu bringen. Das Verfahren empfiehlt sich besonders dann, wenn von einer eindimensionalen Nutzenfunktion ausgegangen werden kann, aber für jedes Objekt gleichzeitig mehrere nutzenrelevante Aspekte zu beachten sind, also z.B. „Szenarien“ oder die Beschreibung „typischer“ Verhaltensbündel von Führungskräften nach Durchführung von Trainings.
- *Paarvergleichsmethoden*; die einzelnen Alternativen werden paarweise vorgegeben und es wird gefragt, welche den jeweils höheren Nutzenwert hat; dieses Verfahren ist dann angezeigt, wenn die Eindimensionalität der Nutzenfunktion fraglich ist, da durch geeignete Auswertungsverfahren die Ergebnisse von Paarvergleichen auch auf mehrdimensionale zugrundeliegende Entscheidungsstrukturen zurückgeführt werden können. Anwendungsgebiete sind etwa Beschreibungen der empirischen Ergebnisse verschiedener Behandlungsformen unter Einbezug von Nebenwirkungen, die Auswahl von Anbietern (z.B. verschiedene Dozenten von Trainings) auf der Basis ihrer empirisch festgestellten Trainingsergebnisse oder die nachträgliche Bewertung von Stellenbesetzungen.

Ausführliche Darstellungen für die hier nur knapp skizzierten Fragen der Nutzenmessung finden sich z.B. in Kaplan und Schwartz (1975) sowie Wendt und Vlek (1975).

Ein besonders für die Evaluation von Interventionswirkungen entwickeltes Verfahren ist die sogenannte *Goal-Attainment-Scale*, die vor allem im klinischen Bereich eingesetzt wird. Die „Nutzenbewertung“ erfolgt in der Form, daß für den konkreten Einzelfall möglichst exakt Dimensionen des angestrebten Erfolges beschrieben und für jede dieser Dimensionen fünf Kategorien des möglichen nutzenbezogenen Ergebnisses (z.B. „-2“: ein besonders unerwünschter Ausgang, „+2“: Beschreibung des

bestmöglichen Ergebnisses) vor Beginn der Intervention festgelegt werden. Nach der Intervention können die erzielten Ergebnisse auf dieser Punkteskala pro Dimension bestimmt und, ggf. unter Verwendung einer unterschiedlichen Gewichtung der einzelnen Dimensionen, zu einem nutzenbezogenen Gesamtwert des Interventionsresultates zusammengefaßt werden. Dadurch wird nicht nur ein Vergleich mit dem für jeden einzelnen Probanden ggf. unterschiedlichen Idealstandard retrospektiv ermöglicht, sondern auch eine die Besonderheiten der einzelnen Probanden berücksichtigende Vergleichsmöglichkeit verschiedener Interventionen (auf der Basis der zusammengefaßten Ergebnisse der den einzelnen Interventionen ausgesetzten Probandengruppe) geschaffen. Eine ausführliche Darstellung dieses Verfahrens gibt etwa Wittmann (1985).

3.3 Nutzenverrechnung

Es gibt bei Evaluationsstudien eine Reihe von Situationen, in denen die Nutzenmessung, als die Bewertung von einzelnen Variablenausprägungen oder von „Ergebnisszenarien“, für die weitere Arbeit nicht ausreicht. Zum einen kann es notwendig sein, Globalurteile für Ergebnisszenarien auf die der Nutzenbewertung zugrundeliegenden Teildimensionen zurückzuführen; zum anderen gibt es Fragestellungen, bei denen die rechnerische Abschätzung des erwartbaren Gesamtnutzens für verschiedene Grade der Ausprägungen der einzelnen nutzenrelevanten Aspekte notwendig ist.

Analyse von Globalurteilen

Die Zerlegung von Globalurteilen in objektivierbare Teil-Nutzenaspekte ist immer dann erforderlich, wenn für die spätere Nutzung der Ergebnisse nach einer expliziten Rechenvorschrift gesucht wird, die nach Messung der Einzelaspekte der Nutzenbewertung die Einholung eines solchen globalen Expertenurteiles überflüssig macht. Dies ist z.B. dann wichtig, wenn man für die Gestaltung eines Trainings vor der konkreten Durchführung abschätzen kann, wie sich als Folge der verschiedenen möglichen Trainingsvarianten die Teil-Nutzenaspekte verändern werden, ohne daß man – da ja noch keine Realisierung vorliegt – ein zusammenfassendes Urteil (z.B. durch die Teilnehmer) einholen kann.

Das Standardverfahren dafür ist die sogenannte *multiattributive Nutzenmessung* (MAUT). Diese geht von einem der Regressionsrechnung entsprechenden Zusammenhang zwischen den einzelnen Teilaspekten und der Gesamtnutzenbewertung aus (diese wird als eine gewichtete Summe der Ausprägungen der Teilnutzen dargestellt). Von den „Experten“ wird sowohl eine Einschätzung des globalen Nutzens als auch der einzelnen Teilnutzenaspekte erhoben und geprüft, ob die Beziehung zwischen diesen Größen tatsächlich der vermuteten linearen Modellform entspricht. Ist dies nicht der Fall, kann dies den Experten rückgemeldet werden (ggf. unter Nutzung von Sozialtechniken wie etwa Delphi, vgl. dazu Abschnitt 2.1) und dieses Verfahren solange fortgeführt werden, bis eine ausreichende Übereinstimmung besteht. Danach kann das Globalurteil durch die Messung der Teilaspekte unter Beachtung der empirischen Gewichtungen bestimmt werden. Ein Anwendungsbeispiel aus dem Evaluationsbereich geben etwa Kasubeck und Aschenbrenner (1978).

Zusammenfassung mehrerer Nutzenaspekte

Für die globale Bewertung von Alternativen auf der Basis von Einzel-Nutzendimensionen können vor allem folgende Modelle angewandt werden:

- Bestimmung der Teilnutzen jeder Dimension auf der gleichen „Einheit“, wobei dann meistens Geldbeträge gewählt werden; die „geldwerten Einschätzungen“ der einzelnen Dimensionen können dann zu Gesamtnutzenwerten addiert werden. Dieses Verfahren erscheint vor allem bei Evaluationsstudien mit stark betriebswirtschaftlichen Aspekten beliebt zu sein, zu Fragen der Bildungsarbeit in Unternehmen siehe dazu etwa Drumm (1992).
- Die gewichtete Zusammenfassung der Teilaspekte, wobei hier besondere Skalenprobleme (Nutzenmessung meist nur auf Rangskalen!) eine Rolle spielen können; in Übereinstimmung mit den Ergebnissen zum Allgemeinen Linearen Modell der parametrischen Statistik (vgl. den gleichlautenden Beitrag von Andres, in diesem Band) kann es von Vorteil sein, statt mit den ordinalen Meßwerten mit entsprechenden Rangplätzen zu arbeiten. Voraussetzung für dieses Vorgehen ist entweder eine direkte Schätzung der Gewichte z.B. durch Experten oder mit Hilfe der im vorhergehenden Absatz beschriebenen Aufklärung von Globalurteilen.
- Verschiedene Formen der Profilvergleiche: Hat man von zwei Alternativen jeweils ein mehrdimensionales Nutzenprofil, scheidet – unabhängig von der Stärke der Skala und einer speziellen Gewichtung der Teildimensionen – eine Alternative prinzipiell dann aus, wenn sie in allen Dimensionen einen schlechteren Wert hat als die überlegene Alternative; entfernt man nach diesem Grundsatz alle Alternativen, die in diesem Sinne von mindestens einer anderen übertroffen werden, aus der Menge der zu bewertenden Möglichkeiten, erhält man die sogenannte *paretooptimale Teilmenge*. Die Entscheidung innerhalb dieser Teilmenge muß dann nach verschiedenen Distanzmaßen erfolgen.

4 Methodische Besonderheiten prospektiver Evaluation

Für die meisten Evaluationsstudien liegt der Nutzen nicht in der nachträglichen Bewertung, sondern in der Gestaltung zukunftsrelevanter Entscheidungen. Hierfür werden alle in den bisherigen Abschnitten besprochenen Aspekte (sinnvolle Vergleichsgrundlagen, Nutzenmessung und Nutzenverrechnung) benötigt. Zusätzlich treten aber methodische Forderungen auf, die sich auf die Verallgemeinerbarkeit der Ergebnisse für die zukünftigen Verhältnisse beziehen und damit strenger sind als bei der rückblickenden Betrachtung.

4.1 Entscheidungsorientierte Variablenauswahl

Um Evaluationsstudien zu vermeiden, die zwar hohes wissenschaftliches Niveau haben, aber aus Sicht des Entscheidungsträgers wenig konkreten Nutzen bieten, sollte sich jeder Evaluator folgende Fragen stellen:

- Ist für jede der von mir aufgenommenen Variablen ein klarer Zusammenhang mit den verfügbaren Alternativen des relevanten Entscheidungsträgers gegeben?
- Welche konkreten Ergebnisse in diesen Variablen würden welche Entscheidung des Auftraggebers zur Folge haben? Hier kann sich das diskussionsmäßige „Durchspielen“ fiktiver Ergebnisse mit dem Entscheidungsträger als sehr nützlich erweisen, da die Gefahr besteht, daß der Evaluator und der Entscheidungsträger sehr unterschiedliche Vorstellungen über die Verwendung bzw. Verwendbarkeit der empirischen Ergebnisse haben.
- Ist die Menge der nach diesen Kriterien aufgenommenen Variablen ausreichend, oder gibt es Gestaltungsvarianten des Entscheidungsträgers, zu denen die bisher berücksichtigten Aspekte keine aussagekräftigen Hilfen zu einer differenzierten Bewertung zulassen?

Der Grund, warum in vielen Evaluationsprojekten die Perspektive des Entscheidungsträgers nicht ausreichend gesehen wird, kann in dem unterschiedlichen Rollenverständnis des Evaluators liegen. Viele verstehen sich aufgrund ihrer im Wissenschaftsbereich geprägten Sozialisation nicht als „Auftragnehmer“, sondern als „Forscher“. Für die Probleme des Rollenverständnisses des Evaluators, die leider auch erhebliche Auswirkungen auf die Leistungsfähigkeit des methodischen Vorgehens haben, siehe Alkin (1990).

4.2 Stichprobenprobleme

Während bei der retrospektiven Evaluation relativ wenige Verallgemeinerungsprobleme entstehen, sind die Fragen einer sachgerechten Stichprobenziehung bei einem prospektiven Ansatz außerordentlich komplex:

- Die in die Untersuchung *aufgenommenen Personen* sollten repräsentativ für die *später* den verschiedenen Maßnahmen ausgesetzten Personen sein; leider ist es außerordentlich schwierig, innerhalb des rechtlichen Rahmens einer demokratischen Gesellschaft diese Forderung zu erfüllen. Auch bei sorgfältigster Untersuchungsplanung ist damit zu rechnen, daß die freiwillige Teilnahme an der Erprobung einer neuen Schulform, einer neuen medizinischen Behandlungsmethode oder an einem neuartigen, auf bisher nicht verwendeten Inhalten aufbauenden Führungskräfte trainings keine zufällige, repräsentative Auswahl für die späteren Nutzer solcher Angebote zur Folge hat.
- Die für die Evaluation probeweise realisierten Alternativen sollten hinsichtlich ihrer Auswirkung mit den späteren Echt-Verhältnissen vergleichbar sein; auch dies ist kaum realisierbar, da zum einen die Neuartigkeit von Ansätzen schon an sich Auswirkungen hat (siehe dazu den *Hawthorne-Effekt*, etwa in Frese, 1988), zum anderen sich aufgrund der für die Probephase bereitstehenden Mittel für die Gestaltung der Alternativen gegenüber der späteren Situation wesentliche Begrenzungen oder Vorteile ergeben (z.B. die Auswirkungen der bei Modellversuchen stets vorgesehenen Evaluation auf die Ergebnisse des Versuches selbst).

- Die als *Durchführende* für die einzelnen Maßnahmenalternativen eingesetzten Personen sollten ebenfalls eine repräsentative Stichprobe der später mit dieser Arbeit betrauten Personen sein. In den allermeisten Fällen findet aber eine systematische Auswahl, im wesentlichen durch Selbstselektion, von Lehrern für verschiedene Schulformen statt, neue Therapieformen werden zunächst meist von ihrem Erfinder oder seinen überzeugten Anhängern praktisch eingeführt, und ebensowenig ist eine zufällige Zuordnung von Trainern oder Unternehmensberatern zu neuen methodischen Vorgehensweisen denkbar.

Die Konsequenz ist, daß der Aufnahme von zusätzlichen unabhängigen bzw. als Moderatorvariablen verwendbaren Aspekten eine besondere Bedeutung zukommt. Zwar kann das statistisch fundierte Verallgemeinern auch dadurch nicht gesichert werden, diese zusätzlichen Informationen können aber entscheidend dazu beitragen, Fehlinterpretationen einzuschränken.

4.3 *Abschätzung zukünftiger Auswirkungen im Rahmen von Entscheidungsprozessen*

Für den Einbau in konkrete Entscheidungsvorgänge ist neben einer (annähernd) zufriedenstellenden Klärung der Verallgemeinerungsmöglichkeit aufgrund der Stichprobengestaltung die Frage wichtig, von welchem Modell zur Abschätzung der zukünftigen Wirkung man ausgeht. Im Prinzip stehen dafür drei Varianten zur Verfügung:

- Der einfache *Induktionsschluß*; man geht davon aus, daß eine Maßnahme in der Zukunft die gleichen Auswirkungen haben wird wie in der Vergangenheit, sofern die wichtigsten Rahmenbedingungen konstant gehalten bzw. kontrolliert werden können. Dies dürfte die häufigste Schlußweise für die prospektive Verwertung von Evaluationsergebnissen sein, allerdings als Folge der sich verändernden Rahmenbedingungen häufig nicht mit ausreichender Aussagekraft.
- Der Einbau der Ergebnisse in ein umfassendes, kausales *Zusammenhangsmodell*; gelingt es, die relevanten Nutzenfunktionen als abhängige Variablen in einem komplexen Wirkungsgefüge ausreichend präzise darzustellen, kann in ein solches Prognosemodell die erwartbare zukünftige Veränderung eingesetzt und damit die später auftretenden Auswirkungen der einzelnen evaluierten Alternativen abgeschätzt werden. Solche Variablenmodelle (je nach Aufwand Regressionsgleichungen, pfadanalytische Modelle oder das LISREL-Verfahren, vgl. Rietz, Rudinger & Andres, in diesem Band) können eine erhebliche Hilfe sein, lassen sich aber aufgrund der Komplexität der Zusammenhänge nicht immer erarbeiten. Hinzu kommt, daß die aus rechentechnischen Gründen unterstellte lineare Struktur nicht immer den empirischen Gegebenheiten gerecht wird.
- In Einzelfällen ist die Nutzung *dynamischer* Modelle möglich, etwa bei der prospektiven Evaluation im Marketingbereich (wo die durch die Maßnahmen erreichte Marktsättigung in der Kernzielgruppe ein wichtiger dynamischer Aspekt ist).

Die Maßnahmenbewertung kann stark von den verwendeten Modellformen abhängen. Je besser das Modell die Realität erfaßt, umso höher ist die Verwertbarkeit der Evaluations-Interpretation. Aus diesem Grund sollte man einen unreflektierten Induktionsschluß vermeiden.

5 Meta-Analyse als Entscheidungshilfe

Gerade die im vorhergehenden Abschnitt besprochene Vielzahl von Störeffekten legt es nahe, Entscheidungen möglichst nicht auf einzelnen prospektiven Evaluationsstudien aufzubauen, sondern möglichst viele in aussagekräftiger Weise zusammenzufassen. Dies wird vor allem dann sinnvoll sein, wenn man eine grobe Bewertung verschiedener Vorgehensweisen (die sich dann im Detail noch immer, etwa aufgrund eigener Evaluationsstudien, gemäß den Besonderheiten des konkreten Anwendungsfalles präzisieren läßt) anstrebt. Die Zusammenfassung vieler Studien ermöglicht auch die Berücksichtigung sehr vieler Maßnahmenvarianten oder Erhebungsinstrumente, was in einer einzelnen Untersuchung schon aus Aufwandsgründen nicht möglich wäre.

Allerdings sind mit der zusammenfassenden Betrachtung einer Vielzahl von unterschiedlichen Studien auch einige typische Probleme verbunden, die ein metaanalytisches Vorgehen berücksichtigen muß.

5.1 Auswahl der „berücksichtigungswürdigen“ Arbeiten

Möchte man für ein konkretes Entscheidungsproblem, z.B. die Auswahl von Instrumenten zur Eignungsdiagnostik in Unternehmen, die bereits vorhandene Information zusammentragen, wird man natürlich bemüht sein, möglichst viele relevante Arbeiten in die Meta-Analyse einzubeziehen. Dabei treten aber folgende Probleme auf:

- Ein unbekannt hoher, aber möglicherweise wichtiger Anteil von Evaluationsvorhaben wird *nicht publiziert*; dies betrifft zum einen Studien ohne signifikante Ergebnisse (Wer publiziert eine Arbeit, in der nicht gezeigt wird, daß sein bevorzugtes Verfahren eine Validität aufwies? Und welche Zeitschrift druckt eine solche „ergebnislose“ Arbeit?), zum anderen die Unterdrückung unerwünschter Resultate in Abhängigkeit von politischen oder wissenschaftlichen „Modeströmungen“ (z.B. negative Befunde zu einer gerade „innovativen“ Schulform, negative Auswirkungen einer dem politischen Zeitgeist entsprechenden Gesetzgebung). Faßt man nur die publizierten Arbeiten zusammen, müssen sich daraus Verzerrungen im Hinblick auf besonders viele positive Befunde vor allem der „modernen“ Bestrebungen ergeben, was tendenziell zu einer „Aufschaukelung“ von Modewellen führt.
- Die *Relevanz* einzelner Studien ist uneinheitlich; so stellt sich die Frage, bis zu welchem Zeitpunkt man zurückgehen soll (ab wann haben sich z.B. die Anforderungen an Führungskräfte so stark verändert, daß ältere Studien heute nicht mehr aussagefähig sind?), welche Kontextbedingungen übertragbar sind (sind Studien über das Führungsverhalten im öffentlichen Dienst auch für die Privatwirtschaft relevant?) oder welche Realisierung von Maßnahmenalternativen noch innerhalb des interessanten Spektrums liegt (ist die Bewertung ei-

nes Führungskräfte trainings, das von einem berühmten „Guru“ durchgeführt wurde, tatsächlich für die Entscheidung zwischen verschiedenen Trainingsformen im eigenen Unternehmen aussagefähig?).

- Die Studien haben *unterschiedliche Qualitäten*; dies kann zum Teil an objektivierbaren Fakten (Stichprobengröße, Design, verwendete Meßinstrumente) festgemacht werden, bleibt aber zum Teil auch „schwammig“ (Art und Ort der Publikation, Stringenz der Darstellung, Eindruck der Glaubwürdigkeit der ermittelten Befunde). Da es ebenso falsch wäre, schlechte Studien in die Meta-Analyse aufzunehmen, wie gute Arbeiten aufgrund eines fragwürdigen subjektiven Eindrucks zu vernachlässigen, sind hier Setzungen erforderlich, die sich nicht immer voll objektivieren lassen.

Ein weiterer Aspekt ist natürlich die Zugänglichkeit der Studien. Manche aussagekräftigen Befunde sind nur in schwer zugänglichen internen Berichten festgehalten, so daß auch die Fähigkeit des Meta-Analytikers zur Beschaffung von interessantem Material das Ergebnis mit beeinflußt.

5.2 Statistische Fragen der Meta-Analyse

Ein in der Literatur ausführlich behandelter Aspekt ist das Ermitteln von statistischen Überschreitungswahrscheinlichkeiten bei Zusammenfassung mehrerer Studien (vgl. dazu etwa Hunter & Schmidt, 1990), der aber für die praktische Nutzung von Meta-Analysen im Evaluationsbereich nur eine geringe Rolle spielt. Von erheblicher praktischer Bedeutung ist die Ermittlung von Effektstärken über viele Studien hinweg, da man ja für den Vergleich verschiedener Alternativen eine Kenngröße für die Auswirkung der einzelnen Variablen benötigt. Die hierbei auftretenden Probleme betreffen insbesondere

- die unterschiedlichen Stichprobengrößen in den einzelnen Untersuchungen,
- die unterschiedliche Aussagekraft der verwendeten Meßinstrumente (Varianz, Reliabilität),
- die Wahl einer Kennziffer für Effektstärke, meist wird ein auf die jeweilige Varianz normiertes Differenzmaß verwendet; es kann aber auch zweckmäßig sein, statt einer Mittelung über die Studien hinweg die Prozenträge der Studien bezüglich der erzielten Effektgrößen anzugeben.

Im Gegensatz zu den stets problematischen Fragen der Auswahl der Studien gibt es für die Lösung der statistischen Probleme inzwischen sehr bewährte Ansätze, die auch programmtechnisch unterstützt sind. Einen umfangreichen Überblick zu dieser Problematik geben Hunter und Schmidt (1990).

5.3 Die Leistungsfähigkeit von Meta-Analysen für die Evaluation

Es liegt inzwischen eine größere Anzahl von relevanten Meta-Analysen vor, etwa zu Testvaliditäten Ones, Chockalingam und Schmidt (1993), Schuler und Funke (1989); zur Organisationsentwicklung Gebert (1974), Neumann, Edwards und Raju (1989),

für einen umfassenden Überblick siehe Gebert (1993, S. 490ff); für den schulischen Unterricht relevante Fragen in Schiefele, Krapp und Schreyer (1993).

Solche Studien sind typische Beispiele für Informationen, die der evaluativen Einschätzung von Experten zugrunde liegen. Insgesamt ist aus methodischer Sicht der Evaluationsprozeß als Kreislauf zu sehen, von einer – vielleicht noch wenig empirisch fundierten – Expertenbewertung über noch überschaubare retrospektive Studien zu den vom methodischen Aufwand her wesentlich größeren prospektiven Arbeiten, die ihrerseits die Grundlage für Meta-Analysen schaffen und damit die Basis für verbesserte Expertenbewertungen sind. Die Methoden sind heute so weit entwickelt, daß mit ihrer Hilfe fundierte Beiträge für die entscheidungsrelevante Evaluationsforschung geleistet werden können, wenn auch nicht immer im Sinne von Ideallösungen.

6 Weiterführende Literatur

Allgemeine Einführungen in den Gegenstandsbereich der Evaluation geben Franklin und Thrasher (1976), Hunter, Schmidt und Jackson (1982), Wottawa und Thierau (1990) sowie Shadish, Cook und Leviton (1991). Spezielle statistische Fragen werden in den Büchern von Cook und Reichardt (1979), Pollard (1986) sowie Morris, Fitz-Gibbon und Lindheim (1987) behandelt.

Interessante Beispiele für Evaluationsstudien in verschiedenen Gegenstandsbereichen aus dem deutschsprachigen Raum geben Lukie (1987), Hager und Hasselhorn (1993), Eckes und Six (1994) sowie Grawe, Donati und Bernauer (1994). Wer konkrete Hilfen für die praktische Durchführung von Evaluationen sucht, z.B. in Form von konkreten Meßinstrumenten oder „Checklisten“, findet diese bei Herman (1988).

Literaturverzeichnis

- Alkin, M. C. (1990). *Debates on evaluation*. Beverly Hills: Sage.
- Borg, I. (1981). *Anwendungsorientierte multidimensionale Skalierung*. New York: Springer.
- Cook, T. D. & Reichardt, C. S. (1979). *Qualitative and quantitative methods in evaluation research*. Beverly Hills: Sage.
- Dienel, P. C. (1978). *Die Planungszelle*. Opladen: Westdeutscher Verlag.
- Drumm, H. J. (1992). *Personalwirtschaftslehre*. Heidelberg: Springer.
- Eckes, T. & Six, B. (1994). Fakten und Fiktionen in der Einstellungs-Verhaltens-Forschung: Eine Meta-Analyse. *Zeitschrift für Sozialpsychologie*, 25, 253-271.
- Erlbaum, L., Vosmadon, S. & Otony, A. (1989). *Similarity analogical reasoning*. Cambridge: Cambridge University Press.
- Franklin, J. L. & Thrasher, J. H. (1976). *An introduction to program evaluation*. New York: Wiley.
- Frese, E. (1988). *Grundlagen der Organisation*. Wiesbaden: Gabler.
- Gebert, D. (1974). *Organisationsentwicklung – Probleme des geplanten organisatorischen Wandels*. Stuttgart: Kohlhammer.
- Gebert, D. (1993). Interventionen in Organisationen. In H. Schuler (Hrsg.), *Organisationspsychologie* (S. 481-494). Bern: Huber.
- Gentner, D. & Stevens, A. L. (Eds.) (1983). *Mental models*. Hillsdale: Erlbaum.
- Grawe, K., Donati, R. & Bernauer, F. (1994). *Psychotherapie im Wandel: Von der Konfession zur Profession*. Göttingen: Hogrefe.

- Groeben, N. & Scheele, B. (1984). *Heidelberger-Struktur-Legetechnik*. Weinheim: Beltz.
- Hager, W. & Hasselhorn, M. (1993). Evaluation von Trainingsmaßnahmen am Beispiel von Klauers Denktraining für Kinder. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 25, 307–321.
- Herman, J. L. (1988). *Program evaluation kit*, Band I – IX. Newbury Park: Sage.
- Hunter, J. E. & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park: Sage.
- Hunter, J. E., Schmidt, F. L. & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills: Sage.
- Internationale Organisation für Normung (1992). *Qualitätsmanagement und Elemente eines Qualitätssicherungssystems. Leitfaden für Dienstleistungen. DIN ISO 9004*. Teil 2. Berlin: Ausschuß Qualitätssicherung und angewandte Statistik (AQS) im DIN Deutschen Institut für Normung e. V.
- Kaplan, M. F. & Schwartz, S. (1975). *Human judgement and decision processes*. New York: Academic Press.
- Kasubeck, W. & Aschenbrenner, K. M. (1978). Optimierung subjektiver Urteile: Anwendung der multiattributiven Nutzentheorie bei medizinischen Therapieentscheidungen. *Zeitschrift für experimentelle und angewandte Psychologie*, 25, 594–616.
- Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 95, 163–182.
- Linstone, H. A. & Turoff, M. (1975). *The Delphi method*. London: Addison-Wesley.
- Lukie, M. (1987). Evaluation innerbetrieblicher Traineeprogramme. In H. Will, A. Winterler & A. Krapp (Hrsg.), *Evaluation in der beruflichen Aus- und Weiterbildung* (S. 169–184). Heidelberg: Sauer.
- Morris, L. L., Fitz-Gibbon, C. T. & Lindheim, E. (1987). *How to measure performance and use tests*. Newbury Park: Sage.
- Neumann, G. A., Edwards, J. E. & Raju, N. S. (1989). Organizational development interventions: A metaanalysis of their effects on satisfaction and other attitudes. *Personnel Psychology*, 42, 461–489.
- Ones, D. S., Chockalingam, V. & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, 78, 679–703.
- Pollard, W. E. (1986). *Bayesian statistics for evaluation research. An introduction*. Beverly Hills: Sage.
- von Reibnitz, U. (1983). Die Szenario-Technik - Ein Instrument der Zukunftsanalyse und der strategischen Planung. In H. Haase & K. Koeppler (Hrsg.), *Fortschritte der Marktpsychologie*. Band 3. Frankfurt: Fachbuchhandlung für Psychologie.
- Rossi, P. H. & Freeman, H. E. (1985). *Evaluation. A systematic approach*. Beverly Hills: Sage.
- Rutman, L. (1977). *Evaluation research methods: A basic guide*. Beverly Hills: Sage.
- Schiefele, U., Krapp, A. & Schreyer, I. (1993). Metaanalyse des Zusammenhangs von Interesse und schulischer Leistung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 25, 120–148.
- Schuler, H. & Funke, U. (1989). Berufseignungsdiagnostik. In E. Roth (Hrsg.), *Organisationspsychologie* (= Enzyklopädie der Psychologie, Serie 3, Band 3, S. 281–320). Göttingen: Hogrefe.
- Shadish, W. R., Cook, T. D. & Leviton, L. E. (1991). *Foundations of program evaluation: Theories of practice*. Beverly Hills: Sage.
- Spada, H. & Reimanns, P. (1988). Wissensdiagnostik auf kognitionswissenschaftlicher Basis. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 9, 183–192.

- Wendt, D. & Vlek, C. (1975). *Utility, probability and human decision making*. Selected proceedings of an interdisciplinary research conference, Rome, 3-6 Sept. 1973. Dordrecht: Riedel.
- Wittmann, W. W. (1985). *Evaluationsforschung. Aufgaben, Probleme und Anwendungen*. Berlin: Springer.
- Wottawa, H. & Hof, A. (1987). Individualmodelle zur Erfassung handlungsrelevanter Strukturen als Hilfsmittel der Verbesserung von Weiterbildungsmaßnahmen. In H. Will, A. Winteler & A. Krapp (Hrsg.), *Evaluation in der Aus- und Weiterbildung*. Heidelberg: Sauer.
- Wottawa, H. & Thierau, H. (1990). *Lehrbuch Evaluation*. Bern: Huber.