Evaluationsforschung und Programmevaluation im Gesundheitswesen

Werner W. Wittmann[1], Rüdiger Nübling[2] und Jürgen Schmidt[3]

1. Aufgaben und Probleme

Das Gesundheitswesen ist immer wieder im Kritikpunkt vieler aktueller politischer Debatten und Auseinandersetzungen. Fragen der Finanzierbarkeit, der Qualität der gesamten Versorgungskette, ärztliche Kunstfehler, adäquate Bedürfnisbefriedigung und Zielgruppenerreichung bestimmen die Debatten. Schon sehr früh hat Donabedian (1966) drei wichtige Qualitätsaspekte, die Struktur-, die Prozess- und die Ergebnisqualität unterschieden.

Qualitätssicherung und Evaluationsforschung hängen eng miteinander zusammen. Evaluationsforschung ist die explizite Verwendung wissenschaftlicher Methoden, die den möglichst kausalen Nachweis der Wirksamkeit und Effizienz einer Intervention erbringen soll, um somit unsäglichen Debatten vorzubeugen, die aus persönlichen Präferenzen und Lobbyarbeit gespeist, zu unterschiedlichen Wertungen führen. Programmevaluationen sind angewandte Forschungsarbeiten, die Kosten-Nutzen und Kosten-Effektivität von einzelnen Interventionen oder ganzen Interventionspaketen zum Ziele haben. Die Sozialwissenschaften und hierin vor allem Psychologie und Soziologie haben vielerlei quantitative und qualitative Verfahren und Methoden entwickelt und bereit gestellt um solche Aufgaben erfüllen zu können. Eines der bekanntesten Lehrbücher des Feldes ist in seiner neuesten Auflage eine Kooperation eines Soziologen und eines Psychologen (Rossi, Freeman & Lipsey 1999). Das Gebäude der Evaluationsforschung und der Programmevaluation ruht auf drei starken Säulen, erstens den Methoden der Versuchsplanung und den dazugehörigen statistischen Datenanalysestrategien, zweitens den Methoden und Verfahren der Datenerhebung (Assessment) und drittens den Planungs-, Zielfindungs-, Bewertungs- und Entscheidungshilfen.

Im Bereich der Medizin haben sich starke Organisationen herausgebildet wie die Cochrane Colloboration, deren Ziel es ist, über Forschungssynthesen und Metaanalysen Bewertungen über die kausale Wirksamkeit und Evidenz von medizini-

1 Universität Mannheim,
2 eqs-Institut Karlsruhe,
3 Karlsruher Sanatorium AG

schen Interventionen und Maßnahmen zu erarbeiten. Grundlage hierfür sind meist randomisierte kontrollierte Vergleichsstudien (RCTs, d.h. randomized clinical trials) aus der Grundlagenforschung, welche die stärkste kausale Evidenz liefern können. Evidenzbasierte Medizin (Stevens et al. 2001) ist ein weiterer Ansatz und ein Schlagwort, das enge Verbindungen zu Evaluation und Programmevaluation hat. Gesundheitsökonomie ist ebenfalls ein verwandtes Gebiet, in dem Verfahren und Strategien aus der Betriebswirtschaftslehre wie Total Quality Management (TQM) und Kosten-Wirksamkeits- bzw. Kosten-Nutzen-Analysen eingesetzt werden (Schöffski und Schulenburg 2000).

2. Lösungsvorschläge zur Indikation adäquater Forschungsstrategien: Die Konzeption der fünf Datenboxen

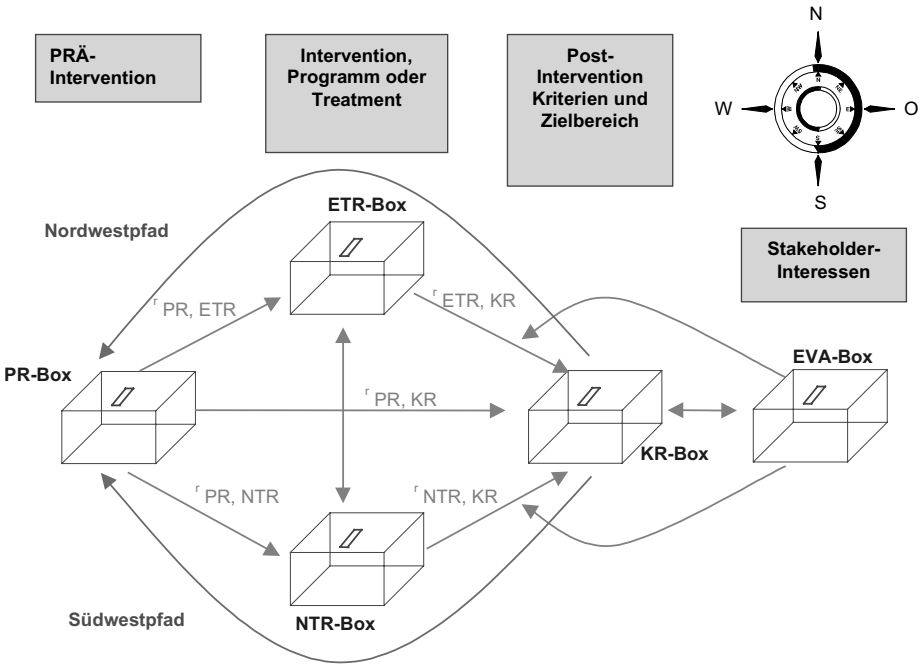
Evaluationsforscher stehen, wie gute Ärzte, vor dem Problem, nach einer ausführlichen Diagnose der Rahmenbedingungen, eine adäquate Indikationsentscheidung zu treffen. Welche Versuchspläne, Datenerhebungen und Bewertungs- und Entscheidungshilfen sind für die Überprüfung einer medizinischen Maßnahme oder eines Programms am besten indiziert, um eine faire Bewertung durchführen zu können?

Wir haben dazu einen konzeptuellen Rahmen vorgeschlagen, der auf viele denkbare Evaluationen hilfreich angewendet werden kann. Die Konzeption umfasst fünf verschiedene Datenboxen, die wichtige notwendige Datenbereiche visualisieren, die es zu erfassen gilt. Die einzelnen Datenboxen können auf unterschiedliche Weisen miteinander verknüpft und ausgewertet werden, was auf unterschiedliche Versuchspläne hinweist. Die Effektgrößen, welche die einzelnen Datenboxen in Beziehung setzen, liefern Grundlagen für Bewertungen und Entscheidungen. Abb. 1 stellt diese Konzeption der fünf Datenboxen dar.

Jede Datenbox umfasst drei Dimensionen, die Personen, die Variablen und die Messzeitpunkte bzw. Situationen. Cattell (1957) hat diese Datenbox als covariation chart im Bereich psychologischer Forschung vorgestellt. Zu Beginn jeder Evaluation sollten wir uns darüber Klarheit verschaffen, wer die so genannten Interessenträger (Stakeholder) eines Programms sind. Hierzu zählen alle direkt Betroffenen wie Patienten und Versorgungs- bzw. Interventionspersonal, aber auch die Kostenträger und Gesundheitspolitiker und die Öffentlichkeit. Diese verschiedenen Stakeholdergruppen haben in der Regel ganz unterschiedliche Interessen und Ziele (EVA-Box), was automatisch zu einer Vielzahl von möglichen Kriterienmaßen führt. Faire Evaluationen müssen diese unterschiedlichen Kriterien möglichst repräsentativ in der Kriterienbox (KR-Box) abbilden. Die Variabilität in den Kriterienmaßen gilt es vorherzusagen und zu erklären. Hierfür haben wir drei Datenboxen vorgesehen. Die experimentelle Treatmentbox (ETR-Box) umfasst die randomisierten Kontrollgruppenpläne, die in der medizinischen Forschung als RCTs bezeichnet werden. Die nichtexperimentelle Treatmentbox (NTR-Box) visualisiert nichtäquivalente Kontrollgruppendesigns oder reine korrelative Studien. Der Doppelpfeil zwischen ETR- und NTR-Box weist auf den fließenden Übergang von RCTs über Quasiexperimente (Cook & Campbell 1979) bis zu den rein korrelativen Designs hin. Die Prädiktorenbox bildet Zustand und Ausgangslage vor jeder Intervention, sei sie nun

experimentell oder nichtexperimentell, ab. Die Windrose soll uns dazu anregen, diese Datenboxkonzeption wie eine geographische Karte zu lesen. Versuchen wir die Kriterienvariablen über den Pfad der über ETR- und PR-Box verläuft, so beschreiten wir den Nordwestpfad, der von der Northwesternschule empfohlen wurde. Campbells (1969) einflussreicher Aufsatz "Reforms as experiments" steht hierfür als Modell und vor allem die oft als Versuchsplanungsbibeln beschriebenen Werke von Campbell und Stanley (1966), bzw. Cook und Campbell (1979). Donald T. Campbell war eine umfassend gebildete charismatische Forscherpersönlichkeit, er hat das Gebiet der Sozialwissenschaften und die Evaluationsforschung maßgeblich geprägt. Seine Schüler, Mitarbeiter und langjährigen Weggefährten sind heute über viele Universitäten in den USA verstreut. Bob Boruch inzwischen an der University of Philadelphia hat mit zahlreichen Kollegen eine Campbell Collaboration gegründet, die in ähnlicher Weise wie die UK-basierte Cochrane Collaboration das Ziel hat, experimentelle Forschungsarbeiten und Evaluationsstudien zu fördern und zu synthetisieren. Gene Glass (1983) hat den Begriff Northwestern Schule geprägt, der eine Richtung bezeichnet, durch RCTs zu möglichst guten kausalen Aussagen zu gelangen. Die methodischen Gütestandards der internen und der statistischen Schlussfolgerungsvalidität spielen dabei eine zentrale Rolle.

Abb. 1: Die Konzeption der fünf Datenboxen



Die Dominanz der von der Northwesternschule gesetzten Standards blieben natürlich nicht unwidersprochen. Viele Programmevaluationsstudien konnten keine oder nur geringe Effekte von Interventionen nachweisen, was Rossi (1978) in seinem bekannten „iron law of program evaluation" zusammenfasste. Cronbach kritisierte

den Fokus den die Northwestern Schule auf den Gütemaßstab interne Validität legte und unterstrich die Bedeutung der externen Validität, der Konstruktvalidität und der Generalisierbarkeit, die bei korrelativen Untersuchungen, die über den Südwestpfad, d.h. über NTR- und PR-Box die Kriterien erklären wollen, höher sein sollte. Das Dilemma dieser Erklärungsstrategie liegt allerdings in der Gefahr der Konfundierung von PR-Box mit der NTR-Box, eine Gefährdung, die als Selektion in das Treatment bekannt ist. Die in vielen Bereichen enttäuschenden Ergebnisse über Effektivitätsnachweise haben viele Evaluationsforscher an der Angemessenheit einer quantitativen Evaluationsforschung zweifeln lassen, und wir finden heute in der American Evaluation Association eine Dominanz qualitativ orientierter Forscher.

Shadish, Cook & Leviton (1991) zeichnen die Geschichte der unterschiedlichen Phasen der Evaluationsforschung mit den jeweils präferierten Methodologien nach. Evaluation hat viele unterschiedliche Facetten und Gesichter. Die einzelnen Schulrichtungen unterscheiden sich meist einfach dadurch, dass bestimmte Phasen im Ablauf und der Durchführung einer umfassenden Programmevaluation besonders betont werden. Für die einzelnen Phasen sind wiederum bestimmte Methodologien besonders indiziert. Für die Planungsphase und die Eruierung von Bedürfnissen und Zielgruppen erscheinen qualitative Methoden besonders geeignet. Die formative Evaluationsphase, bei der die Entwicklung und schrittweise Optimierung im Vordergrund stehen, erfordert ebenfalls häufig qualitative Beobachtungsverfahren, die durch quantitativ orientierte Monitoringsysteme ergänzt werden. Die summative Phase, bei der die abschließende Bewertung eines Programms bezüglich Zielerreichung und Kosten-Nutzen bzw. Kosten-Effektivität den Schwerpunkt bildet, dominieren quantitative Methoden, die den bestmöglichen kausalen Nachweis von Interventionen liefern sollen.

Das gesamte Gesundheitswesen ist ein komplexes System bestehend aus vielen Einrichtungen und Interventionsprogrammen. Wie kann und soll ein solches System am besten und fairsten bewertet werden? Die denkbare Anzahl von Interessengruppen ist hierbei enorm hoch. Wir können die Versichertenperspektive, die Patientenperspektive, die Versorgungsperspektive, die Therapeutenperspektive, die Trägerperspektive, die gesundheitspolitische Perspektive und vieles mehr unterscheiden. Das gesamte System ist entstanden, um bestimmte Bedürfnisse zu befriedigen und gesundheits- ebenso wie gesellschaftspolitische Zielsetzungen zu erreichen. Dennoch können wir hierbei alle Stakeholder in zwei große Gruppen unterteilen. Die eine Gruppe legt einen Schwerpunkt auf monetär bewertbare Kriterien um die Finanzierbarkeit und die optimale Verwendung finanzieller Ressourcen bewerten zu können. Die andere Gruppe betont besonders Kriterien der Lebensqualität, vor allem soziale, psychologische und humanistische Kriterien, die sich einer reinen monetären Bewertung entziehen, aber dennoch als wichtig erachtet werden.

In der Grundlagenforschung zur Wirksamkeit von Interventionen ist die Situation völlig anders. Die Freiheit der Forschung sichert die Auswahl eines meist spezifischen Kriteriums, das nur die Forscherin interessiert. Durch hochkontrollierte Studien werden Störeinflüsse konstant gehalten um wenige Kausalfaktoren effizient testen zu können. Als Kriterien finden wir hier Auseinandersetzungen um das beste Kriterienmaß, das als Goldstandard am genauesten auf die Abbildung des Wirkmechanismus zugeschnitten ist. Die interne Validität wird dadurch maximiert. Metaanalysen synthetisieren dann das kausale Wirkungspotential unter optimalen Bedin-

gungen. Die Cochrane Collaboration veröffentlicht regelmäßig Berichte über solche Wirksamkeitsanalysen. Die Kenntnis der Wirksamkeit unter hochkontrollierten Bedingungen sagt nichts darüber aus, wie groß die Wirksamkeit unter den natürlichen Rahmenbedingungen der realen Versorgung ist. Aus dem Bereich der Umsetzung naturwissenschaftlicher Erkenntnisse in entsprechende Technologien ist es seit langem bekannt, dass der Wirkungsgrad deutlich geringer als theoretisch erwartet ausfällt. Optimierung und Verbesserung im Rahmen der Ingenieurswissenschaften nehmen einen großen Stellenwert ein. Formative Evaluationen zur Produktverbesserung und Erhöhung eines Wirkungsgrades, z.B. Ausschöpfung des Energiepotentials eines Kraftstoffes, sind dort längst Hauptziel und Routine. Im Gesundheitswesen dagegen wissen wir immer noch viel zu wenig über den Wirkungsgrad hinsichtlich Wirkungen und Nebenwirkungen im realen System. Dort ist die Situation zudem wiederum sehr viel komplizierter, da in der Regel ganze Behandlungspakete aus verschiedenen Interventionen geschnürt werden. Wie weit sich die Effekte aus den kontrollierten Studien generalisieren lassen und wie groß deren Wirkungsgrad in der Praxis ausfällt, ist meist auf Grund unzureichender Programmevaluationsforschung völlig unbekannt.

Das National Institute of Mental Health (NIMH) hat aus diesen Gründen seiner Forschungsförderungspolitik eine neue Akzentuierung gegeben. Foxhall (2000) berichtet im APA Monitor über diese neue Politik des Direktors Steven Hyman unter der Schlagzeile: „Research for the real world. NIMH is pumping big money into effectiveness research to move promising treatments into practice." NIMH investiert jährlich mehr als 40 Millionen Dollar, zum ersten Mal in seiner Geschichte, um herauszufinden, was in der klinischen Praxis effektiv umgesetzt wird. Der neue Fokus steht in starkem Gegensatz zu traditionellen Wirksamkeitsstudien mit ausgewählten Patienten in universitären Settings: „It will study large numbers of divers patients in real-world settings, follow them for lengthy periods of time and measure progress by the patients' functioning in school, work and other areas of life."

Das deutsche Rehabilitationssystem als Teilkomponente unseres Gesundheitswesens ist auf Grund seiner Besonderheit und Einmaligkeit schon seit langem in der Kritik und hat seit zwei Jahrzehnten in bemerkenswerter Weise mit systematischen Programmevaluationsstudien auf diese Herausforderungen reagiert. Zahlreiche private Träger und Anbieter von Rehabilitationsmaßnahmen haben aus Eigeninitiative solche Studien finanziert und durchgeführt. Tab. 1 listet eine beindruckende Anzahl von Studien auf. Die Bundesversicherungsanstalt für Angestellte (BfA) in Berlin hat ein Fünf-Punkte-Programm zur Qualitätssicherung aufgestellt, in dem Programmevaluationen einen hohen Stellenwert haben. Die Bundesregierung und der Verband Deutscher Rentenversicherungsträger fördern zahlreiche Forschungsverbünde deren Hauptzielsetzungen ebenfalls geeignete Forschungsmethoden und Messinstrumente für Untersuchungen zur Effektivität des realen Rehabilitationssystems sind.

Tab. 1: Evidenz aus umfassenden Evaluationsstudien zur Ergebnisqualität in der psychosomatischen Rehabilitation mit mindestens einem Katamnesezeitpunkt (Auswahl)

Studie	Messzeitpunkte	n	Publikationen/Autoren (Auswahl)
BKK-Studie	3	148	Zielke 1991
Berus-Studie	4	370	Broda et al. 1996
Zauberberg-I-Studie	4	365	Schmidt 1991, Lamprecht & Schmidt 1990
Zauberberg-II-Studie	3	565	Schmidt et al. 1994, Schmidt & Lamprecht 1992, Nübling 1992
Bliestal-Studie	3	1088	Sandweg et al. 1991
Reinerzauer Katamnese-Studie	4	560	Nübling et al. 1994, 1995, 1999
Bad Herrenalber Katamnese-Studie	3	317	Nübling et al. 2000a, Bürgy et al. 2000
Grönenbacher Studie		767	Mestel et al. 2000a, Mestel et al. 2000b
Bad Kreuznacher Studie	4	376	Schulz et al. 1999, Rüddel et al. 1999
Protos-Studie, Teilstichprobe Psychosomatik	3	884	Dilcher et al. 2000, Gerdes et al. 2000
Gelderland-Studie		345	Kriebel et al. 1999
EQUA-Studie	3	899	Schmidt et al. 2000a
INDIKA-Studie, Teilstichprobe Psychosomatik	3	274	Nübling et al. 2000b
CED-Studie	3	175	Maatz & Schmidt, 1998
GR-Studie	3	292	Amann et al. 1995, Amann 1997
Priener Lehrer-Projekt		61	Hillert et al. 1999, Hillert et al. 2000
Katamnese-Studie Eifelklinik	3	47	Tigiser 1997
Berliner PT-Studie, Teilstichprobe Psychosomatische Rehabilitation	3	132	Rudolf et al. 1991, Wilke et al. 1988
Multizentrische Studie Esstörungen (MZ-ESS), Teilstichprobe Psychosomatische Rehabilitation	4	ca, 500	Kächele et al. 1999
Prä-Post-Studie	6	145	Bischoff et al. 2000
		8310	

Rufen wir uns die drei wichtigen Säulen der Evaluationsforschung und der Programmevaluation in Erinnerung, so stellt sich die Frage, welche Versuchspläne und Datenanalysestrategien, welche diagnostischen Assessmentinstrumente und welche Zielfindungs-, Bewertungs- und Entscheidungshilfen für diese Art von Forschung besonders indiziert und geeignet erscheinen. Randomisierte Kontrollgruppenpläne sind in den realen Versorgungssystemen besonders schwer zu implementieren, da sie erfordern würden, zufällig ausgewählten Patienten bestimmte Maßnahmen zumindest zeitweilig vorzuenthalten und damit juristische und ethische Bedenken aufwerfen. Als besonders indiziert erscheinen deshalb längerfristig angelegte Zeitreihenstudien bzw. follow-up Versuchpläne mit mehreren Messzeitpunkten, die nach Cook und Campbell (1979) als die stärksten quasiexperimentellen Designs gelten. In der Tat wurden solche Versuchspläne überwiegend in den in Tab. 1 aufgelisteten Untersuchungen umgesetzt.

Als Kriterienmaße eignen sich besonders multiple Ergebnismaße, die den unterschiedlichen Stakeholderinteressen entsprechen müssen. Programme des Gesundheitswesen sind unter dem Einfluss unterschiedlicher Interessengruppen entstanden

und versuchen gleichzeitig, mehrere monetär und nichtmonetär bewertbare Ziele zu erreichen.

Zur Bewertung müssen daher Kosten-Nutzen ebenso wie Kosten-Effektivitätsanalysen herangezogen werden. Beide Varianten erfassen den gesamten Aufwand an Personal, Zeit und Ressourcen in monetären Größen. Während bei Kosten-Nutzen-Analysen auch die Ergebnisse in Geldeinheiten gemessen werden, ist dies bei Kosten-Effektivitätsanalysen nicht der Fall. Dort einigt man sich auf psychometrische oder biometrische Skalen, wie z.B. Beschwerde und Symptomchecklisten oder Lebensqualität und analysiert wie viel ein Punktwert, eine Standardabweichung oder eine Effektgröße im Vergleich zu Alternativen kostet.

Eine umfassende Evaluation eines Programms bezeichnen Rossi, Freeman und Lipsey (1999) als „full coverage program evaluation“. Die erste Komponente umfasst eine Analyse der Bedürfnisse, die mit dem Programm gestillt werden sollen. Hierzu sind Methoden des "need assessment" hilfreich. Epidemiologische Daten zu Krankheitsgruppen, differenziert nach Auftretenshäufigkeit sind hierbei wichtige Informationsgrundlagen. Programmevaluationen orientieren sich nun am Ablauf einer Versorgungs- und Interventionskette. Als nächster Schritt steht nun die Entwicklung und Optimierung des Prototypen eines Programms im Vordergrund. Scriven (1988) bezeichnet diese Phase als formative Evaluation. Sobald ein prototypisches Programm aus theoretischen und praktischen Erwägungen entwickelt wurde, konzentriert man sich auf das Programmmonitoring, d.h. läuft das Programm nach Plan und wurde es gemäß theoretischer Vorgaben optimal implementiert. Erst wenn diese Phasen systematisch evaluiert wurden, lohnt sich eine summative Evaluation mit Fokus auf Effektgrößen, Metaanalysen, Analysen der kausalen Wirkmechanismen und den vergleichenden Kosten-Nutzen und Kosten-Effektivitäts-Analysen. Attkisson und Broskowski (1978) beschreiben diesen Prozess als die Erfassung der Bedürfnisse (Needs), des Aufwandes (Input), des Prozesses und der Ergebnisse (Output) und unterteilen ihn in fünf verschiedene Phasen. Der erste Bereich ist die Bestimmung des Aufwandes (effort measurement) der sich auf den Input bezieht. Die zweite Phase konzentriert sich auf den Output als Bestimmung der Leistung und Ausführung des Programms der Angemessenheit (adequacy measurement) als Quotient der Relation des Outputs zu den Bedürfnissen. Die vierte Phase fokussiert auf einem weiteren Quotienten, nämlich das Verhältnis von Output zu Input der als Bestimmung der Leistungsfähigkeit (efficiently measurement) bezeichnet wird. In der fünften und letzten Phase wird der Prozess analysiert (process measurement), bei dem versucht wird, das Ergebnis (outcome) als Funktion des Aufwandes zu erklären.

Kontroversen um die richtige, beste und angemessenste Evaluationsstrategie und Schulrichtung drehen sich im Grunde darum, welche Phase in einem Anwendungsgebiet mit der höchsten Priorität zu versehen ist. Da für die einzelnen Phasen bestimmte Methodologien besser indiziert sind als für andere, ist der Streit der Schulrichtungen nichts anderes als Auseinandersetzungen um die Wichtigkeit einer Phase und das Indikationsproblem, welche Methodologie für welche Phase am besten geeignet ist. Im deutschen Sprachraum finden wir diese Auseinandersetzungen in unterschiedlichen Anwendungsfeldern ebenfalls in einer Reihe von interessanten Abhandlungen beschrieben und diskutiert (Beywl 1988, Stockmann 2000).

3. Abbau von Evaluationsängsten: Das Potenzial von ex ante Kosten-Nutzen Betrachtungen

Evaluationen rufen neben der Hoffnung auf faire Bewertungen und die Optimierung bestehender Programme natürlich auch besonders Ängste hervor, im Spiegel der angelegten Messlatten entweder nicht bestehen zu können oder unfair bewertet zu werden. Im Vorfeld einer Evaluation sind deshalb besonders solche Methoden indiziert, die helfen können Ängste abzubauen.

Unter der Perspektive der Kosten-Nutzen Perspektive haben wir die Strategie a priori einen so genannten Break-Even-Point zu berechnen als besonders hilfreich empfunden und vorgeschlagen (Wittmann 1996). Grundlage dieser Betrachtung ist eine von Brogden (1949) entwickelte Gleichung, die den ökonomischen Nutzen einer Intervention abzuschätzen gestattet. Diese Gleichung wurde von Schmidt, Hunter und Pearlman (1982) aufgegriffen und in ihrer Bedeutung für die Bewertung von Ausbildungs- und Trainingsprogrammen vorgestellt. Sie kann jedoch in gleicher Weise für psychotherapeutische und medizinische Interventionen angewandt werden.

(1) $U= N*T*d*SD_{prod} – N*K$

Die einzelnen Parameter diese Gleichung sind wie folgt definiert:

U = Der Nettonutzen einer Intervention in Geldeinheiten, z.B. Euro,
N = Anzahl der therapierten Patienten,
T = Zeitdauer, wie lange der Therapieeffekt anhält, in Jahren,
d = Die Effektgröße, gemessen als standardisierte Mittelwertsdifferenz, entweder berechnet im Vergleich zu einer unbehandelten Kontrollgruppe, oder als standardisierte Differenz nachher-vorher,
SD_{prod} = Die Standardabweichung der Produktivität einer Vergleichsgruppe die diese Intervention nicht benötigte bzw. erhalten hat auf der Berechnungsgrundlage eines Jahres in Euro,
K = Die Gesamtkosten (direkt plus indirekt) der Intervention pro Patient in Euro.

Vor Durchführung einer Evaluationsstudie kennen wir natürlich die Effektgröße d noch nicht. Wir können aber abschätzen wie groß der Effekt mindestens sein muss, wenn Bruttonutzen und Kosten sich gerade aufwiegen. Dieser Wert ist der so genannte Break-Even-Point einer monetären Investition. An diesem Punkt ist der Nettonutzen U gleich Null. Setzen wir in Gl. 1 U = 0 und lösen die Gleichung nach d auf, so erhalten wir die Effektgröße am Break-Even-Point.

(2) $d_{break\text{-}even} = K/(T*SD_{prod})$

Wir benötigen nur noch die Gesamtkosten K, Dauer des Effektes T und die Standardabweichung der Produktivität, um die Effektgröße am Break-Even-Point berechnen zu können. Diese Größe kann als Vergleichsmaßstab für den später zu erhebenden tatsächlichen empirischen Effekt $d_{empirisch}$ herangezogen werden. Bei allen Investitionen hofft man einen so genannten „Return On Investment (ROI) zu erhalten, der größer als Eins ist. Niemand ist zufrieden, wenn die Investition nach einem

Jahr gerade wieder eingebracht wird. Dieser ROI-Koeffizient ist der Quotient aus der empirischen Effektgröße und derjenigen am Break-Even-Point, oder anders ausgedrückt:

(3) $d_{empirisch} = ROI * d_{break\text{-}even}$

Der Hauptgrund, weshalb diese seit über 50 Jahre bekannte Gleichung nicht angewendet werden konnte, lag in der Schwierigkeit, den Parameter SD_{prod} vernünftig zu schätzen. Man würde erwarten, dass die Betriebswirtschaftslehre hierfür eine Lösung erarbeitet hätte was aber nicht der Fall war. Interessanterweise haben die Psychologen Frank Schmidt und Jack Hunter dieses Problem auf eine verblüffend einfache Weise gelöst und brauchbare Schätzungen entwickelt. Produktivität ist ein Merkmal das durch viele Faktoren beeinflusst wird. Es gibt viele unterschiedliche Wege produktiv zu sein. Solche Merkmale sind normalverteilt. Der Bereich vom 15ten bis zum 85ten Perzentil umfasst hierbei gerade 2 Standardabweichungen, einen Bereich der durch den Mittelwert solcher symmetrischer Verteilungen gerade halbiert wird. Sie befragten daher Experten aus vielen Organisationen und Berufsfeldern, was die Arbeitsleistung eines unterdurchschnittlich produktiven Mitarbeiters am 15ten Perzentil wert sei, was am 50ten Perzentil (Durchschnitt) and was am 85ten Perzentil. Letzteres entspräche einer überdurchschnittlichen Produktivität. Am 85ten Perzentil haben nur noch 15 Prozent der Mitarbeiter eine höhere Produktivität und am 15ten Perzentil nur 15 Prozent eine schlechtere Arbeitsleistung. Zur Stützung der Schätzung empfahlen sie, für die Schätzungen den finanziellen Aufwand heranzuziehen wenn die Arbeitsleistung auf dem freien Markt eingekauft werden müsste. Die Differenz zwischen dem Mittelwert und dem 85ten Perzentil sollte ungefähr genauso groß sein, wie die Differenz zwischen Mittelwert und dem 15ten Perzentil. Die aggregierten Ergebnisse bestätigten diese Annahme. Das Gehalt wiederum sollte ein Spiegel der Produktivität auf einem freien Markt sein, da Organisationen, die zu viel bezahlen, Bankrott gehen, solche die zu wenig bezahlen ihre Mitarbeiter verlieren werden. Sie setzten deshalb die auf zwei Wegen geschätzte Standardabweichung in Relation zum bezahlten Jahresgehalt und fanden als Hauptergebnis , dass SD_{prod} in dem Intervall 40-70% des Jahresgehaltes liegt. Neuere Untersuchungen weisen eher auf die 70% Marke hin.

Nehmen wir nun an, dass eine Klinik evaluiert werden soll, die medizinische Rehabilitationsmaßnahmen für psychosomatische Patienten zur Wiederherstellung der Erwerbsfähigkeit durchführt, die von der Rentenversicherung finanziert wird. Wir gehen davon aus, dass das durchschnittliche Monatsgehalt eines Versicherten 2000 Euro beträgt, bei 13 Monatsgehältern wäre das Jahresgehalt 36000 Euro. Als Schätzung von SD_{prod} erhalten wir dann .70 * 36000 = 25200 Euro. Die durchschnittliche Dauer der Intervention betrage 45 Tage, bzw. 1,5 Monate. Der Pflegesatz sei 110 Euro pro Tag über den der Klinikträger alle Aufwendungen und Gewinne kalkulieren muss. Daraus resultieren 45 * 110 Euro = 4950 Euro an Behandlungskosten plus 1,5 * 2000 Euro = 3000 Euro für die Gehaltsfortzahlung, d.h. die Gesamtkosten K betragen dann 6950 Euro. Wir vermuten, dass der Effekt T z.B. rund 2 Jahre anhält.

Die Effektgröße am Break-Even wäre dann .138. Sensitivitätsanalysen können hier erhellen, wie stark der Effekt variiert, wenn wir unsere Parameter pessimistischer oder optimistischer einschätzen. Hält der Effekt nur ein Jahr an, erhalten wir

.276. Ändern wir SD auf 40% des Jahresgehaltes, d.h. 14400 Euro, so resultiert der Wert .241. Setzen wir zusätzlich wieder T auf ein Jahr herunter, erhalten wir $d_{break-even}$ = .483 als konservativste Schätzung.

Cohen (1992) hat aus den Erfahrungswerten vieler Untersuchungsbereiche folgende Faustregeln zur Bewertung von Effektgrößen empfohlen: d = .20, .50 und .80 als klein, mittel und groß. Mit diesen Faustregeln sehen wir, dass der Break-Even-Point zwischen kleinen und mittelgroßen Effekten liegt. Aus metaanalytischer Forschung (Smith, Glass & Miller 1980) wissen wir, dass die durchschnittlichen Effekte für psychotherapeutische Interventionen bei d = .80 liegen, für die deutschsprachige Psychotherapieforschung liegen ähnliche Effektgrößen vor (Wittmann und Matt 1986). Lipsey und Wilson (1993) haben eine beeindruckende Metaanalyse aller bisher durchgeführten Metaanalysen zu psychologischen, pädagogischen und behavioralen Interventionen vorgelegt, die einen schnellen Überblick über die Wirksamkeit von Interventionen in vielen Anwendungsfeldern erlauben und eine ex-ante Einschätzung des zu erwartenden Nutzens erlauben.

Unter Verwendung von Gl. 3 sehen wir, dass die ROIs zwischen 1,66 bei sehr konservativer Betrachtung und 5,80 bei optimistischer Betrachtung liegen. ROIs von 1,66 sind bereits als sehr gut zu bewerten. Obwohl diese Sensitivitätsanalyse eine breite Spanne aufweist, so liegt sie doch immer in einem Bereich, der keinen Vergleich mit anderen Investitionsentscheidungen zu scheuen braucht.

Eine medizinische Maßnahme, die wenigstens mittelgroße Effekte d > .50 erzielt, erreicht bei diesem Szenario immer ROI's größer als 1. Solche a-priori Analysen können Befürchtungen einer Kosten-Nutzen-Analyse nicht gewachsen zu sein, schon im Vorfeld zerstreuen und helfen die Akzeptanz für systematische Evaluationen zu erhöhen. Im Gegenteil wir erhalten hier sogar den Eindruck, dass der rein monetäre Nutzen solcher Interventionen massiv unterschätzt wird.

4. Ein Beispiel einer umfassenden Programmevaluationsstudie im Gesundheitssystem, die Zauberbergstudie

Zahlreiche Patienten mit unklaren medizinischen Befunden und langen Irrwegen durch das klassische medizinische Versorgungssystem, erhalten die Diagnose psychosomatische Erkrankung oder Störung. Diese Patienten können eine stationäre medizinische Rehabilitationsmaßnahme in einer spezialisierten Klinik beantragen, die im Regelfall 4-6 Wochen dauern kann. Welche Effekte und Ergebnisse werden in einer solchen Einrichtung kurz-, mittel- und langfristig erzielt? Welche Forschungsmethoden sind für die Evaluation eines solchen Behandlungsprogramms besonders geeignet? Die Patienten werden auf Antrag entweder von den Rentenversicherungsträgern oder den Krankenkassen überwiesen. Eine Randomisierung der Patienten zu einzelnen Einrichtungen und Maßnahmen ist von vorneherein ausgeschlossen. Als Versuchsplan ergibt sich daraus nur die Möglichkeit, naturalistische Korrelationsdesigns mit mehreren Messwiederholungen und möglichst umfassenden Datenerhebungen zu Kontrollzwecken einzusetzen. Die multiplen Ergebnismaße haben wir entwickelt, um den vielfältigen Zielsetzungen und der Komplexität dieser Behandlungsprogramme gerecht zu werden. Wir realisieren damit ein wich-

tiges Symmetrieprinzip zwischen Behandlung, Ergebnis und den Stakeholderinteressen, das wir als Brunswik-Symmetrie (Wittmann 1988, 1990) bezeichnet haben. Faire Evaluationen sollten immer darauf achten, dass das, was therapiert und verändert wird, auch eine Entsprechung im Kriterium findet. Nach Berücksichtigung unterschiedlicher Stakeholderinteressen erarbeiteten wir ein Kriterium, das aus 27 unterschiedlichen Kriterienmaßen besteht.

Tab. 2: Multiples Ergebniskriterium zum Zeitpunkt der Katamnese EMEK_27

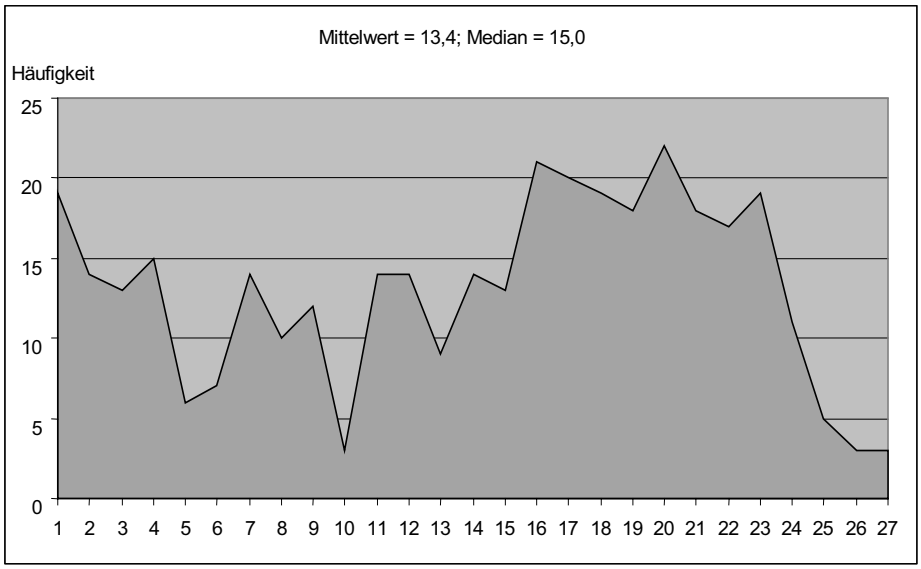
– Veränderung des allgemeinen Zustandes (zum Katamnesezeitpunkt)
– Veränderung „Lebensqualität“
– Veränderung körperliche Verfassung
– Veränderung seelische Verfassung
– Veränderung Allgemeinbefinden
– Veränderung Leistungsfähigkeit
– Veränderung Beschwerden
– Veränderung Gesundheitszustand
– Veränderung Umgang Probleme/alltägl. Belastungen
– Veränderung Beziehung zu Bezugspersonen
– Veränderung Beziehung zum Partner
– Veränderung Familienleben
– Veränderung Arbeitsfähigkeit
– Veränderung Anzahl der Arztbesuche (Jahr vor HB vs. Jahr nach HB)
– Veränderung Wohlbefinden
– Veränderung Umgang mit Problemen
– Veränderung Fähigkeit zur Selbsthilfe
– Veränderung Ertragen von Enttäuschungen
– Veränderung Zurechtkommen mit Arbeit
– Veränderung Belastbarkeit
– Veränderung Auskommen mit Mitmenschen
– Veränderung Leben können mit Einschränkungen/Problemen
– Veränderung Ausgeglichenheit
– Veränderung Tablettenkonsum
– Veränderung Krankenhaustage
– Veränderung Arbeitunfähigkeiten, Fehlzeiten

Anzahl der Komponenten: 27 "Items"
Verrechnung: jede Komponente mit 1/0
1 = positiv bewertbarer Aspekt
0 = neutral bzw. negativ bewertbarer Aspekt
theoretischer Range: 0-27 Punkte

Tab. 2 zeigt die Liste dieser Variablen. Um einen Index zu entwickeln, der auch Stakeholdern, die nicht mit den methodisch-statistischen Datenanalysekonzepten der Forschung vertraut sind, verständlich gemacht werden kann, haben wir eine positive Veränderung mit dem Punktwert 1, keine positive Veränderung mit dem Wert Null skaliert und diese Werte über alle 27 Ergebnisaspekte aufsummiert. Die individuelle Ergebnisqualität kann daher von einem minimalen Wert von Null bis zum

maximalen Erfolgswert von 27 variieren. Abb. 2 zeigt eine solche typische Ergebnisverteilung für eine Kohorte von Patienten aus dem Zauberbergprojekt (Schmidt 1991, Schmidt et al. 1994). Wir sehen, dass der volle Range ausgeschöpft wird. Die große Mehrzahl der Patienten hat sich deutlich verbessert, es gibt aber auch eine kleinere Gruppe von Patienten deren Punktwerte sehr gering sind. Diese Patienten haben offensichtlich von der Behandlung nur wenig profitiert. Aus einer forschungsmethodischen Perspektive stützen diese geringen Punktwerte jedoch die Glaubwürdigkeit der Untersuchung, da wir mit Paracelsus davon ausgehen können, dass es kein Allheilmittel gibt, das allen Patienten in gleicher Weise hilft. Aus differenzierteren Analysen wissen wir, dass zur Gruppe der weniger erfolgreichen Patienten vor allem solche zählen, die lange vor der Rehabilitation bereits einen Rentenantrag gestellt haben. Nach dem Prinzip "Reha vor Rente" müssen sie jedoch erst noch eine Rehamaßnahme durchlaufen, um die Aussichten der Bewilligung zu erhöhen, ein Hinweis darauf, dass dieses gesundheitspolitisch an und für sich vernünftige Prinzip seine Grenzen hat und für manche Rentenversicherte zu spät kommt und contraindiziert ist.

Abb. 2: Verteilung des multiplen Ergebniskriteriums EMEK-27
1 Jahr nach Behandlungsende



N = 367 (Zauberberg-II-Studie)

Der Index EMEK-27 kann natürlich kritisiert werden, Äpfel und Birnen und vieles mehr miteinander zu vermischen. Erinnern wir uns aber daran, wie sinnvoll solche Indizes z.B. in der Volkswirtschaftslehre als Warenkorb, der zum Bruttosozialindex führt, eingesetzt werden, erahnen wir, wie auf solche Kritik reagiert werden kann.

Da die Datenboxkonzeption eine Zeitstruktur beinhaltet, kann nun auch versucht werden, über Pfadanalysen kausale Wirkungsstrukturen zu testen. In unseren

Analysen haben wir uns darauf konzentriert, die Mehrdimensionalität des Interventionsprogramms differenziert abzubilden. Neben der reinen quantitativen Erfassung der Therapiedosis, gemessen in Stunden erhaltener Therapie, haben wir auch einen Fragebogen, bestehend aus 18 Items entwickelt, der aus Sicht der Patienten hilfreiche Aspekte des Kliniksettings am Entlassungszeitpunkt erfragte. Dieser Itempool ließ sich in vier orthogonale Dimensionen faktorisieren. Der erste Faktor erfasste die Qualität der Beziehung zum wichtigsten Bezugstherapeuten, der zweite Faktor die Qualität der Konfrontation und Auseinandersetzung mit den wichtigsten Problemen, der dritte Faktor die Qualität der Beziehung zu den Mitpatienten und der vierte Faktor die Qualität vor allem der nichtpsychotherapeutischen traditionellen Reha- und Kurmaßnahmen. Die Operationalisierung der NTR-Box erfolgte hier deutlich differenzierter als es bei RCTs der Fall ist. Dort wird die randomisierte Interventionsgruppe einer randomisierten Kontroll- oder Vergleichsgruppe gegenübergestellt und diese Unterschiedlichkeit nur über eine Dummy-Variable mit den Werten 1 und 0 abgebildet, eine wahrlich sehr primitive Form der Messung. Frank (1992) hatte als zentrales Konstrukt das Ausmaß der Demoralisierung der Patienten als Kernelement psychologischer Störungen benannt. Erfolgreiche Psychotherapie muss sich daher auf Wirkmechanismen zur Remoralisierung konzentrieren. Nübling (1992) hat deshalb im Kontext der Messung von Psychotherapiemotivation und Krankheitskonzepten eine Skala zur Erfassung des Ausmaßes an Demoralisierung entwickelt und deren Eignung zur Erfassung von differentiellen Änderungen nachweisen können. In diese Skala gingen vor allem Items zu den Aspekten Hoffnung auf Behandlungserfolg und Furcht vor Misserfolg ein.

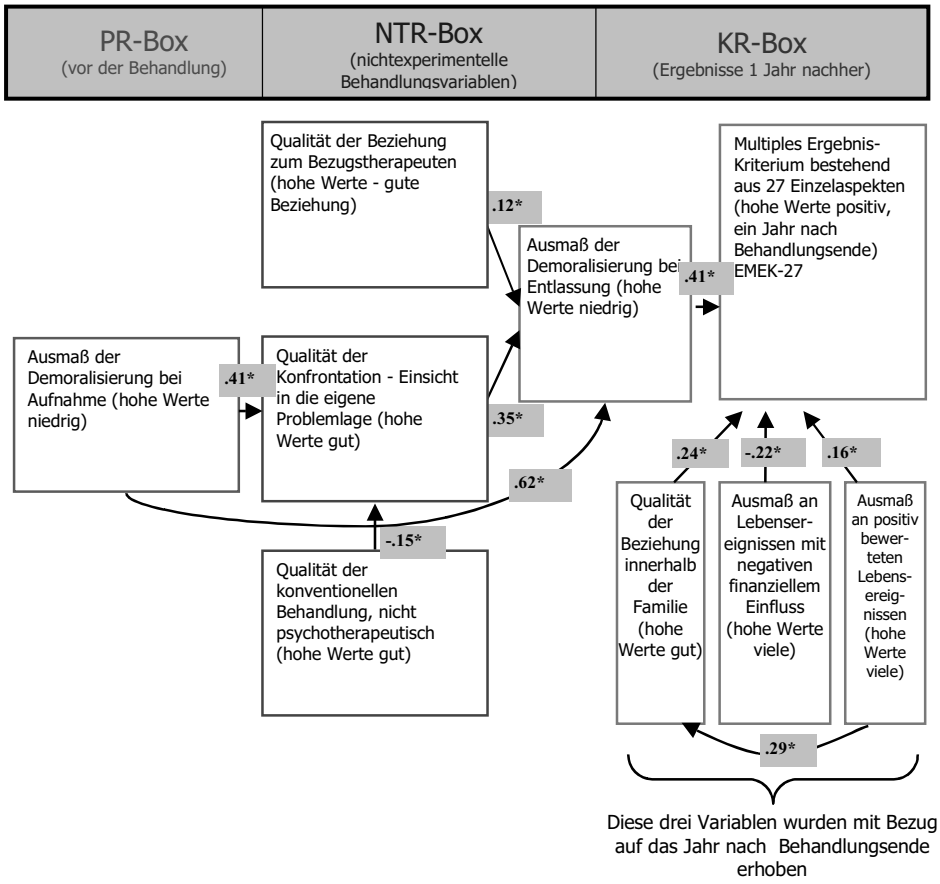
In unserer Pfadanalyse wurde diese Skala zur Erfassung der Ausgangslage (PR-Box) und unmittelbar am Ende der Behandlung, d.h. zum Zeitpunkt der Entlassung (KR-Box) eingesetzt. Das Ausmaß der differentiellen Veränderungen in dieser Skala bezeichnen wir als das Ausmaß der Remoralisierung, das als Funktion der Dimensionalität der Intervention (NTR-Box) erklärt werden soll. Die Remoralisierung sollte sich wiederum positiv auf das multiple Ergebniskriterium ein Jahr nach Ende der Behandlung auswirken. Im Katamnesezeitraum nach der Entlassung können unterschiedliche Lebensereignisse, die unabhängig von den Wirkungen der Intervention betrachtet werden müssen, das multiple Ergebniskriterium positiv oder negativ beeinflussen. Solche Effekte sind als „history“ Effekte oder Effekte des zwischenzeitlichen Geschehens einzustufen. Wir haben diese Gefahren für die interne Validität (Cook & Campbell 1979) ebenfalls als Kontrollvariablen in die Pfadanalyse eingebaut. Die Variablen positiv bewertete Lebensereignisse und Qualität der Beziehung innerhalb der Familie haben einen positiven Effekt, negative Lebensereignisse finanzieller Konsequenzen hingegen einen negativen Effekt auf EMEK-27. Der Pfadkoeffizient von .41 ist der um diese Störvariablen bereinigte direkte Effekt, den die Remoralisierung auf EMEK-27 ausübt. Dieser Effekt kann wiederum als ein eher großer Effekt im Sinne Cohens interpretiert werden.

Die Pfadanalyse erhellt das gesamte Wirkungsgeflecht. Die Qualität der therapeutischen Beziehung und noch viel stärker die Qualität der Auseinandersetzung mit den eigenen Problemen tragen zur Remoralisierung bei, die wiederum den Langzeiterfolg erklärt. Wir sehen auch am negativen Pfadkoeffizienten von -.15, dass Patienten, die nichtpsychotherapeutische Interventionen hinsichtlich ihrer Qualität höher einstufen, sich tendenziell weniger mit den eigenen Problemen auseinander-

setzen und damit eine tendenziell geringere Remoralisierung aufweisen, was einen indirekten negativen Langzeiteffekt auf EMEK-27 zur Folge hat. Bezüglich der Gefahr der Selektion in das Treatment bedeutet der signifikante Pfadkoeffizient von .41, zwischen Demoralisierung vor der Intervention und der Qualität der Auseinandersetzung mit den eigenen Problemen, dass Patienten mit relativ geringerer Demoralisierung sich intensiver mit ihren Problemen auseinandersetzen. Wir erhalten hierdurch einen Hinweis, wie das Programm weiter optimiert werden kann. Für die am stärksten demoralisierten Patienten sollte eine höhere Dosierung oder zusätzliche Interventionen geplant werden, die eine Auseinandersetzung mit den Problemen fördert und die Flucht oder das Ausweichen in nichtpsychotherapeutischen Maßnahmen reduzieren. Das Beispiel zeigt uns, dass bei umfassenden Evaluationsstudien die Trennung zwischen reiner formativer und summativer Evaluation eine künstliche Dichotomie ist. Wir erhalten neben der summativen Bewertung der Effektivität auch wertvolle Hinweise über Struktur und Prozess der Wirkmechanismen, die wiederum zur formativen Optimierung eines bestehenden Programms genutzt werden können.

Nicht alle Stakeholdergruppen werden unser umfassendes Ergebniskriterium EMEK-27 voll akzeptieren und Fragen stellen, ob die positive Gesamtbewertung vor allem durch Komponenten bedingt wurde, an denen sie eigentlich nicht besonders interessiert sind. Als Antwortstrategie können wir den Gesamtindex nur auf die eigentlich interessierenden Komponenten herunterbrechen. Wir haben deshalb auch einen reinen monetären Index aus den fünf Komponenten zur Reduktion der Anzahl der Arbeitsunfähigkeitstage, der Krankenhaustage, der Arztbesuche, des Medikamentenkonsums und der Erwerbstätigkeit des Einjahreszeitraumes vor der Behandlung verglichen mit dem Einjahreszeitraum nach der Behandlung, erfasst am Katamnesesmesszeitpunkt, konstruiert.

Abb. 3: Pfadanalytische (kausale) Modellierung der Behandlungseffekte in der Zauberbergstudie mit dem multiplen Ergebniskriterium EMEK-27. Kausalanalyse über den Südwestpfad der Fünf-Datenbox-Konzeption



Legende:

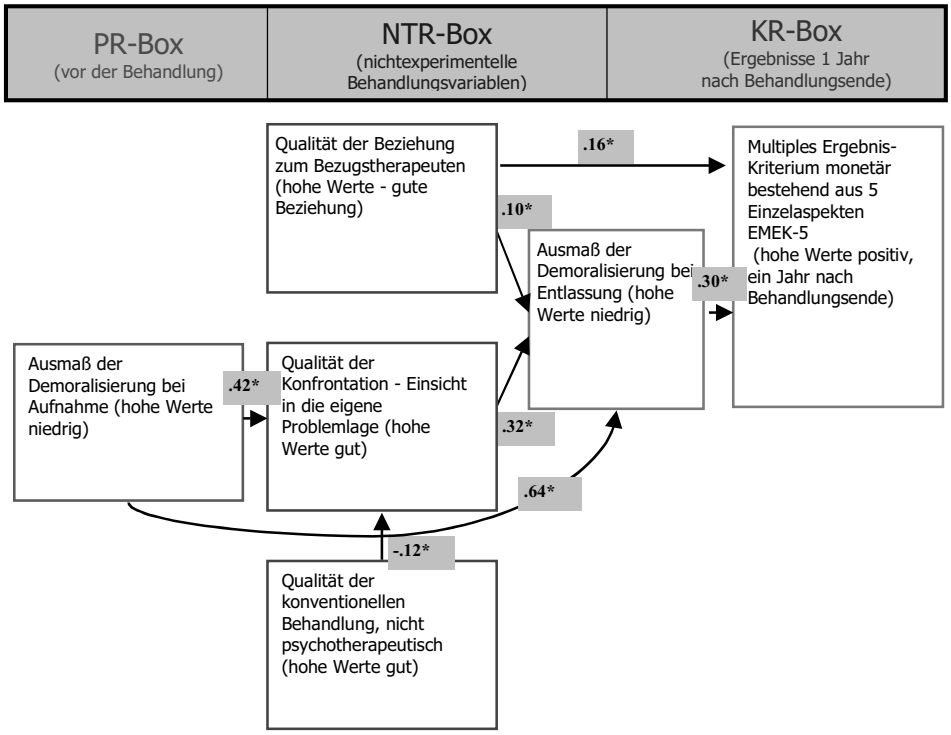
* Signifikante standardisierte Pfadkoeffizienten
$p < .05$

EQS model indices:

CHI-square for the independence model: 319.871 df = 36
CHI-square for postulated model above: 40.20 df = 26 p = .037
Bentler normed fit index: .875
Bentler non normed fit index: .931
Comparative fit index (CFI): .950
RMSEA: .061 (.015 – .095)

R^2_{EMEK27} = .37 (.61)
adj. R^2 = .34
N = 154

Abb. 4: Pfadanalytische (kausale) Modellierung der Behandlungseffekte in der Zauberbergstudie mit einem reinen multiplen monetären Kriterium EMEK-5 Kausalanalyse über den Südwestpfad der Fünf-Datenbox-Konzeption



Legende:
* Signifikante standardisierte Pfadkoeffizienten
$p < .05$

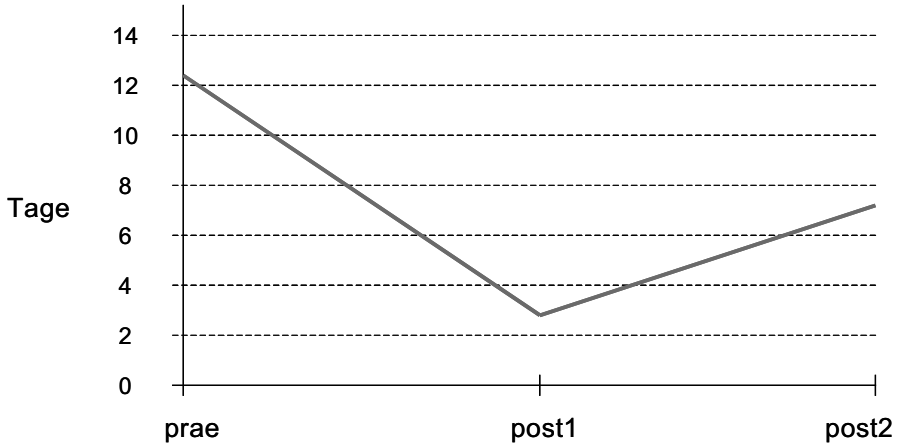
EQS model indices:
CHI-square for the independence model: 228.281 df = 10
CHI-square for postulated model above: 3.233 df = 4 p = .52
Bentler normed fit index: .986
Bentler non normed fit index: 1.006
Comparative fit index (CFI): .950
RMSEA: .000 (.000 – .111)

$R^2_{EMEK5} = .13$ (.36)
adj. $R^2 = .11$
N = 154

In Abb. 4 haben wir nun das Kriterium EMEK-27 durch das reine monetär bewertbare Kriterium EMEK-5 ersetzt und wiederum die pfadanalytische Struktur des Südwestpfades getestet. Die Variablen des zwischenzeitlichen Geschehens aus Abb. 3 hatten nun keine signifikante Effekte mehr, sie wurden deshalb nicht mehr aufgeführt. Alle anderen Pfade blieben jedoch bedeutsam und signifikant. Wir sehen, dass das Ausmaß der Remoralisierung wiederum den wichtigsten Langzeiteffekt auf das monetäre Kriterium ausübt. Neu hinzu kommt noch ein direkter Effekt,

der von der Qualität der Beziehung zum wichtigsten Bezugstherapeuten ausgeht. Diese Ergebnisse erhellen einen bemerkenswerten Sachverhalt, d.h. eine Verbindung von Psychologie und Ökonomie. Die kritische Stakeholdergruppe, die nur monetäre Kriterien akzeptieren würde, muss hier zugestehen, dass die Ergebnisse der Analyse des kausalen Wirkungsgeflechtes auch für das von ihnen bevorzugte Kriterium gelten. Wir können aber noch einen Schritt weitergehen und einzelne monetäre Kriterien heranziehen, an denen die Implikationen einer Kostenreduktion direkt sichtbar werden.

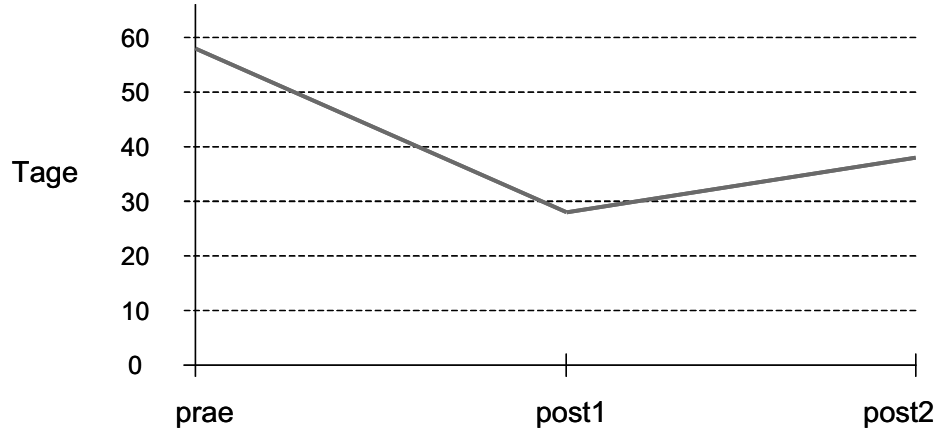
Abb. 5: Vergleich Krankenhaustage in den drei Beobachtungszeiträumen prae, post 1, post 2, N = 139



Zeitraum	x (Tage)	s (Tage)	
prae	12.55	27.57	vor der psychosomatischen Rehabilitation
post1	3.00	12.64	1 Jahr nachher
post2	6.92	15.77	3 Jahre nachher

Abb. 5 und Abb. 6 zeigen den Verlauf der Entwicklung der Krankenhaustage und der Arbeitsunfähigkeitstage direkt über den gesamten Zeitraum der Dreijahreskatamnese. Nach einer starken Reduktion zum Einjahreszeitpunkt steigen die Werte zum dritten Jahr nach Ende der Behandlung wiederum an, erreichen jedoch bei weitem nicht die aggregierten Werte über das Jahr unmittelbar vor der Behandlung. Die Schätzung, dass der Behandlungseffekt T im Schnitt ca. zwei Jahre anhält, ist deshalb mit Sicherheit keine Überschätzung.

Abb. 6: Vergleich Krankschreibungstage in den drei Beobachtungszeiträumen prae, post 1, post 2, N = 101



Zeitraum	x (Tage)	s (Tage)	t-Tests (pairs)
prae	57.70	82.10	a) prae vs post1 T = 4.02; df =100; p = .000 ***
post1	27.56	61.90	b) prae vs post2 T = 2.18; df =100; p = .031 *
post2	38.50	63.28	c) post1 vs post2 T = -1.58; df =100; p = .117 n.s.

5. Ausblick, Konsequenzen und Forderungen

Evaluationsforschung und Programmevaluationen im gesamten Gesundheitswesen liefern vielfältige Möglichkeiten, die brennenden Fragen der Qualitätskontrolle und der Finanzierbarkeit zu beantworten. Die klassischen Wirksamkeitsstudien von medizinischen Interventionen aus den kontrollierten Studien der Grundlagenforschung der Schulmedizin reichen dazu nicht aus. Welcher Wirkungsgrad mit welcher Kosteneffizienz und Kosteneffektivität bei der Umsetzung im realen System erreicht wird und wie er optimiert werden kann, sind Fragen summativer und formativer Evaluation. Die empirischen Sozialwissenschaften haben dazu vielfältige Strategien und Verfahrensweisen angewandter Forschung entwickelt, die systematisch umgesetzt und angewandt werden müssen. Ein entscheidendes Problem besteht darin, wie dieses Wissen in das System des Gesundheitswesen transferiert und umgesetzt werden kann. Die Ärzte sind in der Regel sehr schlecht in den Forschungsmethoden, wie Versuchsplanung und statistischer Datenanalyse ausgebildet. Die medizinische Ausbildung ist bereits durch eine sehr große Stofffülle geprägt, die wenig Raum für zusätzliche Inhalte lässt. Die Lehrstühle und Abteilungen zur medizinischen und statistischen Dokumentation übernehmen meist eine gewisse Beratungs- und Ausbildungsfunktion, die dem eigentlichen Bedarf für systematische Evaluationsforschung aber oft nicht gewachsen ist. In vielen medizinischen Fakultäten sind aber seit einigen Jahrzehnten Professuren und Lehrstühle für medizinische Psycho-

logie und/oder Soziologie eingerichtet worden. Da die meisten Forschungsmethoden zur Programmevaluation aus den empirischen Sozialwissenschaften stammen, könnte gerade dieses Know-how auch in den Curricula dieser Abteilungen über konkrete Projekte vermittelt werden.

Die vielleicht beste Lösung ist allerdings die Einbindung von spezifischen Ausbildungsprogrammen mit fokussierten Abschlüssen in die medizinischen Fakultäten, wie wir sie in den „Schools of Public Health" in den USA finden. In der Bundesrepublik Deutschland hat man zumindest in Ansätzen an einigen Universitäten Gesundheitswissenschaften mit vergleichbaren Zielsetzungen etabliert. Es bleibt abzuwarten, welche Wirkungen für die Evaluationsforschung im Gesundheitswesen insgesamt daraus resultieren. Im System des Rehabilitationswesens finden wir die bemerkenswerte Entwicklung, dass dort bereits der größte Teil der Programmevaluationen und der Evaluationsforschung von Psychologen und Soziologen durchgeführt wird, was an den Beiträgen zu Fachkongressen und Publikationen leicht ablesbar ist. Erfolge und Verbesserungen von medizinischen Therapien resultierten zu einem hohen Anteil aus der Integration der Naturwissenschaften wie Physik, Chemie und Biologie in die Medizin und es ist unschwer zu prognostizieren, dass dieser Trend weiter anhalten wird. Wir wagen jedoch auch die Prognose, dass durch bessere Einbindung der empirischen Sozialwissenschaften die Optimierung und Steuerung des Gesundheitswesens gerade hinsichtlich Effektivität, Wirksamkeit und der Kosten-Nutzen sowie der Kosten-Effektivitätsrelation wichtige und entscheidende Impulse erhalten wird.

Literatur:

Amann, K. (1997): Verlaufsänderung von Gesundheitsverhalten und Risikofaktoren bei Patienten einer psychosomatischen Klinik. Unveröff. Med. Diss., MMH Hannover.

Attkisson, C.C. & Broskowski, A. (1978): Evaluation and the emerging human service concept. In: C.C. Attkisson, W.A. Hargreaves, M.J. Horrowitz & J.E. Sorensen (Eds.): *Evaluation of human service programs.* New York: Academic Press.

Beywl, W. (1988): Zur Weiterentwicklung der Evaluationsmethodologie. Frankfurt: Verlag Peter Lang.

Bischoff, C., Ehrhardt, M., Limbacher, K. & Husen, E. (2000): Ambulante prä- und poststationäre Maßnahmen zur Optimierung zielorientierter psychosomatischer Rehabilitation. In: Bengel, J.; Jäckel, W.H. (Hg.): *Zielorientierung in der Rehabilitation – Rehabilitationswissenschaftlicher Forschungsverbund Freiburg/Bad Säckingen*. Regensburg, Roderer, 39-4895-106.

Broda, M.; Bürger, W.; Dinger-Broda, A. & Massing, H. (1996): *Die Berus-Studie*. Berlin, Bonn, Westkreuz.

Brogden, H.E, (1949): When testing pays off. *Personnel Psychology* 2: 171-183.

Bürgy, R.; Nübling, R.; Meyerberg. J.; Oppl, M.; Kieser, J.; Schmidt, J. & Wittmann, W.W. (2000): Stationäre psychosomatische Rehabilitation im Rahmen eines schulenübergreifenden Behandlungskonzepts: Ergebnisse der 1-Jahres-Katamnese der Bad Herrenalber Katamnesestudie. In: VDR (Hg): *9. Rehabilitationswissenschaftliches Kolloquium, Individualität und Reha-Prozeß.* DRV-Schriften, Band 20: 374.

Campbell, D.T. & Stanley, J.C. (1966): *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

Campbell, D.T. (1969): Reforms as experiments. In: *American Psychologist*, 24: 409-429.

Cattell, R.B. (1957): *Personality and motivation structure and measurement*. New York: World Book Company.

Cohen, J. (1992): A power primer. *Psychological Bulletin*, 112: 155-159.
Cook, T.D. & Campbell, D.T. (1979): Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.
Dilcher, K., Mestel, R., Klingelhöfer, J., Köbel, W., Sprenger, B., Stauss, K. (2000): Psychosomatische Kliniken. In: Gerdes, N., Weidemann, H., Jäckel, W.H. (Hg.): Die Protos-Studie. Darmstadt: Steinkopff: 173-202.
Donabedian, A. (1966) Evaluating the quality of medical care. In: Milbank Memorial Fund Quarterly, 44: 166-203.
Foxhall, K. (2000): Research for the real world. In: Monitor on Psychology, 31, No. 7 July/August. Washington, DC: APA-Publications.
Frank, J. D. (1992): Wirkungsweisen psychotherapeutischer Beeinflussung. Vom Schaminismus bis zu den modernen Therapien. Stuttgart: Klett Cotta.
Glass, G.V. (1983): Evaluation methods synthesized. Review of L.J. Cronbach designing evaluations of educational and social programs. In: Contemporary Psychology, 28: 501-503.
Gerdes, N., Weidemann, H. & Jäckel, W.H. (2000): Die PROTOS-Studie. Darmstadt: Steinkopff.
Hillert, A., Maasche, B., Kretschmer, A. & Fichter, M.M. (2000): Psychosomatisch erkrankte LehrerInnen im poststationären Verlauf: Halbjahreskatamne zum Priener Lehrerprojekt. In: VDR (Hg): 9. Rehabilitationswissenschaftliches Kolloquium, Individualität und Reha-Prozeß. DRV-Schriften, Band 20: 449-451.
Hillert, A., Maasche, B., Kretschmer, A., Ehrig, C., Schmitz, E. & Fichter, M.M. (1999): Psychosomatische Erkrankungen bei LehrerInnen. In: PPmP Psychther Psychosom med Psychol, 49: 375-380.
Kächele, H., für die Studiengruppe MZ-ESS (1999): Eine multizentrische Studie zu Aufwand und Erfolg bei psychodynamischer Therapie von Eßstörungen. In: PPmP Psychther Psychosom med Psychol, 49: 100-108.
Kriebel, R., Schmitz-Buhl, S.M. & Paar, G.H. (1999): Dauer der Behandlungswirkung (Katamnese). In: Kriebel, R. & Paar, G.H. (Hg.): Psychosomatische Rehabilitation: Möglichkeit und Wirklichkeit. Geldern, Verlag Johannes Keuck: 127-148.
Lamprecht, F. & Schmidt, J. (1990): Das Zauberberg-Projekt: Zwischen Verzauberung und Ernüchterung. In: Ahrens: (Hg): Entwicklung und Perspektiven der Psychosomatik in der Bundesrepublik Deutschland. Berlin etc, Springer: 97-115.
Lipsey, M.W. & Wilson, D.B. (1993): The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. In: American Psychologist, 48: 1181-1209.
Maatz, E. & Schmidt, J. (1998): Psychosomatische Rehabilitation von Patienten mit chronisch-entzündlichen Darmerkrankungen – Erste Ergebnisse der Gengenbacher CED-Studie. In: Verband Deutscher Rentenversicherungsträger (Hg): Interdisziplinarität und Vernetzung. DRV-Schriften Band 11. Frankfurt, VDR: 462-463.
Mestel, R., Erdmann, A., Schmidt, M., Klingelhöfer, J., Stauss, K. & Hautzinger, M. (2000b): Verläufe nach stationärer psychosomatischer Rehabilitation von depressiven Patienten. In: VDR (Hg): 9. Rehabilitationswissenschaftliches Kolloquium, Individualität und Reha-Prozeß. DRV-Schriften, Band 20: 385-386.
Mestel, R., Neeb, K., Hauke, B., Klingelhöfer, J. & Stauss, K. (2000a): Zusammenhänge zwischen der Therapiezeitverkürzung und dem Therapieerfolg bei depressiven Patienten. In: Bassler, M. (Hg.): Empirische Forschung in der stationären psychosomatischen Rehabilitation. Psychosozial, Gießen.
Nübling, R., Bürgy, R., Meyerberg, J., Oppl, M., Kieser, J., Schmidt, J. & Wittmann, W.W. (2000a): Stationäre psychosomatische Rehabilitation in der Klinik Bad Herrenalb: Erste Ergebnisse einer Katamnesestudie. In: Bassler, M (Hg): Leitlinien zur stationären Psychotherapie. Gießen, Psychosozial-Verlag (im Druck).
Nübling, R., Puttendörfer, J.; Wittmann, W.W.; Schmidt, J. & Wittich, A. (1995): Evaluation psychosomatischer Heilverfahren. Ergebnisse einer Katamnese-Studie. In: Die Rehabilitation, 34: 74-80.

Nübling, R.; Hafen, K.; Jastrebow, J.; Schmidt J. & Bengel, J. (2000b): Indikation zu psychotherapeutischen und psychosozialen Maßnahmen in der stationären Rehabilitation. In: Bengel, J.; Jäckel W.H. (Hg.): Zielorientierung in der Rehabilitation – Rehabilitationswissenschaftlicher Forschungsverbund Freiburg/Bad Säckingen. Regensburg, Roderer: 95-106.

Nübling, R.; Puttendörfer, J.; Schmidt, J. & Wittmann, W.W. (1994): Längerfristige Ergebnisse psychosomatischer Rehabilitation. In Lamprecht, F; Johnen, R (Hg.): Salutogenese. Ein neues Konzept in der Psychosomatik? Frankfurt/Main, Verlag für Akademische Schriften VAS: 254-270.

Nübling, R.; Schmidt, J. & Wittmann, W.W. (1999): Langfristige Ergebnisse psychosomatischer Rehabilitation. In: PPmP Psychother Psychosom med Psychol, 49: 343-353.

Rossi, P.H. (1978): Issues in the evaluation of human services delivery. In: Evaluation Quarterly, 2: 573-599.

Rossi, P.H., Freeman, H.E. and Lipsey, M.W. (1999): Evaluation. A systematic approach. 6th ed. Thousand Oaks: Sage Publications.

Rüddel, H.; Jürgensen, R.; Lotz-Rambaldi, W. et al. (1999): 3-Jahres-Katamnese von Patienten nach psychosomatischer Rehabilitation. In: Schliehe, F.; Schuntermann, M.F. (Hg): 8. Rehabilitationswissenschaftliches Kolloquium vom 8.-10.3.1999 auf Norderney. Reha-Bedarf – Effektivität – Ökonomie. DRV-Schriften Band 12. Frankfurt, VDR: 375.

Rudolf, G.; Grande, T. & Prosch, U. (1991): Die therapeutische Arbeitsbeziehung. Heidelberg, Berlin etc, Springer.

Sandweg, R.; Sänger-Alt, C. & Rudolf, G. (1991): Erfolge in der stationären Psychotherapie – Ergebnisse eines Katamneseprojekts in einer Fachklinik für psychogene Erkrankungen. In: Öff Gesundh-Wes, 53: 801-809

Schmidt, F.L., Hunter, J.E. & Pearlman, K. (1982): Assessing the economic impact of personnel programs on workforce productivity. In: Personnel Psychology 35. S. 333-347.

Schmidt, J. (1991). Evaluation einer psychosomatischen Klinik. Frankfurt/Main, Verlag für Akademische Schriften VAS.

Schmidt, J.& Lamprecht, F. (1992): Psychosomatische Rehabilitation. Ergebnisse von Verlaufsstudien. In: Bundesarbeitsgemeinschaft für Rehabilitation BAR (Hg): Rehabilitation – Zukunft 2000. Essen, A. Sutter-Messe-Verlag: 261-267.

Schmidt, J., Lamprecht, F., Nübling, R. & Wittmann, W.W. (1994): Veränderungsbeurteilungen von Patienten und von Haus- und Fachärzten nach psychosomatischer Rehabilitation – Ein katamnestischer Vergleich. In: PPmP Psychother Psychosom med Psychol, 44: 108-114.

Schmidt, J.; Karcher, S. & Nübling, R. (1999): Ergebnisevaluation psychosomatischer Rehabilitation. Vortrag Kogreß „Brennpunkte der Psychiatrie: forum Rehabilitation, 6.-8. Mai 1999 in Hamburg.

Schmidt, J.; Karcher, S.; Steffanowski, A.; Nübling, R. & Wittmann, W.W. (2000a): Die EQUA-Studie – Erfassung der Ergebnisqualität stationärer psychosomatischer Rehabilitationsbehandlungen – Vergleich unterschiedlicher Evaluationsstrategien und Entwicklung neuer Messintrumentarien. In: Bengel J.; Jäckel, W.H. (Hg.): Zielorientierung in der Rehabilitation – Rehabilitationswissenschaftlicher Forschungverbund Freiburg/Bad Säckingen. Regensburg, Roderer: 109-118.

Schmidt, J.; Nübling, R. & Wittmann, W.W. (2000b): Ergebqualität stationärer psychosomatischer Rehabilitation nach einem Jahr. Die Patientenperspektive in 5 Programmevaluationsstudien mit auf der Basis von fünf Programmevaluationsstudien mit 2098 Patienten. Prax. Klin. Verhaltensmed. Rehab. (im Druck).

Schöffski, O. & Graf v.d. Schulenburg, J.M. (Hg.) (2000): Gesundheitsökonomische Evaluationen. Berlin: Springer.

Schulz, H.; Lotz-Rambaldi, W.; Koch, U.; Jürgensen, R. & Rüddel, H. (1999): 1-Jahres-Katamnese stationärer psychosomatischer Rehabilitation nach differentieller Zuweisung zu psychoanalytisch und verhaltenstherapeutisch orientierter Behandlung. In: PPmP Psychother Psychosom med Psychol, 49: 114-130.

Scriven, M. (1988): Philosophical inquiry methods in education. In: R.M. Jaeger (Ed.): Complementary Methods for research in education. Washington, DC: American Educational Research Association, pp. 131-183.

Shadish, W. R., Cook, T.D. & Leviton, L. C. (1991): Foundations of program evaluation. Newbury Park, CA: Sage Publications.

Smith, M.L., Glass, G.V. & Miller, T.I. (1980): The benefits of psychotherapy. Baltimore, MD: Johns Hopkins University Press.

Stevens, A. Abrams, K., Brazier, J., Fitzpatrick, R. & Lilford, R. (Eds.) (2001): The Advanced Handbook of Methods in Evidence Based Healthcare. London: Sage.

Stockmann, R. (Hg.) (2000): Evaluationsforschung. Grundlagen und ausgewählte Forschungsfelder. Opladen: Leske u. Budrich.

Tigiser: (1997): Qualitätssicherung in einer psychosomatischen Rehabilitationsklinik am Beispiel der Evaluationsstudie in der Eifelklinik Manderscheid. Unveröff. Diplomarbeit, Universität Koblenz-Landau, Abteilung Landau, Fachbereich 8: Psychologie.

Wilke, S., Grande, T., Rudolf, G. & Porsch, U. (1988): Wie entwickeln sich Patienten im Anschluß an eine stationäre Psychotherapie? In: Zts. Psychosom. Med., 34: 107-124.

Wilson, D.B. & Lipsey, M.W. (2001): The role of method in treatment effectiveness research: Evidence form meta-analysis. In: Psychological Methods, 6: 413-429.

Wittmann, W.W. & Walach, H. (2001): Evaluating complementary medicine: Lessons to be learned from evaluation research. In: G. Lewith, W.B. Jonas & H. Walach (Eds.): Clinical research in complementary therapies. Principles, problems and solutions. London: Churchill Livingston, p. 93-108.

Wittmann, W.W. (1988): Multivariate reliability theory: Principles of symmetry and successful validation strategies. In: Nesselroade, J.R. & Cattell, R.B. (Hg.): Handbook of multivariate experimental psychology. 2nd ed.. New York: Plenum Press: 505-560.

Wittmann, W.W. (1996): Evaluation in der Rehabilitation. Wo stehen wir heute? In: DRV-Schriften, 6: 27-37.

Wittmann, W.W. (1990): Brunswik-Symmetrie und die Konzeption der Fünf-Datenboxen. Ein Rahmenkonzept für umfassende Evaluationsforschung. In: Zeitschrift für Pädagogische Psychologie, 4: 241-251.

Wittmann, W.W. & Matt, G.E. (1986): Meta-Analyse als Integration von Forschungsergebnissen am Beispiel deutschsprachiger Arbeiten zur Effektivität von Psychotherapie. In: Psychologische Rundschau, 37: 20-40.

Zielke, M. (1993): Wirksamkeit stationärer Verhaltenstherapie. München, Psychologie Verlags-Union.