

Enzyklopädie der Psychologie

Grundlagen und
statistische Methoden
der Evaluationsforschung

Evaluation

1



Hogrefe · Verlag für Psychologie
Göttingen · Bern · Toronto · Seattle

Enzyklopädie der Psychologie

ENZYKLOPÄDIE DER PSYCHOLOGIE

In Verbindung mit der
Deutschen Gesellschaft für Psychologie

herausgegeben von

Prof. Dr. Niels Birbaumer, Tübingen
Prof. Dr. Dieter Frey, München
Prof. Dr. Julius Kuhl, Osnabrück
Prof. Dr. Wolfgang Schneider, Würzburg
Prof. Dr. Ralf Schwarzer, Berlin

Themenbereich B

Methodologie und Methoden

Serie IV

Evaluation

Band 1

Grundlagen und statistische Methoden
der Evaluationsforschung



Hogrefe • Verlag für Psychologie
Göttingen • Bern • Toronto • Seattle

Grundlagen und statistische Methoden der Evaluationsforschung

herausgegeben von

Prof. Dr. Heinz Holling, Münster



Hogrefe • Verlag für Psychologie
Göttingen • Bern • Toronto • Seattle

Bibliografische Information der Deutschen Nationalbibliothek

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

© 2009 Hogrefe Verlag GmbH & Co. KG
Göttingen • Bern • Wien • Paris • Oxford • Prag • Toronto
Cambridge, MA • Amsterdam • Kopenhagen • Stockholm
Rohnsweg 25, 37085 Göttingen

<http://www.hogrefe.de>

Aktuelle Informationen • Weitere Titel zum Thema • Ergänzende Materialien



Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Satz: Grafik-Design Fischer, Weimar
Druck und Bindung: AZ Druck und Datentechnik, Kempten/Allgäu
Auf säurefreiem und chlorfrei gebleichtem Papier gedruckt
Printed in Germany

ISBN 978-3-8017-1505-2

Autorenverzeichnis

Prof. Dr. Wolfgang Bilsky

Westfälische Wilhelms-Universität
Münster
Psychologisches Institut IV
Fliednerstraße 21
48149 Münster
E-Mail: bilsky@uni-muenster.de

Prof. Dr. Ingwer Borg

ZUMA – Zentrum für Umfragen,
Methoden und Analysen
Postfach 12 21 55
68072 Mannheim
E-Mail: ingwer.borg@gesis.org

Dr. David Cairns

Macquarie University
Department of Psychology
North Ryde, NSW 2109
Australien
E-Mail: dcairns@psy.mq.edu.au

Dipl.-Stat. Britta Colver

Westfälische Wilhelms-Universität
Münster
Psychologisches Institut IV
Statistik und Methoden
Fliednerstraße 21
48149 Münster
E-Mail: britta-colver@uni-muenster.de

Prof. Dr. Kai S. Cortina

University of Michigan
Department of Psychology
1012 East Hall
530 Church Street
Ann Arbor, MI 48109-1043
USA
E-Mail: kai.cortina@umich.edu

Prof. Dr. Nicola Döring

TU Ilmenau, IfMK
EAZ, Ehrenbergstraße 29
98693 Ilmenau
E-Mail: nicola.doering@tu-ilmenau.de

Prof. Dr. Michael Eid

Freie Universität Berlin
Arbeitsbereich Methoden und
Evaluation
Habelschwerdter Allee 45
14195 Berlin
E-Mail: eid@zedat.fu-berlin.de

Dr. Friedrich Funke

Friedrich-Schiller-Universität Jena
Institut für Psychologie
Humboldtstraße 26
07743 Jena
E-Mail: sff@uni-jena.de

PD Dr. Günther Gediga

Westfälische Wilhelms-Universität
Münster
Psychologisches Institut IV
Statistik und Methoden
Fliednerstraße 21
48149 Münster
E-Mail: gediga@uni-muenster.de

Dr. Heiko Grossmann

Queen Mary University of London
School of Mathematical Sciences
Mile End Road
London E1 4NS
United Kingdom
E-Mail: h.grossmann@qmul.ac.uk

PD Dr. Christian Heumann

Institut für Statistik
LMU München
Ludwigstraße 33
80539 München
E-Mail: christian.heumann@stat.uni-
muenchen.de

Prof. Dr. Heinz Holling

Westfälische Wilhelms-Universität
Münster
Psychologisches Institut IV
Statistik und Methoden
Fliednerstraße 21
48149 Münster
E-Mail: holling@psy.uni-
muenster.de

Prof. Dr. Karl Christoph Klauer

Universität Freiburg
Institut für Psychologie
Sozialpsychologie und Methodenlehre
Engelbergerstraße 41
79085 Freiburg
E-Mail: Christoph.Klauer@
psychologie.uni-freiburg.de

Dipl.-Psych. Martin Löb

Erich-Weinert-Straße 39
10439 Berlin
E-Mail: martin.loeb@helios-
kliniken.de

Prof. Dr. Georg E. Matt

San Diego State University
Department of Psychology
San Diego, CA 92182-4611
USA
E-Mail: gmatt@sciences.sdsu.edu

Prof. Dr. Helfried Moosbrugger

Goethe-Universität Frankfurt
Psychologische Methodenlehre
Evaluation und Forschungs-
methodik
Mertonstraße 17
60054 Frankfurt
E-Mail: moosbrugger@psych.uni-
frankfurt.de

Dr. Thomas Nittner

Hornhaldenstrasse 17
CH-8802 Kilchberg ZH
E-Mail: tnittner@web.de

Dr. Hans Anand Pant

Institut für Schulqualität
der Länder Berlin und Brandenburg,
An-Institut der Freien Universität
Berlin
Otto-von-Simson-Straße 15
14195 Berlin
E-Mail: HansAnand.Pant@
isq-bb.de

Dr. Ivailo Partchev

Friedrich-Schiller-Universität Jena
Institut für Psychologie
Lehrstuhl Methodenlehre und
Evaluationsforschung
Am Steiger 3, Haus 1
07743 Jena
E-Mail: s7paiv@uni-jena.de

Dr. Franziska Perels

Institut für Qualitätsentwicklung
Walter-Hallstein-Straße 5–7
65197 Wiesbaden
E-Mail: f.perels@iq.hessen.de

Dr. Wolfgang Rauch

Goethe-Universität Frankfurt
Institut für Psychologie
Arbeitsbereich Pädagogische
Psychologie
Senckenberganlage 15
60325 Frankfurt
E-Mail: wolfgang.rauch@psych.uni-
frankfurt.de

Prof. em. Dr. Jürgen Rost

Leibniz-Institut für die Pädagogik
der Naturwissenschaften (IPN)
Olshausenstraße 62
24098 Kiel
E-Mail: an@j-rost.de

Prof. Dr. Bernhard Schmitz

Technische Universität Darmstadt
Institut für Psychologie
Alexanderstraße 10
64283 Darmstadt
E-Mail: schmitz@psychologie.tu-
darmstadt.de

Dr. Matthias Siemer

Ernst-Moritz-Arndt-Universität
Greifswald
Institut für Psychologie
Lehrstuhl für Allgemeine Psycho-
logie und Forschungsmethodik
Franz-Mehring-Straße 47
17487 Greifswald
E-Mail: masiemer@mail.uni-
greifswald.de

Prof. Dr. Rolf Steyer

Friedrich-Schiller-Universität Jena
Institut für Psychologie
Lehrstuhl Methodenlehre und
Evaluationsforschung
Am Steiger 3, Haus 1
07743 Jena
E-Mail: rolf.steyer@uni-jena.de

Dr. Thorsten Stumpp

Universität Bielefeld
Fakultät für Psychologie und
Sportwissenschaft
Abteilung für Psychologie
Postfach 10 01 31
33501 Bielefeld
E-Mail: thorsten.stumpp@uni-
bielefeld.de

Prof. Dr. Dr. Helge Toutenburg

Ludwig-Maximilians-Universität
München
Department für Statistik
Ludwigstraße 33
80539 München
E-Mail: helge.toutenburg@stat.uni-
muenchen.de

Prof. Dr. Rainer Westermann

Ernst-Moritz-Arndt-Universität
Greifswald
Institut für Psychologie
Lehrstuhl für Allgemeine Psychologie
und Forschungsmethodik
Franz-Mehring-Straße 47
17487 Greifswald
E-Mail: westerma@uni-
greifswald.de

Prof. Dr. Hans Westmeyer

Freie Universität Berlin
Wissenschaftsbereich Psychologie
Habelschwerdter Allee 45
14195 Berlin
E-Mail: hawest@zedat.fu-berlin.de

Prof. Dr. Werner W. Wittmann

Universität Mannheim
Fakultät für Sozialwissenschaften
Lehrstuhl Psychologie II
Methodenlehre, Diagnostik und
Evaluation
Schloss Ehrenhof Ost
68131 Mannheim
E-Mail: wittmann@tnt.psychologie.
uni-mannheim.de

Dipl.-Psych. Andreas Wolf

Friedrich-Schiller-Universität Jena
Institut für Psychologie
Lehrstuhl Methodenlehre und
Evaluationsforschung
Am Steiger 3, Haus 1
07743 Jena
E-Mail: andreas.wolf@uni-
jena.de

Vorwort

Die Evaluationsforschung ist eine relativ junge Disziplin. Ihr Beginn geht auf die in den sechziger Jahren des letzten Jahrhunderts vorwiegend in den USA durchgeführten sozialwissenschaftlichen Untersuchungen zur Wirksamkeit bildungs- und sozialpolitischer Reformansätze zurück. Seitdem hat die Evaluationsforschung einen regelrechten Boom erlebt. In nahezu allen gesellschaftlichen Bereichen, wie Bildung, Verwaltung, Wirtschaft oder Gesundheit, werden heutzutage viele unterschiedliche Aktivitäten evaluiert. In vielen Staaten wurden eigene Evaluationsgesellschaften gegründet und inzwischen existieren zahlreiche wissenschaftliche Zeitschriften für Evaluationsforschung. Nicht nur in der Psychologie, sondern auch in den Nachbarwissenschaften, wie der Pädagogik, Politologie oder Ökonomie, spielt die Evaluationsforschung eine bedeutende Rolle. Evaluationsforschung ist insofern als ein interdisziplinäres Gebiet anzusehen. Diese zweibändige Serie innerhalb der Enzyklopädie der Psychologie soll einen Überblick über den derzeitigen Stand der Evaluationsforschung geben. Der inhaltliche Fokus liegt auf den für die Psychologie bedeutsamen Themen, jedoch behandeln viele Beiträge methodische und statistische Grundlagen, die ebenso einschlägig für die Evaluationsforschung in anderen wissenschaftlichen Disziplinen sind.

Es gibt zahlreiche unterschiedliche Definitionen der Begriffe Evaluation und Evaluationsforschung. Eine Definition dieses Gebietes wird am Anfang dieses Bandes entwickelt. Sie ist insofern wichtig, als sie die Auswahl der Themen in den beiden Bänden der Enzyklopädie Evaluationsforschung steuert. Im Kontext der Psychologie wird hier Evaluation und gleichbedeutend damit Evaluationsforschung als die Bewertung von psychologischen Interventionen mittels anerkannter wissenschaftlicher Methoden bestimmt. Die wichtigsten psychologischen Interventionsmaßnahmen stammen aus den psychologischen Anwendungsfächern, der Klinischen Psychologie, der Arbeits- und Organisationspsychologie und der Pädagogischen Psychologie, weiterhin aber auch aus anderen Disziplinen, wie der Verkehrspsychologie, Rechtspsychologie oder Markt- und Werbepsychologie. Häufig handelt es sich hier um die multiattributive Bewertung komplexer, längerfristiger und praktisch bedeutsamer Interventionen, über die zudem Entscheidungen zu treffen sind. Daher spielen hier spezifische Methoden, z. B. aus der multiattributiven Nutzentheorie oder der Entscheidungstheorie, eine wichtige Rolle. Da Evaluationsstudien häufig komplexer und

längerfristiger Natur sind, erfordern sie eine besonders sorgfältige methodische Planung und aufgrund großer Datenmengen insbesondere multivariate statistische Auswertungsverfahren. Zudem sind für die Durchführung der in vielen Fällen umfangreichen Evaluationsprojekte Kenntnisse und Kompetenzen aus dem Bereich des Managements gefordert, auch deshalb, weil es gilt, unterschiedlichen Interessen, z. B. der Auftraggeber oder der betroffenen Bevölkerungsgruppen, gerecht zu werden.

Es gibt mittlerweile zahlreiche Abhandlungen zur Evaluationsforschung, von denen viele sehr praxisorientiert ausgerichtet sind. Im Gegensatz dazu konzentrieren sich die Beiträge dieses Enzyklopädiebandes stärker auf die wissenschaftlichen Grundlagen der Evaluationsforschung. Hier stehen grundlegende Fragen zur Definition und wissenschaftstheoretischen Fundierung sowie zur Methodik und statistischen Auswertung im Vordergrund.

Im ersten Kapitel geht es um die Definition sowie spezifische Merkmale dieses Gebiets und damit um Unterscheidungsmerkmale zur allgemeinen Methodenlehre. Das zweite Kapitel thematisiert wissenschaftstheoretische Fragen der Evaluationsforschung, wie die logische Struktur des Evaluationsbegriffs, den empirischen Gehalt von Evaluationshypothesen oder das Problem der Generalisierbarkeit von Evaluationshypothesen. Im anschließenden dritten Kapitel werden die grundlegenden Modelle der Evaluationsforschung expliziert, die ebenso wie die im vierten Kapitel dargestellten Phasenmodelle dazu dienen, die häufig zeitlich ausgedehnten und komplexen Evaluationsprojekte zu strukturieren. Diesem Zweck dient auch die in Kapitel fünf dargestellte Facettentheorie, die sich im Wesentlichen mit der Forschungsplanung komplexer empirischer Studien befasst. Das Aufstellen und die Prüfung von Evaluationshypothesen, z. B. zur Überprüfung fester und zufälliger Effekte, werden im sechsten Kapitel abgehandelt, während im anschließenden siebten Kapitel die Generalisierbarkeitstheorie dargestellt wird, die eine differenzierte Analyse der Zuverlässigkeit der in Evaluationsstudien erzielten Effekte erlaubt.

Die Zusammenfassung der unüberschaubar werdenden empirischen Forschungsergebnisse zu nahezu allen sozialwissenschaftlichen, ökonomischen oder medizinischen Themen erfolgt zunehmend anhand des Verfahrens der Metaanalyse, das in Kapitel 8 dargestellt wird. Da in Evaluationsstudien häufig eine sehr große Anzahl von Variablen untersucht wird, erfordert die Auswertung solcher Studien sehr häufig multivariate statistische Verfahren. Daher werden in den Kapiteln 9 bis 18 alle wesentlichen multivariaten statistischen Verfahren abgehandelt. Dazu zählen das Allgemeine Lineare Modell, hierarchische und generalisierte lineare Modelle sowie Strukturgleichungsmodelle, Verfahren der multidimensionalen Skalierung und Mischverteilungsmodelle. Die anschließend dargestellten Modelle der Veränderungsmessung und Zeitreihenanalysen

werden insbesondere Längsschnittdaten gerecht. Schließlich werden Resampling-Verfahren, die keine restriktiven Verteilungsannahmen erfordern, dargestellt sowie statistische Methoden bei unvollständigen Daten, die gerade bei komplexen empirischen Studien zu erwarten sind.

Auch der zweite Band der Enzyklopädie Evaluationsforschung enthält Übersichten zum Stand der Evaluationsforschung in den wichtigsten psychologischen Disziplinen, wie der Klinischen, der Pädagogischen sowie der Arbeits- und Organisationspsychologie. Daneben richten sich weitere Kapitel auf methodische Grundlagen, wie z. B. Stichprobenmodelle, Versuchsplanung oder spezifische Skalierungsmethoden. Da nicht alle methodischen Gebiete der Evaluationsforschung in einem Band untergebracht werden konnten, sind beide Bände der Enzyklopädie Evaluationsforschung als eine Einheit zu betrachten.

Für alle Beiträge in diesem Band konnten führende Expertinnen und Experten auf dem jeweiligen Gebiet gewonnen werden. Ihnen allen möchte ich hier für die Mitwirkung an diesem Werk ganz herzlich danken. Sie haben nicht nur einen umfassenden Beitrag verfasst, sondern in der Regel auch zu zwei weiteren Beiträgen dieses Bandes ein Review übernommen.

Weiterhin gilt mein Dank Frau Margret Unger für ihre wertvolle Unterstützung bei der Koordination der Arbeiten zu diesem Band, die Überprüfung der Einhaltung der Formalien bei allen Beiträgen sowie die Übernahme der Korrektur von Manuskripten. Ebenfalls danke ich Frau Weidinger und Frau Fries vom Hogrefe Verlag für ihre kompetente Unterstützung bei der Entstehung dieses Bandes.

Münster, im März 2008

Heinz Holling

Inhaltsverzeichnis

1. Kapitel: Grundlagen der Evaluationsforschung Von Heinz Holling

1	Einleitung	1
2	Definitionen von Evaluation und Evaluationsforschung	2
3	Zwei typische Beispiele für Evaluationsforschung	5
3.1	Eine Evaluationsstudie zu Personalentwicklungsmaßnahmen	5
3.2	Eine Evaluationsstudie zur Schulzeitverkürzung	7
4	Gegenstand psychologischer Evaluationsforschung	9
5	Ziele von Evaluationsstudien	11
6	Theoretische Grundlegung psychologischer Evaluationsstudien	13
6.1	Kausalität	14
6.2	Generalisierung	15
7	Gütekriterien von Evaluationsstudien	17
7.1	Allgemeine Standards der Evaluationsforschung	17
7.2	Validität von Evaluationsstudien	19
7.2.1	Statistische Validität	21
7.2.2	Interne Validität	23
7.2.3	Konstruktvalidität	24
7.2.4	Externe Validität	24
8	Nutzenmessung	25
9	Spezifische Bedingungen von Evaluationsstudien	28
9.1	Die Bedeutung der Stakeholder	28
9.2	Formative und summative Evaluation	30
10	Fazit	31
	Literatur	31

2. Kapitel: Wissenschaftstheoretische Grundlagen der Evaluationsforschung

Von Hans Westmeyer

1	Aufgaben der Wissenschaftstheorie	35
2	Zur logischen Struktur des Evaluationsbegriffs	36
3	Handlungsregeln und Evaluationshypothesen	39
4	Zum empirischen Gehalt von Evaluationshypothesen	41
4.1	Ein Beispiel für unterschiedliche begriffliche Einbettungen einer Evaluationshypothese	42
4.2	Ein Beispiel für unterschiedliche methodische Einbettungen einer Evaluationshypothese	44
4.3	Zum Problem der Invarianz des empirischen Gehalts	45
5	Zum Problem der Generalisierbarkeit von Evaluationsresultaten	46
6	Auffassungen über die Beschaffenheit der Welt	49
7	Auffassungen über die Erkennbarkeit der Welt	50
8	Zur Funktion von Theorien in der Evaluationsforschung	53
	Literatur	55

3. Kapitel: Evaluationsmodelle

Von Werner W. Wittmann

1	Einleitung	59
2	Das Modell von Posavac und Carey	63
3	Konzeptuelle Beiträge aus der Urteils- und Entscheidungsforschung zum Indikationsproblem: die Matrix von Kenneth Hammond als ein bisher vernachlässigtes Evaluationsmodell	66
4	Das CIPP-Modell von Daniel Stufflebeam	68
5	Die Konzeption der fünf Datenboxen als umfassendes Evaluationsmodell	71
6	Varianten des Bewertungsfokus unterschiedlicher Evaluationsmodelle	84
7	Fazit	91
	Literatur	92

4. Kapitel: Phasen der Evaluationsforschung

Von Nicola Döring

1	Phasenmodelle in Wissenschaft und Praxis	99
1.1	Phasenmodelle des Forschungsprozesses	99
1.2	Phasenmodelle für Projekte	101
1.3	Phasenmodelle für Evaluationsforschungsprojekte	103
2	Initiierungsphase der Evaluation	105
2.1	Identifikation von Evaluationsmöglichkeiten	105
2.2	Prüfung der Evaluationsziele	106
3	Konzeptionsphase der Evaluation	110
3.1	Exploration der Evaluationsbedingungen	110
3.2	Entwicklung des Evaluationskonzepts	111
3.3	Einbeziehung der Evaluationsstandards	113
3.4	Vorbereitung des Projektmanagements	114
3.5	Formulierung des Evaluationsvertrags	115
4	Planungsphase der Evaluation	116
4.1	Formulierung der Evaluationsfragen und -hypothesen	116
4.2	Untersuchungsplanung	118
4.3	Projektplanung	123
5	Realisierungsphase der Evaluation	123
5.1	Erhebung von Evaluationsdaten	123
5.2	Analyse von Evaluationsdaten	124
5.3	Interpretation von Evaluationsdaten	125
5.4	Projektcontrolling	126
6	Abschlussphase der Evaluation	126
6.1	Abgabe der Evaluationsberichte	127
6.2	Organisation der Evaluationsnutzung	128
6.3	Durchführung einer Metaevaluation	128
6.4	Publikation der Evaluationsstudie	129
7	Schlussfolgerung	131
	Literatur	131

5. Kapitel: Facettentheorie

Von Wolfgang Bilsky und David Cairns

1	Grundidee der Facettentheorie (FT)	135
2	Beispiel: Eine Delphi-Studie	137

3	FT-Design	142
3.1	Beobachtungen, Facetten, Elemente und Struktupel	142
3.2	Personen-, Inhalts- und Reaktionsfacetten	143
3.3	Abbildungssatz und gemeinsamer Bildbereich	144
3.4	Universum und Stichprobe	146
4	FT-Datenanalyse	147
4.1	Regionale Hypothesen und konfirmatorische Analysen	147
4.2	Die Rolle von Facetten für die Entwicklung und Prüfung von Strukturmodellen	149
5	Der Abbildungssatz als Strukturierungshilfe	154
5.1	Programmevaluation	154
5.2	Mitarbeiterbefragungen	157
6	Forschungsstand und Perspektiven	159
	Literatur	160

6. Kapitel: Aufstellen und Prüfen von Hypothesen Von Rainer Westermann und Matthias Siemer

1	Einleitung	163
2	Substanzielle Hypothesen in der Evaluation	164
2.1	Evaluationsbegriffe	164
2.2	Evaluierte Objekte	165
2.3	Ziele von Maßnahmen und Evaluationen	166
2.4	Beteiligte Interessengruppen	167
2.5	Kontrollbedingungen und Vergleichsmaßnahmen	168
2.6	Wissenschaftliche und praktische Ausrichtung	169
2.7	Evaluationsparadigmen	171
3	Standardmodelle zur Prüfung von Evaluationshypothesen	172
3.1	Empirische Hypothesen und Vorhersagen	173
3.2	Validitäts- und Störungshypothesen	174
3.3	Statistische Hypothesen und Effekte	178
4	Weitere Ansätze zur Analyse von Evaluationshypothesen	184
4.1	Hypothesen über zufällige und feste Effekte und ihre statistisch valide Überprüfung	184
4.1.1	Statistische Modelle	185
4.1.2	Zufällige und feste Effekte und generalisierende statistische Inferenz	188
4.1.3	Zufällige und feste Effekte nach randomisierter und nicht randomisierter Zuweisung	189

4.1.4 Umgang mit eingesteten Faktoren ohne zufällige Ziehung und Randomisierung	190
4.1.5 Gekreuzte Faktoren	192
4.2 Evaluationshypothesen auf unterschiedlichen Ebenen der Aggregation	192
4.3 Hypothesen zur efficacy und effectiveness einer Maßnahme und die Gütekriterien der internen und externen Validität	195
4.4 Parametrische und nicht parametrische Hypothesen und die Kontrolle der Teststärke in Randomisierungstests	197
Literatur	201

7. Kapitel: Generalisierbarkeitstheorie Von Thorsten Stumpp und Heiko Großmann

1 Einleitung	207
2 Grundbegriffe	208
3 Generalisierbarkeitsstudie	210
4 Planung von Entscheidungsstudien	216
5 Gütemaße für D-Studien	220
6 Design von G- und D-Studien	225
6.1 Gekreuzte und verschachtelte Designs	226
6.2 Zufällige und Feste Facetten	227
7 Anwendung der Generalisierbarkeitstheorie in der Evaluationsforschung ...	228
8 Softwareprogramme zur Generalisierbarkeitstheorie	230
Literatur	231

8. Kapitel: Metaanalyse in der Evaluationsforschung Von Georg E. Matt

1 Einleitung	235
2 Die Suche nach generalisierbaren kausalen Beziehungen	236
2.1 Deskriptive kausale Beziehungen	236
2.2 Domänen der Generalisierbarkeit	237
2.3 Kernfragen der Generalisierbarkeit	237
2.4 Traditionelle Begründungen von Generalisierbarkeitsansprüchen ...	238

2.5	Cooks (1990, 1993) „Grounded Theory“ generalisierter kausaler Schlussfolgerungen	239
3	Metaanalyse zur Generalisierung kausaler Beziehungen	241
3.1	Explikation der Generalisierungsdomänen: Das Thema	241
3.2	Selektion von Studien	243
3.3	Beschreibung und Kodierung von Studien	245
3.4	Berechnung von Effektstärken	247
3.4.1	Wahl des Effektindikators, T	247
3.4.2	Berechnung von Effekten innerhalb einer Studie	248
3.5	Die statistische Analyse metaanalytischer Daten	252
3.5.1	Fixed-Effekt- und Random-Effekt-Modelle	252
3.5.2	Fixed-Effekt-Modell: Ohne Prädiktoren	255
3.5.3	Random-Effekt-Modell: Ohne Prädiktoren	256
3.5.4	Fixed-Effekt-Modell mit Prädiktoren	258
3.5.5	Random-Effekt- und Mixed-Effekt-Modelle mit Prädiktoren	261
3.5.6	Alternative und komplementäre Entwicklungen	262
3.6	Die Interpretation und Präsentation metaanalytischer Ergebnisse	263
3.6.1	Sensitivitätsanalysen	263
3.6.2	Artefaktkorrekturen	267
3.6.3	Grafische Darstellungen	268
3.7	Validität metaanalytischer Schlussfolgerungen	273
3.7.1	Validität statistischer Schlussfolgerungen	274
3.7.2	Validität der kausalen Schlussfolgerungen	274
3.7.3	Validität der Konstrukte	275
3.7.4	Validität generalisierter Schlussfolgerungen	276
4	Schlussfolgerungen	276
	Literatur	277

9. Kapitel: Das Allgemeine Lineare Modell in der Evaluationsforschung

Von Helfried Moosbrugger und Wolfgang Rauch

1	Einleitung und Überblick	285
1.1	Statistische Modelle zur Evaluation von Effekten	286
1.2	Zur Bewertung von Untersuchungsergebnissen: Effektgrößen, Konfidenzintervalle und Teststärke	287
1.2.1	Effektgrößen zur Wirksamkeitsüberprüfung	287
1.2.2	Konfidenzintervalle zur Beurteilung der Genauigkeit einer Schätzung	288
1.2.3	Teststärke	290

2	Das Allgemeine Lineare Modell (ALM)	291
2.1	Grundsätzliche Formulierung des ALM	292
2.1.1	Modellgleichung	292
2.1.2	Parameterschätzung im ALM	293
2.1.3	Verteilungsannahmen im ALM	295
2.1.4	Varianzzerlegung	296
2.1.5	Eingeschränkte und uneingeschränkte Modelle	298
2.1.6	Hypothesentests	299
2.1.7	Hypothesentest mit der Allgemeinen Linearen Hypothese (ALH)	300
2.1.8	Tests und Konfidenzintervalle für Modellparameter	301
2.1.9	Test einer Subgruppe von Parametern	301
2.1.10	Effektgrößen im ALM	305
2.2	Überlegungen zur Stichprobengröße und Teststärke	306
2.3	Überprüfung der Verteilungsvoraussetzungen, Modelldiagnose und Multikollinearität	307
2.3.1	Verteilungsvoraussetzungen	307
2.3.2	Ungewöhnliche Merkmalsausprägungen	311
2.3.3	Multikollinearität	313
3	Regressionsanalyse	315
3.1	Effektgrößen und Hypothesentests in der Regressionsanalyse	315
3.2	Vorhersage neuer Werte	316
3.3	Nichtlineare Zusammenhänge	317
3.3.1	Kurvilineare Zusammenhänge	317
3.3.2	Moderatormodelle	318
4	Varianzanalyse	319
4.1	Kodiervariablen	319
4.2	Effektgrößen für Varianzanalysen	320
4.3	Strategien zur Erhöhung der Teststärke	322
4.3.1	Verringerung der Fehlervarianz bei varianzanalytischen Designs	323
4.3.2	Erhöhung der Primärvarianz	323
5	Kovarianzanalyse	324
5.1	Kovariate als Kontrollvariable	325
5.2	Kovariate als substanzielle Variable	327
6	Zusammenfassung und Ausblick	328
	Literatur	329

10. Kapitel: Hierarchische Lineare Modelle (HLM)

Von Kai S. Cortina und Hans Anand Pant

1	Einleitung	335
2	Clustersampling in der Evaluationsforschung	336
3	Das Zwei-Ebenen-HLM-Modell	337
	3.1 Intraklassenkorrelation und Design-Effekt	338
	3.2 Formales HLM-Gleichungssystem	339
4	Beispiel eines Zwei-Ebenen-Modells	341
5	Modellspezifikation und -optimierung	343
6	Schätzalgorithmen und Software	345
7	Nicht normalverteilte und nicht intervallskalierte Variablen	346
8	Analyse von Längsschnittdaten im Rahmen von HLM	347
9	Metaanalyse mittels HLM	352
	9.1 Vorzüge der Metaanalyse mit HLM	354
	9.2 Erweiterungen des HLM-Metaanalysemodells	355
	9.3 Bayes'scher Ansatz der hierarchischen Metaanalyse	357
10	Resümee	359
	Literatur	360

11. Kapitel: Generalisierte Lineare Modelle

Von Britta Colver und Heinz Holling

1	Einleitung	363
2	Grundlagen generalisierter linearer Modelle	364
	2.1 Das allgemeine lineare Modell als Ausgangspunkt	364
	2.2 Verallgemeinerung des allgemeinen linearen Modells	366
	2.2.1 Die Linkfunktion	366
	2.2.2 Verteilungen aus der Exponentialfamilie	367
	2.3 Definition generalisierter linearer Modelle	368
	2.4 Maximum-Likelihood-Schätzung der Parameter im generalisierten linearen Modell	369
	2.5 Modellfit	370
	2.6 Anwendungssoftware	372
3	Das allgemeine lineare Modell	373
	3.1 Das Modell	373
	3.2 Modellfit	374
	3.3 Beispiel	375

4	Die logistische Regression	377
4.1	Das Modell	377
4.2	Modellfit	379
4.3	Beispiel	379
5	Die logistische Regression für nominale und ordinale Variablen	381
5.1	Das Modell	382
5.1.1	Nominale abhängige Variablen	382
5.1.2	Ordinale abhängige Variablen	383
5.2	Modellfit	384
5.3	Beispiel	384
6	Das loglineare Modell	386
6.1	Das Modell	387
6.2	Modellfit	388
6.3	Beispiel	389
7	Abschließende Bemerkungen	391
	Literatur	392

12. Kapitel: Strukturgleichungsmodelle

Von Rolf Steyer, Andreas Wolf, Friedrich Funke und
Ivailo Partchev

1	Einleitung	393
2	Pfadmodelle mit Observablen	395
3	Modelle der Klassischen Testtheorie	397
3.1	Grundbegriffe der Klassischen Testtheorie	398
3.2	Das Modell τ -äquivalenter Variablen	400
3.3	Weitere Modelle der KTT	402
3.4	Weiterführende Literatur zur KTT	404
4	Latent-state-trait-Modelle	405
4.1	Grundbegriffe	405
4.2	Das Multistate-Modell	410
4.3	Das Multistate-singletrait-Modell	411
4.4	Weiterführende Literatur	414
5	Latent-change-Modelle und Wachstumskurven-Modelle	415
5.1	Latent-change-Modelle	415
5.2	Wachstumskurven-Modelle	417
6	Multitrait-multimethod-Modelle	418
6.1	Ein Multitrait-multimethod-Modell	419
6.2	Weiterführende Literatur	423

7	Faktorenanalytische Modelle	423
7.1	Darstellung des Modells τ -äquivalenter Variablen in Matrizenform ...	425
7.2	Das Multistate-Modell in Matrizenform	426
7.3	Das Multistate-singletrait-Modell in Matrizenform	427
7.4	Das Multitrait-multimethod-Modell in Matrizenform	428
7.5	Weiterführende Literatur	429
8	Strukturgleichungsmodelle	429
8.1	Parameterschätzung	431
8.2	Modellbewertung	434
9	Ausblick	439
9.1	Weitere Anwendungsbereiche von Strukturgleichungsmodellen	440
9.2	Aktuelle Entwicklungsrichtungen von SEM	441
9.3	Literatur und Informationsquellen über SEM	442
	Literatur	442

13. Kapitel: Multidimensionale Skalierung Von Ingwer Borg

1	Einleitung	449
2	Grundideen der MDS	449
2.1	MDS als psychologisches Modell	449
2.2	MDS zur Visualisierung von Datenstrukturen	452
3	Wichtige technische Aspekte der MDS	454
3.1	Güte der MDS-Darstellung	454
3.2	Bewertung des Stress	456
3.3	MDS-Modelle	458
3.4	MDS-Algorithmen	460
3.5	MDS-Algorithmen und degenerierte Lösungen	461
4	Interpretationsansätze in der MDS	463
4.1	Dimensionen, Richtungen, Regionen und Cluster	463
4.2	MDS-Interpretation mit Hilfe anderer Verfahren	466
5	Prokrustische Transformationen	468
6	Individuelle Unterschiedsmodelle	470
7	Konfirmatorische MDS	471
8	Unfolding	473
8.1	Triviale Lösungen im Unfolding	475
8.2	Mehrdimensionales vs. multiples Unfolding	477
9	MDS und verwandte Verfahren	477
	Literatur	479

14. Kapitel: Mischverteilungsmodelle Von Jürgen Rost und Michael Eid

1	Einleitung	483
2	Die „latent class analysis“ (LCA)	489
	2.1 Die LCA mit Gleichheitsrestriktionen	492
	2.2 Die linear-logistische LCA	497
	2.3 LCA für ordinale Daten	499
	2.4 Das Mixed-Rasch-Modell	501
3	Gemischte Strukturgleichungsmodelle	503
	3.1 Modellerweiterungen	507
	3.2 Mischverteilungsfaktorenanalyse	508
	3.3 Mischverteilungs-Latent-State-Trait-Modelle	509
	3.4 Mischverteilungs-Wachstumskurvenmodelle	511
	3.5 Profilanalyse	514
4	Modelltests und Bestimmung der Klassenanzahl	515
5	Ausblick	519
6	Software	519
	Literatur	520

15. Kapitel: Veränderungsmessung Von Karl Christoph Klauer

1	Einleitung	525
2	Veränderungsmessung: Voraussetzungen und Eingrenzung des Gegenstands- bereichs	526
3	Der Schluss auf die Ursache einer Veränderung	530
4	Die Modellierung individueller Veränderungsverläufe bei vielen Mess- zeitpunkten	535
5	Psychometrische Eigenschaften von Veränderungsindizes	538
6	Veränderungsmessung als Schätzung der Parameter individueller Wachstumskurven	543
	6.1 Parametrisierte Familien von Wachstumskurven	543
	6.2 Statistische Behandlung: Hierarchische lineare Modelle	547
	6.3 Das hierarchische lineare Modell für zwei Messzeitpunkte	548
	6.4 Mehr als zwei Messzeitpunkte: Reliabilität und Hypothesentests	550
	6.5 Hierarchisches Lineares Modell: Verallgemeinerungen	552

6.6	Zeitstrukturierte Daten: Statistische Analyse im Rahmen der Kovarianzstrukturanalyse	553
7	Beiträge der Probabilistischen Testtheorie zur Veränderungsmessung	554
7.1	LLTM und LLRA	556
7.2	Probabilistische Modelle mit interindividueller Varianz in Veränderungen	557
8	Fazit	559
	Literatur	560

16. Kapitel: Zeitreihenanalyse

Von Bernhard Schmitz, Franziska Perels und Martin Löb

1	Einführung	565
2	Grundbegriffe	567
2.1	Grafische Darstellung	568
2.2	Die Lag-Funktion	568
2.3	Autokorrelation	569
2.4	Partialautokorrelation	570
2.5	Stationarität	570
3	Designvarianten	571
3.1	Einzelfallanalyse	571
3.2	AB-Design	573
3.3	Multiples Baselinedesign	574
3.4	Zeitreihendesign	577
3.5	Zeitreihendesign mit Kontrollgruppe	578
3.6	Weitere Designs	578
4	Univariate Zeitreihenanalysen: Grundlagen	578
4.1	Glättung	579
4.2	Trends	579
4.2.1	Linearer Trend	579
4.2.2	Polynomialer Trend	580
4.3	Trendbereinigung	581
4.4	Spektralanalyse	582
5	ARIMA-Modelle	582
5.1	AR-Modell	583
5.1.1	AR(1)-Modell	583
5.1.2	AR(p)-Modell	585
5.2	MA-Modell	585
5.2.1	MA(1)-Modell	586
5.2.2	MA(q)-Modell	587
5.3	ARIMA-Modell	587

5.4	Vorgehen bei einer Modellidentifikation	587
5.4.1	Identifikation des ARIMA-Modelltyps der Zeitreihe	588
5.4.2	Schätzung der Modellparameter	589
5.4.3	Diagnose des Modells	589
6	Interventionsanalysen	589
6.1	Grundlage	590
6.2	Puls- und Stufeninput	590
6.3	Interventionswirkung und Transferfunktion	591
6.4	Noise-Modell	593
6.5	Vorgehen bei der Interventionsanalyse	593
7	Multivariate Zeitreihenanalyse	594
7.1	Kreuzkorrelationen	596
7.2	Prewhitening	597
7.3	Multivariate AR-Modelle	598
8	Ausblick	601
	Literatur	603

17. Kapitel: Resampling

Von Günther Gediga

1	Einleitung	607
2	Sampling und Resampling	608
2.1	Sampling einer einfachen Stichprobe	608
2.2	Verteilungsgebundenes Monte-Carlo-Resampling	609
2.3	Verteilungsfreies Resampling	611
3	Resampling-Techniken	612
3.1	Bootstrap und Jackknifing	612
3.1.1	Die Grundidee des Bootstrap-Verfahrens	613
3.1.2	Bootstrap-Standardfehler	614
3.1.3	Bootstrap-Bias	615
3.1.4	Naive Bootstrap-Perzentile	616
3.1.5	Korrigierte Bootstrap-Konfidenzintervalle	616
3.1.6	Jackknife	618
3.1.7	Voraussetzungen	618
3.1.8	Methodenvergleiche	619
3.2	Permutationstests	620
3.2.1	Die Grundidee der Permutationstests	620
3.2.2	Berechnung der Verteilung einer Teststatistik unter Randomisierungsannahme	622
3.2.3	Voraussetzungen des Permutationstests	623
3.3	Kreuzvalidierung	624

3.4	Modellierung durch die nächsten Nachbarn	625
3.5	Resampling-Ideen: Eine Zusammenfassung	627
4	Wichtige Anwendungsbereiche	629
4.1	Regression	629
4.2	Varianzanalyse	632
4.3	Diskriminanzanalyse	634
4.4	Einzelfallanalyse	635
4.5	Behandlung fehlender Daten	636
4.6	Modellauswahl	638
4.7	Weitere Anwendungsfelder	639
5	Programme mit und für Resampling-Verfahren	639
6	Ausblick	641
	Literatur	642

18. Kapitel: Statistische Methoden bei unvollständigen Daten

Von Helge Toutenburg, Christian Heumann und
Thomas Nittner

1	Einführung	649
1.1	Ein einfaches Beispiel	650
1.2	Fehlendmuster	652
1.3	Fehlendmechanismen	654
1.4	Hinweise und weiterführende Literatur	655
2	Allgemeine Verfahren	656
2.1	Complete Case Analysis (CCA)	656
2.2	Available Case Analysis (ACA)	657
2.3	Imputation	658
2.3.1	Mittelwertimputation	658
2.3.2	Bedingte Mittelwertimputation	660
2.3.3	Single Imputation	660
2.3.4	Nearest Neighbour Imputation	661
2.3.5	Weitere Verfahren	662
2.4	Verteilungsbasierte Modelle	662
2.5	Hinweise und weiterführende Literatur	663
3	Inferenz basierend auf der Likelihood	663
3.1	Maximum Likelihood (ML)	665
3.2	Pattern Mixture-Modelle	666
3.3	Selection-Modelle	668
3.4	Äquivalenz von Pattern Mixture-Modellen und Selection-Modellen in speziellen Fällen	669
3.5	Parametrische Verteilungsannahmen im multivariaten Fall	671

3.6	EM-Algorithmus	672
3.7	Bayes-Inferenz	676
3.8	Hinweise und weiterführende Literatur	677
4	Modelle vom Regressionstyp	679
4.1	Lineares Modell	679
4.1.1	Fehlende Daten im Response	679
4.1.2	Fehlende Werte in der X -Matrix	683
4.2	Nicht parametrische und semiparametrische Modelle	684
4.3	Hinweise und weiterführende Literatur	686
	Literatur	688
	Autorenregister	693
	Sachregister	709

1. Kapitel

Grundlagen der Evaluationsforschung

Heinz Holling

1 Einleitung

Die Evaluationsforschung ist eine relativ junge Disziplin der Psychologie und geht in ihrer heutigen Form insbesondere auf die in den sechziger Jahren vorwiegend in den USA durchgeführten wissenschaftlichen Begleituntersuchungen sozial- und bildungspolitischer Reformansätze zurück (vgl. Beywl, Speer & Kehr, 2004; Rossi, Lipsey & Freeman, 2004). Hier wurden Sozialwissenschaftler in die Beratung, Gestaltung und Erfolgskontrolle komplexer, interdisziplinärer Programme einbezogen. Diese häufig politisch motivierten Programme, wie z. B. das *War on Poverty*-Programm des US-Präsidenten Lyndon B. Johnson, boten zahlreichen Sozialwissenschaftlern die Möglichkeit, die Effektivität und Effizienz der durchgeführten Maßnahmen (z. B. Ausbildungsoffensive für ökonomisch benachteiligte Bevölkerungsgruppen) zu analysieren.

Der Schwerpunkt dieser frühen Evaluationsstudien lag größtenteils auf der zielorientierten Wirksamkeitsüberprüfung anhand sozialwissenschaftlicher Methoden (vgl. Campbell, 1969). Später wurden erweiterte Evaluationsansätze entwickelt, die auch mögliche, in der Zielformulierung nicht explizierte negative Auswirkungen sozialer Programme sowie die Bedürfnisse betroffener Zielgruppen berücksichtigen (vgl. Scriven, 1991). Mittlerweile hat sich die Evaluationsforschung zu einer recht differenzierten eigenständigen Disziplin entwickelt, die in vielen gesellschaftlichen Bereichen Anwendung findet, so z. B. im Bildungs- und Gesundheitswesen, der Verkehrs- und Umweltpolitik oder im Städtebau und Justizvollzug.

Im ökonomischen Bereich hat Evaluation seit langem insbesondere unter dem Begriff „Controlling“ einen festen Platz. Controlling dient dazu, Prozesse

und Projekte während der Umsetzung zu kontrollieren, um sie gegebenenfalls bei auftretenden Schwierigkeiten sofort in die gewünschte Richtung zu steuern. So wird Transparenz geschaffen und die Effektivität der Maßnahmen sichergestellt. Dabei ist das Ziel eines Unternehmens relativ klar: Es ist seine Wirtschaftlichkeit, eindeutig erfassbar über Zahlen. In psychologischen Anwendungsfeldern ist die Ziel- und Nutzenbestimmung in aller Regel weitaus schwieriger.

Was ist nun unter Evaluation und Evaluationsforschung im Rahmen der Psychologie zu verstehen? Um diese Frage zu klären, zitieren wir zunächst einige typische Definitionen und stellen zwei prototypische Evaluationsstudien dar. Dann versuchen wir zu klären, was Gegenstand und Ziele von Evaluation bzw. Evaluationsforschung sind. Anschließend gehen wir auf wesentliche Probleme zu theoretischen und methodischen Grundlagen der Evaluationsforschung ein.

2 Definitionen von Evaluation und Evaluationsforschung

Es gibt zahlreiche unterschiedliche Definitionen der Begriffe Evaluation und Evaluationsforschung. Weiterhin wird mehr oder weniger synonym zum Begriff Evaluation eine Reihe weiterer Begriffe verwendet. So spricht man je nach dem spezifischen Kontext von Qualitätskontrolle, Erfolgskontrolle, Effizienzforschung usw. Zudem trifft man auch auf Begriffe, die auf eine spezielle Form von Evaluation Bezug nehmen, im schulischen Bereich wird beispielsweise häufig von Schulbegleitforschung gesprochen.

Ganz allgemein kann Evaluation als die Bewertung von Sachverhalten definiert werden. Um diesen Begriff bzw. den Begriff *Evaluationsforschung* im Rahmen der Psychologie einzugrenzen, betrachten wir zunächst einige repräsentative Definitionen.

Gemäß einer der ältesten Definitionen von Suchman (1967) ist Evaluation der Prozess der Beurteilung des Wertes eines Produktes, Prozesses oder eines Programms, wobei nicht notwendigerweise systematische Verfahren oder datengestützte Beweise erforderlich sind. Evaluationsforschung zeichne sich dagegen durch die explizite Verwendung wissenschaftlicher Forschungsmethoden und -techniken für den Zweck der Durchführung einer Bewertung aus. Rossi und Freeman (1993) bestimmen in ihrer viel zitierten Definition Evaluationsforschung als die systematische Anwendung empirischer Forschungsmethoden zur Bewertung des Konzeptes, des Untersuchungsplanes, der Implementierung und der Wirksamkeit sozialer Interventionsprogramme. Dieser Definition schließen sich auch Bortz und Döring (2006, S. 676) an, wobei sich

ihrer Meinung nach Evaluationsforschung im weiteren Sinne nicht nur mit sozialen Interventionsprogrammen befasst, sondern auch mit einer Vielzahl anderer Objekte, wie Personen, Umweltfaktoren, Produkte oder Techniken/Methoden. Damit umfasst Evaluationsforschung alle forschenden Aktivitäten, bei denen es um die Bewertung des Erfolges von gezielt eingesetzten Maßnahmen oder um Auswirkungen von Wandel in Natur, Kultur, Technik und Gesellschaft geht.

Wottawa und Thierau (2003, S. 14) halten es angesichts der bestehenden Definitionsvielfalt für sinnvoller, statt eines weiteren Definitionsversuchs, die allgemeinen Kennzeichen wissenschaftlicher Evaluation herauszuarbeiten.

- Ein allgemeiner Konsens, der hier auch schon durch die Wortwurzel von ‚Evaluation‘ vorgezeichnet ist, liegt darin, dass alle solche Tätigkeiten etwas mit ‚Bewerten‘ zu tun haben. Evaluation dient als Planungs- und Entscheidungshilfe und hat somit etwas mit der Bewertung von Handlungsalternativen zu tun (vgl. Wottawa, 1986).
- Evaluation ist ziel- und zweckorientiert. Sie hat primär das Ziel, praktische Maßnahmen zu überprüfen, zu verbessern oder über sie zu entscheiden.
- Es besteht im wissenschaftlichen Sprachgebrauch ebenfalls ein Konsens darüber, dass Evaluationsmaßnahmen dem aktuellen Stand wissenschaftlicher Techniken und Forschungsmethoden angepasst sein sollten“.

Mittag und Hager (2000, S. 103) explizieren den Begriff Evaluationsforschung dahingehend, dass sie „damit die wissenschaftlich fundierte, empirische und hypothesenorientierte Forschung unter systematischer Anwendung sozialwissenschaftlicher Forschungsmethoden meinen. Die Ergebnisse dieser Forschung bilden die wesentliche, wenn auch nicht die einzige Grundlage einer wissenschaftlichen Evaluation oder Bewertung der Konzeption, Ausgestaltung, Umsetzung und des Nutzens sozialer und psychologischer Interventionsprogramme“.

Umfangreichere Maßnahmen werden im Rahmen der Evaluationsforschung zumeist als Programme bezeichnet. Dementsprechend spricht man dann auch von Programmevaluation. Gemäß einer häufig zitierten Definition von Attkisson und Broskowski (1978) ist Programmevaluation ein Prozess der Durchführung vernunftgeleiteter Beurteilungen eines Programms hinsichtlich Aufwand, Effektivität und Angemessenheit auf der Grundlage systematischer Datenerhebung und Datenanalyse.

Die oben zitierten Definitionen stimmen zwar weitgehend im Kern überein, unterscheiden sich aber bei der weiteren Spezifikation der Kriterien zum Gegenstand oder zu den Zielen von Evaluation bzw. Evaluationsforschung. Zunächst

einmal kann Evaluation ganz allgemein als Bewertung jeglicher Sachverhalte bezeichnet werden. Gemäß einer solch allgemeinen Definition ist prinzipiell jeder alltägliche Bewertungsprozess als Evaluation anzusehen, so z. B. die Ermittlung des Geldbetrags in einer Geldbörse, die Benotung eines Schulaufsatzes oder das Schreiben einer Literaturkritik. Grenzt man nun zunächst alltägliche Evaluationen von wissenschaftlichen Evaluationen ab, so sind diese durch die Verwendung wissenschaftlich anerkannter Methoden und Verfahren gekennzeichnet. Wissenschaftliche Evaluation wird daher auch als Evaluationsforschung bezeichnet. Damit meint dieser Terminus nicht, wie man leicht folgern könnte, Forschung über Evaluationen.

Wie lässt sich das Feld *wissenschaftliche Evaluation bzw. Evaluationsforschung* weiter eingrenzen? Dazu wurde in den zitierten Definitionen eine Vielzahl weiterer Kriterien genannt, die wir in den folgenden Abschnitten im Einzelnen diskutieren werden. Zunächst gilt es, den Gegenstandsbereich zu spezifizieren. Hier gehen die obigen Definitionen von unterschiedlichen Sachverhalten aus, so der Beurteilung von z. B. allen möglichen Sachverhalten, gesellschaftlichem oder kulturellem Wandel oder des Wertes von Produkten, Prozessen sowie Programmen bzw. Interventionen. In anderen Definitionen wird spezifischer lediglich der Wert bzw. die Effektivität von Interventionen als Gegenstand von Evaluation aufgeführt oder noch spezifischer die Bewertung der Konzeption, Ausgestaltung, Umsetzung und des Nutzens sozialer und psychologischer Interventionsprogramme.

Weiterhin unterscheiden sich die Definitionen im Wesentlichen hinsichtlich der Ziele von Evaluationsstudien. Geht es nur um die Bewertung oder aber auch notwendigerweise um bestimmte Entscheidungen über die Beibehaltung bzw. Absetzung von Interventionen und/oder um die Verbesserung von Maßnahmen?

Um die folgenden Ausführungen anschaulicher zu gestalten, soll auf zwei prototypische psychologische Evaluationsstudien Bezug genommen werden. In den drei grundlegenden Anwendungsfächern der Psychologie, Klinische Psychologie, Arbeits- und Organisationspsychologie sowie Pädagogische Psychologie sind zahlreiche Evaluationsstudien durchgeführt worden. Wir wählen aus den beiden zuletzt genannten Gebieten jeweils eine Studie aus, da in dem Beitrag von Westmeyer (in diesem Band) sehr ausführlich auf die Evaluationsforschung in der Klinischen Psychologie eingegangen wird. Die beiden ausgewählten Studien erfüllen weitgehend alle Kriterien der oben angeführten Definitionen und können als Prototyp von Evaluationsstudien gelten.

3 *Zwei typische Beispiele für Evaluationsforschung*

3.1 Eine Evaluationsstudie zu Personalentwicklungsmaßnahmen

Untersuchungsgegenstand einer Studie von Morrow, Jarrett und Rupinski (1997) war die Effektivität verschiedener Trainingsformen für Arbeitnehmer eines großen US-amerikanischen Unternehmens. Den Ausgangspunkt bildete dabei die Forderung eines neuen Vorstandsvorsitzenden nach einer Beurteilung aller 18 im Unternehmen eingesetzten Trainingsmethoden unter Kosten-Nutzen-Aspekten (vgl. Tab. 1). Diese Analysen dienten zur Vorbereitung der Entscheidungen, inwieweit die Trainingsmaßnahmen beizubehalten oder zu modifizieren seien und welches Budget den einzelnen Trainingsmaßnahmen zukommen solle. Nachdem die Personalabteilung glaubte, sich dieser Aufgabe – wie schon häufiger – mit einem „anekdotisch“ verfassten Bericht entziehen zu können, bestand jedoch der Vorstandsvorsitzende, der mit der monetären Nutzenanalyse von Personalentwicklungsmaßnahmen vertraut war, auf die Implementierung dieser aussagekräftigen und wissenschaftlich anerkannten Methode.

Zur Erfassung der ökonomischen Konsequenzen der Trainings nutzten die Autoren das Nutzenanalyse-Modell von Raju, Burke und Normand (1990). Dieses Modell setzt im Wesentlichen multiattributive Nutzenanalysen aller Arbeitstätigkeiten sowie Metaanalysen zur Effektivität der einzelnen Trainingsformen voraus. Zur Durchführung dieser sehr aufwendigen Erhebungen wurden entsprechende empirische Studien über einen Zeitraum von vier Jahren mittels quasiexperimenteller Designs durchgeführt. Dabei war aufgrund organisatorischer Randbedingungen eine randomisierte Aufteilung von Mitarbeitern auf die verschiedenen implementierten Trainingsformen nicht möglich, da stets vollständige Kurse untersucht werden mussten. Ebenso konnten aus mehreren Gründen nicht bei allen durchgeführten Trainingsprogrammen Kontrollgruppen eingesetzt werden.

Die multiattributiven Nutzenanalysen der 18 Trainingsprogramme beruhten auf den Angaben von vorgesetzten, untergebenen oder gleichgestellten Unternehmensangehörigen. Zur Ermittlung der Effektstärken der einzelnen Trainings wurden weiterhin Kovarianzanalysen mit Messwiederholungen eingesetzt. Die in Tabelle 1 zusammengefassten Ergebnisse zeigen deutliche Unterschiede in der Wirkung (Effektstärke d) der verschiedenen Trainingsformen auf. Zwar hängt der monetäre Nutzen der einzelnen Trainings von der Effektstärke ab, aber auch Trainings mit einer eher geringen Effektstärke können in ökonomischer Hinsicht sehr erfolgreich sein, wenn sie z. B. sehr häufig eingesetzt werden (hier z. B. Product Sales-Trainings). Auf der Grundlage dieser Ergebnisse konnten dann Entscheidungen zu Modifikationen am gesamten Trainingsprogramm getroffen werden.

Tabelle 1:
Ergebnisse der monetären Nutzenanalyse von Morrow, Jarret und Rupinski (1997)

Titel	Z	N	A	SD _Y	d	G	K	MN	ROI	BE
Executive-2	M	18	0,39	\$20,24	-0,05	-\$16,10	\$271,0	-\$287,1	-105 %	Nie
Leadership skills	M	40	0,19	\$8,34	0,12	\$40,20	\$62,4	-\$22,2	-36 %	1,55
Supervisors-1	M	37	0,24	\$8,06	0,23	\$68,80	\$112,6	-\$43,8	-39 %	1,64
Managers-1	M	18	0,37	\$23,69	0,38	-\$18,30	\$73,2	-\$91,5	125 %	0,44
Managers-2	M	19	0,32	\$11,37	0,76	\$167,00	\$74,0	\$93,0	126 %	0,44
Middle mgrs.-2	M	18	0,51	\$24,30	1,11	\$483,20	\$81,6	\$401,6	492 %	0,17
Lab mgrs.-1	M	18	0,40	\$8,16	-0,09	-\$14,40	\$49,9	-\$64,3	-129 %	Nie
Oral presentations	M	8	0,32	\$11,23	0,26	\$22,60	\$13,4	\$9,2	69 %	0,59
Problem solving	M	14	0,36	\$11,26	0,31	\$47,90	\$41,4	\$6,5	16 %	0,86
Project mgmt.	M	20	0,15	\$4,95	0,37	\$37,20	\$23,2	\$14,0	60 %	0,52
Team bldg.	M	34	0,18	\$6,33	0,12	\$28,20	\$199,2	-\$171,0	-86 %	7,11
Product sales	V	1.100	0,17	\$7,63	0,67	\$5.488,70	\$4.136,5	\$1352,2	33 %	0,75
Territory mgmt.	V	19	0,21	\$9,73	0,54	\$98,00	\$52,9	\$45,1	85 %	0,54
Hazard energy	V	132	0,10	\$2,68	0,90	\$308,90	\$76,2	\$232,7	306 %	0,25
In-house time mgmt.	T	16	0,11	\$4,41	0,89	\$62,60	\$3,0	\$59,6	1.989 %	0,05
Off-shelf time mgmt.	T	88	0,15	\$5,72	0,28	\$142,60	\$69,1	\$73,5	106 %	0,48
Speciality valves	T	41	0,57	\$2,97	0,37	\$41,20	\$17,9	\$23,3	130 %	0,43
Written comm.	T	24	0,13	\$4,51	1,07	\$115,30	\$30,7	\$84,6	275 %	0,37

Anmerkungen:

Titel: Bezeichnung des Managementtrainings wie von Morrow et al. (1997) angegeben.
Weitere Erläuterungen wurden nicht gegeben.

Z: Zielgruppe der Trainings: M = Management, V = Verkaufspersonal, T = Techniker

N: Zahl der Trainingsteilnehmer

A: Anteil der durch das Training betroffenen Kompetenzen an allen Kompetenzen

SD_Y: Standardabweichung der Leistung (soweit durch das Training beeinflusst)

d: Effektstärke

G: Monetärer Gewinn infolge des Trainings

K: Kosten für das gesamte Training, Entwicklung und Durchführung

MN: Monetärer Nutzen des Trainings: $G - K$

ROI: Return of Investment (Verhältnis des monetären Nutzens zu den Kosten)

BE: Break Even (Anzahl der Jahre, bis der Gewinn den Kosten entspricht)

3.2 Eine Evaluationsstudie zur Schulzeitverkürzung

Eine aus der Pädagogischen Psychologie stammende Längsschnittstudie richtet sich auf die Evaluation einer Intervention zur Schulzeitverkürzung unter Leitung von Prof. Heller (vgl. Heller & Rindermann, 1997, 1999; Reimann & Heller, 2004; eine etwas umfangreichere Zusammenfassung dieser aus mehreren Einzelstudien bestehenden Studie findet sich in Holling, 2001 sowie Vock, Holling & Preckel, 2007).

Im Jahr 1991 startete in Baden-Württemberg ein Modellversuch zur Schulzeitverkürzung, bei dem vier baden-württembergische Gymnasien ihr Unterrichtsangebot um einen achtjährigen Bildungsgang, kurz G8 genannt, erweiterten. Dazu wurde ein separater gymnasialer Zweig eingerichtet, der leistungsfähigen Schülerinnen und Schülern die Möglichkeit bot, durch eine Komprimierung des Stoffs der Jahrgangsstufen 5 bis 11 auf die Jahrgänge 5 bis 10 das Abitur bereits nach 12 Jahren (statt nach 13 Jahren) abzulegen. Über die Zulassung zum G8-Zweig wurde anhand eines Gutachtens der Grundschule, Informationen aus Gesprächen mit den Eltern und dem Kind selbst sowie kleinerer Arbeitsproben des Kindes entschieden. Für die aufgenommenen Schülerinnen und Schüler begann mit Beginn der fünften Klasse das schulische Förderprogramm (Reimann & Heller, 2004). Am Anfang der zwölften Klasse wurden die Schülerinnen und Schüler des G8-Zweigs wieder in den G9-Zweig integriert. Am Ende der 13. Jahrgangsstufe absolvierten somit G8- und G9-Schülerinnen und Schüler die gleiche Abiturprüfung.

Das G8-Programm wurde nach der Modellphase weitergeführt und erweitert, sodass schließlich an 82 Gymnasien in Baden-Württemberg G8-Zweige eingerichtet wurden. Gleichzeitig fanden einige Modifikationen des Programms statt (z. B. Änderung der Aufnahme-prozedur, Abschaffung der ursprünglich erhöhten Anzahl an Leistungskursen für Schülerinnen und Schüler im G8-Zweig). Ab dem Schuljahr 2004/2005 wurde das achtjährige Gymnasium in Baden-Württemberg flächendeckend für alle Schülerinnen und Schüler eingeführt, nicht zuletzt aufgrund der Ergebnisse der im Folgenden berichteten Evaluationsstudie.

Die Evaluationsstudie erstreckte sich über den Zeitraum von 1992 bis 2001 und umfasste insgesamt zehn Untersuchungswellen. Die kompletten ersten drei G8-Einschulungsjahrgänge an den vier Gymnasien ($N=117$) wurden jährlich untersucht. Ergänzt wurde die Studie durch Kontrollstichproben an zwei neunjährigen Regelgymnasien, die keine verkürzten Klassen anboten ($N=59-64$, je nach Jahrgangsstufe). Das zentrale Ziel dieser Evaluationsstudie war die Erfassung der Begabungs- und Leistungsentwicklung der Schülerinnen und Schüler. Es wurden zahlreiche unterschiedliche Messinstrumente insbesondere

zur Erfassung der kognitiven, motivationalen und sozialen Entwicklung eingesetzt.

Hinsichtlich der kognitiven Fähigkeiten, gemessen mit dem Kognitiven Fähigkeitstest KFT4-13+ (Heller, Gaedicke & Weinländer, 1985), zeigte sich, dass die Schülerinnen und Schüler des G8-Zweigs den Schülerinnen und Schülern im G9-Zweig überlegen waren. Der Vorsprung der G8-Gruppe von einer halben Standardabweichung zu Beginn weitete sich in der zweiten Hälfte der Gymnasialzeit auf eine ganze Standardabweichung aus. Dieser Unterschied entspricht im direkten Leistungsvergleich etwa einer ganzen Note. Am Ende der Gymnasialzeit bestand ein klarer Vorsprung in den schulischen Leistungen, sichtbar in Form von besseren Zensuren, der zu Beginn lediglich in einigen, nun aber in allen Fächern deutlich wurde. Auch bei der Integration der G8-Schülerinnen und Schüler in das G9-System war trotz bereits besserer Bewertung nochmals ein Notenanstieg bei ihnen zu verzeichnen. Im Abiturzeugnis lagen die G8-Schülerinnen und Schüler durchschnittlich eine dreiviertel Standardabweichung über dem G9-Notenmittelwert, obwohl sie im Schnitt ein Jahr jünger waren und ein Schuljahr weniger absolviert hatten.

Auch für die motivationale und emotionale Entwicklung konnte für die G8-Schülerinnen und -Schüler eine günstigere Entwicklung im Vergleich zu denen des G9-Systems aufgezeigt werden. Die G8-Schülerinnen und Schüler zeigten sowohl eine höhere schulische Selbstwirksamkeit und ein höheres akademisches Selbstkonzept. Keine Unterschiede zwischen den beiden Schülergruppen konnten in Bezug auf das Lern- und Arbeitsverhalten sowie die Entwicklung sozialer Kompetenzen, die in beiden Fällen positiv verlief, gefunden werden (Reimann & Heller, 2004). Am Ende der Gymnasialzeit besaßen die G8-Schülerinnen und Schüler ein höheres Selbstvertrauen, eine größere Veränderungsbereitschaft sowie eine stärkere innere Ausgeglichenheit als die Schülerinnen und Schüler der G9-Klassen.

In die Längsschnittstudie wurden als Kontrollpersonen neben den Schülerinnen und Schülern neunjähriger Regelgymnasien auch besonders begabte Schülerinnen und Schüler mit einbezogen, die für den G8-Zweig zwar geeignet gewesen wären, sich jedoch gegen den G8-Besuch und für das reguläre neunjährige Gymnasium entschieden hatten (als „Extra-Schülerinnen und Schüler“ bezeichnet). Die Ergebnisse zeigen, dass die Extra-Schülerinnen und Schüler ein geringfügig negativeres Selbstkonzept aufwiesen, sich stärker stigmatisiert fühlten und weniger erfolgszuversichtlich waren als ihre vergleichbar begabten Mitschülerinnen und Mitschüler aus den G8-Klassen. Während Heller und Rindermann (1999) die Persönlichkeitsentwicklung der Schülerinnen und Schüler sowohl des G8- als auch des G9-Zweigs insgesamt als positiv bewerten, zeigten sich für die Extra-Schülerinnen und Schüler etwas ungünstigere Ergebnisse.

Weiterhin wurden umfassende Einschätzungen der Lehrkräfte und Eltern vorgenommen, auf die hier nicht im Einzelnen eingegangen werden soll. Die bisherigen Ergebnisse wurden durch diese Analysen bestätigt. Die beteiligten Lehrkräfte zeigten sich sehr zufrieden mit dem G8-Zweig. Als besonders positiv erlebten sie die hohe Motivation und die überdurchschnittlichen Fähigkeiten der Schülerinnen und Schüler und das relativ seltene Auftreten von Disziplinproblemen. Die meisten Eltern der G8-Schülerinnen und Schüler waren insgesamt sehr zufrieden mit dem G8-Zweig und würden ihr Kind – erneut vor die Wahl gestellt – wieder dort anmelden.

Fazit dieser nicht nur in schulpolitischer Hinsicht sehr einflussreichen Evaluationsstudie ist, dass die Schulzeitverkürzung anhand des G8-Konzepts hinsichtlich der erhobenen kognitiven, motivationalen und sozialen Kriterien durchweg als positiv zu bewerten ist. So kommen Heller und Rindermann (1997) aufgrund der erhobenen Daten zu dem Schluss, dass zu Beginn der gymnasialen Laufbahn etwa 20 % der Schülerinnen und Schüler der Gesamtpopulation für den Besuch des G8-Zweigs geeignet sind.

Die beiden dargestellten Evaluationsstudien zeigen deutlich auf, dass in Evaluationsstudien weitaus komplexere Probleme auftreten als in „klassischen“ labor-experimentellen Studien. Hier mag es ausreichen,

- nur wenige Probanden
- mittels einer Methode, z. B. der funktionellen Kernspintomographie (fMRT),
- im Rahmen einer eng umgrenzten Fragestellung, z. B. hinsichtlich der Beteiligung bestimmter Gehirnregionen an der Wiedererkennung bestimmter visueller Muster,
- anhand vollständig randomisierter Versuchspläne

zu analysieren.

In den folgenden Kapiteln werden die besonderen Merkmale von Evaluationsstudien thematisiert. Dabei stellt sich zunächst die Frage nach dem Gegenstand psychologischer Evaluationsstudien.

4 Gegenstand psychologischer Evaluationsforschung

Definiert man Evaluation und gleichbedeutend damit Evaluationsforschung als die Bewertung von Sachverhalten mittels anerkannter wissenschaftlicher Methoden, stellt sich die Frage, welche Sachverhalte Gegenstand der Bewertung sind. Im Rahmen der psychologischen Evaluationsforschung sollte eine Beschränkung auf solche Sachverhalte erfolgen, die Gegenstand der wissen-

schaftlichen Psychologie sind. Dieses erste einschränkende Kriterium allein würde allerdings noch zu einem zu weiten Gebiet führen. So müsste man im Rahmen psychologischer Fragestellungen zahlreiche psychologische Standardmaßnahmen aus den unterschiedlichsten Disziplinen als Evaluation bezeichnen, wie z. B. Eignungsdiagnosen oder Leistungsbeurteilungen von Mitgliedern einer Organisation. Insofern sollten nicht alle möglichen Sachverhalte wie Personen, Prozesse oder Interventionen Gegenstand eines Fachs Evaluationsforschung sein. Weiterhin würde die Einbeziehung der Bewertung der Auswirkungen von Wandel in Natur, Kultur, Technik und Gesellschaft zu einem zu weiten Feld führen. Daher erscheint es sinnvoll, wie es in vielen Definitionen vorgenommen wird, sich auf gezielte Interventionen bzw. Programme als Gegenstand zu beschränken, wobei wir hier den Begriff Intervention präferieren. Hager und Hasselhorn (2000, S. 41) verstehen unter Interventionsmaßnahmen „jede Art von außengesteuerter, zielorientierter und systematischer Beeinflussung von Personen- und/oder Systemmerkmalen. Jede Interventionsmaßnahme besteht mindestens aus einer Menge von zu bearbeitenden Aufgaben bzw. Problemen und mindestens einer Methode der Instruktion“.

Die wichtigsten Interventionsmaßnahmen im Rahmen der Psychologie entstammen den wichtigsten psychologischen Anwendungsfächern, der Klinischen Psychologie, Arbeits- und Organisationspsychologie und der Pädagogischen Psychologie, weiterhin aber auch der Verkehrspsychologie, Markt- und Werbepsychologie und anderer angewandter Disziplinen.

Interventionen in den aufgeführten Disziplinen können sich hinsichtlich ihrer Komplexität bzw. ihres Umfangs sehr unterscheiden. So sind Bewertungen der Akzeptanz eines einstündigen Rhetoriktrainings mittels eines kurzen Fragebogens weitaus einfacher durchzuführen als monetäre Nutzenanalysen oder Interventionen zur Schulzeitverkürzung, wie sie oben beschrieben wurden. Schwerpunkt der Evaluationsforschung sollten jedoch relativ komplexe Studien sein. Akzeptanzanalysen kurzfristiger Trainings oder ähnliche Studien erfordern kaum methodische Kenntnisse und Kompetenzen, die über die gängigen Lehrinhalte der Methodenlehre hinausgehen. Die Etablierung eines eigenen Fachs Evaluationsforschung würde damit kaum zu rechtfertigen sein. Hingegen führt die multiattributive Bewertung komplexer, längerfristiger und praktisch bedeutsamer Interventionen, die zudem die Entscheidungsgrundlage für Veränderungen der Interventionen bilden, zu spezifischen Merkmalen dieser Studien, die Kenntnisse und Kompetenzen bedingen, die über die gängigen empirischen Studien aus der Grundlagen- und angewandten Forschung hinausgehen.

Aus der Komplexität und Langfristigkeit solcher Studien folgt in der Regel, dass die Zielstellung mehrdimensional ist und die Bewertung der Interventionen anhand multipler Kriterien erfolgen muss. Damit ist eine Vielzahl von unab-

hängigen und abhängigen Variablen oftmals auf unterschiedlichen Ebenen einzubeziehen. Folglich sind multivariate statistische Analyseverfahren und Modelle der multiattributiven Nutzentheorie gefordert. Hier sind z. B. Strukturgleichungsmodelle, hierarchische lineare Modelle, Modelle zur multiattributiven Nutzenmessung oder monetären Nutzenanalyse einzusetzen.

Komplexe Interventionen erschweren weiterhin die Bildung von kausalen Annahmen bzw. Theorien. Zudem sind Managementkenntnisse und Kompetenzen zur laufenden Steuerung solcher umfangreichen Projekte gefordert, die im Rahmen der sogenannten formativen Evaluation thematisiert werden. Sehr häufig erlauben erst externe Aufträge Evaluationen umfassender Interventionen. Daher hat man es oftmals zunächst mit einer sehr globalen Fragestellung zu tun, die erst noch anhand bestimmter Methoden zu spezifizieren ist. Weiterhin sind die Interessen der Auftraggeber und anderer Interessengruppen, sogenannte Stakeholder, in Einklang zu bringen (s. u.). Hierzu sind spezifische kommunikative Kompetenzen hilfreich.

5 Ziele von Evaluationsstudien

Psychologische Interventionen können mit unterschiedlichen Zielstellungen bzw. aus verschiedenen Motiven evaluiert werden. Wie die beiden Beispiele oben gezeigt haben, können sie die Grundlage für ökonomische, gesellschaftliche oder politische Entscheidungen bilden. Oftmals geht es um Fragen wie:

- Soll eine bestimmte Intervention eingeführt, aufrechterhalten, modifiziert oder beendet werden?
- Welche von mehreren zur Auswahl stehenden Interventionen verspricht den größten Nutzen? Welche Komponenten einer bestimmten Intervention machen die Wirksamkeit aus?
- Wie kann eine Intervention wirksamer gestaltet werden?
- Unter welchen Bedingungen ist eine Intervention wirksam?

Solche oder ähnliche Ziele liegen häufig konkreten Evaluationsstudien zugrunde. Dabei werden solche Evaluationsstudien oftmals durch öffentliche oder private Auftraggeber finanziert. Insofern ist nicht auszuschließen, dass Evaluationsstudien, wenn auch nicht *expressis verbis*, als strategisches Instrument eingesetzt werden. Entscheidungsträger sollen evtl. vom Erfolg oder auch Misserfolg einer Maßnahme überzeugt werden, es können Marketingzwecke verfolgt werden oder Evaluationen können durchgeführt werden, um die Motivation und Leistungsbereitschaft der Zielgruppe zu erhöhen (motivational evaluation; Scriven, 1991). Evaluation mag hier nicht unbedingt als wertneutrales Erkenntnisinstrument begriffen werden (vgl. Campbell, 1969), sondern als Mittel zum Zweck.

Dabei ist aber die Gefahr gegeben, dass Evaluationsergebnisse einem Bias unterliegen. Hier ist also ein besonderes Augenmerk auf eine fachgerechte Bewertung der Interventionen im Sinne des zuerst aufgeführten Zieles, der multiattributiven Nutzenmessung, zu richten.

Auf der anderen Seite wird z. B. im Ansatz der *utilization-focused evaluation* (Patton, 1997) Evaluation ganz bewusst und intendiert auf die Interessen einer Zielgruppe abgestimmt. Hier wird Evaluation nicht als wertfreies Instrument, sondern als eine Maßnahme im Rahmen konkreter Interessensvorgaben gesehen. So kann die Verbesserung der Lebensqualität bestimmter Bevölkerungsgruppen ein Ziel von Evaluationsstudien darstellen (z. B. Mark, Henry & Julnes, 2000; Patton, 1997; Posavac & Carey, 2003).

Grundsätzlich kann der Nutzen von Interventionen für verschiedene Ziele bestimmt werden. Durch unterschiedliche Zielstellungen ergeben sich auch unterschiedliche Kriterien für die Festlegung des Nutzens, was im Rahmen multiattributiver Nutzenanalysen (s. u.) deutlich wird.

Ein weiteres wichtiges Ziel der Evaluationsforschung ist zweifelsohne die Erweiterung des theoretischen Wissens über das untersuchte inhaltliche Gebiet (knowledge-generating evaluation; Patton, 1996) bzw. weitergehend die Erstellung von Theorien zu bestimmten Interventionen und ihres Nutzens. Die hypothesenorientierte Überprüfung einer der evaluierten Intervention zu Grunde liegenden Theorie ermöglicht über die reine Wirksamkeitsprüfung hinausgehende Erkenntnisse konzeptueller Natur, womit auch in utilitaristischer Hinsicht Vorteile entstehen können. So lassen sich u. U. präzisere Aussagen über zukünftige Entwicklungen treffen, was gerade im Zusammenhang groß angelegter sozialer Programme einen hohen Wert haben kann (Cronbach, 1982), z. B. wenn Pilotprojekte evaluiert werden, um über eine eventuelle flächendeckende Einführung zu entscheiden. Die theoretische Grundlegung von Evaluation ist ein wichtiges Thema, auf das wir im nächsten Kapitel näher eingehen.

Evaluationsstudien können zugleich mehreren der oben genannten Zielen dienen, da sie sich keineswegs gegenseitig ausschließen müssen. Auch wenn bestimmte Evaluationsstudien nicht primär für die Überprüfung bestimmter theoretischer Annahmen konzipiert wurden, so können die Ergebnisse nichtsdestotrotz selbstverständlich zur Erweiterung des theoretischen Wissens dienen. So konnten aus den Ergebnissen der oben berichteten Studie von Morrow, Jarrett und Rupinski (1997) Annahmen zur Trainingsforschung überprüft werden. Die umfangreichen Daten und Ergebnisse der Studie zur Schulzeitverkürzung (s. o.) sind eine wichtige Grundlage für die Weiterentwicklung von Theorien innerhalb der Pädagogischen Psychologie, etwa zur Begabungsforschung oder Schulmotivation.