Evidenz psychosomatischer Rehabilitation im Spiegel multipler Ergebniskriterien

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Sozialwissenschaften
der Universität Mannheim

vorgelegt von
Dipl.-Psych. Andrés Steffanowski

Universität Mannheim
Fakultät für Sozialwissenschaften

April 2008

Dekan der Fakultät für Sozialwissenschaften:

Professor Dr. Josef Brüderl

Gutachter:

Prof. Dr. Werner W. Wittmann (Universität Mannheim)

Prof. Dr. Michael Bosnjak (Universität Mannheim, seit 03/2008 Universität Bozen, Italien)

Eingereicht am:

05.04.2008

Tag der Disputation:

04.09.2008

Vorwort

Die stationäre psychosomatische Rehabilitation ist mit etwa 100.000 behandelten Patienten pro Jahr ein bedeutsamer Versorgungssektor für psychische Erkrankungen in der Bundesrepublik Deutschland. In den vergangenen 20 Jahren wurden hier eine Vielzahl von Studien, davon mehrere umfassende Programmevaluationsstudien, an größeren Patientenstichproben durchgeführt, die insbesondere die Effektivität der Behandlungen zum Gegenstand hatten. Erst kürzlich wurde eine entsprechende meta-analytische Übersicht vorgelegt, welche die Ergebnisqualität der durchgeführten Maßnahmen zur stationären psychosomatischen Rehabilitation eindrucksvoll belegt und deren Bedeutung im medizinischen Versorgungssystem unterstreicht.

Eng verbunden mit der Frage nach der Ergebnisqualität sind neben inhaltlichen Aspekten auch immer methodische Erwägungen, welche evaluative Strategie im Kontext der jeweils gegebenen Versorgungsrealität innerhalb einer Einrichtung am ehesten geeignet ist, um den Behandlungserfolg möglichst zuverlässig und dennoch praxisnah abzubilden. Ein methodischer Zugang besteht dabei in der Aggregation einer Reihe von singulären Ergebniskriterien (z.B. körperliches Befinden, psychisches Befinden, soziale Situation und Alltagsbewältigung) zu einem multiplen Ergebniskriterium. Dies geschieht mit dem Ziel, zu einer umfassenderen Gesamtbewertung des Behandlungserfolges zu gelangen, als dies mit einer isolierten Betrachtung einer Vielzahl von Einzelaspekten möglich wäre.

Der Ansatz der multiplen Ergebniskriterien wurde bereits Ende der 1980er Jahre entwickelt und in den Folgejahren in mehreren Programmevaluationsstudien erfolgreich angewandt, so dass nun eine hinreichend große Datenbasis vorliegt, um den Ansatz einer methodischen Revision und Weiterentwicklung zu unterziehen. Dies ist Thema der hier vorgelegten Dissertation.

Diese Arbeit wäre ohne die Mithilfe Dritter nicht möglich gewesen. Mein Dank richtet sich zunächst an unsere Arbeitsgruppe, namentlich Dr. Rüdiger Nübling, Dr. Jürgen Schmidt, Dr. Manuel Völkle, Dr. Stephanie Lichtenberg und Dipl.-Psych. David Kriz für die langjährige erfolgreiche und inspirierende Zusammenarbeit in mehreren Programmevaluationsstudien. Meinen besonderen Dank möchte ich Professor Werner W. Wittmann widmen, der als mein akademischer Mentor an der Universität Mannheim meine Forschungsarbeit und die in dieser Arbeit verwendeten Programmevaluationsstudien unserer Arbeitsgruppe in den letzten Jahren mit seinem Evaluationsansatz maßgeblich geprägt hat, indem er immer wieder auf die Wichtigkeit einer Synthese wissenschaftlicher Forschungsbefunde und deren anschauliche Vermittlung in die Praxis hingewiesen hat.

Mein Dank geht auch an alle Patienten, Mitarbeiter und Träger der an den Programmevaluationsstudien beteiligten Kliniken für die zeitaufwendige Unterstützung der Studienorganisation und Erhebung der Daten vor Ort. Darüber hinaus möchte ich dem Bundesministerium für Bildung und Forschung (BMBF) sowie der Deutschen Rentenversicherung Bund (DRV) für die Etablierung des Förderschwerpunktes Rehabilitationswissenschaften

danken, ohne dessen finanzielle Förderung mehrere der im Rahmen dieser Arbeit verwendeten Studien in dieser Form nicht realisierbar gewesen wären.

Meiner Familie sowie allen Freunden und Kollegen, die mich bei der Entstehung dieser Arbeit mit ihren Ideen und Anmerkungen begleitet und immer wieder ermutigt haben, möchte ich ebenfalls danken. Mein ganz spezieller Dank richtet sich dabei an meine Frau Juliana und an meinen Sohn Jonas für ihre liebevolle und geduldige Unterstützung. Vielen dafür, dass ihr mir den Rücken freigehalten habt. Ihr habt einen wesentlichen Beitrag dazu geleistet, dass diese Arbeit möglich wurde.

Mannheim, den 5. April 2008 Andrés Steffanowski

Zusammenfassung

Die stationäre psychosomatische Rehabilitation hat in Deutschland einen wichtigen Stellenwert bei der Behandlung von psychischen Störungen. In den letzten 20 Jahren sind eine Reihe von PRÄ-POST-Programmevaluationsstudien durchgeführt worden, die eindrucksvolle Belege für die Effektivität dieses Versorgungsbereiches erbracht haben. Im ersten Abschnitt der vorliegenden Arbeit wird der Bereich der psychosomatischen Rehabilitation mitsamt der aktuellen Forschungslandschaft skizziert.

Zur evaluativen Beurteilung der Ergebnisqualität existieren unterschiedliche methodische Zugänge. So finden im Rahmen der allgemeinen Ergebnismessung direkte, indirekte und quasi-indirekte Veränderungsmessungen Anwendung. Als individualisierte Verfahren der Erfolgsbewertung sind das Goal-Attainment-Scaling sowie die Zielorientierte Ergebnismessung zu nennen. All diese Zugänge sind jeweils mit bestimmten Vorzügen und Nachteilen verbunden. Eine kritische Auseinandersetzung mit den verschiedenen Evaluationsansätzen erfolgt im zweiten Kapitel.

Kernthema der vorliegenden Arbeit sind multiple Ergebniskriterien. Darunter wird die Aggregation einer Reihe von inhaltlich heterogenen singulären Ergebniskriterien zur Beurteilung der Ergebnisqualität verstanden. So hat sich in der Ergebnisforschung der aus der meta-analytischen Methodik stammende Ansatz etabliert, vor Durchführung einer Aggregation eine z-Standardisierung von Mittelwertsdifferenzen in Form von Effektgrößen durchzuführen. Darüber hinaus existiert ein weiterer Ansatz, der im Jahr 1987 von Schmidt, Bernhard, Wittmann und Lamprecht entwickelt wurde. Dieser basiert darauf, „wünschenswerten" Ergebnissen im Sinne einer Verbesserung des Befindens auf Itemebene den Wert Eins, „nicht wünschenswerten" Ergebnissen im Sinne eines unveränderten oder verschlechterten Befindens hingegen den Wert Null zuzuweisen und auf diese Weise codierten singulären Ergebniskriterien sodann aufzusummieren.

Das dritte Kapitel enthält eine entsprechende Reanalyse des von Schmidt et al. (1987) entwickelten 27 Items umfassenden multiplen Ergebniskriteriums anhand der Daten von fünf größeren Programmevaluationsstudien. Hierbei wird neben der herkömmlichen auf dichotomen Items basierenden Variante (EMEK_27a) eine alternative Skalenvariante (EMEK_27b) entwickelt, welche die gesamte Iteminformation ausschöpft. Hintergrund dabei ist die Überlegung, dass durch jede Dichotomisierung Varianz verloren geht. Das Ausmaß des entsprechenden Informationsverlusts auf Item- und Skalenebene bei der herkömmlichen Skala EMEK_27a wird daher im Vergleich zur Variante EMEK_27b untersucht, welche die gesamte verfügbare Iteminformation ausschöpft. Die Datenanalysen zeigen, dass der Informationsverlust auf Skalenebene weniger dramatisch ist als auf Itemebene. Beide Skalenvarianten weisen hinsichtlich Reliabilität und Validität fast vergleichbare Kennwerte auf, wenngleich EMEK_27a im Gegensatz zur neu konstruierten Variante EMEK_27b Boden- und Deckeneffekte aufweist.

Es wird daher die Empfehlung ausgesprochen, in Kontexten, wo es auf möglichst hohe wissenschaftliche Präzision ankommt, etwa bei indikativen Entscheidungen die Variante

EMEK_27b mit vollständiger Informationsausschöpfung zu verwenden. Geht es hingegen um eine möglichst anschauliche Vermittlung der Ergebnisse in die Praxis, etwa im Rahmen eines Routinescreenings der Ergebnisqualität für ein fortlaufendes Qualitätsmanagement, so ist die herkömmliche auf binären Items basierende Variante EMEK_27a durchaus brauchbar und statistisch wenig geschulten Gesprächspartnern wie politischen Entscheidungsträgern in der Praxis sogar eher vermittelbar. Ungelöst bleibt bei der auf direkten Veränderungsmessungen basierenden Skala EMEK_27 ist - unabhängig vom Grad der Informationsausschöpfung durch dichotome oder kontinuierliche Items – das methodische Problem der fehlenden Berücksichtigung des Ausgangszustandes der Patienten zu Beginn der Behandlung. So kann die Information „unverändert" (entspricht einem Nulleffekt) zweierlei beinhalten: die Beibehaltung eines ungünstigen Zustandes oder aber die Beibehaltung eines günstigen Zustandes. Bei näherer Betrachtung tritt dieses Ausgangswertproblem allerdings auch bei der indirekten bzw. quasi-indirekten Veränderungsmessung auf, wenn es sich um reine PRÄ-POST-Untersuchungen ohne Kontrollgruppe handelt. So wird bei einer ausschließlichen Betrachtung von PRÄ-POST-Effektgrößen oder direkten Veränderungsmaßen der präventive Gedanke einer Beibehaltung von erwünschten Zuständen durch die herkömmliche Methodik der Ergebnismessung nur unzureichend berücksichtigt, was nicht im Sinne einer „fairen" Evaluation ist. Zwar wurden in der Vergangenheit zur Behebung dieses Problems mathematische Korrekturverfahren und individualisierte Ansätze zur Ergebnismessung vorgeschlagen, die jedoch zu anderweitigen methodischen Problemen führen wie mangelnde Vergleichbarkeit der berechneten Kennwerte, Selektion von Extremgruppen sowie ungelöste Fragen hinsichtlich der Reliabilität. Im vierten Abschnitt der vorliegenden Arbeit wird daher mit der Konzeption eines Composit-Kriteriums eine innovative Strategie zur Bewertung der Ergebnisqualität anhand von PRÄ-POST-Messungen vorgeschlagen, welche sowohl den präventiven (Beibehaltung von erwünschten Zuständen) als auch rehabilitativen Aspekt (Verbesserung des Befindens) berücksichtigt. Der neue Kennwert setzt sich dabei aus z-standardisierten Status- und Veränderungsinformationen zusammen. Mehrere singuläre Composit-Ergebniskriterien lassen sich wiederum durch Summenbildung zu multiplen Ergebniskriterien aufaggregieren.

Eine erste Erprobung, deren Ergebnisse im fünften Kapitel wiedergegeben sind, liefert vielversprechende Anhaltspunkte dafür, dass das Composit-Kriterium eine sehr differenzierte Bewertung der Ergebnisqualität ermöglicht, die mit den zur Validierung herangezogenen Maßen im Einklang steht. Insbesondere die bei der klassischen indirekten Veränderungsmessung häufig zu beobachtende negative Korrelation zwischen Ausgangs- und Differenzwerten, die zu dem unerwünschten statistischen Phänomen der Regression zur Mitte und damit zu einer Verzerrung der Ergebnisse führt, tritt hier nicht auf. Darüber hinaus zeigt das neue Composit-Kriterium ausgezeichnete statistische Kennwerte und Verteilungseigenschaften, die denen herkömmlicher Statusmessungen oder PRÄ-POST-Differenzen (Effektgrößen) in nichts nachstehen.

Inhaltsverzeichnis

Abkürzungsverzeichnis

AU	Arbeitsunfähigkeit (Krankschreibung)
BE*SD*	Binomial Effect Size Display
BfA	Bundesversicherungsanstalt für Angestellte
BKK	Bundesverband der Betriebskrankenkassen
BMBF	Bundesministerium für Bildung und Forschung
c	Grenzwert, an dem Wahrscheinlichkeit für einen Probanden gleich groß ist, der gesunden und kranken Population anzugehören
CI	Vertrauensintervall (Confidenzintervall)
CR	Kriterium (abhängige Variable)
d	PRÄ-POST-Effektgröße (Mittelwertsdifferenz, dividiert durch die Streuung der PRÄ-Messung)
DIMDI	Deutsches Institut für Medizinische Dokumentation und Information
df	Freiheitsgrade (degrees of freedom)
dVM	Direkte Veränderungsmessung zum POST-Zeitpunkt
EM	Expectation-Maximization-Algorithmus
EMEK	Einmalig gemessenes multiples Ergebniskriterium
EQ	Compositkriterium zur Beurteilung der Ergebnisqualität
EQUA	Multizentrische Studie zur Ergebnisqualität stationärer psychosomatischer Rehabilitation – Vergleich unterschiedlicher Evaluationsstrategien (Schmidt et al., 2003)
ETR	Experimentelles Treatment (randomisierte, kontrollierte Studie)
EVA	Bewertungskriterien der an einer Evaluation interessierten Stakeholder
GAS	Goal Attainment Scaling
GKV	Gesetzliche Krankenversicherung
GRV	Gesetzliche Rentenversicherung
h^2	Kommunalität
ICD-10	Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme
ICF	International Classification of Functioning, Disability and Health
ICIDH-2	International Classification of Impairments Disabilities and Handicaps
IES	Individuelle Effektgröße nach Grawe & Braun (1994).
IRES	Fragebogen zur Erfassung der Indikatoren des Reha-Status
iVM	Indirekte Veränderungsmessung (PRÄ-POST-Differenz)
KTT	Klassische Testtheorie
LVA	Landesversicherungsanstalt (Rentenversicherung für Arbeiter)

M	Mittelwert
MAR	Missing at Random
MCAR	Missing Completely at Random
MD	Anteil fehlender Werte (Missing Data)
MESTA	Meta-Analyse der Effekte stationärer psychosomatischer Rehabilitation (Steffanowski et al., 2007)
N	Stichprobengröße
NTR	Nichtexperimentelles Treatment (naturalistische Studie)
OR	Odds Ratio
p	Inferenzstatistische Wahrscheinlichkeit, dass ein beobachtetes Ergebnis auch durch Zufallseffekte entstanden sein könnte
PR	Prädiktor (unabhängige Variable)
PRÄ	Messung vor der Behandlung
POST	Messung nach der Behandlung
POWC	Randomisierte Kontrollgruppenstudie ohne PRÄ-Test (Post-Only-With-Control)
PPWC	Randomisierte Kontrollgruppenstudie mit PRÄ-Test (Pre-Post-With-Control)
qVM	Quasi-indirekte Veränderungsmessung (RETRO-POST-Differenz)
r	Zusammenhang zwischen zwei Variablen (Produkt-Moment-Korrelation)
r^2	Gemeinsamer Varianzanteil zweier Variablen
r_{it}	Korrigierte Item-Trennschärfe
r_{tt}	Reliabilität (z.B. Cronbachs Alpha)
RC	Reliable Change Index
RETRO	Retrospektive PRÄ-Messung zum POST-Zeitpunkt aus der Erinnerung des Befragten
ROC	Receiver-Operator-Charakteristik zur Bestimmung von Sensitivität und Spezifität
ROI	Return on Investment
RZM	Regression zur Mitte
SD	Standardabweichung (Streuung)
SE	Standardfehler (*SD* dividiert durch Quadratwurzel aus N)
SGB	Sozialgesetzbuch
SGPP	Eingruppen-PRÄ-POST-Studie (Single Group Pre Post)
VDR	Verband Deutscher Rentenversicherungsträger
WHO	World Health Organization
ZOE	Zielorientierte Ergebnismessung

1 Einführung

Psychische Erkrankungen sind in Deutschland trotz rückläufiger Gesamtentwicklung der Krankschreibungstage seit mehreren Jahren auf dem Vormarsch (Bundesverband der Betriebskrankenkassen, 2005) und häufig durch langwierige, progrediente Verläufe gekennzeichnet, was einen steigenden Bedarf an qualifizierten therapeutischen Angeboten impliziert. Die stationäre psychosomatische Rehabilitation hat dabei einen hohen Stellenwert. Mit etwa 100.000 behandelten Patienten pro Jahr handelt es sich um den zweitwichtigsten stationären Versorgungssektor für psychische Erkrankungen in der Bundesrepublik (Statistisches Bundesamt, 2006).

Auf der anderen Seite ist das Gesundheitssystem seit Jahren mit steigenden Ausgaben bei einer gleichzeitigen Verknappung der finanziellen Ressourcen konfrontiert. Damit stellt sich auch für die medizinische Rehabilitation die Frage nach der Effektivität, Effizienz und Angemessenheit der durchgeführten Maßnahmen.

Auf die Herausforderung des steigenden Versorgungsbedarfes bei leerer werdenden öffentlichen Kassen bei einem gleichzeitig wachsenden Druck, mit Krankenhäusern Gewinn erwirtschaften zu müssen, haben Leistungserbringer und –träger in den 1990er Jahren mit einem ganzen Bündel von Maßnahmen zur Qualitätssicherung der angebotenen Behandlungen sowie Belebung der wissenschaftlichen Rehabilitationsforschung reagiert.

So sind im Bereich der psychosomatischen Rehabilitation in den letzten 20 Jahren eine Reihe von Programmevaluationsstudien durchgeführt worden, die eindrucksvolle Belege für die kurz- und mittelfristige Wirksamkeit der durchgeführten Behandlungen erbracht haben. So wurde erst kürzlich eine Meta-Analyse vorgelegt, welche die Ergebnisse von insgesamt 65 Ergebnisstudien zusammenfasst (Steffanowski, Löschmann, Schmidt, Wittmann und Nübling, 2007). Bei dieser Übersichtsarbeit fiel auch die inhaltliche Vielfalt der in den einzelnen Untersuchungen jeweils verwendeten Maße zur Abbildung des Behandlungserfolges auf. Dies ist durch die Komplexität dieses Versorgungsbereiches mit seinen vielfältigen Indikationen bedingt, der zudem ein entsprechend breites Spektrum langjährig gewachsener Behandlungskonzepte mit unterschiedlichen Behandlungsangeboten und –Schwerpunkten beinhaltet.

Wie lässt sich die Ergebnisqualität stationärer psychosomatischer Rehabilitation auf diesem Hintergrund möglichst umfassend und fair abbilden? Welche methodischen Zugänge und Arten von Veränderungsinformationen eignen sich zur Abbildung der Ergebnisqualität? Wie lassen sich bei Betrachtung unterschiedlicher Patienten, Störungsbilder oder Kliniken vergleichbare Aussagen erzielen, ohne dass die individuell relevanten Aspekte dabei verloren gehen? Wie lassen sich die Interessen der verschiedenen Auftraggeber (z.B. Patient, Klinik, Kostenträger, Öffentlichkeit) einer Ergebnisevaluation angemessen berücksichtigen und die vielfältigen Informationsebenen zu einem aussagekräftigen Gesamtbild zusammensetzen?

Zu diesen Themen soll in der vorliegenden Arbeit Stellung bezogen werden. Dabei richtet sich der Fokus zunächst auf eine innovative Vorgehensweise bei der Aggregation von

Einzelinformationen zu Multiplen Ergebniskriterien. Der Ansatz wurde von Schmidt, Bernhard, Wittmann und Lamprecht (1987) erstmals erprobt und in der Folgezeit in mehreren psychosomatischen Evaluationsstudien erfolgreich angewendet. Allerdings bestehen ungelöste methodische Fragen, deren Beantwortung Voraussetzung für eine systematische Weiterentwicklung des Ansatzes der Multiplen Ergebniskriterien ist.

1.1 Stationäre psychosomatische Rehabilitation

Seit 1990 hat der Anteil psychischer Erkrankungen in Deutschland deutlich zugenommen (Bundesverband der Betriebskrankenkassen [BKK], 2006). Besorgniserregend ist dabei die hohe Krankschreibungsdauer pro Fall und die damit einhergehende Gefährdung der Erwerbsfähigkeit der Betroffenen. Chronische Erkrankungen (dazu zählen auch viele psychische Störungen) sind durch langfristige und oft progrediente Verläufe gekennzeichnet, die zu wachsenden Einschränkungen der Funktionsfähigkeit und Lebensqualität im privaten und beruflichen Alltag führen. Mit der medizinischen Rehabilitation verfügt Deutschland über ein vorbildliches Versorgungsangebot für chronische Erkrankungen. Je nach Indikationsbereich werden dabei unterschiedliche Behandlungsansätze verfolgt. So stehen bei der psychosomatischen Rehabilitation die psychotherapeutischen Verfahren im Mittelpunkt.

Nachfolgend wird zunächst die wachsende Bedeutung der psychischen Störungen in Deutschland aufgezeigt, was einen entsprechenden rehabilitativen Bedarf impliziert. Sodann wird die medizinische Rehabilitation skizziert und eine Begriffsdefinition der psychosomatischen Rehabilitation vorgenommen.

1.1.1 Zunahme psychischer Erkrankungen in Deutschland

Einer aktuellen Pressemitteilung des Bundesverbandes der Betriebskrankenkassen (BKK) vom 18.04.2006 ist folgendes zu entnehmen:

> Seit Beginn der Neunzigerjahre hat sich der Krankenstand über alle Krankheitsarten nahezu halbiert. Allein bei den psychischen Erkrankungen gibt es einen unverminderten Zunahmetrend: Ihr Anteil an den Krankheitstagen hat sich seit 1990 mehr als verdoppelt.

Die durchschnittliche Zahl der Krankschreibungstage (AU-Tage) pro Arbeitnehmer ist im Jahr 2005 mit 12,7 Kalendertagen dabei auf den niedrigsten Wert seit Beginn der Statistik 1976 gefallen. Bei der ersten gesamtdeutschen Erhebung 1991 waren die Versicherten noch durchschnittlich 24,7 Tage krankgeschrieben.

Der BKK-Gesundheitsreport (BKK, 2005) berichtet dabei folgende Zahlen: Standen psychische Störungen als Ursache für Arbeitsunfähigkeit (AU) im Jahr 1991 mit 93 Tagen pro 100 Pflichtversicherten (entspricht 3,8% von 2470 AU-Tagen) dabei noch an siebter Stelle, so rangierten diese im Jahr 2004 mit 119 AU-Tagen (9,2% von 1300 AU-Tagen) be-

reits an vierter Stelle (Abbildung 1). Bei Frauen nahmen sie mit 11,9% den dritten Rang ein und bei Arbeitslosen mit 18,2% sogar den zweiten Rang. Als häufigste Einzeldiagnose nach der Internationalen Statistischen Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (ICD-10, Deutsches Institut für Medizinische Dokumentation und Information [DIMDI]) wurden unter den psychischen Erkrankungen depressive Störungen (F32) registriert.

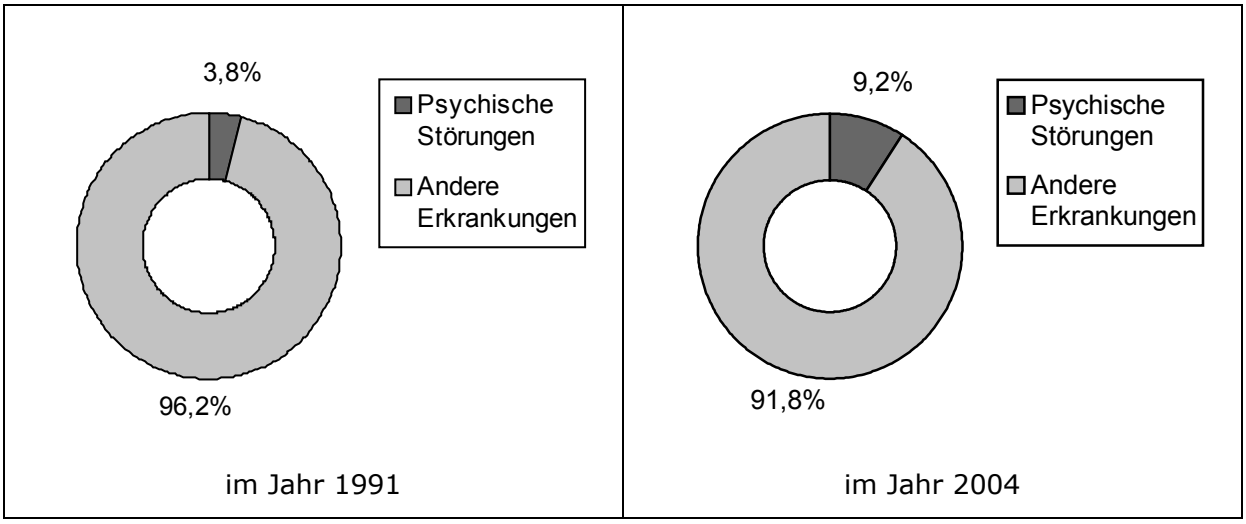


Abbildung 1. Anteil psychischer Störungen an den Krankschreibungstagen (BKK, 2005) bei den Pflichtversicherten inklusive Arbeitslose.

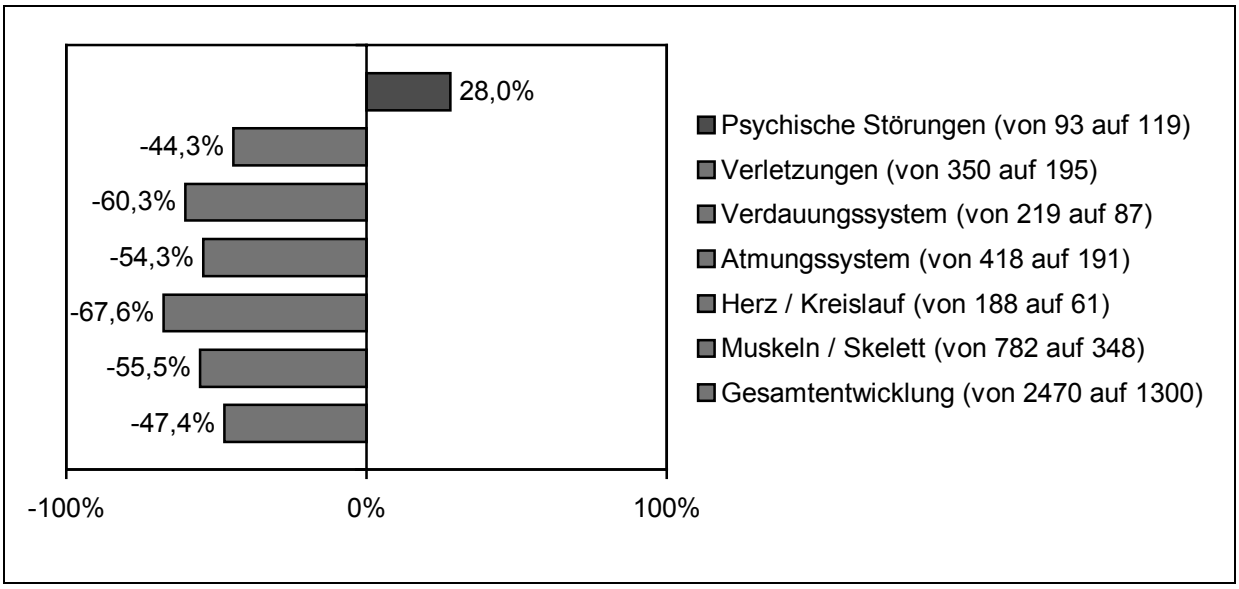


Abbildung 2. Entwicklung der Krankschreibungstage zwischen 1991 und 2004 (Kalendertage pro Jahr je 100 Pflichtversicherte, inklusive Arbeitslose). BKK (2005).

Auffallend ist, dass die Zahl der AU-Tage aufgrund psychischer Störungen nicht nur relativ, sondern auch absolut betrachtet entgegen dem Gesamttrend zugenommen hat. Während bei allen anderen Hauptindikationen eine erhebliche Reduktion der AU-Tage zu ver-

zeichnen ist, haben durch psychische Störungen bedingte AU-Tage je 100 Versicherte von 93 im Jahr 1991 auf 119 im Jahr 2004 zugenommen – ein Plus von 28,0% (vgl. Abbildung 2).

Eine Erklärung für diese Dynamik lässt sich in der Entwicklung des Arbeitsmarktes und der gesamtwirtschaftlichen Situation in Deutschland ausmachen. So sind einerseits die körperlichen Erkrankungen etwa im Produktivgewerbe durch Frühverrentung und verbesserten Arbeitsschutz rückläufig, darüber hinaus lassen sich viele Menschen aus Angst um ihren Job nur noch in gravierenden Fällen krankschreiben. Andererseits haben die psychomentalen Belastungen in den letzten Jahren durch Arbeitsverdichtung, Flexibilisierung, wachsende Unsicherheit der Arbeitsverhältnisse und steigende Arbeitslosigkeit ständig zugenommen (Siegrist, 2005). So ergab eine Studie von Larisch, Joksimovic, Knesebeck, Starke und Siegrist (2003), dass Beschäftigte, die durch ein Ungleichgewicht von Verausgabung und Belohnung bei der Arbeit belastet sind, gegenüber beruflich nicht Belasteten ein sechsfach erhöhtes Risiko aufweisen, von depressiven Symptomen betroffen zu sein.

Die Zunahme der psychischen Erkrankungen ist neben dem Verlust an Lebensqualität für die Betroffenen auch gesundheitspolitisch und volkswirtschaftlich im Hinblick auf die Krankheitskosten besorgniserregend. Im BKK-Gesundheitsreport 2005 wird weiter ausgeführt, dass bei den psychischen Störungen die durchschnittliche Dauer pro AU-Fall mit 28,0 Tagen nach Krebserkrankungen an zweiter Stelle steht, was erhebliche Belastungen der Kassen durch Krankengeldzahlungen mit sich bringt. Darüber hinaus kann bei derart langen Krankschreibungszeiten von einer erheblichen Gefährdung des Arbeitsplatzes der Betroffenen ausgegangen werden.

Im stationären Versorgungsbereich verursachten die psychischen Erkrankungen im Jahr 2004 entgegen des allgemeinen Trends zur Verkürzung der Liegezeiten

> (...) erstmals mit 241 Tagen je 1.000 Versicherte den höchsten Anteil der Krankenhaustage (15,7%) unter allen Krankheitsarten (...). Über alle Altersgruppen haben psychische Erkrankungen weiter zugenommen. Allein von 2003 auf 2004 ist die relative Fallhäufigkeit um gut 10% auf nunmehr 9,8 Fälle je 1.000 Versicherte gestiegen (...). Ein weiteres Augenmerk ist auf die anhaltende Zunahme der psychischen und Verhaltensstörungen bei Kindern und Jugendlichen zu lenken (...). Mit 273 Tagen je 1.000 Versicherte verursachen sie bereits über ein Drittel aller Krankenhaustage dieser Altersgruppe. (BKK 2005, S. 17-18)

Für das Jahr 2004 weist das statistische Bundesamt (Statistisches Bundesamt, 2006a) Gesamtausgaben des Gesundheitssystems in Höhe von 234,0 Milliarden Euro aus. 22,8 Milliarden Euro davon (9,7%) entfielen auf die Kategorie „Psychische Störungen.

Zu einer ähnlichen Quote hinsichtlich der Schätzung des Gesamtausfalls an Bruttowertschöpfung für das Jahr 2004 gelangt die Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (2005). So wird in dem Bericht eine Gesamtzahl von 440,1 Millionen AU-Tagen auf insgesamt 70,0 Milliarden Euro beziffert. 46,3 Millionen AU-Tage bzw. 7,4 Milliarden Euro entfallen dabei auf psychische Erkrankungen, was einem Anteil von 10,6% entspricht.

Als Fazit lässt sich festhalten, dass psychische Erkrankungen mittlerweile etwa 10% aller direkten und indirekten Krankheitskosten verursachen – Tendenz steigend. Neben langen

AU-Zeiten und hohen Krankheitskosten muss dabei auch von einer erheblichen Gefährdung des Arbeitsplatzes und der Erwerbsfähigkeit der Betroffenen ausgegangen werden. So war im Jahr 2003 bei den Männern jede vierte (24%) und bei den Frauen sogar jede dritte (35%) vorzeitige Berentung durch psychische Störungen bedingt (BKK, 2005).

Die Zunahme psychischer Störungen in den letzten Jahren schlägt sich auch in der Statistik der gesetzlichen Rentenversicherung zu den abgeschlossenen Maßnahmen zur stationären medizinischen Rehabilitation nieder (Deutsche Rentenversicherung Bund, 2005). Während der relative Anteil von Erkrankungen des Bewegungsapparates und von Herz-Kreislauf-Erkrankungen seit Jahren rückläufig ist, hat der Anteil psychischer Störungen von 12,7% im Jahr 1991 (99.000 von 779.000 Maßnahmen) auf 17,8% im Jahr 2004 (125.000 von 702.000 Maßnahmen) zugenommen. Abbildung 3 gibt die Entwicklung im Detail wieder. Der medizinischen Rehabilitation kommt als wichtiges Instrument zum Erhalt der Teilhabe in der Gesellschaft bei der Behandlung dieses Indikationsbereiches damit eine wichtige und weiter wachsende Bedeutung zu.

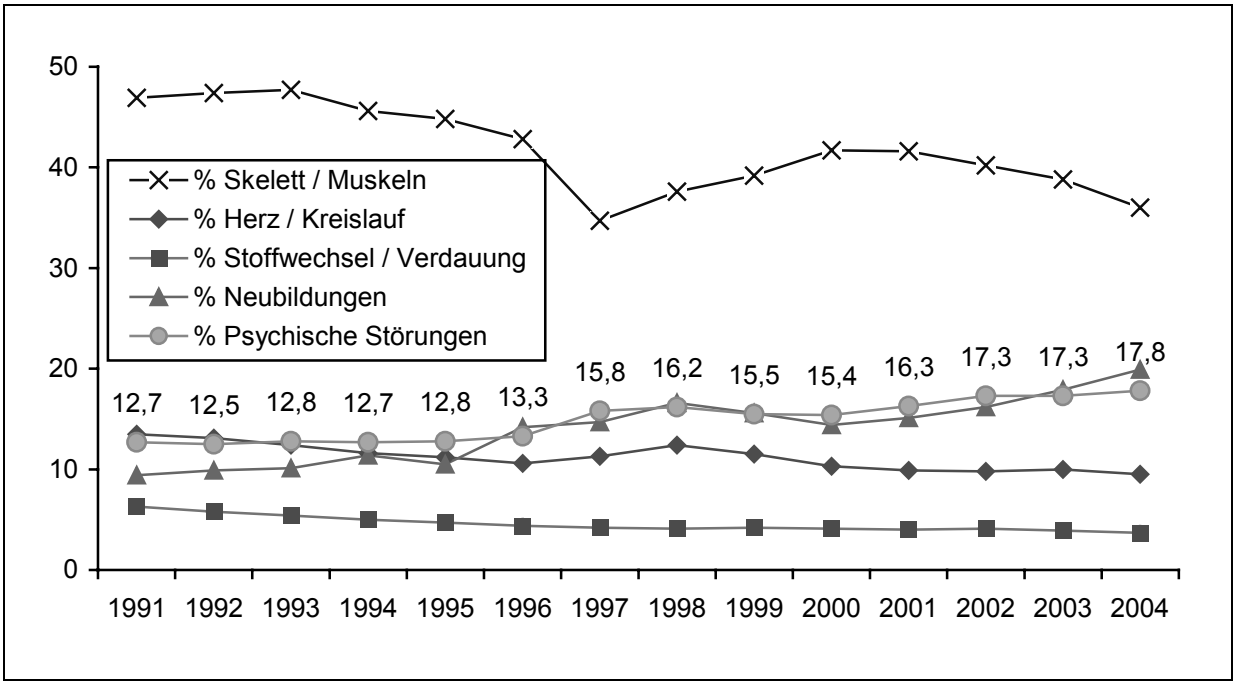


Abbildung 3. Prozentualer Anteil von fünf Indikationsbereichen an der Gesamtheit der durchgeführten Maßnahmen zur medizinischen Rehabilitation in den Jahren 1991 bis 2004 (Deutsche Rentenversicherung Bund, 2005). Der prozentuale Anteil psychischer Störungen ist für jedes Jahr auch in Zahlen angegeben.

1.1.2 Was versteht man unter Rehabilitation?

Eine ausführliche Abhandlung des vielseitigen Gebietes der Rehabilitationsmedizin findet sich bei Delbrück und Haupt (1998), einen umfassenden Überblick zu wichtigen Aspekten der Rehabilitationswissenschaften geben Bengel und Koch (2000). Nachfolgend sollen einige Aspekte kurz beschrieben werden, die zum Verständnis der vorliegenden Arbeit von Bedeutung sind.

Der Rehabilitationsbegriff leitet sich vom Krankheitsfolgen- und Behinderungsmodell der Weltgesundheitsorganisation (World Health Organization, 1993) ab, wie es in der International Classification of Impairments Disabilities and Handicaps (ICIDH-2) definiert ist. Dort wird zwischen körperlichen Strukturen, Aktivitäten und gesellschaftlicher Partizipation differenziert. So kann eine gesundheitliche Schädigung dazu führen, dass eine Person nicht mehr in der Lage ist, eine bestimmte Aktivität auszuführen. Dies wiederum kann zur Folge haben, dass sie bei der Teilhabe in einem bestimmten Lebensbereich behindert wird. Rehabilitation setzt auf allen drei Ebenen an und lässt sich nach Haupt und Delbrück (1998) wie folgt definieren:

> Rehabilitation umfasst die Gesamtheit aller Maßnahmen medizinischer, schulisch-pädagogischer, beruflicher und sozialer Art, die erforderlich sind, um für den Behinderten die bestmöglichen körperlichen, seelischen und sozialen Bedingungen zu schaffen. Diese sollen ihn befähigen, aus eigener Kraft einen möglichst normalen Platz in der Gesellschaft zu behalten und wiederzuerlangen. (S. 35)

Aus dieser Definition wird deutlich, dass bei der Rehabilitation versucht wird, die Gesamtheit des Menschen und seiner Lebensumwelt angemessen zu berücksichtigen, indem ein bio-psycho-soziales Modell als Grundlage von Therapie und Forschung verwendet wird. Dabei spielt die funktionale Ebene der Alltagsbewältigung eine wichtige Rolle, wie (Koch & Bengel, 2000) hervorheben:

> Das allgemeine Ziel der Rehabilitation besteht darin, dass die Betroffenen trotz der bleibenden Gesundheitsschäden und ihrer Folgen möglichst gut mit den Anforderungen des alltäglichen Lebens zurechtkommen und ihre Rollenverpflichtungen in Familie, Beruf und Gesellschaft möglichst weitgehend selbständig erfüllen können. (S. 42)

Der Anspruch auf medizinische Rehabilitation ist im Sozialgesetzbuch (SGB) IX (Bundesregierung, 2001) verbindlich festgeschrieben, als Leistungsträger sind vor allem die Gesetzliche Rentenversicherung (GRV) und Gesetzliche Krankenversicherung (GKV) zu nennen.

Ein wichtiges Prinzip der GRV ist der Grundsatz „Rehabilitation vor Rente", wonach rehabilitative Leistungen bei einer drohenden oder bereits eingetretenen Minderung der Erwerbsfähigkeit zu bewilligen sind. Erst wenn diese erfolglos bleiben bzw. ein Erfolg nicht erwartet werden kann, ist eine vorzeitige Berentung wegen Erwerbsunfähigkeit in Betracht zu ziehen. Ein wesentliches Hauptziel der Rehabilitation durch die GRV besteht

somit in Erhalt und Wiederherstellung der Erwerbsfähigkeit, was bei einer Evaluation entsprechend zu berücksichtigen ist.

Im Bereich der GKV gehört die Rehabilitation zur Krankenbehandlung. Ein Anspruch auf Krankenbehandlung durch die Krankenkasse besteht laut SGB grundsätzlich dann, wenn diese notwendig ist, um eine Krankheit zu erkennen, zu heilen, eine Verschlimmerung zu verhüten oder Beschwerden zu lindern. Dabei wird zwischen akutmedizinischen und rehabilitativen Leistungen differenziert: Medizinische Leistungen zur Rehabilitation können von der GKV nur bei einer drohenden Behinderung oder Pflegebedürftigkeit bzw. zur Stabilisierung des Behandlungserfolges nach akutmedizinischer Behandlung (z.B. Operation) durchgeführt werden.

Angesichts der demografischen Entwicklung in Deutschland und damit korrespondierenden Verlängerung der Lebensarbeitszeit ist davon auszugehen, dass chronische Krankheitsbilder unter den (im Durchschnitt zunehmend älteren Erwerbstätigen) weiter zunehmen werden, was einen entsprechenden Bedarf an qualifizierten rehabilitativen Versorgungsangeboten im Gesundheitssystem impliziert.

Medizinische Rehabilitation findet bislang fast ausschließlich im stationären Setting statt, d.h. die Patienten übernachten in der Einrichtung und nehmen dort auch ihre Mahlzeiten ein. Erst seit wenigen Jahren richtet sich angesichts der aktuellen gesundheitspolitischen Entwicklung das Augenmerk verstärkt auf teilstationäre und ambulante Angebotsformen (Grigoleit, 1998; Rüddel, Jürgensen, Terporten & Mans, 2002). Ein Vorteil nichtstationärer Angebote liegt in der Wohnortnähe, da auf diese Weise das soziale Umfeld und der Alltag vermehrt in die Rehabilitation einbezogen werden können. Auf der anderen Seite bietet eine wohnortfern gelegene stationäre Einrichtung gerade den Vorteil einer Herausnahme des Patienten aus seinem üblichen Milieu mit den damit verbundenen Rollenverpflichtungen, was sich förderlich auf den Genesungsprozess auswirken kann. Darüber hinaus können seltene Krankheitsbilder häufig nur überregional in entsprechend spezialisierten Fachkliniken behandelt werden.

1.1.3 Merkmale von Rehabilitationseinrichtungen

Rehabilitationseinrichtungen im Sinne von § 107 des SGB IX sind Einrichtungen, die

> „1. der stationären Behandlung der Patienten dienen, um (...) eine Krankheit zu heilen, ihre Verschlimmerung zu verhüten oder Krankheitsbeschwerden zu lindern oder im Anschluss an Krankenhausbehandlung den dabei erzielten Behandlungserfolg zu sichern oder zu festigen, auch mit dem Ziel, eine drohende Behinderung oder Pflegebedürftigkeit abzuwenden, zu beseitigen, zu mindern, auszugleichen, ihre Verschlimmerung zu verhüten oder ihre Folgen zu mildern (...).
>
> 2. fachlich-medizinisch unter ständiger ärztlicher Verantwortung und unter Mitwirkung von besonders geschultem Personal darauf eingerichtet sind, den Gesundheitszustand der Patienten nach einem ärztlichen Behandlungsplan vorwiegend durch Anwendung von Heilmitteln einschließlich Krankengymnastik, Bewegungstherapie, Sprachtherapie oder Arbeits- und Beschäftigungstherapie, ferner durch

andere geeignete Hilfen, auch durch geistige und seelische Einwirkungen, zu verbessern und den Patienten bei der Entwicklung eigener Abwehr- und Heilungskräfte zu helfen, und in denen

3. die Patienten untergebracht und verpflegt werden können."

Rehabilitationseinrichtungen weisen im Hinblick auf die im SGB IX vorgenommene Differenzierung zwischen Akut- und Rehabilitationsmedizin einige Besonderheiten auf (Buschmann-Steinhage, 1998; Bürger & Buschmann-Steinhage, 2000):

- Da die meisten Patienten in der Rehabilitation nicht bettlägerig sind, findet ein Großteil der Therapie in den jeweiligen Funktionsräumen und nicht im Krankenzimmer statt. Die Patienten sind meist in wohnlichen Einzelzimmern untergebracht und nach Möglichkeit wird der Eindruck einer „Krankenhausatmosphäre" in der Einrichtung vermieden. Abgesehen von einer ausreichenden Notfallausstattung kann auf die ansonsten in Akuteinrichtungen vorhandene Intensivmedizin und –pflege weitgehend verzichtet werden, was zu deutlich niedrigeren Tagespflegesätzen in der Rehabilitation führt (Statistisches Bundesamt, 2006b): Von 234,0 Milliarden Euro an Gesamtausgaben für das Gesundheitswesen im Jahr 2004 entfielen 60,4 Milliarden Euro auf 2.166 Krankenhäuser. Umgelegt auf insgesamt 146,7 Millionen Pflegetage im gleichen Jahr entspricht dies einem Tagessatz von 412 Euro. Auf die 1.294 Vorsorge- und Rehabilitationseinrichtungen entfielen 7,3 Milliarden Euro. Umgelegt auf 47,4 Millionen Pflegetage entspricht dies einem Tagessatz von 154 Euro.
- Im Gegensatz zum somatisch orientierten Krankheitsverständnis der Akutmedizin werden ganzheitliche Modelle favorisiert, die körperliche, psychische und soziale Faktoren gleichermaßen berücksichtigen. Dementsprechend sind die Behandlungsteams interdisziplinär zusammengesetzt. Die Therapiekonzepte sind in der Regel integrativ ausgerichtet und beinhalten unterschiedliche Angebote wie medizinische Therapie und Pflege, Physiotherapie, Schulung, Funktionstraining, Krankengymnastik, Psychotherapie, Ernährungsberatung, Sporttherapie, Entspannungstraining, balneologische Therapie und Sozialberatung. Je nach Indikation werden dabei unterschiedliche Schwerpunkte gesetzt. Bei der psychosomatischen Rehabilitation spielen psychotherapeutische Verfahren eine zentrale Rolle.
- Die klassische Patientenrolle ist in der Rehabilitation einem Wandlungsprozess unterworfen. So wird vom Rehabilitanden zunehmend erwartet, dass er nicht mehr passiver Leistungsempfänger ist, sondern zum kritischen Experten im Umgang mit seiner Erkrankung wird. So umfasst die Rehabilitation neben umfassender Information zu den Grundlagen der Erkrankung auch das Aufgeben von Krankenrolle und Vermeidungsverhalten, ein Wiedererlangen von Vertrauen in die eigenen Fähigkeiten, gezieltes Training von Fertigkeiten, Rückfallprävention sowie die Förderung von sozialer Kompetenz im Umgang mit schwierigen Lebenssituationen.

1.1.4 Psychosomatische Medizin

Zunächst ist festzustellen, dass es „die" psychosomatische Medizin nicht gibt und eine diesbezügliche Uniformitätsannahme daher zurückzuweisen ist (Schmidt, 1991). Gemäß der Komplexität des Fachgebietes existieren eine ganze Reihe von unterschiedlichen Erklärungsansätzen und darauf aufbauenden Behandlungsmethoden für psychosomatische Erkrankungen. Dabei werden integrative bzw. pluralistische Ansätze favorisiert, wobei u.a. psychoanalytische, lerntheoretische, kognitive, humanistische, psychophysiologische sowie neuroimmunologische Aspekte Berücksichtigung finden. Eine umfassende Darstellung zur psychosomatischen Medizin geben z.B. Uexküll (1996), Deter et al. 1996 oder Ahrens & Schneider (2002).

So wurde vom Deutschen Ärztetag 1992 mit dem „Facharzt für Psychotherapeutische Medizin" eine neue Berufsbezeichnung eingeführt und die Fortbildung in Psychosomatik verbindlich in der medizinischen Ausbildung festgeschrieben (Janssen & Hoffmann, 1994). Dort wird die Psychotherapeutische Medizin wie folgt definiert:

> Die Psychotherapeutische Medizin umfasst die Erkennung, psychotherapeutische Behandlung, die Prävention und Rehabilitation von Krankheiten und Leidenszuständen, an deren Verursachung, deren subjektiver Verarbeitung psychosoziale Faktoren und/oder körperlich-seelische Wechselwirkungen maßgeblich beteiligt sind.

Die Begriffe der „Psychotherapeutischen Medizin" und „Psychosomatischen Medizin" können synonym verwendet werden. Aus der Definition geht hervor, dass körperliche, seelische und soziale Aspekte bei der Therapie berücksichtigt werden. Je nach konkretem Störungsbild und individuellem Behandlungsanliegen können diese drei Bereiche allerdings unterschiedlich stark in den Mittelpunkt rücken. Diese Besonderheit drückt sich in einer entsprechenden Breite der Indikationen und Vielfalt des therapeutischen Angebots von psychosomatischen Fachkliniken aus.

1.1.5 Stationäre psychosomatische Rehabilitation

Um Verwechslungen und Missverständnissen vorzubeugen, soll an dieser Stelle eine Definition und Abgrenzung des Begriffs der stationären psychosomatischen Rehabilitation vorgenommen werden.

- Es findet eine stationäre Behandlung in einer Rehabilitationseinrichtung im Sinne von SGB IX (Bundesregierung, 2001) mit Indikation für psychosomatische Erkrankungen statt. Häufig drückt sich dies durch eine entsprechende Zusatzbezeichnung der behandelnden Einrichtung wie „Fachklinik für psychosomatische Medizin" oder „Abteilung für psychosomatische Rehabilitation" aus, kann sich aber auch hinter Begriffen wie „Stationäre Psychotherapie", „Behandlung von psychovegetativen / psychogenen Erkrankungen" oder einfach auch „Klinik für medizinische Re-

habilitation" (bei mehreren Fachbereichen im Haus) verbergen. Im Zweifelsfall gibt das Behandlungskonzept der betreffenden Einrichtung nähere Auskunft. Eine umfassende, wenn auch nicht mehr aktuelle Auflistung von psychosomatischen Einrichtungen findet sich bei Neun, Dahlmann, Geyer und Potreck-Rose (1994). Ein ausführliches Klinikverzeichnis im Internet mit über 1.000 Rehabilitationseinrichtungen, darunter mehr als 100 Einrichtungen mit Indikation Psychosomatik stellt das Internet Access Center Düsseldorf (2008) unter der Adresse http://www.klinikverzeichnis-online.de zur Verfügung.

- Typische Indikationsbereiche der stationären psychosomatischen Rehabilitation sind Depressive Störungen, Angst- und Zwangsstörungen, Anpassungs-, dissoziative und posttraumatische Störungen, Erschöpfungssyndrome, leichtere psychotische Residualzustände, Essstörungen, Persönlichkeitsstörungen, begleitende Abhängigkeitserkrankungen, vegetative und neurotische Störungen, somatoforme bzw. funktionelle Störungen wie Spannungskopfschmerz, Migräne, Schlafstörungen, Herz-Kreislauf-Störungen, Atemstörungen, Magen-Darm-Beschwerden sowie urogenitale bzw. sexuelle Beschwerden. Typischerweise nicht aufgenommen werden bettlägerige, demente, sozial desintegrierte, manifest suizidale, psychotische oder drogenabhängige Patienten.
- Körperliche, psychische und soziale Aspekte werden in der Therapie berücksichtigt, wobei die Psychotherapie eine zentrale Rolle spielt. Das Behandlungsangebot baut häufig auf einem langjährig gewachsenen integrativen Konzept auf und umfasst je nach Ausrichtung der Klinik neben der allgemeinmedizinischen und pflegerischen Betreuung ein breites Spektrum an verbalen und nonverbalen psychotherapeutischen Methoden unterschiedlicher Schulen. Zum Teil werden indikations- und themenspezifische Gruppen (z.B. zu Angst, Essstörungen oder beruflichen Problemen) gebildet. Darüber hinaus beinhaltet die Rehabilitation, wie bereits in Kapitel 1.1.3 ausgeführt, physio-, sport-, balneo- und sozialtherapeutische Angebote sowie weitere Maßnahmen wie Informationsveranstaltungen, Schulung, Beratung, Entspannungstraining, Meditation und kreativ-gestalterische Angebote. Eine erschöpfende Aufzählung ist an dieser Stelle sicherlich nicht möglich, die genannten Verfahren sollten aber zumindest einen Eindruck von der Vielzahl an kombinierbaren Behandlungsmethoden liefern und den in der Definition geforderten, ganzheitlichen Ansatz in der Rehabilitation illustrieren. Ein Punkt ist noch hervorzuheben, der sich häufig in den Behandlungskonzepten psychosomatischer Kliniken findet und der die Rolle der Mitpatienten betont: Das Konzept der therapeutischen Gemeinschaft (Hilpert, 1979), die einen sozialen Raum mit eigenen Regeln bildet. In diesem wird eine therapeutische Atmosphäre von Offenheit und gegenseitigem Vertrauen geschaffen, in der neue Einsichten und korrigierende Lernerfahrungen möglich sind sowie neue Verhaltensweisen in einem geschützten Rahmen eingeübt werden können.

Nachfolgend wird zur Abgrenzung des Definitionsbereiches gegenüber verwandten Disziplinen auch dargelegt, was „Stationäre psychosomatische Rehabilitation" aus Sicht des Autors nicht ist, wenngleich häufig inhaltliche Überschneidungen bestehen:

- Akutpsychiatrische Behandlung zählt nicht zum Bereich der psychosomatischen Rehabilitation, auch wenn in der offenen Psychiatrie hinsichtlich der Indikationen und Behandlungsmodalitäten zum Teil Überschneidungen bestehen (z.B. stationäre Psychotherapie bei depressiven Störungen). Eine Ausnahme bilden Patienten, die nach einem akutpsychiatrischen Krankenhausaufenthalt eine psychosomatische Rehabilitation durchführen. Der Übergang aus der Akutbehandlung in die Rehabilitation geschieht mitunter fliessend.
- Suchtrehabilitation wie Alkohol- und Drogenentzug ist nicht mit psychosomatischer Rehabilitation gleichzusetzen. Bereits aus der deutlich höheren mittleren Behandlungsdauer geht hervor, dass die Rehabilitation von Abhängigkeits- und Suchterkrankungen unter anderen Voraussetzungen als die psychosomatische Rehabilitation stattfindet. Auch werden zum Teil andere Therapiekonzepte bei der Behandlung sowie andere Ergebniskriterien bei der Evaluation verwendet (vgl. Hollstein, 1998; Süss, 1997). Eine wichtige Ausnahme bilden bereits entwöhnte Patienten, die zur Rehabilitation in eine psychosomatische Klinik aufgenommen werden sowie begleitende Abhängigkeitserkrankungen in der psychosomatischen Rehabilitation, etwa wenn bei einer Angsterkrankung zusätzlich Alkoholmissbrauch als Nebendiagnose festgestellt wird. In der realen Versorgungspraxis stellt sich dabei allerdings die Frage, was Ursache und was Wirkung ist und ob sich nicht hinter mancher, primär als psychosomatischer Erkrankung diagnostizierten Störung nicht doch eigentlich eine Suchterkrankung verbirgt und umgekehrt. Hinsichtlich der zu bearbeitenden Lebensprobleme bestehen häufig Überschneidungen zwischen beiden Indikationsgruppen.
- Rehabilitation bei neurologischen Erkrankungen, etwa nach Schlaganfall, erworbenen Hirnschädigungen oder bei degenerativen Prozessen (z.B. Altersdemenz) gehört ebenfalls nicht zur psychosomatischen Rehabilitation. Eine Ausnahme bilden nichtorganisch bedingte Schmerzsyndrome wie z.B. Migräne im Rahmen psychosomatischer Rehabilitation.
- Rehabilitation bei Kindern und Jugendlichen bildet einen eigenständigen Bereich. Zielgruppe der psychosomatischen Rehabilitation sind erwachsene Patienten im Erwerbsalter, wenngleich bei gegebenen Voraussetzungen vereinzelt auch Jugendliche in eine psychosomatische Klinik aufgenommen werden können.

Ein Grenzbereich stellt die psychosomatische Behandlung in einer Fachabteilung im Akutkrankenhaus dar. Diese ist auf dem Hintergrund der Unterschiede bei den versorgungsstrukturellen Voraussetzungen zwar keine psychosomatische Rehabilitation im Sinne der Definition, andererseits bestehen in der Praxis hinsichtlich Indikationsspektrum und Behandlungsmodalitäten große Ähnlichkeiten zwischen akuter und rehabilitativer psychosomatischer Therapie. Aus diesem Grund dürften die Ausführungen im Rahmen der vorliegenden Arbeit auch für diesen Versorgungsbereich Gültigkeit besitzen.

1.1.6 Aktuelle Zahlen zur psychosomatischen Versorgung

Auf dem Hintergrund der Unterscheidung zwischen akuter und rehabilitativer Versorgung wird der Anteil der psychosomatischen Medizin in beiden stationären Versorgungsbereichen vergleichend dargestellt.

Tabelle 1. Stationäre Akuteinrichtungen im Jahr 2004.

Abteilung	Betten	%	Fälle	%	Dauer (Tage)	Pflegetage (Millionen)	%
Psychotherapeutische Medizin / Psychosomatik	4.412	0,8	35.310	0,2	41,4	1,5	1,0
Psychiatrie und Psychotherapie	53.021	10,0	712.533	4,2	24,7	17,6	12,0
Kinder-/Jugendlichen-Psychiatrie / Psychoth.	4.835	0,9	36.770	0,2	43,6	1,6	1,1
Alle anderen Akut-Abteilungen	469.065	88,3	16.017.036	95,3	7,9	126,0	85,9
Alle Abteilungen in allen Akut-Krankenhäusern	531.333	100,0	16.801.649	100,0	8,7	146,7	100,0

Tabelle 2. Stationäre Rehabilitations- und Vorsorgeeinrichtungen im Jahr 2004.

Abteilung	Betten	%	Fälle	%	Dauer (Tage)	Pflegetage (Millionen)	%
Psychotherapeutische Medizin / Psychosomatik	13.371	7,6	93.658	5,0	40,0	3,7	7,9
Psychiatrie und Psychotherapie	12.477	7,1	64.874	3,4	62,8	4,1	8,6
Alle anderen Reha-Indikationsbereiche	150.625	85,3	1.730.830	91,6	22,9	39,6	83,5
Alle Abteilungen in allen Reha-Einrichtungen	176.473	100,0	1.889.362	100,0	25,1	47,4	100,0

Für das Jahr 2004 liegen folgende Daten vor (Statistisches Bundesamt, 2005): Von insgesamt 531.333 Betten in 2.166 Akut-Krankenhäusern standen 4.412 Betten in psychosomatischen Fachabteilungen (0,8%). Insgesamt wurden 35.310 Patienten (0,2% von 16,8 Millionen) hier behandelt, was bei einer durchschnittlichen Behandlungsdauer von 41,4 Tagen einem Anteil von 1,0% aller dokumentierten Pflegetage (1,5 Millionen von

146,7 Millionen) entspricht. Damit spielt die Psychosomatische Medizin im Bereich der Akutversorgung eine untergeordnete Rolle (vgl. Tabelle 1).

Aus der Bundesstatistik geht weiter hervor (Statistisches Bundesamt, 2006b), dass im Jahr 2004 insgesamt 1.294 Vorsorge- und Rehabilitationseinrichtungen mit 176.473 Betten in Deutschland existierten. Für den Bereich der psychotherapeutischen Medizin waren 13.371 Betten vorhanden. Insgesamt wurden 93.658 Patienten behandelt, die durchschnittliche Aufenthaltsdauer lag bei 40,0 Tagen und die Gesamtzahl der dokumentierten Pflegetage bei 3,8 Millionen (vgl. Tabelle 2).

Abbildung 4 veranschaulicht die Unterschiede zwischen dem akuten und rehabilitativen Versorgungssektor. Während der Gesamtanteil aller psychischen Störungen bei Aufschlüsselung der Pflegetage nach Fachabteilungen in beiden Bereichen in etwa vergleichbar ist, hat die psychosomatische Medizin nur in der Rehabilitation eine bedeutenden Anteil. Zur Erklärung dieser Diskrepanz sollten die historisch gewachsenen Versorgungsstrukturen in Rechnung gestellt werden. Nicht jedes Kreiskrankenhaus verfügt über eine eigene psychosomatische Abteilung, so dass Patienten mit einer psychischen Problematik in der Akutmedizin vorwiegend in psychiatrische Behandlung überwiesen werden.

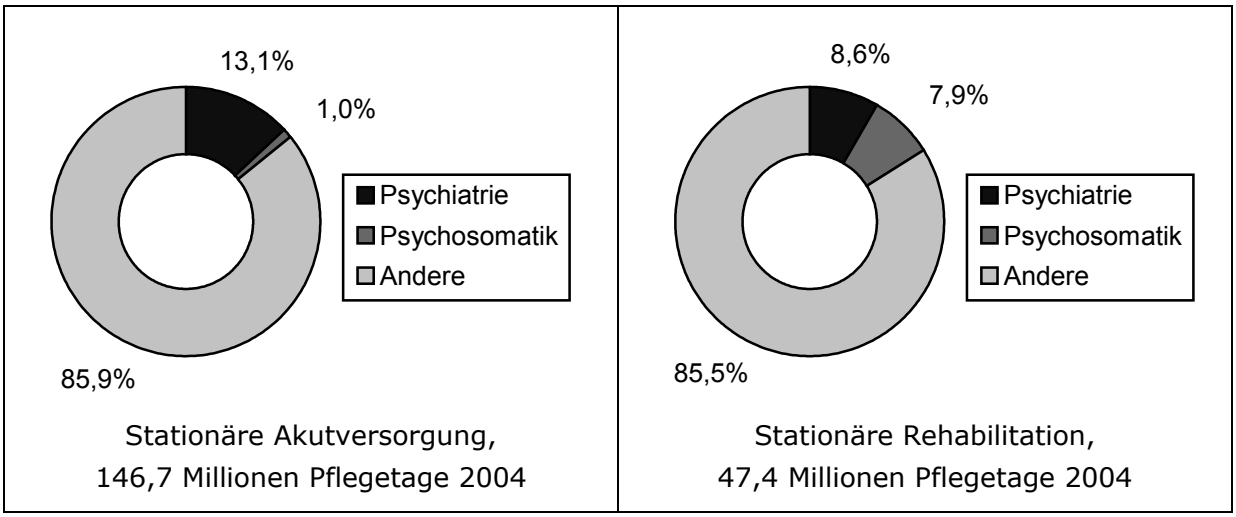


Abbildung 4. Anteil psychiatrischer und psychosomatischer Behandlung an der Gesamtzahl der Pflegetage im akuten und rehabilitativen Setting (Statistisches Bundesamt, 2005).

Setzt man auf dem Hintergrund der Ausführungen in Abschnitt 1.1.3 für die 3,7 Millionen Pflegetage in der Rehabilitation einen Tagessatz von 154 Euro an, entspricht dies einem Gesamtausgabevolumen von 569,8 Millionen Euro pro Jahr. Auch wenn dieser Betrag nur einen geringen Anteil am Gesamtvolumen aller Ausgaben im Gesundheitswesen ausmacht, stellt sich auf dem Hintergrund der gegenwärtigen Ressourcenverknappung dennoch die Frage, ob sich diese Investition für alle Beteiligten langfristig „lohnt" und ob die durchgeführten Behandlungen wirksam sind. Eine Antwort auf diese Frage kann die sozialwissenschaftliche Disziplin der Programmevaluation liefern.

1.2 Evaluation in der Rehabilitation

In den letzten Jahren ist ein Trend zur verstärkten Evidenzbasierung in der Medizin festzustellen. Antworten auf die Frage nach der bestmöglichen Evidenz kann die Evaluationsforschung liefern. Im Bereich der Sozialwissenschaften hat sich dabei der Begriff der „Programmevaluation" durchgesetzt, wenn es um die umfassende wissenschaftlich gestützte Bewertung etwa eines ganzen „Interventionspakets" geht, wie dies bei einer medizinischen Rehabilitationsmaßnahme der Fall ist. Je nach theoretischer Ausrichtung sind dabei unterschiedliche Forschungsstrategien möglich, welche eher die interne oder externe Validität betonen. Bei jeder Evaluation sind unterschiedliche Stakeholder zu berücksichtigen, was sich in einer entsprechenden Auswahl der Assessmentverfahren und Bewertungskriterien niederschlägt. Im Bereich der medizinischen Rehabilitation ist seit den 1990er Jahren eine deutliche Verstärkung der Forschungsaktivitäten zu beobachten, wobei das Qualitätssicherungsprogramm der Deutschen Rentenversicherung sowie der Förderschwerpunkt „Rehabilitationswissenschaften" besonders hervorzuheben sind.

1.2.1 Evidenzbasierung in der Medizin

Seit Ende der 1990er Jahre taucht der Begriff der Evidenzbasierung in der Medizin (EBM) zunehmend in den Medien, unter Klinikern und in der gesundheitspolitischen Diskussion auf (Luber & Geene, 2004) und ist damit auch für die medizinische Rehabilitation relevant. Mittlerweile existieren eine Reihe von Aus- und Weiterbildungsprogrammen zur Praxis und Lehre der EBM und in Großbritannien haben sich bereits Zentren für evidenzbasierte medizinische Praxis in mehreren Fachbereichen etabliert. Sackett und Rosenberg (1997) definieren EBM wie folgt:

> EBM ist der gewissenhafte, ausdrückliche und vernünftige Gebrauch der gegenwärtig besten externen, wissenschaftlichen Evidenz für Entscheidungen in der medizinischen Versorgung individueller Patienten. Die Praxis der EBM bedeutet die Integration individueller klinischer Expertise mit der bestmöglichen externen Evidenz aus systematischer Forschung. (S. 644)

Eng mit der EBM verbunden ist die Diskussion um ein leitlinienorientiertes Vorgehen. Hierzu existieren kontroverse Positionen (Petermann, 2005). Während Befürworter argumentieren, dass EBM und Leitlinien zu mehr Transparenz im Gesundheitswesen und einer Neubewertung bisher unreflektiert akzeptierter medizinischer Verfahren und zur Steigerung der Qualität der Behandlungen führt, gehen die Kritiker davon aus, dass EBM und Leitlinien als Instrumente zur profitmotivierten Durchsetzung von Sparplänen und Kontrolle missbraucht werden, zu Einschränkungen der ärztlichen Handlungsfreiheit führen und im übrigen der Individualität des Patienten mit seiner Krankengeschichte nicht gerecht werden („Kochbuchmedizin"). Insgesamt ist davon auszugehen, dass der EBM-Ansatz am ehesten fruchtbare Ergebnisse bringt, wenn die Interessen aller Zielgruppen

und Beteiligten einer Intervention angemessen berücksichtigt werden (Rosenbrock, 2004).

Donabedian unterschied bereits 1966 zwischen Struktur-, Prozess- und Ergebnisqualität. Während die Strukturqualität für die räumlichen, technischen und personellen Voraussetzungen einer Einrichtung zur Leistungserbringung steht, bezieht sich die Prozessqualität auf das, was während der Intervention geschieht. Die Ergebnisqualität hat schließlich mit der Effektivität (Wirksamkeit) und Effizienz (erzielter Effekt im Verhältnis zum Aufwand) des evaluierten Programms zu tun. Eine hohe Struktur- und Prozessqualität stellt eine notwendige, aber nicht hinreichende Voraussetzung für eine hohe Ergebnisqualität dar. Eine zu starke Betonung der Struktur- und Prozessqualität unter Vernachlässigung der Ergebnisseite birgt die Gefahr der Kostenexplosion. Oder, um Wittmann (2003) zu zitieren:

> Ergebnisqualität ist nicht Alles, aber ohne Ergebnisqualität ist alles Nichts!

Diese Aussage ist im Hinblick auf die zunehmende Evidenzorientierung von großer Bedeutung, da man davon ausgehen kann, dass künftig nur noch Maßnahmen und Medikamente finanziert werden, die einen wissenschaftlichen Wirksamkeitsnachweis erbracht haben. Ein wichtiger Aspekt bei der EBM ist somit die Frage, mit welchen Methoden sich am ehesten Evidenz für alle Beteiligten herstellen lässt (Tschuschke, 2005; Schmacke, 2006). Antworten auf solche Fragen kann die sozialwissenschaftliche Disziplin der Evaluationsforschung bzw. Programmevaluation liefern.

1.2.2 Evaluation, Evaluationsforschung und Programmevaluation

Am weitesten ist der Begriff der „Evaluation" gefasst und steht nach Wottawa & Thierau (1998) generell für die zunächst qualitative Bewertung eines Produktes oder Prozesses:

> Prozeß der Beurteilung des Wertes eines Produktes, Prozesses oder eines Programms, was nicht notwendigerweise systematische Verfahren oder datengestützte Beweise zur Untermauerung einer Beurteilung erfordert. (S. 13)

„Evaluationsforschung" impliziert die Verwendung von quantitativen wissenschaftlichen Methoden zur Durchführung der Bewertung und wird von Wittmann (1985) wie folgt definiert:

> Evaluationsforschung ist hier die explizite Verwendung wissenschaftlicher Forschungsmethoden und –techniken für den Zweck der Durchführung einer Bewertung. Evaluationsforschung bezieht sich dabei auf jene Verfahren, die die Möglichkeit des Beweises anstelle der reinen Behauptung bezüglich des Wertes und Nutzens einer bestimmten sozialen Aktivität erhöhen.

Für den Praxisbezug ist noch der Begriff der „Programmevaluation" wichtig, der nach Wittmann (1985) nicht so weit gefasst ist wie „Evaluation", aber breiter gefasst ist als „Evaluationsforschung". Rossi, Freeman und Lipsey (1999) definieren Programmevaluation wie folgt:

> Program evaluation is the use of social research procedures to systematically investigate the effectiveness of social intervention programs. (S. 4)

Der Begriff „Social Intervention Programs" bezieht sich dabei auf alle Aktivitäten zur Verbesserung der sozialen Lebensbedingungen bzw. Milderung sozialer Probleme. Wittmann (1985) hebt hervor, dass Programmevaluation immer anwendungsbezogen ist und dem Praktiker Grundlagen für rationale Bewertungen und Entscheidungen liefern soll. Neben streng wissenschaftlichen Versuchsplänen werden dabei auch „weichere" quasi-experimentelle Forschungsstrategien in Kontexten verwendet, wo ein randomisiertes Kontrollgruppenexperiment z.B. aus ethischen Gründen nicht möglich ist, die aber dennoch einen möglichst hohen Grad an Wissenschaftlichkeit und Objektivität haben. Wottawa & Thierau (1998) führen aus, dass eine wissenschaftliche Evaluation angesichts einer Vielfalt von gebräuchlichen Definitionen folgende Kennzeichen aufweist:

- Ein allgemeiner Konsens, der hier auch schon durch die Wortwurzeln ‚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." (S. 14).

Auch wenn eine ausführliche Darstellung aller Positionen aus Platzgründen hier nicht möglich ist, sollen nachfolgend einige Vertreter kurz genannt werden, die stellvertretend für eine häufig anzutreffende Dialektik in der Evaluationsforschung stehen:

Donald Campbell (Campbell, 1957; Campbell & Stanley, 1966; Campbell, 1969) steht für die experimentelle Position, bei der Wert auf eine Gewinnung von kausalen Aussagen und hohe interne Validität gelegt wird. Nur durch randomisierte Kontrollgruppenexperimente und eine standardisierte Vorgehensweise bei der Datenerhebung lässt sich ein beobachteter Effekt unter Ausschluss alternativer Erklärungen kausal auf eine bestimmte Intervention zurückführen. Ein gravierender Nachteil dieser Position besteht allerdings in der Vernachlässigung von Aspekten der externen Validität und damit Generalisierbarkeit der Ergebnisse, da eine strikte experimentelle Kontrolle oder gar künstliche Laborsituation nicht den Gegebenheiten im Versorgungsalltag entspricht (Fydrich & Schneider, 2007).

Lee J. Cronbach (Cronbach, 1980; Cronbach, 1982) betont hingegen die externe Validität und Anwendungsorientierung als unverzichtbare Voraussetzung für die Glaubwürdigkeit der Evaluation und wies eine einseitige Betonung experimenteller Designs zurück. Thomas Cook (Cook & Campbell, 1979; Shadish, Cook & Campbell, 2002) konzentrierte sich auf das quasi-experimentelle Design mit Vor-Test (PRÄ) und Nach-Test (POST) als Alternative zum echten randomisierten Experiment, wenn dieses z.B. aus ethischen Gründen nicht realisierbar ist. Cook hebt in seinen Arbeiten die Bedeutung von Kontextfaktoren für die Forschungsergebnisse hervor, plädiert für Methodenpluralität je nach Fragestellung der Evaluation und weist auf die Wichtigkeit hin, verschiedene Interessengruppen bereits

früh bei der Formulierung von Evaluationsfragestellungen einzubeziehen, damit die gewonnenen Informationen später auch von Nutzen sind.

Bezugnehmend auf Donabedian (1966) ergibt sich aus der Darstellung dieser Positionen die Notwendigkeit für die medizinische Evaluation im weitesten Sinne, integrativ sowohl Prozesse und Struktur als auch Ergebnisse zu betrachten. Nur alle diese Forschungsobjekte gemeinsam erlauben Ergebnisse der Evaluationsforschung, welche sowohl bezüglich der internen als auch der externen Validität zufrieden stellen können.

1.2.3 Integration unterschiedlicher Forschungsparadigmen

Letztendlich sind alle Forderungen nach interner und externer Validität im Feld kaum umsetzbar (vgl. zu dieser Debatte auch Leichsenring, 2004a, 2004b; Fydrich & Schneider, 2007; Shadish, Matt, Navarro & Siegle, 1997; Kächele 2006). Peter Rossi (Rossi et al., 1999) erweiterte seinen Blick über diese Dialektik hinaus und formulierte das Konzept der „Maßgeschneiderten Evaluation", wonach es bei Auswahl der geeigneten Evaluationsstrategie zur Beantwortung bestimmter Fragestellungen auf die richtige „Passung" zwischen Evaluation und Programm ankommt und durchaus auch mehrgleisige Ansätze verfolgt werden können, die sowohl qualitative als auch quantitative Komponenten beinhalten.

Auch Wittmann verfolgt hinsichtlich der zu verwendenden Forschungsstrategien einen integrativen Ansatz und hat ein umfassendes methodisches Rahmenkonzept zur Programmevaluation entwickelt (Wittmann, 1985, 1990, 1995; Wittmann, Nübling & Schmidt, 2002):

> Programmevaluation ist ein Prozess der Durchführung rational- und vernunftgeleiteter Beurteilungen eines Programms hinsichtlich Aufwand, Effektivität, Wirksamkeit, Angemessenheit auf der Grundlage systematischer Datenerhebung und Datenanalyse, konzipiert für die Verwendung beim Programm-Management, beim Rechenschaftsbericht für Auftraggeber oder Öffentlichkeit, Zukunftsplanung. (Wittmann, 1985, S. 23-24)

Je nach Interesse des Auftraggebers der Evaluation kann dabei die Programmplanung, die Prozessevaluation oder aber die Ergebnisevaluation im Vordergrund stehen (Posavac & Carey, 1980). Die drei wichtigsten methodischen Säulen, auf denen eine systematische Programmevaluation aufbaut, sind nach Wittmann dabei

- Techniken der Versuchsplanung
- Verfahrensweisen des Assessments (Mess- und Diagnoseinstrumente)
- Zielbestimmungs-, Bewertungs- und Entscheidungshilfen

Zur Visualisierung evaluativer Forschungsstrategien verwendet Wittmann ein Modell mit fünf Datenboxen (Abbildung 5). Jede Datenbox in dem Modell stellt ein Covariation-Chart im Sinne von Cattell (1957, 1966) dar und erlaubt eine Aggregation von Daten über die drei Dimensionen Personen, Variablen und Situationen (Messzeitpunkte):

> Die einzelnen Datenboxen enthalten die Operationalisierungen der jeweils theoretisch interessierenden Konstrukte, hinsichtlich Ursachen (PR-, ETR- und NTR-Box) und Effekten (CR-Box), die an ausgewählten Personengruppen über ausgewählte Zeitpunkte hinweg gemessen werden. Die Bewertungsbox (EVA) enthält jedoch nur solche Variablen, die einen evaluativen Charakter haben wie z.B. gut vs. schlecht, ökonomisch vs. unökonomisch, sozial vs. unsozial, demokratisch vs. undemokratisch, fortschrittlich vs. rückständig, gerecht vs. ungerecht, wissenschaftlich, vs. unwissenschaftlich usw. (Wittmann 1990, S. 243)

Bei einer Evaluation sind in der Regel unterschiedliche Interessenten (Stakeholder) zu berücksichtigen. Dies wird hier durch die Evaluations-Box (EVA-Box) symbolisiert. Die unterschiedlichen Interessen und Ziele der Stakeholder sind bei Auswahl und Operationalisierung der zu erhebenden Ergebniskriterien in der Kriteriums-Box (CR-Box) maßgebend.

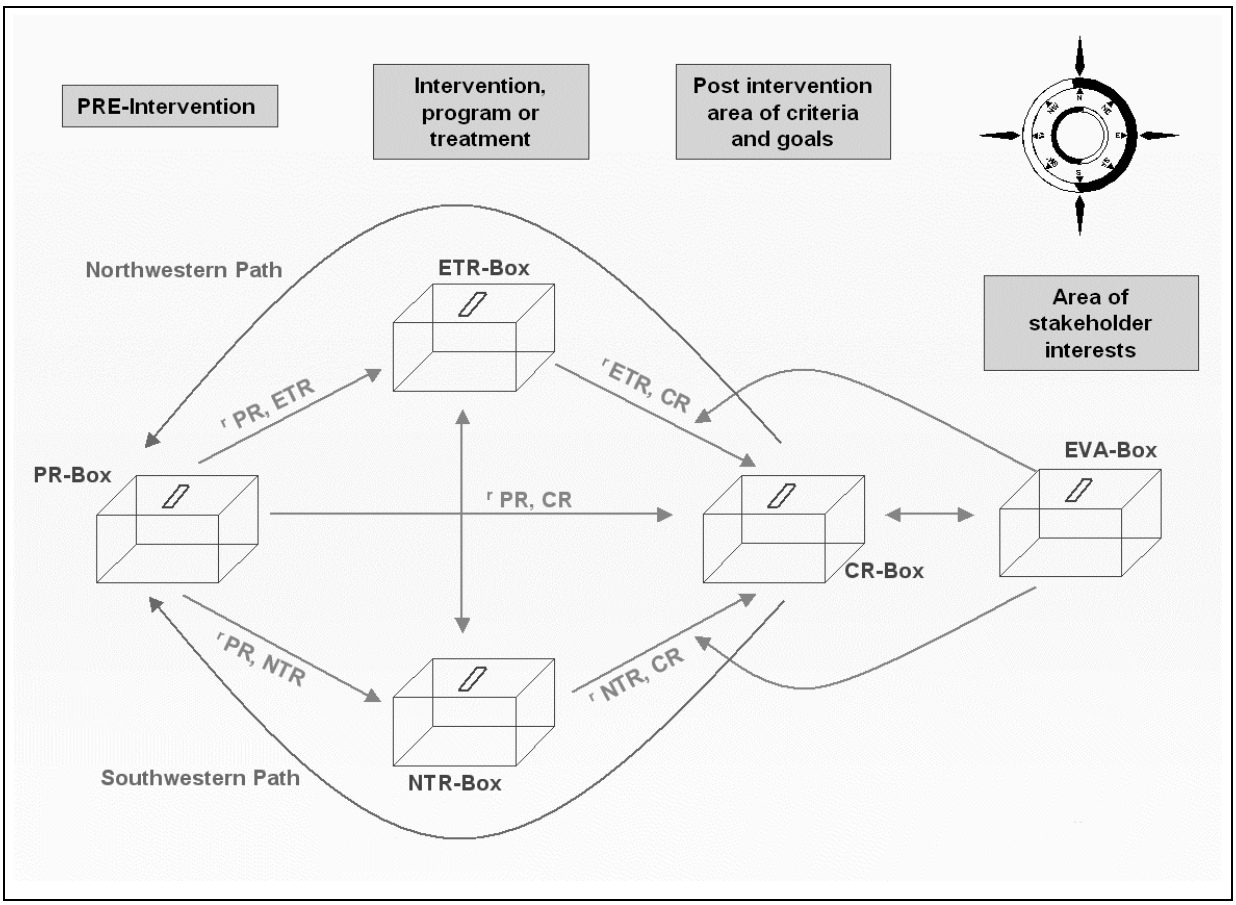


Abbildung 5. Das Modell der fünf Datenboxen nach Wittmann (1990). PR-Box = Prädiktor-Box, ETR-Box = Experimentelle Treatment-Box, NTR-Box = Nichtexperimentelle Treatment-Box, CR-Box = Kriterienbox, EVA-Box = Bewertungs-Box.

Die Prädiktor-Box (PR-Box) enthält eine Reihe von Vorhersagevariablen wie z.B. soziodemografische Angaben, Diagnosen, Krankengeschichte und Schweregrad. Diese können dazu verwendet werden, Unterschiede in der Kriteriumsausprägung vorherzusagen (symbolisiert durch den Pfeil zwischen PR- und CR-Box).

Die experimentelle Treatment-Box (ETR-Box) steht für randomisierte Kontrollgruppendesigns. Diese erlauben im Idealfall kausale Aussagen darüber, ob die Effekte in der CR-Box tatsächlich auf die Behandlung zurückzuführen sind, was durch den Pfeil zwischen ETR-Box und CR-Box veranschaulicht wird. Der Zusammenhang zwischen PR-Box und ETR-Box ist bei erfolgreicher Randomisierung für alle Prädiktorvariablen gleich Null, d.h. es bestehen keine systematischen Unterschiede zwischen Interventions- und Kontrollgruppe hinsichtlich der Ausgangsbedingungen wie soziodemografische Merkmale, Störungsbild oder Erkrankungsschwere.

Der „Südwest-Pfad" über die nichtexperimentelle Treatment-Box (NTR-Box) hebt hingegen mehr die Generalisierbarkeit der Ergebnisse im täglichen Versorgungsalltag und damit die externe Validität hervor. Hier finden sich korrelative bzw. quasi-experimentelle Versuchspläne. Bei der Ergebnisforschung in der psychosomatischen Rehabilitation war bislang ausschließlich die Realisierung von quasi-experimentellen Designs möglich. Ohne Verwendung einer Kontrollgruppe besteht allerdings immer die Gefahr einer Konfundierung zwischen PR- und NTR-Box, auch als „Selektion in das Treatment" bekannt (Wittmann et al., 2003). So erscheint es plausibel, dass motivierte Patienten zum einen bessere Behandlungsergebnisse erzielen und zum anderen auch eher bereit sind, die Fragebögen zur Messung eben dieser Behandlungsergebnisse auszufüllen. Dies konnte für psychosomatische Rehabilitanden im Rahmen einer Studie zur Ergebnisqualität stationärer psychosomatische Rehabilitation (EQUA-Studie, Schmidt, Steffanowski, Nübling, Lichtenberg & Wittmann, 2003) anhand von Fremdeinschätzungen der behandelnden Therapeuten belegt werden. So wurden von den 858 Studienteilnehmern 76,2% arbeitsfähig entlassen, 60,4% erhielten zum Entlass-Zeitpunkt eine günstige Prognose hinsichtlich der Hauptsymptomatik und lediglich 6,9% beendeten die Behandlung nicht regulär. Bei den 652 Patienten, die eine Teilnahme an der EQUA-Studie von vorneherein ablehnten, wurden lediglich 62,5% arbeitsfähig entlassen, nur 48,1% wiesen eine günstige Prognose zum Entlass-Zeitpunkt auf und die Abbrecherquote war mit 12,2% fast doppelt so groß.

Besteht keine Möglichkeit, randomisierte Untersuchungen durchzuführen, gibt es dennoch eine Reihe von Optionen zur Stärkung der Evidenz von Aussagen, die mit quasi-experimentellen Designs gewonnen wurden:

- Mehrpunkt-Erhebung: Es kann ein PRÄ-Test durchgeführt werden, um den Ausgangszustand vor der Behandlung zu erheben. Nach der Behandlung erfolgt dann die POST-Messung und die Differenz zwischen beiden Messungen wird zur Abbildung des Behandlungserfolges verwendet. Darüber hinaus kann zur Beurteilung der Stabilität des Therapieerfolgs eine katamnestische Erhebung z.B. ein Jahr nach Ende der Behandlung durchgeführt werden.
- Multizentrische Erhebung: Werden die Daten an mehreren Kliniken gleichzeitig erhoben, wird die Gefahr von Stichprobeneffekten verringert. Zeigen sich vergleichbare Ergebnisse in allen Substichproben, so steigt die Glaubwürdigkeit der gewonnenen Aussagen.
- Multimodale Erhebung: Die Einbeziehung unterschiedlicher Datenquellen wie Selbstangaben der Patienten, Fremdangaben von Klinktherapeuten, Hausärzten und Krankenkassen ist eine weitere Möglichkeit, um die Evidenz zu verbessern.

Auch die Erhebung unterschiedlicher Datenbereiche im Sinne des bio-psycho-sozialen Modells wie körperliches Befinden, psychische Symptomatik, soziale Kompetenz, Lebenszufriedenheit, Arbeitsfähigkeit und sozialmedizinisch relevante Kriterien lässt sich hier anführen.

- Multimethodale Erhebung: Die simultane Verwendung unterschiedlicher Erhebungsmethoden wie indirekte und direkte Veränderungsmessung, standardisierte Testverfahren, Therapiezielerreichungsskalierung, klinisches Interview kann sich förderlich auf die Validität der gewonnenen Aussagen auswirken. Die konvergente und diskriminante Konstruktvalidität lässt sich mit einer korrelativen Multitrait-Multimethod-Matrix (Campbell & Fiske, 1959) überprüfen. Verschiedene Methoden zu einem Konstrukt sollten dabei hoch, unähnliche Konstrukte, die mit einer bestimmten Methode gewonnen wurden, hingegen niedrig miteinander korrelieren (Amelang, Zielinski, Fydrich & Moosbrugger, 1997).
- Meta-Analyse: Mehrere ähnliche Einzeluntersuchungen lassen sich meta-analytisch integrieren, um auf diese Weise eine Steigerung der Evidenz zu erreichen (Farin, 1997). Neben der Ermittlung des durchschnittlichen Gesamteffektes erlauben Meta-Analysen darüber hinaus die Identifikation von Moderatorvariablen (z.B. Patienten- und Methodenmerkmale), welche Unterschiede bei der Effektausprägung erklären (Bosnjak, 2007). Auf diese Weise lassen sich mit der meta-analytischen Methodik Fragestellungen beantworten, die im Rahmen einer einzelnen Untersuchung nicht lösbar sind.

1.2.4 Aggregation und Symmetrie

Ein Hauptproblem des multivariaten Ansatzes bei Erhebung einer Vielzahl von Ergebniskriterien besteht in der Parameter- bzw. Faktoreninflation. Wittmann (1985) schlägt zur Lösung die Strategie der Aggregation vor:

> Aggregation bzw. Aufsummierung bringt meist den Effekt der Reliabilitätssteigerung unserer Meßinstrumente. Fehlervarianz wird dabei abgeschwächt und gemeinsame bzw. wahre Varianz in Relation zur Gesamtvarianz vergrößert. (S. 111)

Hierbei ist zu beachten, dass eine Aggregation in erster Linie bei korrelierten Variablen indiziert ist, wobei das Ziel verfolgt wird, redundante Informationen auf eine überschaubare Anzahl von zugrundeliegenden Dimensionen zu reduzieren. Erreicht werden kann dies durch die Anwendung faktorenanalytischer Verfahren, die dem Forscher sinnvolle Anhaltspunkte für die einer Datenstruktur zugrundeliegenden inhaltlichen Dimensionen liefern können.

Ein anderes Problem der multivariaten Forschung bezieht sich auf die Symmetrie zwischen Ursachen und Effekten. So hat man es in der sozialwissenschaftlichen Forschung häufig mit komplexen, hierarchisch aufgebauten Konstrukten zu tun. In der stationären psychosomatischen Rehabilitation kann das Konstrukt "Allgemeinbefinden" zusammengesetzt aus einer somatischen, psychischen und sozialen Komponente verstanden werden. Das psychische Befinden lässt sich wiederum in mehrere Unterbereiche wie Selbstver-

trauen", "Depressivität" oder "Angst" gliedern, die ihrerseits schließlich durch verschiedene Einzelindikatoren (singuläre Beobachtungskriterien), etwa durch Einzelfragen in einem Angstfragebogen, operationalisierbar sind. Ein weiterer wichtiger Bestandteil des Rahmenkonzeptes nach Wittmann ist daher das Symmetrieprinzip (Wittmann, 1985; Wittmann & Matt, 1986b; Wittmann 1990; Wittmann, 2002; Brunswik, 1955):

> Die Zusammenhänge unserer Operationalisierungen von Ursachen und Effekten werden nur dann maximal, wenn das Modell der Ursachen (PR-, ETR- und NTR-Box) zum Modell des Kriteriums (CR-Box) symmetrisch ist. Ein besonders wichtiger Aspekt besteht in der Möglichkeit, daß auf verschieden komplexen Hierarchieebenen jeweils hohe Zusammenhänge bestehen. Die Relationen sinken umso stärker, je asymmetrischer beide Modelle zueinander werden, unabhängig vom jeweiligen Generalitätsniveau. (Wittmann, 1985, S. 247)

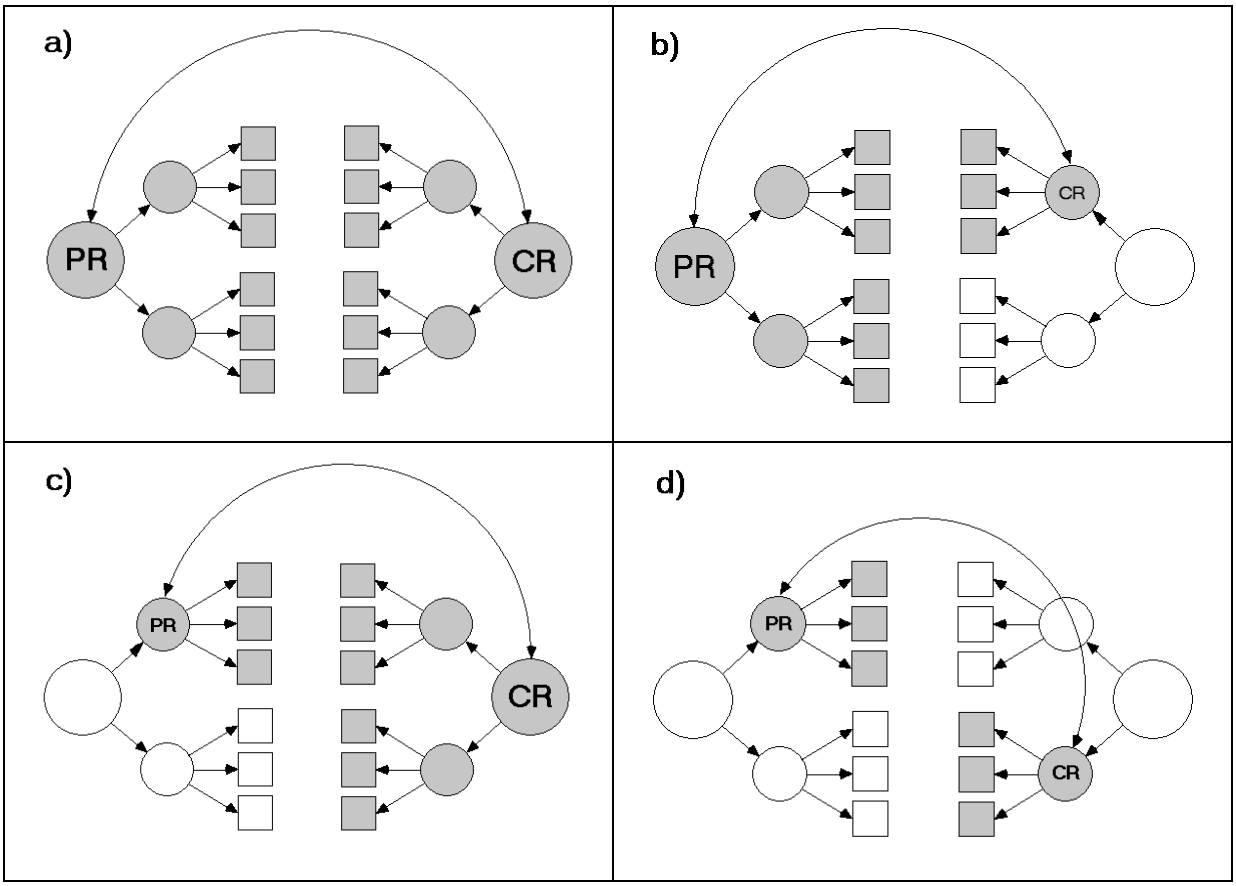


Abbildung 6. Veranschaulichung des Symmetrieprinzips. PR = Prädiktor, CR = Kriterium.

Abbildung 6 veranschaulicht das Symmetrieprinzip (Wittmann, 1990). Oben links in der Grafik (a) ist das Ideal einer vollständigen Symmetrie zwischen Prädiktor und Kriterium dargestellt. Oben rechts (b) ist die Situation veranschaulicht, dass ein enges Kriterium aus einem breiten Prädiktor vorhergesagt werden soll. Dies wäre z.B. dann der Fall, wenn die Veränderung des bio-psycho-sozialen Gesamtbefindens aller Patienten einer psychosomatischen Klinik ausschließlich mit einer Angstskala abgebildet werden soll. Auch der umgekehrte Fall eines engen Prädiktors und breiten Kriteriums (c) ist denkbar, etwa

wenn ein Patient mit einem umschriebenen Behandlungsanliegen (z.B. Kopfschmerzen) eine ganze Testbatterie mit Fragen ausfüllen muss, von denen die meisten für ihn irrelevant sind. Schließlich besteht noch die Möglichkeit, dass sich sowohl Prädiktor als auch Kriterium nur teilweise oder gar nicht überlappen (d). Dies wäre z.B. dann der Fall, wenn die Erfolgsbeurteilung bei Angstpatienten ausschließlich mit einem Depressionsfragebogen durchgeführt wird (wenngleich häufig Korrelationen zwischen depressiven Symptomen und Angstsymptomen bestehen). Eine Verletzung des Symmetrieprinzips äußert sich dergestalt, dass zwischen Prädiktoren und Kriterien nur geringe Zusammenhänge bzw. bei den Kriterien nur geringe Effektgrößen zu beobachten sind. Die hier ausgeführten Symmetrieüberlegungen spielen eine wichtige Rolle für die spätere Unterscheidung zwischen allgemeinen, gruppenbezogenen und individualisierten Ansätzen zur Evaluation der Ergebnisqualität.

1.2.5 Domänen der psychosomatischen Rehabilitation

Rossi et al. (1999) unterscheiden zwischen formativer und summativer Evaluation:

- Formative Evaluation umfasst vor allem Aktivitäten zur Verbesserung von Konzept und Umsetzung eines Programms. Damit bezieht sich die formative Evaluation eher auf die Struktur- und Prozessqualität.
- Summative Evaluation befasst sich vorwiegend mit der Frage, ob das Programm seine Ziele erreicht und versucht zu einer zusammenfassenden Bewertung der Ergebnisqualität zu gelangen.

Hieraus ist ersichtlich, dass eine formative Evaluation vorab geplant werden muß, um programmbegleitend die relevanten struktur- und prozessbezogenen Variablen zu erheben. Eine rein summative Evaluation kann auch post-hoc durchgeführt werden.

Mit der formativen und summativen Evaluation werden nach Rossi fünf Hauptdomänen eines Programms abgedeckt:

- Der Bedarf („Need") für ein Programm (Was ist das Problem und wen betrifft es auf welche Weise?)
- Das Programmdesign (Welche Interventionen sind geeignet und durchführbar? An welche Zielpopulation richtet sich die Intervention?)
- Die Programmimplementierung (Erreicht eine bestimmte Intervention die Zielpopulation? Ist die Intervention gut implementiert?)
- Die Ergebnisse bzw. Effekte des Programms (Ist die Intervention effektiv bei der Erreichung der gewünschten Ziele? Gibt es unerwünschte Nebenwirkungen?
- Die Effizienz des Programms (Wie viel kostet das Programm? Sind die Kosten des Programms im Verhältnis zum Nutzen gerechtfertigt?)

Bezieht man diese fünf Hauptdomänen auf die psychosomatische Rehabilitation, so lässt sich folgendes festhalten:

- Der Rehabilitationsbedarf ergibt sich aus den Vorgaben des SGB IX, nämlich Erhalt der Arbeitsfähigkeit und Teilhabe in der Gesellschaft sowie Linderung und

Heilung gesundheitlicher Beschwerden („Reha vor Rente" bzw. „Reha vor Pflege"). Bei der gesetzlichen Rentenversicherung erfolgt die Feststellung der Rehabilitationsbedürftigkeit einzelfallbezogen durch die sozialmedizinische Untersuchung, bei der auch eine Prognose abgegeben wird, ob die Rehabilitation in einer bestimmten Klinik bei dem gegebenen Beschwerdebild erfolgsversprechend ist. Auch Fragen der Über- bzw. Unterinanspruchnahme, der rechtzeitigen Zuweisung und angemessenen Behandlungsdauer werden hier angeschnitten.

- Das Programmdesign steht in enger Verbindung zum Einrichtungs- und Behandlungskonzept der jeweiligen psychosomatischen Klinik. Auch die therapeutische Ausrichtung (eher tiefenpsychologisch oder eher verhaltenstherapeutisch?) und angewandte Therapietheorie (warum eine bestimmte Intervention bei einem bestimmten Störungsbild indiziert ist und wie diese wirkt) gehört dazu.
- Bei der Programmimplementierung stellt sich die Frage, ob die Therapie auch tatsächlich im Sinne des Behandlungskonzeptes realisiert wird. Hier bestehen eindeutige Bezüge zur Prozessqualität. Diese lässt sich z.B. durch fortlaufende Routineerhebungen feststellen wie Basisdokumentationen oder Patientenbefragungen und durch entsprechende Maßnahmen zur Qualitätssicherung optimieren.
- Fragt man nach den Ergebnissen und Effekten der psychosomatischen Rehabilitation, so geht es hier zunächst um die Wirksamkeit und darum, ob mit dem Klinikaufenthalt wichtige Therapieziele erreicht werden (z.B. Besserung von Angst und Depressivität, Steigerung des Sebstvertrauens und der sozialen Kompetenz etc.). Auch die Untersuchung möglicher Nebenwirkungen und Kontraindikationen gehört zur Ergebnisevaluation. Als Ergebniskriterien werden hier vor allem Maße des subjektiven Befindens und der Lebenszufriedenheit eingesetzt.
- Bei Untersuchung der Effizienz wird der Aufwand in Beziehung zum erzielten Effekt gesetzt und damit die Frage gestellt, ob sich die Investition in die Behandlung „gelohnt" hat. Hierzu werden Dosis-Wirkungs- und Kosten-Nutzen-Berechnungen angestellt und auch monetäre Kriterien herangezogen wie z.B. Reduktion von Krankschreibungszeiten, Verbleib im Erwerbsleben, Vermeidung vorzeitiger Berentung, Verringerung des Inanspruchnahmeverhaltens hinsichtlich akutmedizinischer Leistungen und Medikamentenkonsum.

Im Rahmen der vorliegenden Arbeit sind vor allem die letzten beiden Punkte relevant, d.h. die Frage nach der Effektivität und Effizienz der durchgeführten Rehabilitationsmaßnahmen und damit nach der Ergebnisqualität.

1.2.6 Stakeholder der psychosomatischen Rehabilitation

Ein weiterer entscheidender Faktor bei jeder Evaluation sind nach Rossi et al. (1999) die „Stakeholder". Darunter sind sämtliche Personen, Individuen und Organisationen zu verstehen, die ein Interesse am Funktionieren des evaluierten Programms haben. Eine „faire" Evaluation sollte grundsätzlich immer die Interessen und Fragestellungen der wichtigsten Stakeholdergruppen abbilden, wenn die Ergebnisse der Evaluation eine breite Akzeptanz finden sollen. Bei der psychosomatischen Rehabilitation lassen sich folgende Stakeholdergruppen identifizieren:

- Patient selbst
- Privates Umfeld des Patienten (Familie, Partner, Freunde)
- Berufliches Umfeld des Patienten (Kollegen, Arbeitgeber)
- Klinikmitarbeiter (Ärzte, Pfleger, Therapeuten, Servicemitarbeiter)
- Klinikleitung (Ärztliche Leitung und Verwaltungsleitung)
- Klinikträger (Öffentliche Hand, GmbH oder Aktiengesellschaft)
- Kostenträger (Rentenversicherung, Krankenkasse)
- Politiker (Entscheidungsträger im Gesundheitswesen)
- Öffentlichkeit (Mediale Berichterstattung, Wirksamkeit der Behandlung bei bestimmten Störungsbildern wie z.B. Depressionen oder Ängsten, Weiterempfehlung einer Einrichtung, Werbung)
- Wissenschaft (Entwicklung neuer Behandlungskonzepte und Therapieverfahren als Antwort auf veränderte gesellschaftliche Bedingungen, Grundlagenforschung, Weiterentwicklung der Forschungsmethodik)
- Auftraggeber bzw. Geldgeber der Evaluation

Je nach Stakeholder können bei der Ergebnisevaluation unterschiedliche Bewertungskriterien für eine „erfolgreiche Behandlung" angelegt werden, was bei der Datenerhebung entsprechend zu berücksichtigen ist. Die Interessen unterschiedlicher Stakeholder können sich in manchen Fällen sogar widersprechen, etwa wenn ein Patient mit einem Rentenbegehren (Reha-Ziel: Vorzeitige Berentung) von der Rentenversicherung (Reha-Ziel: Vermeidung der vorzeitigen Berentung) in die Klinik geschickt wird.

1.3 Forschungsstand in der Psychosomatik

Inspiriert von den Ergebnissen der amerikanischen Psychotherapieforschung entstanden Anfang der 1990er Jahre die ersten Programmevaluationsstudien, in denen die Ergebnisqualität der stationären psychosomatischen Rehabilitation untersucht wurde.

Zur zuverlässigen Abschätzung der Effektivität einer Behandlungsform sollten allerdings nicht nur einzelne Studien betrachtet werden, sondern die Befunde der Einzelstudien

darüber hinaus systematisch sekundäranalytisch integriert werden. Die inzwischen häufigste Form solcher Sekundäranalysen ist die Meta-Analyse. Eine entsprechende metaanalytische Integration der mittlerweile in größerer Zahl vorliegenden empirischen Studien zu den Ergebnissen stationärer psychosomatischer Rehabilitation wurde erst kürzlich abgeschlossen. Dabei ergaben sich Therapieeffekte in mittlerer Größenordnung, die auch katamnestisch stabil sind.

1.3.1 Rehabilitationswissenschaftliche Evaluationsforschung

Auf die Frage nach dem Zweck evaluativer Ergebnisforschung geben Kordy und Scheibler (1984a, 1984b) folgende Antworten:

1. Legitimation der Behandlung (Nachweis, dass die Behandlung wirksam ist und keine schädlichen Nebenwirkungen hat)
2. Absicherung therapierelevanter Entscheidungen (Nachweis, für welche Störungsbilder eine bestimmte therapeutische Intervention indiziert ist und die größten Erfolgsaussichten hat).

An diesen beiden Hauptaufgaben der Ergebnisevaluation hat sich bis heute wenig geändert, allerdings ist angesichts der Vielzahl von positiven Befunden zur generellen Wirksamkeit psychotherapeutischer Interventionen (vgl. Abschnitt 1.3.3) der zweite Aspekt (differenzielle Forschung) in den letzten 20 Jahren mehr in den Vordergrund getreten. Ungeachtet dessen sollte jede neue Behandlungsform einen generellen Wirksamkeitsnachweis für die intendierte Zielpopulation erbringen, bevor über weiter differenzierende Fragestellungen nachgedacht wird.

Die rehabilitationswissenschaftliche Forschung in Deutschland ist noch eine vergleichsweise junge Disziplin. Nachdem in den 1980er Jahren die gesundheitspolitische Debatte zur Finanzierbarkeit des Gesundheitswesens einsetzte, mehrten sich kritische Stimmen, welche die Effektivität des Rehabilitationswesens insgesamt in Frage stellten. Auf diese Kritik haben Kostenträger, Kliniken und Wissenschaftler mit einer deutlichen Belebung und Intensivierung der Forschung reagiert.

So wurden in den 1990er Jahren auf Basis der Empfehlungen der Reha-Kommission (Verband Deutscher Rentenversicherungsträger, 1992) eine Reihe von Maßnahmen zur Intensivierung der Forschung, Verbesserung der Qualitätssicherung und Weiterentwicklung im Bereich der medizinischen Rehabilitation eingeleitet (Koch & Bengel, 2000; Haupt & Delbrück, 1998; Haaf & Schliehe, 1998). Hervorzuheben sind dabei folgende Punkte:

- Entwicklung von Mindestanforderungen für Rehabilitationseinrichtungen
- Entwicklung eines Rahmenkonzepts für die medizinische Rehabilitation und Ausarbeitung indikationsspezifischer Behandlungskonzepte (Verband Deutscher Rentenversicherungsträger, 1992)
- Entwicklung von Verfahren zur routinemäßigen Erfolgskontrolle (Schmidt, Nübling & Lamprecht, 1992)

- Einführung eines verbindlichen Qualitätssicherungsprogramms („5-Punkte-Programm der gesetzlichen Rentenversicherung", Verband Deutscher Rentenversicherungsträger, 1994)
- Gründung von neuen rehabilitationsmedizinischen Fachgesellschaften, Stiftungslehrstühlen und rehabilitationswissenschaftlichen Instituten
- Förderung rehabilitationswissenschaftlicher Untersuchungen, so insbesondere durch den vom Bundesministerium für Bildung und Forschung (BMBF) sowie Verband Deutscher Rentenversicherungsträger (VDR) eingerichteten Förderschwerpunkt „Rehabilitationswissenschaften" (Koch et al., 1995; Buschmann-Steinhage, Gerwinn, Klosterhuis & Mitreiter, 1998; Zwingmann, Buschmann-Steinhage, Gerwinn & Klosterhuis, 2004).

Das „5-Punkte-Programm" ist ein umfassendes Qualitätssicherungsprogramm für die stationäre medizinische Rehabilitation (Jäckel, Protz, Maier-Riehle & Gerdes, 1997; Protz, Gerdes, Maier-Riehle & Jäckel; Paar, 1997; Koch & Tiefensee, 1998). Hierbei wurden in Zusammenarbeit mit wissenschaftlichen Instituten und Rehabilitationsexperten neue Verfahren und Instrumente der Qualitätssicherung entwickelt, erprobt und flächendeckend in die Versorgungsroutine integriert. Das „5-Punkte-Programm" umfasst die Komponenten

- Klinikkonzepte
- Patiententherapiepläne
- Qualitätsscreening (Peer-Review)
- Patientenbefragung
- Qualitätszirkel

und deckt damit die drei Qualitätsdimensionen Struktur, Prozess und Ergebnis nach Donabedian (1966) ab. An dem Programm waren im Jahr 2001 insgesamt 961 Rehabilitationseinrichtungen beteiligt, davon 76 psychosomatische Einrichtungen. Die stationäre medizinische Rehabilitation ist damit der erste Versorgungsbereich im deutschen Gesundheitswesen, in dem systematisch und flächendeckend eine externe Qualitätssicherung betrieben wird. Die für die Qualitätssicherung benötigten Daten werden von der Deutschen Rentenversicherung ausgewertet und regelmäßig in Form von Qualitätsberichten an die Kliniken zurückgemeldet. Darin werden die individuellen Ergebnisse der Einrichtung vergleichend in Bezug zur jeweiligen Referenzgruppe dargestellt.

Der Förderschwerpunkt „Rehabilitationswissenschaften" (Koch et al., 1995; Buschmann-Steinhage et al., 1998; Zwingmann et al., 2004) wurde im Jahr 1998 gemeinsam vom Bundesministerium für Bildung und Forschung und der Deutschen Rentenversicherung ins Leben gerufen. Das Finanzvolumen lag bei 40,9 Millionen Euro und wurde von beiden Förderern je hälftig aufgebracht. Acht regionale Forschungsverbünde mit insgesamt 150 Forschungsprojekten wurden über einen zweiphasigen Zeitraum (1998-2002, 2001-2005) gefördert. Darüber hinaus wurden wissenschaftliche Geschäftsstellen und methodische Querschnittsprojekte initiiert, um die Infrastruktur der Verbünde zu stärken.

Zielsetzung des Förderschwerpunktes war die Steigerung von Qualität und Umfang anwendungsorientierter Forschung auf dem Gebiet der Rehabilitation. Auch die strukturelle Verankerung der Rehabilitationswissenschaften an den Universitäten sollte nachhaltig

gestärkt werden, um auch über die Förderphase hinaus eine wissenschaftlich begründete Weiterentwicklung der Rehabilitation sicherzustellen. So wurden an zwei Standorten (Würzburg und Halle) Stiftungsprofessuren eingerichtet und durch die Rentenversicherung gefördert.

Auch die EQUA-Studie (Schmidt, Steffanowski, Nübling, Lichtenberg & Wittmann, 2003), auf deren Daten die vorliegende Arbeit unter anderem basiert, entstand im Rahmen des Förderschwerpunktes „Rehabilitationswissenschaften" im Forschungsverbund Freiburg / Bad Säckingen. Leitthema des Verbundes ist „Zielorientierung in Diagnostik, Therapie und Ergebnismessung" (Bengel & Jäckel, 2000). Wichtige Verbundthemen sind die Erkennung von reha-relevanten Problemlagen im somatischen, psychosozialen und funktionalen Bereich, die Ermittlung von spezifischen operationalen Therapiezielen, die Messung des Grades der Zielerreichung sowie eine stärkere konzeptuelle Ausrichtung des Rehabilitationsprozesses an den Therapiezielen. Damit liegt ein Schwerpunkt des Forschungsverbundes Freiburg / Bad Säckingen bei der Ergebnisevaluation.

Durch den Förderschwerpunkt wurde eine positive und nachhaltige Entwicklung in der Rehabilitationsforschung initiiert (Zwingmann et al., 2004). Um einen weiteren Transfer der Ergebnisse in die Praxis zu gewährleisten, wurde im Sommer 2006 eine dritte Förderphase „Versorgungsnahe Forschung – chronische Krankheiten und Patientenorientierung" ausgeschrieben (Bundesministerium für Bildung und Forschung, 2006).

Neben diesen extern koordinierten Aktivitäten zur Belebung der Rehabilitationsforschung entstanden bereits in den frühen 1990er Jahren erste Programmevaluationsstudien, die selbständig von Rehabilitationseinrichtungen bzw. ihren Trägern initiiert wurden (Schmidt, 1991; Zielke, 1993), darüber hinaus begannen viele Kliniken mit der Durchführung von kontinuierlichen Basisdokumentationen und Patientenbefragungen (Schmidt et al., 1992; Mans, 1995; Pollmann, 1998; Nübling & Schmidt, 1998). So sollten auch erfolgreich laufende Therapieprogramme einem kontinuierlichen Routinemonitoring unterzogen werden, um auf eine Verschlechterung der Struktur-, Prozess- und Ergebnisqualität rechtzeitig reagieren zu können. Nübling und Schmidt (1998) skizzieren in ihrem „Zweigleisigen Modell der Qualitätssicherung", wie sich umfangreichere und zeitlich begrenzte Programmevaluationsstudien mit einem fortlaufenden Routinemonitoring einer minimalen Auswahl von qualitätsrelevanten Patienten-, Behandlungs- und Ergebnismerkmalen kombinieren lassen. Solche Routineerhebungen sollten ökonomisch durchführbar sein, um nach Möglichkeit die Gesamtheit aller Patienten zu erreichen und lassen sich z.B. durch die Einführung einer einheitlichen Basisdokumentation und Patientenbefragung realisieren (Schmidt, Nübling, Lamprecht & Wittmann, 1994; Steffanowski, Nübling, Schmidt & Löschmann, 2006; Nübling et al., 2007).

1.3.2 Ergebnisqualität in der psychosomatischen Rehabilitation

Die Frage, mit welchen Methoden und anhand welcher Kriterien der "Erfolg" psychosomatisch-psychotherapeutischer Behandlungen beurteilt werden soll, hat eine lange und kontroversenreiche Geschichte, die im Umfeld der wissenschaftlichen Psychotherapiefor-

schung, Ergebnisevaluation bzw. Ergebnisforschung angesiedelt ist (Kordy & Kächele, 1996). Theoretisch kann Ergebnisqualität als deskriptives Konstrukt betrachtet werden, das über verschiedene empirische Indikatoren erfasst werden kann (Nübling & Schmidt, 2000). Maßgeblich ist Ergebnisqualität über den „Therapieerfolg" definiert.

Im Laufe der Jahre gab es eine Vielzahl von Vorschlägen, wie man zu bedeutungsvollen Aussagen über die Effektivität und Effizienz psychotherapeutischer Maßnahmen kommen kann z.B. (Schulte, 1993; Grawe & Braun, 1994; Fydrich, Laireiter, Saile & Engberding, 1996; Deck & Röckelein, 1999; Muthny & Bullinger, 1999). Im Rahmen der wissenschaftlichen Ergebnisevaluation besteht heute Konsens darüber, dass die Datensammlung multidimensional, multimodal und multimethodal anzulegen ist (Kordy & Kächele, 1996; Wittmann, 1987). Weiterhin wird heute gefordert, dass der Behandlungserfolg durch katamnestische Untersuchungen in einem ausreichenden zeitlichen Abstand nach Abschluss der Behandlung überprüft wird, um die Stabilität des Behandlungsergebnisses beurteilen zu können. Die Erfassung der Ergebnisqualität erfolgt dabei im Hinblick auf die PRÄ-POST-Veränderung des Gesundheitszustandes des Patienten, wobei auch die Zufriedenheit des Patienten mit der Behandlung berücksichtigt werden sollte (Swart & Philbert-Hasucha, 1998).

Die „International Classification of Functioning, Disability and Health (ICF)" lässt sich dabei als konzeptueller Rahmen für die Klassifikation unterschiedlicher inhaltlicher Ebenen von Ergebnisqualität verwenden. Wichtige Neuerungen und Erweiterungen der aktuellen Fassung (DIMDI, 2005) betreffen den Übergang vom defizitorientierten Krankheitsfolgenmodell zum bio-psycho-sozialen Modell der funktionalen Gesundheit. Dieses berücksichtigt auch die individuelle Lebensumwelt einer Person sowie vorhandene Ressourcen. Im SGB IX zur Rehabilitation und Teilhabe behinderter Menschen wurden wesentliche Aspekte der neuen ICF aufgenommen (Bundesregierung, 2001). Zentraler Oberbegriff in der ICF ist die Funktionsfähigkeit, die drei Aspekte umfasst:

- Körperstrukturen und -funktionen (Körperteile und physiologische Funktionen von Körpersystemen; umfasst auch die psychologischen Funktionen; Störungsbegriff: Schädigung, Funktionsstörung)
- Aktivitäten (Durchführung einer Aufgabe oder Handlung durch einen Menschen; Störungsbegriff: Beeinträchtigung der Aktivität)
- Teilhabe (Partizipation, Einbezogensein in eine Lebenssituation; Störungsbegriff: Beeinträchtigung der Teilhabe)

Die Funktionsfähigkeit und Behinderung eines Menschen wird in der ICF als dynamische Interaktion zwischen dem jeweiligen Gesundheitsproblem und personen- bzw. umweltbezogenen Kontextfaktoren verstanden. So hängt für einen Rollstuhlfahrer die Teilhabe am Arbeitsleben z.B. entscheidend davon ab, inwieweit sein Arbeitsplatz für ihn barrierefrei zugänglich ist.

Auch wenn die Entwicklung von ICF-basierten Assessmentinstrumenten derzeit noch in der Erprobungsphase ist, wird deutlich, dass eine umfassende Evaluation eine Reihe von unterschiedlichen inhaltlichen Ebenen berücksichtigen muss. So sollte die Ergebnisbewertung im Bereich der Symptomatik, der Aktivitäten und der gesellschaftlichen Partizipation ansetzen sowie körperliche, psychische, soziale und leistungsbezogene Komponenten

einbeziehen. Im Interesse einer „fairen" Evaluation sind dabei grundsätzlich verschiedene Stakeholdergruppen zu berücksichtigen. Die daraus resultierenden Kriterienmaße lassen sich zwei Hauptgruppen zuordnen (Wittmann et al., 2002):

- monetär bewertbare Kriterien wie Finanzierbarkeit, Kosten/Nutzen der Behandlung, Verbleib im Erwerbsleben, Krankschreibungszeiten, Inanspruchnahme des Gesundheitssystems
- nicht monetär bewertbare Kriterien wie Patientenzufriedenheit, subjektives Befinden auf der körperlichen, psychischen, sozialen und funktionalen Ebene, Lebenszufriedenheit in verschiedenen Bereichen sowie Selbsterfahrung, Problemlösekompetenz und Sinnfindung.

Bei der Ergebnisevaluation ist darüber hinaus zwischen Effektivität und Effizienz zu differenzieren (Zielke, 1993). Während sich der Begriff der Effektivität auf die Wirksamkeit einer Behandlungsmaßnahme bezieht, werden beim Begriff der Effizienz zusätzlich Kosten-Nutzen-Überlegungen eingeführt, d.h. die beobachteten Effekte werden ins Verhältnis zum Aufwand (z.B. Behandlungsdauer, Therapiedosis, Behandlungskosten) gesetzt (vgl. Abschnitt 2.2.3).

1.3.3 Meta-Analysen zur Psychotherapieforschung

Seit Veröffentlichung der Pionierarbeit von Smith und Glass (1977), in der erstmals die Ergebnisse von 375 Psychotherapiestudien meta-analytisch integriert wurden, sind im angloamerikanischen Sprachraum mehr als einhundert Meta-Analysen vorgelegt worden, die sich mit der Effektivität von psychotherapeutischen Programmen befassen. In einer breit angelegten Übersichtsarbeit von Lipsey und Wilson (1993) wurden unter anderem auch 124 Meta-Analysen zum Thema „Psychische Gesundheit" zusammengefasst, die sich auf mehr als 6000 einzelne Studien beziehen. Überprüft wurde dort eine breite Palette von Behandlungsprogrammen wie allgemeine Psychotherapie, Depressions-, Angst- und Abhängigkeitsbehandlung, Bulimie, Individual-, Gruppen- und Familientherapie, kognitive Verhaltenstherapie, psychoedukatives Training, Verhaltenstherapie bei Delinquenten, Meditation, Entspannung, präventive Beratung, soziales Training, Gesundheitserziehung, Biofeedback, Raucherentwöhnung sowie psychologische Schmerzbehandlung. Keine einzige dieser 124 Meta-Analysen gelangt zu einer negativen Gesamteffektgröße und der über alle Arbeiten gemittelte Wert beträgt 0,58, was in der Terminologie von Cohen (1992) einem „mittleren" Effekt entspricht. Wenn man die Auswahl auf die 41 Meta-Analysen einschränkt, welche sich explizit auf Psychotherapie im engeren Sinne beziehen, so liegt die gemittelte Effektgröße sogar bei 0,68. Die Debatte um die generelle Wirksamkeit psychotherapeutischer Verfahren gilt angesichts der positiven Ergebnisse seit mehreren Jahren als entschieden und aktuelle Arbeiten befassen sich daher eher mit der Frage, unter welchen Bedingungen eine bestimmte Interventionsform für einen bestimmten Patienten wirksam ist und welche therapeutischen Prozessmerkmale dafür verantwortlich sind (differenzielle Indikation).

Auch im deutschsprachigen Raum wurden einige Meta-Analysen zur allgemeinen Effektivität von Psychotherapie durchgeführt, so beispielsweise von Wittmann und Matt (1986a) unter besonderer Berücksichtigung von methodischen Aspekten der einbezogenen 85 Studien. Ebenfalls viel Beachtung hat die Meta-Analyse von Grawe (Grawe, Donati und Bernauer, 1994; Orlinsky, 1994). Insbesondere die hinsichtlich der Therapieschulen differenzierenden und bewertenden Ergebnisse Grawes haben eine kontroverse methodische und gesundheitspolitische Diskussion ausgelöst (Fäh & Fischer, 1998; Leichsenring, 1996; Tschuschke, Kächele & Hölzer, 1994).

Speziell für die medizinische Rehabilitation in Deutschland sind hingegen bislang nur wenige Meta-Analysen vorhanden. Dieses Forschungsdefizit liegt in der noch relativ jungen wissenschaftlichen Tradition dieses Versorgungsbereichs begründet. Indikationsspezifische Meta-Analysen in der medizinischen Rehabilitation wurden hinsichtlich der Behandlung von Alkoholabhängigkeit (Süss, 1995), chronischem Tinnitus (Olderog, 1999; Schilter, 2000), der Therapie von Essstörungen (Herzog & Hartmann, 1997; Jacobi, Dahme & Rustenbach, 1997), der Effektivität von Rückenschulen (Maier-Riehle & Härter, 1996), Behandlung von chronischen Rückenschmerzen (Hüppe & Raspe, 2005) sowie zur Wirksamkeit von psychoonkologischen Behandlungsansätzen (Stump & Koch, 1998) vorgelegt.

1.3.4 Evaluationsstudien in der psychosomatischen Rehabilitation

Neben einer Vielzahl von kleineren Untersuchungen sind seit den 1990er Jahren eine Reihe von größeren Programmevaluationsstudien in der stationären psychosomatischen Rehabilitation durchgeführt worden. Dabei wurden diagnostisch heterogene Patientengruppen im Längsschnittverlauf untersucht, wobei ein ganzes Bündel von unterschiedlichen Ergebniskriterien zu mehreren Messzeitpunkten, zumeist Aufnahme (A), Entlassung (E) und 1-Jahres-Katamnese (K) erfasst wurde. Tabelle 3 enthält eine exemplarische Auswahl dieser Untersuchungen mit entsprechenden Publikationen. Die Ergebnisse dieser Studien liefern eine eindrucksvolle Dokumentation der positiven Behandlungseffekte dieses Versorgungsbereichs für eine Vielzahl von Patienten, die häufig mit einer langen Krankengeschichte, einer starken psychischen Belastung und ausgeprägten Multimorbidität eine stationäre psychosomatische Rehabilitation begonnen haben. Neben ausgeprägten Effekten zum unmittelbaren Entlass-Zeitpunkt ist bei einem großen Anteil der Patienten eine Stabilisierung der Behandlungseffekte in der 1-Jahres-Katamnese festzustellen, wobei das persönliche Umfeld, das Vorhandensein von Nachsorgeangeboten sowie positive und negative Lebensereignisse im Jahr nach der Rehabilitation eine wichtige Moderatorfunktion haben. Während die Symptomatik zum Entlass-Zeitpunkt besonders starke Besserungseffekte zeigt, um im Katamnese-Zeitraum durch die Exposition gegenüber den alltäglichen Belastungen wieder etwas anzusteigen (ohne dabei allerdings das Ausgangsniveau zu erreichen), werden hinsichtlich bewältigungsrelevanter Merkmale wie Selbsteffektivität, sozialer Kompetenz und Fähigkeit zur Alltagsbewältigung zum Entlass-

Zeitpunkt zunächst kleine bis mittlere Effekte berichtet, die in der Katamnese dann allerdings zu einer weiteren leichten Zunahme tendieren, was für die erfolgreiche Anwendung der in der Rehabilitation erworbenen Einsichten und neu erlernten Verhaltensstrategien spricht, was ganz im Sinne des Grundgedankens der Rehabilitation ist, chronisch kranken Menschen Fähigkeiten zur Selbsthilfe in ihrem privaten und beruflichen Leben zu vermitteln.

Tabelle 3. PRÄ-POST-Studien mit Katamnese (Auswahl)

Studie	**Publikation**	***N***
Zauberberg-Studie I	Schmidt (1991)	364
Bliestal-Studie	Sandweg, Sänger-Alt & Rudolf (1991)	1088
Zauberberg-Studie II	Nübling (1992)	565
BKK-Studie	Zielke (1993)	148
Berus-Studie	Broda, Bürger, Dinger-Broda & Massing (1996)	370
Reinerzauer Katamnese-Studie	Nübling, Schmidt & Wittmann (1999)	560
TWK-Studie	Nosper (1999)	297
Bad Kreuznacher Studie	Schulz, Lotz-Rambaldi, Koch, Jürgensen & Rüddel (1999)	376
Gelderland-Studie	Kriebel & Paar (2000)	345
Bad Grönenbacher Studie	Mestel et al. (2000)	800
Bad Herrenalber Studie	Nübling et al. (2000)	317
PROTOS-Studie	Gerdes, Weidemann & Jäckel (2000)	879
PRÄ-POST-Projekt	Bischoff et al. (2003)	144
EQUA-Studie	Schmidt et al. (2003)	858
INDIKA-Studie	Nübling et al. (2004)	324
Gesundheitsökonomie-Studie	Zielke et al. (2004a)	338

Anmerkung. *N* = Patientenanzahl pro Studie zu Beginn der stationären Behandlung.

1.3.5 Meta-Analyse stationärer psychosomatischer Rehabilitation

Erst kürzlich wurde eine meta-analytische Integration der mittlerweile vorliegenden Evaluationsstudien in der stationären psychosomatischen Rehabilitation vorgelegt (MESTA-Studie), die sich auf 65 Primärstudien mit einer Ausgangsstichprobe von insgesamt fast 30.000 Patienten bezieht (Steffanowski et al., 2007). Die Durchführung der MESTA-Studie orientierte sich an den Empfehlungen von Lipsey & Wilson, 2001), zum Ablaufschema vgl. auch Cooper (1982). Das Projekt wurde im Förderschwerpunkt Rehabilitationswissenschaften (vgl. Abschnitt 1.3.1) realisiert. Anhand der MESTA-Studie wird nachfolgend der Stand zur Ergebnisforschung in der stationären psychosomatischen Ergebnisforschung etwas detaillierter skizziert.

Die drei Hauptfragestellungen der MESTA-Studie waren:

- Abbildung der Forschungslandschaft, welche Studien wurden mit welchen Methoden bislang durchgeführt?
- Ermittlung der kurz- und längerfristigen Effektivität der stationären psychosomatischen Rehabilitation
- Erklärung von Unterschieden bei der Effektausprägung durch Moderatorvariablen wie Patientenmerkmale oder methodische Güte der einbezogenen Untersuchungen

Einschlusskriterien für die Studienauswahl bei der Literaturrecherche waren:

- Die Studie wurde in einer stationären psychosomatischen Rehabilitationseinrichtung in Deutschland durchgeführt.
- Es wurden empirische Behandlungsergebnisse berichtet, also keine rein korrelative Studien oder Untersuchungen, die sich ausschließlich auf Prozessmerkmale beziehen.
- Mindestens eine Vergleichsbedingung zur Abschluss- bzw. Katamnese-Messung war vorhanden, also entweder eine Kontrollgruppe ohne Behandlung oder aber ein Vor-Test zu Beginn der Behandlung (PRÄ-POST-Design).

Bis einschließlich 2004 konnten insgesamt 65 Primärstudien identifiziert werden, welche diese drei Einschlusskriterien erfüllten. Die Untersuchungen basieren in der Terminologie von Carlson und Schmidt (1999) durchweg auf naturalistischen Eingruppen-PRÄ-POST-Designs (SGPP), während experimentelle randomisierte Kontrollgruppenstudien mit Vor-Test (PPWC) bzw. ohne Vor-Test (POWC) fehlen. Aufgrund der besonderen gesetzlichen und rechtlichen Rahmenbedingungen in der rehabilitativen Versorgung in Deutschland, aber auch aus ethischen Gründen war es bislang nicht möglich, echte randomisierte Kontrollgruppen mit unbehandelten Patienten zu bilden (vgl. Abschnitt 0). Dennoch wäre es wünschenswert, hier künftig über mehr Evidenz zu verfügen; denkbar wären z.B. Wartelistenkontrollgruppen, falls eine echte Randomisierung nicht realisierbar sein sollte.

Ausgehend von den Studienfragestellungen wurde ein ausführliches Codierschema entwickelt und eine datenbankgestützte Eingabemaske erstellt. Erhoben wurden folgende Datenbereiche:

- Stichprobenmerkmale (Alter, Geschlecht, Schulbildung, Diagnosen etc.)
- Behandlungsbezogene Merkmale (Aufenthaltsdauer, Therapierichtung, etc.)
- Methodische Merkmale (Vollständigkeit, Nachvollziehbarkeit etc.)
- PRÄ-POST-Effekte (in Form von standardisierten d-Effektgrößen)
- Kriterienbezogene Merkmale (Eigenschaften der verwendeten Ergebnismaße)

Die überwiegende Mehrzahl der Studien (57 bzw. 87,7% von 65) stammt aus den Jahren 1995-2004. Bezogen auf die Messung zum Aufnahme-Zeitpunkt (A) in die Klinik wurden in den 65 Studien insgesamt 29.777 Patienten befragt. Dies entspricht einer Ausgangsstichprobe von durchschnittlich 458 Patienten pro Studie. 56 Studien berichten für insgesamt 24.682 Patienten Ergebnisse zum Entlass-Zeitpunkt (E). 46 Untersuchungen berichten katamnestische Ergebnisse (K, meist 1-Jahres-Katamnese) für insgesamt 10.080 Patienten. (Ausgangsstichprobe hier 16.486 Patienten, was einer durchschnittlichen

Rücklaufquote von 62,1% entspricht). Bei immerhin 32 Studien handelt es sich um Dreipunkterhebungen (also A-, E- und K-Messung vorhanden).

Hinsichtlich der Indikationsbreite wurden bei 67,7% aller einbezogenen Studien (44 von 65) heterogene Stichproben untersucht (z.B. alle aufgenommenen Patienten). Bei 32,3% (21 von 65) aller Untersuchungen wurde nur eine bestimmte Indikationsgruppe, etwa ausschließlich Patienten mit depressiven Störungen, befragt.

Der durchschnittliche Altersmittelwert über alle Studien beträgt 41,8 Jahre. Die Streuung der Mittelwerte beträgt 5,2 Jahre (Anmerkung zum Verständnis: Bei einer Meta-Analyse werden keine Messungen an einzelnen Personen, sondern statistische Kennwerte ganzer Studien, hier Mittelwerte, zusammengefasst). Hinsichtlich der Geschlechtsverteilung sind Frauen mit 64,0% stärker vertreten als Männer mit 36,0%. Was den Bildungsgrad angeht, so überwiegen Hauptschulabschlüsse (48,8%), gefolgt von Realschulabschlüssen (28,9%) und (Fach-)Abitur mit 22,3%.

Der Anteil verheirateter Patienten liegt bei 46,7%. Hinsichtlich der beruflichen Stellung sind 52,2% Angestellte, 23,6% Arbeiter, 10,6% Beamte und 4,5% Selbstständige. 9,2% waren nie erwerbstätig. Zum Zeitpunkt der Aufnahme waren 66,1% erwerbstätig und 16,6% arbeitslos. Kostenträger der Rehabilitation war in 65,8% aller Fälle die gesetzliche Rentenversicherung. Bei 28,2% finanzierte die gesetzliche Krankenversicherung und bei 6,0% eine private Krankenversicherung die Behandlung.

Hinsichtlich der Hauptdiagnosen nach ICD-10 sind depressive (29,9%) und somatoforme Störungen (18,2%) am häufigsten vertreten, gefolgt von Anpassungs- (8,3%) und Angststörungen (8,1%). Auch posttraumatische Belastungsstörungen (6,7%), Persönlichkeitsstörungen (5,6%), körperliche Erkrankungen (5,3%), Essstörungen (4,7%), Zwangsstörungen (4,6%) und Substanzmissbrauch (3,2%) finden sich unter den Hauptdiagnosen.

Die mittlere Erkrankungsdauer beträgt 8,0 Jahre (*SD* = 2,9 Jahre). 12,2% aller Patienten hatten zum Zeitpunkt der Aufnahme in die Klinik bereits einen Antrag auf vorzeitige Berentung gestellt. Damit ist ein erheblicher Grad an Chronifizierung festzustellen, was die dringende Frage nach einer Verbesserung der Früherkennung und Zuweisungspraxis aufwirft. 52,3% der Patienten hatten zum A-Zeitpunkt bereits Erfahrung mit ambulanter und 30,8% mit stationärer psychotherapeutischer Behandlung.

Bei Betrachtung der behandlungsbezogenen Merkmale ist auffallend, dass nur selten verwertbare Daten zur Art (z.B. Teilnehmeranteil an Entspannungsverfahren) und zum Umfang (z.B. Wochenstunden) der therapeutischen Anwendungen für die Studienstichproben dokumentiert wurden. Häufiger sind allgemeine Informationen zur Palette des therapeutischen Angebotes und zum Behandlungskonzept der jeweiligen Klinik. Von der therapeutischen Ausrichtung her ist das Klinikkonzept bei 69,5% eher tiefenpsychologisch und bei den übrigen 30,5% eher verhaltenstherapeutisch orientiert. Allerdings haben die meisten Einrichtungen ein integratives Konzept, das verschiedene therapeutische Schulen und Verfahren je nach individuellem Behandlungsanliegen der Patienten miteinander kombiniert. Die durchschnittliche Klinikgröße liegt über alle Studien gemittelt bei 194 Betten.

Im Durchschnitt wurden 91,1% aller Behandlungen regulär beendet. Die durchschnittliche mittlere Behandlungsdauer beträgt 56,5 Tage (*SD* = 15,8), bei fallendem Trend in den letzten Jahren auf dem Hintergrund des steigenden Kostendrucks im Gesundheitswesen (vgl. Abschnitt 1.1.6).

Neben den Stichprobenmerkmalen wurde auch die methodische Qualität der einbezogenen Studien analysiert. Die meisten Studien weisen eine gute methodische Qualität auf, wenngleich nicht immer alle Standards eingehalten wurden.

Insgesamt wurden 531 unterschiedliche PRÄ-POST-Kriterienmaße bzw. –skalen in den 65 Studien verwendet. Diese Vielfalt und Heterogenität an verwendeten Ergebnismaßen verdeutlicht die Notwendigkeit einer verbesserten Koordination der Forschungsbemühungen und Konsensfindung hinsichtlich der geeigneten Outcome-Maße. Durch den Förderschwerpunkt Rehabilitationswissenschaften ist hier eine positive Entwicklung in den letzten Jahren festzustellen.

Sämtliche Effektgrößen wurden in der MESTA-Studie nach Gleichung 1 berechnet, indem die PRÄ-POST-Mittelwertsdifferenz durch die bei Aufnahme gemessene Streuung geteilt wurde. Gemäß einer Einteilung nach Cohen (1992) lassen sich Effektgrößen ab 0,20 als klein, ab 0,50 als mittel und ab 0,80 als groß interpretieren.

$$d = \frac{M_{post} - M_{prä}}{SD_{prä}} \qquad (1)$$

Insgesamt gingen mehr als 2.000 Einzeleffekte in die Meta-Analyse ein. Diese wurden zunächst zu mittleren Studien-Effekten aggregiert und diese sodann zu einem über alle Studien gewichteten Gesamteffekt zusammengefasst. Dieser beträgt für den Vergleich zwischen A- und E-Messung 0,51 (k = 56 Studien) und für den Vergleich zwischen A- und K-Messung 0,41 (k = 46 Studien). Bezieht man nur durchgängige Messungen über alle drei Zeitpunkte in die Analyse ein (k = 32 Studien), so beträgt der gewichtete Gesamteffekt bei Entlassung 0,57 und bei Katamnese 0,49. Die Abnahme um nur 0,08 Punkte im Jahr nach Entlassung deutet darauf hin, dass die Behandlungseffekte relativ stabil sind und die Patienten in der Rehabilitation neu erworbene Fertigkeiten über einen reinen „Erholungseffekt" hinaus offenbar gut im Alltag umsetzen können. Der Größenordnung nach entsprechen die Effekte in der Terminologie von Cohen einem mittelgroßen Effekt und kommen den Befunden aus der Übersichtsarbeit von Lipsey und Wilson (1993), vgl. Abschnitt 1.3.3 recht nahe.

Berechnet man die gewichteten Gesamteffekte nur für bestimmte Indikationsgruppen, so zeigen sich für Patienten mit depressiven Störungen sogar große Effekte (d = 0,84 zum E-Zeitpunkt bzw. d = 0,67 zum K-Zeitpunkt). Schränkt man darüber hinaus die Auswahl der verwendeten Ergebniskriterien im Sinne des Symmetrieprinzips (Wittmann & Matt, 1986) auf störungsspezifische Messinstrumente ein, so beträgt der gewichtete Gesamteffekt über alle Studien bei depressiven Patienten zum E-Zeitpunkt 1,07 bzw. in der 1-Jahres-Katamnese 0,76.

Zur Identifikation von potentiell relevanten Moderatorvariablen wurde eine gewichtete multiple Regression berechnet (zur Methodik vgl. Lipsey & Wilson, 2001). Insgesamt konnten mehr als 80% der Effektvarianz durch Stichproben- und Methodenmerkmale der einbezogenen Untersuchungen und Outcome-Maße aufgeklärt werden! Als stärkster Einzelprädiktor erwies sich dabei die Behandlungsdauer: Jede zusätzliche Behandlungswoche lässt eine um 0,15 höhere katamnestischen Gesamteffektgröße erwarten. Dies lässt eine weitere Verkürzung der Aufenthaltsdauern mit kurzsichtigen Kostenargumenten im Hinblick auf die damit verbundene Gefährdung des gesamten Behandlungserfolges sehr fragwürdig erscheinen. Aber auch der Bildungsgrad, die Diagnosegruppe und die Erkrankungsdauer wiesen signifikante Partialgewichte auf. So steht ein geringer Bildungsgrad, eine hohe Chronifizierung sowie ein hoher Anteil von somatoformen Störungen negativ mit dem Behandlungserfolg in Relation. Auch methodische Merkmale haben offenbar einen Einfluss auf den gemessenen Behandlungserfolg: 10 Prozentpunkte mehr Rücklauf in der Katamnese lassen eine um etwa 0,06 Punkte geringere Effektgröße erwarten. Patienten, die erst im zweiten oder dritten Anlauf mit Erinnerungsschreiben oder –anrufen Auskunft erteilen, schätzen ihren Behandlungserfolg also kritischer ein als Patienten, die sofort auf das erste Anschreiben antworten. Aus diesem Grund empfiehlt es sich im Interesse der Repräsentativität bei katamnestischen Befragungen grundsätzlich ein Erinnerungsschreiben zu verschicken. Durch diese einfache Maßnahme lässt sich bereits häufig eine Erhöhung der Rücklaufquote in einer Größenordnung von 10-15% realisieren.

Nimmt man eine gesundheitsökonomische Perspektive ein, so ergibt sich folgendes Bild aus der MESTA-Studie: Immerhin 67,4% der Patienten sind ein Jahr nach der Behandlung wieder bzw. noch erwerbstätig, was auf dem Hintergrund der erheblichen Gefährdung der Erwerbsfähigkeit zu Beginn der Behandlung einen beachtlichen Erfolg darstellt. Der Anteil der vorzeitig berenteten Patienten hat sich im gleichen Zeitraum von 2,8% auf 5,6% erhöht. 16 Studien berichten Krankschreibungszeiten (AU) für den 12-Monats-Zeitraum vor und nach dem stationären Klinikaufenthalt (bezogen auf Patienten, die im gesamten Zeitraum sozialversicherungspflichtig erwerbstätig waren). Die gewichtete Gesamteffektgröße beträgt 0,30, was auf den ersten Blick vergleichsweise gering erscheint. AU-Zeiten weisen allerdings die Eigenschaft auf, stark rechtsschief verteilt zu sein, was allgemein zu hohen Streuungen und dadurch zu niedrigeren Effektgrößen als bei normal verteilten Daten führt. Absolut betrachtet nehmen die AU-Zeiten im Schnitt von 47,2 auf 31,9 AU-Tage pro Jahr ab, was einem Rückgang um 15,3 Tage bzw. einer Reduktion um 32,4% entspricht. Auch hinsichtlich der Entwicklung von Inanspruchnahme des Gesundheitssystems durch die Patienten (Krankenhausaufenthalte, Arztkontakte und Medikamentenkonsum) ergeben sich aus der MESTA-Studie positive Evidenzen.

Ausgehend von den berechneten Gesamteffekten wurde im Rahmen der MESTA-Studie auch eine Kosten-Nutzen-Analyse (Wittmann et al., 2002, zur Methodik vgl. Abschnitt 2.2.3) durchgeführt, wobei sowohl die direkten (Tagespflegesatz) als auch indirektem (Arbeitsausfall) Kosten des Klinikaufenthaltes berücksichtigt wurden. Der „Return on Investment" beträgt nach dieser Schätzung bereits 2 Jahre nach Ende der Rehabilitation 2:1 und steigert sich nach 7 Jahren auf 4:1, wenn der Therapieeffekt unter Zugrundelegung einer linearen Abnahme um 0,08 pro Jahr nach 7 Jahren „aufgebraucht" ist:

> Hochgerechnet auf 93.658 Behandlungsfälle pro Jahr (...) entspricht dies einem gesamtgesellschaftlichen Nutzen der stationären psychosomatischen Rehabilitation von 3,0 Milliarden Euro pro therapiertem Patientenjahrgang! (Steffanowski et al., 2007, S. 118)

Zu Ergebnissen in ähnlicher Größenordnung beim Kosten-Nutzen-Verhältnis gelangen auch Zielke (1993) bzw. Zielke et al. (2004b). Die Ergebnisse verdeutlichen, dass beim Einnehmen einer langfristigen Perspektive der gesellschaftliche Nutzen psychotherapeutischer Interventionen offenbar massiv unterschätzt wird und das in den 1980er Jahren vorgebrachte Argument der „Milliardenverschwendung" (Kanzow, 1986) nicht haltbar ist. Die in Abschnitt 1.1.1 erwähnte Zunahme psychischer Erkrankungen erfordert im Gegenteil den Ausbau von qualifizierten Behandlungsangeboten. So gehen Wittchen und Jacobi (2006) davon aus, dass nur etwa ein Viertel aller Betroffenen durch das Versorgungssystem erreicht werden. Das Argument der „Milliardenverschwendung" lässt sich auf diesem Hintergrund durchaus auch umgekehrt verwenden, wenn man nach den Opportunitätskosten für eine nicht bzw. zu spät erfolgte adäquate psychotherapeutische Behandlung fragt. Mit der stationären psychosomatischen Rehabilitation verfügt die Bundesrepublik über einen wichtigen und effizienten Versorgungszweig für Patienten mit psychischen Störungen.

2 Strategien der Ergebnisevaluation

Kordy und Scheibler (1984a) führen aus, dass die Evaluation von Psychotherapie zwei Schritte umfasst:

- (...) Sammeln von Informationen (Daten) über den Patienten oder dem Messen von Eigenschaften;
- (...) Bewertung der Daten anhand a priori festgelegter Kriterien nach mehr oder weniger festen Regeln. (S. 220).

Auch wenn beide Aspekte in der Praxis nicht streng voneinander zu trennen sind, sollten diese doch voneinander differenziert werden, um Missverständnissen vorzubeugen. So fließen bereits bei Auswahl der zu messenden Ergebnisaspekte auch immer bestimmte Stakeholderinteressen ein bzw. die Bewertungsmöglichkeiten hängen auch immer von der vorher verwendeten Messmethode zur Gewinnung der Daten ab. Je nachdem, welcher Weg bei der Messung und Bewertung der Behandlungsergebnisse eingeschlagen wird, resultieren verschiedene Evaluationsstrategien, die sich hinsichtlich des Individualisierungsgrades voneinander unterscheiden.

2.1 Ergebnismessung

Ein bestimmtes Ergebniskriterium lässt sich unabhängig von seiner inhaltlichen Ausgestaltung durch unterschiedliche methodische Herangehensweisen erfassen. Als Alternative zur häufig eingesetzten indirekten Veränderungsmessung (PRÄ-POST-Messung) bieten sich im Rahmen von Einpunkterhebungen die quasi-indirekte und direkte Veränderungsmessung an. Darüber hinaus existieren zielorientierte Ansätze zur Ergebnismessung, welche sich durch einen hohen Individualisierungsgrad der Datenerhebung auszeichnen.

Eine wichtige Voraussetzung für die einheitliche Ergebnisbewertung unabhängig von der Skalierung des jeweils verwendeten Assessmentinstruments ist im Rahmen von Meta-Analysen die Standardisierung der Messungen. Dies ist unter anderem durch die Berechnung von Effektgrößen möglich, diese Vorgehensweise hat sich durch die meta-analytische Forschung mittlerweile als Standard etabliert.

Bei jeder Veränderungsmessung stellt sich auch die Frage nach dem Ausgangszustand der Befragten. So wurde in der Vergangenheit zur Lösung dieses Problems häufig vorgeschlagen, Veränderungswerte anhand des Ausgangszustandes statistisch zu korrigieren. Mit dem zielorientierten Messansatz wird hingegen eine andere Vorgehensweise verfolgt, indem für die spätere Erfolgsbewertung nur Aspekte bei einem bestimmten Patienten berücksichtigt werden, bei denen zu Beginn der Behandlung eine auffällige Belastung vorliegt.

2.1.1 Indirekte, quasi-indirekte und direkte Veränderungsmessung

Fragen zur allgemeinen Veränderungsmessung spielen in der Ergebnisevaluation eine zentrale Rolle. Zur Erfassung von Veränderungen bzw. zur Einschätzung des Therapieerfolgs wird in der Rehabilitationsforschung aus Gründen der Praktikabilität und Ökonomie meist auf die Fragebogenmethode zurückgegriffen. Eine entsprechende Übersicht zu gängigen Instrumenten sowie Empfehlungen geben Muthny & Bullinger (1999) Hinsichtlich des Individualisierungsgrades lassen sich dabei generische und indikations- bzw. störungsspezifische Instrumente unterscheiden.

Bei der Veränderungsmessung lassen sich drei Zugangsweisen unterscheiden, die allesamt mit unterschiedlichen Vor- und Nachteilen verbunden sind (Stieglitz & Baumann, 1994; Kohlmann & Raspe, 1998; Stieglitz, 1990; Schmidt & Nübling et al., 2001; Schmidt et al., 2003):

- die indirekte Veränderungserfassung (iVM)
- die quasi-indirekte Veränderungserfassung (qVM) mit retrospektivem PRÄ-Test
- die direkte Veränderungserfassung (dVM)

Die iVM stellt als klassisches „PRÄ-POST-Design" den gebräuchlichsten Ansatz der Erfassung von Veränderungsinformationen dar. Bei diesem Verfahren wird die Ausprägung eines interessierenden Merkmals jeweils vor und nach der Behandlung (wiederholte Statusmessung) erfasst. Die durch Subtraktion bestimmbaren Messwertdifferenzen (Vergleiche zwischen PRÄ- und POST-Test) stellen somit indirekt gewonnene Veränderungsmaße dar (Abbildung 7).

Der Vorteil dieser Methode besteht darin, dass zeitnahe Informationen über den Ausgangszustand des noch unbehandelten Patienten gewonnen werden. Allerdings erfordert die iVM mehrere Messzeitpunkte, was einen hohen organisatorischen Aufwand bei der Datenerhebung mit sich bringt. Dieser Aspekt ist insbesondere im Rahmen von Routinebefragungen zur Qualitätssicherung problematisch. Darüber hinaus bestehen methodische Probleme wie Regression zur Mitte, die Frage nach der Reliabilität von Differenzmaßen sowie mögliche Veränderungen im Bezugssystem des Patienten, welche die Vergleichbarkeit wiederholter Statusmessungen generell in Frage stellt (Petermann, 1978). Dies impliziert die Gefahr der Interpretation von statistischen Artefakten bei der Berechnung von Korrelationsmaßen zwischen Differenzmaßen und anderen Ergebnisindikatoren.

Die qVM erfordert nur einen Messzeitpunkt, bei dem neben dem POST-Status die notwendige PRÄ-Messung retrospektiv, also im Nachhinein aus der Erinnerung, erhoben wird. Auch hier werden die Veränderungsinformationen durch Differenzbildung zwischen beiden Statusangaben, also zwischen der POST-Messung und der retrospektiven PRÄ-Messung (RETRO) gewonnen. Abbildung 8 veranschaulicht die Vorgehensweise.

Indirekte Veränderungsmessung (iVM)

Differenzbildung zwischen
PRÄ-Status und POST-Status

iVM = POST – PRÄ

PRÄ-Messung (vor der Behandlung) Mein Gesundheitszustand ist ...				
1	2	3	4	5
sehr gut	gut	mittel-mäßig	schlecht	sehr schlecht

POST-Messung (nach der Behandlung) Mein Gesundheitszustand ist ...				
1	2	3	4	5
sehr gut	gut	mittel-mäßig	schlecht	sehr schlecht

Abbildung 7. Messrational der indirekten Veränderungsmessung.

Quasi-Indirekte Veränderungsmessung (qVM)

Differenzbildung zwischen erinnertem PRÄ-Status und aktuellem POST-Status

qVM = POST – RETRO

Retrospektive PRÄ-Messung (nach der Behandlung) Mein Gesundheitszustand **war** ...				
1	2	3	4	5
sehr gut	gut	mittel-mäßig	schlecht	sehr schlecht

POST-Messung (nach der Behandlung) Mein Gesundheitszustand **ist** ...				
1	2	3	4	5
sehr gut	gut	mittel-mäßig	schlecht	sehr schlecht

Abbildung 8. Messrational der quasi-indirekten Veränderungsmessung.

Ein Vorteil besteht in der ökonomischen Anwendbarkeit dieser Methode, da nur eine einmalige Messung nach Ende der Behandlung benötigt wird und dennoch Informationen über den Ausgangszustand des Patienten gewonnen werden. Damit eignet sich der Ansatz in allen Kontexten, in denen ein echter PRÄ-Test nicht möglich ist. Allerdings tauchen hier die gleichen methodischen Probleme (z.B. Regression zur Mitte) bei der Verwendung von Differenzwerten auf wie bei der klassischen iVM. Darüber hinaus stellt sich die Frage nach der Validität von retrospektiven Angaben, da ein Bias durch Gedächtniseffekte, d.h. eine Beeinflussung der retrospektiven Messung durch das aktuelle Befinden, nicht ausgeschlossen werden kann. Andererseits besteht der Vorteil der quasi-indirekten Messung darin, dass beide Einschätzungen zum gleichen Zeitpunkt aus erfolgen und somit nur eine Fehlerkomponente durch situative Faktoren oder Tagesform gegeben ist.

In der Literatur wird berichtet, dass retrospektive Angaben ein pessimistischeres Bild vom Ausgangszustand vermitteln als tatsächliche PRÄ-Messungen (Stieglitz, 1990; Steffanowski, Lichtenberg, Nübling, Wittmann & Schmidt, 2003), was zu einer Überschätzung der Behandlungserfolge führt (Response-Shift). Andererseits kann aber auch argumentiert werden, dass ein Jahr nach der Behandlung eine vermehrte Problemeinsicht des Patienten vorhanden ist, was zu einer realistischeren Bewertung des tatsächlichen Ausmaßes der damaligen Beeinträchtigung führt. Damit würde die klassische Methode der indirekten Veränderungsmessung die tatsächlichen Behandlungserfolge unterschätzen. Es hat sich gezeigt, dass retrospektive und reale Einschätzungen von Befindlichkeiten in mittlerer Größenordnung korrelieren und Veränderungswerte, die auf der qVM basieren, sogar höher mit Fremdeinschätzungen korrelieren können als Veränderungswerte, die anhand der iVM gewonnen wurden (Stieglitz, 1990).

Auch die dVM erfordert nur einen Messzeitpunkt nach der Intervention. Bei diesem Zugang wird die subjektiv erlebte Veränderung direkt im Sinne einer Vergleichsaussage eingestuft (Abbildung 9).

Direkte Veränderungsmessung (dVM)				
POST-Messung (nach der Behandlung) Mein Gesundheitszustand hat sich im Vergleich zu vorher ...				
1	2	3	4	5
deutlich gebessert	etwas gebessert	nicht verändert	etwas verschlechtert	deutlich verschlechtert

Abbildung 9. Messrational der direkten Veränderungsmessung.

Vorteile bestehen wie bei der qVM in der besonderen Ökonomie des Verfahrens, welches sich damit für Routinebefragungen und Untersuchungen eignet, bei denen ein PRÄ-Test zu aufwendig oder überhaupt nicht mehr zu realisieren ist. Die dVM steht dem Prozesscharakter des menschlichen Erlebens nahe und entspricht der kognitiven Eigenschaft, Urteile durch Vergleichsrelationen (hier „besser-schlechter") zu bilden. Das Hauptproblem der dVM besteht darin, dass bei ihrer ausschließlichen Anwendung keinerlei Information über den Ausgangszustand des Patienten verfügbar ist. So kann die Aussage „nicht verändert" sowohl die Beibehaltung eines erwünschten Zustandes als auch das Fortbestehen eines unerwünschten Zustandes beinhalten, was keine eindeutige Erfolgsbewertung zulässt. Darüber hinaus wird diskutiert, ob der Ansatz tatsächlich Veränderung und nicht etwa einfach das momentane Befinden erfasst (Kastner & Basler, 1997). Kohlmann und Raspe (1998) berichten lediglich mäßige Übereinstimmungen zwischen direkten und indirekten Veränderungsmaßen und kommen zu dem Schluss, dass die dVM eine eigenständige Veränderungsdimension abbildet. Schmidt et al. (2003) konnten hingegen zeigen, dass sich aus indirekten, quasi-indirekten und direkten Veränderungsmaßen ein gemeinsamer Veränderungsfaktor bilden lässt und berichten beachtliche Übereinstimmungen zwischen den drei methodischen Ansätzen.

2.1.2 Regression zur Mitte und Residual Gain Scores

Zur Methodik der iVM existiert eine umfangreiche Diskussion in der Literatur, die sich vor allem mit der Reliabilität von Veränderungsmessungen sowie der Abhängigkeit der POST-Messung bzw. der berechneten Differenzwerte von der PRÄ-Messung auseinandersetzt. Insbesondere die Problematik der Regression zur Mitte (RZM) wird kontrovers diskutiert, was auch Ausgangspunkt für die Entwicklung entsprechender Korrekturverfahren zur Adjustierung von PRÄ-POST-Differenzen war.

Überlegungen hinsichtlich einer möglichen Überschätzung von Behandlungserfolgen durch RZM sollten vor allem bei quasi-experimentellen PRÄ-POST-Designs ohne Kontrollgruppe angestellt werden. Bei einem randomisierten Design stellt sich dieses Problem nicht. So weisen bei erfolgreicher Randomisierung Behandlungs- und Kontrollgruppe die gleichen Ausgangsbedingungen auf. Verglichen werden hier die POST-Messungen beider Gruppen und nicht die PRÄ-POST-Differenzen. So gehen Cronbach und Furby (1970) sogar soweit, die Empfehlung auszusprechen, auf die Verwendung von Differenzwerten zu verzichten:

> There appears to be no need to use measures of change as dependent variables and no virtue in using them. If one is testing the null hypothesis that two treatments have the same effect, the essential question is whether Post-Test Y_{oo} Scores vary from group to group. Assuming that errors of measurement of Y are random, Y is an entirerly suitable dependent variable. (S. 78)

Im Eingruppen-PRÄ-POST-Design besteht diese Option allerdings nicht. Eine wirklich aussagekräftige indirekte Veränderungsmessung wäre hier eigentlich nur unter der Prämisse möglich, dass alle Patienten zu Beginn der Behandlung die gleichen Ausgangsbedingungen aufweisen. Dies ist in der Praxis natürlich nie der Fall, so werden Patienten mit unterschiedlichem Beeinträchtigungsgrad in die Klinik aufgenommen. Je schlechter es einem Patienten zu Beginn der Behandlung aber geht, desto größer ist die Wahrscheinlichkeit, dass allein aufgrund von Zufallseffekten eine positive Veränderung beobachtet wird. Zur Erklärung dieses Phänomens können Boden- bzw. Deckeneffekte herangezogen werden: Befindet sich ein Patient bei der PRÄ-Messung bereits am oberen Ende einer Symptomskala, so ist die ansonsten gültige Annahme, dass Messfehler gleichermaßen nach oben und unten auftreten und sich somit zu Null ausmitteln, außer Kraft gesetzt. Da sich der Patient in diesem Fall nur noch in eine Richtung verändern kann, resultiert bei der POST-Messung durch Zufallseffekte eine einseitige Veränderung auf der betreffenden Skala weg vom oberen Ende. Diese Beziehung zeigt sich in einer negativen Korrelation zwischen der PRÄ-Messung und dem berechneten Differenzwert und wird auch als Regression zur Mitte (RZM) bezeichnet. Je größer die Abweichung vom statistischen Mittelwert, desto stärker ist in der Regel die entsprechende Regression zur Mitte (Furby, 1973; Nesselroade, Stigler & Baltes, 1980). Im Rahmen der EQUA-Studie (Schmidt et al., 2003) konnte gezeigt werden, dass die iVM bei Therapiebeginn stark beeinträchtigte Patienten

bei der Erfolgsbewertung leicht begünstigt, während die dVM die Ergebnisse weniger stark beeinträchtigter Patienten in einem etwas günstigeren Licht darstellt.

Angesichts der Problematik der RZM wurden zur Schätzung „wahrer Differenzwerte", „Residual Gain Scores" bzw. „Base-Free Measure of Changes" in der Vergangenheit vor allem regressionsanalytische Ansätze vorgeschlagen (Lord, 1956; Dubois, 1957; Lord, 1958a, 1958; McNemar, 1958). Tucker, Damarin und Messick (1966) postulieren dabei, dass sich die empirisch gemessene PRÄ-POST-Differenz aus zwei Komponenten zusammensetzt, wobei die eine Komponente komplett abhängig von der PRÄ-Messung und die andere Komponente komplett unabhängig von der PRÄ-Messung ist. Letztere soll als „True Independent Gain Score" das Ausmaß an tatsächlicher „wahrer Veränderung" repräsentieren.

Die Berechnung von Veränderungsresiduen („Residual Gain Score") zur Korrektur der RZM bei indirekten Veränderungsmessungen soll an dieser Stelle zur Veranschaulichung der Diskussion exemplarisch dargestellt werden. So basiert die Regressionsmethode zur Berechnung von Veränderungsresiduen darauf, dass die POST-Messung zunächst aus der PRÄ-Messung vorhergesagt wird. Die Abweichung der POST-Messung von dem vorhergesagten Wert (bzw. von der Regressionsgeraden) stellt somit den „Residual Gain Score" dar, aus dem alle anhand des PRÄ-Wertes linear vorhersagbaren Anteile entfernt wurden. So lässt sich der POST-Wert x_2 aus dem PRÄ-Wert x_1 nach Gleichung 2 schätzen.

$$\hat{x}_2 = b \cdot x_1 + a \tag{2}$$

Der individuelle Residualwert D_{iRes} der Veränderung für einen bestimmten Patienten *i* berechnet sich dann nach Gleichung 3, indem man den individuellen PRÄ-Wert x_{i1} und POST-Wert x_{i2} des betreffenden Patienten nach Bestimmung der Parameter *a* und *b* in die Regressionsgleichung einsetzt.

$$D_{i\mathrm{Res}} = x_{i2} - \hat{x}_{i2} = x_{i2} - b \cdot x_{i1} + a \tag{3}$$

Cronbach und Furby (1970) kritisieren an diesem Ansatz, dass durch die Entfernung der vorhersagbaren Varianz aus dem Residual Gain Score auch wichtige Aspekte tatsächlicher Veränderung verloren gehen:

> Residualizing removes from the Post-Test score, and hence from the gain, the portion that could have been predicted linearily from pretest status. One cannot argue that the residualized score is a "corrected" measure of gain, since in most studies the portion discarded includes some genuine and important change in the person. The residualized score ist primarily a way of singling out individuals who changed more (or less) then expected. (S. 74)

Cronbach und Furby (1970) führen weiter aus, dass die Anwendung einer derartigen Korrektur impliziert, dass die PRÄ- und POST-Werte in irgendeiner Weise systematisch über- oder unterschätzt sind oder aber die berechneten PRÄ-POST-Differenzen durch andere

Wirkvariablen als die Behandlung zustande gekommen sind – eine Annahme, die sich nur schwer rechtfertigen lässt.

Für das Eingruppen-PRÄ-POST-Design, bei dem es um die Frage geht, ob eine bestimmte Behandlung zu einer statistisch signifikanten Veränderung bzw. einer bestimmten Effektgröße führt, empfehlen Cronbach und Furby (1970) daher die Verwendung der unkorrigierten PRÄ- und POST-Werte:

> An estimate of true gain might appear to be pertinent. But it is not. For if one were to estimate D_{oo} for each individual, and average, he would arrive back at the sample mean of observed gain. A significance test need only to ask whether μ_Y is reliably different from μ_X. The difference in sample means for X and Y ist the best available estimate of the mean D." (S. 79)

Eine andere vorgebrachte Kritik an der indirekten Veränderungsmessung kann auf die Aussage von Lord (1956) zurückverfolgt werden, wonach Differenzwerte zwischen zwei Statusmessungen sehr viel unreliabler seien als die Statusmessungen selbst. Rogosa und Willett (1983) führen hierzu aus, dass diese Annahme nur in Extremfällen gilt, wenn sich alle Individuen einer Gruppe in einem fast identischen Ausmaß verschlechtern oder verbessern. Nur in diesem Fall ist das Differenzmaß nicht geeignet, interindividuelle Unterschiede hinsichtlich der Veränderung zu erfassen und entsprechend unreliabel. Darüber hinaus muss Regression zur Mitte bei Veränderungsmessungen nicht zwangsläufig auftreten, sondern ist nach Nesselroade et al. (1980) nur dann zu erwarten, wenn die Korrelation zwischen PRÄ-Messung und Differenzwert negativ ist (vgl. auch Rogosa (1995).

Insgesamt entsteht bei Sichtung der damaligen stark mathematisch dominierten Diskussion der Eindruck, dass alle Schätzverfahren und Korrekturformeln die Abbildung von Ergebnisqualität letztendlich künstlich verkomplizieren, anstatt diese einfacher zu gestalten. Problematisch ist auch die Abhängigkeit der Veränderungsresiduen von der jeweiligen Stichprobe (der Mittelwert von Residualwerten beträgt per Definition Null), so dass nach Durchführung der Korrektur keine Vergleichbarkeit mit anderen Stichproben, Studien oder Einrichtungen mehr gegeben ist.

Unbeantwortet bleibt dabei immer die Frage, wie bestimmte Veränderungen im Hinblick auf den Ausgangs- und Endzustand zu bewerten sind. Das eigentliche Problem bezieht sich somit nicht auf den Messaspekt und damit verbundene Fragen der Reliabilität, sondern auf die evaluative Frage nach der unterschiedlichen Bedeutung eines bestimmten Messergebnisses (z.B. einer PRÄ-POST-Differenz von 0,00) bei unterschiedlichen Voraussetzungen (z.B. gutes oder schlechtes Befinden) auf einem bestimmten Beurteilungsaspekt zu Beginn der Behandlung im Hinblick auf die Ergebnisqualität nach der Behandlung.

Eine ausschließliche Verwendung von PRÄ-POST-Differenzen zur Bewertung der Ergebnisqualität erscheint auf diesem Hintergrund nicht ratsam, diese Kritik richtet sich auch gegen die unkritische Publikation von PRÄ-POST-Effektgrößen ohne ergänzende Informationen zum Ausgangszustand der untersuchten Patientenstichprobe. Bei angemessener Berücksichtigung des PRÄ-Status bzw. POST-Status bei der evaluativen Bewertung liefern die gemessenen PRÄ-POST-Differenzen jedoch wertvolle und unerlässliche Informationen zur Beurteilung des Behandlungserfolges.

Eine wichtige Voraussetzung für ein umfassenderes Verständnis von Veränderungsprozessen ist auch die Anzahl der verwendeten Messzeitpunkte. So stellen Nesselroade et al. (1980) fest, dass erst die Verwendung von mehr als zwei Messzeitpunkten eine sinnvolle Abschätzung von Veränderungsprozessen und Regressionseffekten ermöglicht. Rogosa (1995) führt aus, dass zwei Messzeitpunkte zwar eine Schätzung der Größe der Veränderung, nicht aber der individuellen Wachstumskurve erlauben. So ist eine beliebige Vielzahl von exponentiellen oder logistischen Wachstumskurven denkbar, die durch die beiden Punkte der PRÄ- und POST-Messung verlaufen könnten.

In den letzten Jahren hat eine erhebliche Weiterentwicklung der Methoden zur Erfassung von Veränderungsprozessen stattgefunden, wobei ein wichtiger Fortschritt in der zunehmenden Verwendung von Strukturgleichungsmodellen (Latent Growth Curve Models) zu sehen ist, die eine Integration der klassischen faktoren-, regressions- und pfadanalytischen Methoden ermöglichen. Für den interessierten Leser sei an dieser Stelle auf die entsprechende Literatur verwiesen, so z.B. auf die Herausgeberwerke von Gottman (1995), Collins und Sayer (2001) und Moskowitz und Hershberger (2002).

2.1.3 Standardisierte Ergebnisdarstellung mit Effektgrößen

Zur Ergebnisdarstellung bei PRÄ-POST-Vergleichen (indirekte Veränderungsmessung) wird zunehmend auf die Effektgrößenmetrik als Ergänzung zum klassischen Signifikanztest zurückgegriffen. Ein wichtiger Ausgangspunkt dieser Entwicklung war die Meta-Analyse von Smith und Glass (1977). Zur Berechnung der Effektgröße wird der beobachtete Unterschied zwischen zwei Mittelwerten durch die Streuung geteilt, d.h. die Differenz zwischen Behandlungs- und Kontrollgruppe (bzw. zwischen PRÄ- und POST-Messung) wird in Form von z-standardisierten Werten dargestellt. Hauptvorteil dieser Vorgehensweise ist, dass auf diese Weise auch unterschiedlich skalierte Testverfahren direkt miteinander vergleichbar werden.

Ein weiteres Argument für die Verwendung von Effektgrößen ist der Umstand, dass Signifikanzaussagen von der Größe der jeweils verwendeten Stichprobe abhängig sind und somit kaum über das tatsächliche Ausmaß einer beobachteten Veränderung Auskunft geben (Leonhart, 2004). Diese einseitige Dominanz des Signifikanztests in der Forschung führt zu Fehlinterpretationen. So werden bei sehr großen Stichproben auch sehr kleine, für die praktische Anwendung vielleicht irrelevante Effekte „hoch signifikant", was bei Entscheidungsträgern zu Fehlinvestitionen führen kann. Umgekehrt werden bei kleinen Stichproben aufgrund zu geringer Teststärke nur große Effekte statistisch signifikant, was dazu führt, dass vielversprechende Ansätze nicht mehr weiterverfolgt bzw. nicht anhand einer größeren Stichprobe noch einmal überprüft werden.

Innerhalb der Literatur werden für PRÄ-POST-Vergleiche ohne Kontrollgruppe (konventionelle indirekte Veränderungsmessung) verschiedene Berechnungsvarianten von Effektgrößen diskutiert. Anlass zur Kontroverse gibt vor allem das Streuungsmaß, durch welches die Mittelwertsdifferenzen zu dividieren sind (vgl. Hartmann & Herzog, 1995; Maier-Riehle & Zwingmann, 2000):

- Division der mittleren PRÄ-POST-Differenz durch die Standardabweichung der PRÄ-Messung bei Aufnahme (Gleichung 1 in Abschnitt 1.3.5), vgl. Kazis, Anderson & Meenan, 1989)
- Division der mittleren PRÄ-POST-Differenz durch die gepoolten Standardabweichungen der PRÄ- und POST-Messung (Gleichung 4, vgl. Hedges & Olkin, 1985)
- Division der mittleren PRÄ-POST-Differenz durch die Standardabweichung der Differenzwerte (Gleichung 5, vgl. Gerdes, 1998).

$$d = \frac{M_{post} - M_{prä}}{SD_{pool}} = \frac{M_{post} - M_{prä}}{\sqrt{\frac{(N_{prä} - 1) \cdot SD_{prä}^2 + (N_{post} - 1) \cdot SD_{post}^2}{N_{prä} + N_{post} - 2}}} \quad (4)$$

$$d = \frac{M_{post} - M_{prä}}{SD_{diff}} = \frac{M_{post} - M_{prä}}{\sqrt{SD_{prä}^2 + SD_{post}^2 - 2 \cdot r \cdot SD_{prä} \cdot SD_{post}}} = \frac{M_{post} - M_{prä}}{SD_{prä} \cdot \sqrt{2 \cdot (1 - r)}} \quad (5)$$

Während bei Gleichung 1 (siehe Abschnitt 1.3.5) lediglich die PRÄ-Streuung im Nenner steht, fließt bei Gleichung 4 eine zweite Größe, die POST-Streuung in die Berechnung ein. Eine Varianzerweiterung durch die Behandlung führt zu niedrigeren, eine Varianzverringerung hingegen zu höheren Effektgrößen (Hartmann & Herzog, 1995). Maier-Riehle und Zwingmann (2000) merken hierzu kritisch an, dass jegliche Varianzreduktion durch höhere Effektgrößen „belohnt" wird, etwa auch dann, wenn sich sämtliche Patienten auf einem suboptimalen Niveau stabilisieren. Umgekehrt bedeutet eine Varianzerweiterung nicht automatisch, dass sich das Befinden bei einzelnen Patienten verschlechtert haben muss, theoretisch können alle Befragten von der Behandlung profitiert haben, wenn auch in unterschiedlichem Ausmaß. Eine Verwendung von gepoolten Streuungen mag jedoch sinnvoll sein, wenn die Stichprobe eine stark eingeschränkte PRÄ-Streuung aufweist und damit die Gefahr einer Überschätzung der Effektgrößen bei Anwendung von Gleichung 1 besteht.

Bei Gleichung 5 wirkt sich darüber hinaus die Korrelation zwischen PRÄ- und POST-Messung auf die Höhe der resultierenden Effektgröße aus. Bei einem hohen Zusammenhang zwischen PRÄ- und POST-Messung kann die Effektgröße deutlich höher ausfallen (bei einer Korrelation von 1.00 zwischen PRÄ- und POST-Messung würde sie theoretisch gegen Unendlich gehen). Es stellt sich daher die Frage, mit welcher Begründung ein stärkerer Effekt postuliert werden sollte, nur weil eine höhere PRÄ-POST-Korrelation vorliegt.

Das Hauptproblem der beiden Berechnungsvarianten in Gleichung 4 und Gleichung 5 bei mehr als zwei Messzeitpunkten besteht darin, dass die Streuung der Differenzmaße dazu tendiert, bei kurzem zeitlichen Abstand zwischen den Messungen eher gering zu sein und im Katamnese-Zeitraum zuzunehmen, was

> selbst bei anhaltendem Behandlungserfolg, d.h. konstanten Mittelwerten bei den Nacherhebungszeitpunkten – mit zunehmender zeitlicher Distanz zu abnehmenden Effektstärken führt (...). (Maier-Riehle & Zwingmann, 2000, S.196)

Dies widerspricht der Logik von Effektgrößemaßen als Schätzer der zentralen Tendenz für den durchschnittlichen Therapieerfolg einer Population. Die PRÄ-Streuung der noch unbehandelten Patientengruppe ist die beste Schätzung für die Streuung einer hypothetischen unbehandelten Kontrollgruppe im Sinne des klassischen Ansatzes von Smith und Glass (1977), da bei gelungener Randomisierung keine Unterschiede der Ausgangsverteilung zwischen Interventions- und Kontrollgruppe zu erwarten sind.

Ein methodisches Problem bleibt allerdings bei allen drei hier skizzierten Berechnungsvarianten bestehen: Die Stichprobenabhängigkeit der jeweils zur Standardisierung verwendeten Streuungsvariante und damit die Frage nach der adäquaten Kalibrierung der Messergebnisse (Sechrest, McKnight & McKnight, 1996). So resultiert bei gleicher Mittelwertsdifferenz rein rechnerisch ein höhere Effektgröße und damit eine optimistischere Erfolgsbewertung, wenn die Stichprobenstreuung gering ist. Diese Gefahr der Überschätzung von Effekten besteht vor allem bei Extremgruppen mit besonders hoher Ausgangsbelastung und entsprechenden Deckeneffekten.

Ein Lösungsansatz zum Umgang mit dieser Problematik könnte darin bestehen, grundsätzlich die Streuung einer repräsentativen klinischen (bzw. gesunden) Referenzstichprobe zur Berechnung der Effektgrößen heranzuziehen, sofern eine entsprechende Normierung verfügbar ist (Gleichung 6). Während dies bei publizierten klinischen Standardtestverfahren häufig der Fall ist, kann bei selbstentwickelten Instrumenten in der Regel nicht vom Vorhandensein entsprechender Bezugsparameter ausgegangen werden.

$$d = \frac{M_{post} - M_{prä}}{SD_{Norm}} \tag{6}$$

Die Berechnung von Effektgrößen soll dem Leser anhand eines hypothetischen Datenbeispiels veranschaulicht werden: Zwei Depressionsbehandlungen werden im Hinblick auf ihre Wirksamkeit miteinander verglichen. Zur Erfolgskontrolle wird die Depressivität jeweils vor (PRÄ) und nach (POST) der Behandlung gemessen. In Studie 1 wird das Beck-Depressionsinventar BDI (Hautzinger, Bailer, Worall & Keller, 1994) und in Studie 2 die Depressivitätsskala der Symptom-Checkliste SCL-90-R (Franke, 2002) verwendet. In Studie 1 habe sich im BDI eine Abnahme des Gruppenmittelwerts von 24 auf 20 Punkte ergeben. Die Streuung der PRÄ-Messung betrage 10 Punkte. In Studie 2 habe sich in der SCL-90-R eine Abnahme von 1,25 auf 0,75 Punkte ergeben, die Streuung der PRÄ-Messung betrage hier 0,65 Punkte.

Welche Behandlung ist nun wirksamer? Auf den ersten Blick erscheint die PRÄ-POST-Differenz in Studie 1 größer als in Studie 2. Allerdings haben die beiden Testverfahren einen unterschiedlichen Wertebereich, weshalb die Ergebnisse nicht direkt miteinander vergleichbar sind. Die Lösung des Problems besteht nun darin, die PRÄ-POST-Differenzen

aus den beiden Untersuchungen anhand der jeweiligen Streuung zu standardisieren (Tabelle 4).

Tabelle 4. Berechnung von PRÄ-POST-Effektgrößen mit zwei hypothetischen Beispielen

Allgemeine Formel für *d* bei PRÄ-POST-Vergleichen	Studie 1: Messung mit BDI	Studie 2: Messung mit SCL-90-R
$d = \frac{M_{PRÄ} - M_{POST}}{SD_{PRÄ}}$	$d = \frac{24 - 20}{10} = \frac{4}{10} = 0{,}40$	$d = \frac{1{,}25 - 0{,}75}{0{,}65} = \frac{0{,}50}{0{,}65} = 0{,}77$

Die in Tabelle 4 berechneten Effektgrößen lassen sich jetzt direkt miteinander vergleichen. So fällt die Verbesserung des Befindens in Studie 2 mit einer Effektgröße von 0,77 fast doppelt so groß aus wie in Studie 1 mit 0,40.

Wie lässt sich *d* praktisch interpretieren? Eine gebräuchliche Einteilung stammt von Cohen (1992). Demnach lassen sich Effekte ab 0,20 als klein, ab 0,50 als mittel und ab 0,80 als groß klassifizieren. Letztlich kann dies nur als Richtschnur dienen, in der Praxis erscheint es angebracht, eine bestimmte Effektgröße immer im Kontext zu den berichteten Effektgrößen aus vergleichbaren Untersuchungen zu bewerten. Immerhin korrespondiert die Einteilung von Cohen recht gut mit den Befunden von Lipsey und Wilson (1993, 2001), die anlässlich einer Integration von mehr als 300 Meta-Analysen von Interventionen im Gesundheits- und Bildungswesen zu folgender Effektgrößenverteilung gelangten: Eine Effektgröße von 0,30 entspricht dem Übergang zwischem dem ersten und zweiten Quartil (bis hier reichen somit die „schlechtesten" 25% aller einbezogenen Studien). Eine Effektgröße von 0,50 entspricht dem Median der Verteilung (50% aller einbezogenen Studien liegen ober- bzw. unterhalb dieses Wertes) und eine Effektgröße von 0,67 entspricht dem Übergang zwischen dem dritten und vierten Quartil (Beginn der „besten" 25% aller einbezogenen Studien).

Eine weitere, ergänzende Darstellungsweise der Effektgrößen ist als „Binomial-Effect-Size-Display" (BE*SD*) in meta-analytischen Untersuchungen gebräuchlich und erlaubt eine anschauliche Interpretation der Effekte im Sinne von prozentualen „Erfolgswahrscheinlichkeiten" (Rosenthal & Rubin, 1983; Lipsey & Wilson, 2001). Ausgangspunkt ist die Annahme, dass bei einer unbehandelten Patientengruppe die Wahrscheinlichkeit gleich groß ist, sich im Befinden zu verbessern (50%) oder zu verschlechtern (50%). Kontrastiert man die Verteilung der unbehandelten Kontrollgruppe mit der Verteilung der behandelten Gruppe, so lässt sich die Erfolgswahrscheinlichkeit für einen „statistisch durchschnittlichen" Patienten anhand der Standard-Normalverteilung ermitteln. Eine Effektgröße von 0,50 bedeutet demnach, dass sich die behandelte Gruppe um 0,50 Standardabweichungseinheiten in ihrem Befinden gebessert hat und es einem behandelten Patienten im statistischen Durchschnitt besser geht als 69% aller unbehandelten Patienten (Abbildung 10).

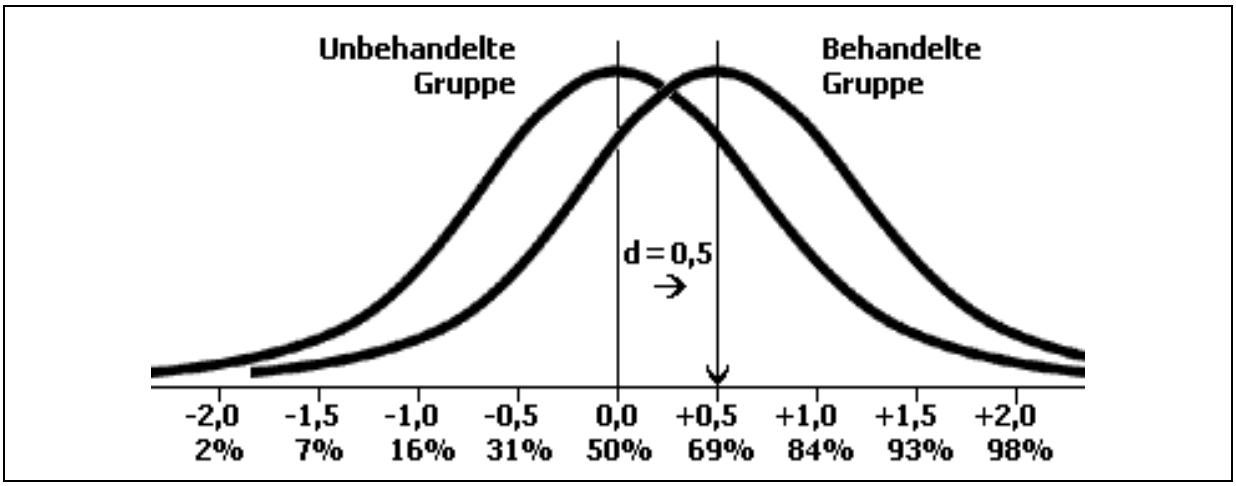


Abbildung 10. Veranschaulichung von d anhand der Standard-Normalverteilung.

Überträgt man die Erfolgswahrscheinlichkeiten für eine Verbesserung bzw. Verschlechterung des Befindens für beide Gruppen in eine Vierfeldertafel, (Tabelle 5) so lässt sich darüber hinaus das vor allem in der medizinischen Forschung gebräuchliche Odds-Ratio nach Gleichung 7 berechnen (Schwarzer, Türp & Antes, 2004). Für die hier im Beispiel angenommene Effektgröße von 0,50 mit einer korrespondierenden Erfolgswahrscheinlichkeit von 69% resultiert demnach ein Odds-Ratio (*OR*) für den „durchschnittlichen Patienten" von 2,23, wenn man die Erfolgswahrscheinlichkeiten aus Tabelle 5 in Gleichung 7 einsetzt. Praktisch bedeutet Odds Ratio, dass die Chance für eine Verbesserung des Befindens bei einem durchschnittlichen behandelten Patienten 2,23 mal und damit mehr als doppelt so groß ist wie für einen durchschnittlichen unbehandelten Patienten, vorausgesetzt, dass sich das Befinden bei 50% aller unbehandelten Patienten verbessert und bei 50% aller unbehandelten Patienten verschlechtert.

Tabelle 5. Erfolgswahrscheinlichkeiten bei einer Effektgröße von 0,50

	Wahrscheinlichkeit für eine Verbesserung des Befindens	**Wahrscheinlichkeit für eine Verschlechterung des Befindens**
Behandelte Gruppe	a = 69%	b = 31%
Unbehandelte Gruppe	c = 50%	d = 50%

$$OR = \frac{a}{b} \Big/ \frac{c}{d} = \frac{a \cdot d}{b \cdot c} \qquad (7)$$

2.1.4 Das Ausgangswertproblem

Ein entscheidendes methodisches Problem bei Verwendung der allgemeinen Veränderungsmessung zur Erfolgsbewertung der Rehabilitation besteht darin, dass keine angemessene Berücksichtigung des Ausgangszustandes erfolgt. Dies betrifft alle drei in Abschnitt 2.1.1 skizzierten Varianten:

Bei der klassischen indirekten Veränderungsmessung (iVM) wird der Zustand vor (PRÄ) und nach (POST) der Behandlung gemessen. Die Mittelwertsdifferenz bzw. die daraus abgeleitete Effektgröße wird sodann zur Bewertung des Behandlungserfolges verwendet. Bei dieser Vorgehensweise wird der entscheidende Aspekt vernachlässigt, dass die Patienten mit unterschiedlichen Voraussetzungen die Behandlung beginnen. So kann ein Nulleffekt (d = 0,00) bedeuten, dass ein Patient in einem schlechten Zustand verharrt, was in diesem Fall als Behandlungsmisserfolg zu bewerten wäre. Die gleiche gemessene Differenz von 0,00 kann aber auch bedeuten, dass ein Patient auf dem betreffenden Ergebnisparameter sowohl vor als auch nach der Behandlung ein positives Befinden aufweist, etwa mit seinem Familienleben oder mit seiner beruflichen Situation zufrieden ist. In diesem Fall müsste das gleiche Messergebnis als Behandlungserfolg bewertet werden, wenn man davon ausgeht, dass sich das Befinden des Patienten auf dem betreffenden Parameter ohne die Intervention möglicherweise verschlechtert hätte.

Für die quasi-indirekte Veränderungsmessung (qVM) gilt diesbezüglich das Gleiche wie für die iVM. Auch hier werden Mittelwertsdifferenzen zur Bewertung der Behandlungserfolge herangezogen, der einzige Unterschied besteht darin, dass der erinnerte PRÄ-Zustand anstelle der echten PRÄ-Messung vom POST-Zustand subtrahiert wird.

Bei der direkten Veränderungsmessung (dVM) stellt sich das Problem der mangelnden Berücksichtigung des Ausgangszustandes noch dringlicher, da bei ausschließlicher Verwendung dieses Ansatzes etwa im Rahmen von Einpunktmessungen überhaupt keine Anhaltspunkte über das Befinden zu Beginn der Behandlung verfügbar sind. Die Ankreuzung „unverändert" kann Verschiedenes beinhalten: Die Stagnation auf einem dysfunktionalen Niveau „unverändert schlecht", oder aber die Beibehaltung eines erwünschten Zustandes „unverändert gut".

Selbst wenn mit der iVM, qVM oder dVM rein rechnerisch eine Verbesserung des Befindens festgestellt wurde, so geht daraus nicht hervor, ob die Veränderung in klinischer Hinsicht groß genug ist, um von einem Behandlungserfolg im Sinne einer „Heilung" sprechen zu können. Neben dem in Abschnitt 2.1.1 skizzierten Versuch, dem Ausgangswertproblem durch mathematische Korrekturverfahren zu begegnen, setzen sich andere Ansätze mit dem Individualisierungsgrad der Messung auseinander.

2.1.5 Allgemeine, gruppenspezifische und individuelle Messung

Hinsichtlich der zur Erfassung der Veränderungs- bzw. Ergebnisinformationen verwendenden Messstrategie differenzieren Kordy und Scheibler zwischen drei Ansätzen und diskutieren verschiedene Vor- und Nachteile, die damit verbunden sind:

- Bei der allgemeinen Messstrategie werden bei allen untersuchten Patienten die gleichen Eigenschaften gemessen, indem eine möglichst große Zahl von Variablen und Methoden eingesetzt wird. Hierdurch wird versucht, zu einem möglichst umfassenden und objektiven Bild der Ergebnisqualität zu gelangen. Als Vorteile sind die Verwendbarkeit von vorhandenen standardisierten Messinstrumenten und die daraus resultierende Vergleichbarkeit über verschiedene Stichproben hinweg zu nennen. Als Nachteile sind der hohe Abstraktionsgrad der aggregierten Gesamtmaße sowie die auch von Gerdes, Bengel und Jäckel (2000) angesprochene Irrelevanz-Problematik zu nennen, wonach viele der einbezogenen Variablen für die Beurteilung der Ergebnisse eines einzelnen Patienten keine Bedeutung haben.
- Bei der gruppenspezifischen Messstrategie werden bei umschriebenen Patientengruppen störungsspezifische Messverfahren eingesetzt, etwa wenn nur Patienten mit einer diagnostizierten depressiven Störung ein Depressionsinventar vor und nach der Behandlung ausfüllen. Durch diese Vorgehensweise ist eine bessere Beachtung von Symmetrieprinzip (vgl. Wittmann, 2002) und Relevanzproblematik (vgl. Gerdes et al., 2000) gegeben und die zu erwartenden Effektgrößen sind höher als bei der allgemeinen Messstrategie. Problematisiert wird die Tatsache, dass der gruppenspezifische Messansatz gewisse Anforderungen an die Homogenität und Vergleichbarkeit der damit untersuchten Gruppen stellt, was nicht immer ohne weiteres vorausgesetzt werden kann.
- Die individuelle Messstrategie zielt darauf ab, für jeden einzelnen Patienten ein spezifisches Messinstrument bzw. eine Auswahl von denjenigen Variablen zu verwenden, bei denen eine Veränderung durch die Behandlung intendiert ist. Ersteres entspricht dem Goal Attainment Scaling (GAS), letzteres entspricht der Zielorientierten Ergebnismessung (ZOE). Auch wenn hier versucht wird, das individuelle Behandlungsanliegen optimal zu berücksichtigen, so ergeben sich methodische Probleme hinsichtlich der Vergleichbarkeit und Berechnung von statistischen Parametern. Darüber hinaus nimmt mit zunehmendem Individualisierungsgrad die Ökonomie der Verfahren und damit deren Anwendbarkeit im Rahmen von Routineerhebungen und größeren Studien, etwa im Rahmen von Meta-Analysen oder Klinikvergleichen, ab.

2.2 Ergebnisbewertung

In der Praxis stellt sich die Frage, wie die gemessenen Veränderungen und Behandlungsergebnisse im Hinblick auf die Interessen der Stakeholder einer Evaluation zu bewerten sind. Neben der Definition von Cut-Off-Kriterien zur Unterscheidung zwischen klinischen und gesunden Stichproben sind auch statistische Ansätze wie der Reliable Change Index zur Unterscheidung zwischen einem veränderten und unveränderten Befinden gebräuchlich. Kosten-Nutzen-Analysen können dazu dienen, die gewonnenen Veränderungsinformationen in monetär bewertbare Einheiten zu „übersetzen", um politischen Entscheidungsträgern entsprechende Bewertungshilfen zur Verfügung zu stellen. Eine angemessene Berücksichtigung des Ausgangswertproblems ist durch die Kombination von Status- und Veränderungsinformationen möglich, wobei eine gebräuchliche Methode darin besteht, PRÄ- und POST-Wert in einem Scatterplot darzustellen und die entsprechenden Cut-Off-Werte zur Klassifikation des Behandlungserfolges der Patienten in die Grafik einzutragen.

2.2.1 Zielorientierte Ergebnismessung

„Zielorientierung in Diagnostik, Therapie und Ergebnismessung" ist übergeordnetes Leitthema des Forschungsverbundes Freiburg Bad Säckingen im Förderschwerpunkt „Rehabilitationswissenschaften" des Bundesministeriums für Bildung und Forschung sowie der Deutschen Rentenversicherung (vgl. Abschnitt 1.3.1). Gerdes et al. (2000) führen zum Förderschwerpunkt aus, dass sich eine wissenschaftlich entwickelte Rehabilitation unter anderem durch folgende Merkmale auszeichnet:

- Die rehabilitationsrelevanten Problemlagen individueller Patienten können auf den verschiedenen Dimensionen (somatisch, funktional, psychosozial und edukativ) detailliert beschrieben und Fallgruppen zugeordnet werden, die durch jeweils ähnliche Problemlagen gekennzeichnet sind. (...).
- Für die verschiedenen Fallgruppen können auf den rehabilitationsrelevanten Dimensionen kurz-, mittel-, und langfristige Therapieziele bestimmt werden, deren Erreichung gemessen oder nach transparenten Kriterien beobachtet werden kann. (...).
- Für die verschiedenen Therapieziele sind therapeutische Verfahren bzw. kurz-, mittel- und langfristige Therapieprogramme verfügbar, die mit einer angebbaren Wahrscheinlichkeit zur Zielerreichung führen. (...). (S. 3-4)

Die große Komplexität der Rehabilitationsthematik impliziert bei der Ergebnismessung eine besondere methodische Herausforderung: Da die Ziele körperliche, psychosoziale, funktionale und edukative Aspekte beinhalten können, muss eine entsprechend große Anzahl von möglicherweise relevanten Ergebnisaspekten erfasst werden. Für einen be-

stimmten Patienten sind in der Regel jedoch nur einige wenige dieser Aspekte individuell relevant.

Wie in Abschnitt 2.1.1 und 2.1.4 ausgeführt, taucht im PRÄ-POST-Design die Problematik auf, dass die Patienten unterschiedliche Ausgangsbedingungen zu Beginn der Behandlung aufweisen und dadurch bereits rein rechnerisch eine unterschiedliche Chance haben, sich auf den berechneten Differenzmaßen zu verbessern. Kordy und Kächele (1996) führen zum Ausgangswertproblem folgendes aus:

> Die Anwendung des Konzeptes für Fragen der Erfolgskontrolle in der Psychotherapie basiert auf der Überlegung, daß für die Patienten, die in einer Klinik ein Therapieangebot erhalten, unterschiedliche Ausgangsbedingungen gegeben sind, die sich etwa in der Variation der Schwere und/oder Art der Beschwerden oder Störungen ausdrücken. Unter dieser Prämisse sind dann auch Veränderungen im Befinden oder Verhalten der Patienten nach der Therapie unterschiedlich zu bewerten. Das heißt, die Bewertung eines Behandlungsergebnisses als Erfolg oder Mißerfolg soll in Abhängigkeit von den individuellen Ausgangsbedingungen erfolgen. (S. 224)

In der traditionellen Vorgehensweise des Goal Attainment Scaling (GAS) von Kiresuk und Sherman (1968), vgl. auch Kiresuk, Smith und Cardillo (1994), werden zu Beginn der Therapie gemeinsam vom Therapeuten und Patienten individuelle Therapieziele formuliert und hinsichtlich ihrer Relevanz gewichtet. Weiterhin werden für jedes Ziel abgestufte konkrete Verhaltenserwartungen definiert. Nach der Behandlung schätzt der Patient ein, inwieweit er seine Ziele erreicht hat. Kordy und Scheibler (1984b) setzen die Methodik des GAS so ein, dass sie a priori bis zur fünften Therapiestunde gemeinsam mit dem Patienten mehrere Therapieziele definieren und in Anlehnung an die Vorgehensweise des GAS möglichst konkrete Verhaltens-/Erlebensalternativen definieren, die den Bewertungskriterien „verschlechtert", „unverändert", „gebessert", „gut gebessert" und „optimal gebessert" stehen. Den Bewertungskriterien werden Zahlen von Null (verschlechtert) bis Vier (optimal gebessert) zugeordnet und für alle definierten Therapieziele zu einem Gesamtindex aufsummiert, der – durch die maximal mögliche Punktzahl dividiert – eine Schätzung für die prozentuale Therapiezielerreichung ergibt.

Als Bestandteil des therapeutischen Prozesses liefert das GAS eine wertvolle Hilfestellung und wird daher von Therapeuten und Patienten in der Praxis gut akzeptiert. Auch wenn das GAS die individuellen Belange der Patienten berücksichtigt und versucht, die individuellen Therapieerfolge in Form von quantifizierbaren Bewertungskriterien abzubilden, so bleibt doch das Problem der fehlenden Standardisierung ungelöst. Kordy und Scheibler (1984b) führen selbst aus, dass aufgrund der Einmaligkeit der konstruierten Erfolgsmaße für jeden Patienten eine Schätzung von statistischen Gütemaßen ausgeschlossen ist und zudem der Arbeitsaufwand bei dieser Methode sehr hoch ist. Dies macht das GAS in der Anwendung für Forschung und Routinemonitoring unökonomisch.

Eine andere Vorgehensweise wurde mit dem Verfahren der „Zielorientierten Ergebnismessung" (ZOE) von Gerdes (1998) entwickelt. So führt das allgemein übliche Verfahren der Ergebnisevaluation, bei dem für jeden Patienten alle Outcome-Parameter in die Mes-

sung und Erfolgsbeurteilung der Rehabilitation einbezogen werden, nach Gerdes et al. (2000) zu folgendem Problem:

- Auf der Ebene des einzelnen Patienten enthält dessen individueller Summenscore eine Vielzahl von Outcome-Parametern, die für diesen Patienten nicht relevant sind, d.h. schon zu Beginn der Rehabilitation unauffällige Werte aufweisen.
- Auf der Ebene der Stichprobe enthält jeder Mittelwert für einen bestimmten Ergebnisaspekt eine Vielzahl von Patienten, für die der betreffende Outcome-Parameter nicht relevant ist somit schon zu Beginn der Rehabilitation unauffällige Werte aufweist.

Irrelevante, d.h. zu Beginn der Behandlung bereits unauffällig ausgeprägte Parameter lassen keine Verbesserung durch die Rehabilitation erwarten. Dies führt zu einer Nivellierung und somit Unterschätzung der Behandlungseffekte sowohl beim Summenwert für den einzelnen Patienten als auch beim Mittelwert für die Stichprobe. Gerdes et al. (2000) schlagen zur Lösung dieses Problems das Verfahren der „Zielorientierten Ergebnismessung" (ZOE) vor. Hierbei werden zu Beginn der Rehabilitation individuelle Therapieziele definiert, indem auffällige Skalen in einem entsprechenden Fragebogen zur Erfassung der Indikatoren des Reha-Status (IRES) zunächst computergestützt durch einen Vergleich mit Bezugsnormen gesunder Patienten markiert und sodann gemeinsam von Patient und Therapeut als behandlungsrelevante Bereiche ausgewählt werden (Gerdes, Jäckel & Fliedner, 1991; Zwingmann, 2003). In die Erfolgsbeurteilung und Berechnung von PRÄ-POST-Effekten werden dann nur noch diejenigen Parameter einbezogen, die zuvor als Therapieziel ausgewählt wurden. Es hat sich in der Praxis gezeigt, dass mit der ZOE-Methode deutlich höhere Effektgrößen resultieren als mit dem herkömmlichen Ansatz der allgemeinen indirekten Veränderungsmessung (Gerdes, 1998; Gerdes et al., 2000).

Tabelle 6. Veranschaulichung der zielorientierten Ergebnismessung (ZOE)

	Herkömmlicher Ansatz			**Zielorientierte Ergebnismessung (ZOE)**			
Skala	**PRÄ**	**POST**	**Differenz**	**Relevantes Reha-Ziel?**	**PRÄ**	**POST**	**Differenz**
001	64	52	-12	ja	64	52	-12
002	57	55	-2	nein			
003	72	59	-13	ja	72	59	-13
004	65	60	-5	ja	65	60	-5
005	52	54	+2	nein			
Mittelwert	**62**	**56**	**-6**		**67**	**57**	**-10**
Effektgröße			**0,60**				**1,00**

Anmerkung. Hypothetische T-Werte für einen einzelnen Patienten. T-Werte > 60 wurden in diesem Beispiel als „auffällig" und somit als „Relevantes Reha-Ziel" definiert. Die Effektgröße wurde unter Annahme einer Streuung der T-Werte von *SD* = 10 berechnet.

Tabelle 6 veranschaulicht das Prinzip der ZOE anhand eines hypothetischen Beispiels. Durch die Ausblendung der zu Beginn der Behandlung unauffälligen Skalen 002 und 005 resultiert bei der ZOE eine deutlich höhere Effektgröße gegenüber der herkömmlichen indirekten Veränderungsmessung.

2.2.2 Kritische Betrachtung der zielorientierten Ergenismessung

Vorteilhaft bei Verwendung des ZOE-Ansatzes ist die Tatsache, dass die individuelle Ausgangslage der Patienten berücksichtigt wird. Insbesondere bei Patienten, die mit einem eng umschriebenen Behandlungsanliegen in die Rehabilitation kommen und somit nur auf wenigen der erhobenen Skalen eine Beeinträchtigung zum Zeitpunkt der PRÄ-Messung aufweisen, resultiert eine optimistischere Bewertung der Therapieergebnisse als bei Anwendung des herkömmlichen allgemeinen Ansatzes.

Durch die vorgegebene Auswahl von Zielparametern und stärker standardisierte Vorgehensweise ist der Aufwand im Vergleich zum GAS (Kiresuk & Sherman, 1968; Kiresuk et al., 1994) weniger groß. Die mit dem ZOE-Ansatz berechneten PRÄ-POST-Veränderungen lassen sich darüber hinaus in Form der allgemein gebräuchlichen standardisierten Effektgrößenmetrik darstellen. Damit stellt die ZOE einen Kompromiss zwischen der allgemeinen und individuumsorientierten Erfolgsbewertung (vgl. Kordy & Scheibler, 1984a) dar.

Problematisch bei Verwendung des ZOE-Ansatzes (zur ausführlichen Kritik an der ZOE vgl. Zwingmann, 2003) ist die Praxis, bestimmte Skalen angesichts bestimmter Ausgangswerte als „irrelevant" zu klassifizieren und von der weiteren Analyse auszuschließen. Gesetzlicher Auftrag der Rehabilitation ist neben einer Verbesserung des Befindens auch die Beibehaltung von erwünschten Zuständen und Fähigkeiten des Erlebens, Verhaltens und des sozialen Eingebundenseins (vgl. Abschnitt 1.1.2). Dieser wichtige präventive Teilaspekt der Rehabilitation wird beim ZOE-Ansatz durch den Ausschluss der betreffenden (zu Beginn der Behandlung „unauffälligen") Skalen komplett ausgeblendet. Auch die Möglichkeit, dass sich ein Patient auf einer nicht markierten Skala verschlechtern kann, wird aufgrund dieser Vorgehensweise von der ZOE nicht berücksichtigt.

Der mehr oder weniger willkürliche Ausschluss von Skalen aus der Analyse führt zu einem erheblichen Informationsverlust, was im Hinblick auf die Validität der mittels ZOE gewonnenen Veränderungsmaße problematisch ist. So resultieren bei Anwendung der ZOE stark reduzierte Stichprobengrößen, was zu einer entsprechenden Verringerung der Teststärke führt und immer die Gefahr von systematischen Selektionseffekten beinhaltet.

Da bei jedem Patienten unterschiedliche Skalen in unterschiedlicher Anzahl in das aggregierte Gesamtmaß einfließen können, ist eine Berechnung von statistischen Kennwerten wie Reliabilitätsschätzungen für die Gesamtstichprobe entweder überhaupt nicht möglich, oder aber wie im Fall von aggregierten Mittelwerten mit Unwägbarkeiten behaftet. Eine inhaltliche Vergleichbarkeit der gewonnenen Maße zwischen verschiedenen Patienten bzw. Stichproben ist kaum möglich.

Durch Anwendung des ZOE-Ansatzes wird eine Selektion von hochbelasteten Patienten auf der betreffenden Skala vorgenommen und damit künstlich eine Extremgruppe er-

zeugt. Dies führt zu entsprechenden Deckeneffekten der PRÄ-Werte und zu einer stark eingeschränkten PRÄ-Streuung. Aufgrund der hierdurch verstärkt zu erwartenden Regression zur Mitte (Petermann, 1978) kann davon ausgegangen werden, dass mit dem ZOE-Ansatz die Behandlungseffekte überschätzt werden. Würde man die PRÄ-POST-Effektgrößen anhand der eingeschränkten PRÄ-Streuung berechnen, so würde daraus eine weitere künstliche Aufblähung der Effektgrößen resultieren. Immerhin versuchen Gerdes et al. (2000) diesem Problem zu begegnen, indem sie anstelle der PRÄ-Streuung die Streuung der Differenzwerte zur Standardisierung verwenden. Diese Vorgehensweise ist allerdings mit anderen methodischen Unwägbarkeiten behaftet (vgl. hierzu die Ausführungen in Abschnitt 2.1.3). Problematisch ist dabei insbesondere die Einbeziehung der Korrelation zwischen PRÄ- und POST-Messung in die Standardisierung der PRÄ-POST-Differenzen, weshalb die damit berechneten Effektgrößen nicht mit den Effektgrößen im Sinne von Smith und Glass (1977) oder Cohen (1992) äquivalent sind.

Als Fazit lässt sich feststellen, dass die bei der ZOE im Gegensatz zur herkömmlichen allgemeinen indirekten Veränderungsmessung realisierte Berücksichtigung des Ausgangszustandes durch einen hohen Informationsverlust erkauft wird. Durch die Nichtberücksichtigung von „irrelevanten" Skalen werden gerade diejenigen wünschenswerten Zustände des Befindens, Erlebens und Verhaltens aus der Analyse ausgeschlossen, deren Erhalt neben einer Verbesserung von problematischen Aspekten ebenfalls Aufgabe der medizinischen Rehabilitation ist. Durch diese einseitige Betonung von dysfunktionalen Aspekten erfolgt keine Berücksichtigung der vorhandenen Ressourcen und protektiven Faktoren des Patienten, denen in der Literatur eine wachsende Bedeutung für die Gesundheit beigemessen wird (Bengel, Strittmatter & Willmann, 1998; Antonovsky & Franke 1997). Darüber hinaus bleibt unberücksichtigt, dass sich Patienten auf Skalen, die zu Beginn der Reha unauffällig sind, auch verschlechtern können. Auf der anderen Seite ist positiv hervorzuheben, dass durch die prospektive und zielorientierte Vorgehensweise nur solche Bewertungsaspekte in die Evaluation eingehen, deren Verbesserung explizit zu Beginn der Therapie indendiert wurde, was durchaus im Sinne einer fairen Evaluation ist.

2.2.3 Kosten-Nutzen-Analysen

Kosten-Nutzen-Überlegungen spielen in Zeiten eines hohen Kostendrucks im Sozialwesen bei der Erfolgsbewertung eine wichtige gesundheitspolitische Rolle. Methodisch existieren hier eine Reihe von Zugängen (Schöffski & Graf von der Schulenburg, 2000), die mit verschiedenen Vor- und Nachteilen verbunden und an bestimmte Voraussetzungen geknüpft sind. Wie lässt sich die Effizienz der stationären psychosomatischen Rehabilitation ermitteln? Zunächst lassen sich direkte und indirekte Behandlungskosten direkt in Beziehung zu monetär bewertbaren Größen wie Verbleib im Erwerbsleben, Reduktion von Krankschreibungszeiten oder Inanspruchnahme von Leistungen des Gesundheitssystems setzen. So gelangen Zielke (1993) und Zielke et al. (2004) zu einem Return of Investment in einer Größenordnung von mehr als 3:1, d.h. jeder in die Rehabilitation investierte Euro bringt einen Ertrag von drei Euro durch eine Verringerung von Sozialausgaben. Eine aus-

schließliche Reduktion auf kostenrelevante Aspekte bei der Ergebnisbewertung erscheint angesichts der Komplexität der psychosomatischen Rehabilitation allerdings unangemessen und würde einen Großteil von relevanten Ergebnisaspekten außer Acht lassen. Darüber hinaus ist die monetäre Bewertung einer Reihe von Merkmalen wie z.B. Medikamentenkonsum methodisch problematisch (Zielke et al., 2004). Von Wittmann et al. (2002) wird daher ein alternativer Zugang vorgeschlagen, der auf die Effektgrößenmetrik zurückgreift. So lässt sich der Nettonutzen *U* einer Intervention nach Schmidt, Hunter und Pearlman (1982) anhand Gleichung 8 schätzen, wobei mit *N* die Patientenanzahl, mit *T* das Anhalten des Effektes in Jahren, mit *d* die PRÄ-POST-Effektgröße, mit SD_{prod} die Standardabweichung der Produktivität in Euro und *K* die (direkten und indirekten) Gesamtkosten der Behandlung in Euro einzusetzen sind. SD_{prod} entspricht nach Schmidt et al. (1982) etwa 40-65% der Jahresproduktivität. Leider wurde dieser Ansatz bei gesundheitsökonomischen Evaluationen trotz seiner Einfachheit bislang kaum in der Literatur aufgegriffen, weshalb Vergleichszahlen mit konventionellen Berechnungen noch weitgehend fehlen. Aus diesem Grund wird SD_{prod} konservativ mit 40% der Jahresproduktivität angesetzt.

$$U = N \cdot T \cdot d \cdot SD_{prod} - N \cdot K \tag{8}$$

Nachfolgend wird das Konzept anhand einer Modellrechnung expliziert. Das Statistische Bundesamt (2006c) geht für das Jahr 2005 von einer Bruttowertschöpfung der Wirtschaft in Höhe von 2.022,5 Milliarden Euro aus. Umgerechnet auf 38,8 Millionen Erwerbstätige im gleichen Jahr entspricht dies einem Betrag von durchschnittlich 52.126 Euro pro Jahr und Patient. 40% davon (konservative Schätzung gemäß der Annahme, dass psychosomatische Rehabilitanden häufiger durch Arbeitslosigkeit, soziale Probleme und Armut belastet sind als der Bevölkerungsdurchschnitt) entsprechen SD_{prod} = 20.850 Euro. Geht man davon aus, dass bei einem Patienten (*N* = 1) ein mittelgroßer Behandlungseffekt (*d* = 0,50) erzielt wurde und dieser mindestens zwei Jahre lang (*T* = 2) anhält, so errechnet sich ein Bruttonutzen von 1 x 2 x 0,50 x 20.850 Euro = 20.850 Euro.

Um den Nettonutzen *U* zu ermitteln, sind vom Bruttonutzen noch die Kosten der Rehabilitationsmaßnahme abzuziehen. Geht man bei einer durchschnittlichen Behandlungsdauer von derzeit 40 Tagen (Statistisches Bundesamt, 2006b, vgl. Abschnitt 1.1.6) in der psychosomatischen Rehabilitation von einem Tagespflegesatz in Höhe von 154 Euro aus, so errechnen sich direkte Behandlungskosten in Höhe von 6.160 Euro. Die indirekten Behandlungskosten ergeben sich aus dem Produktivitätsausfall für die Zeit der stationären Behandlung mit (40 Tage / 365 Tage) x 52.126 Euro = 5.712 Euro. Die Gesamtkosten betragen damit *K* = 4.400 + 5.712 = 11.872 Euro. Es errechnet sich ein Nettonutzen von *U* = 20.850 Euro - 11.872 Euro = 8.978 Euro pro Patient. Dies entspricht einem Return on Investment (*ROI*) in Höhe von 20.850 : 11.872 = 1,76 : 1.

Durch einfaches Umstellen von Gleichung 8 lässt sich darüber hinaus die zu erreichende Mindesteffektgröße am „Break-Even-Point" ermitteln, bei der sich Kosten und Nutzen

gegenseitig aufwiegen (U = 0) und sich die Investition in die Rehabilitationsmaßnahme „zu lohnen" beginnt (Gleichung 9).

$$d = \frac{K}{T \cdot SD_{prod}} \tag{9}$$

Geht man davon aus, dass der Behandlungseffekt mindestens zwei Jahre lang (T = 2) anhält und setzt man die Behandlungskosten mit K = 11.872 Euro an, so resultiert nach Gleichung 9 eine Effektgröße am Break-Even-Point von d = 11.872 / (2 x 20.850) = 0,28. Dies bedeutet, dass sich bereits ein kleiner Effekt in der Terminologie von Cohen (1992) unter Kosten-Nutzen-Gesichtspunkten lohnt. Derartige Modellrechnungen eignen sich somit gut, um im Vorfeld einer Evaluation eventuell vorhandenen Widerständen und Ängsten bei den Auftraggebern hinsichtlich der Effizienz der von ihnen angebotenen Behandlungsform zu begegnen (Wittmann et al., 2002).

Für die MESTA-Studie (Steffanowski et al., 2007, vgl. Abschnitt 1.3.5) wurde ausgehend von einer Effektgröße von 0,57 zum Entlass-Zeitpunkt und von 0,49 zum Zeitpunkt der 1-Jahres-Katamnese ein *ROI* von 4 : 1 für einen 7-Jahres-Zeitraum errechnet, wobei von einem weiteren linearen Nachlassen des Effektes um 0,08 Punkte pro Jahr ausgegangen wurde. Dies bedeutet, dass jeder in die Rehabilitation investierte Euro auf lange Sicht einen Bruttogewinn von vier Euro (bzw. Nettogewinn von drei Euro) erbringt, da die Patienten seltener krank geschrieben oder vorzeitig berentet werden, das Gesundheitssystem seltener in Anspruch nehmen und insgesamt produktiver sind – von der Verbesserung der Gesundheit und subjektiven Lebensqualität einmal abgesehen, die sich nicht in Geld aufwiegen lässt.

2.2.4 Erfolg und Mißerfolg in der Therapie

Im Versorgungsalltag wird dem Forscher von Praktikern und Entscheidungsträgern häufig die Frage entgegengebracht, wie hoch der Anteil gebesserter Patienten in Prozent ausgedrückt ist bzw. wie hoch die Erfolgsaussichten einer Intervention beim Vorliegen von bestimmten Voraussetzungen sind. Zur Beantwortung dieser Frage existieren verschiedene Bewertungs- und Entscheidungshilfen.

Betrachtet man einfach das Vorzeichen der berechneten Differenzwerte bei einem Vergleich zwischen PRÄ- und POST-Messung, so lässt sich zwar ermitteln, wie viele Patienten ein verbessertes Befinden aufweisen, allerdings erlaubt dies noch keine Aussage darüber, ob die beobachtete Veränderung bei einem Patienten ausreichend groß genug ist, um wirklich von einer klinisch relevanten Verbesserung oder Verschlechterung sprechen zu können. Gleiches gilt bei isolierter Betrachtung von statistischer Signifikanz (z.B. t-Test für Messwiederholungen) zur Erfolgsbewertung bei einer Patientengruppe, da bei entsprechend großer Stichprobe auch sehr kleine und damit praktisch kaum relevante Effekte statistisch signifikant werden.

Es stellt sich daher die Frage, welche Bewertungs- und Entscheidungshilfen zur klassifikatorischen Beurteilung von Behandlungsergebnisse verfügbar sind. So lässt sich zur Bewertung des PRÄ- und POST-Status unabhängig vom beobachteten Ausmaß der Veränderung ein Grenzwert bestimmen, bei dessen Überschreitung ein Patient als gesund bzw. krank klassifiziert wird. So werden bei standardisierten Testverfahren ausgehend von praktischen Erfahrungen oder statistischen Erwägungen mitunter entsprechende Cut-Off-Werte angegeben. Bei der Symptom-Checkliste SCL-90-R (Franke, 2002) wird für Screening-Zwecke beispielsweise ein T-Norm-Wert von 62,8 als Grenzwert zur Klassifikation eines Patienten als psychisch auffällig empfohlen, was dem 90. Perzentil der bevölkerungsrepräsentativen Normierungsverteilung entspricht und beim Beck-Depressions-Inventar (Hautzinger et al., 1994) wird 18 Punkten von einer klinisch relevanten depressiven Symptomatik ausgegangen.

Ein anderer Ansatz basiert darauf, dass Referenzwerte für Stichproben von gesunden und kranken Probanden vorhanden sind (Jacobson & Truax, 1991). Der Cut-Off-Wert *c* befindet sich dann an der Stelle, wo die Wahrscheinlichkeit gleich groß ist, zur funktionalen bzw. dysfunktionalen Population zu gehören (Abbildung 11) und bestimmt sich nach Gleichung 10. Hierbei stehen M_1 und SD_1 für Mittelwert und Streuung der dysfunktionalen Gruppe, M_2 und SD_2 hingegen für Mittelwert und Streuung der gesunden Gruppe. Voraussetzung ist, dass der betreffende Test in der Lage ist, zwischen Gesunden und Kranken zu differenzieren und die beiden Stichprobenstreuungen der Größe nach einigermaßen vergleichbar sind (Normalverteilungsvoraussetzung). Auch wenn Schmitz und Davies-Osterkamp (1997) kritisieren, dass dies nicht immer gegeben ist, so ermöglicht der Cut-Off-Wert *c* unter Sensitivitäts- und Spezifitätsgesichtspunkten dennoch am ehesten eine faire Klassifikation der Patienten.

$$c = \frac{s_1 \cdot M_2 + s_2 \cdot M_1}{s_1 + s_2} \tag{10}$$

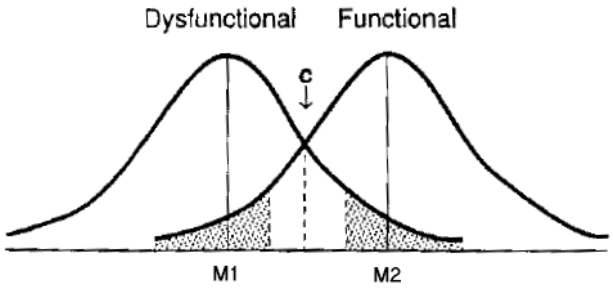


Abbildung 11. Cut-Off-Wert nach Jacobson & Truax (1991) zur Klassifikation eines Probanden als gesund bzw. krank.

Zur Bewertung von PRÄ-POST-Veränderungen existiert ebenfalls ein statistischer Ansatz. Dieser baut auf dem Konzept der Reliabilität auf und ist als Reliable Change Index (*RC*)

bekannt. So kann anhand der Skalenstreuung und Reliabilität eine Mindestdifferenz bestimmt werden, welche erreicht werden muss, damit die beobachtete Veränderung bei einem einzelnen Patienten mit 95%-tiger Wahrscheinlichkeit nicht mehr durch Messfehler erklärbar ist (Amelang et al., 1997). Die Bestimmung der „kritischen Differenz" D_{krit} geschieht nach Gleichung 11, wobei *SD* für die Skalenstreuung, r_{tt} für die Reliabilität (z.B. Cronbachs α) und z_α für den kritischen Wert zur Absicherung einer bestimmten Irrtumswahrscheinlichkeit (für das 5%-Niveau liegt dieser bei 1,96). Patienten mit PRÄ-POST-Differenz-Beträgen innerhalb des Vertrauensintervalls werden als „unverändert" klassifiziert. Patienten mit größeren PRÄ-POST-Differenz-Beträgen gelten als „gebessert" bzw. „verschlechtert".

$$D_{krit} = \pm z_\alpha \cdot SD \cdot \sqrt{2 \cdot (1 - r_{tt})} \qquad (11)$$

Das Verfahren weist durch die mathematische Abhängigkeit von der Reliabilität allerdings den Nachteil auf, dass die Klassifikation keine Aussage über das tatsächliche quantitative Ausmaß einer Veränderung darstellt. Nimmt man in Gleichung 11 für die Reliabilität einen Wert von .90 an, was allgemein bereits als relativ hoch angesehen wird und setzt man die PRÄ-Streuung gleich 1,00, so beträgt die kritische Differenz bei einer Irrtumswahrscheinlichkeit von 5% (z_α = 1.96) bereits 0,88 Standardabweichungseinheiten. Dies bedeutet, dass lediglich Patienten mit einem großen Effekt (Cohen, 1992) als „gebessert" klassifiziert werden. Bei einer Reliabilität von .80 beträgt D_{krit} bereits 1,24 Standardabweichungseinheiten. Die Verwendung des Kriteriums der kritischen Differenzen zur Unterscheidung zwischen „gebesserten" und „nicht gebesserten" Rehabilitanden stellt somit extrem hohe Anforderungen an die Reliabilität bzw. Homogenität der jeweils verwendeten Skala, was unter dem Aspekt einer fairen Bewertung bei Beachtung von Sensitivitäts- und Spezifitätsgesichtspunkten als fragwürdig einzustufen ist. Aufgrund der rechnerischen Abhängigkeit der kritischen Differenzen von der Skalenreliabilität ist das Verfahren zur Identifikation erfolgreicher Behandlungsverläufe wenig geeignet, da keine standardisierte Entscheidungsregel ausgehend vom tatsächlichen quantitativen Ausmaß der Veränderung gegeben ist.

Problematisch an jeder klassifikatorischen Erfolgsbewertung im Sinne von „gebessert" versus „nicht gebessert" ist grundsätzlich der damit einhergehende Informationsverlust, da eine kontinuierliche Variable künstlich dichotomisiert wird. Hier gerät der methodische Anspruch, eine möglichst hohe Informationsausschöpfung zu erreichen, in Widerspruch zum praktischen Anspruch, klare Beurteilungs-, Zuweisungs- und Selektionskriterien festlegen zu müssen.

Eine Alternative zur Festlegung der zu erreichenden Mindestdifferenz, um einen Patient als „gebessert" bzw. „verschlechtert" zu klassifizieren, bietet die Effektgrößenmetrik (vgl. Abschnitt 2.1.3). Ausgehend von praktischen Überlegungen oder einer in der Planungsphase durchgeführten Kosten-Nutzen-Analyse zur Bestimmung einer zur Erreichung des Break-Even-Points (vgl. Abschnitt 2.2.3) notwendigen Mindesteffektgröße lässt sich ein entsprechender Grenzwert definieren, bei dessen Überschreiten ein Patient als „gebes-

sert" bzw. „nicht gebessert" klassifiziert wird. Im Rahmen der EQUA-Studie (Schmidt et al., 2003) wurde im Hinblick auf die Ergebnisse von Lipsey (1993) pragmatisch ein Mindestwert von 0,50 (mittelgroßer Effekt) angesetzt.

Für die Bewertung eines einzelnen Patienten können individuelle Effektgrößen (*IES*, vgl. Grawe & Braun, 1994) berechnet werden. Hierbei wird die individuelle PRÄ-POST-Differenz des Patienten standardisiert. Die Verwendung der Effektgrößenmetrik als Entscheidungshilfe zur Erfolgsbewertung setzt dabei die Einigung auf ein adäquates Streuungsmaß zur Standardisierung der Mittelwertsdifferenzen voraus (vgl. Abschnitt 2.1.3). Unter leichter Abwandlung von Gleichung 1 kann *IES* nach Gleichung 12 berechnet werden, wenn man für $X_{PRÄ}$ und X_{POST} den individuellen PRÄ- und POST-Wert eines Patienten einsetzt. Die individuellen Effektgrößen können in Anlehnung an die Terminologie von Cohen (1992) interpretiert werden.

$$IES = \frac{X_{PRÄ} - X_{POST}}{SD_{PRÄ}} \tag{12}$$

2.2.5 Kombination von Status- und Veränderungsinformationen

Um dem Ausgangswertproblem bei der Verwendung von Veränderungsinformationen zur Erfolgsbewertung zu begegnen, schlägt Kordy (1997) vor, Status- und Veränderungsinformationen miteinander zu kombinieren und durch Festlegung von Cut-Off-Werten klinisch bedeutsame Evaluationskriterien zu definieren.

Die Vorgehensweise soll anhand der Symptomcheckliste SCL-90-R (Franke, 2002) veranschaulicht werden. Verwendung findet hierzu die Skala „Depressivität" aus Studie C und D (vgl. Abschnitt 3.1.3 und 3.2.6). Die PRÄ-Messung der 664 Patienten umfassenden Stichprobe (A-, E- und K-Messung vorhanden) ergab einen Mittelwert von 1,50 (*SD* = 0,87). Der Mittelwert der POST-Messung bei Entlassung aus der Klinik beträgt 0,88 (*SD* = 0,76). Berechnet man die Effektgröße nach Gleichung 1 in Abschnitt 2.1.3, so resultiert ein Wert von 0,71. Diese PRÄ-POST-Verbesserung der depressiven Symptomatik der Patientenstichprobe ist statistisch signifikant (t = 20,66; df = 663; p<.001). Die Skalenreliabilität in der Studienstichprobe beträgt .90 (Cronbachs α, Aufnahme-Messung). Abbildung 12 enthält den Scatterplot der Aufnahme- und Entlassmessung.

Ausgehend von den Kennwerten der A-Messung (M_1 = 1,50; SD_1 = 0,87) und den Kennwerten der bevölkerungsrepräsentativen Normierungsstichprobe aus dem Testmanual (M_2 = 0,44; SD_2 = 0,51; Franke, 1995) lässt sich der Cut-Off-Wert c zur Unterscheidung zwischen „gesunden" und „kranken" Probanden anhand Gleichung 10 berechnen. Es resultiert ein Wert von 0,83, der als entsprechende horizontale und vertikale Linie in Abbildung 12 eingezeichnet ist. Patienten mit Werten ab 0,83 werden als „krank" hinsichtlich der Ausprägung depressiver Symptome klassifiziert. Dies trifft zu Beginn der Behandlung auf 72,9% aller Patienten zu, am Ende der Behandlung hingegen nur noch auf 43,1%.

Patienten, die auf der mittleren Diagonalen liegen, haben zu Beginn und am Ende der Behandlung jeweils exakt den gleichen Score auf der Skala „Depressivität" in der SCL-90-R erzielt und sich damit nicht verändert. Die obere und untere Diagonale markiert Verschlechterungen bzw. Verbesserungen ab einer Größenordnung, die dem 95%-Konfidenzintervall des Standardmessfehlers entspricht. Diese kritische Differenz berechnet sich nach Gleichung 11. Die Reliabilität (Cronbachs Alpha) der SCL-90-R-Skala „Depressivität" der hier verwendeten Teilstichprobe der EQUA-Studie (*N* = 487) beträgt .90. Wenn man neben der Reliabilität die Skalenstreuung von 0,87 sowie den mit dem 95%-Konfidenzintervall korrespondierenden z-Wert von 1,96 in Gleichung 11 einsetzt, so resultiert eine zu erreichende kritische Mindestdifferenz bei 0,76, um von einer individuell statistisch signifikanten Veränderung sprechen zu können. Bei Verwendung dieses Kriteriums zur Beurteilung der PRÄ-POST-Veränderungen werden 41,4% aller Patienten als gebessert klassifiziert, 56,3% als unverändert und 2,3% als verschlechtert.

Kombiniert man diese Status- und Veränderungsinformationen zu einem evaluativen Gesamtbild der Entwicklung der depressiven Symptomatik der Patientenstichprobe, so gelangt man zu folgendem Ergebnis: Unten links in der Grafik finden sich Patienten, die sowohl zu Beginn als auch zum Ende der Behandlung keine auffälligen Werte aufweisen. Dies entspricht nach Kordy (1987) einer klinisch irrelevanten Veränderung (23,0% aller Patienten). Unten rechts, unterhalb der horizontalen Linie außerhalb der Diagonale rechts unten finden sich Patienten, die zu Beginn der Behandlung auffällig waren, es nach der Behandlung nicht mehr sind und sich darüber hinaus um mindestens 0,76 Punkte verbessert haben. Dies entspricht einer klinisch relevanten Verbesserung und trifft hier auf 27,0% zu. Analog werden klinisch relevant verschlechterte Patienten oben links in der Grafik dargestellt (lediglich 0,9%). Darüber hinaus finden sich bei 14,0% zwar signifikante Verbesserungen von mindestens 0,76 Punkten, die aber hinsichtlich des Bewertungskriteriums *c* nicht ausreichen, um von einer vollständigen Remission der Symptomatik sprechen zu können. Dies bezeichnet Kordy (1987) als klinisch nicht ausreichende Verbesserung.

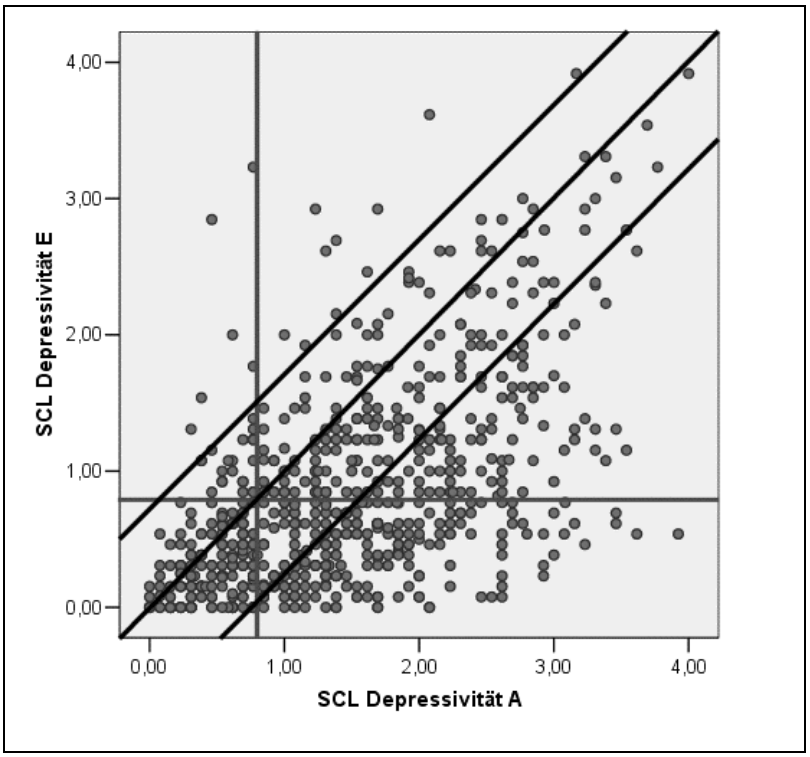


Abbildung 12. Definition klinisch relevanter Evaluationskriterien mit Cut-Off-Werten. EQUA**-Studienstichprobe, Vergleich zwischen Aufnahme- und Entlass-Zeitpunkt, SCL-90-R, Skala „Depressivität". *N* = 664.**

Positiv ist bei diesem Ansatz der Grad an Differenziertheit zu bewerten, mit dem die Status- und Veränderungsinformationen hier kombiniert werden. Allerdings hängt die Häufigkeitsverteilung stark von den jeweils verwendeten Cut-Off-Kriterien ab. Bereits kleine Verschiebungen der Grenzwerte führen zu gravierenden Änderungen bei der Häufigkeitsverteilung in den Feldern von Abbildung 12, so dass der Einwand einer gewissen Willkürlichkeit vorgebracht werden kann. Setzt man im hier gewählten Beispiel den Grenzwert zur Unterscheidung zwischen „auffälligen" und „unauffälligen" Rehabilitanden zu niedrig an, so werden viele (tatsächlich gesunde) Patienten fälschlicherweise als krank klassifiziert und die tatsächlichen Behandlungserfolge werden somit unterschätzt. Setzt man den Grenzwert hingegen zu hoch an, so werden viele (tatsächlich kranke) Patienten fälschlicherweise als gesund klassifiziert, so dass ein zu positives Bild von der Ergebnisqualität der Behandlung vermittelt wird. Auch die Verwendung des Reliable Change Index zur Klassifikation von Patienten als „verändert" versus „unverändert" im Hinblick auf die Sensitivität und Spezifität problematisch, da sich die kritische Differenz aus der Streuung und Reliabilität, nicht aber aus der tatsächlich gemessenen Veränderung ableitet.

Unter Fairness-Gesichtspunkten könnte gefordert werden, dass der Cut-Off-Wert zur Beurteilung von Status- und Veränderungsinformationen so festgesetzt wird, dass beide Fehlerarten (falsch positive und falsch negative Zuordnungen) möglichst ausgeglichen sind und gleichzeitig die Anzahl der korrekten Zuordnungen (Sensitivität und Spezifität) maximiert wird. Der Cut-Off-Wert *c* stellt diesbezüglich eine sinnvolle Entscheidungshilfe zur Beurteilung von Status-Messungen dar. Für den Fall, dass keine Bezugsnormen zur Berechnung von *c*, dafür aber objektive Validierungskriterien verfügbar sind, bietet sich

mit der Receiver-Operator-Charakteristik (ROC) noch eine weitere Option zur Berechnung von „fairen" Cut-Off-Werten. So ermöglicht die ROC-Analyse eine Bestimmung der Diskriminationsleistung eines Tests in Bezug auf das gewählte Außenkriterium in Abhängigkeit von einem bestimmten Cut-Off-Wert. Hierbei wird unter Berücksichtigung von Sensitivitäts- und Spezifitätsaspekten der Anteil an Fehlklassifikationen minimiert (Zweig & Campbell, 1993; Rundel, 2001). Für die Klassifikation von Veränderungsinformationen kann hingegen auch auf die Effektgrößenmetrik (Abschnitt 2.2.4) zurückgegriffen werden.

2.3 Multiple Ergebniskriterien

Je nach Aggregationsniveau der zur Erfolgsbewertung erhobenen Variablen kann zwischen singulären und multiplen Ergebniskriterien unterschieden werden, wobei multiple Ergebniskriterien im Gegensatz zu isolierten Einzelaspekten eine breitere Datengrundlage für die Evaluation bereitstellen.

So sieht man sich bei der Bewertung der Ergebnisqualität angesichts der Komplexität der stationären psychosomatischen Rehabilitation mit einer Vielfalt von relevanten Kriterien konfrontiert. Wie lässt sich der Behandlungserfolg möglichst umfassend und aussagekräftig abbilden, ohne dabei im „Dschungel" der vielen mittlerweile vorliegenden Einzelstudien und Einzelbefunde den Überblick zu verlieren?

Eine methodische Antwort auf diese Problematik liefert die meta-analytische Methodik mit ihrer Effektgrößenmetrik, bei der alle in einer Studie berichteten Mittelwerts-Differenzen im Interesse der Vergleichbarkeit in einem ersten Schritt zunächst z-standardisiert und diese Effektgrößen sodann zu einem Studiengesamteffekt aggregiert werden. Auf diese Weise können inhaltlich bzw. von der Metrik her unterschiedliche Einzelaspekte zu einem Gesamtbild integriert werden.

Ein anderer Ansatz der standardisierten Ergebnisbewertung besteht darin, das Erreichen eines bestimmten Erfolgskriteriums mit dem Wert Eins, das Nichterreichen hingegen mit dem Wert Null zu codieren. Die auf diese Weise gewonnenen singulären Ergebniskriterien lassen sich wiederum zu einem multiplen Gesamtindex aufaggregieren. Die methodenkritische Untersuchung und Weiterentwicklung dieses von Schmidt et al. (1987) entwickelten Ansatzes der multiplen Ergebniskriterien ist Thema der vorliegenden Arbeit.

2.3.1 Singuläre und multiple Bewertungskriterien

Multiple Indexmaße erfreuen sich im Alltag aufgrund ihrer praktischen Anwendbarkeit als Bewertungs- und Entscheidungshilfe großer Beliebtheit – man denke einmal an den umgangssprachlich als „Warenkorbindex" bezeichneten Verbraucherpreisindex (Egner, 2003), das Bruttosozialprodukt, den ifo-Geschäftsklimaindex (ifo Institut für Wirtschaftsforschung e.V. an der Universität München, 2008) oder den Deutschen Aktienindex (DAX)

an der Börse. All diesen Maßen ist gemeinsam, dass inhaltlich heterogene singuläre Einzelbewertungen zu einem multiplen Gesamtindex zusammengefasst werden. Auch in Automobil- und Computerzeitschriften finden sich beim Vergleichstest unterschiedlicher Produkte regelmäßig multiple Gesamtwertungen, die sich aus Aspekten wie Benutzerfreundlichkeit, Funktionalität, Sicherheit, Design und Preis-Leistungs-Verhältnis zusammensetzen. Häufig wird auch eine mehr oder weniger beliebige Gewichtung bestimmter Aspekte bei Berechnung der Gesamtnote vorgenommen.

Auch wenn für die aus dem Alltag geläufigen multiplen Bewertungsmaße keine statistischen Gütekriterien zu deren Reliabilität und Validität mitgeteilt werden, lässt sich das Beispiel doch gut auf die Evaluation im klinischen Kontext übertragen. So entspricht die skizzierte Vorgehensweise einer anteiligen Berücksichtigung mehrerer Stakeholder-Interessen (Patient, Angehörige, Arbeitgeber, Kostenträger) oder Ebenen des Gesamtbefindens (körperlich, psychisch, sozial, funktional) bei Bildung eines multiplen Kriteriums der zusammenfassenden Bewertung der Ergebnisqualität.

Obwohl damit die Gefahr einer Übersimplifizierung verbunden ist, kann doch davon ausgegangen werden, dass sich zufällige Schwankungen bei den Einzelbewertungen gegenseitig aufheben und sich erst eine gleichsinnige Verbesserung oder Verschlechterung auf mehreren Einzelmerkmalen als Trend auf der Gesamtskala bemerkbar machen wird. Auf diese Weise lassen sich komplexe Zustände und Entwicklungen mit nur einem einzigen Globalmaß abbilden. Die Aggregation mehrerer zu einem bestimmten Konstrukt gehörigen Merkmale bringt zudem den Effekt einer Reliabilitätssteigerung mit sich und verringert das Risiko von Fehlinterpretationen bei isolierter Betrachtung einer Vielzahl von Einzelvariablen aufgrund von Alpha-Fehler-Inflation.

Unberührt vom methodischen Ansatz der Standardisierung bleiben dabei inhaltliche Fragen nach der Art der aggregierten Einzelbefunde bestehen. So zielt eine Hauptkritik an der meta-analytischen Methodik auf das Inkommensurabilitätsproblem ab, wonach es unzulässig sei, „Äpfel und Birnen" miteinander zu vermischen (Lösel, 1987). Diese Argumentation trifft sicherlich auch auf die hier skizzierte multiple Indexbildung zu. Auf der anderen Seite kann argumentiert werden, dass die Aggregation gerade dann indiziert ist, wenn man die Ergebnisqualität auf einem höheren Globalisierungsniveau („Obst", um bei der gewählten Analogie zu bleiben) abbilden möchte. Bei Auswahl der zu aggregierenden Einzelbefunde sind unter Fairnessgesichtspunkten in der Evaluation daher immer die jeweiligen Stakeholderinteressen a priori genau zu definieren und in Form entsprechender Bewertungskriterien angemessen zu berücksichtigen.

2.3.2 Entwicklung eines multiplen Ergebniskriteriums

Das Konzept der multiplen Ergebniskriterien wurde im Rahmen der „Zauberberg-Studie", einer breit angelegten Programmevaluationsstudie in der psychosomatischen Klinik Schömberg (Schmidt et al., 1987; Schmidt, 1991), erstmals erprobt und danach in mehreren weiteren Studien (vgl. Abschnitt 3.1.3) erfolgreich eingesetzt. Anlass für die Entwicklung der multiplen Ergebniskriterien war damals die Frage, wie sich die Vorhersag-

barkeit der Ergebnisse von psychotherapeutischen Behandlungen verbessern lässt (Luborsky, 1980; Kächele & Fiedler, 1985). Schmidt et al. (1987) führen den mangelnden Zusammenhang zwischen Prädiktoren und Ergebniskriterien u.a. darauf zurück, dass in der Ergebnisforschung häufig nur einzelne, isolierte Outcome-Aspekte (singuläre Ergebniskriterien) verwendet werden, die

> (...) in ihrer Repräsentativität so eng und begrenzt sind, dass sie nicht in der Lage sind abzubilden, ob ein Patient – in der Summe betrachtet – mehr oder weniger profitiert hat. (S. 294)

Im Bereich der psychosomatischen Rehabilitation werden heterogene Patientenstichproben mit einer breiten Palette von Problemen behandelt, so dass bei Beachtung des Symmetrieprinzips (Wittmann, 1990) die alleinige Verwendung von singulären Ergebniskriterien unangemessen ist. Schmidt et al. (1987) schlagen daher die Verwendung von multiplen Ergebniskriterien vor, die auf der Vorstellung basieren,

> (...) dass Behandlungsergebnisse viele, unterschiedliche Facetten haben und dass von Patient zu Patient auch unterschiedliche Facetten angesprochen werden können. Multiple Ergebniskriterien entsprechen deshalb der – einmaligen oder wiederholten – Erfassung mehrerer Outcome-Aspekte, die zu einem Index verknüpft werden, um dadurch eine umfassendere und zuverlässigere Abbildung der individuellen Ergebnisseite zu erreichen. (S. 294)

In der Praxis funktioniert dies so, dass Veränderungen mehrerer Einzelmerkmale wie z.B. Symptombelastung, Lebenszufriedenheit, Arbeitsfähigkeit, Medikamentenkonsum, Beziehungszufriedenheit oder Problemlösefähigkeit zu einem multiplen Gesamtindex der Ergebnisqualität zusammengefasst werden. Je höher der Punktwert auf der resultierenden Gesamtskala, desto größer ist der Behandlungserfolg. Dieser Ansatz entspricht auch einer multidimensionalen Betrachtungsweise mit Einbeziehung von somatischen, psychischen, sozialen und funktionalen Aspekten, wie dies in der modernen rehabilitationswissenschaftlichen Forschung favorisiert wird (Bengel & Koch, 2000; Delbrück & Haupt, 1998).

Inspiriert wurde die Idee der Multiplen Ergebniskriterien von Fishbein und Ajzen (Fishbein & Ajzen 1975; Ajzen & Fishbein, 1980), die mit ihrer „Theory of Reasoned Action" bei der Vorhersage von menschlichem Verhalten beachtliche Erfolge in verschiedenen Anwendungsbereichen erzielt haben. Insbesondere der von Fishbein und Ajzen verwendete innovative methodische Ansatz bei der Codierung und Aggregation von Einzelinformationen erwies sich bei der Entwicklung der Multiplen Ergebniskriterien als hilfreich.

Ajzen und Fishbein (1980) geben folgende Definition für singuläre Verhaltenskriterien:

> A single act is a specific behavior performed by an individual. To be able to measure a single action, we have to define it clearly enough so that we can determine whether or not it has been performed. (S. 31)

Der zweite für die Entwicklung der Multiplen Ergebniskriterien bedeutsame Aspekt in den Arbeiten von Fishbein und Ajzen ist die Hervorhebung der Wichtigkeit einer Aggregation von singulären zu multiplen Verhaltenskriterien:

> Regularities, patterns, or tendencies cannot be discerned in single instances of behavior. Rather, to obtain a measure of a behavioral tendency, we must aggregate observations made on different occasion. When we compute the average behavioral tendency over repeated occasions, the influence of factors that vary from one occasion to another tends to cancel out. (Ajzen, 1988, S. 46)

Tabelle 7 enthält eine Klassifikation verschiedener Verhaltenskriterien (Fishbein & Ajzen, 1975, adaptiert von Wittmann, 1985). So wird zum einen zwischen singulären und multiplen und zum anderen zwischen einmalig und wiederholt (= über mehrere Zeitpunkte und Situationen) gemessenen Verhaltenskriterien differenziert. Es resultieren vier mögliche Kriterienarten.

Tabelle 7. Taxonomie singulärer und multipler Verhaltenskriterien (Wittmann, 1985)

SVK	**Singuläres Verhaltenskriterium** Einmalige Beobachtung eines einzelnen Verhaltensaktes
BWK	**Beobachtungswiederholungs-Kriterium** Wiederholte Beobachtung eines einzelnen Verhaltensaktes
EMVK	**Einmaliges multiples Verhaltenskriterium** Einmalige Beobachtung multipler Verhaltensakte
WMVK	**Wiederholtes multiples Verhaltenskriterium** Wiederholte Beobachtung multipler Verhaltensakte

Ajzen und Fishbein (1980) verfolgen bei der Bewertung und Aggregation der Einzelinformationen einen ebenso einfachen wie pragmatischen Ansatz, indem sie zunächst jedes singuläre Verhaltenskriterium binär im Sinne von „vorhanden" bzw. „nicht vorhanden" codieren und alle auf diese Weise gebildeten Einzelitems dann zu einem multiplen Verhaltenskriterium aufsummieren.

Übertragen auf evaluative Fragestellungen bedeutet dies, dass man bereits auf der Ebene der Einzelitems zu einer möglichst klaren Bewertung gelangen sollte, ob der Patient auf dem entsprechenden Ergebniskriterium „erfolgreich" oder „nicht erfolgreich" ist.

Diese Grundidee wurde auf die „Zauberberg-Studie", einer breit angelegten Programmevaluationsstudie in der Klinik Schömberg (Schmidt, 1991) übertragen. Die Datenerhebungen im Rahmen des Projekts beinhalteten

> (...) eine Fülle singulärer Ergebniskriterien (12 Monate nach Entlassung), die allesamt den Nachteil haben, daß sie lediglich – in Form von Gruppenstatistiken – einen isolierten Ausschnitt aus dem Outcomespektrum vermitteln. Die Frage, wie einzelne Mosaiksteinchen in der Summe aussehen, sollte durch die Konstruktion eines explorativen multiplen Ergebniskriteriums beantwortet werden. (Schmidt et al., 1987, S. 295)

Die Konstruktion eines multiplen Ergebniskriteriums nach Schmidt et al. (1987) beinhaltet drei Schritte:

1. Auswahl und Messung der Einzelkomponenten gemäß der jeweiligen evaluativen Fragestellung, in der Regel auf mehrstufigen Ratingskalen

2. Codierung der Veränderungen mit Eins („gebessert") bzw. Null („nicht gebessert"), hierzu Bildung eines neuen Items für jede Einzelkomponente
3. Aufsummierung der neu gebildeten dichotomen Items zu einer Gesamtskala.

Ausschlaggebend für die Verwendung der dichotomen Codierung waren damals Überlegungen der Praktikabilität, wobei eine möglichst hohe Transparenz und anschauliche Vermittelbarkeit der Behandlungsergebnisse für statistisch nicht geschulte klinische Praktiker und politische Entscheidungsträger (z.B. Kostenträger) angestrebt wurde.

Tabelle 8 veranschaulicht das Prinzip der Codierung anhand des Items „Beurteilung des Gesundheitszustandes im Vergleich zur Zeit vor der Behandlung". Hier wurden die Patienten um eine direkte Einschätzung der von ihnen subjektiv wahrgenommenen Veränderungen (im Sinne von „besser – schlechter" ein Jahr nach der stationären Behandlung gebeten. Das Original-Item sieht fünf abgestufte Antwortalternativen vor. Positive Veränderungen werden mit dem Wert Eins, eine Nichtveränderung bzw. Verschlechterung des Gesundheitszustandes hingegen mit dem Wert Null codiert. Tabelle 9 enthält die entsprechende Transformationsvorschrift für das Item „Medikamentenkonsum". Hier handelt es sich um ein dreifach gestuftes Item.

Tabelle 8. Binäre Codierung singulärer Ergebniskriterien (Beispiel 1)

Veränderung des Gesundheitszustandes	Originalcodierung	Binärcodierung
deutlich verbessert	1	1
etwas verbessert	2	
unverändert	3	0
etwas verschlechtert	4	
deutlich verschlechtert	5	

Tabelle 9. Binäre Codierung singulärer Ergebniskriterien (Beispiel 2)

Veränderung des Medikamentenkonsums	Originalcodierung	Binärcodierung
weniger Medikamente	1	1
etwa gleich viel	2	0
mehr Medikamente	3	

Diese beiden Beispiele sollen genügen, um die Vorgehensweise zu veranschaulichen. Generell ist es mit dem hier beschriebenen Ansatz also möglich, mit unterschiedlichen Skalierungen gewonnene Ergebnisinformationen nach deren Neucodierung mit Null bzw. Eins zu einer neuen Skala zusammenzufassen.

Das explorativ zunächst aus 17 (Schmidt et al., 1987; Schmidt, 1991) solchen singulären Outcome-Aspekten gebildete (einmalig zum Katamnese-Zeitpunkt gemessene multiple Ergebniskriterium) erhielt in Anlehnung an die in Tabelle 7 wiedergegebene Taxonomie

die Bezeichnung EMEK1 und wies eine hohe interne Konsistenz auf (Cronbachs α = .86, *N* = 223). Der Mittelwert von EMEK1 lag bei 11,2 (*SD* = 4,3), d.h. die Patienten in der Zauberbergstudie I gaben ein Jahr nach Entlassung aus der stationären psychosomatischen Rehabilitation im Durchschnitt bei 11 von 17 singulären Ergebniskriterien eine positive Bewertung bzw. Entwicklung ihres Erlebens und Verhaltens an.

Bei der Validierung von EMEK1 ergaben sich höhere Zusammenhänge mit einer Reihe von Prädiktorvariablen als bei der Verwendung von singulären Ergebnismaßen. Darüber hinaus erwiesen sich die mit der neu entwickelten Skala in der 1-Jahres-Katamnese erfassten Behandlungsergebnisse auch in der 3-Jahres-Katamnese als stabil (Schmidt, Lamprecht, Bernhardt & Nübling, 1989).

Aufgrund der positiven Erfahrungen in der Zauberberg-I-Studie wurde in den Folgejahren der Ansatz in fünf weiteren Programmevaluationsstudien (vgl. Abschnitt 3.1.3) zur stationären psychosomatischen Rehabilitation angewandt, wobei jeweils 27 Einzelaspekte zur Skala EMEK_27 aggregiert wurden. Die Item-Inhalte der Skala EMEK_27 sind in Tabelle 10 in Abschnitt 3.1.1 detailliert wiedergegeben.

2.3.3 Kritische Betrachtung der Skala EMEK_27

Die Einzelkomponenten von EMEK_27 werden gewonnen, indem auf Item-Ebene eine Verbesserung des Befindens mit dem Wert Eins, ein unverändertes bzw. verschlechtertes Befinden hingegen mit dem Wert Null in eine neue Variable codiert wird.

Vorteile:

- Inhaltlich heterogene Veränderungsinformationen können unabhängig von der jeweils verwendeten Skalierung zu einem Gesamtindex zusammengefasst werden. Dies entspricht der Forderung, eine möglichst breite Palette von bio-psycho-sozialen Ergebnisaspekten bei der Evaluation zu berücksichtigen.
- Der evaluative Prozess des Messens und Bewertens (Kordy & Scheibler, 1984a) wird bereits auf Einzel-Item-Ebene vollzogen und definiert für jeden singulären Ergebnisaspekt klare Bewertungskriterien im Sinne von „vorhanden" versus „nicht vorhanden". Dadurch wird eine präzise und gleichberechtigte Abbildung aller erfassten Stakeholder-Interessen in der EVA-Box im Datenmodell von Wittmann (1990) ermöglicht.
- Die Mittelwerte der 27 neu gebildeten dichotomen Items lassen sich direkt im Sinne von Prozentwerten interpretieren. Der Summenwert der Gesamtskala kann ähnlich interpretiert werden. Erreicht z.B. ein Patient 20 Punkte auf der Skala EMEK_27, so entspricht dies einer „Erfolgsquote" von 20 / 27 = 74,1%. Diese Darstellungsform erleichtert erheblich den Transfer der Ergebnisse in die Praxis. Damit ist dieser Ansatz für alle Anwendungsfelder ideal, wo es auf eine anschauliche Vermittlung der Ergebnisse ankommt, etwa bei der Erstellung von Routineauswer-

tungen, Präsentationen und Qualitätsberichten oder im Dialog mit klinischen Praktikern und politischen Entscheidungsträgern.

Nachteile:

- Durch jede Dichotomisierung geht Varianz und damit wertvolle Information verloren. Cohen (1983) beziffert das Ausmaß des Informationsverlustes und der damit einhergehenden Reduktion der statistischen Power wie folgt: Geht man von zwei bivariat normal verteilten kontinuierlichen Variablen und dem ursprünglich ermittelten Zusammenhang *r* aus und führt man die Dichotomisierung einer der beiden Variablen am Mittelwert durch, so beträgt die resultierende Korrelation der künstlich dichotomisierten mit der kontinuierlichen Variable .798, was einem Varianzverlust und damit Informationsverlust von $1 - .798^2 = 36{,}3\%$ entspricht. Cohen führt weiter aus, dass der Informationsverlust um so größer wird, je weiter die Dichotomisierung vom Item-Mittelwert entfernt durchgeführt wird (etwa wenn Patienten mit extremer Belastung mit allen übrigen Patienten kontrastiert werden). Dieser Informationsverlust wird vor allem dann zum Problem, wenn der Behandlungserfolg anhand bestimmter Prädiktoren vorhergesagt werden soll, etwa zu Zwecken der indikativen Zuweisung von Patienten zur Behandlung mit der größten Erfolgsaussicht. Eine wichtige Voraussetzung hierfür ist die möglichst umfassende Ausschöpfung der vorhandenen Varianz. Cohen argumentiert, dass angesichts der Verfügbarkeit von statistischen Verfahren zur Ausschöpfung der gesamten Information kein Grund dafür besteht, diese nicht auch zu nutzen.
- Ein anderes methodisches Problem des herkömmlichen EMEK-Ansatzes besteht darin, dass zwischen Patienten mit einem verschlechterten und solchen mit einem unveränderten Befinden nicht differenziert wird. Hierdurch entsteht eine Unschärfe bei der Ergebnisbewertung, die eigentlich nicht mit der ursprünglichen Grundidee vereinbar ist, bereits auf Itemebene zu einer möglichst klaren evaluativen Aussage zu gelangen, ob ein bestimmtes Resultat vorliegt oder nicht: Während eine Verschlechterung in jedem Fall ein unerwünschtes und damit negativ zu bewertendes Ergebnis darstellt, kann die Aussage „unverändert" auch bedeuten, dass ein erwünschter Zustand beibehalten wurde (z.B. Zufriedenheit mit der Partnerschaft). So besteht der gesetzliche Auftrag der Rehabilitation neben der Heilung und Besserung von Beschwerden auch darin, wünschenswerte Zustände aufrecht zu erhalten bzw. eine weitere Verschlechterung des Befindens dort zu verhindern, wo eine Heilung nicht mehr möglich ist.

Es stellt sich daher die Frage, wie sich die herkömmliche Skala EMEK_27 einerseits so modifizieren lässt, dass die gesamte verfügbare Information auf Itemebene ausgeschöpft wird, andererseits aber die Vorteile wie die Integrierbarkeit unterschiedlicher Itemformate, gleichberechtigte Abbildung von Stakeholder-Interessen sowie anschauliche Vermittelbarkeit in die Praxis erhalten bleiben.

3 Reanalyse der Skala EMEK_27

Es wird eine Reanalyse der Skala EMEK_27 anhand der vorliegenden Daten von fünf Programmevaluationsstudien durchgeführt. Hierbei werden zwei Skalenvarianten berechnet, wobei die eine Variante auf den bisherigen dichotomisierten Items, die andere Variante hingegen auf z-standardisierten Items beruht, welche die gesamte Information der Original-Items ausschöpfen.

Beide Skalenvarianten werden einer vergleichenden Analyse hinsichtlich der Itemkennwerte, der Verteilungseigenschaften, der dimensionalen Struktur und der Validität unterzogen. Hierbei wird der Frage nachgegangen, inwieweit durch die hier neu entwickelte Variante eine Verbesserung der bisherigen Variante möglich ist.

Die Datenanalysen zeigen, dass durchaus eine leichte Verbesserung der Informationsausschöpfung mit der neu entwickelten Variante von EMEK_27 möglich ist, wobei sich der durch die Dichotomisierung der Items bedingte Informationsverlust auf Skalenebene bei der herkömmlichen Skala EMEK_27 offenbar nicht so gravierend auswirkt wie zunächst angenommen.

3.1 Methodik zur Durchführung der Reanalyse

Insgesamt 27 dichotome singuläre Ergebniskriterien werden zu dem multiplen Ergebniskriterium EMEK_27 aufsummiert, das somit einen möglichen Wertebereich von 0 bis 27 Punkte umfasst. Trotz der methodischen und inhaltlichen Heterogenität der einbezogenen singulären Ergebnisaspekte hat die Skala in den bisherigen Studien eine hohe Reliabilität mit einem Cronbachs Alpha von über .90 und damit eine hohe Messgenauigkeit bewiesen. Ursprünglich wurde die Skala explorativ gebildet und dann in unveränderter Form eingesetzt. Mittlerweile stehen genug Daten zur Verfügung, um den Ansatz sowohl inhaltlich als auch methodisch weiter zu entwickeln. Aus diesem Grund wird im ersten Schritt eine Reanalyse der vorhandenen Daten durchgeführt, um einen Gesamteindruck von den Itemkennwerten, den Verteilungseigenschaften sowie der faktoriellen Struktur der herkömmlichen Skala EMEK_27 zu gewinnen.

Auf dem Hintergrund der kritisierten Dichotomisierung von kontinuierlichen Merkmalen erfolgt die Reanalyse der herkömmlichen Skala EMEK_27 im Vergleich zu einer hier alternativ entwickelten EMEK_27-Variante, bei der auf Ebene der Einzelitems jeweils die gesamte verfügbare Information ausgeschöpft wird. Hierzu werden die Daten von fünf Programmevaluationsstudien (vgl. Kapitel 3.1.3) zusammengeführt, die eine Ausgangsstichprobe von insgesamt 2.624 Patienten umfassen und bei denen sämtliche Einzelkomponenten der Skala EMEK_27 erhoben wurden.

3.1.1 Dichotomisierte und standardisierte Items

Für die Reanalyse werden die Daten aus fünf Programmevaluationsstudien mit einer Ausgangsstichprobe von insgesamt 2.624 Patienten verwendet.

Nach Behandlung der Missing-Data-Problematik bei den 27 singulären Veränderungsmaßen des Ausgangsmaterials werden die einzelnen EMEK-Komponenten für zwei Item-Varianten a und b in jeweils 27 neue Items codiert:

- Die nachfolgend als *Variante a* bezeichnete Transformation entspricht der klassischen dichotomen Codierung und führt zu binären Items mit zwei möglichen Ausprägungen 0 "nicht gebessert" und 1 "gebessert". Diese lassen sich direkt im Sinne von Erfolgsprozenten bzw. Patientenanteilen mit einem gebesserten Befinden interpretieren.
- Die nachfolgend als *Variante b* bezeichnete Transformation berücksichtigt die gesamte verfügbare Iteminformation und führt zu intervallskalierten Items. Im Gegensatz zur Variante a werden bei der Variante b durch Lineartransformation sämtliche „Zwischentöne" zugelassen. Darüber hinaus wird die Möglichkeit berücksichtigt, dass auch eine Verschlechterung des Befindens eintreten kann. Dies wird durch eine Standardisierung aller Einzelitems anhand der jeweiligen Itemstreuung erreicht. Der Nullwert für die Standardisierung wird dabei so gewählt, dass dieser der Kategorie „unverändert" entspricht. Verbesserungen des Befindens werden somit durch ein positives Vorzeichen, Verschlechterungen des Befindens hingegen durch ein negatives Vorzeichen ausgedrückt. Der zu erwartende Wertebereich der Items und der Gesamtskala dürfte gemäß z-Verteilung zwischen –3 und +3 liegen, ist durch die Abhängigkeit von der Itemstreuung theoretisch jedoch nach oben und unten hin offen.

Tabelle 10 gibt alle 27 Items des „Ausgangsmaterials" sowie deren Originalcodierung wieder. Mit Ausnahme von Item 1 sowie Item 16-18 handelt es sich durchweg um direkte Veränderungsinformationen. Dabei sind die Items 9-15 dreistufig, die Items 1-8 fünfstufig, die Items 19-27 siebenstufig und die Items 16-18 dreizehnstufig skaliert. Bei Item 1 handelt es sich um eine Statusmessung („Wie geht es Ihnen derzeit, also etwa ein Jahr nach Ende der Behandlung, insgesamt?") mit einem Wertebereich des Originalitems von 1 („gut") über 3 („weder noch") bis 5 („schlecht"). Für Variante b wurde als Nullpunkt für die Transformation daher ersatzweise der Wert 3 („weder noch") gewählt. Bei den Items 16, 17 und 18 handelt es sich um quasi-indirekte Veränderungsmessungen, d.h. es wurde zum Katamnese-Zeitpunkt neben dem aktuellen Status auch retrospektiv der damalige Status bei Aufnahme aus dem Gedächtnis auf siebenstufig skalierten Items erfragt. Die Differenz aus beiden Angaben bildet den Rohwert der Items 16, 17 und 18 und kann somit einen Wertebereich von –6 (maximale Abnahme von Arztbesuchen, AU-Zeiten und Krankenhauszeiten) über 0 (keine Veränderung) bis +6 (maximale Zunahme) umfassen.

Die mit Variante a überschriebene Spalte gibt die Transformationsvorschrift der Items für die herkömmliche Skala EMEK_27a wieder, so wie dies in den fünf einbezogenen Studien

durchgeführt wurde. Die Skalenvariante EMEK_27a wird durch Aufsummierung der 27 Einzelitems gebildet und kann somit einen Wertebereich zwischen 0 und 27 Punkten umfassen.

Tabelle 10. Singuläre Ergebniskriterien der Skala EMEK_27a und ihre Codierung

Nr	EMEK_27-Komponente	Originalcodierung	Variante a
1	Befinden zum POST-Zeitpunkt (Statusmessung)	1 gut 2 eher gut 3 weder noch 4 eher schlecht 5 schlecht	1 1 0 0 0
2 3 4 5 6 7 8	Lebensqualität Körperliches Befinden Seelisches Befinden Allgemeinbefinden Leistungsfähigkeit Beschwerden Gesundheitszustand	1 deutlich besser 2 etwas besser 3 unverändert 4 etwas schlechter 5 deutlich schlechter	1 1 0 0 0
9	Umgang mit Alltagsbelastungen	1 eher besser 2 unverändert 3 eher schlechter	1 0 0
10	Gesundheitsbewusste Lebensführung	1 lebe gesünder 2 keine Veränderung 3 lebe weniger gesund	1 0 0
11	Medikamentenkonsum	1 weniger Medikamente 2 etwa gleich viel 3 mehr Medikamente	1 0 0
12 13 14 15	Beziehungen Bezugspersonen Beziehung zum Partner Familienleben Arbeitsfähigkeit	1 verbessert 2 keine Veränderung 3 verschlechtert	1 0 0
16 17 18	Arztbesuche Krankschreibungszeiten (AU) Krankenhaustage quasi-indirekte Veränderungsmessung (qVM = POST-Messung minus Retrospektive PRÄ-Messung)	-6 maximal verbessert -5 (...) 0 unverändert (...) +5 +6 maximal verschlechtert	1 1 1 0 0 0 0
19 20 21 22 23 24 25 26 27	Wohlbefinden Umgang mit Problemen Fähigkeit zur Selbsthilfe Umgang mit Enttäuschungen Zurechtkommen mit Arbeit Belastbarkeit Auskommen Mitmenschen Leben mit Einschränkungen Ausgeglichenheit	1 sehr stark verbessert 2 deutlich verbessert 3 etwas verbessert 4 unverändert 5 etwas verschlechtert 6 deutlich verschlechtert 7 sehr stark verschlechtert	1 1 1 0 0 0 0

Für die Itemtransformation der Variante b muss zunächst eine angemessene Vorgehensweise gefunden werden, um die unterschiedlich skalierten Items direkt miteinander ver-

gleichbar zu machen. Hierzu bietet sich die z-Standardisierung an. Verbesserungen und Verschlechterungen des Befindens sollten in Anlehnung zu der im Rahmen von Meta-Analysen bei Mittelwertsvergleichen verwendete Effektgrößenmetrik dabei durch ein positives bzw. negatives Vorzeichen kenntlich gemacht werden. Da die Items des Ausgangsmaterials fast ausschließlich auf direkten Veränderungsmessungen basieren, sind weder PRÄ-POST-Differenzen noch Streuungen einer unbehandelten Patientenstichprobe zur Berechnung von Effektgrößen verfügbar. Aus diesem Grund stellt sich die Frage nach den adäquaten Bezugspunkten zur Standardisierung der Items. Geht man davon aus, dass sich das Befinden eines unbehandelten Patienten nicht verändert, so bietet sich bei direkten Veränderungsmaßen der entsprechende Erwartungswert *E* für ein unverändertes Befinden an, der vom gemessenen Item-Mittelwert M_{dVM} subtrahiert wird. Zur Standardisierung dieser Mittelwertsdifferenz wird die Item-Streuung SD_{dVM} der Stichprobe verwendet (Gleichung 13).

$$d_{dVM} = \frac{M_{dVM} - E}{SD_{dVM}} \tag{13}$$

Die Berechnung soll anhand des Items „Gesundheitszustand" exemplarisch veranschaulicht werden. Der Wertebereich des Items erstreckt sich von 1,00 „deutlich verbessert" über 3,00 „unverändert" bis 5,00 „deutlich verschlechtert". Der gemessene Item-Mittelwert beträgt 2,33 und die Item-Streuung beträgt 1,13. Setzt man diese Werte in Gleichung 13 ein und geht man von einem Erwartungswert von 3,00 („unverändert") aus, so beträgt die standardisierte Mittelwertsdifferenz -0,60 (Gleichung 14).

$$d_{dVM} = \frac{2{,}33 - 3{,}00}{1{,}33} = -0{,}60 \tag{14}$$

Codiert man die beobachtete Verbesserung des Befindens entsprechend der Konvention bei der Effektgrößenberechnung mit einem positiven Vorzeichen, so resultiert ein Wert von +0,60. Verfolgt man die Analogie zu den Effektgrößen weiter, so würde dieser Betrag nach Cohen (1992) zwischen einem mittleren (0,50) und großen (0,80) Effekt liegen. Die Gesamtskala EMEK_27 für Variante b wird gemäß dem ersten Aggregationsschritt bei einer meta-analytischen Studie (Lipsey & Wilson, 2001) mittels Aufsummierung der 27 neu gewonnenen standardisierten Einzelitems und anschließende Division durch 27 berechnet. Der resultierende Wert entspricht dem durchschnittlichen *Studieneffekt*, der alle in einer bestimmten Untersuchung berichteten Einzeleffekte mit gleicher Gewichtung berücksichtigt (bei einer Meta-Analyse würden die Studieneffekte mehrerer Untersuchungen dann im zweiten Aggregationsschritt zu einem gewichteten Gesamteffekt zusammengesetzt werden).

Die Standardisierung hat den Vorteil, dass nach Durchführung der Transformation alle Items eine einheitliche Streuung von 1,00 aufweisen. Bei Aggregation der Items gehen

somit sämtliche Einzelaspekte mit gleicher Gewichtung in die Gesamtskala ein. Außerdem können mit standardisierten Werten über die korrespondierenden Perzentile der z-Normalverteilung auch Aussagen im Sinne von Erfolgswahrscheinlichkeiten gemacht werden (vgl. Abschnitt 2.1.3).

3.1.2 Fragestellungen zur Reanalyse

Wie stellen sich Item- und Verteilungskennwerte, Dimensionalität und Vorhersagbarkeit der herkömmlichen, aus dichotomen Einzelitems gebildeten Skala EMEK_27 (nachfolgend als EMEK_27a bezeichnet) im Vergleich zu der hier entwickelten alternativen Variante EMEK_27b dar, bei der die gesamte verfügbare Iteminformation ausgeschöpft wird?

Fragestellung 1: Itemkennwerte

Wie hoch sind die Mittelwerte, Streuungen und Trennschärfen der 27 EMEK-Items bei Variante a und b? Zeigen sich Unterschiede im Vergleich? Wie hoch sind die paarweisen Korrelationen zwischen den Items der Varianten a und b?

Hypothesen zu Fragestellung 1:

- Die binären Items der Variante a können entweder den Wert 0 oder 1 annehmen, wodurch der mögliche Wertebereich der Item-Mittelwerte festgelegt ist. Insgesamt ist angesichts der bislang vorliegenden Erfahrungen mit der Skala EMEK_27 davon auszugehen, dass sich auf allen einbezogenen Ergebnisaspekten im Durchschnitt positive Veränderungen ($M > 0$) zeigen. Die Item-Mittelwerte der Variante b sind aufgrund der Standardisierung und der daraus resultierenden anderen Metrik nicht vergleichbar mit Variante a. Auf dem Hintergrund der Resultate von Lipsey und Wilson (1993) sowie Steffanowski et al. (2007) werden im Durchschnitt mittelgroße Ausprägungen von 0,50 erwartet - vorausgesetzt, dass standardisierte Abweichungen vom Erwartungswert bei direkten Veränderungsmaßen der Größe nach ähnliche Beträge wie PRÄ-POST-Effektgrößen erreichen (Schmidt et al., 2003).
- Die Item-Streuungen *SD* der dichotomen Variante a lassen sich rechnerisch direkt aus dem Item-Mittelwert M nach Gleichung 15 ableiten und haben somit einen möglichen Wertebereich von 0,00 bis 0,50. Solange keine extrem hohen oder niedrigen Item-Mittelwerte bei Variante a auftreten, sollten die Item-Streuungen typischerweise im Bereich zwischen 0,40 bis 0,50 liegen. Die Item-Streuungen der Variante b betragen aufgrund der Standardisierung durchgängig exakt 1,00.

$$SD = \sqrt{M \cdot (1 - M)} \tag{15}$$

- Bei den Item-Trennschärfen wird erwartet, dass diese bei Variante b aufgrund der besseren Informationsausschöpfung deutlich höher ausfallen als bei Variante a.

- Hinsichtlich der paarweisen Item-Korrelationen zwischen Variante a und b erwartet, dass ein Informationsverlust in der von Cohen angegebenen Größenordnung auftritt, d.h. die binären und z-standardisierten Items sollten in einer Größenordnung von .80 miteinander korrelieren.

Fragestellung 2. Skalenkennwerte

Wie hoch sind die Mittelwerte, Streuungen und Reliabilitäten der beiden Skalen-Varianten von EMEK_27? Welche Form weisen die Skalen-Verteilungen auf (Schiefe, Kurtosis und Abweichung von der Normalverteilung)? Wie hoch korrelieren EMEK_27a und EMEK_27b miteinander?

Hypothesen zu Fragestellung 2:

- Wie bei den Einzel-Items wird positive Veränderung der Patientenstichprobe auf beiden Skalenvarianten erwartet. Für Variante b wird wie bei den Items eine mittelgroße Veränderung von 0,50 Standardabweichungseinheiten erwartet.
- Bei EMEK_27b wird die Streuung im Gegensatz zu den Einzelitems kleiner als 1,00 ausfallen, da bedingt durch die Aggregation die systematische Varianz stärker als die Fehlervarianz zunimmt.
- Für die Skalen-Reliabilität wird erwartet, dass diese aufgrund der besseren Informationsausschöpfung bei Variante b höher ausfällt als bei Variante a.
- In Bezug auf die Verteilungsform kann bei EMEK_27a angesichts der bereits vorliegenden Befunde aus den Einzelstudien auch für die Gesamtstichprobe ein bimodaler Verlauf der Häufigkeiten erwartet werden. Dies bedeutet, dass neben einem großen Patientenanteil mit einem hohen Punktwert auf der Skala, was einer Verbesserung in vielen Bereichen entspricht, eine zweite, kleinere Gruppe existiert, die einen sehr geringen Punktwert erreicht und somit nur in wenigen Bereichen eine Verbesserung ihres Befindens berichtet. Bei EMEK_27b wird erwartet, dass die Verteilungsform eher der Normalverteilung entspricht und weniger extrem ausfällt als bei EMEK_27a. So sind bei normalverteilten Merkmalen mittlere Ausprägungen wahrscheinlicher als extreme Ausprägungen und für die Rohwert-Items des „Ausgangsmaterials" der Skala EMEK_27 kann angenommen werden, dass diese annähernd normalverteilt sind. Dichotomisiert man die Items nun mit den beiden zwei Extremwerten Null und Eins, so resultiert eine künstliche Varianzerweiterung der neu codierten Items der Variante a, die sich in der Verteilungsform der Gesamtskala entsprechend widerspiegelt.
- Bei der Korrelation zwischen EMEK_27a und EMEK_27b wird wie bereits bei den Einzel-Items erwartet, dass diese deutlich geringer als 1,00 ausfällt, d.h. es zeigt sich auch auf Skalenebene ein spürbarer Informationsverlust durch die vorangegangene Dichotomisierung bei der Variante a.

Fragestellung 3: Dimensionalität von EMEK_27

Welche inhaltliche Struktur zeigt sich bei einer Faktorenanalyse der 27 Einzelkomponenten? Wie viele Faktoren lassen sich extrahieren? Wie groß ist die Varianzaufklärung der Ladungsmatrix? Gibt es Unterschiede zwischen Variante a und b? Falls sich mehrere Faktoren extrahieren lassen: Wie hoch sind die Reliabilitäten der ausgehend vom Ladungsmuster gebildeten Subskalen? Welche Interkorrelationen weisen die Subskalen auf?

Hypothesen zu Fragestellung 3:

- Hinsichtlich der Gesamtvarianzaufklärung wird bei Variante b eine höhere Ausschöpfung durch die Faktorenanalyse als bei Variante a erwartet.
- Die Skala EMEK_27 hat wiederholt eine sehr hohe interne Konsistenz von über .90 bewiesen. Aus diesem Grund ist anzunehmen, dass sich bei Variante a und b jeweils ein Generalfaktor „Veränderung des gesundheitlichen Befindens" identifizieren lässt.
- Im Itempool von EMEK_27 sind zum einen Items enthalten, die auf Veränderungen des subjektiven Befindens abzielen. Andere Items erfragen Aspekte, die mehr mit Fertigkeiten zur Bewältigung des Alltags sowie mit wichtigen interpersonalen Beziehungen zusammenhängen. Eine weitere Itemgruppe beinhaltet sozialmedizinisch relevante Kriterien wie Krankschreibungszeiten und Inanspruchnahme des Gesundheitssystems. Auch die funktionale Ebene mit Einschätzungen zur subjektiv wahrgenommenen Belastbarkeit und Arbeitsfähigkeit ist im Itempool enthalten. Es wird erwartet, dass sich diese inhaltlichen Dimensionen auch in den Daten identifizieren lassen, wenn eine Extraktion und Varimax-Rotation von mehreren Faktoren durchgeführt wird.
- Die Ladungsmuster der beiden Varianten a und b sollten sich inhaltlich entsprechen und keine großen Unterschiede aufweisen, mit einer Ausnahme: Bei Variante b sind im Durchschnitt deutlich höhere Ladungsbeträge und Kommunalitäten als bei Variante a zu erwarten.
- Sofern sich mehrere Faktoren extrahieren und inhaltlich sinnvoll interpretieren lassen, kann davon ausgegangen werden, dass die entsprechend gebildeten Subskalen je nach Itemzahl zufriedenstellende bis gute Reliabilitäten aufweisen. Die Höhe der Skaleninterkorrelationen dürfte je nach dem Grad der durch die Faktorenanalyse erzielten Einfachstruktur (jedes Item sollte möglichst nur auf einem Faktor hoch und auf allen anderen Faktoren niedrig laden) variieren. Da sich Gesundheit, Leistungsfähigkeit im Alltag und Inanspruchnahme des Gesundheitssystems zum Teil gegenseitig bedingen, ist zumindest mit Skaleninterkorrelationen in mittlerer Höhe zu rechnen. Die Korrelationen zwischen den Subskalen der Variante a dürften aufgrund der geringeren Informationsausschöpfung wegen der vorangegangenen Dichotomisierung der Einzelitems dabei insgesamt niedriger ausfallen als die Korrelationen zwischen den Subskalen der Variante b.

Fragestellung 4: Vorhersagbarkeit von EMEK_27 aus Stichprobenmerkmalen

In welchem Ausmaß lässt sich der Skalenwert von EMEK_27 durch Stichprobenzugehörigkeit und Patientenmerkmale vorhersagen? Welche Prädiktoren spielen eine besonders wichtige Rolle zur Vorhersage der Ergebnisqualität? Gibt es Unterschiede zwischen Variante a und b?

Hypothesen zu Fragestellung 4:

- Hinsichtlich der Gesamtvarianzaufklärung durch die Prädiktoren wird bei Variante b aufgrund der höheren Informationsausschöpfung eine höhere multiple Korrelation mit der Skala EMEK_27 erwartet als bei Variante a.
- In der MESTA-Studie (Steffanowski et al., 2007) wird berichtet, dass der Bildungsgrad und die Behandlungsdauer positiv, die Erkrankungsdauer und der Anteil von Patienten mit somatoformen Störungen hingegen negativ mit dem Outcome korrelieren. Alters- und Geschlechtseffekte wurden nicht gefunden. Entsprechende Resultate sind auch bei der Reanalyse von EMEK_27 zu erwarten. Darüber hinaus wurde berichtet, dass das Vorliegen eines Antrages auf vorzeitige Berentung ein negativer Prädiktor für die Behandlungsergebnisse ist. Auch bei Behandlungsabbruch (vorzeitige Entlassung) kann ein schlechteres Therapieergebnis erwartet werden. Entsprechende Partialzusammenhänge sollten sich sowohl bei Variante a als auch bei Variante b nachweisen lassen.
- Darüber hinaus könnten sich Unterschiede hinsichtlich der Ergebnisqualität zwischen den fünf Teilstichproben zeigen, die unabhängig von den einbezogenen Patientenmerkmalen sind und auf andere Faktoren zurückzuführen sind (Unterschiede bei nicht einbezogenen Patientenmerkmalen oder unterschiedliche Wirksamkeit des Behandlungskonzeptes der einzelnen psychosomatischen Kliniken, in denen die Teilstichproben gewonnen wurden). Diesbezüglich wird die Hypothese formuliert, dass nach Auspartialisierung der Patientenmerkmale die mit EMEK_27 gemessene Ergebnisqualität in den fünf Teilstichproben keine Unterschiede mehr aufweisen sollte.
- Hinsichtlich der inhaltlichen Gewichtung einzelner Prädiktoren sind kaum Unterschiede zwischen a und b anzunehmen, mit einer Ausnahme: Bei Variante b sind im Durchschnitt höhere Partialgewichte als bei Variante a zu erwarten.

Fragestellung 5: Vorhersagbarkeit von EMEK_27 aus Prozessmerkmalen

In welchem Ausmaß lässt sich der Skalenwert von EMEK_27 durch Prozessmerkmale wie Qualität der therapeutischen Arbeitsbeziehung sowie Problemeinsicht und Grad an Demoralisierung unmittelbar bei Behandlungsende vorhersagen? Welche Rolle spielt dabei das Auftreten von signifikanten positiven bzw. negativen Lebensereignissen im Jahr nach der Rehabilitation? Gibt es Unterschiede zwischen Variante a und b?

Hypothesen zu Fragestellung 5:

- Wittmann et al. (2002) haben den therapeutischen Prozess mit einem entsprechenden Strukturgleichungsmodell anhand von 154 Patienten aus der Teilstichprobe A modelliert und berichten eine Gesamtvarianzaufklärung von 37% an der Skala EMEK_27a durch die einbezogenen Prädiktoren (Grad an Demoralisierung zu Beginn und am Ende der Behandlung, Qualität der therapeutischen Arbeitsbeziehung, Grad an Problemeinsicht bei Entlassung sowie Auftreten von positiven und negativen Lebensereignissen im Jahr nach Entlassung aus der Rehabilitation). Es wird angenommen, dass eine Replikation des multiplen Zusammenhangs in ähnlicher Höhe resultieren wird.
- Bei EMEK_27b sollten die Pfadkoeffizienten sowie der multiple Gesamtzusammenhang aufgrund der anzunehmenden besseren Informationsausschöpfung generell höher als bei EMEK_27a ausfallen. Ansonsten sollten inhaltlich weitgehende Entsprechungen zwischen den beiden Skalenvarianten bestehen.
- Inhaltlich wird für die Skalen des HAQ angenommen, dass die Zufriedenheit mit der therapeutischen Beziehung einen deutlichen Einfluss auf den Grad an Problemeinsicht bei Entlassung ausübt und sich letzteres wiederum positiv auf den Grad an Demoralisierung bei Entlassung und darüber hinaus auch auf das multiple Ergebniskriterium EMEK_27 auswirkt, da von einer effektiveren Alltagsbewältigung aufgrund der verbesserten Problemlösekompetenz ausgegangen werden kann. Positive Lebensereignisse im Jahr nach der Rehabilitation dürften sich positiv, negative Lebensereignisse hingegen negativ auf das multiple Ergebniskriterium EMEK_27 auswirken.
- Wittmann et al. (2002) berichten auch pfadanalytische Ergebnisse für ein multiples Ergebniskriterium, das sich ausschließlich aus kostenrelevanten Einzelaspekten (Veränderung der Krankschreibungszeiten sowie Inanspruchnahme des Gesundheitssystems) zusammensetzt und somit in erster Linie die Interessen der entsprechenden Stakeholdergruppe berücksichtigt. Bei Verwendung dieser „harten" Bewertungskriterien wird eine Varianzaufklärung von 13% berichtet. Für die hier durchgeführte Reanalyse wird ein multipler Zusammenhang in ähnlicher Größenordnung erwartet.

Zur Beantwortung der Fragestellungen wird eine Itemanalyse und Faktorenanalyse der beiden Skalenvarianten a und b für eine aus fünf Programmevaluationsstudien gebildeten Gesamtstichprobe durchgeführt. Zur Prognose von EMEK_27 werden entsprechende regressions- und pfadanalytische Untersuchungen für die beiden Skalenvarianten durchgeführt.

3.1.3 Merkmale der fünf Programmevaluationsstudien

Für die Reanalyse der Skala EMEK_27 werden die Daten von fünf Studien herangezogen, bei denen eine Ausgangsstichprobe von insgesamt 2.624 psychosomatischen Rehabilitanden befragt wurde:

- Studie A: Zauberbergstudie II (Nübling, 1992; Schmidt, Lamprecht, Nübling & Wittmann, 1994; Wittmann et al., 2002; Nübling et al., 2002; *N* = 565)
- Studie B: Reinerzauer Katamnesestudie (Nübling, Puttendörfer, Schmidt & Wittmann, 1994; Nübling, Puttendörfer, Wittmann, Schmidt & Wittich, 1995; Nübling et al., 1999; *N* = 560)
- Studie C: Bad Herrenalber Katamnesestudie (Nübling et al., 2000; Steffanowski, Oppl, Meyerberg, Schmidt & Wittmann, 2001; *N* = 317)
- Studie D: EQUA-Studie (Schmidt, Karcher, Steffanowski, Nübling & Wittmann, 2000; Schmidt et al., 2003; Nübling, Steffanowski, Wittmann & Schmidt, 2004; Schmidt, Nübling, Steffanowski, Lichtenberg & Wittmann, 2006; multizentrisch mit Teilstichproben aus vier Kliniken, *N* = 858)
- Studie E: INDIKA-Studie (Nübling et al., 2004), multizentrisch mit drei Indikationsbereichen aus insgesamt sieben Kliniken: Psychosomatik (drei Kliniken, *N* = 324), Kardiologie (zwei Kliniken, *N* = 370) und Orthopädie (zwei Kliniken, *N* = 270). Berücksichtigung findet im Rahmen dieser Arbeit allerdings ausschließlich die Teilstichprobe Psychosomatik mit 324 Patienten.

Die fünf Programmevaluationsstudien weisen folgende gemeinsame Merkmale auf:

- Prospektive Eingruppen-PRÄ-POST-Verlaufsstudie ohne Kontrollgruppe
- Drei Messzeitpunkte (Aufnahme, Entlassung und 1-Jahres-Katamnese)
- Heterogene Patientenstichprobe der stationären psychosomatischen Rehabilitation (alle Patienten, die im Studienzeitraum behandelt wurden, konnten teilnehmen)
- Freiwillige Studienteilnahme
- Datenquelle: Angaben der Patienten sowie Angaben der Kliniktherapeuten
- Sämtliche Items der Skala EMEK_27 wurden zum Katamnese-Zeitpunkt von den Patienten erhoben.

Tabelle 11 enthält die Stichprobeneigenschaften für die fünf Studien. Alle Merkmale beziehen sich auf die Aufnahme-Messung. Es handelt sich um heterogene Patientenstichproben, wie sie für den Bereich der stationären psychosomatischen Rehabilitation charakteristisch sind. Zwischen den Studien bestehen zum Teil erhebliche Unterschiede hinsichtlich der Zusammensetzung der Patientenmerkmale (Case-Mix).

Tabelle 11. Stichprobenmerkmale der fünf Programmevaluationsstudien

Studie	A	B	C	D	E	p
Stichprobe						
Basiserhebung (Jahr)	1989	1990	1997	2000	2001	
Ausgangsstichprobe (*N*)	565	560	317	858	324	
1-Jahres-Katamnese-Antworter (*N*)	367	401	191	569	210	
Rücklaufquote in %	65,0%	71,6%	60,3%	66,3%	64,8%	.011
Geschlecht						
weiblich	51,9%	57,0%	59,6%	56,8%	63,3%	.016
Alter in Jahren						
M	42,2	41,4	39,9	43,8	41,4	<.001
SD	10,2	10,2	11,6	10,6	11,6	
Schulbildung						
bis Hauptschule	62,3%	70,0%	28,5%	35,7%	53,3%	<.001
Realschule / Mittl. Reife / Sonstiges	24,4%	18,4%	34,5%	28,7%	25,2%	
Fachabitur / Abitur	13,3%	11,6%	37,0%	35,6%	21,5%	
Familienstand						
Ledig	18,9%	19,5%	37,2%	25,5%	28,0%	<.001
Verheiratet	65,0%	62,0%	40,1%	55,1%	50,6%	
Getrennt / Geschieden	12,4%	15,0%	20,2%	15,9%	16,4%	
Verwitwet	3,7%	3,6%	2,5%	3,5%	5,0%	
Kostenträger						
LVA (Landesversicherungsanstalt)	33,6%	55,3%	16,1%	19,2%	36,2%	<.001
BfA (Bundesvers. für Angestellte)	41,6%	33,8%	38,8%	49,1%	12,6%	
GKV (Gesetzliche Krankenvers.)	15,2%	8,0%	25,2%	18,3%	39,1%	
PKV (Private Krankenversicherung)	9,6%	2,9%	17,7%	10,4%	4,5%	
Sonstige	0,0%	0,0%	2,2%	2,9%	7,4%	
Beruflicher Status						
Vollzeit erwerbstätig	65,8%	68,0%	43,3%	52,5%	48,1%	<.001
Teilzeit erwerbstätig	9,5%	14,7%	20,6%	16,6%	16,2%	
Arbeitslos	11,3%	9,0%	15,1%	14,8%	16,5%	
Nicht erwerbstätig	13,4%	8,3%	20,9%	16,1%	19,2%	
Antrag auf vorzeitige Berentung						
Anteil der Rentenantragsteller	4,6%	2,0%	3,9%	5,1%	6,1%	.029

Anmerkung. *p* – Unterschiede zwischen den Stichproben wurden beim Alter mit einfaktorieller Varianzanalyse, ansonsten mit Chiquadrat-Test auf statistische Signifikanz überprüft.

Tabelle (Fortsetzung). Stichprobenmerkmale der fünf Programmevaluationsstudien

Studie	A	B	C	D	E	p
Erkrankungsdauer in Jahren						
M	7,5	4,8	7,1	6,4	5,4	<.001
SD	5,6	3,5	3,5	5,4	5,1	
Hauptdiagnose nach ICD-10						
Depressive Stör. (F32, F33, F34.1)	24,6%	17,5%	44,1%	35,9%	43,9%	<.001
Angststörungen (F40, F41)	4,8%	9,3%	5,9%	10,7%	13,5%	
Somatoforme Störungen (F45)	25,0%	23,2%	2,6%	5,0%	1,4%	
Anpassungsstörungen (F43.2, F48)	11,0%	16,6%	7,0%	32,5%	19,9%	
Belast./ dissoz. Stör. (F43.0/1, F44)	6,2%	7,7%	1,5%	2,4%	2,5%	
Persönlichkeitsstörungen (F60)	3,5%	5,7%	8,9%	3,2%	2,1%	
Essstörungen (F50)	1,9%	1,3%	6,7%	2,4%	3,9%	
Substanzmissbrauch (F1)	0,5%	0,4%	3,3%	1,3%	1,1%	
Sonstige psych. Störungen (Fxx.x)	8,7%	10,4%	19,6%	2,4%	10,3%	
Somatische Diagnose (A-E, G-Z)	13,8%	8,0%	0,4%	4,0%	1,8%	
Nebendiagnosen						
Nebendiagnose(n) vorhanden	79,5%	68,2%	70,7%	75,0%	58,7%	<.001
Behandlungsdauer						
M	52,3	47,6	54,9	49,6	48,5	<.001
SD	19,8	20,7	21,2	17,7	22,7	
Entlassung aus der Rehabilitation						
Vorzeitig (Behandlungsabbruch)	6,5%	7,3%	9,5%	6,9%	11,1%	.111

Anmerkung. *p* - Unterschiede zwischen den Stichproben wurden bei der Erkrankungs- und Behandlungsdauer mit einfaktorieller Varianzanalyse, ansonsten mit Chiquadrat-Test auf statistische Signifikanz überprüft.

3.1.4 Repräsentativität der Katamnese-Antworter

Die Daten aus den fünf Studien wurden zu einem Gesamtdatensatz zusammengeführt, der eine Ausgangsstichprobe von insgesamt 2.624 Patienten umfasst. Insgesamt 1.738 davon haben sich an der 1-Jahres-Katamnese beteiligt und die ausgefüllten Fragebögen zurückgeschickt, was einer Rücklaufquote von 66,2% entspricht. Es wurde überprüft, inwieweit diese Antworterstichprobe repräsentativ für die Ausgangsstichprobe ist. Tabelle 12 enthält die Ergebnisse.

Tabelle 12. Vergleich zwischen Antwortern und Nichtantwortern (Gesamtstichprobe)

Katamnese-Antworter	**ja**	**nein**	**Gesamt**	**p**
Anzahl *N*	1738	886	2624	
Anteil in %	66,2%	33,8%	100,0%	
Geschlecht				
weiblich	57,0%	56,8%	56,9%	.959
Alter in Jahren				
M	42,6	41,3	42,2	.003
SD	10,5	11,2	10,8	
Schulbildung				
bis Hauptschule	49,1%	51,9%	50,1%	.395
Realschule / Mittl. Reife / Sonstiges	26,2%	25,1%	25,8%	
Fachabitur / Abitur	24,6%	23,0%	24,1%	
Familienstand				
Ledig	23,2%	27,2%	24,5%	.001
Verheiratet	54,0%	45,6%	51,2%	
Getrenntlebend / Geschieden	19,8%	22,3%	20,6%	
Verwitwet	3,0%	4,9%	3,6%	
Kostenträger				
LVA	29,9%	35,3%	31,7%	.005
BfA	41,2%	33,7%	38,6%	
GKV	18,1%	20,0%	18,7%	
PKV	8,7%	9,0%	8,8%	
Sonstige	2,1%	2,0%	2,1%	
Beruflicher Status				
Vollzeit erwerbstätig	57,6%	56,1%	57,1%	.347
Teilzeit erwerbstätig	15,5%	14,2%	15,1%	
Arbeitslos	12,3%	14,5%	13,0%	
Nicht erwerbstätig	14,6%	15,2%	14,6%	
Antrag auf vorzeitige Berentung				
Anteil der Rentenantragsteller	4,4%	3,9%	4,2%	.696
Erkrankungsdauer in Jahren				
Mittelwert	6,3	6,3	6,3	.981
Streuung	5,0	5,1	5,0	

Anmerkung. p - Stichprobenunterschiede wurden beim Alter und bei der Erkrankungsdauer mit einfaktorieller Varianzanalyse, ansonsten mit Chiquadrat-Test auf statistische Signifikanz überprüft.

Tabelle (Fortsetzung). Vergleich zwischen Antwortern und Nichtantwortern

Katamnese-Antworter	ja	nein	Gesamt	p
Anzahl N	1738	886	2624	
Anteil in %	66,2%	33,8%	100,0%	
Hauptdiagnose nach ICD-10				
Depressive Störungen (F32, F33, F34.1)	31,0%	30,8%	31,0%	.371
Angststörungen (F40, F41)	9,2%	8,2%	8,9%	
Somatoforme Störungen (F45)	13,4%	11,9%	12,9%	
Anpassungsstörungen (F43.2, F48)	19,3%	21,1%	19,9%	
Belast./ dissoz. Stör. (F43.0/1, F44)	4,2%	4,8%	4,4%	
Persönlichkeitsstörungen (F60)	4,2%	4,6%	4,3%	
Essstörungen (F50)	2,6%	2,9%	2,7%	
Substanzmissbrauch (F1)	0,8%	1,8%	1,1%	
Sonstige psych. Störungen (Fxx.x)	8,5%	8,0%	8,4%	
Somat. Diagnose (A-E, G-Z)	6,8%	5,8%	6,5%	
Nebendiagnosen				
Nebendiagnose vorhanden	73,2%	70,3%	72,2%	.145
Behandlungsdauer				
Mittelwert	51,0	48,7	50,2	.007
Streuung	18,9	21,7	19,9	
Entlassung aus der Rehabilitation				
Vorzeitig (Behandlungsabbruch)	5,4%	12,2%	7,7%	<.001

Anmerkung. p - Stichprobenunterschiede wurden bei der Behandlungsdauer mit einfaktorieller Varianzanalyse, ansonsten mit Chiquadrat-Test auf statistische Signifikanz überprüft.

Deutliche Unterschiede zwischen Antwortern und Nichtantwortern ergeben sich beim Anteil der vorzeitig entlassenen Patienten: Während die Abbrecherquote bei den Antwortern nur bei 5,7% liegt, beträgt diese bei den Nichtantwortern 12,2%. Dies erklärt auch die durchschnittlich um zwei Tage geringere Behandlungsdauer bei den Nichtantwortern. Darüber hinaus sind unter den Antwortern mehr BfA-Patienten und Verheiratete. Auch das Durchschnittsalter der Antworter liegt etwas höher. Keine statistisch signifikanten Unterschiede zwischen Anwortern und Nichtantwortern bestehen hinsichtlich Geschlecht, Schulbildung, Erwerbsstatus, Rentenantrag, Krankheitsdauer und Hauptdiagnose.

3.2 Ergebnisse der Reanalyse von EMEK_27

Nach Behandlung der Missing-Data-Problematik wurden die Item- und Skalen-Kennwerte, die dimensionale Struktur sowie die Validität der beiden Varianten von E-MEK_27 untersucht.

3.2.1 Itemkennwerte der Ausgangsdaten und fehlende Werte

Nicht alle Katamnese-Teilnehmer haben sämtliche 27 Rohwert-Items beantwortet. Im Gegenteil: Abbildung 13 veranschaulicht, dass häufig bis zu drei Angaben pro Patient fehlen (Missing-Data). Lediglich 401 (23,1% von 1.738) Patienten haben alle 27 Items durchgängig beantwortet. Bei 459 Patienten fehlt ein Item, bei 423 Patienten fehlen zwei Items, bei 258 Patienten fehlen drei Items und 78 Patienten haben vier Items nicht beantwortet. Auffällig an der Verteilung ist darüber hinaus, dass diese ein zweites, kleineres Maximum bei 11 fehlenden Items aufweist.

Schaut man sich näher an, welche Items besonders häufig unbeantwortet bleiben (vgl. Tabelle 14), so stehen die Krankschreibungszeiten mit 45,8% fehlenden Angaben an erster Stelle, gefolgt von der Beziehung zum Partner mit 38,1% und dem Familienleben mit Kindern (35,7%). Diese Fragen sind für einen erheblichen Anteil von Patienten irrelevant, weil nicht alle erwerbstätig sind, in einer Partnerschaft leben oder eine Familie mit Kindern haben. Auch Fragen nach dem Medikamentenkonsum (8,5%) und nach Krankenhausaufenthalten (8,1%) bleiben eher unbeantwortet als andere Items.

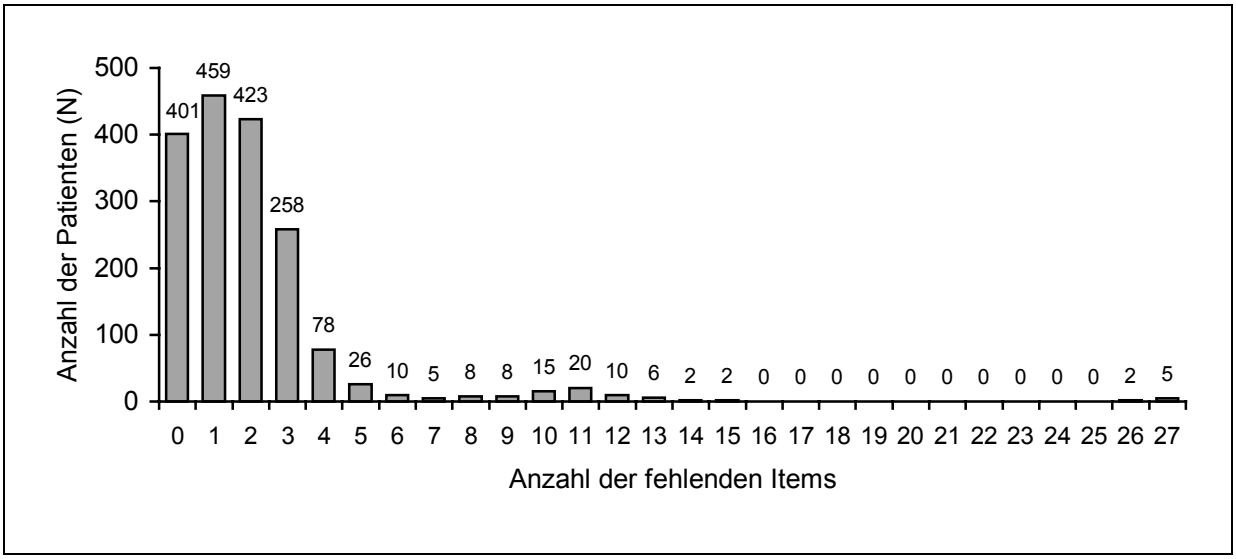


Abbildung 13. Missing-Data bei den Ausgangsdaten für die Skala EMEK_27. *N* = 1.738; *M* = 2,05; *SD* = 2,74.

Für jedes der 27 Rohwert-Items wurde zur Analyse der fehlenden Werte eine Indikatorvariable gebildet, in der das Vorhandensein bzw. Fehlen des betreffenden Items im katamnestischen Datensatz codiert wurde. Eine Hauptkomponentenanalyse der 27 Indikatorvariablen ergab, dass folgende Items besonders häufig gemeinsam fehlen: Item 19-27 (1. Faktor), Item 3-8 (2. Faktor), Item 15-18 (3. Faktor) sowie Item 13 und 14 (4. Faktor). Bei den Items 19-27 handelt es sich um die 9 Items der Veränderungsskala VM9, die in den fünf Studien jeweils als zusammengehöriger Item-Block katamnestisch erhoben wurden. Bei einigen Patienten in den fünf Programmevaluationsstudien fehlte das entsprechende Blatt in der Fragebogenmappe bzw. der Itemblock wurde bei der Beantwortung komplett ausgelassen, was das zweite Maximum in der Verteilung in Abbildung 13 erklärt. Die Items 3-8 bilden ebenfalls einen eigenen Itemblock. Bei den Items 15-18 handelt es sich um vier sozialmedizinisch relevante Variablen und bei den Items 13 und 14 schließlich um die beiden Fragen nach Partnerschaft und Familienleben.

Für die Reanalyse der Daten muss ein Umgang mit den fehlenden Werten gefunden werden, will man nicht unangemessen viel Information „verschenken". Hierzu existieren eine Reihe von Verfahrensweisen (vgl. Rubin & Little, 1987; Müller 2002; Wirtz, 2004). Als Standardverfahren hat sich in den letzten Jahren der EM-Algorithmus durchgesetzt, der auf einer Maximum-Likelihood-Schätzung fehlender Werte beruht. Andere bekannte Verfahren wie Fallweiser Ausschluss, Paarweiser Ausschluss, Ersetzung durch Mittelwerte oder die Regressionsmethode werden aufgrund der Nachteile dieser Verfahren nicht mehr empfohlen (Rabung, 2007). Bildet man die Skala EMEK_27 zunächst nur für die 401 Patienten mit vollständiger Messung auf allen 27 Items, so beträgt Cronbachs Alpha .94. Eine derart hohe Konsistenz weist darauf hin, dass hohe Interkorrelationen zwischen den 27 Einzelitems bestehen und sich die Werte einzelner fehlender Items relativ zuverlässig aus den Werten der vorhandenen Items schätzen lassen. Die Ausprägungen der fehlenden Items in den Ursprungsdaten wurden daher mit dem EM-Algorithmus aus den übrigen Items geschätzt und in den Datensatz eingesetzt. Der Little's MCAR-Test weist ein Chiquadrat von 18.931 aus (df = 14.874, p < .001). Dies weist darauf hin, dass die fehlenden Daten nicht „Completely Missing at Random" (MCAR) sind. Es ist daher von „Missing at Random" (MAR) auszugehen, d.h. zwischen bestimmten Subgruppen von Patienten gibt es Unterschiede hinsichtlich der Missing-Verteilung, was hier zu erwarten war. Wie erwähnt, sind Krankschreibungszeiten, Partnerschaft und familiäre Situation für bestimmte Patienten nicht relevant und die entsprechenden Items bleiben dann unbeantwortet. Tabelle 13 gibt das „Ausgangsmaterial" der 27 Items für die weiteren Analysen vor und nach Durchführung der EM-Schätzung fehlender Werte wieder.

Abbildung 14 gibt die Häufigkeitsverteilungen der 27 Originalitems nach Durchführung der EM-Schätzung fehlender Werte wieder. Aus diesen werden später die Varianten a und b der singulären Ergebniskriterien gebildet. Niedrige Werte stehen bei allen 27 Items hier für eine Verbesserung des Befindens, die mittlere Kategorie steht für ein unverändertes Befinden und hohe Werte stehen für eine Verschlechterung des Befindens. Eine Ausnahme bildet die Frage nach dem aktuellen Befinden (Item 1), die keine Veränderungsmessung, sondern eine Statusmessung beinhaltet (1 = gut, 2 = eher gut, 3 = weder noch, 4 = eher schlechter, 5 = schlecht). Die Häufigkeitsverteilungen der Items deuten durchgängig auf ein positives Bild der Ergebnisqualität hin.

Tabelle 13. Item-Kennwerte der Ausgangsdaten vor und nach EM-Schätzung

		vor EM-Schätzung				nach EM-Schätzung			
Nr.	**Item**	***N***	***MD***	***M***	***SD***	***N'***	***MD'***	***M'***	***SD'***
1	Befinden bei Katamnese	1713	1,4%	2,52	1,16	1738	0,0%	2,52	1,16
2	Lebensqualität	1713	1,4%	2,11	1,08	1738	0,0%	2,12	1,09
3	Körperliches Befinden	1721	1,0%	2,35	1,05	1738	0,0%	2,35	1,05
4	Seelisches Befinden	1713	1,4%	2,28	1,13	1738	0,0%	2,27	1,13
5	Allgemeinbefinden	1709	1,7%	2,30	1,06	1738	0,0%	2,30	1,06
6	Leistungsfähigkeit	1714	1,4%	2,39	1,10	1738	0,0%	2,39	1,10
7	Beschwerden	1715	1,3%	2,23	1,06	1738	0,0%	2,23	1,06
8	Gesundheitszustand	1718	1,2%	2,32	1,13	1738	0,0%	2,32	1,13
9	Umgang Alltagsbelastungen	1720	1,0%	1,40	0,63	1738	0,0%	1,40	0,63
10	Gesundheitsbewusstes Leben	1692	2,6%	1,62	0,56	1738	0,0%	1,62	0,56
11	Medikamentenkonsum	1591	8,5%	1,79	0,71	1738	0,0%	1,78	0,71
12	Beziehungen allgemein	1688	2,9%	1,65	0,64	1738	0,0%	1,65	0,64
13	Beziehung zum Partner	1076	38,1%	1,77	0,65	1738	0,0%	1,79	0,66
14	Familienleben mit Kindern	1117	35,7%	1,73	0,62	1738	0,0%	1,72	0,62
15	Arbeitsfähigkeit	1669	3,9%	1,66	0,72	1738	0,0%	1,66	0,72
16	Arztbesuche	1638	5,8%	-0,46	1,42	1738	0,0%	-0,47	1,42
17	Krankschreibungszeiten (AU)	942	45,8%	-0,80	2,00	1738	0,0%	-0,74	2,05
18	Krankenhaustage	1598	8,1%	-0,38	1,84	1738	0,0%	-0,38	1,84
19	Wohlbefinden	1655	4,8%	3,09	1,42	1738	0,0%	3,10	1,42
20	Umgang mit Problemen	1658	4,6%	3,13	1,30	1738	0,0%	3,14	1,30
21	Selbsthilfe	1663	4,3%	3,00	1,24	1738	0,0%	3,00	1,24
22	Umgang mit Enttäuschungen	1663	4,3%	3,32	1,26	1738	0,0%	3,33	1,26
23	Zurechtkommen mit Arbeit	1636	5,9%	3,40	1,41	1738	0,0%	3,41	1,41
24	Belastbarkeit	1657	4,7%	3,42	1,43	1738	0,0%	3,43	1,43
25	Auskommen Mitmenschen	1655	4,8%	3,19	1,16	1738	0,0%	3,19	1,16
26	Leben mit Einschränkungen	1663	4,3%	3,24	1,33	1738	0,0%	3,25	1,33
27	Ausgeglichenheit	1713	4,4%	3,28	1,31	1738	0,0%	3,29	1,31

Anmerkungen. *n*, *MD*, *M* und *SD*: Stichprobengröße, Anteil der mittels EM-Algorithmus zu ersetzenden fehlenden Antworten, Mittelwert und Streuung der Items vor Schätzung fehlender Werte. *n'*, *MD'*, *M'* und *SD'* : Stichprobengröße, Anteil fehlender Antworten, Mittelwert und Streuung nach Schätzung und Imputation fehlender Werte mit EM-Algorithmus.

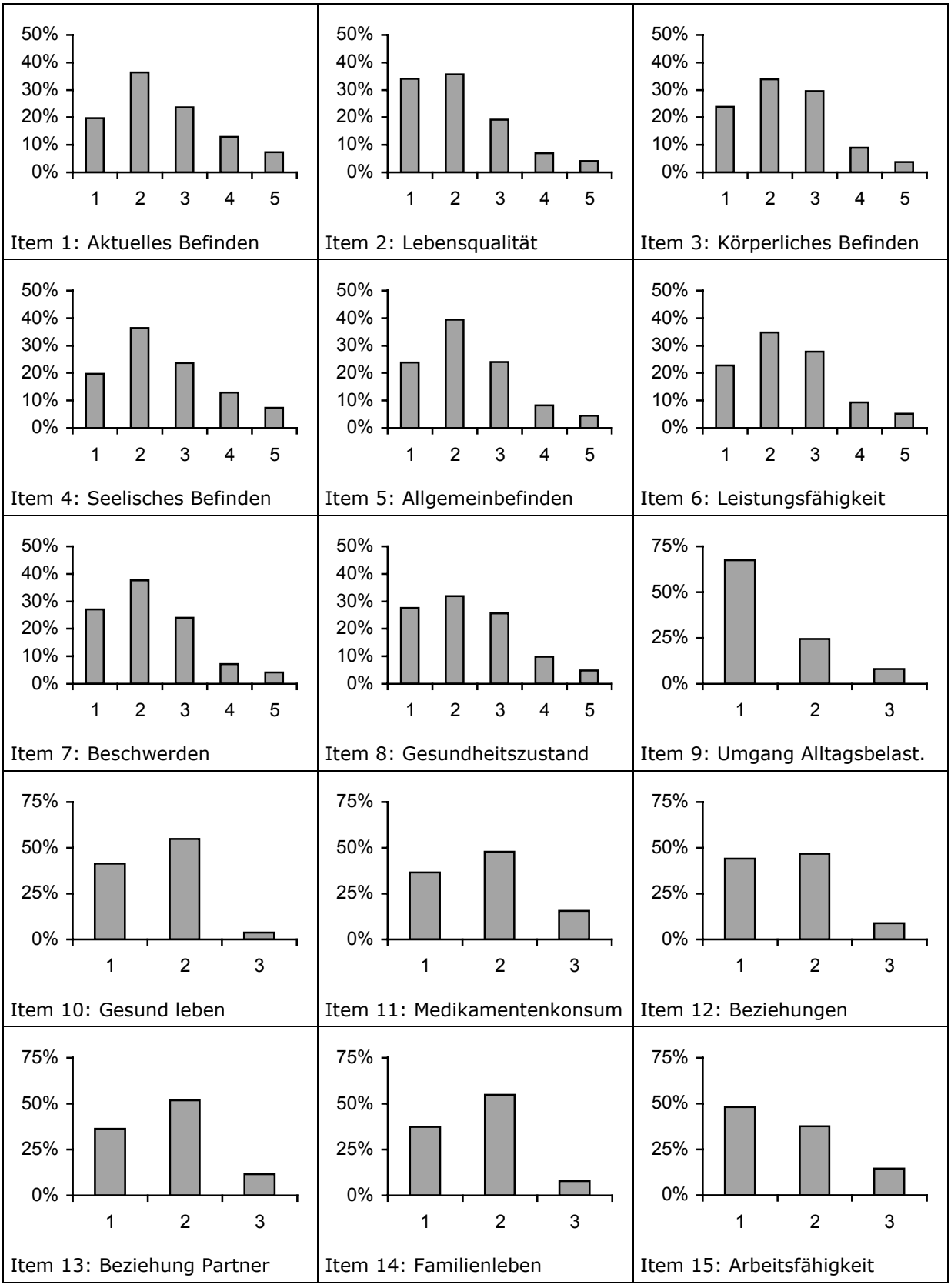


Abbildung 14. Häufigkeitsverteilungen der 27 Rohwert-Items, aus denen die singulären Ergebniskriterien der Skala EMEK_27 codiert werden (*N* = 1.738).

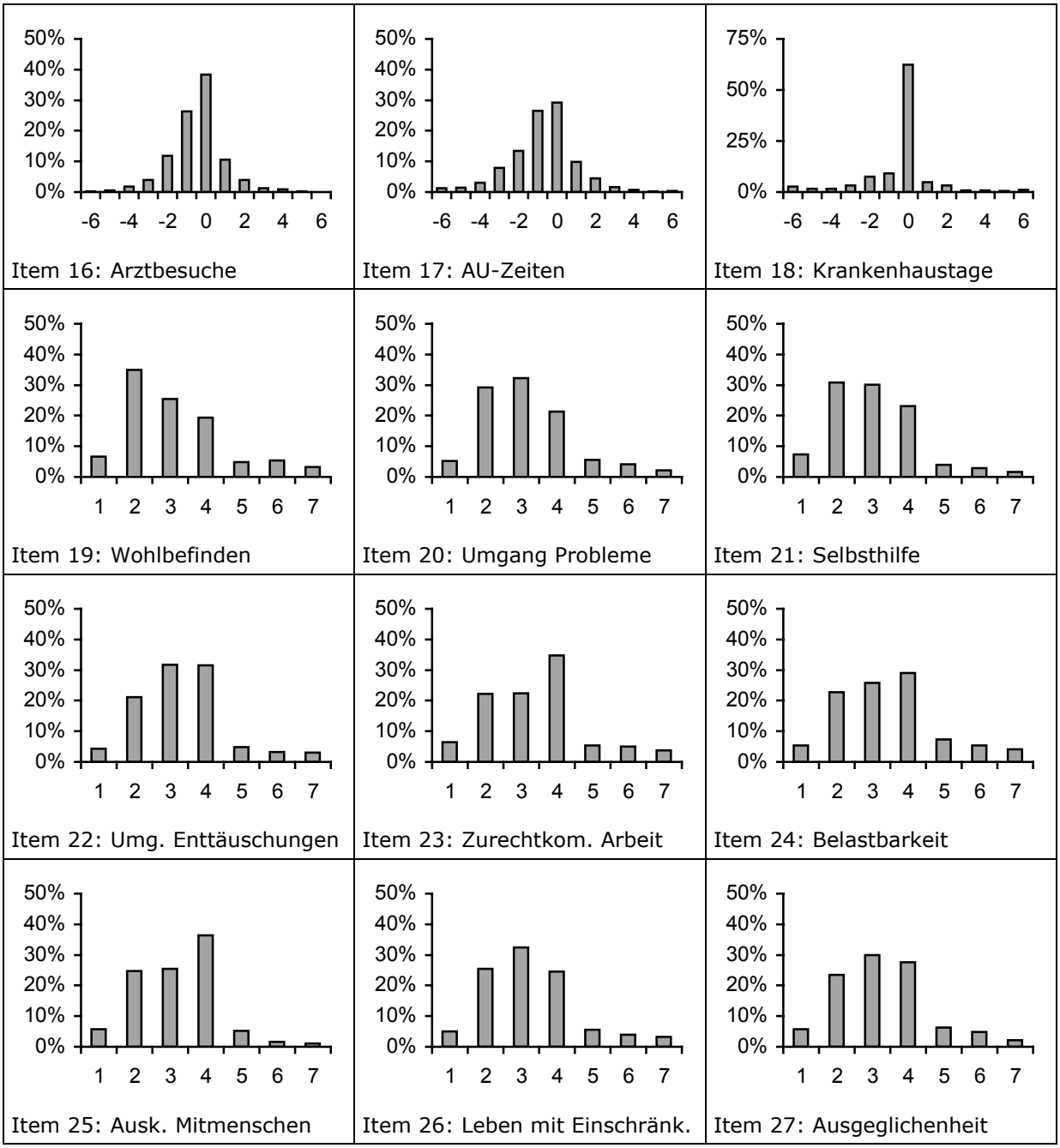


Abbildung (Fortsetzung). Häufigkeitsverteilungen der 27 Rohwert-Items, aus denen die singulären Ergebniskriterien der Skala EMEK_27 codiert werden (*N* = 1.738).

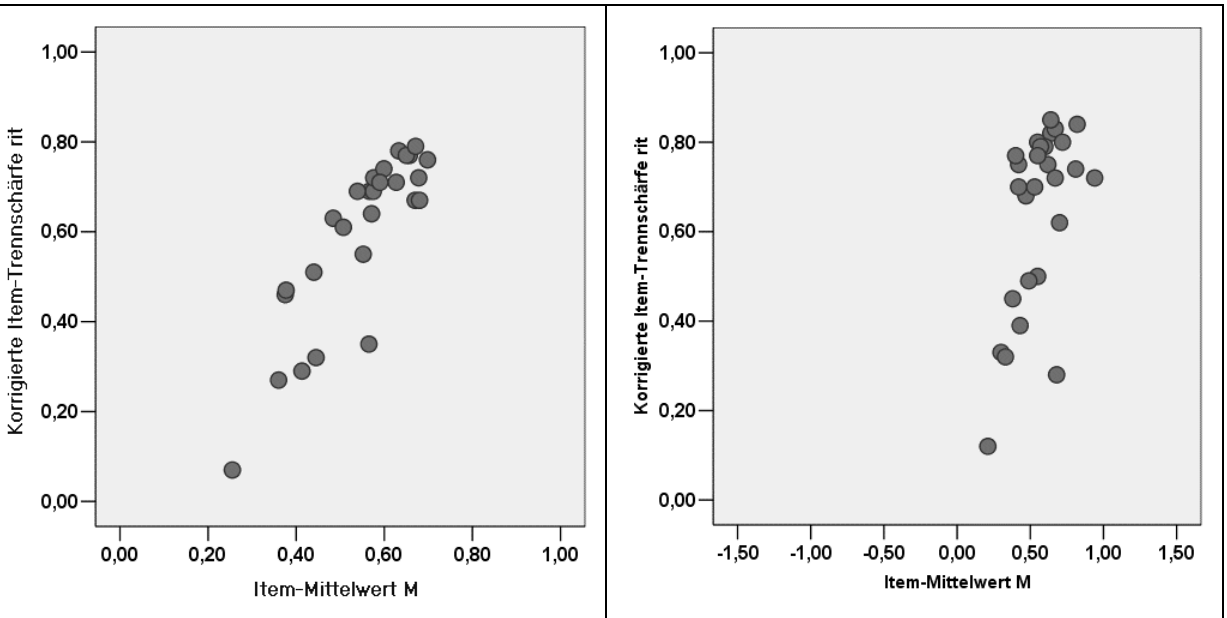
3.2.2 Itemkennwerte der Skala EMEK_27

Nach EM-Schätzung der fehlenden Werte und Vervollständigung der Rohdaten wurden die 27 EMEK-Items der Variante a anhand des Codierschemas in Tabelle 10 (Abschnitt 3.1.1) gebildet. In Tabelle 14 sind die Kennwerte aller 27 singulären Ergebniskriterien wiedergegeben. Die Mittelwerte der 27 binär mit Null und Eins codierten singulären Ergebniskriterien bewegen sich zwischen 0,26 (Veränderung Krankenhaustage) und 0,70 (Verände-

rung Lebensqualität). Der gemittelte Wert über alle Items beträgt 0,55 (95%-Konfidenzintervall: +/- 0,02). Die Item-Mittelwerte der Variante a lassen sich direkt im Sinne von „Erfolgsprozenten" interpretieren: 55% aller Patienten berichten auf den einbezogenen Einzelkriterien im Durchschnitt also eine Verbesserung des Befindens. Die über alle Items gemittelte Streuung beträgt 0,48 (Range: 0,44 - 0,50). Die korrigierten Item-Trennschärfen bewegen sich in einem Bereich zwischen .07 und .79 (gemittelter Wert: .59).

Die Itemkennwerte der Variante b sind in Tabelle 15 wiedergegeben. Diese wurden nach Gleichung 13 (vgl. Abschnitt 3.1.1) berechnet. Die über alle Items gemittelte standardisierte Abweichung vom Erwartungswert beträgt 0,56 (95%-Konfidenzintervall: +/- 0,05). Der Range bewegt sich zwischen 0,21 (Veränderung Krankenhaustage) und 0,94 (Veränderung im Umgang mit Alltagsbelastungen). Die Item-Streuungen sind nicht wiedergegeben, da diese aufgrund der Standardisierung durchgängig 1,00 betragen. Die korrigierten Itemtrennschärfen sind bei Variante b mit einem gemittelten Wert von .64 (Range: .12 bis .85) im Durchschnitt etwas höher als bei Variante a. Die 27 paarweisen Item-Korrelationen zwischen Variante a und Variante b liegen in einem Bereich zwischen .71 und .94 (gemittelter Wert: .84).

Abbildung 15 veranschaulicht die Ergebnisse der Itemanalyse für beide Varianten vergleichend in grafischer Form. Im linken Teil der Abbildung sind die Ergebnisse für EMEK_27a als Scatterplot wiedergegeben. Der rechte Teil enthält die entsprechende Darstellung für EMEK_27b. Mit Ausnahme von Item 18 („Veränderung Krankenhaustage") liegen alle Items oberhalb der von Lienert und Raatz (1998) angegebenen Mindestgrenze von .20, wenngleich auch die Items 10 „Gesundheitsbewusste Lebensführung", 11 „Medikamentenkonsum", 16 „Arztbesuche" und 17 „Krankschreibungszeiten" nur sehr mäßige Werte unterhalb .40 erreichen.



Variante a | Variante b

Abbildung 15. Itemanalyse der Skala EMEK_27a und EMEK_27b (*N* = 1.738)

Tabelle 14. Kennwerte der 27 singulären Ergebniskriterien der Variante a

	Item	***M***	***SD***	r_{it}
1	Befinden zum Katamnese-Zeitpunkt	0,56	0,50	.68
2	Lebensqualität	0,70	0,46	.76
3	Körperliches Befinden	0,58	0,49	.69
4	Seelisches Befinden	0,66	0,48	.77
5	Allgemeinbefinden	0,63	0,48	.78
6	Leistungsfähigkeit	0,58	0,49	.72
7	Beschwerden	0,65	0,48	.76
8	Gesundheitszustand	0,59	0,49	.74
9	Umgang mit Alltagsbelastungen	0,68	0,47	.71
10	Gesundheitsbewusste Lebensführung	0,41	0,49	.29
11	Medikamentenkonsum	0,37	0,48	.27
12	Beziehungen zu Bezugspersonen	0,44	0,50	.51
13	Beziehung zum Partner	0,36	0,48	.45
14	Familienleben mit Kindern	0,37	0,48	.47
15	Arbeitsfähigkeit	0,48	0,50	.62
16	Arztbesuche	0,45	0,50	.32
17	Krankschreibungszeiten (AU)	0,54	0,50	.34
18	Krankenhaustage	0,26	0,44	.07
19	Wohlbefinden	0,67	0,47	.79
20	Umgang mit Problemen	0,67	0,47	.68
21	Selbsthilfe	0,68	0,47	.67
22	Umgang mit Enttäuschungen	0,57	0,45	.65
23	Zurechtkommen mit Arbeit	0,51	0,50	.62
24	Belastbarkeit	0,54	0,50	.70
25	Auskommen Mitmenschen	0,56	0,50	.56
26	Leben mit Einschränkungen	0,63	0,48	.71
27	Ausgeglichenheit	0,59	0,49	.72
	Gemittelter Wert über alle Items	**0,55**	**0,48**	**.59**

Anmerkungen. N = 1.738 Katamnese-Antworter in fünf Programmevaluationsstudien. Codierung der Items bei Variante a: 0=Keine Veränderung bzw. Verschlechterung, 1=Verbesserung.
M – Mittelwert, SD – Streuung, r_{it} – Korrigierte Item-Trennschärfe.

Tabelle 15. Kennwerte der 27 singulären Ergebniskriterien der Variante b

	Item	M_{dVM}	SD_{dVM}	Min	E	Max	z_{dVM}	r_{it}	r_{ab}
1	Befinden bei Katamnese	2,52	1,16	1	3	5	0,42	.75	.85
2	Lebensqualität	2,12	1,08	1	3	5	0,82	.84	.85
3	Körperliches Befinden	2,35	1,05	1	3	5	0,62	.75	.85
4	Seelisches Befinden	2,28	1,13	1	3	5	0,64	.82	.85
5	Allgemeinbefinden	2,30	1,06	1	3	5	0,67	.83	.84
6	Leistungsfähigkeit	2,39	1,09	1	3	5	0,55	.80	.84
7	Beschwerden	2,24	1,05	1	3	5	0,72	.80	.84
8	Gesundheitszustand	2,33	1,13	1	3	5	0,60	.79	.85
9	Umgang Alltagsbelastungen	1,40	0,63	1	2	3	0,94	.72	.92
10	Gesundheitsbewusstes Leben	1,62	0,56	1	2	3	0,68	.28	.94
11	Medikamentenkonsum	1,79	0,69	1	2	3	0,30	.33	.87
12	Beziehungen allgemein	1,65	0,64	1	2	3	0,55	.50	.90
13	Beziehung zum Partner	1,75	0,65	1	2	3	0,38	.45	.88
14	Familienleben mit Kindern	1,71	0,61	1	2	3	0,49	.49	.90
15	Arbeitsfähigkeit	1,66	0,71	1	2	3	0,47	.68	.89
16	Arztbesuche	-0,46	1,39	-6	0	+6	0,33	.32	.76
17	Krankschreibungszeiten (AU)	-0,74	1,73	-6	0	+6	0,43	.39	.75
18	Krankenhaustage	-0,38	1,78	-6	0	+6	0,21	.12	.71
19	Wohlbefinden	3,10	1,41	1	4	7	0,64	.85	.83
20	Umgang mit Problemen	3,14	1,28	1	4	7	0,67	.72	.81
21	Selbsthilfe	3,00	1,23	1	4	7	0,81	.74	.81
22	Umgang mit Enttäuschungen	3,33	1,25	1	4	7	0,53	.70	.79
23	Zurechtkommen mit Arbeit	3,41	1,39	1	4	7	0,42	.70	.80
24	Belastbarkeit	3,43	1,41	1	4	7	0,40	.77	.81
25	Auskommen Mitmenschen	3,19	1,15	1	4	7	0,70	.62	.83
26	Leben mit Einschränkungen	3,25	1,32	1	4	7	0,57	.79	.80
27	Ausgeglichenheit	3,29	1,30	1	4	7	0,55	.77	.81
	Gemittelter Wert alle Items						**0,56**	**.64**	**.84**

Anmerkungen. N = 1.738 Katamnese-Antworter in fünf Studien. M_{dVM} und SD_{dVM} – Mittelwert und Streuung der Rohwerte nach Schätzung und Ersetzung fehlender Werte mit EM-Algorithmus. *Min* – kleinstmöglicher Itemwert. *Max* – größtmöglicher Itemwert. E – Erwartungswert bei unverändertem Befinden. Z-standardisierte Abweichung vom Erwartungswert bei direkter Veränderungsmessung: $z_{dVM} = (M_{dVM} - E) / SD_{dVM}$. Verbesserungen des Befindens sind bei z_{dVM} mit einem positiven Vorzeichen dargestellt. Item 1 beinhaltet anstelle der dVM eine Statusmessung. Item 16-18 beinhalten anstelle der dVM eine quasi-indirekte Veränderungsmessung. r_{ab} = Korrelation zwischen den jeweiligen Items der Variante a und b. r_{it} – Korrigierte Item-Trennschärfe

Nr	Item	*M* Variante a	*M* Variante b
1	Befinden bei Katamnese	0,56	0,42
2	Lebensqualität	0,70	0,82
3	Körperliches Befinden	0,58	0,62
4	Seelisches Befinden	0,66	0,64
5	Allgemeinbefinden	0,63	0,67
6	Leistungsfähigkeit	0,58	0,55
7	Beschwerden	0,65	0,72
8	Gesundheitszustand	0,59	0,60
9	Umgang Alltagsbelastungen	0,68	0,94
10	Gesundheitsbewusste Lebensführung	0,41	0,68
11	Medikamentenkonsum	0,37	0,30
12	Beziehungen allgemein	0,44	0,55
13	Beziehung zum Partner	0,36	0,38
14	Familienleben mit Kindern	0,37	0,49
15	Arbeitsfähigkeit	0,48	0,47
16	Arztbesuche	0,45	0,33
17	Krankschreibungszeiten	0,54	0,43
18	Krankenhaustage	0,26	0,21
19	Wohlbefinden	0,67	0,64
20	Umgang mit Problemen	0,67	0,67
21	Selbsthilfe	0,68	0,81
22	Umgang mit Enttäuschungen	0,57	0,53
23	Zurechtkommen mit Arbeit	0,51	0,42
24	Belastbarkeit	0,54	0,40
25	Auskommen Mitmenschen	0,56	0,70
26	Leben mit Einschränkungen	0,63	0,57
27	Ausgeglichenheit	0,59	0,55
	Gemittelter Wert	**0,55**	**0,56**

Abbildung 16. Item-Mittelwerte der Skala EMEK_27a und EMEK_27b im Vergleich

Vergleicht man die beiden Varianten hinsichtlich der Rangfolge der 27 Item-Mittelwerte (Abbildung 16), so ergibt sich folgendes Bild: Die drei Einzelaspekte „Lebensqualität", „Selbsthilfe" und „Umgang mit Alltagsbelastungen" liegen bei beiden Berechnungsvarianten auf den vorderen drei Plätzen. Bei anderen Items gibt es hingegen deutliche Unterschiede bei der Rangfolge zwischen Variante a und b, so z.B. bei dem Item „Gesundheitsbewusste Lebensführung". Aus der unterschiedlichen Rangfolge der Mittelwerte wird

deutlich, dass bei Aggregation aller 27 Items EMEK_27a und EMEK_27b inhaltlich nicht vollständig äquivalent sind, sondern je nach Abschneiden eines bestimmten Patienten auf den singulären Bewertungskriterien und Größe der Item-Streuung bestimmte Aspekte bei der Skalenberechnung mehr in den Vordergrund treten können als andere.

3.2.3 Skalenkennwerte von EMEK_27

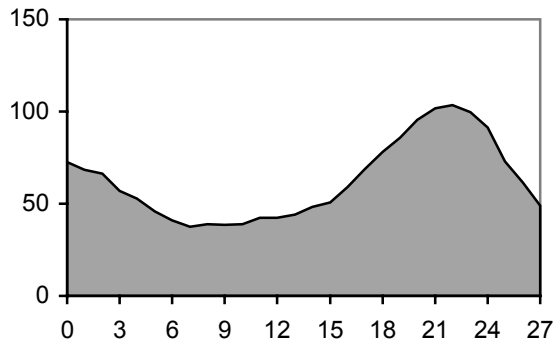
Berechnet man aus den 27 Einzelkomponenten der Variante a die Gesamtskala EMEK_27a, so reicht der mögliche Skalenrange von 0 bis 27 Punkte. 0 Punkte bedeuten, dass auf keinem der Einzelaspekte eine positive Veränderung berichtet wird, 27 Punkte bedeuten, dass der betreffende Patient auf sämtlichen Einzelkomponenten eine positive Entwicklung wahrnimmt. Der Mittelwert von EMEK_27a liegt bei 14,72, die Streuung beträgt 8,24. Die interne Konsistenz der Skala erreicht für die Gesamtstichprobe der 1.738 Patienten ein Cronbachs Alpha von .94.

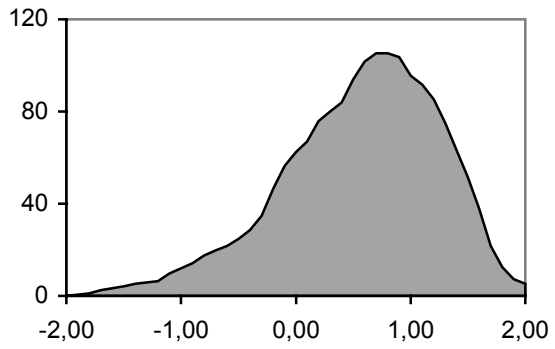
Die Häufigkeitsverteilung der Skala EMEK_27a (linker Teil von Abbildung 17) weist ein größeres Maximum im Bereich von 18-24 Punkten und eine zweites, kleineres Maximum im Bereich von 0-4 Punkten auf. Dies bedeutet, dass neben einer vergleichsweise großen Gruppe von Patienten, die in sehr vielen Bereichen positive Entwicklungen berichten können, eine zweite Gruppe von Patienten existiert, die nur in sehr wenigen Bereichen positive Veränderungen angeben. Die Häufigkeitsverteilung ist linksschief mit einer Skewness von -0,42, hat eine negative Wölbung mit einer Kurtosis von -1,19 und weicht von der Normalverteilung deutlich ab (K-S-Anpassungstest: $Z = 5{,}80$, $p<.001$). Die Verteilung weist sowohl einen Bodeneffekt als auch einen Deckeneffekt auf. Diese bimodale Verteilungsform ist für EMEK_27a charakteristisch und in allen fünf Programmevaluationsstudien beobachtbar, mit Ausnahme von Studie C, wo das zweite Maximum im Bereich von 0-4 Punkten nur wenig ausgeprägt ist. Die Häufigkeitsverteilungen für alle fünf Studien sind im Anhang 8.1.1 wiedergegeben.

Summiert man die 27 standardisierten Items der Variante b auf und dividiert man das Ergebnis durch 27, so resultiert ein Skalen-Mittelwert von 0,56. Die Skalen-Streuung beträgt 0,67 und ist damit deutlich kleiner als bei den Items, die aufgrund der vorangegangenen Standardisierung durchgängig eine Streuung von 1,00 aufweisen. Hier zeigt sich das Aggregationsprinzip nach Wittmann und Matt (1986), wonach die systematische Varianz bei einer Aggregation mehrerer zum gleichen Konstrukt gehörigen Items stärker zunimmt als die Fehlervarianz, was mit einer entsprechenden Reliabilitätserhöhung und geringeren Streuung der Gesamtskala korrespondiert.

Die Verteilungsform von EMEK_27b (rechter Teil von Abbildung 17) entspricht eher der Normalverteilung als die von EMEK_27a. Richtung und Größe der Veränderungen können der Grafik direkt entnommen und im Sinne von z-Standardwerten (bzw. in Analogie zur Effektgrößenmetrik) interpretiert werden. Ein Großteil der Patienten berichtet deutliche Verbesserungen. Einige Patienten haben im Durchschnitt hingegen kaum profitiert bzw. sich in der Summe auch verschlechtert und weisen damit ein besonders kritisches katamnestisches Gesamtbild auf. Die Wölbung ist mit einem Kurtosis-Wert von +0,19 im

Gegensatz zur Skala EMEK_27a nicht mehr auffällig, allerdings besteht nach wie vor eine deutliche Linksschiefe mit einem Skewness-Wert von –0,64 (K-S-Anpassungstest: Z = 2,28; p<.001). Cronbachs Alpha von EMEK_27b beträgt .95 und ist damit kaum höher als bei EMEK_27a. Im Gegensatz zur Variante a ist bei Variante b kein Boden- und Deckeneffekt mehr festzustellen.





Skala EMEK_27a	**Skala EMEK_27b**
N = 1.738 Patienten	*N* = 1.738
Min = 0; *Max* = 27	*Min* = -1,74; *Max* = 1,95
M = 14,72; *SD* = 8,24; *SE* = 0,20	*M* = 0,56; *SD* = 0,67; *SE* = 0,02
Schiefe = -0,42; Wölbung = -1,19	Schiefe = -0,64; Wölbung = +0,19
Cronbachs Alpha = .94	Cronbachs Alpha = .95

Abbildung 17. Häufigkeitsverteilung und Skalenkennwerte von EMEK_27a und EMEK_27b. M – Mittelwert, *SD* – Streuung, SE – Standardfehler.

Die Korrelation zwischen EMEK_27a und EMEK_27b beträgt .925, was einem gemeinsamen Varianzanteil von 85,6% (r^2 = .856) entspricht. Abbildung 18 veranschaulicht den Zusammenhang zwischen beiden Skalen in Form eines Scatterplots. Vor allem der Bodeneffekt der Skala EMEK_27a ist deutlich sichtbar: Der Range der Skala EMEK_27b erstreckt sich bei einem Wert von EMEK_27a = 0 auf einen Bereich von 0,00 bis –1,95. Bei mittleren Ausprägungen von EMEK_27a streuen die Werte von EMEK_27b hingegen weniger stark. Bei höheren Werten von EMEK_27a nimmt die Streubreite der Werte von EMEK_27b wieder zu.

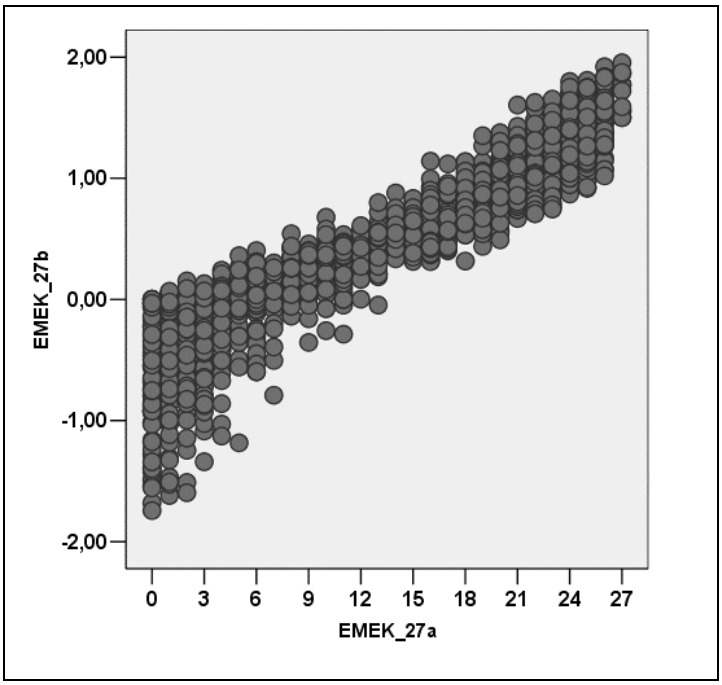


Abbildung 18. Scatterplot von EMEK_27a und EMEK_27b (*N* = 1.738).

3.2.4 Dimensionale Struktur von EMEK_27

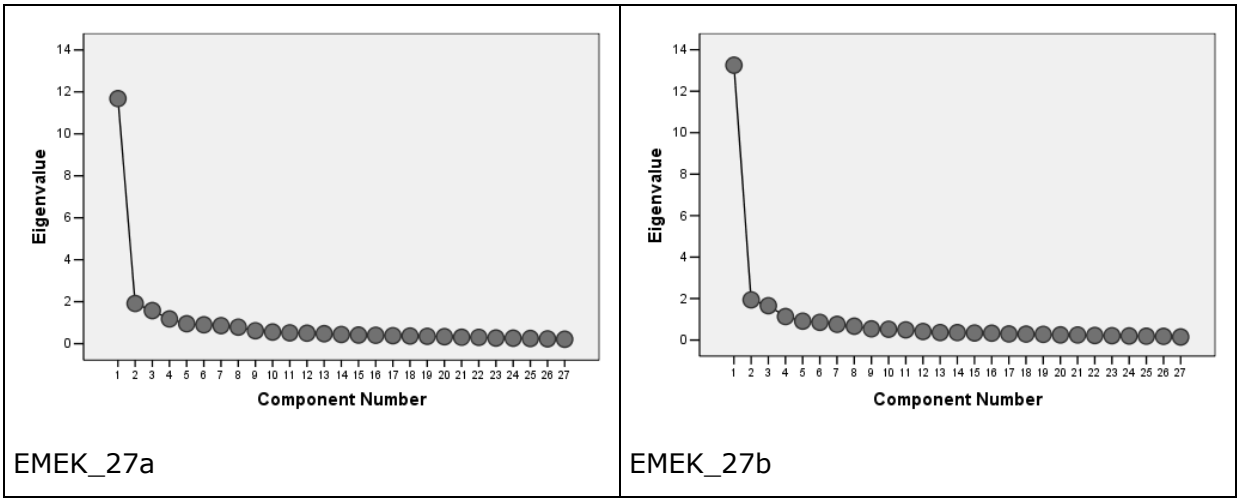


Abbildung 19. Faktorenanalyse von EMEK_27: Screeplot der unrotierten Faktorenlösung

Zur Überprüfung der inhaltlichen Struktur von EMEK_27 wurde eine Hauptkomponenten-analyse der 27 singulären Ergebniskriterien durchgeführt. Abbildung 19 gibt den Screeplot für Variante a und b vergleichend wieder. Zunächst ist festzustellen, dass sich ein deutlicher Generalfaktor zeigt, so dass die Bildung einer entsprechenden Gesamtskala aus allen 27 Items sinnvoll ist. Sowohl bei Variante a als auch bei Variante b haben vier Faktoren einen Eigenwert von größer als Eins (Kaiser-Kriterium) und der Screeplot lässt Anhaltspunkte dafür erkennen, dass eine Extraktion von drei bis vier Faktoren trotz der varianzstarken ersten Komponente vielversprechend erscheint.

Tabelle 16. Faktorenanalyse von EMEK_27 mit Varimax-Rotation (dreifaktorielle Lösung)

		Variante a				Variante b			
Nr	**EMEK_27-Komponente**	**F_I**	**F_{II}**	**F_{III}**	***h²***	**F_I**	**F_{II}**	**F_{III}**	***h²***
19	Wohlbefinden	**.82**	.15	.12	.71	**.85**	.21	.13	.79
5	Allgemeinbefinden	**.81**	.13	.16	.70	**.84**	.15	.18	.76
4	Seelisches Befinden	**.79**	.16	.14	.67	**.82**	.22	.14	.74
2	Lebensqualität	**.78**	.17	.13	.66	**.82**	.23	.17	.75
7	Beschwerden	**.77**	.10	.26	.67	**.80**	.10	.26	.73
8	Gesundheitszustand	**.76**	.08	.22	.63	**.80**	.11	.22	.70
27	Ausgeglichenheit	**.75**	.18	.02	.60	**.79**	.20	.08	.67
26	Leben mit Einschränkungen	**.74**	.20	.02	.59	**.80**	.22	.07	.69
6	Leistungsfähigkeit	**.74**	.07	.23	.60	**.80**	.09	.26	.72
9	Umgang mit Alltagsbelastungen	**.74**	.18	.07	.58	**.73**	.20	.09	.58
24	Belastbarkeit	**.72**	.15	.11	.56	**.80**	.12	.15	.67
1	Befinden zum Katamnese-Zeitpunkt	**.71**	.11	.18	.54	**.74**	.17	.20	.62
20	Umgang mit Problemen	**.71**	.17	.07	.53	**.74**	.19	.07	.59
22	Umgang mit Enttäuschungen	**.70**	.17	-.02	.51	**.74**	.18	.00	.58
3	Körperliches Befinden	**.69**	.10	.29	.56	**.75**	.08	.29	.65
21	Selbsthilfe	**.69**	.18	.04	.52	**.75**	.24	.05	.62
23	Zurechtkommen mit Arbeit	**.65**	.12	.09	.45	**.73**	.10	.12	.56
15	Arbeitsfähigkeit	**.60**	.12	.27	.45	**.67**	.08	.29	.54
25	Auskommen Mitmenschen	**.58**	.30	-.13	.45	**.63**	.28	-.04	.47
10	Gesundheitsbewusste Lebensführung	**.27**	.17	.05	.10	**.27**	.18	.00	.11
13	Beziehung zum Partner	.19	**.89**	.09	.84	.19	**.90**	.08	.85
14	Familienleben mit Kindern	.22	**.87**	.07	.80	.26	**.84**	.06	.78
12	Beziehungen zu Bezugspersonen	.27	**.84**	.07	.79	.26	**.85**	.09	.80
17	Krankschreibungszeiten (AU)	.18	.03	**.76**	.61	.18	.06	**.84**	.74
16	Arztbesuche	.18	.06	**.66**	.47	.16	.08	**.66**	.46
18	Krankenhaustage	-.06	-.01	**.56**	.32	-.04	-.04	**.68**	.46
11	Medikamentenkonsum	.16	.10	**.47**	.26	.23	.10	**.43**	.25
	Varianzaufklärung in %	38,3	10,3	7,6	56,2	43,1	10,7	8,7	62,4

Anmerkungen. *N* = 1.738 Patienten mit katamnestischer Messung aus 5 Studien. Die jeweils höchste Faktorladung in einer Zeile ist durch Fettdruck hervorgehoben und die Items wurden anhand der drei Faktoren gruppiert. *h²* - Kommunalität der Items (Summe der quadrierten Faktorladungen).

Zunächst wurde eine Extraktion und Varimax-Rotation von drei Faktoren durchgeführt. Tabelle 16 enthält die Ergebnisse. Die Varianzaufklärung der dreifaktoriellen Lösung beträgt bei Variante a insgesamt 56,2% und bei Variante b insgesamt 62,4%. Es liegt Einfachstruktur im Sinne von Thurstone (1954) vor, d.h. mit Ausnahme von Item 10 laden alle Items auf genau einem Faktor hoch und kein Item weist hohe Ladungen auf mehr als

einem Faktor auf. Ausgehend von den jeweils höchsten Ladungen lassen sich die drei Faktoren inhaltlich wie folgt charakterisieren:

- Faktor I: Veränderung des Allgemeinbefindens (AL)
- Faktor II: Veränderung der Beziehungen zu nahen Personen (BZ)
- Faktor III: Veränderung von kostenrelevanten Aspekten (KO)

Faktor I beinhaltet sowohl Veränderungen der gesundheitlichen Allgemeinverfassung als auch der Fähigkeit zur Alltagsbewältigung. Faktor II beinhaltet Veränderungen im Bereich wichtiger interpersonaler Beziehungen und steht damit für den Bereich der sozialen Unterstützung. Faktor III enthält kostenrelevante sozialmedizinische Aspekte wie Krankschreibungszeiten und Inanspruchnahme des Gesundheitssystems. Inhaltlich zeigen sich kaum Unterschiede zwischen den Ladungsmustern der Variante a und b, wenn man davon absieht, dass die Faktorladungen und Kommunalitäten bei Variante b meist etwas höher ausfallen als bei Variante a.

Eine getrennte Faktorenanalyse für die fünf Studien A, B, C, D und E führte zu dem Ergebnis, dass sich die dreidimensionale Lösung in allen Teilstichproben im wesentlichen gut replizieren lässt. Lediglich in Studie C zeigten sich bei Variante b Abweichungen dahingehend, dass die Items zur Veränderung des körperlichen Befindens und zur subjektiv empfundenen Leistungs- und Arbeitsfähigkeit die höchste Ladung auf der sozialmedizinisch relevanten Dimension aufweisen. Die rotierten Ladungsmuster für die fünf Teilstichproben sind hier aus Platzgründen nicht wiedergegeben, sondern finden sich im Anhang 8.1.2.

Eine Extraktion und anschließende Varimax-Rotation von vier Faktoren führte zu dem Ladungsmuster, das in Tabelle 17 abgebildet ist. Die Varianzaufklärung der vierfaktoriellen Lösung beträgt bei Variante a insgesamt 60,6% und bei Variante b insgesamt 66,6%. Ausgehend von den jeweils höchsten Faktorladungen lassen sich die vier Faktoren inhaltlich wie folgt charakterisieren:

- Faktor I: Veränderung des Gesundheitszustandes (GS)
- Faktor II: Veränderung der Fähigkeit zur Selbsthilfe (SH)
- Faktor III: Veränderung der Beziehungen zu nahen Personen (BZ)
- Faktor IV: Veränderung von kostenrelevanten Aspekten (KO)

Während Faktor I die Veränderung der Beschwerden und Symptomatik des Patienten sowie der subjektiv wahrgenommenen Leistungsfähigkeit abbildet, ist Faktor II eher bewältigungsorientiert und steht für die erfolgreiche Anwendung von psychosozialen Fertigkeiten im intra- und interpersonalen Bereich. Faktor III beinhaltet Veränderungen im Bereich naher Beziehungen (soziale Unterstützung). Faktor IV enthält schließlich die sozialmedizinisch relevanten Aspekte externer Stakeholder wie Arbeitgeber oder Kostenträger der Behandlung. Inhaltlich gesehen zeigen sich mit Ausnahme von Item 11 kaum Unterschiede zwischen den Ladungsmustern der Variante a und b, wenn man davon absieht, dass die Faktorladungen und Kommunalitäten bei Variante b wie bei der dreifaktoriellen Lösung auch hier höher ausfallen als bei Variante a.

Im Vergleich zur dreidimensionalen Lösung verteilt sich die ursprüngliche erste latente Dimension nun auf Faktor I und II, während der „Beziehungsfaktor" und der „sozialmedi-

zinische Faktor" nahezu unverändert als Faktor III und IV wieder auftauchen. Angesichts vieler Mehrfachladungen auf Faktor I und II liegt keine Einfachstruktur vor, d.h. Gesundheitszustand und Fähigkeit zur Selbsthilfe sind eng miteinander assoziiert.

Tabelle 17. Faktorenanalyse von EMEK_27 mit Varimax-Rotation (vierfaktorielle Lösung)

		Variante a					Variante b				
Nr	**EMEK_27-Komponente**	**F_I**	**F_{II}**	**F_{III}**	**F_{IV}**	***h²***	**F_I**	**F_{II}**	**F_{III}**	**F_{IV}**	***h²***
7	Beschwerden	**.79**	.27	.12	.14	.73	**.82**	.30	.12	.13	.79
5	Allgemeinbefinden	**.77**	.35	.14	.06	.74	**.79**	.38	.16	.06	.80
4	Seelisches Befinden	**.75**	.34	.17	.04	.71	**.76**	.39	.22	.03	.77
3	Körperliches Befinden	**.74**	.20	.13	.17	.64	**.81**	.22	.11	.14	.74
8	Gesundheitszustand	**.74**	.31	.09	.12	.67	**.80**	.31	.12	.09	.76
2	Lebensqualität	**.73**	.36	.17	.04	.69	**.72**	.44	.22	.07	.76
6	Leistungsfähigkeit	**.72**	.30	.09	.14	.64	**.79**	.32	.10	.14	.76
1	Befinden bei Katamnese	**.69**	.29	.12	.09	.58	**.70**	.35	.17	.10	.65
19	Wohlbefinden	**.69**	.47	.14	.06	.71	**.68**	.54	.18	.07	.79
9	Umgang mit Alltagsbelastungen	**.56**	.49	.15	.04	.58	**.54**	.51	.17	.05	.58
15	Arbeitsfähigkeit	**.55**	.30	.12	.22	.45	**.62**	.33	.08	.21	.55
22	Umgang mit Enttäuschungen	.31	**.72**	.09	.05	.62	.38	**.72**	.10	.03	.67
21	Selbsthilfe	.32	**.71**	.11	.11	.63	.39	**.73**	.16	.09	.71
25	Auskommen Mitmenschen	.18	**.71**	.21	-.04	.58	.25	**.70**	.20	.02	.59
26	Leben mit Einschränkungen	.39	**.70**	.13	.07	.66	.48	**.69**	.16	.07	.74
20	Umgang mit Problemen	.40	**.63**	.11	.10	.58	.44	**.65**	.13	.08	.64
27	Ausgeglichenheit	.48	**.61**	.13	.03	.62	.51	**.63**	.15	.07	.69
24	Belastbarkeit	.48	**.57**	.10	.12	.58	.56	**.59**	.08	.12	.69
23	Zurechtkommen mit Arbeit	.39	**.55**	.07	.12	.48	.49	**.57**	.05	.11	.58
10	Gesundheitsbewusstes Leben	.10	**.31**	.14	.08	.13	.03	**.40**	.12	.06	.18
13	Beziehung zum Partner	.16	.16	**.89**	.05	.85	.16	.16	**.90**	.04	.86
14	Familienleben	.16	.20	**.86**	.04	.81	.20	.22	**.83**	.02	.79
12	Beziehungen zu Bezugspersonen	.19	.24	**.83**	.05	.79	.18	.24	**.84**	.06	.80
17	Krankschreibungszeiten	.16	.13	.01	**.79**	.67	.22	.11	.05	**.84**	.77
16	Arztbesuche	.20	.08	.06	**.66**	.48	.21	.07	.08	**.65**	.47
18	Krankenhaustage	-.11	.08	-.04	**.63**	.42	-.06	.09	-.06	**.74**	.57
11	Medikamentenkonsum	.28	-.05	.13	**.42**	.27	**.42**	-.10	.15	.32	.31
	Varianzaufklärung in %	25,7	18,4	9,6	6,9	60,6	29,2	20,2	9,9	7,3	66,6

Anmerkungen. *N* = 1.738 Patienten mit katamnestischer Messung aus 5 Studien. Die 27 Items sind absteigend nach Höhe der Faktorenladungen (Variante a) geordnet. Die jeweils höchste Faktorladung der Items ist durch Fettdruck hervorgehoben. *h²* - Kommunalität der Items (Summe der quadrierten Faktorladungen).

Führt man eine getrennte Faktorenanalyse mit Extraktion von vier Faktoren für die fünf Teilstichproben durch, so ergibt sich folgendes Bild: Die Faktoren III (Beziehungen) und

IV (sozialmedizinische Aspekte) lassen sich in allen fünf Studien replizieren, während es bei Faktor I und II vor allem bei den Teilstichproben D und E deutliche Abweichungen im Ladungsmuster gibt. Auch die entsprechenden Ladungsmuster der vierfaktoriellen Lösung für die fünf Teilstichproben finden sich im Anhang 8.1.2.

Ausgehend von den Ergebnissen der Faktorenanalysen wurden jeweils fünf Subskalen für die beiden Varianten a und b berechnet. Die Items mit der jeweils höchsten Ladung auf einem bestimmten Faktor wurden zu einer entsprechenden Subskala zusammengefasst. Einzige Ausnahme ist Item 11 „Medikamentenkonsum" bei der vierfaktoriellen Lösung von Variante b, das durchgängig der Skala KO zugeordnet wurde. Tabelle 18 gibt die Einzelheiten zur Skalenbildung wieder. Tabelle 19 enthält die Kennwerte der Subskalen für Varianten a und b im Vergleich. Alle Subskalen mit Ausnahme der Skala KO weisen gute bis ausgezeichnete interne Konsistenzen auf. Die Skalenmittelwerte der Variante b bewegen sich in einem Bereich zwischen 0,32 (Skala KO) und 0,64 (Skala GS).

Tabelle 18. Alle Skalen des multiplen Ergebniskriteriums EMEK_27

Kürzel	Skalenbezeichnung	Items
27	**Multiples Ergebniskriterium** (entspricht der aus allen 27 Items gebildeten Gesamtskala des Globalfaktors)	1 – 27
AL	**Veränderung des Allgemeinbefindens** (entspricht dem 1. Faktor der dreidimensionalen Lösung)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 19, 20, 21, 22, 23, 24, 25, 26, 27
GS	**Veränderung des Gesundheitszustandes** (entspricht dem 1. Faktor der vierdimensionalen Lösung)	1, 2, 3, 4, 5, 6, 7, 8, 9, 15, 19
SH	**Veränderung der Fähigkeit zur Selbsthilfe** (entspricht dem 2. Faktor der vierdimensionalen Lösung)	10, 20, 21, 22, 23, 24, 25, 26, 27
BZ	**Veränderung der Beziehungen zu nahen Personen** (entspricht dem 3. Faktor der vierdimensionalen Lösung)	12, 13, 14
KO	**Veränderung kostenrelevanter Aspekte** (entspricht dem 4. Faktor der vierdimensionalen Lösung)	11, 16, 17, 18

Tabelle 19. Kennwerte aller Skalen des multiplen Ergebniskriteriums EMEK_27

	Variante a						*Variante b*				
	Min	***Max***	***M***	***SD***	α	***Min***	***Max***	***M***	***SD***	α
EMEK_27	0	27	14,72	8,24	.94	-1,74	1,95	0,56	0,67	.95
EMEK_AL	0	20	11,93	7,04	.95	-1,97	2,01	0,62	0,77	.97
EMEK_GS	0	11	6,77	4,25	.94	-1,80	1,80	0,64	0,84	.96
EMEK_SH	0	9	5,16	3,16	.88	-2,27	2,27	0,59	0,76	.91
EMEK_BZ	0	3	1,18	1,32	.89	-1,59	1,59	0,47	0,90	.89
EMEK_KO	0	4	1,61	1,26	.56	-2,41	3,01	0,32	0,69	.63

N = 1.738 Patienten mit katamnestischer Messung.

Mit den Subskalen der Variante b lassen sich entsprechende „Erfolgsprofile" für einzelne Patienten bzw. Gruppen von Patienten erstellen. Zwischen den Hauptdiagnosegruppen zeigen sich keine statistisch signifikanten Unterschiede auf den Skalen, mit einer Ausnahme: Patienten mit der Hauptdiagnose „Somatoforme Störung" berichten im Vergleich zu den übrigen Patienten auf der Skala BZ eine weniger starke Verbesserung ihrer mitmenschlichen Beziehungen, dafür aber eine stärkere Verbesserung auf der Skala KO (und damit auf der für diese Gruppe besonders relevanten Dimension kostenrelevanter Aspekte wie Inanspruchnahme des Gesundheitssystems und Krankschreibungszeiten).

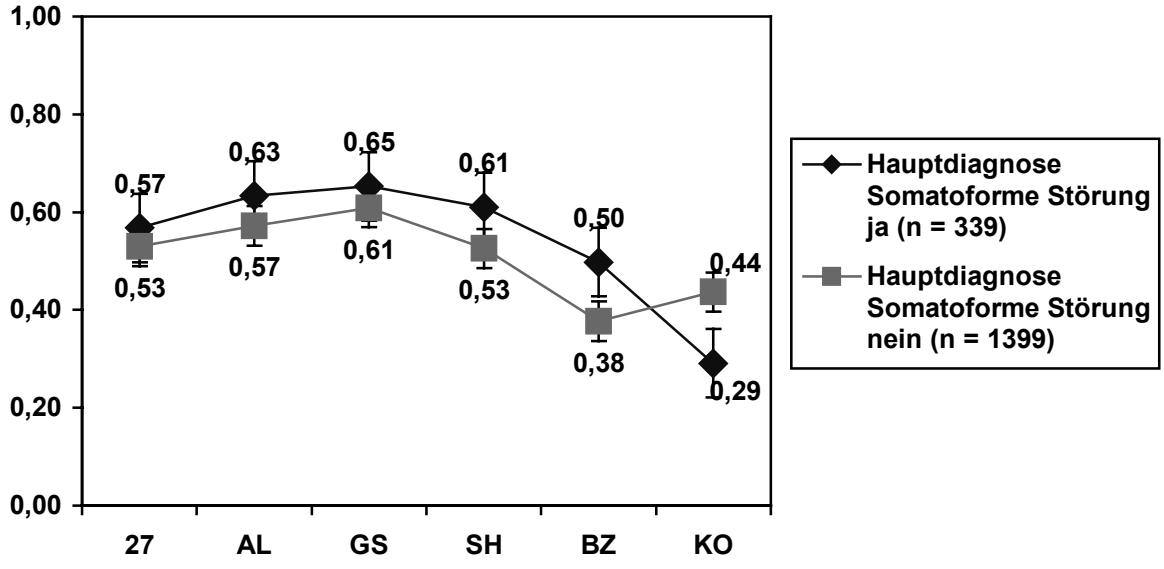


Abbildung 20. Skalenprofil der EMEK-Variante b in Abhängkeit von der Hauptdiagnose. Die Mittelwertsunterschiede auf den Skalen BZ (p=.028) und KO (p<.001) sind statistisch signifikant (95%-Konfidenzintervalle sind eingezeichnet).

Tabelle 20 enthält die Korrelationen zwischen den Subskalen. Am höchsten fällt der Zusammenhang zwischen den Skalen GS und SH mit Werten > .80 aus, was angesichts des großen Anteils an Doppelladungen bei der vierfaktoriellen Lösung auf den ersten beiden Dimensionen zu erwarten war. Die übrigen Korrelationen erreichen moderate Höhe, lediglich der Zusammenhang zwischen den Skalen KO und BZ ist mit Werten <.20 nur schwach ausgeprägt.

Tabelle 20. Interkorrelationen zwischen den Subskalen von Skala EMEK_27

Variante a					Variante b				
	GS	SH	BZ	KO		GS	SH	BZ	KO
AL			.45	.34	AL			.47	.37
GS	.	.80	.42	.35	GS	.	.84	.45	.39
SH		.	.44	.29	SH		.	.46	.31
BZ			.	.17	BZ			.	.17

Anmerkung. Alle Korrelationen zwischen den Skalen sind statistisch signifikant (p<.001). Zwischen AL und den beiden Skalen GS bzw. SH wurden keine Korrelationen berechnet, da deren Items in AL enthalten sind. *N* = 1.738 Patienten mit katamnestischer Messung.

3.2.5 Vorhersage von EMEK_27 aus Stichprobenmerkmalen

Zur Aufklärung unterschiedlicher Skalenausprägungen auf der Skala EMEK_27 lassen sich zwei Gruppen von Prädiktoren unterscheiden:

- Stichprobenzugehörigkeit (Studie A, B, C, D oder E)
- Patientenmerkmale („Case-Mix", z.B. Alter, Geschlecht und Schulabschluss)

Zunächst wurde die Stichprobenzugehörigkeit anhand von vier Dummy-Variablen codiert. Für die aufgeklärte Varianz spielt es keine Rolle, welche der fünf Studien „gegen" die übrigen vier getestet wird, da die Dummy-Variablen zusammengenommen eine Prädiktorvariable bilden. Hier wurde die Stichprobe mit dem niedrigsten EMEK_27-Mittelwert (Studie E) ausgewählt.

Tabelle 21 gibt die Ergebnisse zum multiplen Zusammenhang für beide Skalenvarianten im Detail wieder. Tabelle 22 zeigt die Partialgewichte der beiden Regressionsgleichungen bei simultaner Einbeziehung aller Prädiktoren.

Bezieht man zunächst nur die Stichprobenzugehörigkeit in die Regressionsgleichung ein, so wird bei EMEK_27a ein signifikanter Varianzanteil von 3,3% am Kriterium aufgeklärt (p<.001). Nimmt man die Patientenmerkmale (vgl. Abschnitt 3.1.3) hinzu, so steigt die aufgeklärte Varianz um 6,7% auf insgesamt 10,0% an. Auch dieser inkrementelle Zuwachs ist statistisch signifikant (p<.001). Beginnt man die Analyse hingegen mit den Patientenmerkmalen, so werden hierdurch bereits 7,7% von der Gesamtvarianz aufgeklärt (p<.001), der inkrementelle Beitrag der Stichprobenzugehörigkeit beträgt dann 2,3% (p<.001). Damit liefern beide Varianzquellen einen eigenen inkrementellen Beitrag, wobei der deutlich größere Teil auf die Patientenmerkmale entfällt.

Tabelle 21. Multiple Regression von EMEK_27: Aufgeklärte Varianz

	EMEK_27a					EMEK_27b			
Prädiktorblock und inkrementeller Varianzzuwachs	***R***	***R²***	***R² adj***	***R² diff***		***R***	***R²***	***R² adj***	***R² diff***
1: Nur Stichprobenzugehörigkeit	.183	.033	.031	.033		.163	.027	.024	.027
1: Stichprobenzugehörigkeit und 2: Patientenmerkmale	.317	.100	.092	.067		.301	.091	.082	.064
2: Nur Patientenmerkmale	.277	.077	.070	.077		.264	.070	.063	.070
2: Patientenmerkmale und 1: Stichprobenzugehörigkeit	.317	.100	.092	.023		.301	.091	.082	.021

Anmerkungen. N = 1.738 Patienten mit katamnestischer Messung aus 5 Studien. R^2_{diff}: Inkrementeller Varianzzuwachs bei Hinzunahme des jeweiligen Prädiktorblocks. Alle R^2_{diff} sind statistisch signifikant (p<.001).

In Prädiktorblock 1 (Stichprobenzugehörigkeit) fällt das Beta-Gewicht von .19 (p<.001) bei Stichprobe C ins Auge. Das zugehörige B-Gewicht von 5,11 bedeutet inhaltlich, dass – selbst bei Konstanthaltung aller in Prädiktorblock 2 einbezogenen Patientenmerkmale - in Stichprobe C ein um 5,11 Punkte höherer EMEK-Wert zu erwarten ist als in Stichprobe E (gegen die hier getestet wird, weshalb für Studie E keine Dummy-Variable vorhanden ist). Die Mittelwertsunterschiede auf der Skala EMEK_27a zwischen Studie C und E lassen sich also nicht nur einfach auf Unterschiede in der Zusammensetzung der Patientenstichproben zurückführen, sondern in Studie C wird offenbar tatsächlich eine höhere Ergebnisqualität berichtet). Natürlich ist nicht auszuschließen, dass weitere (hier nicht einbezogene) Patientenmerkmale existieren könnten, die für die unterschiedliche Ausprägung von EMEK_27 in den fünf Studien verantwortlich sind. Ebenso könnten aber auch behandlungsbezogene Merkmale für die höhere Ergebnisqualität verantwortlich sein. Auch Studie A (β = .08, p = .020) und D (β = .13, p = .001) schneiden statistisch signifikant „besser" auf der Skala EMEK_27 ab, Studie B unterscheidet sich diesbezüglich hingegen nicht statistisch signifikant von Studie E (β = .05, p = .147).

In Prädiktorblock 2 (Patientenmerkmale) finden sich ebenfalls eine Reihe von signifikanten Partialgewichten. Weibliche, jüngere und höher gebildete Patienten erzielen höhere Werte auf der Skala EMEK_27a. Das Vorliegen eines Rentenantrages bei Aufnahme lässt hingegen ein schlechteres Ergebnis erwarten (B = -3,88 Punkte). Weitere negative Prädiktoren für den Behandlungserfolg sind eine lange Krankheitsdauer, das Vorhandensein von Nebendiagnosen (Komorbidität) sowie ein vorzeitiges Therapieende (Behandlungsabbruch). Patienten der gesetzlichen Rentenversicherung schneiden etwas schlechter (B = -1,18) ab als Patienten, bei denen eine private oder gesetzliche Krankenkasse die Kosten für die Behandlung trägt. Keine statistisch signifikanten Partialzusammenhänge von EMEK_27a bestehen mit Familienstand, Erwerbsstatus, Hauptdiagnose und Behandlungsdauer.

Bei der Skala EMEK_27b (rechter Teil von Tabelle 21 und Tabelle 22) ergibt sich sowohl beim multiplen Gesamtzusammenhang als auch bei den Beträgen der Beta-Partialgewichte ein ähnliches Bild wie bei der Skala EMEK_27a. Die Gesamtvarianzaufklärung beträgt hier 9,1%. An dieser Stelle sei noch erwähnt, dass sich die B-Gewichte bei EMEK_27b direkt im Sinne von z-Werten bzw. in Analogie zur Effektgrößenmetrik interpretieren lassen. So lässt das Vorliegen eines Rentenantrages ein schlechteres Ergebnis auf der Skala EMEK_27b erwarten (*B* = -0,304 Standardabweichungseinheiten). Auch ein Therapieabbruch ist mit einer kritischeren Prognose hinsichtlich des in der 1-Jahres-Katamnese zu erwartenden multiplen Gesamtbildes verbunden (*B* = -0,224).

Tabelle 22. Multiple Regression von EMEK_27: Partialgewichte der Prädiktoren

	EMEK_27a						EMEK_27b				
Prädiktorvariable	***B***	***SE B***	***Beta***	***t***	***p***		***B***	***SE B***	***Beta***	***t***	***p***
Regressionskonstante	14,57	1,41		10,32	<.001		0,727	0,116		6,28	<.001
Prädiktorblock 1: Studie											
Zugehörigkeit zu Studie A	1,69	0,73	**.08**	2,33	.020		0,179	0,059	**.11**	3,01	.003
Zugehörigkeit zu Studie B	1,03	0,71	**.05**	1,45	.147		0,104	0,058	**.07**	1,79	.074
Zugehörigkeit zu Studie C	5,11	0,81	**.19**	6,34	<.001		0,398	0,066	**.19**	6,03	<.001
Zugehörigkeit zu Studie D	2,26	0,66	**.13**	3,43	.001		0,153	0,054	**.11**	2,85	.004
Prädiktorblock 2: Merkmale											
Geschlecht weiblich	1,89	0,39	**.11**	4,85	<.001		0,136	0,032	**.10**	4,27	<.001
Alter in Jahren	-0,05	0,02	**-.07**	-2,62	.009		-0,005	0,002	**-.07**	-2,89	.004
Schulbildung (1=HS, 2=RS, 3=Abi)	0,81	0,26	**.08**	3,18	.002		0,064	0,021	**.08**	3,06	.002
Familienstand verheiratet	0,39	0,41	**.02**	0,95	.344		0,015	0,034	**.01**	0,44	.663
Erwerbstätig bei Aufnahme	0,22	0,45	**.01**	0,49	.622		0,025	0,037	**.02**	0,67	.503
Kostenträger (0=KV, 1=RV)	-1,18	0,46	**-.07**	-2,55	.011		-0,120	0,038	**-.08**	-3,17	.002
Rentenantrag bei Aufnahme	-3,88	0,97	**-.09**	-4,02	<.001		-0,304	0,079	**-.09**	-3,85	<.001
Krankheitsdauer in Jahren	-0,18	0,04	**-.10**	-4,17	<.001		-0,015	0,003	**-.11**	-4,39	<.001
Somatoforme Hauptdiagnose	0,64	0,52	**.03**	1,23	.219		0,021	0,043	**.01**	0,49	.623
Nebendiagnose(n) vorhanden	-1,01	0,44	**-.05**	-2,31	.021		-0,072	0,036	**-.05**	-2,01	.044
Behandlungsdauer in Tagen	0,01	0,01	**.02**	0,85	.393		-0,002	0,001	**-.04**	-1,76	.078
Vorzeitige Entlassung	-2,58	0,90	**-.07**	-2,88	.004		-0,224	0,073	**-.07**	-3,05	.002

Anmerkungen. *N* = 1.738 Patienten mit katamnestischer Messung aus 5 Studien.
SE B: Standardfehler der B-Gewichte. Codierung der Prädiktoren: Wenn keine weitere Erläuterung vorhanden ist, dann ist das Vorhandensein des betreffenden Merkmals mit 1, das Nichtvorhandensein hingegen mit 0 codiert. KV – Krankenversicherung, RV – Rentenversicherung. HS – bis Hauptschule, RS – mittlere Reife, Abi – (Fach)-Abitur.

Insgesamt fallen sowohl die multiplen Zusammenhänge als auch die Partialgewichte relativ klein aus und sollten trotz statistischer Signifikanz angesichts der großen Stichprobe nicht überinterpretiert werden. Berechnet man die Regressionsgleichungen unter Einbeziehung der Patientenmerkmale getrennt für alle fünf Teilstichproben (vgl. Anhang 8.1.3), so lässt sich am ehesten noch das Vorliegen eines Rentenantrages zum Aufnahme-Zeitpunkt bei vier der fünf Teilstichproben fast durchgängig als negativer Prädiktor

für das katamnestische Behandlungsergebnis identifizieren. Bei den übrigen Patientenmerkmalen zeigen sich weniger einheitliche Ergebnisse.

3.2.6 Vorhersage von EMEK_27 aus Prozessmerkmalen

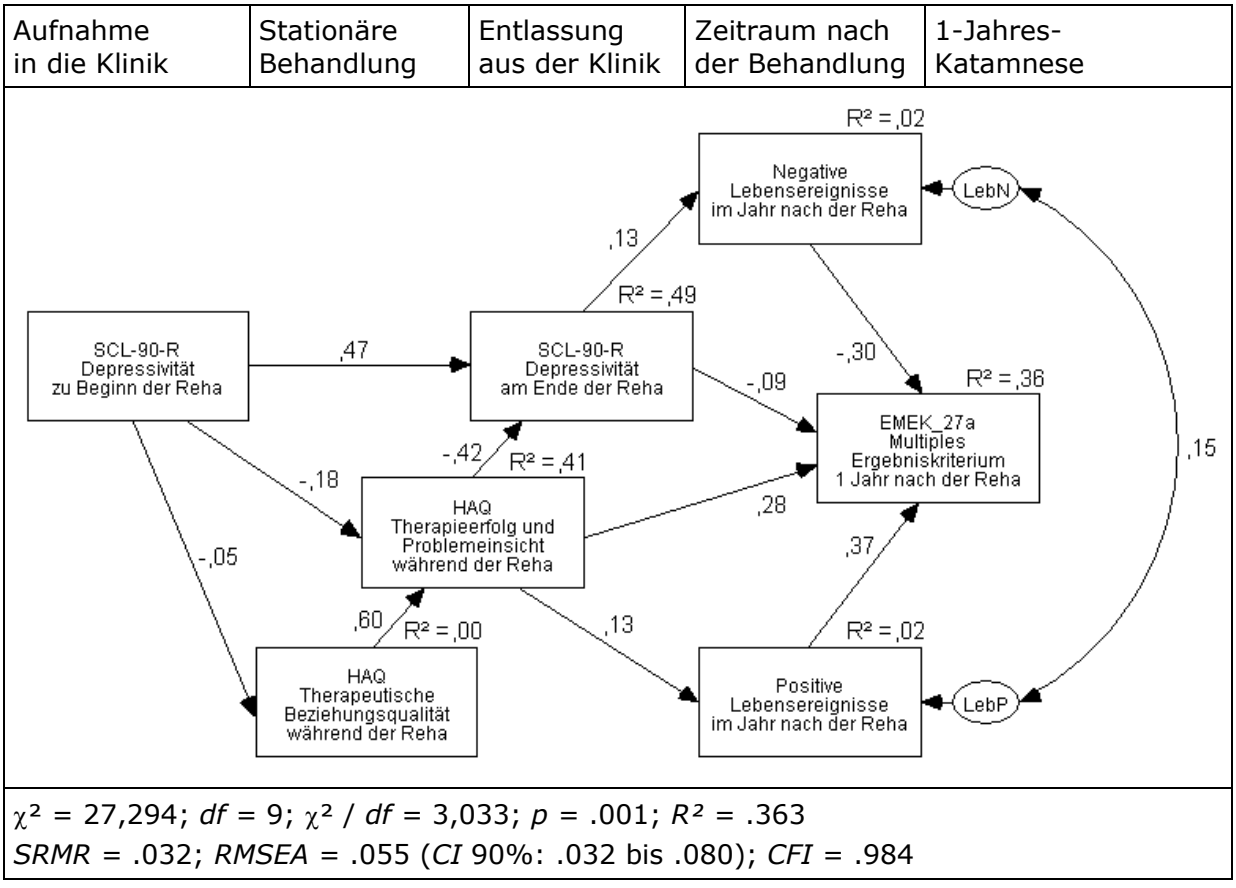
Mit dem Helping Alliance Questionaire (HAQ, Bassler, Potratz & Krauthauser, 1995) wurden in zwei der zur Reanalyse von EMEK_27 verwendeten Studien (Studie C und Studie D, vgl. Abschnitt 3.1.3) mit den beiden Skalen Beziehungs- und Erfolgszufriedenheit zwei behandlungsbezogene Aspekte zum Entlass-Zeitpunkt erhoben. Die Skala „Beziehungszufriedenheit" bezieht sich auf wichtige Teilaspekte der Prozessqualität zu Beginn und im weiteren Verlauf der Behandlung. Sie steht für die Etablierung einer vertrauensvollen therapeutischen Arbeitsbeziehung, die von Offenheit und Empathie geprägt ist (Beispielitem: Ich habe das Gefühl, dass ich mich auf meinen Therapeuten verlassen konnte). Diese vertrauensvolle Beziehung stellt eine wichtige Voraussetzung zur Problemeinsicht und therapeutischen Modifikation von affektiven, kognitiven und behavioralen Mustern dar, was im HAQ mit der Skala „Erfolgszufriedenheit" erfasst wird (Beispielitem: Ich kann absehen, dass ich die Probleme vielleicht bewältigen kann, wegen derer ich in die Behandlung gekommen bin). Die Skala „Erfolgszufriedenheit" knüpft somit zeitlich bei der Skala „Beziehungszufriedenheit" an und steht für Teilaspekte der Prozessqualität, die mit dem weiteren Therapieverlauf bis hin zum unmittelbaren Therapieergebnis bei Entlassung verbunden sind. Damit können die beiden vom HAQ erfassten Konstrukte als Bindeglied zwischen bestimmten Aspekten der Prozessqualität und der unmittelbaren Ergebnisqualität bei Therapieende verstanden werden.

Ein anderer wichtiger Aspekt ist das Ausmaß an Demoralisierung, welche ein Patient zu Beginn und am Ende der Behandlung berichtet. Wie in Abschnitt 3.1.3 expliziert, kommt ein erheblicher Anteil der Patienten mit einer depressiven Symptomatik in die stationäre psychosomatische Rehabilitation. Das Konzept der Demoralisierung ist mit dem Konstrukt der Depressivität verwandt und steht für Hoffnungslosigkeit und fehlendes Vertrauen in die eigenen Fähigkeiten (Frank, 1992). Demoralisierung zeichnet sich durch ein subjektiv erfahrenes Ungleichgewicht zwischen den Anforderungen der Umwelt und den eigenen Bewältigungsmöglichkeiten aus. Die Erfahrung, alltägliche Probleme nicht oder nur unzureichend lösen zu können, führt zu Distress, mangelndem Selbstvertrauen, negativen Kognitionen und zu psychischer Belastung. Eine erfolgreiche psychotherapeutische Behandlung zielt daher auch immer auf eine Remoralisierung durch die Vermittlung von Hoffnung, Aktivierung von Ressourcen und Vermittlung von Problemlösekompetenzen beim Patienten ab. Zur Abbildung der Demoralisierung zu Therapiebeginn und am Ende der Behandlung lässt sich in Studie C und D die Skala „Depressivität" aus der Symptom-Checkliste SCL-90-R (Franke, 1995, 2002) verwenden. So erfassen die Items der Skala „Depressivität" im SCL-90-R unter anderem auch Teilaspekte des Konstruktes der Demoralisierung wie Item 53 „Gefühl der Hoffnungslosigkeit angesichts der eigenen Zukunft" oder Item 71 „Gefühl, dass alles sehr anstrengend ist". Der Skala „Depressivität" aus der SCL-90-R wurde hier der Skala „Hoffnung" aus dem Fragebogen zur Psychotherapiemoti-

vation (FPTM, Nübling & Schulz, 2002) vorgezogen, da der FPTM in der EQUA-Studie nur zu Therapiebeginn erhoben wurde.

Zur Vorhersage der beiden Varianten von EMEK_27 wurde ein Pfadmodell entwickelt, wobei die Version von Wittmann et al. (2002) als Grundlage verwendet und leicht abgewandelt wurde. Neben dem Ausmaß an Depressivität zu Beginn des stationären Klinikaufenthaltes und bei Entlassung aus der Klinik sind die beiden zum E-Zeitpunkt erhobenen HAQ-Skalen Beziehungs- und Erfolgszufriedenheit in dem Modell enthalten sowie das Eintreten von positiven bzw. negativen Lebensereignissen im Jahr nach Entlassung (zeitgleich mit dem Kriterium EMEK_27 zum Zeitpunkt der 1-Jahres-Katamnese erhoben).

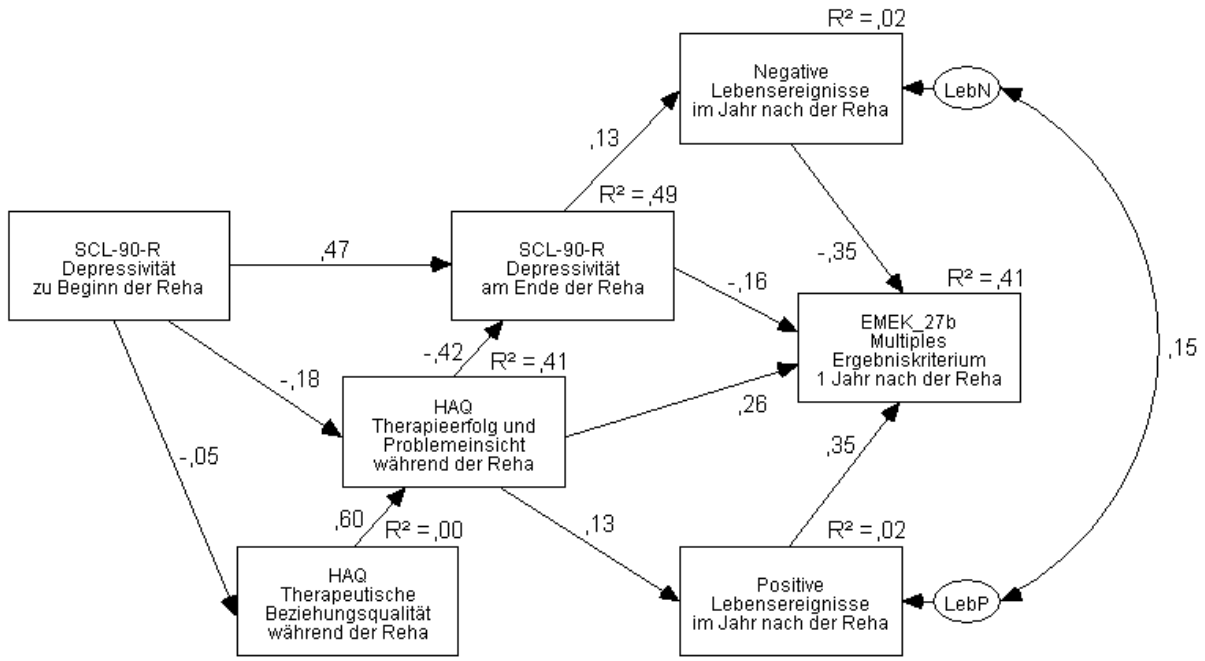
Der Modell-Fit wurde für beide Skalenvarianten EMEK_27a und EMEK_27b mit der Software AMOS 5 überprüft. Zusammengenommen stehen aus den Studien C und D für eine Stichprobe von insgesamt 664 Patienten neben den katamnestischen Daten auch entsprechende Messungen zum E-Zeitpunkt für den HAQ und SCL zur Verfügung. In Abbildung 21 ist das Pfadmodell für die Skala EMEK_27a und in Abbildung 22 das Modell für die Skala EMEK_27b wiedergegeben.



χ^2 = 27,294; *df* = 9; χ^2 / *df* = 3,033; *p* = .001; R^2 = .363
SRMR = .032; *RMSEA* = .055 (*CI* 90%: .032 bis .080); *CFI* = .984

Abbildung 21. Pfadmodell zur Vorhersage von EMEK_27a. *N* = 664 Patienten.

Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 25,459; *df* = 9; χ^2 / *df* = 2,829; *p* = .003; R^2 = .409
SRMR = .031; *RMSEA* = .053 (*CI* 90%: .029 bis .077); *CFI* = .986

Abbildung 22. Pfadmodell zur Vorhersage von EMEK_27b. *N* = 664 Patienten.

Die Modelle zur Vorhersage von EMEK_27a und EMEK_27b weisen ansonsten weitgehende inhaltliche Übereinstimmungen auf, nachfolgend wird das Pfadmodell für EMEK_27b (Abbildung 22, Pfadkoeffizienten jeweils in Klammern) interpretiert:

Erwartungsgemäß zeigt sich zunächst, dass der Grad an Depressivität zu Beginn der Behandlung deutlich mit dem Grad an Depressivität bei Entlassung (+.47) korrespondiert. Ein hohes Ausmaß an Depressivität zu Therapiebeginn wirkt sich zudem auf die mit dem HAQ gemessene Problemeinsicht (-.18) negativ aus. Die beiden Pfadkoeffizienten veranschaulichen somit den Einfluss des Schweregrades der Erkrankung auf das Therapiegeschehen und die Tatsache, dass bei einer ausgeprägten Symptomatik nicht so leicht eine vollständige Remission zu erzielen ist wie in weniger gravierenden Fällen.

Die Qualität der therapeutischen Beziehung übt erwartungsgemäß einen deutlichen Effekt auf die therapiebedingte Problemeinsicht (+.60) aus. Über die Etablierung einer vertrauensvollen therapeutischen Arbeitsbeziehung ist es somit am ehesten möglich, die Patienten zu erreichen, um im nächsten Schritt eine Modifikation von problematischen Denk- und Verhaltensmustern anzustreben. Hinsichtlich der klinischen Forschung und Praxis stellt sich damit die Frage nach der möglichst optimalen Zuweisung der Patienten zu „ihrem" Bezugstherapeuten („Matching").

Bemerkenswert ist, dass sich die Zufriedenheit mit der therapeutischen Beziehung als unabhängig vom Ausprägungsgrad der depressiven Symptomatik erwies. So erwies sich

der ursprünglich angenommene negative Einfluss der Depressivität bei Aufnahme auf die Qualität der therapeutischen Beziehung (-.05) als statistisch nicht signifikant.

Der Grad an therapeutisch erzielter Problemeinsicht reduziert die depressive Symptomatik (-.42), was auf eine erfolgreiche Modifikation von dysfunktionalen depressiven Denk- und Erlebensweisen schließen lässt. Sowohl Problemeinsicht (+.26) als auch die Reduktion der depressiven Symptomatik (-.16) stellen günstige Prädiktoren für eine hohe Ergebnisqualität in der 1-Jahres-Katamnese und damit den langfristigen Therapieerfolg dar.

Das Modell veranschaulicht allerdings auch die Grenzen des Einflusses der Rehabilitationsmaßnahme, wenn der Patient die Klinik wieder verlassen hat. So üben sowohl positive (+.35) als auch negative (-.35) Lebensereignisse im Jahr nach der Rehabilitation einen spürbaren Einfluss auf den langfristigen Behandlungserfolg aus. Mit diesem Befund werden Frage der Nachsorge zur Stabilisierung des Behandlungserfolges angeschnitten, etwa durch weitere Beratung oder ambulante Psychotherapie. Die Zahl der berichteten positiven und negativen Lebensereignisse korreliert schwach miteinander (r = .15), offenbar sind hier manche Patienten generell mitteilsamer als andere.

Zunächst nicht erwartet, aber nachträglich in das Modell aufgenommen wurden zwei weitere Pfade: Je höher das Ausmaß an Depressivität am Ende der Behandlung, desto mehr negative Lebensereignisse im Jahr nach Entlassung werden berichtet (+.13). Möglicherweise zeigt sich hier der depressive Attributionsstil, wonach bestimmte Ereignisse generell negativ verzerrt wahrgenommen und bewertet werden. Eine andere Interpretation könne darin bestehen, dass sich die Betroffenen durch depressive Verhaltensweisen tatsächlich vermehrt Schwierigkeiten im privaten und beruflichen Bereich schaffen. Darüber hinaus zeigte sich eine positive Relation zwischen der erzielten Problemeinsicht bei Ende der Behandlung und positiven Lebensereignissen im Jahr nach der Rehabilitation (+.13). Offenbar führt ein Zuwachs an Problemlösekompetenz dazu, dass die Patienten Verbesserungen ihrer privaten und beruflichen Lebenssituation initiieren und aufrechterhalten können, was wiederum zu einem vermehrten Auftreten von positiv bewerteten Lebensereignissen führt. Möglicherweise ist es aber auch die Korrektur von bestimmten Denkmustern, die zu einer positiveren Sichtweise und optimistischeren Bewertung wichtiger Vorkommnisse im privaten und beruflichen Leben führt.

Auch die in Abschnitt 3.2.4 aus dem Itempool Skala EMEK_27 berechneten Subskalen AL, GS, SH, BZ und KO für Variante a und b wurden einer entsprechenden Validierung unterzogen. Die Strukturgleichungsmodelle für beide Varianten der fünf Skalen mit allen Pfadkoeffizienten sind im Anhang 8.1.4 wiedergegeben. Tabelle 23 enthält eine Zusammenfassung der Kennwerte zum Modellfit. Im Vergleich zur Gesamtskala EMEK_27 ergeben sich kaum Abweichungen, der Modellfit ist durchgehend akzeptabel.

Tabelle 23. Kennwerte der Pfadmodelle zur Vorhersage der Subskalen von EMEK_27

	Variante a					Variante b				
Skala	p	χ^2	*SRMR*	*CFI*	*RMSEA*	p	χ^2	*SRMR*	*CFI*	*RMSEA*
EMEK_27	.001	27,294	.032	.984	.055	.003	25,459	.031	.986	.053
EMEK_AL	.005	23,563	.030	.987	.049	.005	23,740	.030	.987	.050
EMEK_GS	.001	26,625	.032	.983	.054	.002	26,042	.031	.985	.053
EMEK_SH	.022	19,456	.028	.990	.042	.015	20,536	.030	.990	.044
EMEK_BZ	.001	26,344	.032	.982	.054	.010	21,796	.029	.987	.046
EMEK_KO	.002	26,622	.030	.980	.054	.003	24,666	.029	.983	.051

Anmerkung. *N* = 664 Patienten aus den Studien C und D, bei denen entsprechende Messungen zum Entlass- und Katamnese-Zeitpunkt vorlagen. *df* = 9. χ^2 ist statistisch signifikant (p <.001) für alle 10 Modelle (grafische Wiedergabe im Anhang 8.1.4).

In Tabelle 24 sind neben der Gesamtvarianzaufklärung an der betreffenden EMEK-Skala auch alle Pfadkoeffizienten eingetragen, die auf die betreffende EMEK-Skala zeigen. Vergleicht man zunächst die beiden Berechnungsvarianten, so ergibt sich mit Variante b für alle fünf Subskalen eine etwas höhere Varianzaufklärung als mit Variante a.

Vergleicht man die Subskalen untereinander, so ergibt sich folgendes Bild: Während die Varianzaufklärung bei den Skalen AL, GS und SH zwischen 30% und 38% beträgt, ist die aufgeklärte Varianz bei den Skalen BZ und KO deutlich geringer.

Bei der Skala GS tritt Depressivität etwas stärker als Prädiktor in Erscheinung, bei der Skala SH hingegen die Problemeinsicht bei Entlassung. Bei der Skala BZ haben positive Lebensereignisse einen stärkeren Einfluss auf das Outcome als negative Lebensereignisse, bei der Skala KO verhält es sich umgekehrt.

Tabelle 24. Pfadkoeffizienten der Modelle zur Vorhersage der Subskalen von EMEK_27

	Variante a					Variante b				
Skala	R^2	HAQE	DEPE	LebP	LebN	R^2	HAQE	DEPE	LebP	LebN
EMEK_27	.363	+.28	-.09	+.37	-.30	.409	+.26	-.16	+.35	-.35
EMEK_AL	.338	+.26	-.12	+.35	-.28	.377	+.25	-.18	+.34	-.32
EMEK_GS	.323	+.22	-.14	+.35	-.29	.353	+.21	-.19	+.33	-.30
EMEK_SH	.297	+.29	-.07	+.33	-.24	.360	+.27	-.15	+.32	-.32
EMEK_BZ	.187	+.27	+.07	+.31	-.15	.204	+.24	+.02	+.29	-.25
EMEK_KO	.078	+.09	+.01	+.10	-.25	.092	+.10	-.02	+.09	-.26

Anmerkung. *N* = 664 Patienten aus den Studien C und D, bei denen entsprechende Messungen zum Entlass- und Katamnese-Zeitpunkt vorlagen. Pfadkoeffizienten zur Prognose von EMEK_27 (Prädiktoren): HAQE – Zufriedenheit mit Problemeinsicht und unmittelbarem Therapieerfolg bei Entlassung, DEPE – Mit SCL-90-R-Skala gemessene Depressivität (Demoralisierung) bei Entlassung, LebP – Auftreten von positiven Lebensereignissen im Jahr nach Entlassung, LebN – Auftreten von negativen Lebensereignissen im Jahr nach Entlassung. Alle Pfadkoeffizienten > |.05| sind statistisch signifikant. R^2 ist für alle Skalen statistisch signifikant (p<.05). Alle 10 Modelle sind in Anhang 8.1.4 grafisch wiedergegeben.

3.2.7 Beantwortung der Fragestellungen

Fragestellung 1: Item-Kennwerte

Wie hoch sind die Mittelwerte, Streuungen und Trennschärfen der 27 EMEK-Items bei Variante a und b? Zeigen sich Unterschiede im Vergleich? Wie hoch sind die paarweisen Korrelationen zwischen den Items der Varianten a und b?

Mit beiden Berechnungsvarianten ergibt sich für die 27 Einzelitems im Durchschnitt ein positives Bild von der Ergebnisqualität im Sinne einer Verbesserung des Befindens. Die stärksten Verbesserungen zeigen sich auf bewältigungsorientierten Items wie z.B. im Umgang mit Alltagsbelastungen.

Die Mittelwerte der Variante a betragen im Durchschnitt 0,55, was einer „Erfolgsquote" von 55% entspricht. Die standardisierten Mittelwerte der Variante b erreichen im Durchschnitt einen Wert von 0,56 und entsprechen damit wie erwartet einer mittelgroßen Effektausprägung, wenn man in Analogie zur Terminologie von Cohen (1992) einen standardisierten Wert von 0,20 als kleine, 0,50 als mittlere und 0,80 als große Veränderung interpretiert.

Die Item-Trennschärfen liegen bei Variante b mit einem über alle Items gemittelten Wert von .64 zwar höher als bei Variante a (gemittelter Wert: .59), der Unterschied ist jedoch nicht so dramatisch wie zunächst angenommen.

Die Korrelationen zwischen den Einzelitems der Varianten a und b bewegen sich zwischen .71 und .94. Besonders hohe Korrelationen ergeben sich bei den dreistufigen Items 9-15, besonders gering fallen die Korrelationen hingegen bei den drei zwölfstufigen Items 16, 17 und 18 aus. Die über alle Items gemittelte Korrelation beträgt .84, was in etwa der nach Cohen (1983) zu erwartenden Größenordnung entspricht.

Fragestellung 2. Verteilungs-Kennwerte der Skala

Wie hoch sind Mittelwerte, Streuungen und Reliabilitäten der beiden Skalen-Varianten von EMEK_27? Welche Form weisen die Skalen-Verteilungen auf (Schiefe, Kurtosis und Abweichung von der Normalverteilung)? Wie hoch korrelieren EMEK_27a und EMEK_27b miteinander?

Der Skalenmittelwert von EMEK_27a liegt bei 14,72 Punkten, d.h. im Durchschnitt berichten die Patienten bei mehr als der Hälfte aller 27 singulären Ergebnisaspekte eine positive Entwicklung. Die Streuung beträgt 8,24.

Der Skalenmittelwert von EMEK_27b beträgt 0,56. Die Streuung beträgt wie erwartet im Gegensatz zu den Einzelitems nicht 1,00, sondern bedingt durch die Aggregierung 0,67.

Die Reliabilitäten der beiden Skalenvarianten unterscheiden sich kaum voneinander und erreichen ausgezeichnete Werte mit einem Cronbachs Alpha von .94 bei EMEK_27a bzw. .95 bei EMEK_27b.

Beide Skalenvarianten weisen eine deutliche Linksschiefe auf. Insgesamt weicht die Verteilungsform der Skala EMEK_27b aber erwartungsgemäß deutlich weniger von der Nor-

malverteilung ab als bei EMEK_27a und weist im Gegensatz zu EMEK_27a zudem keinen Boden- und Deckeneffekt auf. Die Korrelation zwischen EMEK_27a und EMEK_27b ist mit r = .925 höher als zunächst erwartet.

Fragestellung 3: Dimensionalität von EMEK_27

Welche inhaltliche Struktur zeigt sich bei einer Faktorenanalyse der 27 Einzelkomponenten? Wie viele Faktoren lassen sich extrahieren? Wie groß ist die Varianzaufklärung der Ladungsmatrix? Gibt es Unterschiede zwischen Variante a und b? Falls sich mehrere Faktoren extrahieren lassen: Wie hoch sind die Reliabilitäten der ausgehend vom Ladungsmuster gebildeten Subskalen? Welche Interkorrelationen weisen die Subskalen auf?

Eine Hauptkomponentenanalyse führte bei beiden Varianten a und b zu dem Resultat, dass sich ein varianzstarker Generalfaktor zeigt. Dies deutet darauf hin, dass die Aggregation aller 27 Items zu einem Gesamtindex der Ergebnisqualität inhaltlich sinnvoll ist. Darüber hinaus ergab der Eigenwertverlauf der unrotierten Faktorenlösung Anhaltspunkte dafür, dass eine Extraktion von mehr als einem Faktor vielversprechend erscheint. So führte eine Extraktion von drei Faktoren zu dem Resultat, dass bei Variante a ein Varianzanteil von 56,2% und bei Variante b von 62,4% aufgeklärt wurde. Das varimaxrotierte Faktorenmuster weist Einfachstruktur auf (keine Doppelladungen, die einen Betrag von .30 überschreiten) und lässt sich gut in allen fünf Teilstichproben bei Durchführung einer getrennten Faktorenanalyse für die Studien A bis E replizieren. Die Faktoren lassen sich inhaltlich wie folgt benennen: Faktor I „Veränderung des Allgemeinbefindens", Faktor II „Veränderung der Beziehungen zu nahen Personen" und Faktor III „Veränderung kostenrelevanter Aspekte". Eine Extraktion von vier Faktoren ergab, dass der varianzstarke erste Faktor der dreifaktoriellen Lösung in zwei Komponenten „Veränderung des Gesundheitszustandes" und „Veränderung der Fähigkeit zur Selbsthilfe" ausdifferenzierte. Die vierfaktorielle Lösung erklärt bei Variante a insgesamt 60,6% und bei Variante b insgesamt 66,6% der Gesamtvarianz. Allerdings weist die vierfaktorielle Lösung viele Mehrfachladungen auf und ließ sich bei einer getrennten Faktorenanalyse mit Ausnahme der Beziehungsdimension und der kostenrelevanten Dimension in den Substichproben C bis E nicht replizieren. Mit Ausnahme der Items 10 („Veränderung gesundheitsbewusste Lebensweise") und 11 („Veränderung Medikamentenkonsum") weisen alle Items bei der dreifaktoriellen (bzw. vierfaktoriellen) Lösung akzeptable bis gute Kommunalitäten auf.

Als Fazit lässt sich festhalten, dass sowohl bei der dreifaktoriellen als auch bei der vierfaktoriellen Lösung die Variante b gegenüber der Variante a etwa 6% mehr Varianz aufklärt. Damit ist die Informationsausschöpfung bei Variante b erwartungsgemäß etwas besser als bei Variante a. Die varimax-rotierten Ladungsmuster der Varianten a und b zeigen inhaltlich weitgehende Entsprechungen, wobei die Ladungsbeträge bei der Variante b insgesamt etwas höher ausfallen als bei der Variante a.

Inhaltlich bilden die extrahierten Faktoren damit erwartungsgemäß zum einen Veränderungen des gesundheitlichen Befindens und zum anderen die Veränderung von bewältigungsorientierten Ressourcen im Alltag ab. Als weitere inhaltliche Dimensionen ließen sich die Veränderungen der Qualität von wichtigen persönlichen Beziehungen sowie von kostenrelevanten Aspekten identifizieren. Items, die sich mit der subjektiv empfundenen

Belastbarkeit sowie Leistungs- und Arbeitsfähigkeit befassen, bildeten lediglich in Studie C einen eigenen Faktor.

Entsprechend aus den Items gebildete Subskalen zeigen mit Ausnahme der Skala zur Veränderung sozialmedizinisch relevanter Aspekte hohe interne Konsistenzen. Insgesamt fallen die Reliabilitätsschätzungen bei Variante b etwas höher als bei Variante a aus.

Die Interkorrelationen zwischen den Subskalen erreichen moderate Ausprägungen mit Ausnahme des Zusammenhangs zwischen Gesundheitszustand und Fähigkeit zur Alltagsbewältigung, der mit .80 (Variante a) bzw. .84 (Variante b) sehr hoch ausfällt. Der Zusammenhang zwischen Veränderungen in wichtigen persönlichen Beziehungen und Veränderung sozialmedizinischer Aspekte fällt mit Korrelationen unter .20 für beide Varianten hingegen besonders gering aus. Insgesamt fallen die Zusammenhänge zwischen den Subskalen bei Variante b erwartungskonform etwas höher aus als bei Variante a.

Fragestellung 4: Vorhersagbarkeit von EMEK_27 aus Stichprobenmerkmalen

In welchem Ausmaß lässt sich der Skalenwert von EMEK_27 durch Stichprobenzugehörigkeit und Patientenmerkmale vorhersagen? Welche Prädiktoren spielen eine besonders wichtige Rolle zur Vorhersage der Ergebnisqualität? Gibt es Unterschiede zwischen Variante a und b?

Insgesamt lassen sich etwa 10% der Varianz durch Stichprobenmerkmale bzw. Unterschiede zwischen den fünf Teilstichproben A bis E aufklären. Im Gegensatz zur eingangs formulierten Hypothese ist die durch die Prädiktoren in der Regressionsgleichung aufgeklärte Gesamtvarianz bei EMEK_27b nicht höher als bei EMEK_27a. Auch die Partialgewichte sind bei Variante b nicht höher als bei Variante a.

Inhaltlich zeigen sich erwartungsgemäß Bildungseffekte sowie der Befund, dass eine vorzeitige Beendigung der Behandlung, Rentenantragstellung bei Aufnahme und eine lange Erkrankungsdauer mit schlechteren Ergebnissen assoziiert ist. Nicht erwartungsgemäß sind die gefundenen Geschlechts- und Alterseffekte für die Gesamtstichprobe sowie die Tatsache, dass Patienten mit einer somatoformen Störung gleich gut abschneiden wie Patienten mit einer anderen Hauptdiagnose. Die Muster der gefundenen Partialzusammenhänge unterscheiden sich bei den beiden Varianten inhaltlich erwartungsgemäß kaum voneinander.

Im Gegensatz zur eingangs formulierten Hypothese lässt sich neben dem „Case-Mix", der 7% der Gesamtvarianz am multiplen Ergebniskriterium EMEK_27 aufklärt mit der Stichprobenzugehörigkeit eine zweite Varianzquelle mit weiteren 3% Varianzaufklärung identifizieren.

Insgesamt bleibt festzuhalten, dass die beobachteten multiplen Gesamt- und Partialzusammenhänge klein sind und daher nicht überbewertet werden sollten. So erreicht bei den Patientenmerkmalen kaum ein Beta-Gewicht einen Betrag von mehr als .10. Die Varianzaufklärung am Kriterium EMEK_27 beträgt bei Einbeziehung aller Prädiktorvariablen 10%.

Fragestellung 5: Vorhersagbarkeit von EMEK_27 aus Prozessmerkmalen

In welchem Ausmaß lässt sich der Skalenwert von EMEK_27 durch Prozessmerkmale wie Qualität der therapeutischen Arbeitsbeziehung sowie Problemeinsicht und dem Grad an Demoralisierung unmittelbar bei Behandlungsende vorhersagen? Welche Rolle spielt dabei das Auftreten von signifikanten positiven bzw. negativen Lebensereignissen im Jahr nach der Rehabilitation? Gibt es Unterschiede zwischen Variante a und b?

Bei der Skala EMEK_27a konnte ein Varianzanteil von 35,9% aufgeklärt werden. Die Varianzaufklärung bei EMEK_27b ist mit 40,3% erwartungsgemäß höher als bei EMEK_27a. Der Höhe nach entspricht der am Kriterium aufgeklärte Varianzanteil in etwa den Befunden von Wittmann et al. (2002).

Was die Pfadkoeffizienten im Modell betrifft, das einen akzeptablen Fit aufweist, so entsprechen die Ergebnisse insgesamt den Erwartungen. Allerdings überrascht der nicht signifikante Zusammenhang zwischen Depressivität bei Aufnahme und der Qualität der therapeutischen Beziehung, was als gute Nachricht für das Behandlungsteam der einbezogenen Kliniken gewertet werden kann: Auch bei stark demoralisierten (depressiven) bzw. dem gängigen Vorurteil nach „schwierigen Patienten" gelingt es offenbar, eine therapeutische Beziehung aufzubauen, was eine wichtige Grundlage für den weiteren Behandlungsprozess schafft. Allerdings stellt eine ausgeprägte depressive Symptomatik einen ungünstigen Faktor für den weiteren therapeutischen Prozesses und die daraus resultierende Problemeinsicht dar, was wiederum die Frage nach geeigneten Interventions- bzw. Motivationsmaßnahmen zur Behandlung dieser Patienten aufwirft (Nübling, 1992).

Deutlicher als in der Untersuchung von Wittmann et al. (2002) tritt die Rolle von positiven und negativen Lebensereignissen in den 12 Monaten nach Entlassung aus der Klinik für die Stabilisierung des Behandlungserfolges hervor. Damit ist die Schnittstellenproblematik einer adäquaten Nachsorge angesprochen. Das Jahr nach der Reha-Maßnahme stellt offenbar eine sensible Phase für die Patienten dar, in der sich neu erworbene Einsichten und Verhaltensmuster erst noch festigen und im Alltag bewähren müssen.

4 Diskussion

Die methodenkritische Untersuchung der Skala EMEK_27 führt zu dem Ergebnis, das bei einer Verwendung von z-standardisierten Items eine etwas höhere Informationsausschöpfung als bei der Verwendung von dichotomisierten Items resultiert. Dennoch hat die Verwendung der herkömmlichen Variante EMEK_27a angesichts guter teststatistischer Kennwerte eine gewisse Berechtigung, wenn es um eine möglichst anschauliche Vermittlung der Ergebnisse in der Praxis geht. Für wissenschaftliche Fragestellungen, bei denen es um eine möglichst hohe prognostische Validität, etwa im Rahmen von Indikationsentscheidungen geht, ist hingegen die Variante EMEK_27b zu bevorzugen.

Ungelöst bleibt bei beiden Varianten das Ausgangswertproblem, das im Rahmen von direkten Veränderungsmessungen im Rahmen von Einpunkterhebungen besonders gravierend ist. Eine mögliche Option ist daher die Verwendung von quasi-indirekten Veränderungsmessungen mit retrospektivem PRÄ-Test.

Bei der gesamten Diskussion zur Veränderungsmessung im Rahmen von Prä-Post-Untersuchungen ohne Kontrollgruppe bleibt das Ausgangswertproblem jedoch ungelöst. Dies betrifft auch Effektgrößen, die auf klassischen indirekten Veränderungsmessungen oder auch quasi-indirekten Veränderungsmessungen beruhen. So kann ein Nulleffekt sowohl auf die Beibehaltung eines erwünschten Zustandes als auch auf die Beibehaltung eines unerwünschten Zustandes zurückzuführen sein.

4.1 Vergleich zwischen EMEK_27a und EMEK_27b

Vielversprechend an dem Ansatz der multiplen Ergebniskriterien nach Schmidt et al. (1987) ist die Grundidee, bereits auf Item-Ebene zu einer möglichst klaren evaluativen Aussage im Sinne von Messung und Bewertung (Kordy & Scheibler, 1984a) zu gelangen. Der zweite wichtige Aspekt besteht in der Aggregation einer breiten Palette von bio-psycho-sozialen Ergebniskriterien unter Berücksichtigung unterschiedlicher Stakeholderhinteressen, wobei durch die standardisierte Ergebnisbewertung mit Null und Eins auch uneinheitlich skalierte (z.B. dreistufige, fünfstufige und siebenstufige) Items zu einem Index zusammengefasst werden. Drittens ist die berechnete multiple Ergebnisskala gut in der Praxis vermittelbar, da diese allgemeinverständliche Aussagen im Sinne des prozentualen Anteils gebesserter Patienten ermöglicht. Die Items des herkömmlichen (hier mit Variante a bezeichneten) multiplen Ergebniskriteriums wurden nach deren Messung im Sinne von „gebessert" mit dem Wert Eins bzw. bei Nichtveränderung oder Verschlechterung mit dem Wert Null „nicht gebessert" codiert und sodann zur Skala EMEK_27a aufsummiert.

Problematisch an dieser Vorgehensweise ist die Tatsache, dass

1. durch die Dichotomisierung wertvolle Varianz verschenkt wird, was zu einem Informationsverlust führt.
2. Patienten mit einem unveränderten Befinden solchen mit einem verschlechterten Befinden gleichgesetzt werden.

Auf diese Weise entsteht eine Unschärfe bei der Ergebnisbewertung, die im Widerspruch zu dem Anspruch steht, bereits auf Itemebene zu einer möglichst präzisen evaluativen Aussage zu gelangen.

Weitere Kritikpunkte an der herkömmlichen Skala EMEK_27, die nichts mit der Dichotomisierung zu tun haben, beziehen auf die Einbeziehung von Items, die auf unterschiedlichen Veränderungsinformationen (direkt und quasi-indirekt) beruhen, sowie auf das Ausgangswertproblem, das allerdings nicht nur bei direkten Veränderungsmessungen, sondern auch bei indirekten (bzw. quasi-indirekten) Veränderungsmessungen im Rahmen von Prä-Post-Studien ohne Kontrollgruppe ungelöst ist.

4.1.1 Inhaltliche Aspekte

Aus den Faktorenanalysen geht hervor, dass mit beiden Varianten der Skala EMEK_27 die Interessen von drei wichtigen Stakeholdergruppen abgebildet werden:

- Patient selbst mit den beiden Teilaspekten Gesundheit und Alltagsbewältigung
- Wichtige Bezugspersonen im privaten Bereich
- Gesellschaft (Arbeitgeber und Kostenträger im Gesundheits- und Sozialwesen)

An dieser Stelle ist anzumerken, dass sämtliche Angaben aus der subjektiven Sicht der Patienten (Selbstangaben) erhoben wurden. Um objektive Daten zu erhalten, müßten eigentlich die betreffenden Stakeholder (z.B. Angehörige des Patienten) ebenfalls befragt werden, was häufig gar nicht oder nur mit hohem Aufwand realisierbar ist. In der EQUA-Studie (Schmidt et al., 2003) wurde eine entsprechende Befragung der Krankenkassen und Hausärzte der Patienten realisiert und es ergaben sich beachtliche Übereinstimmungen zwischen Selbst- und Fremdangaben.

Wie steht es um die Gleichgewichtung der Interessen der drei Stakeholdergruppen im Sinne einer „fairen Evaluation"? Aus den Faktorenanalysen geht hervor, dass sich die meisten Items der explorativ gebildeten Skala EMEK_27 vorwiegend auf die subjektiv empfundene Gesundheit und Fähigkeit zur Alltagsbewältigung des Patienten beziehen. Bei einer gleichgewichteten Einbeziehung aller Items in die Gesamtskala EMEK_27 werden die Interessen des Patienten somit entsprechend stark gewichtet. Auf die anderen beiden Stakeholdergruppen entfallen lediglich drei (Veränderung der Beziehungen) bzw. vier (Veränderung kostenrelevanter Aspekte) Items. Diese Ungleichgewichtung ist bei Interpretation der Skala EMEK_27 zu beachten. Ein sinnvoller Umgang mit dieser Problematik kann in einer getrennten Betrachtung der faktorenanalytisch ermittelten Subskalen bestehen, die sich ausschließlich aus den Items der jeweils zugehörigen Stakeholdergruppe zusammensetzen.

Um für eine zusammenfassende Bewertung der Ergebnisqualität mit *einem* Kennwert ein ausgewogenes Gesamtbild zu erhalten, das die drei Stakeholder Patient, Bezugspersonen und Gesellschaft gleichermaßen berücksichtigt, wäre folgende Vorgehensweise denkbar: Selektion redundanter Items bei der Skala AL bzw. Verlängerung der Skalen BZ und KO. Mit dieser Strategie könnte man ein multiples Ergebniskriterium konstruieren, das für alle einbezogenen Stakeholdergruppen die gleiche Itemzahl umfasst. Insoweit steht eine inhaltliche Weiterentwicklung der Skala EMEK_27 noch aus. Sinnvoll wäre dabei eine Einbeziehung von aktuellen ICF-Kriterien (vgl. Abschnitt 1.3.2).

Erfreulich ist die Tatsache, dass sich in der Reanalyse besonders deutliche Verbesserungen auf bewältigungsorientierten Items wie „Fähigkeit zur Selbsthilfe" oder „Umgang mit Alltagsbelastungen" zeigen. Dies entspricht dem gesetzlichen Auftrag der Rehabilitation (vgl. Abschnitt 1.1.2) und deutet darauf hin, dass die während des stationären Klinikaufenthaltes erworbenen Einsichten im Jahr nach der Rehabilitation erfolgreich im privaten und beruflichen Alltag umgesetzt werden. Damit gehen die Behandlungsergebnisse über einen reinen Erholungseffekt hinaus. Die Datenanalysen haben ferner ergeben, dass die verbesserte Alltagsbewältigung sehr eng mit einer Verbesserung des Gesundheitszustandes und Wohlbefindens korrespondiert. Auch hinsichtlich des gesundheitlichen Befindens berichten die Patienten ein Jahr nach Entlassung aus der Klinik noch deutliche Verbesserungen im Vergleich zur Zeit vor der Rehabilitation. Eine dritte Itemgruppe befasst sich mit interpersonalen Aspekten, hier beobachten die Patienten eine Verbesserung der Qualität wichtiger Beziehungen zu nahestehenden Personen wie Partner, Familie und Freunde. Die Fähigkeit, sich ein unterstützendes soziales Netzwerk aufzubauen und nahe Beziehungen zu anderen Menschen einzugehen und aufrechtzuerhalten ist nach Schmidt und Strauss (1996) wichtig für die psychische Gesundheit und hat therapierelevante Bedeutung. So beschreibt die Bindungstheorie bereits bei Kindern verschiedene Beziehungsmuster und betont die Wichtigkeit der Bindungsfähigkeit für das eigene Wohlergehen während der gesamten Lebensspanne (Ainsworth, Blehar, Waters & Wall, 1978). Hier haben sich im Jahr nach Entlassung aus der Klinik offenbar positive Entwicklungen hin zu einer sicheren Bindungsstrategie ergeben, was auch in einer der hier zur Reanalyse einbezogenen Teilstichprobe (Studie C) mit einem entsprechenden Bindungsfragebogen belegt werden konnte (Steffanowski et al. 2001). Eine vierte Itemgruppe befasst sich mit sozialmedizinischen Kriterien wie Krankschreibungszeiten und Inanspruchnahme des Gesundheitssystems. Auch hier zeigen sich, wenn auch in einem geringeren Ausmaß, positive Entwicklungen. Wie in Abschnitt 2.2.3 ausgeführt, können sich unter Kosten-Nutzen-Gesichtspunkten jedoch auch kleine Effekte lohnen und die Investition in eine Rehabilitationsmaßnahme rechtfertigen.

Einzelne Items weisen auf beiden Varianten der Skala EMEK_27 nur geringe Trennschärfen auf. So ist die Formulierung „gesundheitsbewusst leben" bei Item 10 mehrdeutig und kann unterschiedlich z.B. im Sinne von Ernährung, Rauchen, Alkoholkonsum, Bewegungsverhalten, Stress oder Befolgung ärztlicher Verordnungen aufgefasst werden. Bei Item 11 ist die subjektive Einschätzung der Veränderung des Medikamentenkonsums im Sinne von „mehr" versus „weniger" problematisch, da sich dies auf Häufigkeit oder Dosis der Einnahme beziehen kann.

4.1.2 Informationsverlust

Die hier vorgeschlagene alternative Variante b verwendet im Gegensatz zur Variante a die gesamte Iteminformation, lässt also sämtliche „Zwischentöne" zu und berücksichtigt darüber hinaus die Möglichkeit, dass sich das Befinden auf einem bestimmten Ergebnisaspekt im Jahr nach der Behandlung auch verschlechtern kann. Die Höhe der paarweisen Itemkorrelationen zwischen Variante a und b lassen eindeutige Hinweise auf einen Informationsverlust durch die Dichotomisierung erkennen, der auf Itemebene in etwa der von Cohen (1983) angegebenen und kritisierten Größenordnung entspricht.

Bei der aggregierten Gesamtskala ist der Informationsverlust bei der aus dichotomen Items gebildeten Variante EMEK_27a gegenüber der aus lineartransformierten standardisierten Items gebildeten Variante EMEK_27b weniger dramatisch als zunächst angenommen: So beträgt die Korrelation zwischen beiden Skalenvarianten .925, was einem Informationsverlust von lediglich 1-0,925² = 14,4% entspricht. Zur Erklärung der (im Gegensatz zu den Befunden bei den Items) überraschend hohen Korrelation zwischen EMEK_27a und EMEK_27b kann mit der Klassischen Testtheorie (KTT) argumentiert werden, wonach bei Aggregation von mehreren Items die systematische Varianz stärker zunimmt als die Fehlervarianz. So lässt sich die zu erwartende Reliabilitätserhöhung r'_{tt} bei einer Testverlängerung um das k-fache anhand der bislang gemessenen Reliabilität r_{tt} mit der Spearman-Brown-Formel (Gleichung 16) schätzen. Zu den Grundannahmen der KTT und zur Kritik an der KTT sei an dieser Stelle auf die entsprechende Literatur (Amelang et al., 1997) verwiesen.

$$r'_{tt} = \frac{k \cdot r_{tt}}{1 + (k-1) \cdot r_{tt}} \qquad (16)$$

Wenn man davon ausgeht, dass durch die Dichotomisierung ein Verlust an Reliabilität auf Item-Ebene in der von Cohen (1983) angenommenen Größenordnung resultiert, so wirkt sich die Verlängerung um den Faktor 27 bei Aufaggregation der singulären Ergebniskriterien zum multiplen Ergebniskriterium reliabilitätssteigernd aus, so dass der durch die Dichotomisierung bedingte Informationsverlust auf Skalenebene nicht mehr so stark ins Gewicht fällt wie auf Item-Ebene.

Beide Skalenvarianten weisen tatsächlich eine sehr hohe und dem Betrag nach fast identische interne Konsistenz von .94 (EMEK_27a) bzw. .95 (EMEK_27b) auf und die Vorhersage des Behandlungserfolges anhand verschiedener Patienten- und Stichprobenmerkmale führte in beiden Fällen zu einer fast identischen Varianzaufklärung von etwa 10% durch die einbezogenen Prädiktoren. Eine Faktorenanalyse des Itempools ergab bei Variante b hingegen eine um etwa 6% höhere Informationsausschöpfung durch die extrahierten Faktoren als bei Variante a. Auch die Vorhersage durch Prozessmerkmale in der Pfadanalyse führte bei EMEK_27b zu einer 5% höheren Varianzaufklärung als bei

Pfadanalyse führte bei EMEK_27b zu einer 5% höheren Varianzaufklärung als bei EMEK_27a.

Auf der Ebene der Gesamtskala ist der Informationsverlust aufgrund der durch die Aggregation bedingten Reliabilitätssteigerung somit weniger dramatisch als auf der Ebene der Einzelitems, aber immer noch substanziell. Eine künstliche Dichotomisierung von kontinuierlichen Informationen, wie dies bei der Itembildung für die Skala EMEK_27a vollzogen wird, ist angesichts dieser Ergebnisse in Anwendungfeldern, wo es um ein einfaches, zusammenfassendes und gut vermittelbares Screening der Ergebnisqualität geht, durchaus sinnvoll. Geht es hingegen um Forschungsfragen, etwa um die Erzielung einer möglichst hohen prognostischen Validität bei indikativen Entscheidungen, wird hingegen der Argumentation von Cohen (1983) gefolgt, wonach die gesamte verfügbare Information bei der Datenanalyse auch Verwendung finden sollte.

Folgt man der Argumentation von Cohen (1983), so lässt sich folgendes festhalten: Alle Items einer multiplen Kriterienskala sollten im Interesse einer möglichst hohen Ausschöpfung der vorhandenen Information kontinuierlich skaliert sein (keine nachträgliche Dichotomisierung von vorher kontinuierlich erhobenen Items).

Natürlich existieren Variablen, die auf den ersten Blick dichotom skaliert sind, so etwa die Frage, ob jemand Raucher oder Nichtraucher zu einem bestimmten Zeitpunkt ist. Dennoch kann in solchen Fällen häufig eine differenziertere Erhebungsweise gewählt werden, im hier gewählten fiktiven Raucherbeispiel könnte alternativ die Anzahl der pro Woche gerauchten Zigaretten erfragt werden.

Auch wenn die standardisierten direkten Veränderungsmaße der Variante b keine Effektgrößen im eigentlichen Sinne darstellen, lassen sie sich aufgrund der Standardmetrik dennoch gut interpretieren. Verbesserungen des Befindens werden sowohl beim Einzelitem als auch bei der Gesamtskala EMEK_27b mit einem positiven Vorzeichen, Verschlechterungen hingegen mit einem negativen Vorzeichen ausgedrückt. Die größte Einzelveränderung zeigte sich dabei auf Item 9 „Veränderung im Umgang mit Alltagsbelastungen" mit einem standardisierten Wert von 0,95, was ein erfreuliches Ergebnis im Hinblick auf den gesetzlichen Auftrag der Rehabilitation darstellt. Darüber hinaus wird durch die Standardisierung eine Gleichgewichtung aller Items bei der Skalenberechnung von EMEK_27b erreicht. Diese Vorgehensweise entspricht dem methodischen Prozedere bei einer Meta-Analyse, bei der für jede einbezogene Studie zunächst ein durchschnittlicher Studieneffekt aus allen Items ermittelt wird. Die aggregierte Gesamtskala EMEK_27b erreicht einen Mittelwert von 0,56.

Die standardisierte Variante b stellt dabei gewisse Anforderungen an die Verteilungseigenschaften der Rohwert-Items. So besteht bei stark eingeschränkter Streuung der Rohwert-Items die Gefahr einer Überschätzung der Effekte. Dieses Problem ist allerdings nicht spezifisch für den hier untersuchten Ansatz der multiplen Ergebniskriterien, sondern stellt ein allgemeines Problem bei jeder Berechnung von Effektgrößen bzw. standardisierten Maßen dar: Geht die zur Standardisierung verwendete Streuung gegen Null, so geht die berechnete Effektgröße gegen Unendlich. Aus diesem Grund ist an die Rohwert-Items die Forderung zu stellen, dass diese zumindest annähernd normalverteilt sein sollten. Abbildung 14 in Abschnitt 3.2.1 enthält die entsprechenden Häufigkeitsverteilungen der

27 Rohwert-Items, aus denen die beiden Varianten a und b gebildet werden. Lediglich Item 18 „Veränderung von Krankenhauszeiten" weist eine stark eingeschränkte Varianz auf, da dies ein relativ seltenes Ereignis für die Patienten ist und daher sowohl vor als auch nach der Reha häufig mit Null angegeben wird.

4.1.3 Boden- und Deckeneffekte

Der zweite hier vorgebrachte Kritikpunkt an der Dichotomisierung bezieht sich auf den Anspruch einer fairen Bewertung bei der Ergebnisevaluation. So ist nicht nachvollziehbar, warum Patienten mit einem verschlechterten Befinden solchen mit einem unveränderten Befinden gleichgesetzt werden. Die Bewertung „unverändert" kann sowohl die Beibehaltung eines günstigen als auch eines ungünstigen Zustandes implizieren, während eine Verschlechterung in jedem Fall ein negativ zu bewertendes Ergebnis darstellt. Der durch die Dichotomisierung bedingte Informationsverlust führt diesbezüglich zu einer Unschärfe bei der Erfolgsbewertung. Diese Problematik zeigt sich in den Daten anhand der Skalenverteilung von EMEK_27a in einem deutlichen Bodeneffekt mit dem zweiten Skalenmaximum auf dem Wert Null. Dies bedeutet, dass die Skala EMEK_27a im unteren Bereich im Sinne von kritischen Verläufen wenig differenziert (vgl. Abbildung 17 in Abschnitt 3.2.3). Eine wichtige Teilaufgabe der evaluativen Forschung bzw. eines Routinemonitorings der Ergebnisqualität besteht jedoch darin, neben erfolgreichen Patienten auch kritische Verläufe im Sinne eines Aufmerksamkeitssignals (Kordy & Hannöver, 1998) zu identifizieren, das auf Nebenwirkungen der Behandlung, Kontraindikation oder andere Ursachen für den unbefriedigenden Ausgang der Rehabilitationsmaßnahme bei bestimmten Patienten hindeuten kann. Auch im oberen Bereich sehr positiver Veränderungen und damit sehr guter Ergebnisqualität zeigt die herkömmliche Variante EMEK_27a Schwächen hinsichtlich der Differenzierungsfähigkeit (Deckeneffekt). Die alternativ berechnete Skalenvariante EMEK_27b, welche die gesamte Iteminformation ausschöpft, weist die Problematik eines Boden- und Deckeneffektes nicht auf und entspricht – abgesehen von einer gewissen Linksschiefe – eher der Normalverteilungsform als EMEK_27a.

4.1.4 Fehlende Informationen über den Ausgangszustand

Ein wichtiger Aspekt bei Entwicklung der Skala EMEK_27 war der Anspruch, die Ergebnisqualität im Rahmen von Einpunkterhebungen (d.h. ohne vorhergehenden PRÄ-Test) abbilden zu können, etwa bei katamnestischen Routinebefragungen. Hierdurch könnte der Erhebungsaufwand minimiert werden, da ein organisatorisch aufwendiges Längsschnitt-Design entbehrlich wäre. Aus diesem Grund werden vor allem direkte Veränderungsmessungen eingesetzt. In Abschnitt 2.1.1 bzw. 2.1.4 wurde bereits ausgeführt, dass eine ausschließliche Verwendung von direkten Veränderungsmaßen zu einer Unschärfe bei der Erfolgsbewertung führt, da keine Information über den Ausgangszustand verfügbar ist und die Ankreuzung „unverändert" bei der Einschätzung der Veränderung eines bestimm-

ten Einzelaspektes sowohl eine positiv als auch negativ zu bewertende Information darstellen kann.

Eine ausschließliche Verwendung von direkten Veränderungsmaßen zur Bewertung der Ergebnisqualität ist im Hinblick auf das Ausgangswertproblem nicht ratsam. Wenn eine prospektive PRÄ-POST-Messung nicht realisierbar ist, sollte zumindest mit dem Ansatz der quasi-indirekten Veränderungsmessung (retrospektiver PRÄ-Test) der erinnerte Status vor der Behandlung erfragt werden. Wenn immer möglich, sollte allerdings ein echtes PRÄ-POST-Design (indirekte Veränderungsmessung) realisiert werden, um möglichst zeitnahe Informationen über den jeweiligen Zustand der Patienten zu gewinnen. Optimal im Sinne der bestmöglichen Evidenz wäre dabei ein PRÄ-POST-Kontrollgruppenplan mit Vorher-Nachher-Messungen einer behandelten und unbehandelten Patientengruppe. Dies ist eine Forderung, die aufgrund der rechtlichen und versorgungsstrukturellen Rahmenbedingungen im Bereich der medizinischen Rehabilitation bislang nicht realisiert werden konnte. Dennoch wäre beispielsweise eine Wartelistenkontrollgruppe denkbar, um etwaige Effekte aufgrund spontaner Remission oder anderer nicht rehabilitationsspezifischer Faktoren besser abschätzen zu können.

4.1.5 Zusammenfassende Gesamtbewertung

Die Faktorenanalysen führten zu dem Ergebnis, dass durch beide Varianten von EMEK_27 Veränderungen in vier inhaltlich reha-relevanten Bereichen (Gesundheitszustand, Fähigkeit zur Selbsthilfe, Qualität der Beziehungen zu anderen Menschen sowie kostenrelevante Aspekte) abgebildet werden. Gesundheitszustand und Fähigkeit zur Selbsthilfe hängen dabei sehr eng miteinander zusammen. Sowohl die aus allen 27 Items gebildete Gesamtskala als auch vier der fünf berechneten Subskalen weisen bei Variante a und b hohe interne Konsistenzen und damit eine hohe Messgenauigkeit auf.

Die im Rahmen der Reanalyse alternativ konstruierte Variante b, welche die gesamte verfügbare Iteminformation berücksichtigt, weist dabei insgesamt etwas bessere Verteilungskennwerte und ein höheres Maß an Differenzierungsfähigkeit auf als die herkömmliche Variante a, die auf künstlich dichotomisierten Einzelitems basiert.

Ungelöst bleibt sowohl bei Skalenvariante a als auch bei Skalenvariante b das Ausgangswertproblem, da bei direkten Veränderungsmessungen keinerlei Information über den Ausgangszustand verfügbar ist. Zu diesem Punkt wird in Abschnitt 4.2 ausführlich Stellung bezogen.

Dennoch haben direkte Veränderungsmaße sowie die daraus abgeleitete Skala EMEK_27 angesichts der hohen Ökonomie (nur ein Messzeitpunkt nötig), guter teststatistischer Kennwerte (hohe Reliabilität) und des beachtlichen Zusammenhangs mit indirekten Veränderungsmaßen (vgl. Nübling et al., 2004) durchaus ihre Berechtigung, wenn es um ein routinemäßiges Screening der Ergebnisqualität etwa im Rahmen von fortlaufenden Katamnesen geht. Die Skala EMEK_27 kann und soll dabei selbstverständlich kein umfangreiches PRÄ-POST-Assessment ersetzen.

Das herkömmliche multiple Ergebniskriterium EMEK_27a weist trotz der angesprochenen methodischen Schwächen dabei ein hohes Maß an Praktikabilität bei der Umsetzung in die Praxis auf, während die etwas höhere Informationsausschöpfung der Variante E-MEK_27b diese in allen Anwendungsfeldern geeigneter erscheinen lässt, wo es um eine möglichst hohe prognostische Validität geht, etwa wenn vom Zustand bei Entlassung auf katamnestisch erhobene dVM geschlossen werden soll.

Eine Möglichkeit zur Lösung des Problems der fehlenden Informationen über den Ausgangszustand im Rahmen von Einpunkterhebungen kann die quasi-indirekte Veränderungsmessung und die Entwicklung einer entsprechenden auf qVM beruhenden Skalenvariante von EMEK_27 bieten, da diese wie die dVM mit nur einem Messzeitpunkt nach der Behandlung auskommt. Allerdings muss dieser Vorteil in der Praxis durch eine Verdoppelung der Itemzahl erkauft werden, da jeder Aspekt bei der qVM im Gegensatz zur dVM zwei Mal (retrospektive PRÄ-Messung und aktuelle POST-Messung) erfragt werden muss.

4.2 Methodische Weiterentwicklung

Bei reinen Prä-Post-Studien ohne Kontrollgruppe ergibt sich bei allen drei Varianten der Veränderungsmessung (also iVM, qVM und dVM) das gleiche methodische Problem, wenn zur Bewertung der Ergebnisqualität ausschließlich PRÄ-POST-Differenzen, RETRO-POST-Differenzen bzw. direkte Veränderungseinschätzungen herangezogen werden (Abschnitt 2.1.4): Das tatsächliche Befinden zum POST-Zeitpunkt wird nicht berücksichtigt. Dies führt dazu, dass die Beibehaltung von erwünschten Zuständen ebenso mit einer Effektgröße von Null codiert wird wie die Beibehaltung von unerwünschten Zuständen. Damit wird die tatsächliche Gesamtergebnisqualität unterschätzt.

Gerdes (1998) versucht dieses Problem mit der zielorientierten Ergebnismessung zu lösen. Diese berücksichtigt zwar den Ausgangszustand der Patienten, führt durch die einseitige Selektion belasteter Patienten aber zu anderweitigen methodischen Problemen wie erhebliche Reduktion der Stichprobengröße, Überschätzung der Effekte aufgrund Regression zur Mitte sowie Nichtberücksichtigung der Beibehaltung von positiven Zuständen sowie Nichtberücksichtigung der Möglichkeit, dass sich das Befinden auf einem vorher unauffälligen Kriterium auch verschlechtern kann (Zwingmann, 2003). Die Anwendung von Korrekturverfahren zur Auspartialisierung des Ausgangsniveaus mit dem Ziel der Gewinnung sogenannter „wahrer Veränderungswerte" ist durch die Abhängigkeit von der jeweils verwendeten Stichprobe wiederum mit anderen Unwägbarkeiten verbunden (Abschnitt 2.1.2).

4.2.1 Lösungsvorschlag für das Ausgangswertproblem

Es stellt sich die Frage, wie der Behandlungserfolg bei *PRÄ-POST-Untersuchungen ohne Kontrollgruppe* so bewertet werden kann, dass die mit der ZOE verbundenen methodischen Probleme nicht auftreten und neben dem rehabilitativen Gedanken einer Verbesserung des Befindens auch der präventive Aspekt im Sinne der Erhaltung von wünschenswerten Zuständen abgebildet wird. So berücksichtigt Kordy (1997) mit seinem Konzept der klinischen Signifikanz sowohl Status- als auch Veränderungsinformationen bei der Ergebnisevaluation (Abschnitt 2.2.5). Dieser Grundgedanke soll hier weiterverfolgt werden. Im Hinblick auf eine möglichst große Konzentration der Information wird dabei der Versuch unternommen, die Ergebnisqualität mit nur *einem* Kennwert so abzubilden, dass

- weder das Ausgangswertproblem der klassischen Effektgrößen
- noch das Problem des Informationsverlustes der ZOE durch Selektion

auftritt. Nachfolgend wird daher eine neue und innovative Methodik zur Erfolgsbewertung vorgeschlagen, bei der ausgehend von der klassischen indirekten Veränderungsmessung sowohl das Befinden zum Zeitpunkt der POST-Messung als auch die PRÄ-POST-Veränderungen in die Erfolgsbewertung einfließen und zu einem entsprechenden Index zusammengefasst werden. Das Verfahren lässt sich auch im Rahmen von Einpunkt-Erhebungen anwenden, wenn quasi-indirekte Veränderungsmessungen vorhanden sind.

Folgende Überlegungen sind bei der Konzeption des neuen Kennwertes ausschlaggebend:

- In die Ergebnisevaluation sollten für jedes singuläre Ergebniskriterium neben Veränderungsinformationen auch Informationen über das tatsächliche Befinden des Patienten nach der Rehablilitation (Statusmessung) einfließen. Dies bedeutet, dass die Erfolgsbewertung zweier Patienten mit gleicher Veränderung (PRÄ-POST-Effektgröße) auf dem neuen Ergebnismaß unterschiedlich ausfallen muss, wenn der eine Patient nach der Behandlung ein gutes Befinden und der andere lediglich ein mäßiges Befinden aufweist.
- Eine positive Bewertung der Ergebnisqualität sollte durch ein positives Vorzeichen, eine negative Bewertung hingegen durch ein negatives Vorzeichen kenntlich gemacht werden. Sowohl PRÄ- und POST-Messung als auch die PRÄ-POST-Differenzen werden daher zunächst anhand der PRÄ-Streuung der Stichprobe standardisiert (z-Standardisierung bzw. Berechnung von d-Effektgrößen).
- Zu diskutieren ist der Nullpunkt, an dem sich die z-Standardisierung des PRÄ- und POST-Status orientieren soll. Hierzu erscheint der Cut-Off-Wert *c* von Jacobson und Truax (1991) ratsam, da dieser den Punkt angibt, an dem die Wahrscheinlichkeit für eine Person gleich groß ist, zur nichtklinischen und zur klinischen Population zu gehören (vgl. Abschnitt 2.2.4).

4.2.2 Veranschaulichung anhand eines Datenbeispiels

Zur Veranschaulichung der Vorgehensweise wird das Beispiel zur Symptom-Checkliste SCL-90-R aus Abschnitt 2.2.5 aufgegriffen. Tabelle 25 enthält die Skalenkennwerte zu Beginn und am Ende der stationären Behandlung. Die Skala „Depressivität" kann Rohwerte von 0,00 (minimale Symptombelastung) bis 4,00 (maximale Symptombelastung) annehmen.

Tabelle 25. Kennwerte der Skala „Depressivität" in der SCL-90-R

N	$M_{PRÄ}$	$SD_{PRÄ}$	M_{POST}	SD_{POST}	d	t	p
664	1,50	0,87	0,88	0,76	+0,71	20,66	<.001

Anmerkung. Effektgröße $d = (M_{PRÄ} - M_{POST}) / SD_{PRÄ}$. t-Test für Messwiederholungen.

Berechnet man die Differenz zwischen der PRÄ-Messung $x_{PRÄ}$ eines Patienten und dem Cut-Off-Wert c und standardisiert man das Resultat anhand der PRÄ-Streuung $SD_{PRÄ}$, so erhält man die z-standardisierte PRÄ-Messung $z_{PRÄ}$ des Patienten zu Beginn der Behandlung (Gleichung 17). Werte kleiner als Null bedeuten, dass ein Patient mit größerer Wahrscheinlichkeit zur klinischen Population gehört und somit eher als „krank" einzustufen ist. Werte größer als Null bedeuten, dass ein Patient mit größerer Wahrscheinlichkeit zur nichtklinischen Population gehört und somit eher als „gesund" einzustufen ist.

$$z_{\mathrm{Pr}\ddot{a}} = \frac{c - x_{\mathrm{Pr}\ddot{a}}}{SD_{\mathrm{Pr}\ddot{a}}} \tag{17}$$

Auch die POST-Messung wird anhand c und $SD_{\mathrm{PRÄ}}$ entsprechend standardisiert. Damit erhält man die z-standardisierte POST-Messung z_{POST} (Gleichung 18).

$$z_{Post} = \frac{c - x_{Post}}{SD_{\mathrm{Pr}\ddot{a}}} \tag{18}$$

Der Grenzwert c für die Skala „Depressivität" wurde in Abschnitt 2.2.5 mit 0,83 bestimmt. Setzt man diesen Wert sowie die gemessene PRÄ-Streuung ($SD_{PRÄ}$ = 0,87) in Gleichung 17 bzw. Gleichung 18 ein, so resultiert für $z_{PRÄ}$ ein Mittelwert von –0,77 und für z_{POST} ein Mittelwert von -0,06. Die Differenz zwischen $z_{PRÄ}$ und z_{POST} entspricht der Effektgröße d und beträgt +0,71.

Während die Patienten zu Beginn der Behandlung im Durchschnitt somit noch deutlich im Bereich der klinischen („kranken") Population liegen, so liegt der Mittelwert am Ende der

Behandlung nahe Null und somit im Grenzbereich zwischen der klinischen und nichtklinischen Population.

Wie stellt sich die Bewertung der Ergebnisqualität nun nach der herkömmlichen indirekten Veränderungsmessung für verschiedene Patienten dar? Abbildung 23 enthält den Scatterplot von $z_{PRÄ}$ und z_{POST}. Ein Großteil der Patienten liegt oberhalb der Diagonale und weist auf der POST-Messung im Vergleich zur PRÄ-Messung somit eine Verbesserung des Befindens auf, was sich in der positiven Effektgröße von +0,71 entsprechend ausdrückt. Einzelne Patienten sind in der Grafik mit den Buchstaben a. bis i. gekennzeichnet. Negative z-Werte stehen für eine Symptombelastung, die oberhalb des Cut-Off-Wertes c und damit eher im Bereich der klinischen Population liegt. Ein z-Wert von Null entspricht exakt dem Wert *c* und z-Werte größer als Null stehen für eine Symptombelastung, die unterhalb von *c* und damit im eher Bereich der bevölkerungsrepräsentativen Population liegt.

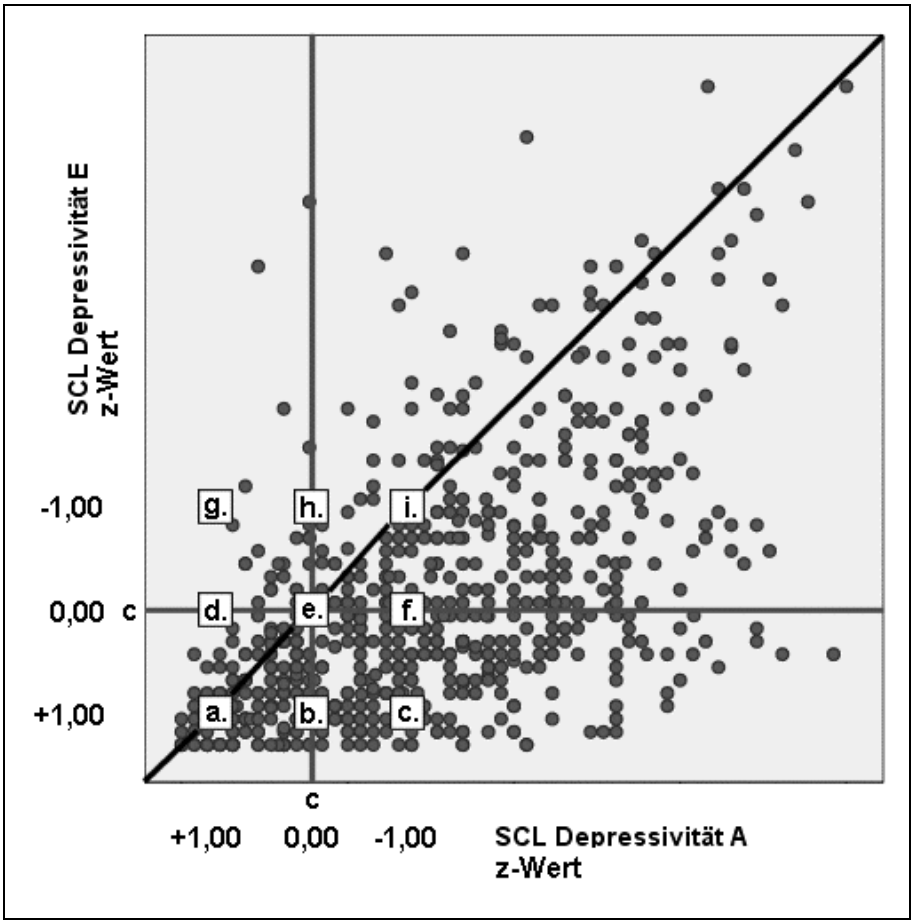


Abbildung 23. Scatterplot der z-standardisierten GSI-Skala der SCL-90-R (EQUA**-Studie, Vergleich zwischen Aufnahme (PRÄ) und Entlassung (POST). *N* = 664 Patienten mit vorhandener PRÄ- und POST-Messung.**

Die Problematik einer ausschließlichen Verwendung von PRÄ-POST-Effektgrößen zur Erfolgsbewertung zeigt sich deutlich, wenn man in der Abbildung 23 die beiden auf der Diagonale liegenden Patienten a. und i. miteinander vergleicht. Beide Patienten weisen eine

individuelle Effektgröße von 0,00 auf. Ist daraus nun der Schluss zu ziehen, dass beide Patienten am Ende der Behandlung eine identische Ergebnisqualität aufweisen? Wohl kaum. Während bei Patient a. bereits zu Beginn der Rehabilitation ein gutes Befinden ($z_{PRÄ}$ = +1,00) festzustellen ist und dies auch am Ende der Behandlung noch so ist (z_{POST} = +1,00), muss bei Patient i. die Beibehaltung eines schlechten Befindens ($z_{PRÄ}$ = -1,00 und z_{POST} = -1,00) und damit ein Versagen der intendierten Therapie konstatiert werden.

Wie stellt sich die Bewertung des Behandlungserfolges von Patienten dar, dessen Befinden sich im Therapieverlauf verändert hat? Auch hier führt die ausschließliche Verwendung von Effektgrößen zu einer Unschärfe bei der Ergebnisbewertung. So ist sowohl bei Patient b. als auch bei Patient f. eine Verbesserung von +1,00 und damit nach Cohen ein „großer Effekt" festzustellen. Während Patient b. jedoch mit einer vergleichsweise moderaten Beeinträchtigung ($z_{PRÄ}$ = 0,00) die Therapie beginnt und am Ende der Behandlung gesund ist (z_{POST} = +1,00), beginnt Patient f. die Behandlung mit einer starken Beeinträchtigung ($z_{PRÄ}$ = -1,00) und weist am Ende der Therapie noch immer eine gewisse Beeinträchtigung (z_{POST} = 0,00) auf, was Kordy (1997) mit dem Begriff „klinisch nicht ausreichende Verbesserung" umschreiben würde. Auch hier sollte die Ergebnisqualität der beiden Patienten somit unterschiedlich bewertet werden, was durch eine ausschließliche Verwendung der PRÄ-POST-Effektgröße *d* nicht möglich ist.

4.2.3 Berechnung eines Composit-Kriteriums

Wie lassen sich Status- und Veränderungsinformationen zu einem einheitlichen Bild zusammensetzen, das eine faire Beurteilung der Ergebnisqualität ermöglicht? Nachfolgend wird hierzu eine ebenso einfache wie pragmatische Vorgehensweise vorgeschlagen, die auf der Überlegung basiert, dass ein Maß zur Bewertung der Ergebnisqualität

1. das tatsächliche Befinden des Patienten nach der Behandlung und damit auch den Erhalt erwünschter Zustände (POST-Messung)
2. die während der Behandlung erzielte PRÄ-POST-Veränderung

berücksichtigen sollte. Das neue Ergebniskriterium stellt somit ein Composit-Maß dar, das sowohl den Zustand am Ende der Behandlung als auch die beobachtete Veränderung in einem neuen Kennwert vereint. Die Berechnung erfolgt nach Gleichung 19. Das neue singuläre Ergebniskriterium wird in Anlehnung an den Begriff der Ergebnisqualität mit EQ bezeichnet. Setzt man den z-standardisierten Mittelwert der POST-Messung (z_{POST} = -0,06) sowie die Effektgröße (*d* = +0,71) in Gleichung 19 ein, so resultiert ein Mittelwert von +0,33 (*SD* = 0,72).

$$EQ = (z_{post} + d)/2 \tag{19}$$

Um Missverständnissen vorzubeugen, ist an dieser Stelle zu anzumerken, dass es sich bei EQ keinesfalls um ein „Veränderungsmaß" oder die Schätzung einer „wahren" oder „resi-

dualen" Veränderung wie in Abschnitt 2.1.3 ausgeführt handeln soll. In den neu berechneten Index fließen zwei unterschiedliche Informationen (Status- und Veränderungswerte) ein, was in der Summe eine zusammenfassende Bewertung der Ergebnisqualität ermöglichen soll.

Formt man Gleichung 19 um, indem man die Beziehung $d = z_{POST} - z_{PRÄ}$ berücksichtigt, so resultiert Gleichung 20. Es wird deutlich, dass der Zustand nach der Behandlung bei EQ rechnerisch gesehen doppelt so starke Berücksichtigung findet wie der Zustand vor der Behandlung. Dies ist durchaus im Sinne einer fairen Evaluation, bei der einerseits vor allem das durch die Intervention beeinflussbare Ergebnis nach der Behandlung interessiert, andererseits aber auch das Ausgangsniveau nicht vernachlässigt wird.

$$EQ = (z_{post} + z_{post} - z_{prä}) / 2 = (2 \cdot z_{post} - z_{prä}) / 2 = z_{post} - 0{,}5 \cdot z_{prä} \qquad (20)$$

Das neu gebildete Composit-Kriterium EQ nimmt bei der Ergebnisevaluation somit gewissermaßen eine Mittelstellung zwischen der reinen POST-Messung (entspricht z_{POST} – 0,0 $z_{PRÄ}$) und dem klassischen indirekten Veränderungsmaß *d* (entspricht z_{POST} – 1,0 $z_{PRÄ}$) ein, was durch Tabelle 26 veranschaulicht wird.

Tabelle 26. Evaluation der Ergebnisqualität mit Status- und Veränderungsinformationen

Regel zur Evaluation der Ergebnisqualität	Berechnung	Kennwert
Ausschließliche Berücksichtigung des POST-Status	$z_{post} - 0{,}0 \cdot z_{prä}$	POST-Messung z_{POST}
Gleichberechtigte Berücksichtigung von POST-Status und PRÄ-POST-Veränderung	$z_{post} - 0{,}5 \cdot z_{prä}$	Composit-Maß EQ
Ausschließliche Berücksichtigung der PRÄ-POST-Veränderung	$z_{post} - 1{,}0 \cdot z_{prä}$	Effektgröße *d*

4.2.4 Bewertung der Ergebnisqualität mit dem Composit-Kriterium

In Tabelle 27 wird anhand des Datenbeispiels veranschaulicht, wie sich die Ergebnisbewertung mit EQ darstellt. Zunächst wird das Befinden bei der POST-Messung betrachtet. Die Patienten a., b. und c. sind am Ende der Rehabilitation „gesund" (keine Depressivität, z_{POST} = +1,00). Die Patienten d., e. und f. befinden sich am Ende der Rehabilitation im Grenzbereich zwischen „krank" und „gesund" (moderate Depressivität, z_{POST} = 0,00). Die Patienten g., h. und i. sind am Ende der Rehabilitation „krank" (ausgeprägte Depressivität, z_{POST} = –1,00).

Betrachtet man als nächstes die in Abbildung 23 auf der Diagonale liegenden Patienten a., e. und i. mit einer Effektgröße von 0,00 so wird folgendes deutlich: Der Index EQ erlaubt eine differenziertere Bewertung der Ergebnisqualität als die bloße Betrachtung der

PRÄ-POST-Effektgrößen. Patient i. ist am Ende der Rehabilitation unverändert „krank" ($z_{PRÄ}$ = −1,00 und z_{POST} = −1,00). Dieser negative Befund schlägt sich in einem EQ-Punktwert von −0,50 nieder. Patient e verharrt unverändert im Grenzbereich zwischen „krank" und „gesund" (EQ = 0,00). Bei Patient a wird die Bewahrung eines gesunden Zustandes ($z_{PRÄ}$ = +1,00 und z_{POST} = +1,00) hingegen mit einem Wert von EQ = +0,50 honoriert.

Patient b. hat mit einer moderaten Beeinträchtigung ($z_{PRÄ}$ = 0,00) die Behandlung begonnen und ist zum Behandlungsende gesund. Zusätzlich zum positiven Ergebnis bei Therapieende (z_{POST} = +1,00) ist also eine Verbesserung des Befindens um einen Punkt (d = +1,00) zu vermerken, was mit einem EQ-Wert von +1,00 entsprechend berücksichtigt wird. Auch das Befinden von Patient d. hat sich zwischen PRÄ- und POST-Messung um einen Punkt verbessert. Allerdings hat Patient d. ungünstigere Ausgangsvoraussetzungen und das erreichte Ergebnis bei Entlassung reicht nicht aus, um von einer völligen Gesundung sprechen zu können. Der Punktwert EQ beträgt +0,50, so dass einerseits die erreichte Verbesserung zwar honoriert wird, andererseits aber auch berücksichtigt wird, dass der Status bei Entlassung noch nicht optimal ist.

Auch die Ergebnisqualität der übrigen Patienten in Tabelle 27 lässt sich dementsprechend durch Kombination der Information zum POST-Status mit der Information zur PRÄ-POST-Veränderung differenziert bewerten.

Tabelle 27. Bewertung der Ergebnisqualität mit dem Composit-Kriterium EQ **unter Berücksichtigung von Status-** ***und*** **Veränderungsinformationen.**

Patient	$z_{PRÄ}$	z_{POST}	d	EQ	**Bewertung**
a.	+1,00	+1,00	0,00	+0,50	Beibehaltung eines günstigen Zustandes
b.	0,00	+1,00	+1,00	+1,00	Von Grenzbereich ausgehend verbessert
c.	-1,00	+1,00	+2,00	+1,50	Klinisch relevante deutliche Verbesserung
d.	+1,00	0,00	-1,00	-0,50	Von gutem Zustand ausgehend verschlechtert
e.	0,00	0,00	0,00	0,00	Beibehaltung eines uneindeutigen Zustandes
f.	-1,00	0,00	+1,00	+0,50	Klinisch nicht ausreichende Verbesserung
g.	+1,00	-1,00	-2,00	-1,50	Klinisch relevante deutliche Verschlechterung
h.	0,00	-1,00	-1,00	-1,00	Von Grenzbereich ausgehend verschlechtert
i.	-1,00	-1,00	0,00	-0,50	Beibehaltung eines ungünstigen Zustandes

4.2.5 Vergleich mit dem Konzept der Residual Gain Scores

In Tabelle 28 sind die Kennwerte der Skala „Depressivität" aus dem Datenbeispiel wiedergegeben. Zusätzlich zu den bereits berechneten Kennwerten wurden standardisierte Veränderungsresiduen (z_{RES}) aus der Regression der POST-Messung auf die PRÄ-Messung berechnet und deren Kennwerte ebenfalls mit in Tabelle 28 aufgenommen. z_{RES} wurde nach Gleichung 2 und 3 in Abschnitt 2.1.2 berechnet. Der Mittelwert von z_{RES} beträgt 0,00, da sich Residualwerte in einer Stichprobe per Definition grundsätzlich immer gegenseitig zu Null ausmitteln.

Tabelle 28. Standardisierte Werte zur Berechnung des Composit-Kriteriums EQ**.**

	Min	**Max**	**M**	***SD***
$z_{PRÄ}$	-0,77	+0,95	-0,77	1,00
z_{POST}	-0,06	+0,95	-0,06	0,87
d	-2,83	+3,89	+0,71	0,89
EQ	-2,79	+2,11	+0,33	0,72
z_{RES}	-2,42	+4,33	0,00	1,00

Anmerkung. SCL-Skala „Depressivität" aus Studie C und D, Vergleich zwischen Aufnahme- und Entlass-Messung. N = 664.

Tabelle 29 enthält die Interkorrelationen zwischen den Kennwerten. Zunächst ist festzustellen, dass ein mittlerer Zusammenhang zwischen PRÄ- und POST-Status besteht (r = .56). Bei Patienten mit einer starken Depressivität zu Beginn der Behandlung ist also davon auszugehen, dass es schwieriger ist, eine vollständige Gesundung zu erreichen als bei Patienten, die mit einer lediglich leicht ausgeprägten Depressivität die Behandlung beginnen. Der negative Zusammenhang zwischen $Z_{PRÄ}$ und d von -.59 verdeutlicht die Problematik der Regression zur Mitte bei iVM: Je schlechter es einem Patienten zu Beginn der Behandlung geht, desto eher wird eine große PRÄ-POST-Differenz d beobachtet. Damit überschätzt d den Behandlungserfolg bei stark beeinträchtigten Patienten bzw. unterschätzt ihn bei nur moderat beeinträchtigten Patienten. Das hier neu eingeführte Composit-Maß EQ weist diese Abhängigkeit von der PRÄ-Messung nicht auf, d.h. es ermöglicht eher eine faire Bewertung des Behandlungserfolges der Patienten unabhängig vom Ausgangsniveau. Die Zusammenhänge zwischen EQ und z_{POST} mit .81 bzw. zwischen EQ und d in Höhe von .82 veranschaulichen, dass sowohl der POST-Status als auch die PRÄ-POST-Veränderung in etwa gleicher Gewichtung in das neue Bewertungskriterium EQ eingegangen sind. Auffällig ist der perfekte Zusammenhang von 1,00 zwischen EQ und z_{RES}. Dies bedeutet zunächst, dass EQ inhaltlich das Gleiche erfasst wie z_{RES}. Die einfache Kombination der POST-Messung mit den PRÄ-POST-Differenzen führt also offensichtlich zu einem ähnlichen Resultat wie die Berechnung von Residualwerten mit einer entsprechenden Regressionsgleichung.

Tabelle 29. Korrelationen zwischen den standardisierten Kennwerten.

	$z_{PRÄ}$	z_{POST}	d	EQ
z_{POST}	.56	--		
d	-.59	.34	--	
EQ	-.02	.81	.82	--
z_{RES}	.00	.83	.81	1.00

Anmerkung. *N* = 664. Mit Ausnahme des Zusammenhangs zwischen $z_{PRÄ}$ und EQ bzw. $z_{PRÄ}$ und z_{RES} sind statistisch signifikant (p<.001).

Ein deutlicher Vorteil von EQ gegenüber den Veränderungsresiduen besteht allerdings darin, dass EQ durch die vorangegangene Normierung anhand von c im Gegensatz zu den Veränderungsresiduen klare Orientierungspunkte für die Ergebnisbewertung bereitstellt, was sich in den unterschiedlichen Mittelwerten der beiden Maße zeigt. Liegen dokumentierte klinische und nichtklinische Normen vor, die zur Berechnung von c verwendet werden können, so lassen sich mit EQ unterschiedliche Studien oder Einrichtungen direkt miteinander vergleichen. Diesen Vorteil besitzen die Veränderungsresiduen nicht, da sie sich immer nur auf die jeweilige Stichprobe beziehen und rechnerisch immer einen Mittelwert von 0,00 aufweisen.

An dieser Stelle sei noch einmal betont, dass EQ hier nicht im Sinne von „wahrer Veränderung", sondern im Sinne einer evaluativen Kombination von Status- (d.h. des Befindens zum POST-Zeitpunkt) *und* Veränderungsinformationen (PRÄ-POST-Veränderung des Befindens) interpretiert und verstanden sein will. Unter Berücksichtigung dieser Vorgabe steht mit EQ möglicherweise ein ebenso einfacher wie praktikabler Ansatz für eine faire Bewertung der Ergebnisqualität zur Verfügung, der nicht die methodischen Probleme der iVM oder ZOE aufweist und damit eine Lösung für die Ausgangswertproblematik darstellen könnte.

5 Erprobung des neuen Ansatzes

Das neu entwickelte Composit-Kriterium EQ wird nachfolgend einer ersten Erprobung und vergleichenden Validierung mit einfachen POST-Messungen, herkömmlichen indirekten, quasi-indirekten sowie direkten Veränderungsmessungen unterzogen. Hierzu werden zwei Varianten von EQ berechnet: Eine Variante, die auf iVM und eine Variante, die auf qVM basiert. Letzteres ist im Hinblick auf eine mögliche Eignung von EQ im Rahmen von Einpunkterhebungen, bei denen eine echte Prä-Messung nicht möglich ist, von Bedeutung.

5.1 Methodik

Zur Erprobung des neu entwickelten Composit-Kriteriums EQ wird Studie D (EQUA-Studie) verwendet. Wichtige Eckdaten und Stichprobenmerkmale zur EQUA-Studie wurden in Abschnitt 3.1.3 mitgeteilt. Für weitere Details sei an dieser Stelle auf die Publikation von Schmidt et al. (2003) verwiesen, wo das gesamte Projekt umfassend dargestellt wird.

5.1.1 Status- und Veränderungsitems der Skala GB13

Eine wichtige Teilfragestellung der EQUA-Studie befasst sich mit der Frage, inwieweit direkte (dVM), indirekte (iVM) und quasi-indirekte (qVM) Veränderungsmessungen (zur Methodik der drei Ansätze vgl. Abschnitt 2.1.1) miteinander korrelieren und zu einer vergleichbaren Bewertung der Ergebnisqualität führen. Insgesamt ergab sich eine recht gute Übereinstimmung der drei Methoden bei der Erfolgsbewertung in einer Größenordnung von bis zu 80% übereinstimmender Klassifikationen. Die Ergebnisse der Datenanalysen finden sich bei Schmidt et al. (2003) bzw. Nübling et al. (2004) und Schmidt et al. (2006).

Für die vergleichende Analyse der iVM, qVM und dVM wurde in der EQUA-Studie eigens eine Skala zum gesundheitlichen Befinden (GB13) entwickelt, die 13 Aspekte des körperlichen und psychosozialen Wohlbefindens umfasst und darüber hinaus auch alltagsrelevante Fragen wie zur Problemlösekompetenz oder Kontaktfähigkeit enthält. GB13 lässt sich damit als multiples Ergebniskriterium verstehen, das inhaltlich unterschiedliche Bewertungseben zu einem Gesamtindex aufaggregiert.

GB13 weist (im Gegensatz zur Skala EMEK_27) den Vorteil auf, dass alle Items symmetrisch sowohl in Form von Statusmessungen („Wie ist Ihr derzeitiges Befinden“) vor und nach der Rehabilitation als auch in Form von direkten Veränderungsmessungen nach der Rehabilitation erhoben wurden („Wie hat sich Ihr Befinden verändert?“), um einen Ver-

gleich der verschiedenen Methoden zu ermöglichen. Darüber hinaus wurden die Patienten zum Katamnese-Zeitpunkt um eine retrospektive Einschätzung ihres Status („Wie war Ihr Befinden damals unmittelbar vor Beginn der Behandlung?") gebeten. Dies bedeutet, dass 13 Items zum Katamnese-Zeitpunkt drei Mal bearbeitet werden mussten: Einmal in Form der POST-Messung des aktuellen Befindens, dann in Form der retrospektiven Statusmessung (erinnerter PRÄ-Zustand) und schließlich ein weiteres Mal in Form der dVM. Damit sind für alle 13 Items vier Messungen verfügbar:

- Status bei Aufnahme (PRÄ)
- Status bei Katamnese (POST)
- Retrospektive PRÄ-Messung bei Katamnese (RETRO)
- Direkte Veränderungsmessung bei Katamnese (dVM)

Tabelle 30 gibt die Rohwerte der 13 Originalitems für alle vier Messungen wieder. Einzelne fehlende Werte wurden mit dem EM-Algorithmus geschätzt (zur Methodik vgl. Abschnitt 3.2.1).

Berücksichtigung fanden alle Patienten, für die neben der Aufnahme- und Katamnese-Messung von GB13 auch eine Entlassmessung auf den beiden HAQ-Skalen sowie der Depressivitäts-Skala im SCL-90-R vorliegt, um eine Validierung des hier neu entwickelten Ergebniskriteriums EQ im Sinne der Ausführungen in Abschnitt 3.2.5 und 3.2.6 zu ermöglichen. Die Analysestichprobe mit vollständiger Dreipunktmessung (Aufnahme, Entlassung und Katamnese) liegt bei N = 487 Patienten, was einem Anteil von 56,8% an der Ausgangsstichprobe von N = 858 Patienten in Studie D entspricht. Eine ausführliche Analyse zur Repräsentativität dieser Teilstichprobe findet sich bei Schmidt et al. (2003). Mit Ausnahme des Alters ist die Analysestichprobe von N = 487 Patienten hinsichtlich der in Abschnitt 3.1.3 präsentierten Merkmale weitgehend repräsentativ für die Ausgangsstichprobe von N = 858. So ist die Analysestichprobe im Durchschnitt mit 45,0 Jahren etwas älter als die Ausgangsstichprobe mit 43,8 Jahren).

Die in Tabelle 30 dargestellten Itemkennwerte bilden das „Ausgangsmaterial" für alle weiteren Indexbildungen und Analysen in diesem Abschnitt. Sämtliche Standardisierungen der Statusmessungen sowie der daraus abgeleiteten Indices für die iVM und qVM werden anhand der Item-Streuungen der PRÄ-Messung durchgeführt. Hierbei sind zwei Überlegungen ausschlaggebend:

- Die PRÄ-Messung stellt die beste Schätzung für die Ausgangssituation der unbehandelten Patienten dar (nähere Ausführungen hier in Abschnitt 2.1.3).
- Die Statusmessungen sowie die daraus abgeleiteten Indices sollten für die hier durchgeführte erste Erprobung untereinander vergleichbar sein, was nur durch die Verwendung einer einheitlichen Streuung zur Standardisierung realisierbar ist.

Bei einer ausschließlichen qVM sind keine echten PRÄ-Messungen und entsprechenden Streuungen verfügbar, so dass in diesem Fall die Item-Streuungen der qVM Verwendung finden können.

Bei den dVM werden aufgrund der abweichenden fünfstufigen Metrik die dVM-Item-Streuungen zur Berechnung von standardisierten Abweichungen vom Erwartungswert im Sinne der EMEK-Variante b aus Kapitel 3 verwendet.

Tabelle 30. Status- und Veränderungsmessung mit der Skala GB13

	Gesundheitliches Befinden Skala GB13	**Status PRÄ**		**Status POST**		**Status RETRO**		**dVM POST**	
Nr.	Item	$M_{PRÄ}$	$SD_{PRÄ}$	M_{POST}	SD_{POST}	M_{RETRO}	SD_{RETRO}	M_{dVM}	SD_{dVM}
1	Gesundheitszustand	3,95	1,18	3,26	1,14	4,34	1,07	2,18	1,07
2	Körperliches Befinden	3,93	1,17	3,25	1,13	4,29	1,03	2,35	1,02
3	Seelisches Befinden	4,71	1,07	3,49	1,27	5,09	0,85	2,18	1,14
4	Körperliche Belastbarkeit	4,08	1,18	3,33	1,19	4,45	1,06	2,39	1,03
5	Emotionale Belastbarkeit	4,60	1,10	3,60	1,18	5,01	0,79	2,26	1,05
6	Allgemeinbefinden	4,14	1,02	3,32	1,16	4,62	0,83	2,18	1,08
7	Körperliche Leistungsfähigkeit	4,03	1,20	3,31	1,16	4,29	1,07	2,47	1,03
8	Geistige Leistungsfähigkeit	3,67	1,25	2,99	1,14	4,08	1,14	2,32	1,00
9	Entspannungsfähigkeit	4,59	1,15	3,58	1,20	4,94	0,85	2,39	0,95
10	Kontaktfähigkeit	3,71	1,31	3,16	1,21	4,20	1,22	2,41	0,96
11	Selbstvertrauen	4,32	1,30	3,43	1,22	4,78	1,08	2,37	1,00
12	Umgang mit Enttäuschungen	4,60	1,18	3,72	1,17	4,88	0,97	2,51	0,94
13	Umgang mit Alltagsbelastungen	4,23	1,14	3,47	1,18	4,63	0,97	2,37	1,04

Anmerkungen. *N* = 487 Patienten. *M* – Mittelwert. *SD* – Streuung.

Status PRÄ: Messung bei Aufnahme in die Klinik. Status POST: Messung zum Zeitpunkt der 1-Jahreskatamnese. Codierung: „Wie würden Sie Ihr derzeitiges gesundheitliches Befinden einschätzen?" mit den Ankreuzalternativen 1 – sehr gut, 2 – gut, 3 – befriedigend, 4 – ausreichend, 5 – schlecht, 6 – sehr schlecht.

Status RETRO: Retrospektive PRÄ-Messung zum Zeitpunkt der 1-Jahres-Katamnese. Codierung: „Wie würden Sie Ihr damaliges gesundheitliches Befinden einschätzen?" mit den Ankreuzalternativen 1 – sehr gut, 2 – gut, 3 – befriedigend, 4 – ausreichend, 5 – schlecht, 6 – sehr schlecht.

dVM POST: Direkte Veränderungsmessung zum Zeitpunkt der 1-Jahres-Katamnese. Codierung: „Wenn Sie Ihr derzeitiges Befinden ein Jahr nach der Reha ganz kritisch mit Ihrem Befinden vor Aufnahme in die Klinik vergleichen: Welche Veränderungen haben sich ergeben?" mit den Ankreuzalternativen 1 – deutlich verbessert, 2 – etwas verbessert, 3 – unverändert, 4 – etwas verschlechtert, 5 – deutlich verschlechtert.

5.1.2 Standardisierung der iVM, qVM und dVM

Zunächst wurden klassische Effektgrößen zur indirekten Veränderungsmessung (iVM) nach Gleichung 13 berechnet. Für den PRÄ-POST-Vergleich von Item 1 „Gesundheitszustand" resultiert eine exemplarische Beispielberechnung für den Vergleich der Mittelwerte nach Gleichung 21.

$$d_{iVM} = \frac{M_{\mathrm{Pr}\ddot{a}} - M_{Post}}{SD_{\mathrm{Pr}\ddot{a}}} = \frac{3{,}95 - 3{,}26}{1{,}18} = +0{,}58 \qquad (21)$$

Die Effektgrößen zur Standardisierung der Mittelwertsdifferenzen zur quasi-indirekten Veränderungsmessung (qVM) wurden für Item 1 dementsprechend nach Formel 22 berechnet.

$$d_{qVM} = \frac{M_{Retro} - M_{Post}}{SD_{Prä}} = \frac{4{,}34 - 3{,}26}{1{,}18} = +0{,}92 \qquad (22)$$

Die direkten Veränderungsmessungen (dVM) wurden ebenfalls einer z-Standardisierung unterzogen. Dies entspricht der Berechnungsvariante b in Abschnitt 3.1.1. Der Erwartungswert *E* für die Ankreuzung „unverändert" beträgt dabei für sämtliche 13 Items 3,00. Für Item 1 resultiert die Gleichung 23 wiedergegebene exemplarische Berechnung der über alle Patienten gemittelten standardisierten Abweichung vom Erwartungswert.

$$z_{dVM} = \frac{E - M_{dVM}}{SD_{dVM}} = \frac{3{,}00 - 2{,}18}{1{,}07} = +0{,}77 \qquad (23)$$

Für sämtliche 13 Items in Tabelle 30 wurden entsprechende Effektgrößen bzw. z-Werte anhand der Item-Mittelwerte und –Streuungen berechnet.

5.1.3 Berechnung der Composit-Kriterien für die 13 Items

Für die Skala GB13 sollen auf Item-Ebene 13 Composit-Werte (EQ) zur Bewertung der Ergebnisqualität berechnet werden. Allerdings stehen für die 13 Items der Skala GB13 noch keine Bezugsnormen für kranke und gesunde Stichproben zur Berechnung des Cut-Off-Wertes *c* nach Jacobson zur Verfügung, weshalb für die Berechnung von *c* hier ersatzweise die PRÄ- und POST-Mittelwerte sowie die PRÄ- und POST-Streuungen der Analyse-Stichprobe verwendet werden.

Setzt man für Item 1 die Werte aus Tabelle 30 in Gleichung 10 ein, so resultiert die Beispielberechnung in Gleichung 24. Diese wurde für alle 13 Items entsprechend der jeweiligen Item-Kennwerte der PRÄ- und POST-Messung durchgeführt.

$$c = \frac{M_{POST} \cdot SD_{PRÄ} + M_{PRÄ} \cdot SD_{POST}}{SD_{PRÄ} + SD_{POST}} = \frac{1{,}18 \cdot 3{,}26 + 1{,}14 \cdot 3{,}95}{1{,}18 + 1{,}14} = 3{,}60 \qquad (24)$$

Im nächsten Schritt werden die drei Statusmessungen (PRÄ-Messung, POST-Messung sowie RETRO) unter Verwendung von *c* einer z-Standardisierung unterzogen. Die exemplarische Beispiel-Berechnung erfolgt wieder jeweils für Item 1 „Gesundheitszustand" aus

Tabelle 30. Es resultiert nach Gleichung 25 ein gemittelter z-Wert der PRÄ-Messung von –0,30 für die Patientenstichprobe.

$$z_{\mathrm{Pr}\ddot{a}} = \frac{c - M_{\mathrm{Pr}\ddot{a}}}{SD_{\mathrm{Pr}\ddot{a}}} = \frac{3{,}60 - 3{,}95}{1{,}18} = -0{,}30 \tag{25}$$

Für die POST-Messung resultiert nach Gleichung 26 ein über alle Patienten gemittelter z-Wert von +0,29.

$$z_{Post} = \frac{c - M_{Post}}{SD_{\mathrm{Pr}\ddot{a}}} = \frac{3{,}60 - 3{,}26}{1{,}18} = +0{,}29 \tag{26}$$

Für die retrospektive PRÄ-Messung wird der entsprechende Item-Mittelwert sowie die Streuung der echten PRÄ-Messung zur Standardisierung verwendet (Gleichung 27).

$$z_{Post} = \frac{c - M_{\mathrm{Re}tro}}{SD_{\mathrm{Pr}\ddot{a}}} = \frac{3{,}60 - 4{,}34}{1{,}18} = -0{,}63 \tag{27}$$

Die Differenz zwischen $z_{PRÄ}$ und z_{POST} entspricht rechnerisch der für Item 1 berechneten Effektgröße zur klassischen indirekten Veränderungsmessung von d_{iVM} = +0,59. Die Differenz zwischen z_{RETRO} und z_{POST} entspricht der Effektgröße zur quasi-indirekten Veränderungsmessung von d_{qVM} = +0,92.

Das Composit-Kriterium EQ für die iVM berechnet sich nach Gleichung 28 (exemplarische Beispielberechnung wieder anhand von Item 1).

$$EQ_{iVM} = z_{post} - 0{,}5 \cdot z_{prä} = +0{,}29 - 0{,}5 \cdot (-0{,}30) = +0{,}44 \tag{28}$$

Auch für die qVM wurde ein entsprechendes Composit-Kriterium EQ berechnet (exemplarische Beispielberechnung in Gleichung 29).

$$EQ_{qVM} = z_{post} - 0{,}5 \cdot z_{retro} = +0{,}29 - 0{,}5 \cdot (-0{,}63) = +0{,}61 \tag{29}$$

Für sämtliche 13 Items in Tabelle 30 wurden entsprechende Cut-Off-Werte *c*, standardisierte z-Werte sowie Composit-Kriterien EQ wie hier expliziert berechnet.

5.1.4 Aggregation zu multiplen Ergebniskriterien

Die 13 Items von GB13 sowie die daraus abgeleiteten z-standardisierten Werte, Effektgrößen und Composit-Maße stellen singuläre Ergebniskriterien im Sinne von Schmidt et al. (1987) dar. Zur Bildung eines multiplen Ergebniskriteriums wird eine Aggregation der 13 Einzelaspekte zu einer Gesamtskala für jede hier behandelte methodische Variante vorgenommen.

Die Skalenberechnung erfolgt dabei jeweils durch Aufsummierung der 13 Einzelkomponenten und anschließende Division durch 13. Eine Behandlung von fehlenden Werten ist nicht mehr erforderlich, dieser Schritt wurde bereits bei der Aufbereitung des Rohmaterials mittels EM-Schätzung vollzogen, so dass für sämtliche 487 Patienten durchgängig vollständige Messungen vorliegen.

Insgesamt werden acht Skalenvarianten von GB13 berechnet, die im Rahmen der weiteren Datenanalysen entsprechende Bezeichnungen erhalten:

- Die aus den 13 z-standardisierten PRÄ-Messungen ($z_{PRÄ}$) berechnete Skala zur PRÄ-Messung bei Aufnahme in die Klinik wird mit PRÄ_13 bezeichnet.
- Die aus den 13 z-standardisierten POST-Messungen (z_{POST}) berechnete Skala zur katamnestischen POST-Messung wird mit POST_13 bezeichnet.
- Die aus den 13 z-standardisierten retrospektiven PRÄ-Messungen (z_{RETRO}) zur rückwirkenden Einschätzung des Befindens berechnete Skala wird mit RETRO_13 bezeichnet.
- Die Skala zur indirekten Veränderungsmessung, die sich aus den 13 klassisch berechneten PRÄ-POST-Effektgrößen (d_{iVM}) zusammensetzt, wird nachfolgend mit iVM_13 bezeichnet.
- Die Skala zur quasi-indirekten Veränderungsmessung, die sich aus den 13 entsprechend berechneten Effektgrößen (d_{qVM}) zusammensetzt, wird nachfolgend mit qVM_13 bezeichnet.
- Die Skala zur direkten Veränderungsmessung, die sich aus den 13 z-standardisierten Abweichungen vom Erwartungswert (z_{dVM}) zusammensetzt, wird nachfolgend mit dVM_13 bezeichnet.
- Das multiple Composit-Kriterium, das auf der iVM beruht und die 13 entsprechenden EQ_{iVM} – Items zusammenfasst, wird nachfolgend mit EQ_{iVM}_13 bezeichnet.
- Das multiple Composit-Kriterium, das auf der qVM beruht und die 13 entsprechenden EQ_{qVM} – Items zusammenfasst, wird nachfolgend mit EQ_{qVM}_13 bezeichnet.

5.1.5 Fragestellungen zur Analyse von EQ

Wie stellen sich die Kennwerte sowie die Vorhersagbarkeit der beiden hier neu gebildeten Composit-Kriterien EQ_{iVM}_13 bzw. EQ_{qVM}_13 im Vergleich zu einfachen POST-Messungen, zu herkömmlichen iVM, qVM sowie dVM dar?

Fragestellung 1: Item- und Skaleneigenschaften

Wie hoch sind die Mittelwerte, Streuungen und korrigiertem Item-Trennschärfen der 13 Einzelkomponenten der hier analysierten acht Skalenvarianten von GB13? Wie hoch sind die Mittelwerte, Streuungen und Reliabilitäten der acht Gesamtskalen? Wie sind die Verteilungseigenschaften der Skalen? Wie stellt sich die Bewertung der Ergebnisqualität der Patienten mit den unterschiedlichen methodischen Ansätzen dar? Wie hoch sind die Zusammenhänge (Interkorrelationen) zwischen den Skalen?

Hypothesen zu Fragestellung 1:

- Auf dem Hintergrund der bislang vorliegenden und in der EQUA-Studie berichteten Ergebnisse zur Skala GB13 (Schmidt et al., 2003) sowie zur iVM, qVM und dVM sind für alle acht hier betrachteten Status- und Veränderungsmaße und damit auch für das neu gebildete Composit-Maß EQ_{iVM}_13 bzw. EQ_{qVM}_13 hohe Werte für die Item-Trennschärfen und Skalenreliabilitäten zu erwarten. Der Wertebereich der standardisierten Maße dürfte sich gemäß der Standardmetrik dabei im Bereich –3 bis +3 bewegen.
- Für die retrospektive Skalenvariante wird in der EQUA-Studie eine deutlich kritischere rückwirkende Einschätzung des Befindens als bei der echten PRÄ-Messung berichtet. qVM führten in der EQUA-Studie daher zu einer günstigeren Erfolgsbewertung als iVM. Entsprechende Resultate sind rechnerisch auch für die aus den iVM und qVM abgeleiteten beiden Composit-Skalen EQ_{iVM}_13 und EQ_{qVM}_13 zu erwarten.
- Zwischen den unterschiedlichen Varianten (iVM, qVM und dVM) zur Ergebnismessung wurden in der EQUA-Studie Korrelationen in mittlerer Größenordnung berichtet. Zwischen EQ_{iVM}_13 und POST_13 bzw. iVM_13 dürften sich hohe Skalen-Interkorrelationen ergeben. Gleiches gilt für die Variante von EQ, die auf qVM beruht. Interessant dürfte die Antwort auf die Frage ausfallen, wie hoch die Korrelation zwischen EQ_{iVM}_13 und PRÄ_13 sein wird. So werden zwischen PRÄ- und POST-Messung positive, zwischen PRÄ-Messung und iVM hingegen negative Zusammenhänge berichtet. Es wird daher vermutet, dass sich die beiden gegenläufigen Effekte aufheben und sich EQ_{iVM}_13 wie bereits in 4.2.5 beschrieben als relativ unabhängig von der PRÄ-Messung erweisen dürfte. Für die qVM-Variante von EQ wird das gleiche Resultat erwartet.

Fragestellung 2: Validierung

Inwieweit lassen sich die hier berechneten unterschiedlichen multiplen Ergebniskriterien durch Stichproben- und Methodenmerkmale vorhersagen? Zur Validierung werden entsprechende multiple Regressionsgleichungen (vgl. Abschnitt 3.2.5) bzw. Pfadanalysen (vgl. Abschnitt 3.2.6) berechnet. Darüber hinaus werden globale Einschätzungen der Patienten zum Zeitpunkt der 1-Jahres-Katamnese hinsichtlich der Therapiezielerreichung zur Validierung verwendet.

Hypothesen zu Fragestellung 2:

- Für die Vorhersage der verschiedenen Skalenvarianten von GB13 aus Stichprobenmerkmalen werden ähnliche Ergebnisse wie bei der Validierung der Skala EMEK_27 erwartet (vgl. Abschnitt 3.2.5). Die Varianzaufklärung sollte bei Einbeziehung der gleichen Prädiktoren etwa 10% erwarten. Insbesondere Rentenantragstellung bei Aufnahme dürfte sich als ungünstiger Prädiktor für die katamnestische Ergebnisqualität erweisen.
- Für die Vorhersage aus den Prozessmerkmalen in den Pfadanalysen (vgl. hierzu die Ergebnisse in Abschnitt 3.2.6) wird ausgehend von den Ergebnissen bei der Validierung von EMEK_27 eine Varianzaufklärung von etwa 40% erwartet.
- Es wird erwartet, dass bei den beiden hier neu eingeführten Varianten EQ_{iVM}_13 und EQ_{qVM}_13 aufgrund der angenommenen besseren Differenzierungsfähigkeit der neuen Methode bei der Erfolgsbewertung insgesamt eine etwas höhere Varianzaufklärung gelingt als bei den Skalen POST_13, dVM_13, iVM_13 sowie qVM_13.

5.2 Ergebnisse

Neben den Item- und Skalenkennwerten der hier berechneten Varianten von GB13 werden auch die Skaleninterkorrelationen sowie Validitätsmaße wie Vorhersagbarkeit durch Patientenmerkmale oder Prozessmerkmale berichet. Die Ergebnisdarstellung zur Erprobung von EQ schließt mit einer Darstellung der Skalenmittelwerte in Abhängigkeit vom Grad der globalen Therapiezielerreichung, den die Patienten ein Jahr nach Ende der Behandlung angeben.

5.2.1 Item- und Skalenkennwerte

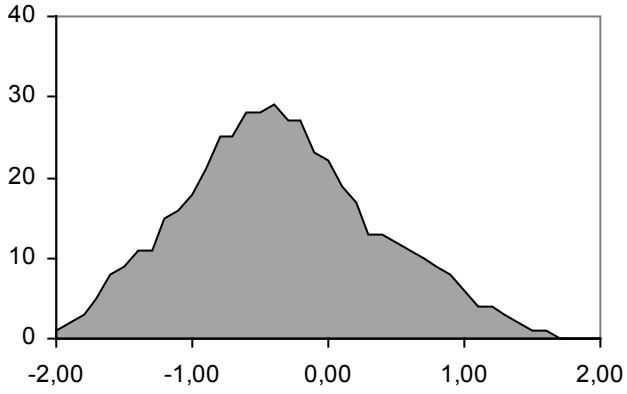
Tabelle 31 enthält die Kennwerte für die z-standardisierten Items der Skala PRÄ_13. Die Item-Mittelwerte liegen zwischen –0,22 (Kontaktfähigkeit) und –0,52 (Seelisches Befinden). Erwartungsgemäß berichten die Patienten zum Aufnahme-Zeitpunkt damit vor allem hinsichtlich ihres psychischen Befindens einen besonders kritischen Zustand. Die Item-Streuungen betragen aufgrund der Standardisierung anhand der gleichen Messung

durchgängig 1,00. Die korrigierten Item-Trennschärfen liegen zwischen .51 (Kontaktfähigkeit) und .80 (Allgemeinbefinden) und weisen damit ausgezeichnete Werte auf. Entsprechend hoch ist die interne Konsistenz der Gesamtskala mit einem Cronbachs Alpha von .91. Häufigkeitsverteilung und Skalenkennwerte von PRÄ_13 sind in Abbildung 24 wiedergegeben.

Tabelle 31. Itemkennwerte von PRÄ_13

Nr.	Item	*Min*	*Max*	*M*	*SD*	r_{it}
1	Gesundheitszustand	-2,03	1,35	-0,30	1,00	.62
2	Körperliches Befinden	-2,07	2,22	-0,30	1,00	.62
3	Seelisches Befinden	-1,73	2,96	-0,52	1,00	.67
4	Körperliche Belastbarkeit	-1,95	2,30	-0,32	1,00	.64
5	Emotionale Belastbarkeit	-1,71	1,93	-0,44	1,00	.70
6	Allgemeinbefinden	-2,19	1,72	-0,38	1,00	.80
7	Körperliche Leistungsfähigkeit	-1,94	2,22	-0,30	1,00	.62
8	Geistige Leistungsfähigkeit	-2,14	1,85	-0,28	1,00	.62
9	Entspannungsfähigkeit	-1,66	2,68	-0,43	1,00	.52
10	Kontaktfähigkeit	-1,97	1,86	-0,22	1,00	.51
11	Selbstvertrauen	-1,63	2,19	-0,35	1,00	.65
12	Umgang mit Enttäuschungen	-1,56	2,67	-0,37	1,00	.60
13	Umgang mit Alltagsbelastungen	-1,88	2,50	-0,33	1,00	.71

Anmerkungen. *N* = 487 Patienten. Aktuelles Befinden bei Aufnahme in die Klinik. *Min* – Kleinster Wert, *Max* – Größter Wert, *M* – Mittelwert, *SD* – Streuung, r_{it}: Korrigierte Item-Trennschärfe.



Skala PRÄ_13

Messung des aktuellen Befindens bei Aufnahme in die Klinik

N = 487
Min = -1,88; *Max* = +1,53
M = -0,35; *SD* = 0,70; *SE* = 0,03
Schiefe = +0,27; Wölbung = -0,34
Cronbachs Alpha = .91

Abbildung 24. Skalenkennwerte von PRÄ_13

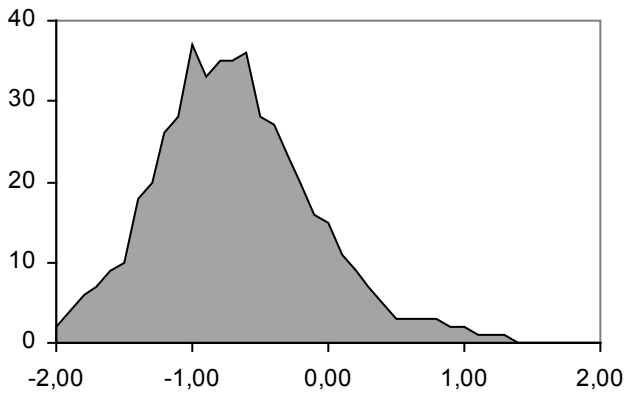
Tabelle 32 enthält die Kennwerte für die z-standardisierten Items der Skala RETRO_13. Die Item-Mittelwerte liegen zwischen –0,52 (körperliche Leistungsfähigkeit) und –0,88 (Seelisches Befinden). Erwartungsgemäß berichten die Patienten zum Aufnahme-

Zeitpunkt damit vor allem hinsichtlich ihres psychischen Befindens einen besonders kritischen Zustand. Die Item-Streuungen liegen zwischen 0,72 und 0,92 und sind somit geringer als die Streuungen der echten PRÄ-Messung. Die korrigierten Item-Trennschärfen liegen zwischen .46 (Kontaktfähigkeit) und .76 (Allgemeinbefinden). Die interne Konsistenz der Gesamtskala liegt .91. Häufigkeitsverteilung und Skalenkennwerte von RETRO_13 sind in Abbildung 25 wiedergegeben.

Tabelle 32. Itemkennwerte von RETRO_13

Nr.	Item	*Min*	*Max*	*M*	*SD*	r_{it}
1	Gesundheitszustand	-2,03	2,20	-0,63	0,90	.64
2	Körperliches Befinden	-2,07	2,22	-0,60	0,89	.62
3	Seelisches Befinden	-1,73	2,02	-0,88	0,80	.65
4	Körperliche Belastbarkeit	-1,95	2,30	-0,63	0,90	.59
5	Emotionale Belastbarkeit	-1,71	1,93	-0,81	0,72	.64
6	Allgemeinbefinden	-2,19	1,72	-0,84	0,81	.76
7	Körperliche Leistungsfähigkeit	-1,94	2,22	-0,52	0,89	.63
8	Geistige Leistungsfähigkeit	-2,14	1,85	-0,61	0,91	.62
9	Entspannungsfähigkeit	-1,66	1,81	-0,74	0,74	.52
10	Kontaktfähigkeit	-1,97	1,86	-0,59	0,93	.46
11	Selbstvertrauen	-1,63	2,19	-0,70	0,82	.58
12	Umgang mit Enttäuschungen	-1,56	1,82	-0,61	0,82	.58
13	Umgang mit Alltagsbelastungen	-1,88	1,63	-0,67	0,85	.70

Anmerkungen. *N* = 487 Patienten. Retrospektive PRÄ-Messung. *Min* – Kleinster Wert, *Max* – Größter Wert, *M* – Mittelwert, *SD* – Streuung, r_{it}: Korrigierte Item-Trennschärfe.



Skala RETRO_13

Retrospektive Messung des damaligen Befindens bei Aufnahme in die Klinik (erhoben 1 Jahr nach Ende der Reha)

N = 487
Min = -1,88; *Max* = +1,41
M = -0,68; *SD* = 0,57; *SE* = 0,03
Schiefe = +0,57; Wölbung = +0,59
Cronbachs Alpha = .90

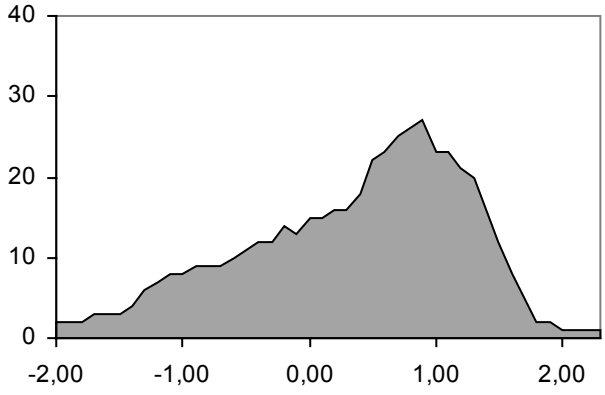
Abbildung 25. Skalenkennwerte von RETRO_13

Tabelle 33 enthält die Kennwerte für die z-standardisierten Items der Skala POST_13. Die Item-Mittelwerte liegen zwischen +0,20 (Kontaktfähigkeit) und +0,62 (Seelisches Befinden). Die Item-Streuungen liegen zwischen 0,91 und 1,13. Die korrigierten Item-Trennschärfen liegen zwischen .72 und .89. Die interne Konsistenz der Gesamtskala ist sehr hoch und liegt bei .97. Häufigkeitsverteilung und Skalenkennwerte von POST_13 sind in Abbildung 26 wiedergegeben.

Tabelle 33. Itemkennwerte von POST_13

Nr.	Item	*Min*	*Max*	*M*	*SD*	r_{it}
1	Gesundheitszustand	-2,03	2,20	0,29	0,97	.83
2	Körperliches Befinden	-2,07	2,22	0,29	0,97	.82
3	Seelisches Befinden	-1,73	2,96	0,62	1,19	.88
4	Körperliche Belastbarkeit	-1,95	2,30	0,32	1,01	.83
5	Emotionale Belastbarkeit	-1,71	2,84	0,47	1,07	.86
6	Allgemeinbefinden	-2,19	2,69	0,43	1,13	.89
7	Körperliche Leistungsfähigkeit	-1,94	2,22	0,30	0,97	.79
8	Geistige Leistungsfähigkeit	-2,14	1,85	0,26	0,91	.81
9	Entspannungsfähigkeit	-1,66	2,68	0,44	1,04	.78
10	Kontaktfähigkeit	-1,97	1,86	0,20	0,93	.72
11	Selbstvertrauen	-1,63	2,19	0,33	0,94	.82
12	Umgang mit Enttäuschungen	-1,56	2,67	0,37	0,99	.80
13	Umgang mit Alltagsbelastungen	-1,88	2,50	0,34	1,03	.86

Anmerkungen. *N* = 487 Patienten. Aktuelles Befinden ein Jahr nach Entlassung aus der Klinik. *Min* – Kleinster Wert, *Max* – Größter Wert, *M* – Mittelwert, *SD* – Streuung, r_{it}: Korrigierte Item-Trennschärfe.



Skala POST_13

Messung des aktuellen Befindens
ein Jahr nach Entlassung aus der Klinik

N = 487
Min = -1,88; *Max* = +2,33
M = +0,36; *SD* = 0,86; *SE* = 0,04
Schiefe = -0,55; Wölbung = -0,37
Cronbachs Alpha = .97

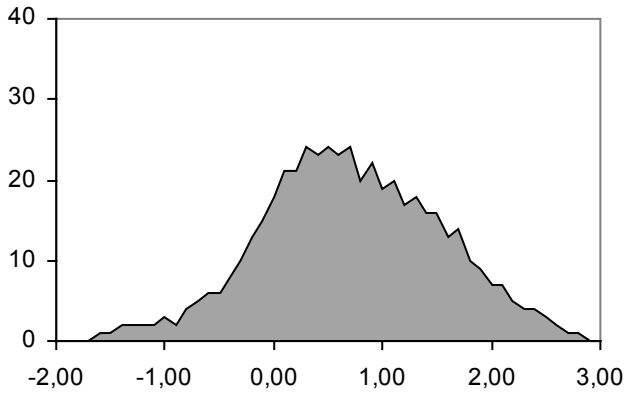
Abbildung 26. Skalenkennwerte von POST_13

Tabelle 34 enthält die Kennwerte für die klassischen PRÄ-POST-Effektgrößen (iVM_13). Die Item-Mittelwerte liegen zwischen +0,42 (Kontaktfähigkeit) und +1,15 (Seelisches Befinden). Die Item-Streuungen liegen zwischen 1,00 und 1,29. Die korrigierten Item-Trennschärfen liegen zwischen .57 und .80. Die interne Konsistenz der Gesamtskala liegt bei .93. Häufigkeitsverteilung und Skalenkennwerte von iVM_13 sind in Abbildung 27 wiedergegeben.

Tabelle 34. Itemkennwerte von iVM_13

Nr.	Item	*Min*	*Max*	*M*	*SD*	*rit*
1	Gesundheitszustand	-2,54	3,38	0,59	1,05	.63
2	Körperliches Befinden	-3,43	3,43	0,58	1,06	.62
3	Seelisches Befinden	-2,81	4,69	1,15	1,29	.76
4	Körperliche Belastbarkeit	-2,55	3,40	0,64	1,05	.70
5	Emotionale Belastbarkeit	-2,73	4,55	0,91	1,21	.71
6	Allgemeinbefinden	-1,96	3,91	0,81	1,20	.80
7	Körperliche Leistungsfähigkeit	-3,33	3,33	0,60	1,02	.66
8	Geistige Leistungsfähigkeit	-2,39	3,98	0,54	1,01	.63
9	Entspannungsfähigkeit	-3,47	4,34	0,88	1,20	.63
10	Kontaktfähigkeit	-3,07	3,83	0,42	1,00	.57
11	Selbstvertrauen	-3,06	3,82	0,68	1,02	.70
12	Umgang mit Enttäuschungen	-2,54	3,38	0,74	1,12	.63
13	Umgang mit Alltagsbelastungen	-2,63	3,50	0,67	1,11	.71

Anmerkungen. *N* = 487 Patienten. Indirekte Veränderungsmessung (standardisierte Mittelwertsdifferenzen zwischen Aufnahme- und Katamnese-Status). *Min* – Kleinster Wert, *Max* – Größter Wert, *M* – Mittelwert, *SD* – Streuung, r_{it}: Korrigierte Item-Trennschärfe.



Skala iVM_13

Indirekte Veränderungsmessung (Vergleich zwischen Aufnahme- und Katamnese-Messung des Befindens)

N = 487
Min = -1,45; *Max* = +2,77
M = +0,71; *SD* = 0,81; *SE* = 0,04
Schiefe = -0,04; Wölbung = -0,22
Cronbachs Alpha = .93

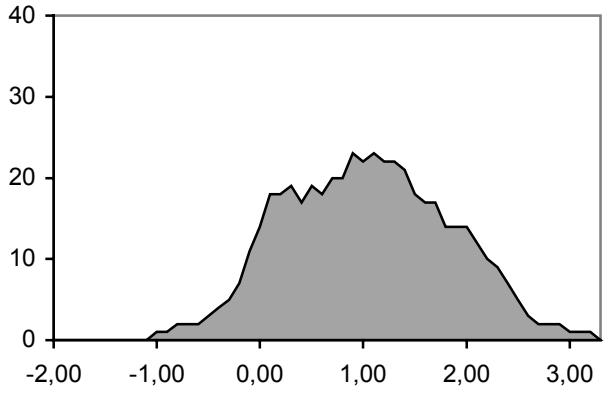
Abbildung 27. Skalenkennwerte von iVM_13

Tabelle 35 enthält die Kennwerte für die standardisierten Mittelwertsdifferenzen zwischen retrospektiver PRÄ-Messung und katamnestischer Messung (qVM_13). Die Item-Mittelwerte liegen zwischen +0,79 (Kontaktfähigkeit) und +1,51 (Seelisches Befinden) und sind damit erwartungsgemäß deutlich höher als bei iVM_13. Die Item-Streuungen liegen zwischen 0,87 und 1,28. Die korrigierten Item-Trennschärfen liegen zwischen .65 und .83. Die interne Konsistenz der Gesamtskala liegt bei .93. Häufigkeitsverteilung und Skalenkennwerte von qVM_13 sind in Abbildung 28 wiedergegeben.

Tabelle 35. Itemkennwerte von qVM_13

Nr.	Item	*Min*	*Max*	*M*	*SD*	*rit*
1	Gesundheitszustand	-2,54	3,38	0,92	1,06	.69
2	Körperliches Befinden	-3,43	3,43	0,89	1,01	.71
3	Seelisches Befinden	-1,87	4,69	1,51	1,28	.81
4	Körperliche Belastbarkeit	-2,55	4,25	0,96	1,09	.74
5	Emotionale Belastbarkeit	-1,82	4,55	1,28	1,16	.78
6	Allgemeinbefinden	-1,96	3,91	1,27	1,15	.83
7	Körperliche Leistungsfähigkeit	-2,50	4,17	0,81	0,99	.72
8	Geistige Leistungsfähigkeit	-1,59	3,98	0,87	0,95	.72
9	Entspannungsfähigkeit	-1,74	4,34	1,18	1,03	.68
10	Kontaktfähigkeit	-2,30	3,83	0,79	0,87	.65
11	Selbstvertrauen	-2,29	3,82	1,03	0,91	.75
12	Umgang mit Enttäuschungen	-1,69	3,38	0,98	0,96	.67
13	Umgang mit Alltagsbelastungen	-1,75	3,50	1,02	1,01	.74

Anmerkungen. *N* = 487 Patienten. Quasi-indirekte Veränderungsmessung (standardisierte Mittelwertsdifferenzen zwischen retrospektiver Aufnahmemessung und aktueller Messung des Befindens zum Katamnese-Zeitpunkt). *Min* – Kleinster Wert, *Max* – Größter Wert, *M* – Mittelwert, *SD* – Streuung, r_{it}: Korrigierte Item-Trennschärfe.



Skala qVM_13

Quasi-indirekte Veränderungsmessung (Vergleich zwischen erinnertem Befinden bei Aufnahme und aktuellem Befinden bei Katamnese)

N = 487
Min = -1,47; *Max* = +3,25
M = +1,04; *SD* = 0,81; *SE* = 0,04
Schiefe = +0,06; Wölbung = -0,26
Cronbachs Alpha = .94

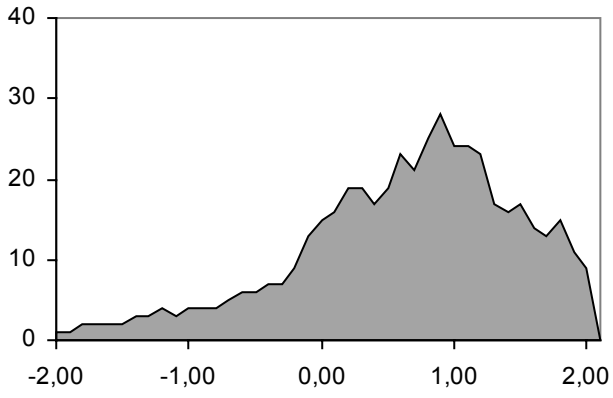
Abbildung 28. Skalenkennwerte von qVM_13

Tabelle 36 enthält die Kennwerte für die standardisierten Mittelwertsdifferenzen der direkten Veränderungsmessungen (dVM_13). Die Item-Mittelwerte liegen zwischen +0,52 (Körperliche Leistungsfähigkeit) und +0,76 (Gesundheitszustand). Die Item-Streuungen liegen aufgrund der Standardisierung anhand der gleichen Messung einheitlich bei 1,00. Die korrigierten Item-Trennschärfen liegen zwischen .65 und .83. Die interne Konsistenz der Gesamtskala liegt bei .97. Häufigkeitsverteilung und Skalenkennwerte von dVM_13 sind in Abbildung 29 wiedergegeben.

Tabelle 36. Itemkennwerte von dVM_13

Nr.	Item	*Min*	*Max*	*M*	*SD*	*rit*
1	Gesundheitszustand	-1,87	1,87	0,76	1,00	.82
2	Körperliches Befinden	-1,96	1,96	0,64	1,00	.81
3	Seelisches Befinden	-1,76	1,76	0,72	1,00	.86
4	Körperliche Belastbarkeit	-1,95	1,95	0,60	1,00	.83
5	Emotionale Belastbarkeit	-1,90	1,90	0,70	1,00	.87
6	Allgemeinbefinden	-1,84	1,84	0,75	1,00	.90
7	Körperliche Leistungsfähigkeit	-1,94	1,94	0,52	1,00	.83
8	Geistige Leistungsfähigkeit	-1,99	1,99	0,68	1,00	.84
9	Entspannungsfähigkeit	-2,11	2,11	0,64	1,00	.79
10	Kontaktfähigkeit	-2,09	2,09	0,61	1,00	.70
11	Selbstvertrauen	-2,00	2,00	0,63	1,00	.83
12	Umgang mit Enttäuschungen	-2,14	2,14	0,52	1,00	.83
13	Umgang mit Alltagsbelastungen	-1,93	1,93	0,61	1,00	.87

Anmerkungen. *N* = 487 Patienten. Direkte Veränderungsmessung zum Katamnese-Zeitpunkt. *Min* – Kleinster Wert, *Max* – Größter Wert, *M* – Mittelwert, *SD* – Streuung, r_{it}: Korrigierte Item-Trennschärfe.



Skala dVM_13

Direkte Veränderungsmessung des Befindens ein Jahr nach Entlassung aus der Klinik

N = 487
Min = -1,96; *Max* = +1,96
M = +0,64; *SD* = 0,86; *SE* = 0,04
Schiefe = -0,71; Wölbung = +0,27
Cronbachs Alpha = .97

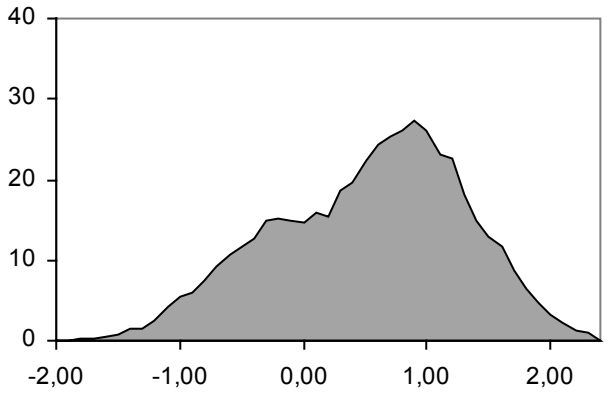
Abbildung 29. Skalenkennwerte von dVM_13

Tabelle 37 enthält die Kennwerte für das erste hier neu eingeführte Composit-Maß E-Q_{iVM}_13 zur Bewertung der Ergebnisqualität. Die Item-Mittelwerte liegen zwischen +0,31 (Kontaktfähigkeit) und +0,89 (Seelisches Befinden). Die Item-Streuungen erreichen Werte zwischen 0,82 und 1,14. Die korrigierten Item-Trennschärfen liegen zwischen .69 und .85. Die interne Konsistenz der Gesamtskala liegt bei .96 und ist damit sehr hoch. Häufigkeitsverteilung und Skalenkennwerte von EQ_{iVM}_13 sind in Abbildung 30 wiedergegeben. Neben einem großen Anteil von Patienten, die eine positive Ergebnisqualität mit Werten um +1,00 aufweisen, existiert auch eine Gruppe von Patienten, die offenbar kaum von der Behandlung profitiert haben (Werte um 0,00)

Tabelle 37. Itemkennwerte von EQ_{iVM}_13

Nr.	Item	*Min*	*Max*	*M*	*SD*	*rit*
1	Gesundheitszustand	-2,28	2,37	0,44	0,87	0,76
2	Körperliches Befinden	-2,75	2,39	0,43	0,88	0,75
3	Seelisches Befinden	-2,27	3,82	0,89	1,14	0,85
4	Körperliche Belastbarkeit	-2,25	2,85	0,48	0,90	0,80
5	Emotionale Belastbarkeit	-2,22	3,70	0,69	1,03	0,80
6	Allgemeinbefinden	-2,08	2,81	0,62	1,05	0,85
7	Körperliche Leistungsfähigkeit	-2,64	2,78	0,45	0,86	0,76
8	Geistige Leistungsfähigkeit	-1,87	2,91	0,40	0,82	0,76
9	Entspannungsfähigkeit	-2,13	3,51	0,66	1,00	0,74
10	Kontaktfähigkeit	-2,52	2,85	0,31	0,82	0,69
11	Selbstvertrauen	-2,35	3,00	0,51	0,84	0,80
12	Umgang mit Enttäuschungen	-2,09	2,79	0,54	0,86	0,80
13	Umgang mit Alltagsbelastungen	-2,25	3,00	0,51	0,95	0,80

Anmerkungen. *N* = 487 Patienten. Composit-Kriterium, basierend auf der POST-Messung und der PRÄ-POST-Differenz. *Min* – Kleinster Wert, *Max* – Größter Wert, *M* – Mittelwert, *SD* – Streuung, r_{it}: Korrigierte Item-Trennschärfe.



Skala EQ_{iVM}_13

Composit-Kriterium EQ, basierend auf der Status-Messung bei Katamnese und den Differenzwerten der iVM

N = 487
Min = -1,55; *Max* = +2,27
M = +0,53; *SD* = 0,76; *SE* = 0,03
Schiefe = -0,30; Wölbung = -0,51
Cronbachs Alpha = .96

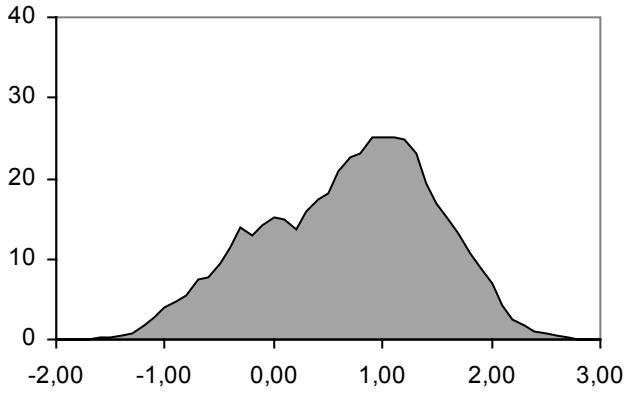
Abbildung 30. Skalenkennwerte von EQ*iVM*_13

Tabelle 38 enthält die Kennwerte für das zweite hier neu eingeführte Composit-Maß EQ_{qVM}_13 zur Bewertung der Ergebnisqualität. Die Item-Mittelwerte liegen zwischen +0,77 (Kontaktfähigkeit) und +1,07 (Seelisches Befinden). Die Item-Streuungen erreichen Werte zwischen 0,77 und 1,17. Die korrigierten Item-Trennschärfen liegen zwischen .76 und .87. Die interne Konsistenz der Gesamtskala liegt bei .97 und ist damit sehr hoch. Häufigkeitsverteilung und Skalenkennwerte von EQ_{qVM}_13 sind in Abbildung 31 wiedergegeben. Die Form der Verteilung ähnelt weitgehend der von EQ_{iVM}_13, lediglich der Mittelwert ist höher.

Tabelle 38. Itemkennwerte von EQ_{qVM}_13

Nr.	Item	*Min*	*Max*	*M*	*SD*	*rit*
1	Gesundheitszustand	-1,86	2,79	0,60	0,91	0,79
2	Körperliches Befinden	-2,75	2,82	0,59	0,89	0,80
3	Seelisches Befinden	-1,80	3,82	1,07	1,17	0,87
4	Körperliche Belastbarkeit	-2,25	3,27	0,64	0,95	0,82
5	Emotionale Belastbarkeit	-1,31	3,70	0,88	1,06	0,84
6	Allgemeinbefinden	-1,59	3,30	0,85	1,07	0,87
7	Körperliche Leistungsfähigkeit	-1,81	3,20	0,55	0,87	0,79
8	Geistige Leistungsfähigkeit	-1,87	2,91	0,57	0,81	0,81
9	Entspannungsfähigkeit	-1,26	3,51	0,81	0,97	0,76
10	Kontaktfähigkeit	-2,14	2,85	0,50	0,77	0,76
11	Selbstvertrauen	-1,58	3,00	0,68	0,83	0,84
12	Umgang mit Enttäuschungen	-1,28	2,40	0,66	0,83	0,83
13	Umgang mit Alltagsbelastungen	-1,38	3,00	0,68	0,93	0,82

Anmerkungen. *N* = 487 Patienten. Composit-Kriterium, basierend auf der POST-Messung und der quasi-indirekten Veränderungsmessung (Differenz zwischen erinnertem Befinden bei Aufnahme in die Klinik und aktuellem Befinden zum Katamnese-Zeitpunkt). *Min* – Kleinster Wert, *Max* – Größter Wert, *M* – Mittelwert, *SD* – Streuung, r_{it}: Korrigierte Item-Trennschärfe.



Skala EQ_{qVM}_13

Composit-Kriterium EQ, basierend auf der Status-Messung bei Katamnese und den Differenzwerten der qVM

N = 487
Min = -1,37; *Max* = +2,68
M = +0,70; *SD* = 0,78; *SE* = 0,04
Schiefe = -0,24; Wölbung = -0,51
Cronbachs Alpha = .97

Abbildung 31. Skalenkennwerte von EQ_{qVM}_13

5.2.2 Zusammenhänge zwischen den Skalen

Tabelle 39 gibt die Korrelationen zwischen allen acht berechneten Skalen von GB13 wieder. Die Zusammenhänge zwischen den Status- und Veränderungsskalen sind mit den bereits bei Schmidt et al. (2003) publizierten Korrelationen identisch, zur Interpretation und Diskussion der Ergebnisse sei daher auf das entsprechende Kapitel im Buch verwiesen. Erwähnenswert ist hier allerdings zum einen die negative Korrelation zwischen der iVM und der PRÄ-Messung, was für eine systematische Bevorzugung von schwerer beeinträchtigten Patienten durch die Methodik der iVM im Sinne einer systematischen Regression zur Mitte bei der Ergebnisbewertung spricht (vgl. Abschnitt 4.2.5). Zum anderen korrelieren dVM positiv mit der PRÄ-Messung, so dass die Behandlungsergebnisse von moderat beeinträchtigten Patienten durch die dVM eher in einem günstigeren Licht dargestellt werden.

Von besonderem Interesse sind die Zusammenhänge der beiden neu konstruierten Composit-Maßen EQ_{iVM}_13 und EQ_{qVM}_13 mit allen übrigen Skalen von GB13.

Zunächst fällt auf, dass EQ_{iVM}_13 keinen statistisch signifikanten Zusammenhang zur Baseline (PRÄ_13) aufweist. Damit bevorzugt der hier neu vorgeschlagene Ansatz zur Bewertung der Ergebnisqualität mit einem Composit-Maß im Gegensatz zur iVM oder dVM keine Patienten aufgrund ihres Beeinträchtigungsgrades zum Aufnahme-Zeitpunkt. Bei der Variante EQ_{qVM}_13 besteht hingegen, ähnlich wie bei der dVM, ein positiver Zusammenhang mit PRÄ_13.

EQ_{iVM}_13 korreliert sehr hoch sowohl mit der POST-Messung (r = .91) als auch mit der iVM (r = .90), was rein rechnerisch zu erwarten war, da sowohl POST-Messung als auch iVM gleichberechtigt in das Composit-Maß EQ_{iVM}_13 auf Item-Ebene eingegangen sind. Gleiches gilt analog für die Zusammenhänge des anderen Composit-Maßes EQ_{qVM}_13 mit der POST-Messung und der qVM. Hier fallen die Zusammenhänge mit .94 sogar noch höher aus, was dadurch bedingt sein mag, dass alle für die qVM erforderlichen Daten zum gleichen Zeitpunkt (1 Jahr nach der Reha) erhoben wurden. Auch die beiden Composit-Maße weisen mit Werten von .94 einen entsprechend engen Zusammenhang auf, da in beiden die POST-Messung enthalten ist.

Eine grafische Repräsentation der Zusammenhänge in Form eines Matrix-Scatterplots ist in Abbildung 32 wiedergegeben. Die starken positiven Korrelationen zwischen den multiplen Ergebniskriterien POST_13, iVM_13, qVM_13, dVM_13 sowie den beiden neuen Composit-Maßen zeigen sich in entsprechend schmalen von links oben nach rechts oben steigenden Ellipsen der Punktwolken.

Tabelle 39. Korrelationen zwischen den berechneten Skalen von GB13

	PRÄ_13	RETRO_13	POST_13	iVM_13	qVM_13	dVM_13	EQ_{iVM}_13
RETRO_13	.67	.					
POST_13	.48	.43	.				
iVM_13	-.35	-.12	.65	.			
qVM_13	.04	-.26	.77	.78	.		
dVM_13	.20	.05	.80	.69	.82	.	
EQ_{iVM}_13	**.08**	**.17**	**.91**	**.90**	**.85**	**.82**	**.**
EQ_{qVM}_13	**.28**	**.10**	**.94**	**.76**	**.94**	**.86**	**.94**

Anmerkungen. *N* = 487 Patienten. Zusammenhänge < |.10| sind statistisch nicht signifikant (5%-Niveau).

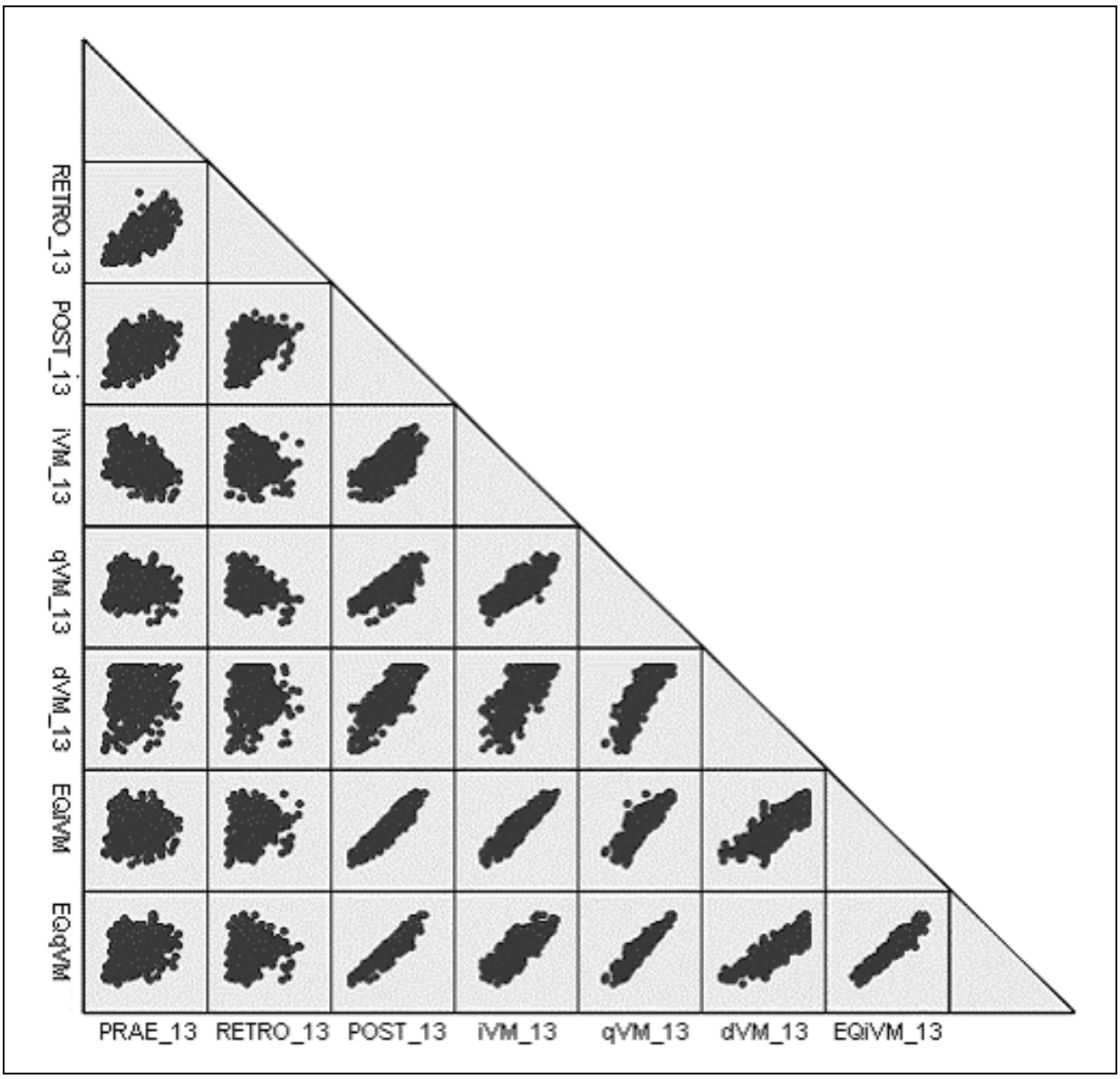


Abbildung 32. Matrix-Scatterplot der acht Skalen von GB13

5.2.3 Varianzaufklärung durch Patientenmerkmale

Anhand der in Abschnitt 3.2.5 zur Validierung von EMEK_27 einbezogenen Patientenmerkmale als Prädiktoren wurden entsprechende Regressionsgleichungen für die sechs multiplen Ergebniskriterien (POST_13, iVM_13, qVM_13, dVM_13, EQ_{iVM}_13 und EQ_{qVM}_13) der Skala GB13 berechnet. Tabelle 40 gibt die standardisierten Beta-Gewichte sowie den multiplen Korrelationskoeefizient R zur Vorhersage aller sechs Skalen wieder.

Die Gesamtvarianzaufklärung durch die Patientenmerkmale beträgt zwischen 6% (iVM_13) und 15% (POST_13). Als deutlichste Einzelprädiktoren mit jeweils schwachen Partialgewichten in einer Größenordnung von etwa -.12 kristallisieren sich hier Rentenantragstellung bei Aufnahme, die Krankheitsdauer (Chronifizierungsgrad), das Vorhandensein von Nebendiagnosen sowie die Behandlungsdauer in Tagen heraus (letzteres wohl aufgrund der Tatsache, dass ein starker Beeinträchtigungsgrad mit entsprechend längeren Aufenthalten in der Klinik verbunden ist).

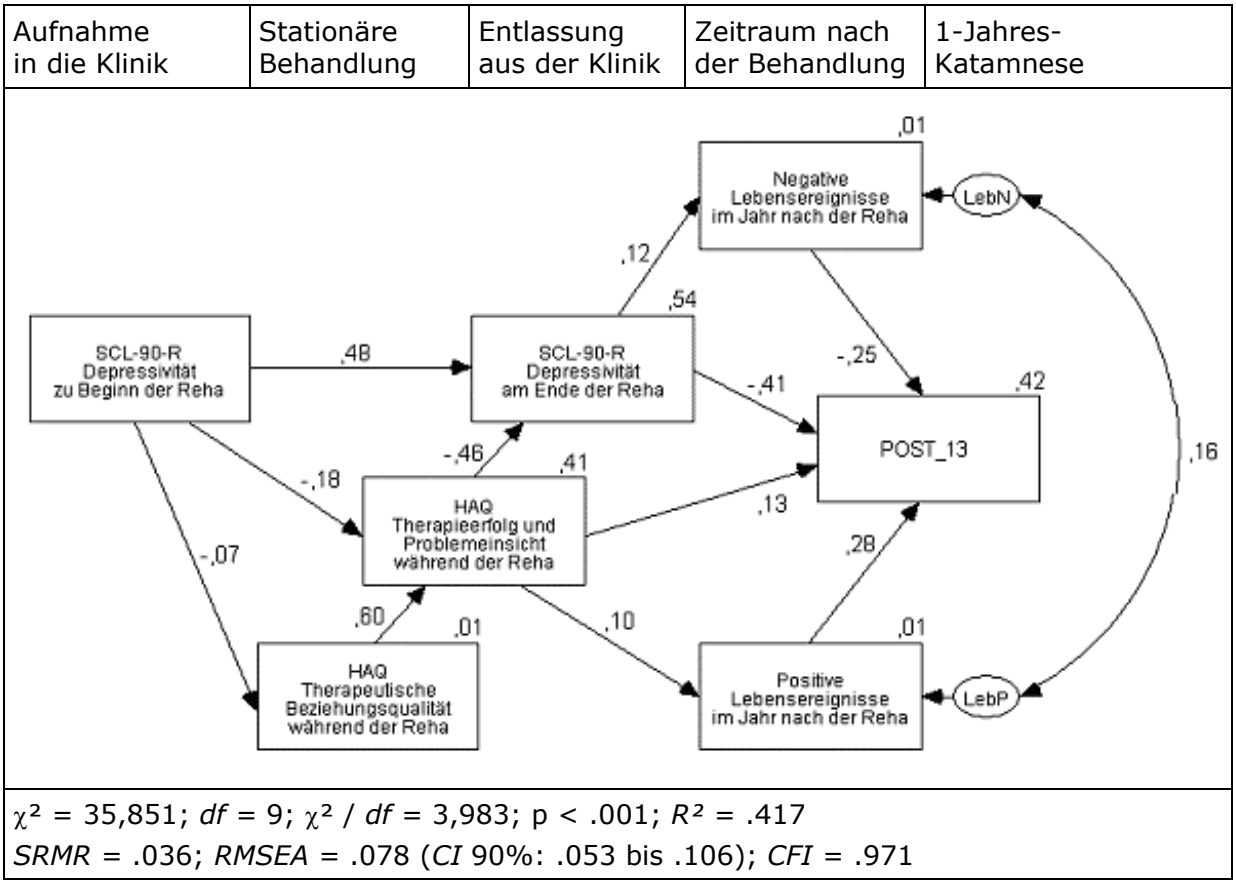
Tabelle 40. Multiple Regression zur Vorhersage der multiplen Ergebniskriterien von GB13

Skalenvariante von GB13	**POST_13**	**iVM_13**	**qVM_13**	**dVM_13**	**EQ_{iVM}_13**	**EQ_{qVM}_13**
	β	β	β	β	β	β
Geschlecht (0=m, 1=w)	.02	.12	.13	.09	.08	.08
Alter in Jahren	-.09	-.02	-.01	-.05	-.05	-.05
Schule (1=HS, 2=RS, 3=Abi)	.10	.04	.04	.07	.08	.07
Familienstand verheiratet	.02	-.01	.05	.04	.00	.04
Erwerbstätig bei Aufnahme	.08	.06	.07	.07	.07	.08
Kostenträger (0=KV, 1=RV)	-.08	-.11	-.08	-.11	-.10	-.09
Rentenantrag bei Aufnahme	-.14	-.08	-.10	-.16	-.12	-.13
Krankheitsdauer in Jahren	-.12	-.09	-.13	-.10	-.12	-.13
Somatoforme Hauptdiagnose	-.01	-.00	.01	.03	.01	.02
Nebendiagnose(n) vorhanden	-.13	-.07	-.08	-.12	-.11	-.11
Behandlungsdauer in Tagen	-.22	-.05	-.06	-.10	-.15	-.16
Vorzeitige Entlassung	-.07	-.07	-.04	-.09	-.08	-.06
Multiple Korrelation *R*	**.38**	**.24**	**.28**	**.33**	**.32**	**.34**
R^2	.15	.06	.08	.11	.11	.11
R^2 adj.	.12	.03	.05	.08	.08	.09

Anmerkung. Koeffizienten ab einem Betrag von |.09| sind statistisch signifikant (p<.05). Alle multiplen Zusammenhänge R sind statistisch signifikant (p<.05). *N* = 487.

5.2.4 Varianzaufklärung durch Prozessmerkmale

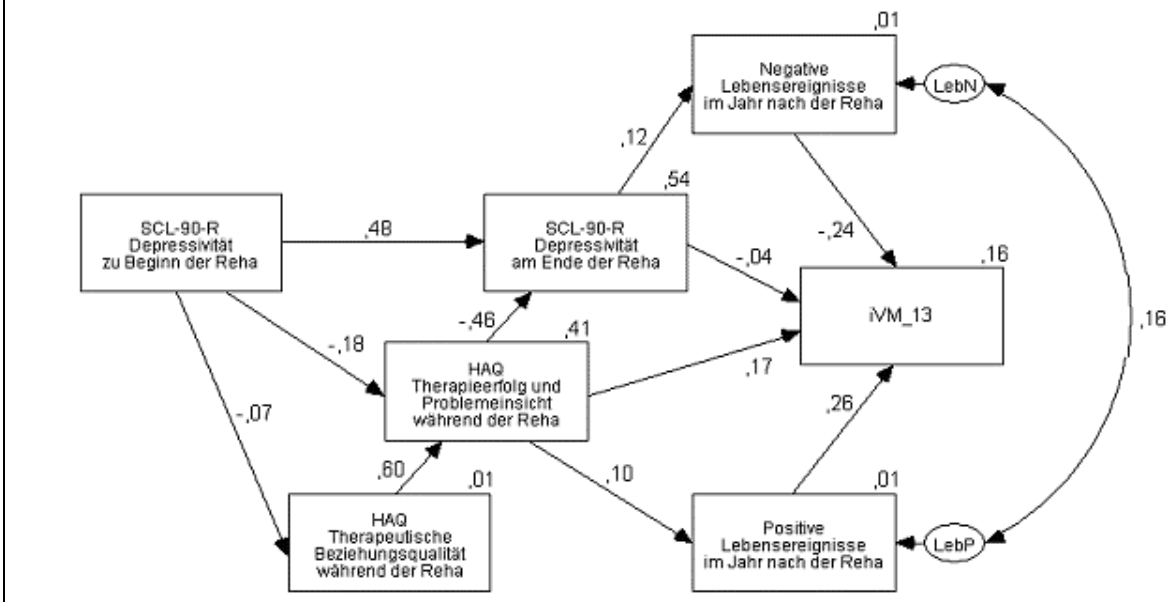
Alle Pfadmodelle (Abbildung 33 bis Abbildung 39) weisen akzeptable bis gute Fit-Indices auf, mit einer Ausnahme: Das postulierte Pfadmodell für iVM_13 deckt sich als einziges nicht mit den empirischen Daten und weist einen deutlichen Mis-Fit auf (Abbildung 34). Dieser ist vor allem durch einen fehlenden Pfad zwischen SCLASK4 und iVM13 bedingt, was wieder die starke Abhängigkeit der iVM von der Baseline veranschaulicht. Ergänzt man den fehlenden Pfad, so liegen die Modell-Fit-Indices auch hier im akzeptablen Bereich (Abbildung 35). Der Partialzusammenhang zwischen SCLASK4 und iVMSK_13 beträgt .49 und ist damit sogar stärker als der Partialzusammenhang zwischen SCLESK4 und iVMSK_13 mit -.36. Inhaltlich gibt es hinsichtlich der Pfadkoeffizienten und Gesamtvarianzaufklärung mit Ausnahme der Skala iVM_13 wenig Abweichungen zu den Befunden in Abschnitt 3.2.6.



χ^2 = 35,851; *df* = 9; χ^2 / *df* = 3,983; p < .001; R^2 = .417
SRMR = .036; *RMSEA* = .078 (*CI* 90%: .053 bis .106); *CFI* = .971

Abbildung 33. Pfadmodell zur Vorhersage von POST_13. *N* = 487 Patienten.

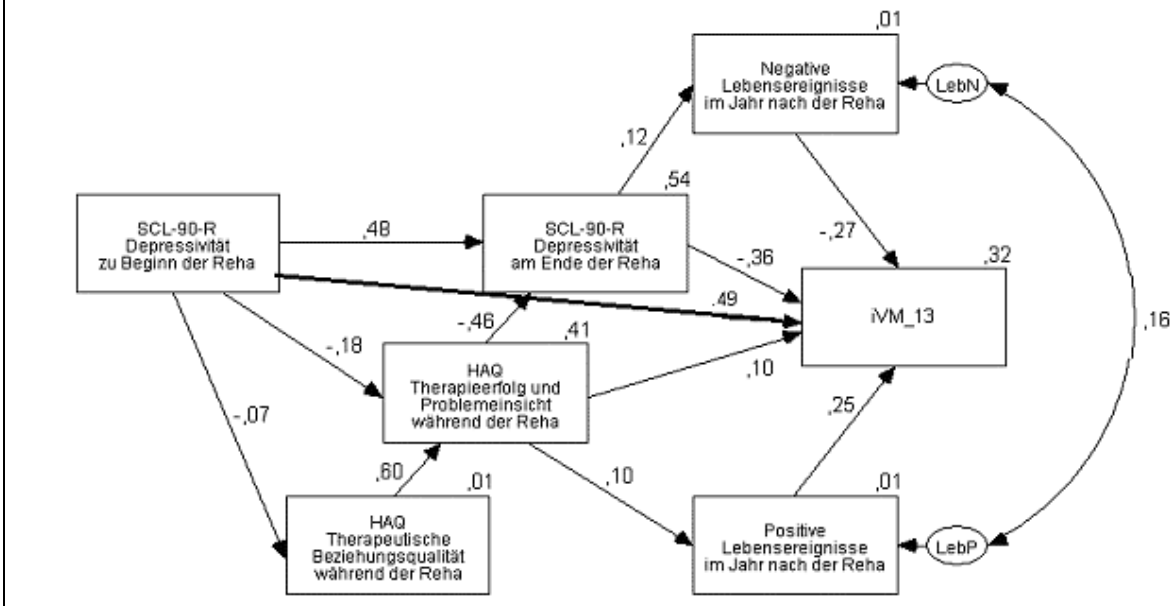
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 123,615; *df* = 9; χ^2 / *df* = 13,735; *p* = .001; R^2 = .164
SRMR = .066; *RMSEA* = .162 (*CI* 90%: .137 bis .188); *CFI* = .865

Abbildung 34. Pfadmodell zur Vorhersage von iVM_13. *N* = 487 Patienten.

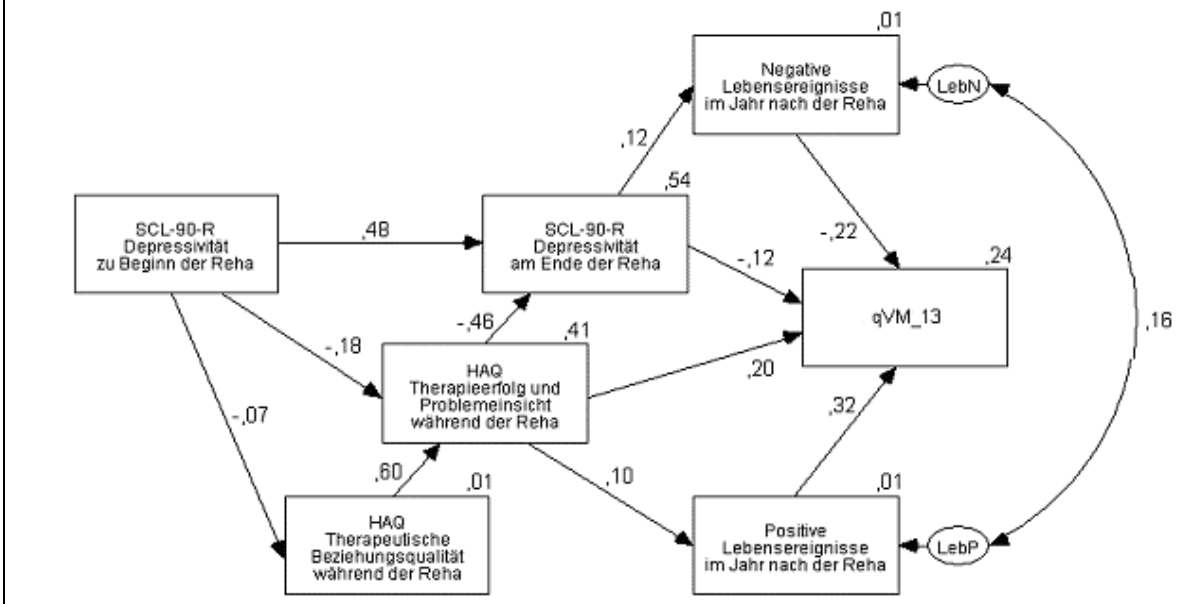
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 24,244; *df* = 9; χ^2 / *df* = 3,031; *p* = .001; R^2 = .164
SRMR = .032; *RMSEA* = .065 (*CI* 90%: .036 bis .095); *CFI* = .981

Abbildung 35. Pfadmodell zur Vorhersage von iVM_13. Modifizierte Version des Modells.

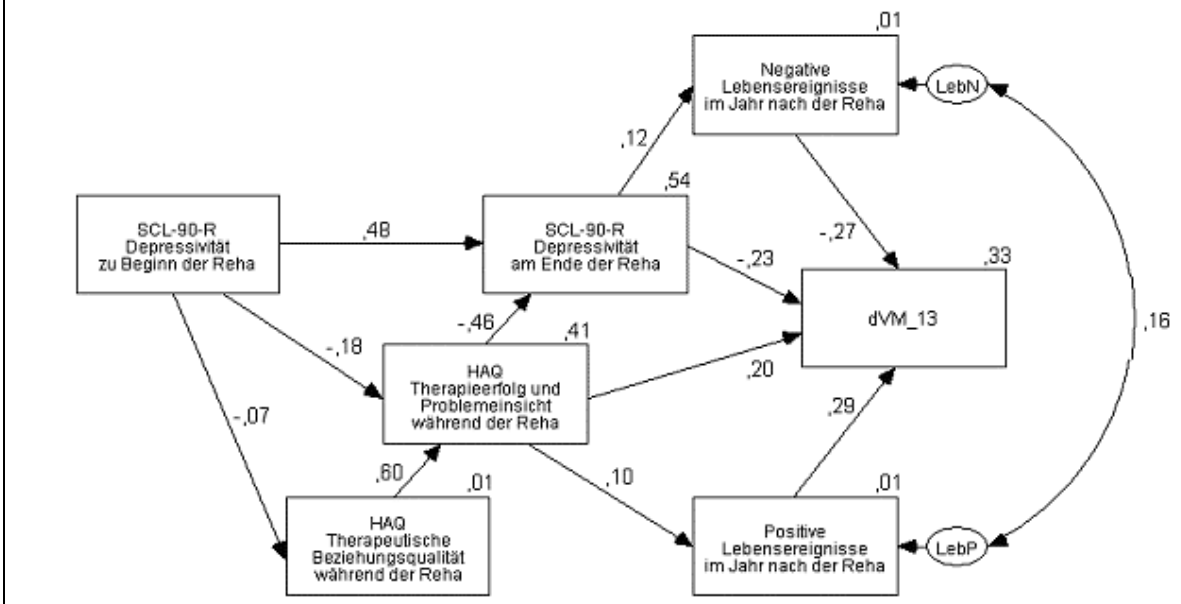
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 42,705; *df* = 9; χ^2 / *df* = 4,745; p < .001; R^2 = .242
SRMR = .039; *RMSEA* = .088 (*CI* 90%: .062 bis .115); *CFI* = .958

Abbildung 36. Pfadmodell zur Vorhersage von qVM_13. *N* = 487 Patienten.

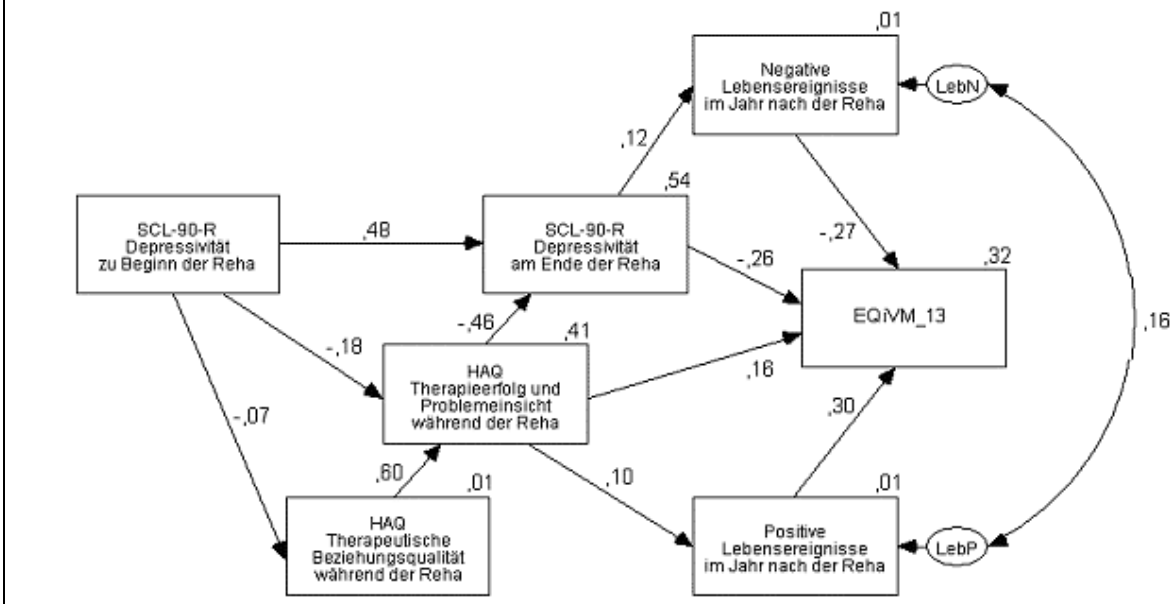
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 26,409; *df* = 9; χ^2 / *df* = 2,934; *p* = .002; R^2 = .326
SRMR = .030; *RMSEA* = .063 (*CI* 90%: .036 bis .092); *CFI* = .980

Abbildung 37. Pfadmodell zur Vorhersage von dVM_13. *N* = 487 Patienten.

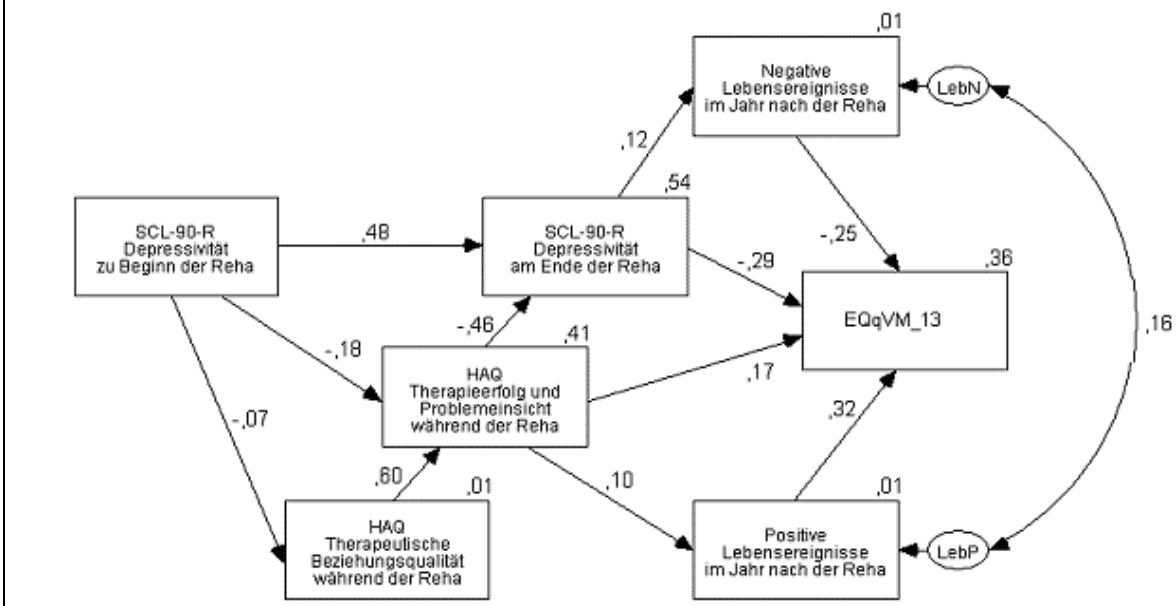
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 47,488; *df* = 9; χ^2 / *df* = 5,276; p < .001; R^2 = .320
SRMR = .041; *RMSEA* = .094 (*CI* 90%: .069 bis .121); *CFI* = .956

Abbildung 38. Pfadmodell zur Vorhersage von EQ_{iVM}_13. *N* = 487 Patienten.

Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 27,648; *df* = 9; χ^2 / *df* = 3,072; *p* = .001; R^2 = .357
SRMR = .032; *RMSEA* = .065 (*CI* 90%: .038 bis .094); *CFI* = .979

Abbildung 39. Pfadmodell zur Vorhersage von EQ_{qVM}_13. *N* = 487 Patienten.

5.2.5 Therapiezielerreichung

In Abbildung 41 sind die beiden hier neu konstruierten multiplen Composit-Ergebniskriterien in Abhängigkeit von der globalen Einschätzung der Therapiezielerreichung der Patienten zum Katamnesezeitpunkt wiedergegeben. So konnten die Patienten auf einem fünffach gestuften Item angeben, inwieweit sie ihre anfänglich formulierten Therapieziele durch die Rehabilitation errreicht haben. Die Grafik lässt sich wie folgt interpretieren: 35,9% aller Patienten gaben an, ihre Therapieziele zu „50 Prozent" und somit „teilweise erreicht" zu haben, 30,6% aller Patienten gaben an, ihre Therapieziele zu „75 Prozent" und somit „größtenteils erreicht" erreicht zu haben etc. Bildet man den Mittelwert aus den Zielerreichungskategorien (0%, 25%, 50%, 75%, 100%), so ergibt sich ein durchschnittlicher Zielerreichungsgrad von 51,7% (*SD* = 25,1%).

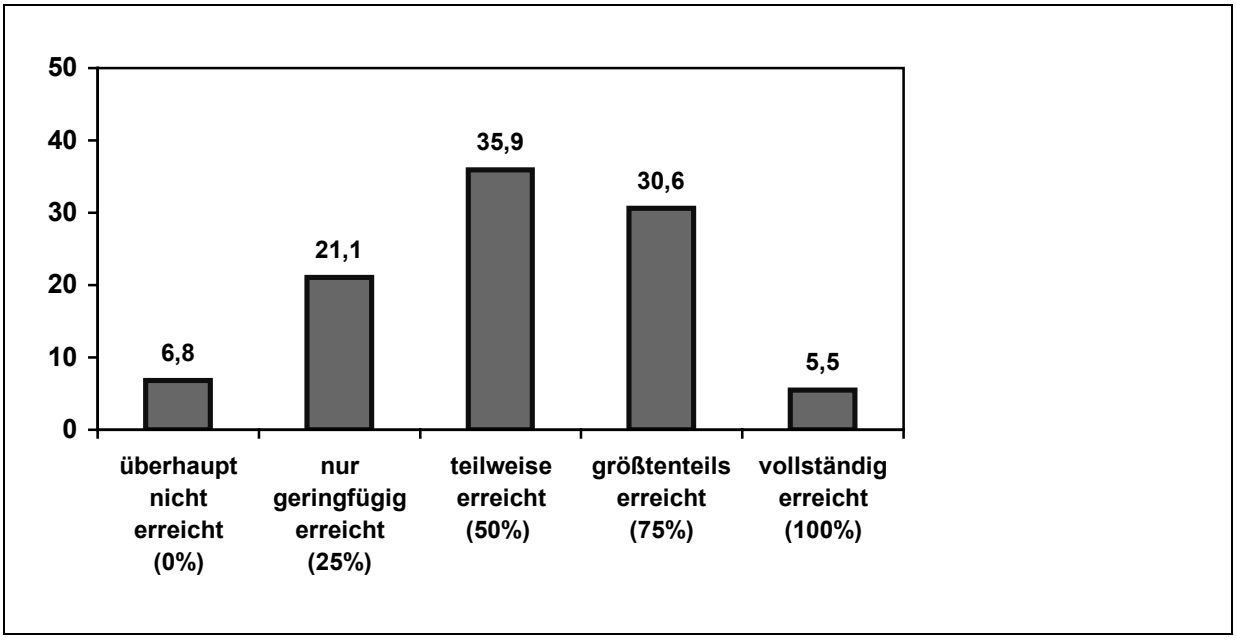


Abbildung 40. Globale Einschätzung der Therapiezielerreichung. Angaben in Prozent. Originalformulierung des Items: „Alles in allem: In welchem Ausmaß sind Ihre Therapieziele durch die Rehabilitation vor einem Jahr erreicht worden?" *N* = 478; *M* = 51,7; *SD* = 25,1.

Für die sechs multiplen Ergebniskriterien der Skala GB13 wurden die Mittelwerte in Abhängigkeit von den fünf Zielerreichungskategorien berechnet. Nachfolgend werden die Mittelwerte für EQ_{iVM}_13 (Werte für EQ_{qVM}_13 jeweils in Klammern) interpretiert (vgl. Abbildung 41): Patienten, die angeben, ihre Ziele „vollständig erreicht" zu haben, weisen einen Mittelwert auf der Skala EQ_{iVM}_13 in Höhe von 1,25 (1,66) auf. Wurden die Therapieziele „größtenteils" erreicht, so erreicht EQ_{iVM}_13 einen Mittelwert von 0,93 (1,12). Bei „teilweiser" Zielerreichung beträgt der Mittelwert 0,52 (0,68). Patienten, die angaben ihre Ziele „nur geringfügig" erreicht zu haben, nähern sich bereits dem Nullniveau auf den neuen Composit-Kriterien an (EQ_{iVM}_13 = 0,05 bzw. EQ_{qVM}_13 = 0,16). Bemerkenswert ist, dass die kleine Patientengruppe, die angab, ihre Ziele „überhaupt nicht erreicht" zu

haben, auf den beiden Composit-Skalen sogar negative Werte (EQ_{iVM}_13 = -0,27 bzw. EQ_{qVM}_13 = 0,19) erreicht. Zum Vergleich: Der Mittelwertsverlauf für die Skala iVM_13 (entspricht der klassischen PRÄ-POST-Effektgröße *d*) in Abhängkeit von der Therapiezielerreichung ist weniger ausgeprägt.

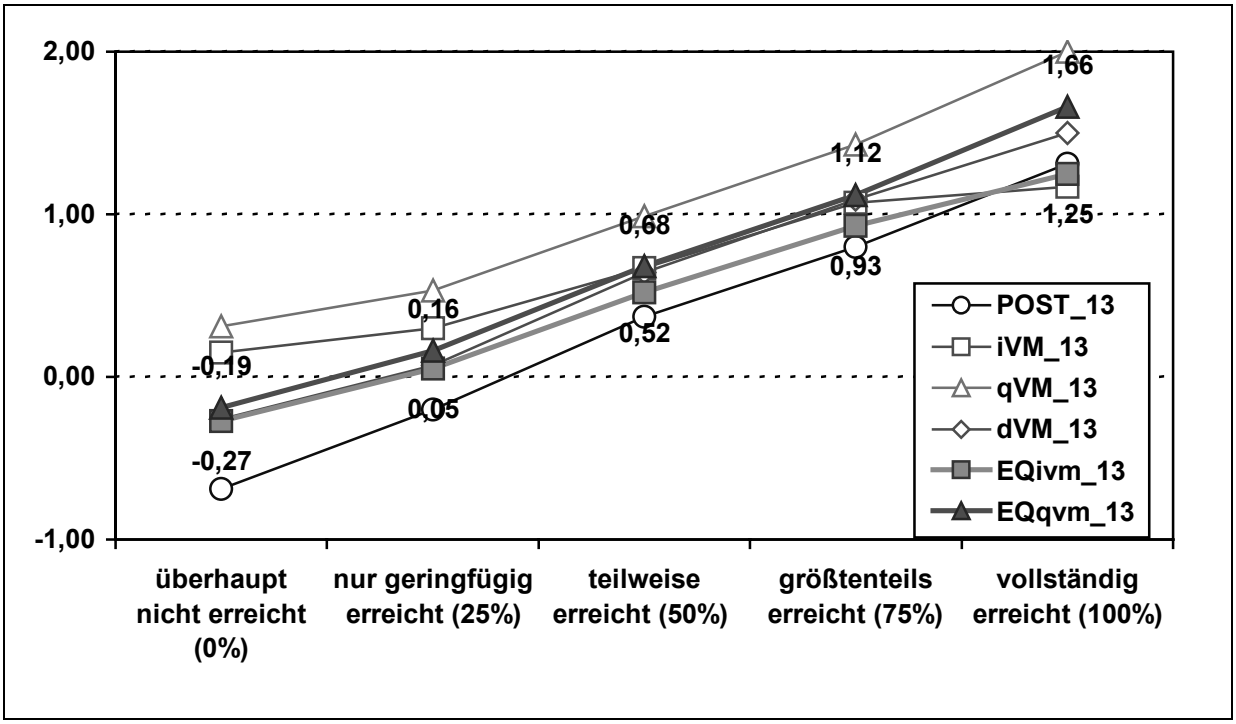


Abbildung 41. Multiple Ergebniskriterien von GB13 in Abhängigkeit von der globalen katamnestischen Therapiezielerreichung (*N* = 478). Die Mittelwerte sind für die Skala EQ_{iVM}**_13 (unterhalb der Kurve) sowie** EQ_{qVM}**_13 (oberhalb der Kurve) auch in Zahlen angegeben.**

5.3 Diskussion

Hinsichtlich der teststatistischen Kennwerte hat die erste Validierung des multiplen Ergebniskriteriums EQ sowohl für die iVM- als auch für die qVM-Variante vielversprechende Hinweise dahingehend ergeben, dass mit dem hier explorativ gebildeten Composit-Kriterium eine differenzierte und faire Bewertung der Ergebnisqualität möglich ist.

5.3.1 Beantwortung der Fragestellungen

Fragestellung 1: Item- und Skaleneigenschaften

Wie hoch sind die Mittelwerte, Streuungen und korrigiertem Item-Trennschärfen der 13 Einzelkomponenten der hier analysierten acht Skalenvarianten von GB13? Wie hoch sind die Mittelwerte, Streuungen und Reliabilitäten der acht Gesamtskalen? Wie sind die Ver-

teilungseigenschaften der Skalen? Wie stellt sich die Bewertung der Ergebnisqualität der Patienten mit den unterschiedlichen methodischen Ansätzen dar? Wie hoch sind die Zusammenhänge (Interkorrelationen) zwischen den Skalen?

Alle hier untersuchten Skalen-Varianten weisen sehr hohe interne Konsistenzen von >.90 bzw. sogar >.95 auf. Damit haben sowohl die hier eingesetzten Status- als auch Veränderungsmaße sowie die daraus neu gebildeten Composit-Kriterien eine hohe Messgenauigkeit bewiesen.

Inhaltlich zeichnen alle untersuchten singulären und multiplen Ergebniskriterien der Skala GB13 ein positives Bild der katamnestischen Behandlungsresultate. Besonders deutlich tritt dabei die positive Bewertung des Outcomes beim Gesundheitszustand (dVM) sowie vor allem beim seelischen Befinden (alle anderen methodischen Varianten) hervor.

Hinsichtlich der Skalenkorrelationen fallen die hohen Zusammenhänge zwischen allen sechs multiplen Ergebniskriterien (POST_13, iVM_13, qVM_13, dVM_13, EQ_{iVM}_13 und EQ_{qVM}_13 auf. Dies bedeutet, dass die unterschiedlichen methodischen Zugänge zur Abbildung der Ergebnisqualität einen beträchtlichen gemeinsamen Varianzanteil erfassen, wie bereits bei Schmidt et al. (2003) ausgeführt.

Das hier neu konstruierte Composit-Maß EQiVM_13 weist keine statistisch signifikante Abhängigkeit von der PRÄ-Messung der Patienten auf. Für EQ_{qVM}_13 wurde entgegen der eingangs formulierten Erwartung allerdings eine leichte positive Abhängigkeit von der Baseline festgestellt, d.h. moderat beeinträchtigte Patienten werden durch diesen Zugang zur Bewertung der Ergebnisqualität eher begünstigt als schwerer beeinträchtigte Patienten.

Fragestellung 2: Validierung

Inwieweit lassen sich die hier berechneten unterschiedlichen multiplen Ergebniskriterien durch Stichproben- und Methodenmerkmale vorhersagen? Zur Validierung werden entsprechende multiple Regressionsgleichungen (vgl. Abschnitt 3.2.5) bzw. Pfadanalysen (vgl. Abschnitt 3.2.6) berechnet. Darüber hinaus werden globale Einschätzungen der Patienten zum Zeitpunkt der 1-Jahres-Katamnese hinsichtlich der Therapiezielerreichung zur Validierung verwendet.

Insgesamt ergibt sich eine Varianzaufklärung in einer Größenordnung von 6% (Skala iVM_13) bis 15% (Skala POST_13) durch Stichprobenmerkmale. Bei den beiden neuen Compositkriterien beträgt der aufgeklärte Varianzanteil jeweils 11%. Bei den Skalen erweist sich fast durchgängig neben der Rentenantragstellung bei Aufnahme auch eine lange Erkrankungsdauer (Chronifizierung) als ungünstiger Prädiktor für die Ergebnisqualität.

Bei den Pfadanalysen zeigt sich eine Varianzaufklärung zwischen 24% (qVM_13) und 42% (POST_13), wobei alle Skalen einen akzeptablen bis guten Modell-Fit aufweisen. Lediglich bei der Skala iVM_13 ergaben sich deutliche Hinweise darauf, dass sich das postulierte Modell nicht mit der von den Daten abgebildeten Realität deckt und die Demoralisierung bei Aufnahme offenbar einen erheblichen Einfluss auf das multiple Ergebniskriterium iVM_13 ausübt: Je stärker der Grad an Depressivität zu Behandlungsbeginn, desto größer fallen die PRÄ-POST-Differenzen der iVM aus. Alle

desto größer fallen die PRÄ-POST-Differenzen der iVM aus. Alle anderen Skalen zeigten diese Anfälligkeit für eine Verzerrung durch die Baseline-Messung nicht.

Die Modellgüte-Indices für die beiden neu konstruierten multiplen Ergebniskriterien EQ_{iVM}_13 und EQ_{qVM}_13 weisen ebenfalls akzeptable Werte auf (wenngleich der Modellfit für EQ_{iVM}_13 sich an der unteren Grenze des Tolerierbaren bewegt). Die Varianzaufklärung durch die einbezogenen Prozessmerkmale beträgt 32% (EQ_{iVM}_13) bzw. 36% (EQ_{qVM}_13).

Stellt man die verschiedenen hier berechneten multiplen Ergebniskriterien von GB13 in Abhängigkeit von der katamnestisch erfragten globalen Therapiezielerreichung dar, so zeigen sich bei den neuen Composit-Maßen deutliche Mittelwertsverläufe. Eine hohe Therapiezielerreichung ist jeweils mit positiven Werten von 1,00 und mehr auf den standardisierten Skalen verbunden, während bei geringer Therapiezielerreichung Werte nahe 0,00 bzw. sogar deutlich kleiner als 0,00 resultieren.

5.3.2 Bewertung der Composit-Kriterien

Die beiden hier neu gebildeten Composit-Kriterien EQ_{iVM}_13 sowie EQ_{qVM}_13 weisen eine hohe Reliabilität auf. Die in der Validierung beobachteten Zusammenhänge zwischen den Composit-Kriterien und Prozess- und Ergebnismerkmalen sind zum Teil sogar höher als die Korrelationen zwischen herkömmlichen PRÄ-POST-Effektgrößen und den Prozess- und Ergebnismerkmalen.

Bemerkenswert ist der Befund, dass EQ_{iVM}_13 keine statistisch signifikante Abhängigkeit von der PRÄ-Messung aufweist. So weisen iVM häufig negative, dVM hingegen positive Zusammenhänge mit der Baseline auf. Dies bedeutet, dass die iVM schwer beeinträchtigte Patienten systematisch positiver beurteilt als moderat beeinträchtigte Patienten (Regression zur Mitte). Bei der dVM ist es umgekehrt: Hier entsteht bei moderat beeinträchtigten Patienten ein positiver Bias bei der Bewertung der Ergebnisqualität. Dies ist nicht im Sinne einer fairen Evaluation, die eine möglichst objektive Beurteilung des Outcome unabhängig vom Vorliegen bestimmter Ausgangsvoraussetzungen ermöglichen sollte. Die hier erprobte Skala EQ_{iVM}_13 weist diese Problematik einer Verzerrung der Bewertung in Abhängigkeit vom Ausgangsniveau offenbar nicht auf.

Im Gegensatz zu EQ_{iVM}_13 zeigte sich bei EQ_{qVM}_13 allerdings eine positive Abhängigkeit vom Ausgangsniveau PRÄ_13. Abgesehen davon gelang aber auch hier die Validierung anhand der einbezogenen Merkmale. Damit mag diese zweite Composit-Variante möglicherweise eine Option in Kontexten darstellen, wo eine echte PRÄ-Messung nicht realisierbar ist, so z.B. im Rahmen von als Einpunkterhebung angelegten Routinekatamnesen.

6 Schlusswort

Multiple Ergebniskriterien eröffnen faszinierende Möglichkeiten, inhaltlich heterogene Einzelinformationen im Sinne des bio-psycho-sozialen Modells zu einem Gesamtindex zu verdichten und damit unterschiedliche Facetten der Ergebnisqualität abzubilden.

Methodisch ist dabei die Ausschöpfung der gesamten Iteminformation sinnvoll, wenn es auf eine möglichst hohe prognostische Validität ankommt. Dies kann recht einfach durch die im Rahmen von Meta-Analysen verwendete Effektgrößenmetrik bzw. z-Standardisierung realisiert werden. Für ein routinemäßiges Screening der Ergebnisqualität, etwa im Rahmen von fortlaufenden katamnestischen Einpunkt-Erhebungen hat sich jedoch auch der Ansatz von Schmidt et al. (1987) bewährt, bei dem jedes Item zunächst einer dichotomen Ergebnisbewertung im Sinne von Eins (erfolgreich) versus Null (nicht erfolgreich) unterzogen wird und diese binären singulären Ergebniskriterien sodann zu einem multiplen Gesamtindex aggregiert werden.

Im Zuge der Reanalyse zur Skala EMEK_27 hat sich mit dem Ausgangswertproblem eine von der Forschung bislang ungelöste Frage herauskristallisiert, die charakteristisch für die in PRÄ-POST-Studien ohne Kontrollgruppe berechneten Ergebnismaße ist. So bringt die Verwendung der gesamten Iteminformation gegenüber der auf dichotomen Items basierenden Variante von EMEK_27 zwar eine Verbessung der Informationsausschöpfung, das Ausgangswertproblem bleibt jedoch ungelöst. Bei näherer Betrachtung tritt dieses methodische Problem nicht nur bei der direkten Veränderungsmessung, sondern auch bei der klassischen indirekten sowie bei der quasi-indirekten Veränderungsmessung auf, solange keine Kontrollgruppe vorhanden ist.

Im Rahmen von randomisierten Studien stellt sich das Ausgangswertproblem nicht, da durch die Kontrollgruppe eine unmittelbare Vergleichsmöglichkeit gegeben ist. Aus diesen Grund wäre es wünschenswert, künftig auch in der Rehabilitationsforschung entsprechende Studien realisieren zu können, was aufgrund der versorgungsstrukturellen Vorgaben bislang leider nicht möglich war. Die Ausführungen der vorliegenden Arbeit beziehen sich auf das Eingruppen-PRÄ-POST-Design und sollen mögliche Lösungswege für die angeschnittenen methodischen Probleme aufzeigen.

Mit dem hier neu gebildeten Composit-Kriterium EQ_{iVM}_13 wurde daher ein Kompromiss zwischen der Berücksichtigung von Status- und Veränderungsinformationen vorgeschlagen, der möglicherweise einen Lösungsansatz für das Ausgangswertproblem darstellt, ohne sich dabei die methodischen Nachteile der ausschließlichen Verwendung von PRÄ-POST-Effektgrößen oder der zielorientierten Ergebnismessung zur Erfolgsbewertung einzuhandeln. So erscheinen die Befunde der ersten Validierung zu dem hier erstmals erprobten multiplen Ergebniskriterium, das auf singulären Composit-Kriterien beruht, vielversprechend, weshalb der Ansatz in weiteren Studien ausführlicher untersucht werden sollte. So lässt sich der hier entwickelte Berechnungsansatz zur Bildung von Composit-Kriterien ohne weiteres auf geläufige PRÄ-POST-Maße wie z.B. standardisierte Testverfahren übertragen, was in Abschnitt 4.2 demonstriert wurde.

Darüber hinaus bestehen noch offene Fragen zum Thema der multiplen Ergebniskriterien, deren nähere Erörterung den Rahmen der vorliegenden Arbeit sprengen würde. Diese Fragen sind eher inhaltlicher Natur. So stellt sich die Frage nach der adäquaten Gewichtung mehrerer singulärer Bewertungsaspekte bei Bildung der Gesamtskala. Im Verlauf der hier durchgeführten Analysen wurde bei der Skalenbildung grundsätzlich eine Gleichgewichtung aller Einzelaspekte vorgenommen. Es wäre aber durchaus denkbar, a priori bestimmte Stakeholdergruppen stärker oder weniger stark zu berücksichtigen, was sich in einer entsprechenden Gewichtung der entsprechenden Ergebnisaspekte ausdrücken könnte.

Eine andere Frage betrifft die Einigung auf eine geeignete Auswahl hinsichtlich der Einzelaspekte, die bei der evaluativen Bewertung der Ergebnisqualität psychosomatischer Rehabilitation einzubeziehen sind. Die aus dem bio-psycho-sozialen Modell sowie den gesetzlichen Vorgaben zur Rehabilitation resultierende Forderung, dass somatische, psychische, soziale und funktionale Aspekte in die Ergebnisbewertung Eingang finden sollten, stellt zumindest schon einmal einen gewissen Konsens sicher, ist aber noch zu wenig konkret. Wünschenswert wäre die Einigung auf ein Standardinstrumentarium, das von verschiedenen Stakeholdern (Kostenträgern, Rehabilitationskliniken, Patienten, Wissenschaftlern) gleichermaßen akzeptiert wird. Diesbezügliche Vorschläge wurden in der Vergangenheit bereits unterbreitet, so z.B. von Gerdes et al. (1991) bzw. Bührlen, Gerdes & Jäckel (2005) mit dem IRES-Fragebogen, von Koch & Tiefensee (1998) im Rahmen des 5-Punkte-Programms zur Qualitätssicherung in der gesetzlichen Rehabilitation mit einem einheitlichen Patientenfragebogen, der von der DRV bundesweit zur Beurteilung der Behandlung eingesetzt wird, sowie von Deck & Röckelein (1999) und Muthny & Bullinger (1999) mit einer Kernbatterie in der Rehabilitationsforschung zu verwendenden soziodemographischen Variablen und Erhebungsinstrumente.

Es bleibt zu hoffen, dass der Dialog zwischen den verschiedenen Stakeholdergruppen in der Versorgungsforschung künftig weiter intensiviert wird und auf diese Weise eine entsprechende Konsensbildung möglich ist. Hierbei ist es von Bedeutung, zwischen der methodischen und inhaltlichen Weiterentwicklung der Ergebnisevaluation zu differenzieren. So wurde in der vorliegenden Arbeit demonstriert, dass mit inhaltlich identischen Bewertungskriterien durchaus heterogene methodische Ansätze bei der Datenerhebung und –bewertung Anwendung finden können. Diese Methodenpluralität ist ganz im Sinne einer maßgeschneiderten Evaluation, die je nach Fragestellung und vertretbarem Aufwand bei der Datenerhebung unterschiedliche Aspekte der Ergebnisqualität abbilden kann.

7 Literaturverzeichnis

Ahrens, S. & Schneider, W. (2002). Lehrbuch der Psychotherapie und Psychosomatischen Medizin. Stuttgart: Schattauer.

Ainsworth, M. D., Blehar, M. C., Waters, E., & Wall, S. (1978). Patterns of Attachment. A psychological Study of the strange situation. New York: Earlbaum.

Ajzen, I. (1988). Attitudes, Personality and Behavior. Chicago: The Dorsey Press.

Ajzen, I. & Fishbein, M. (1980). Understanding Attitudes and predicting social Behaviour. Englewood Cliffs: Prentice Hall.

Amelang, M., Zielinski, W., Fydrich, T., & Moosbrugger, H. (1997). Psychologische Diagnostik und Intervention. Berlin: Springer.

Antonovsky, A. & Franke, A. (1997). Salutogenese. Tübingen: Deutsche Gesellschaft für Verhaltenstherapie.

Bassler, M., Potratz, B., & Krauthauser, H. (1995). Der "Helping Alliance Questionnaire" (HAQ) von Luborsky. Psychotherapeut, 40, 23-32.

Beauducel, A. & Wittmann, W. W. (2005). Simulation Study on Fit Indexes in CFA Based on Data With Slightly Distorted Simple Structure. Structural EQuation Modeling, 12, 41-75.

Bengel, J. & Jäckel, W. H. (Hrsg.). (2000). Zielorientierung in der Rehabilitation. Regensburg: Roderer.

Bengel, J. & Koch, U. (Hrsg.). (2000). Grundlagen der Rehabilitationswissenschaften. Berlin: Springer.

Bengel, J., Strittmatter, R., & Willmann, H. (1998). Was erhält Menschen gesund? Antonovskys Modell der Salutogenese - Diskussionsstand und Stellenwert. Köln: Bundeszentrale für gesundheitliche Aufklärung.

Bischoff, C., Gönner, S., Ehrhardt, M., Limbacher, K., Husen, E., & Jäger, R. S. (2003). Das PräPostProjekt. Lengerich: Pabst.

Bosnjak, M. (2007, Oktober). Die Methode der Meta-Analyse zur Evidenzbasierung von Gesundheitsrisiken: Mögliche Beiträge der Psychologie. Vortrag auf dem internationalen Symposium zum Thema „Tabakrauch am Arbeitsplatz: Gesundheitsrisiken und Kausalität", Universität Mannheim.

Broda, M., Bürger, W., Dinger-Broda, A., & Massing, H. (1996). Die Berus-Studie. Bad Münstereifel: Westkreuz-Verlag.

Browne, M. & Cudeck, R. (1993). Alternative ways of assessing EQuation model fit. Newbury Park: Sage.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. Psychological Review, 62, 193-217.

Bührlen, B., Gerdes, N., & Jäckel, W. H. (2005). Entwicklung und psychometrische Testung eines Patientenfragebogens für die medizinische Rehabilitation (IRES-3). *Die Rehabilitation, 44,* 63-74.

Bundesanstalt für Arbeitsschutz und Arbeitsmedizin. (2005). Volkswirtschaftliche Kosten durch Arbeitsunfähigkeit 2004. Dortmund: Bundesanstalt für Arbeitsschutz und Arbeitsmedizin.

Bundesministerium für Bildung und Forschung. (2006). Pressemitteilung vom 07.07.2006: Bündnis für die Versorgungsforschung. Bonn: Bundesministerium für Bildung und Forschung.

Bundesregierung. (2001). Sozialgesetzbuch IX - Rehabilitation und Teilhabe behinderter Menschen vom 19. Juni 2001 (BGBl. I S. 1046, 1047), zuletzt geändert durch Artikel 1 des Gesetzes vom 23. April 2004 (BGBl. I S. 606). Bonn: Bundesministerium für Gesundheit und soziale Sicherung (BMGS).

Bundesverband der Betriebskrankenkassen. (2005). BKK Gesundheitsreport 2005. Essen: Eigenverlag.

Bundesverband der Betriebskrankenkassen. (2006). Pressemitteilung vom 18.04.2006. Essen: Eigenverlag.

Buschmann-Steinhage, R. (1998). Einrichtungen der Rehabilitation und ihre Aufgaben. In H. Delbrück & E. Haupt (Hrsg.), Rehabilitationsmedizin (S. 93-108). München: Urban und Schwarzenberg.

Buschmann-Steinhage, R., Gerwinn, H., Klosterhuis, H., & Mitreiter, R. (1998). Der Förderschwerpunkt "Rehabilitationswissenschaften" - ein Förderprogramm und seine Umsetzung. Die Rehabilitation, 37, 71-77.

Bürger, W. & Buschmann-Steinhage, R. (2000). Rehabilitative Angebotsformen. In J. Bengel & U. Koch (Hrsg.), Grundlagen der Rehabilitationswissenschaften (S. 139-162).

Byrne, B. M. (2001). Structural EQuation modeling with AMOS: Basic concepts, applications, and programming. Mahwah: Lawrence Erlbaum Associates, Publishers.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 54, 297-312.

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

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

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

Carlson, K. D. & Schmidt, F. L. (1999). Impact of experimental design on effect size: Findings from the research literature on training. Journal of Applied Psychology, 84, 851-862.

Cattell, R. B. (1957). Personality and motivation structure and measurement. Oxford: World Book Co.

Cattell, R. B. (1966). Handbook of Multivariate Experimental Psychology. Rand McNally: Chicago.

Cohen, J. (1983). The cost of dichotomization. Applied Psychological Measurement, 7, 249-253.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Collins, L. M. & Sayer, A. G. (Eds.). (2001). New methods for the analysis of change. Washington: American Psychological Association.

Cook, D. J. & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field setting. Chicago: Rand McNelly.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. Review of Educational Research, 52, 291-302.

Cronbach, L. J. (1980). Toward reform of program evaluation: Aims, methods, and institutional arrangements. San Francisco: Jossey-Bass.

Cronbach, L. J. (1982). Designing Evaluations of educational and social programs. San Francisco: Jossey-Bass.

Cronbach, L. J. & Furby, L. (1970). How we should measure "change": Or should we? Psychological Bulletin, 74, 68-80.

Deck, R. & Röckelein, F. (1999). Zur Erhebung soziodemographischer und sozialmedizinischer Indikatoren in den rehabilitationswissenschaftlichen Forschungsverbunden. DRV-Schriften, 16, 84-102.

Delbrück, H. & Haupt, E. (Hrsg.). (1998). Rehabilitationsmedizin. München: Urban und Schwarzenberg.

Deter, H. C., Albrecht, H., Berghaus, A., Ferszt, R., Gieler, U., Hontschik, B., Klosinski, G., Knispel, H., Köhnlein, B. & Kretz, F.J. (Hrsg.). (1996). Angewandte Psychosomatik. Stuttgart: Thieme.

Deutsche Rentenversicherung Bund. (2005). Abgeschlossene Leistungen zur Rehabilitation 2005. Frankfurt a. M.: VDR.

Deutsches Institut für Medizinische Dokumentation und Information. (Hrsg.). (2005). Internationale Klassifikation der Funktionsfähigkeit, Behinderung und Gesundheit (ICF). Köln: Deutsches Institut für Medizinische Dokumentation und Information.

Deutsches Institut für Medizinische Dokumentation und Information. (Hrsg.). (2007). Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (ICD-10). Köln: Deutsches Institut für Medizinische Dokumentation und Information.

Donabedian, A. (1966). Evaluating the quality of medical care. Milbank Memorial Fund Quarterly, 44, 166-203.

Dubois, P. H. (1957). Multivariate correlational analysis. Oxford: Harper.

Egner, U. (2003). Umstellung des Verbraucherpreisindex auf Basis 2000. Statistisches Bundesamt - Wirtschaft und Statistik, 5, 423-432.

Farin, E. (1997). Metaanalysen: Methodologische Grundlagen und praktische Durchführung. In B. Strauss & J. Bengel (Hrsg.), Forschungsmethoden in der Medizinischen Psychologie (S. 161-180). Göttingen: Hogrefe.

Fäh, M. & Fischer, G. (1998). Sinn und Unsinn in der Psychotherapieforschung.

Fishbein, M. & Ajzen, I. (1975). Belief, attitude, intention and behavior: An introduction to theory and research. Reading: Addison-Wesley.

Frank, J. D. (1992). Wirkungsweisen psychotherapeutischer Beeinflussung. Vom Schamanismus bis zu modernen Therapien. Stuttgart: Klett-Cotta.

Franke, G. H. (1995). SCL-90-R. Symptom-Checkliste von Derogatis. Weinheim: Beltz.

Franke, G. H. (2002). SCL-90-R. Die Symptom-Checkliste von Derogatis. Göttingen: Hogrefe.

Furby, L. (1973). Interpreting regression toward the mean in developmental research. Developmental Psychology, 8, 172-179.

Fydrich, T., Laireiter, A. R., Saile, H., & Engberding, M. (1996). Diagnostik und Evaluation in der Psychotherapie: Empfehlungen zur Standardisierung. Zeitschrift für klinische Psychologie, 25, 161-168.

Fydrich, T. & Schneider, W. (2007). Evidenzbasierte Psychotherapie. Psychotherapeut, 1, 55-68.

Gerdes, N. (1998). Rehabilitationseffekte bei "zielorientierter Ergebnismessung" - Ergebnisse der IRES-ZOE-Studie 1996/97. Deutsche Rentenversicherung, 3-4, 217-238.

Gerdes, N., Bengel, J., & Jäckel, W. H. (2000). Zielorientierung in Diagnostik, Therapie und Ergebnismessung. In J. Bengel & W. H. Jäckel (Hrsg.), Zielorientierung in der Rehabilitation. Rehabilitationswissenschaftlicher Forschungsverbund Freiburg/Bad Säckingen (pp. 3-12). Regensburg: Roderer.

Gerdes, N., Jäckel, W. H., & Fliedner, T. M. (1991). "IRES" - Ein Fragebogen zur Messung von Rehabilitationsbedürftigkeit und Rehabilitationserfolg. Mitteilungen der LVA Württemberg, 3, 72-77.

Gerdes, N., Weidemann, H., & Jäckel, W. (2000b). Die PROTOS-Studie. Darmstadt: Steinkopf.

Gottman, J. M. (Ed.). (1995). The analysis of change. Hillsdale: Lawrence Erlbaum Associates.

Grawe, K. & Braun, U. (1994). Qualitätskontrolle in der Psychotherapiepraxis. Zeitschrift für klinische Psychologie, 23, 242-267.

Grawe, K., Donati, R., & Bernauer, F. (1994). Psychotherapie im Wandel. Göttingen: Hogrefe.

Grigoleit, H. (1998). Ambulante und teilstationäre Rehabilitation im Netzwerk der rehabilitativen Einrichtungen und Funktionen. In H. Delbrück & E. Haupt (Hrsg.), Grundlagen der Rehabilitationsmedizin (S. 109-116). München: Urban und Schwarzenberg.

Haaf, B. & Schliehe, F. (1998). Zur Effektivität und Effizienz der medizinischen Rehabilitation. In H. Delbrück & E. Haupt (Hrsg.), Rehabilitationsmedizin (S. 225-236). München: Urban und Schwarzenberg.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. (1998). Multivariate Data Analysis. New York: Prentice Hall.

Hartmann, A. & Herzog, T. (1995). Varianten der Effektstärkenberechnung in Meta-Analysen: Kommt es zu variablen Ergebnissen? Zeitschrift für klinische Psychologie, 24, 337-343.

Haupt, E. & Delbrück, H. (1998). Grundlagen der Rehabilitation. In H. Delbrück & E. Haupt (Hrsg.), Rehabilitationsmedizin (S. 35-44). München: Urban und Schwarzenberg.

Hautzinger, M., Bailer, M., Worall, H., & Keller, F. (1994). BDI. Beck-Depressions-Inventar. Bern: Huber.

Hedges, L. V. & Olkin, I. (1985). Statistical Methods for Meta-Analysis. Orlando: Academic Press.

Herzog, T. & Hartmann, A. (1997). Psychoanalytisch orientierte Behandlung der Anorexia nervosa. Psychotherapie, Psychosomatik, Medizinische Psychologie, 47, 299-315.

Hilpert, H. R. (1979). Therapeutische Gemeinschaft in einer psychotherapeutischen Klinik. Psychotherapie, Medizinische Psychologie, 29, 46-53.

Hollstein, H. (1998). Alkohol-, Medikamenten- und Drogenabhängigkeit. In H. Delbrück & E. Haupt (Hrsg.), Rehabilitationsmedizin (S. 678-713). München: Urban und Schwarzenberg.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural EQuation Modeling, 6, 1-55.

Hüppe, A. & Raspe, H. (2005). Die Wirksamkeit stationärer medizinischer Rehabilitation in Deutschland bei chronischen Rückenschmerzen: Eine systematische Literaturübersicht 1980-2001. Die Rehabilitation, 42, 143-154.

ifo Institut für Wirtschaftsforschung e.V. an der Universität München. (2008). ifo Geschäftsklima in Deutschland - Ergebnisse des ifo-Konjunkturtests im März 2008. München: ifo Institut.

Internet Access Center Düsseldorf (2008). www.klinikverzeichnis-online.de. Düsseldorf: Internet Access Center Düsseldorf.

Jacobi, C., Dahme, B., & Rustenbach, S. (1997). Vergleich kontrollierter Psycho- und Pharmakotherapiestudien bei Bulimia und Anorexia nervosa. Psychotherapie, Psychosomatik, Medizinische Psychologie, 47, 346-364.

Jacobson, N. S. & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology, 59, 12-19.

Janssen, P. L. & Hoffmann, S. O. (1994). Der Facharzt für Psychotherapeutische Medizin. Psycho, 20, 322-333.

Jäckel, W. H., Protz, W., Maier-Riehle, B., & Gerdes, N. (1997). Qualitäts-Screening im Qualitätssicherungsprogramm der gesetzlichen Rentenversicherung. Deutsche Rentenversicherung, 9-10, 575-591.

Kanzow, U. (1986). Nun kurt mal schön! Deutsches Ärzteblatt, 83, 2487.

Kastner, S. & Basler, H. D. (1997). Messen Veränderungsfragebögen wirklich Veränderungen? Der Schmerz, 11, 254-262.

Kazis, L. E., Anderson, J. J., & Meenan, R. F. (1989). Effect sizes for interpreting changes in health status. Medical Care, 27, 178-189.

Kächele, H. (2006). Ziele der Behandlung: Von Micro- über Meso- zu Macro-Outcome.

Kächele, H. & Fiedler, I. (1985). Ist der Erfolg einer psychotherapeutischen Behandlung vorhersagbar? Psychotherapie, Psychosomatik, Medizinische Psychologie, 35, 201-206.

Kiresuk, T. J. & Sherman, R. E. (1968). Goal Attainment Scaling: A general method for evaluating comprehensive community mental health programs. Community Mental Health Journal, 4, 443-453.

Kiresuk, T. J., Smith, A., & Cardillo, J. E. (1994). Goal attainment scaling: Applications, theory, and measurement. Hillsdale: Lawrence Erlbaum Associates.

Koch, U. & Bengel, J. (2000). Definition und Selbstverständnis der Rehabilitationswissenschaften. In J. Bengel & U. Koch (Hrsg.), Grundlagen der Rehabilitationswissenschaften (S. 3-18). Berlin: Springer.

Koch, U., Gerdes, N., Jäckel, W., Müller-Fahrnow, W., Raspe, H. H., Schian, H. M., Schliehe, F., Wallesch, C.W., & Lotz,W. (1995). Verbundforschung Rehabilitationswissenschaften. Deutsche Rentenversicherung, 7-8, 491-513.

Koch, U. & Tiefensee, J. (1998). Das 5-Punkte-Programm zur Qualitätssicherung in der stationären medizinischen Rehabilitation der Rentenversicherung. In D. Hell, J. Bengel, & M. Kirsten-Krüger (Hrsg.), Qualitätssicherung der psychiatrischen Versorgung. Modelle und Projekte in der Schweiz und in Deutschland (S. 45-52). Basel: Karger.

Kohlmann, T. & Raspe, A. (1998). Zur Messung patientennaher Erfolgskriterien in der Rehabilitation: Wie gut stimmen "indirekte" und "direkte" Methoden der Veränderungsmessung überein? Die Rehabilitation, 37, 30-37.

Kordy, H. (1997). Das Konzept der klinischen Signifikanz in der Psychotherapieforschung. In B. Strauss & J. Bengel (Hrsg.), Forschungsmethoden in der Medizinischen Psychologie (S. 129-145). Göttingen: Hogrefe.

Kordy, H. & Hannöver, W. (1998). Beobachten, Dokumentieren, Bewerten, Steuern: Qualitätsmanagement in der stationären Psychotherapie. In A. R. Laireiter & H. Vogel (Hrsg.), Qualitätssicherung in der Psychotherapie und psychosozialen Versorgung. Ein Werkstattbuch (S. 355-373). Tübingen: DGVT Deutsche Gesellschaft für Verhaltenstherapie.

Kordy, H. & Kächele, H. (1996). Ergebnisforschung in Psychotherapie und Psychosomatik. In J.M.Adler, J. M. Herrmann, K. Köhle, O. W. Schonecke, T. von Uexküll, & W. Wesiack (Hrsg.), Psychosomatische Medizin (pp. 490-501). München: Urban und Schwarzenberg.

Kordy, H. & Scheibler, D. (1984a). Individuumsorientierte Erfolgsforschung: Erfassung und Bewertung von Therapieeffekten anhand individueller Behandlungsziele - Teil 1:

Gibt es in der Ergebnisforschung eine "Lücke" für individuumsorientierte Verfahren? Zeitschrift für Klinische Psychologie, Psychopathologie und Psychotherapie, 32, 218-233.

Kordy, H. & Scheibler, D. (1984b). Individuumsorientierte Erfolgsforschung: Erfassung und Bewertung von Therapieeffekten anhand individueller Behandlungsziele - Teil 2: Anwendungs- und Auswertungsaspekte. Zeitschrift für Klinische Psychologie, Psychopathologie und Psychotherapie, 32, 309-318.

Kriebel, R. & Paar, G. H. (2000). Psychosomatische Rehabilitation: Möglichkeit und Wirklichkeit. Zehn-Jahres-Bericht der Gelderland-Klinik für Psychosomatik und Psychotherapie. Geldern: Johannes Keuck.

Larisch, M., Joksimovic, L., von dem Knesebeck, O., Starke, D., & Siegrist, J. (2003). Berufliche Gratifikationskrisen und depressive Symptome. Eine Querschnittsstudie bei Erwerbstätigen im mittleren Erwachsenenalter. Psychotherapie, Psychosomatik, Medizinische Psychologie, 53, 223-228.

Leichsenring, F. (1996). Zur Meta-Analyse von Grawe und Mitarbeitern. Gruppenpsychotherapie und Gruppendynamik, 32, 205-234.

Leichsenring, F. (2004a). Randomized controlled versus naturalistic studies: A new research agenda. Bulletin of the Menninger Clinic, 68, 137-151.

Leichsenring, F. (2004b). "Empirically supported treatments": Wissenschaftstheoretische und methodische Aspekte kontrollierter vs. naturalistischer Studien. Zeitschrift für Klinische Psychologie, Psychiatrie und Psychotherapie, 52, 209-222.

Leonhart, R. (2004). Effektgrössenberechnung bei Interventionsstudien. Die Rehabilitation, 43, 241-246.

Lienert, G. A. & Raatz, U. (1998). Testaufbau und Testanalyse. Weinheim: Beltz.

Lipsey, M. W. & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. American Psychologist, 48, 1181-1209.

Lipsey, M. W. & Wilson, D. B. (2001). Practical meta-analysis. Thousand Oaks: Sage Publications.

Lord, F. M. (1956). The measurement of growth. Educational and Psychological Measurement, 16, 421-437.

Lord, F. M. (1958a). The utilization of unreliable difference scores. Journal of Educational Psychology, 49, 150-152.

Lord, F. M. (1958b). Further problems in the measurement of growth. Educational and Psychological Measurement, 18, 437-451.

Lösel, F. (1987). Methodik und Problematik von Meta-Analysen - Mit Beispielen der Psychotherapieforschung. Gruppendynamik, 18, 323-343.

Luber, E. & Geene, R. (2004). Qualitätssicherung und Evidenzbasierung in der Gesundheitsförderung.

Luborsky, L. (1980). Predicting the outcome of psychotherapy: Findings of the Penn Psychotherapy Project. Archives of General Psychiatry, 37, 471-481.

Maier-Riehle, B. & Härter, M. (1996). Die Effektivität von Rückenschulen aus empirischer Sicht. Zeitschrift für Gesundheitspsychologie, 4, 197-219.

Maier-Riehle, B. & Zwingmann, C. (2000). Effektstärkevarianten beim Eingruppen-PRÄ-POST-Design: Eine kritische Betrachtung. Die Rehabilitation, 39, 189-199.

Mans, E. J. (1995). Die Patientenzufriedenheit als Kriterium der Qualitätssicherung in der stationären psychosomatischen Rehabilitation. Das Gesundheitswesen, 52, 63-68.

McNemar, Q. (1958). On growth measurement. Educational and Psychological Measurement, 18, 47-55.

Mestel, R., Erdmann, A., Schmid, M., Klingelhöfer, J., Stauss, K., & Hautzinger, M. (2000). 1-3-Jahres-Katamnesen bei 800 depressiven Patienten nach stationärer Psychosomatischer Rehabilitation. In M. Bassler (Hrsg.), Leitlinien zur stationären Psychotherapie - Pro und Contra. Mainzer Werkstatt über empirische Forschung von stationärer Psychotherapie 1999 (S. 243-273). Giessen: Psychosozial-Verlag.

Moskowitz, D. S. & Hershberger, S. L. (Eds.). (2002). Modeling intraindividual variability with repeated measures data: Methods and applications. Mahwah: Lawrence Erlbaum Associates Publishers.

Muthny, F. A. & Bullinger, M. (1999). Variablen und Erhebungsinstrumente in der rehabilitationswissenschaftlichen Forschung - Würdigung und Empfehlungen. DRV-Schriften, 16, 84-102.

Müller, J. M. (2002). Umgang mit fehlenden Werten. In A. Reusch (Hrsg.), Empfehlungen zum Umgang mit Daten in der Rehabilitationsforschung (Regensburg: Roderer.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. Psychological Bulletin, 88, 622-637.

Neun, H., Dahlmann, W., Geyer, M., & Potreck-Rose, F. (1994). Psychosomatische Einrichtungen. Was sie (anders) machen und wie man sie finden kann. Göttingen: Vandenhöck und Ruprecht.

Nosper, M. (1999). Psychosomatische Rehabilitation: Untersuchungen zur Ergebnis- und Prozessqualität stationärer Einzel- und Gruppenpsychotherapien. Berlin: Logos-Verlag.

Nübling, R. (1992). Psychotherapiemotivation und Krankheitskonzept. Frankfurt a. M.: VAS Verlag für Akademische Schriften.

Nübling, R., Bürgy, R., Meyerberg, J., Oppl.-M., Kieser, J., Schmidt, J. & Wittmann, W. W. (2000). Stationäre psychosomatische Rehabilitation in der Klinik Bad Herrenalb: Erste Ergebnisse einer Katamnesestudie. In M. Bassler (Hrsg.), Leitlinien zur stationären Psychotherapie - Pro und Contra. Mainzer Werkstatt über empirische Forschung von stationärer Psychotherapie 1999 (S. 274-300). Giessen: Psychosozial-Verlag.

Nübling, R., Hafen, K., Jastrebow, J., Körner, M., Löschmann, C., Rundel, M., Schmidt, J., Wirtz, M., Bengel, J. (2004). Indikation zu psychotherapeutischen und psychosozialen Massnahmen im Rahmen stationärer medizinischer Rehabilitation. Regensburg: Roderer.

Nübling, R., Puttendörfer, J., Schmidt, J., & Wittmann, W. W. (1994). Längerfristige Ergebnisse psychosomatischer Rehabilitation. In F. Lamprecht & R. Johnen (Hrsg.), Salutogenese. Ein neues Konzept in der Psychosomatik? Kongressband der 40. Jahrestagung des Deutschen Kollegiums für Psychosomatische Medizin (S. 254-270). Frankfurt a. M.: VAS Verlag für Akademische Schriften.

Nübling, R., Puttendörfer, J., Wittmann, W. W., Schmidt, J., & Wittich, A. (1995). Evaluation psychosomatischer Heilverfahren. Die Rehabilitation, 34, 74-80.

Nübling, R. & Schmidt, J. (1998). Interne Qualitätssicherung in der stationären psychosomatischen Rehabilitation: Erfahrungen mit einem "zweigleisigen Modell". In A. R. Laireiter & H. Vogel (Hrsg.), Qualitätssicherung in der Psychotherapie und psychosozialen Versorgung. Ein Werkstattbuch (S. 335-353). Tübingen: DGVT Deutsche Gesellschaft für Verhaltenstherapie.

Nübling, R. & Schmidt, J. (2000). Methodische Grundlagen der Ergebnisevaluation. In J. Bengel & U. Koch (Hrsg.), Grundlagen der Rehabilitationswissenschaften (S. 323-346). Berlin: Springer.

Nübling, R., Schmidt, J., & Wittmann, W. W. (1999). Langfristige Ergebnisse Psychosomatischer Rehabilitation. Psychotherapie, Psychosomatik, Medizinische Psychologie, 49, 343-353.

Nübling, R. & Schulz, H. (2002). FPTM - Fragebogen zur Psychotherapiemotivation. In E. Brähler, J. Schumacher, B. Strauss (Hrsg.), Diagnostische Verfahren in der Psychotherapie (S. 141-145). Göttingen: Hogrefe.

Nübling, R., Steffanowski, A., Körner, M., Rundel, M., Kohl, C., Löschmann, C. et al. (2007). Kontinuierliche Patientenbefragung als Instrument für das interne Qualitätsmanagement in Einrichtungen der Gesundheitsversorgung. Gesundheitsökonomie und Qualitätsmanagement, 12, 44-50.

Nübling, R., Steffanowski, A., Wittmann, W. W., & Schmidt, J. (2004). Strategien der Ergebnismessung am Beispiel der psychosomatischen Rehabilitation. Praxis Klinische Verhaltensmedizin und Rehabilitation, 17, 35-44.

Olderog, M. (1999). Metaanalyse zur Wirksamkeit psychologisch fundierter Behandlungskonzepte des chronischen dekompensierten Tinnitus. Zeitschrift für Medizinische Psychologie, 8, 5-18.

Orlinsky, D. E. (1994). "Learning from Many Masters". Ansätze zu einer wissenschaftlichen Integration psychotherapeutischer Behandlungsmodelle. Psychotherapeut, 39, 2-9.

Paar, G. H. (1997). Das Qualitätssicherungsprogramm der Rentenversicherungsträger. Psychotherapeut, 42, 156-162.

Petermann, F. (1978). Veränderungsmessung. Stuttgart: Kohlhammer.

Petermann, F. (2005). Evidenzbasierte Psychotherapie und Leitlinien - Hilfe oder Fessel? Verhaltenstherapie und Verhaltensmedizin, 26, 452-469.

Pollmann, H., Wilms, E., Hillmer, J., Hübner, P., Borrmann, H., & Zillessen, E. (1998). Patientenbefragung in der Rehabilitationsklinik als Instrument des internen Qualitätsmanagements. LVA Rheinprovinz Mitteilungen, 4/1998, 157-163.

Posavac, E. J. & Carey, R. G. (1980). Program evaluation. Methods and case studies. Englewood Cliffs: Prentice Hall.

Protz, W., Gerdes, N., Maier-Riehle, B., & Jäckel, W. H. (1998). Therapieziele in der medizinischen Rehabilitation. Die Rehabilitation, 37, 24-29.

Rabung, S. (2007). Qualitätssicherung durch (faire) Einrichtungsvergleiche? Zum Umgang mit dem Problem fehlender Werte im Kontext der einrichtungsvergleichenden Qualitätssicherung medizinischer Rehabilitation.

Rogosa, D. R. (1995). Myths and Methods: "Myths about Longitudinal Research" plus Supplemental Questions. In J. M. Gottman (Ed.), The Analysis of Change (pp. 3-66). Mahwah: Lawrence Erlbaum Associates.

Rogosa, D. R. & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. Journal of Educational Measurement, 20, 335-343.

Rosenbrock, R. (2004). Qualitätssicherung und Evidenzbasierung - Herausforderungen und Chancen für die Gesundheitsförderung. In E. Luber & R. Geene (Hrsg.), Qualitätssicherung und Evidenzbasierung in der Gesundheitsförderung. Wer weiss, was gut ist: Wissenschaft, Wirtschaft, Politik, BürgerInnen? (S. 59-73). Frankfurt a. M.: Mabuse-Verlag.

Rosenthal, R. & Rubin, D. B. (1983). A simple, general purpose display of magnitude of experimental effect. Journal of Educational Psychology, 74, 166-169.

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). Evaluation: A systematic approach (6th Edition).

Rubin, D. B. & Little, R. J. (1987). Statistical Analysis with Missing Data. New York: Wiley.

Rundel, M. (2001). Sensitivität und Spezifität der Screening Instrumente HADS, GHQ-12 und SSQ zur Entdeckung psychischer Störungen in der kardiologischen Rehabilitation.

Rüddel, H., Jürgensen, R., Terporten, G., & Mans, E. (2002). Vergleich von Rehabilitationsergebnissen aus einer psychosomatischen Fachklinik mit integriertem vollstationären und teilstationären Rehabilitationskonzept. Die Rehabilitation, 41, 189-191.

Sackett, D. L. & Rosenberg, S. E. (1997). Was ist Evidenz-basierte Medizin und was nicht? Münchner Medizinische Wochenzeitschrift, 139, 644-645.

Sandweg, R., Sänger-Alt, C., & Rudolf, G. (1991). Erfolge in der stationären Psychotherapie. Das öffentliche Gesundheitswesen, 53, 801-809.

Schilter, B. (2000). Therapie des chronischen subjektiven Tinnitus. Frankfurt a. M.: VAS Verlag für Akademische Schriften.

Schmacke, N. (2006). Evidenzbasierte Medizin und Psychotherapie: die Frage nach den angemessenen Erkenntnismethoden. Psychotherapie, Psychosomatik, Medizinische Psychologie, 56, 202-209.

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

Schmidt, J. (1991). Evaluation einer Psychosomatischen Klinik. Frankfurt a. M.: VAS - Verlag für Akademische Schriften.

Schmidt, J., Bernhard, P., Wittmann, W. W., & Lamprecht, F. (1987). Die Unterscheidung zwischen singulären und multiplen Ergebniskriterien. In F. Lamprecht (Ed.), Spezialisierung und Integration in Psychosomatik und Psychotherapie. Deutsches Kollegium für psychosomatische Medizin, 6.-8. März 1986 (pp. 293-299). Berlin: Springer.

Schmidt, J., Karcher, S., Steffanowski, A., Nübling, R., & Wittmann, W. W. (2000). Die EQUA-Studie - Erfassung der Ergebnisqualität stationärer psychosomatischer Rehabilitationsbehandlungen. In J. Bengel & W. H. Jäckel (Hrsg.), Zielorientierung in der Rehabilitation. Rehabilitationswissenschaftlicher Forschungsverbund Freiburg/Bad Säckingen (pp. 109-117). Regensburg: Roderer.

Schmidt, J., Lamprecht, F., Bernhard, P., & Nübling, R. (1989). Zur Nachgeschichte stationär psychosomatisch behandelter Patienten. In H. Speidel & B. Strauss (Hrsg.), Zukunftsaufgaben der psychosomatischen Medizin. Deutsches Kollegium für psychosomatische Medizin, 12.-14. November 1987 (pp. 432-444). Berlin: Springer.

Schmidt, J., Lamprecht, F., Nübling, R., & Wittmann, W. W. (1994). Veränderungsbeurteilungen von Patienten und von Haus- und Fachärzten nach psychosomatischer Rehabilitation. Psychotherapie, Psychosomatik, Medizinische Psychologie, 44, 108-114.

Schmidt, J., Nübling, R., & Lamprecht, F. (1992). Möglichkeiten klinikinterner Qualitätssicherung (QS) auf der Grundlage eines Basis-Dokumentations-Systems sowie erweiterter Evaluationsstudien. Das Gesundheitswesen, 54, 70-80.

Schmidt, J., Nübling, R., Lamprecht, F., & Wittmann, W. W. (1994). Patientenzufriedenheit am Ende psychosomatischer Reha-Behandlungen. Zusammenhänge mit Behandlungs- und Ergebnisvariablen und prognostische Bedeutung. In F. Lamprecht & R. Johnen (Hrsg.), Salutogenese. Ein neues Konzept in der Psychosomatik? Kongressband der 40. Jahrestagung des Deutschen Kollegiums für Psychosomatische Medizin (pp. 271-283). Frankfurt a. M.: VAS Verlag für Akademische Schriften.

Schmidt, J., Nübling, R., Steffanowski, A., Lichtenberg, S., & Wittmann, W. W. (2006). Assessment of the Outcome Qualitäy of Inpatient Psychosomatic Rehabilitation. In W.H.Jäckel, J. Bengel, & J. Herdt (Eds.), Research in Rehabilitation - Results from a Research Network in Southwest Germany (Stuttgart.

Schmidt, J., Steffanowski, A., Nübling, R., Lichtenberg, S., & Wittmann, W. W. (2003). Ergebnisqualität stationärer psychosomatischer Rehabilitation. Regensburg: Roderer.

Schmidt, S. & Strauss, B. (1996). Die Bindungstheorie und ihre Relevanz für die Psychotherapie. Psychotherapeut, 41, 139-150.

Schmitz, N. & Davies-Osterkamp, S. (1997). Klinische und Statistische Signifikanz - diskutiert am Beispiel der Symptom Check Liste (SCL-90-R). Diagnostica, 43, 80-96.

Schöffski, O. & Graf von der Schulenburg, J. N. (Hrsg.). (2000). Gesundheitsökomomische Evaluationen. Berlin: Springer.

Schulte, D. (1993). Wie soll Therapieerfolg gemessen werden? Zeitschrift für klinische Psychologie, 22, 374-393.

Schulz, H., Lotz-Rambaldi, W., Koch, U., Jürgensen, R., & Rüddel, H. (1999). 1-Jahres-Katamnese stationärer psychosomatischer Rehabilitation nach differentieller Zuweisung zu psychoanalytisch oder verhaltenstherapeutisch orientierter Behandlung. Psychotherapie, Psychosomatik, Medizinische Psychologie, 49, 114-130.

Schwarzer, G., Türp, J. C., & Antes, G. (2004). Das Odds Ratio in Interventionsstudien. Deutsche Zahnärztliche Zeitschrift, 59, 10-10.

Sechrest, L., McKnight, P., & McKnight, K. (1996). Calibration of measures for psychotherapy outcome studies. American Psychologist, 51, 1065-1071.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin Co.

Shadish, W. R., Matt, G. E., Navarro, A. M., & Siegle, G. (1997). Evidence that therapy works in clincally representative conditions. Journal of Consulting and Clinical Psychology, 65, 355-365.

Siegrist, J. (2005). Stress am Arbeitsplatz. In R. Schwarzer (Hrsg.), Gesundheitspsychologie (S. 303-318). Göttingen: Hogrefe.

Smith, M. L. & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760.

Statistisches Bundesamt (2005). Grunddaten der Krankenhäuser 2004. Wiesbaden: Statistisches Bundesamt. Wiesbaden: Eigenverlag.

Statistisches Bundesamt (2006a). Statistisches Jahrbuch 2006 für die Bundesrepublik Deutschland. Wiesbaden: Eigenverlag.

Statistisches Bundesamt (2006b). Grunddaten der Vorsorge- oder Rehabilitationseinrichtungen 2004. Wiesbaden: Eigenverlag.

Statistisches Bundesamt (2006c). Lohnkosten 2005 (Pro-Kopf und Pro-Stunde-Angaben). Wiesbaden: Eigenverlag.

Steffanowski, A., Lichtenberg, S., Nübling, R., Wittmann, W. W., & Schmidt, J. (2003). Individuelle Ergebnismessung - Vergleich zwischen prospektiven und retrospektiven Problemangaben in der psychosomatischen Rehabilitation. Die Rehabilitation, 42, 22-29.

Steffanowski, A., Löschmann, C., Schmidt, J., Wittmann, W. W., & Nübling, R. (2007). Meta-Analyse der Effekte stationärer psychosomatischer Rehabilitation - MESTA-Studie. Bern: Huber.

Steffanowski, A., Nübling, R., Schmidt, J., & Löschmann, C. (2006). Patientenbefragungen in der medizinischen Rehabilitation - Computergestütztes Routinemonitoring der Struktur-, Prozess- und Ergebnisqualität. Praxis Klinische Verhaltensmedizin und Rehabilitation, 71, 35-46.

Steffanowski, A., Oppl, M., Meyerberg, J., Schmidt, J., Wittmann, W. W., & Nübling, R. (2001). Psychometrische Überprüfung einer deutschsprachigen Version des Relationship Scale Questionaire (RSQ). In M. Bassler (Hrsg.), Störungsspezifische Ansätze in der stationären Psychotherapie. Mainzer Werkstatt über empirische Forschung von stationärer Psychotherapie 2000 (S. 320-342). Giessen: Psychosozial-Verlag.

Stieglitz, R. D. (1990). Validitätsstudien zum retrospektiven Vortest in der Therapieforschung. Zeitschrift für klinische Psychologie, 19, 144-150.

Stieglitz, R. D. & Baumann, U. (1994). Veränderungsmessung. In R. D. Stieglitz & U. Baumann (Hrsg.), Psychodiagnostik psychischer Störungen (S. 21-36). Stuttgart: Enke.

Stump, S. & Koch, U. (1998). Behandlungsansätze und Interventionsstudien in der Psychoonkologie - Ergebnisse einer Metaanalyse. In U. Koch & J. Weis (Hrag.), Krankheitsbewältigung bei Krebs und Möglichkeiten der Unterstützung. Der Förderschwerpunkt "Rehabilitation von Krebskranken" (S. 357-370). Stuttgart: Schattauer.

Süss, H. M. (1995). Zur Wirksamkeit der Therapie bei Alkoholabhängigen: Ergebnisse einer Meta-Analyse. Psychologische Rundschau, 46, 248-266.

Süss, H. M. (1997). Methoden der Evaluation von Suchttherapie. In B. Strauss & J. Bengel (Hrsg.), Forschungsmethoden in der Medizinischen Psychologie (S. 244-256). Göttingen: Hogrefe.

Swart, E. & Philbert-Hasucha, S. (1998). Qualitätsmanagement im Krankenhaus. In E. Pinter (Hrsg.), Praxis Umfassendes Qualitätsmanagement (S. 282-307). Frankfurt a. M.: pmi-Verlagsgruppe.

Thurstone, L. L. (1954). An analytical method for simple structure. Psychometrika, 19, 173-182.

Tschuschke, V. (2005). Die Psychotherapie in Zeiten evidenzbasierter Medizin. Fehlentwicklungen und Korrekturvorschläge. Psychotherapeutenjournal, 4, 106-115.

Tschuschke, V., Kächele, H., & Hölzer, M. (1994). Gibt es unterschiedlich effektive Formen von Psychotherapie? Psychotherapeut, 39, 281-297.

Tucker, L. R., Damarin, F., & Messick, S. (1966). A Base-Free Measure of Change. Psychometrika, 31, 457-473.

Uexküll, T. (1996). Psychosomatische Medizin.

Verband Deutscher Rentenversicherungsträger. (1992). Rahmenkonzept für die medizinische Rehabilitation in der gesetzlichen Rentenversicherung. Deutsche Rentenversicherung Bund, 7-8, 441-467.

Verband Deutscher Rentenversicherungsträger. (1994). Das Reha-Qualitätssicherungsprogramm der gesetzlichen Rentenversicherung - Perspektiven und Ziele. Deutsche Rentenversicherung Bund, 11, 745-750.

World Health Organization. (1993). International Classification of Impairments, Activities and Participation (ICIDH-2). Genf: WHO.

Wirtz, M. (2004). Über das Problem fehlender Werte: Wie der Einfluss fehlender Informationen auf Analyseergebnisse entdeckt und reduziert werden kann. Die Rehabilitation, 43, 109-115.

Wittchen, H. U. & Jacobi, F. (2006). Psychische Störungen in Deutschland und der EU. Grössenordnung und Belastung. Verhaltenstherapie und psychosoziale Praxis, 38, 189-192.

Wittmann, W. W. (1985). Evaluationsforschung. Berlin: Springer.

Wittmann, W. W. (1987). Grundlagen erfolgreicher Forschung in der Psychologie: Multimodale Diagnostik, Multiplismus, multivariate Reliabilitäts- und Validitätstheorie. Diagnostica, 33, 209-226.

Wittmann, W. W. (1990). Brunswik-Symmetrie und die Konzeption der Fünf-Datenboxen. Zeitschrift für Pädagogische Psychologie, 4, 241-251.

Wittmann, W. W. (1995). Evaluation in der Rehabilitation: Methoden, Ergebnisse und Folgerungen für die Praxis. In K. Siek, F. W. Pape, W. Blumenthal, & M. Schmollinger (Hrsg.), Erfolgsbeurteilung in der Rehabilitation - Begründungen, Möglichkeiten, Erfahrungen (S. 77-88). Ulm: Universitätsverlag.

Wittmann, W. W. (2002). Brunswik-Symmetrie: Ein Schlüsselkonzept für erfolgreiche psychologische Forschung. In M. Myrtek (Hrsg.), Die Person im biologischen und sozialen Kontext (S. 163-186). Göttingen: Hogrefe.

Wittmann, W. W. (2003). Effektivität und Effizienz der medizinischen Rehabilitation von psychosomatischen Erkrankungen unter gesundheitsökonomischen Perspektiven aus dem Blickwinkel verschiedener Stakeholder. Düsseldorf: Expertentagung "Gesundheitsökonomie in der Psychosomatischen Rehabilitation" am 03.12.2003.

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

Wittmann, W. W. & Matt, G. E. (1986b). Aggregation und Symmetrie. Diagnostica, 32, 309-329.

Wittmann, W. W., Nübling, R., & Schmidt, J. (2002). Evaluationsforschung und Programmevaluation im Gesundheitswesen. Zeitschrift für Evaluation, 1, 39-60.

Wottawa, H. (1986). Evaluation. In B. Weidenmann, A. Krapp, M. Hofer, G. L. Huber, & H. Mandl (Hrsg.), Pädagogische Psychologie (S. 703-733). München: Urban und Schwarzenberg.

Wottawa, H. & Thierau, H. (1998). Lehrbuch Evaluation. Bern: Huber.

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

Zielke, M., Borgart, E. J., Carls, W., Herder, F., Lebenhagen, J., Leidig, S., Limbacher, K., Meermann, R., Reschenberg, I., & Schwickeratz, J. (2004a). Evaluation stationärer verhaltensmedizinischer Behandlung und Rehabilitation auf der Basis objektiv erfassbarer Krankheitsdaten (Arbeitsunfähigkeitsgeschehen, Aufenthalte im Akutkrankenhaus) bei psychischen und psychosomatischen Erkrankungen. Praxis Klinische Verhaltensmedizin und Rehabilitation, 17, 169-192.

Zielke, M., Borgart, E. J., Carls, W., Herder, F., Lebenhagen, J., Leidig, S., Limbacher, K., Meermann, R., Reschenberg, I., & Schwickeratz, J. (2004b). Ergebnisqualität und Gesundheitsökonomie verhaltensmedizinischer Psychosomatik in der Klinik. Lengerich: Pabst.

Zielke, M., Borgart, E. J., Carls, W., Herder, F., Lebenhagen, J., Leidig, S., Limbacher, K., Meermann, R., Reschenberg, I., & Schwickeratz, J. (2004c). Krankheitsverhalten, Ressourcenverbrauch und sozialmedizinische Problemstellungen bei Patienten mit

psychosomatischen Erkrankungen im Vorfeld stationärer verhaltensmedizinischer Behandlungen. Praxis Klinische Verhaltensmedizin und Rehabilitation, 17, 125-154.

Zweig, M. H. & Campbell, G. T. (1993). Receiver-Operator-Characteristic (ROC) Plots: A fundamental Evaluation Tool in Clinical Medicine. Clinical Chemistry, 39, 561-577.

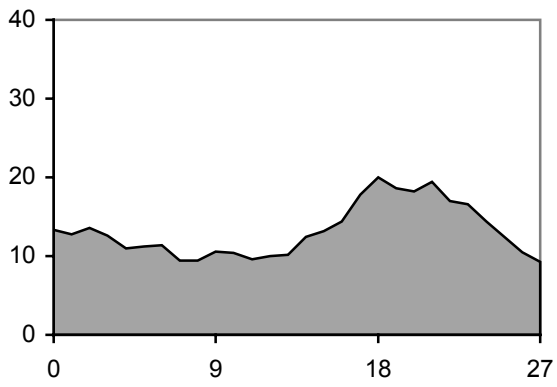
Zwingmann, C. (2003). Zielorientierte Ergebnismessung (ZOE) mit dem IRES-Patientenfragebogen: Eine kritische Zwischenbilanz. Die Rehabilitation, 42, 226-235.

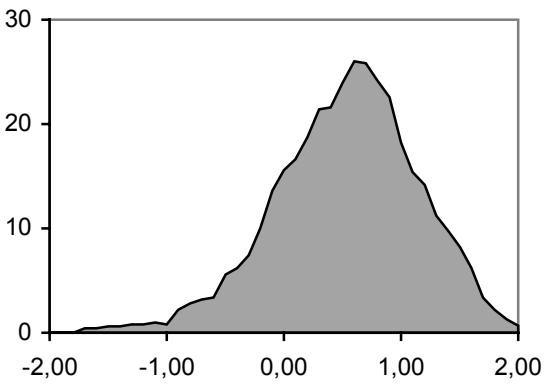
Zwingmann, C., Buschmann-Steinhage, R., Gerwinn, H., & Klosterhuis, H. (2004). Förderschwerpunkt "Rehabilitationswissenschaften": Ergebnisse - Umsetzung - Erfolge und Perspektiven. Die Rehabilitation, 43, 260-270.

8 Anhang

8.1 Getrennte Analysen für die Studien

8.1.1 Skalenkennwerte von EMEK_27





EMEK_27a	EMEK_27b
Min = 0; *Max* = 27	*Min* = -1,54; *Max* = +1,95
M = 14,08; *SD* = 7,89	*M* = 0,53; *SD* = 0,60
Schiefe = -0,30	Schiefe = -0,46
Wölbung = -1,18	Wölbung = 0,34
Z = 2,28; p <.001	Z = 0,92; p=.361
Cronbachs Alpha = .93	Cronbachs Alpha = .94

Abbildung 42. Studie A: Zauberberg II, *N* = 367 Katamnese-Antworter.

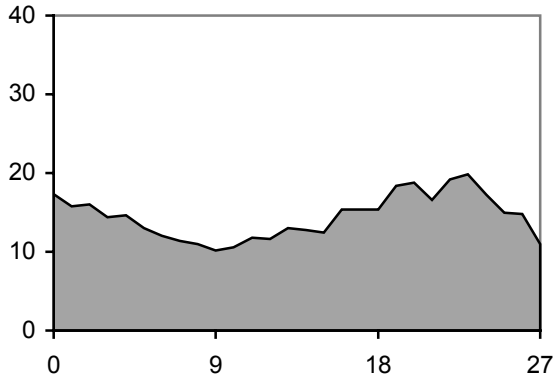
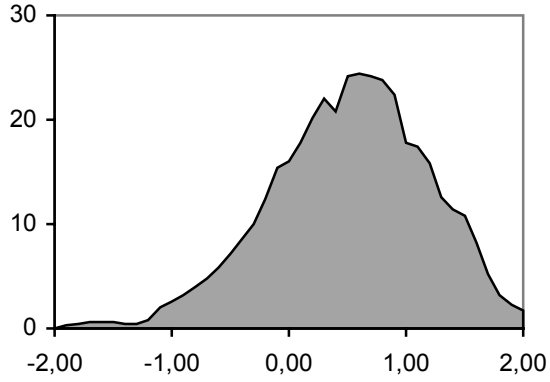
	
EMEK_27a *Min* = 0; *Max* = 27 *M* = 13,84; *SD* = 8,20 Schiefe = -0,18; Wölbung = -1,29 Z = 2,31; p<.001 Cronbachs Alpha = .94	EMEK_27b *Min* = -1,68; *Max* = +1,82 *M* = 0,51; *SD* = 0,65 Schiefe = -0,40 Wölbung = 0,03 Z = 0,81; p=.532 Cronbachs Alpha = .95

Abbildung 43. Studie B: Reinerzau, *N* = 401 Katamnese-Antworter.

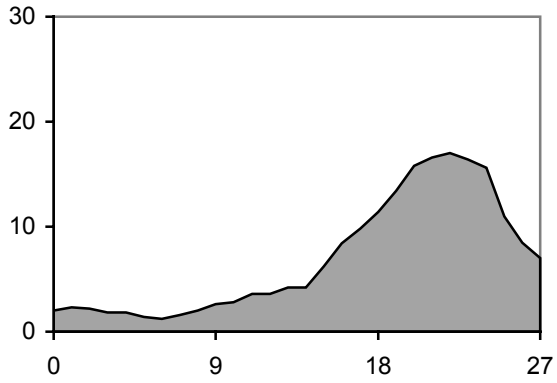
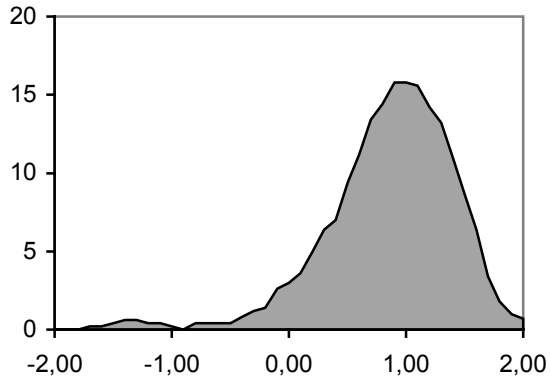
	
EMEK_27a *Min* = 0; *Max* = 26 *M* = 18,62; *SD* = 6,01 Schiefe = -1,30 Wölbung = 1,29 Z = 2,15; p<.001 Cronbachs Alpha = .89	EMEK_27b *Min* = -1,50; *Max* = +1,79 *M* = 0,86; *SD* = 0,54 Schiefe = -1,36 Wölbung = 3,43 Z = 1,18; p=.119 Cronbachs Alpha = .92

Abbildung 44. Studie C: Bad Herrenalb, *N* = 191 Katamnese-Antworter.

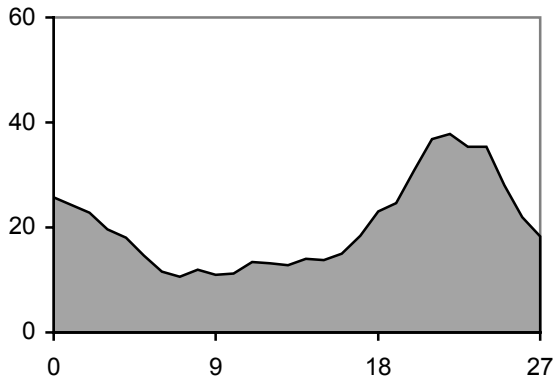
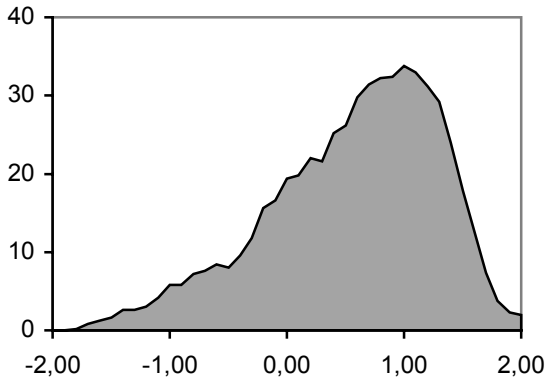
	
EMEK_27a *Min* = 0; *Max* = 27 *M* = 15,02; *SD* = 8,51 Schiefe = -0,45 Wölbung = -1,24 Z = 3,77 p<.001 Cronbachs Alpha = .95	EMEK_27b *Min* = -1,60; *Max* = +1,92 *M* = 0,56; *SD* = 0,71 Schiefe = -0,68 Wölbung = -0,05 Z = 1,81; p=.003 Cronbachs Alpha = .96

Abbildung 45. Studie D: EQUA, *N* = 569 Katamnese-Antworter.

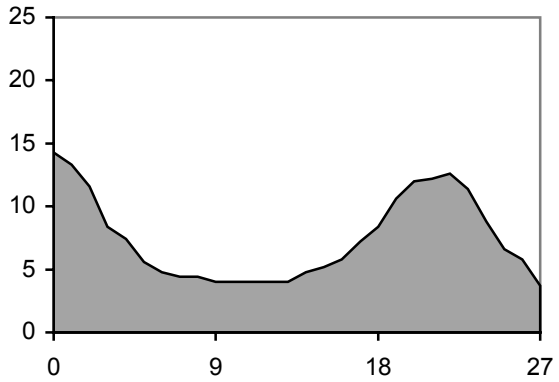
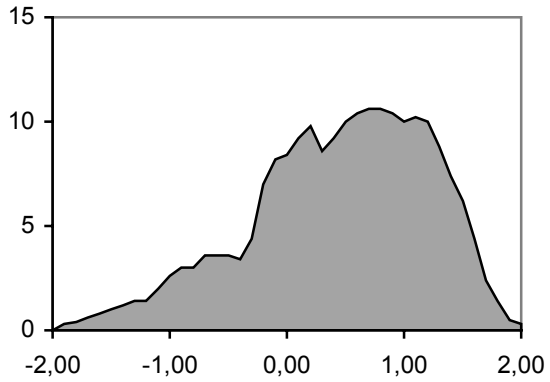
	
EMEK_27a *Min* = 0; *Max* = 27 *M* = 13,22; *SD* = 8,84 Schiefe = -0,22 Wölbung = -1,50 Z = 2,19; p<.001 Cronbachs Alpha = .95	EMEK_27b *Min* = -1,74; *Max* = +1,83 *M* = 0,45; *SD* = 0,75 Schiefe = -0,58 Wölbung = -0,17 Z = 0,92; p=.367 Cronbachs Alpha = .96

Abbildung 46. Studie E: INDIKA, *N* = 210 Katamnese-Antworter.

8.1.2 Faktorenanalyse von EMEK_27

Tabelle 41. Faktorenanalyse von EMEK_27a, dreifaktorielle Lösung

		Gesamtstichprobe				Studie A				Studie B				Studie C				Studie D				Studie E			
		F I	F II	F III	h²	F I	F II	F III	h²	F I	F II	F III	h²	F I	F II	F III	h²	F I	F II	F III	h²	F I	F II	F III	h²
1.	Befinden zum Katammesezeitpunkt	**.71**	.11	.18	.54	**.66**	.00	.30	.53	**.73**	.11	.24	.61	**.53**	.36	.10	.42	**.72**	.12	.13	.55	**.70**	.07	.25	.56
2.	Lebensqualität	**.78**	.17	.13	.66	**.73**	.16	.33	.67	**.79**	.12	.21	.68	**.62**	.32	.14	.50	**.82**	.18	.03	.71	**.74**	.19	.12	.60
3.	Körperliches Befinden	**.69**	.10	.29	.56	**.67**	.08	.34	.57	**.73**	.10	.26	.61	**.58**	.08	.31	.44	**.66**	.15	.37	.60	**.78**	.06	.13	.63
4.	Seelisches Befinden	**.79**	.16	.14	.67	**.69**	.20	.32	.62	**.75**	.18	.20	.63	**.61**	.28	.18	.48	**.85**	.11	.05	.74	**.84**	.15	.10	.74
5.	Allgemeinbefinden	**.81**	.13	.16	.70	**.76**	.11	.33	.71	**.84**	.08	.18	.75	**.64**	.17	.11	.45	**.80**	.15	.13	.69	**.86**	.17	.09	.78
6.	Leistungsfähigkeit	**.74**	.07	.23	.60	**.61**	.07	.39	.53	**.74**	.13	.21	.61	**.60**	-.09	.34	.48	**.84**	.11	.18	.76	**.74**	.11	.05	.56
7.	Beschwerden	**.77**	.10	.26	.67	**.71**	.15	.37	.67	**.75**	.07	.31	.66	**.67**	.19	.21	.53	**.80**	.08	.26	.72	**.84**	.12	.16	.75
8.	Gesundheitszustand	**.76**	.08	.22	.63	**.73**	.06	.39	.69	**.68**	.08	.22	.51	**.71**	.09	.24	.57	**.77**	.09	.26	.67	**.87**	.14	.14	.80
9.	Umgang mit Alltagsbelastungen	**.74**	.18	.07	.58	**.74**	.26	.12	.63	**.70**	.15	.04	.51	**.66**	.24	-.10	.51	**.78**	.15	.08	.64	**.76**	.15	.07	.61
10.	Gesundheitsbewußte Lebensführung	**.27**	.17	.05	.10	**.32**	.03	.08	.11	**.28**	.20	-.14	.14	**.27**	.23	-.12	.14	.17	**.28**	.25	.17	**.28**	.02	.09	.09
11.	Medikamentenkonsum	.16	.10	**.47**	.26	.06	.16	**.48**	.26	.23	.10	**.53**	.34	.13	-.07	**.32**	.12	.21	.14	**.51**	.32	**.17**	.04	.13	.05
12.	Beziehungen zu Bezugspersonen	.27	**.84**	.07	.79	.28	**.82**	.13	.77	.22	**.85**	.16	.79	.27	**.79**	.05	.70	.23	**.86**	.02	.79	.27	**.82**	.03	.75
13.	Beziehung zum Partner	.19	**.89**	.09	.84	.10	**.89**	.17	.82	.20	**.85**	.14	.79	.10	**.89**	.07	.80	.18	**.91**	.03	.86	.21	**.88**	.09	.84
14.	Familienleben mit Kindern	.22	**.87**	.07	.80	.21	**.84**	.13	.76	.18	**.86**	.12	.80	.18	**.85**	.04	.76	.20	**.85**	.04	.76	.20	**.89**	.11	.84
15.	Arbeitsfähigkeit	**.60**	.12	.27	.45	**.44**	.03	.43	.37	**.60**	.22	.25	.47	**.57**	.05	.35	.45	**.64**	.11	.28	.50	**.65**	.26	.10	.51
16.	Arztbesuche	.18	.06	**.66**	.47	.20	.13	**.61**	.43	.34	.03	**.56**	.43	-.04	-.02	**.69**	.47	.14	.08	**.63**	.42	.10	.06	**.75**	.58
17.	Krankschreibungszeiten (AU)	.18	.03	**.76**	.61	.11	.03	**.69**	.49	.26	.06	**.64**	.47	.02	.14	**.82**	.70	.20	.03	**.75**	.60	.26	.07	**.78**	.68
18.	Krankenhaustage	-.06	-.01	**.56**	.32	-.04	.01	**.51**	.27	-.16	.03	**.56**	.34	-.01	.07	**.44**	.20	-.08	-.11	**.52**	.29	.07	.05	**.72**	.53
19.	Wohlbefinden	**.82**	.15	.12	.71	**.76**	.25	.13	.66	**.81**	.12	.16	.69	**.69**	.22	.11	.53	**.85**	.13	.13	.76	**.87**	.18	.12	.80
20.	Umgang mit Problemen	**.71**	.17	.07	.53	**.60**	.08	.07	.37	**.62**	.24	.10	.46	**.56**	.26	-.04	.39	**.78**	.18	.07	.64	**.80**	.22	.18	.73
21.	Selbsthilfe	**.69**	.18	.04	.52	**.74**	.13	-.04	.56	**.66**	.26	-.10	.51	**.52**	.08	.06	.28	**.69**	.23	.12	.54	**.74**	.26	.15	.64
22.	Umgang mit Enttäuschungen	**.70**	.17	-.02	.51	**.68**	.11	-.13	.49	**.61**	.29	-.10	.47	**.61**	.02	-.12	.38	**.74**	.21	.07	.60	**.73**	.17	.13	.58
23.	Zurechtkommen mit Arbeit	**.65**	.12	.09	.45	**.58**	-.01	.10	.34	**.58**	.25	.01	.40	**.50**	.10	.06	.27	**.71**	.13	.17	.55	**.73**	.24	.10	.60
24.	Belastbarkeit	**.72**	.15	.11	.56	**.63**	.19	.04	.44	**.63**	.21	.15	.47	**.71**	.02	.06	.51	**.79**	.18	.19	.70	**.74**	.18	.14	.60
25.	Auskommen Mitmenschen	**.58**	.30	-.13	.45	**.53**	.32	-.25	.44	**.49**	.46	-.21	.50	**.52**	.16	-.17	.33	**.64**	.29	-.03	.49	**.63**	.29	.01	.48
26.	Leben mit Einschränkungen	**.74**	.20	.02	.59	**.71**	.25	-.09	.58	**.64**	.31	.04	.50	**.69**	.02	-.15	.50	**.81**	.16	.07	.69	**.71**	.22	.17	.58
27.	Ausgeglichenheit	**.75**	.18	.02	.60	**.69**	.18	.00	.51	**.69**	.25	.01	.53	**.64**	.30	-.09	.51	**.80**	.19	.05	.68	**.84**	.12	.16	.75
	Varianzaufklärung in %	*38,3*	*10,3*	*7,6*	*56,2*	*32,8*	*10,1*	*10,0*	*52,9*	*35,3*	*11,6*	*7,4*	*54,3*	*27,4*	*10,8*	*7,8*	*46,0*	*42,4*	*10,8*	*7,7*	*60,9*	*43,1*	*10,8*	*7,8*	*61,7*

Tabelle 42. Faktorenanalyse von EMEK_27b, dreifaktorielle Lösung

		Gesamtstichprobe				Studie A				Studie B				Studie C				Studie D				Studie E			
		F I	F II	F III	h²	F I	F II	F III	h²	F I	F II	F III	h²	F I	F II	F III	h²	F I	F II	F III	h²	F I	F II	F III	h²
1.	Befinden zum Katamnesezeitpunkt	**.74**	.17	.20	.62	**.71**	.16	.23	.58	**.77**	.15	.19	.66	**.64**	.28	.24	.55	**.74**	.20	.18	.62	**.79**	.11	.19	.67
2.	Lebensqualität	**.82**	.23	.17	.75	**.79**	.29	.21	.75	**.81**	.23	.20	.74	**.73**	.23	.30	.67	**.85**	.20	.12	.78	**.81**	.16	.21	.73
3.	Körperliches Befinden	**.75**	.08	.29	.65	**.73**	.06	.31	.64	**.79**	.09	.29	.72	.49	**.55**	-.05	.54	**.77**	.11	.25	.67	**.81**	.07	.17	.69
4.	Seelisches Befinden	**.82**	.22	.14	.74	**.75**	.26	.24	.69	**.79**	.27	.18	.73	**.65**	.33	.11	.55	**.87**	.21	.07	.80	**.87**	.20	.10	.80
5.	Allgemeinbefinden	**.84**	.15	.18	.76	**.82**	.19	.25	.78	**.84**	.15	.23	.79	**.64**	.43	-.04	.59	**.84**	.19	.10	.76	**.91**	.13	.11	.85
6.	Leistungsfähigkeit	**.80**	.09	.26	.72	**.77**	.06	.32	.70	**.80**	.08	.32	.75	.55	**.58**	-.10	.65	**.88**	.16	.16	.83	**.76**	.13	.23	.65
7.	Beschwerden	**.80**	.10	.26	.73	**.77**	.18	.31	.72	**.79**	.10	.29	.72	**.61**	.48	-.05	.61	**.84**	.10	.22	.77	**.88**	.12	.15	.81
8.	Gesundheitszustand	**.80**	.11	.22	.70	**.81**	.10	.35	.79	**.77**	.06	.27	.67	**.69**	.36	.10	.62	**.82**	.11	.22	.73	**.88**	.15	.15	.82
9.	Umgang mit Alltagsbelastungen	**.73**	.20	.09	.58	**.64**	.37	.05	.54	**.68**	.18	.08	.50	**.69**	.07	.05	.48	**.80**	.21	.09	.68	**.76**	.17	.06	.61
10.	Gesundheitsbewußte Lebensführung	**.27**	.18	.00	.11	**.26**	.18	-.02	.10	**.30**	.19	-.06	.13	**.40**	-.09	.11	.18	.19	**.28**	.16	.14	**.26**	-.04	.02	.07
11.	Medikamentenkonsum	.23	.10	**.43**	.25	.13	.10	**.43**	.21	.31	.16	**.43**	.31	.19	**.40**	-.08	.21	.27	.13	**.49**	.33	**.19**	.03	.17	.07
12.	Beziehungen zu Bezugspersonen	.26	**.85**	.09	.80	.24	**.82**	.17	.76	.23	**.85**	.13	.79	.31	.12	**.75**	.67	.22	**.88**	.03	.83	.21	**.83**	.06	.74
13.	Beziehung zum Partner	.19	**.90**	.08	.85	.13	**.86**	.22	.81	.18	**.88**	.12	.82	.19	.07	**.86**	.78	.20	**.91**	.01	.86	.12	**.88**	.02	.79
14.	Familienleben mit Kindern	.26	**.84**	.06	.78	.23	**.80**	.19	.72	.23	**.84**	.06	.76	.27	.08	**.78**	.69	.25	**.85**	.00	.78	.28	**.87**	.02	.83
15.	Arbeitsfähigkeit	**.67**	.08	.29	.54	**.58**	-.07	.37	.47	**.66**	.12	.30	.54	**.48**	**.48**	.16	.49	**.73**	.08	.26	.61	**.71**	.22	.24	.61
16.	Arztbesuche	.16	.08	**.66**	.46	.24	.14	**.65**	.50	.31	.08	**.59**	.45	-.23	**.56**	.17	.40	.17	.05	**.63**	.42	.12	.18	**.72**	.56
17.	Krankschreibungszeiten (AU)	.18	.06	**.84**	.74	.13	.10	**.82**	.70	.24	.08	**.79**	.69	-.04	**.78**	.25	.67	.20	.06	**.83**	.74	.30	.05	**.80**	.73
18.	Krankenhaustage	-.04	-.04	**.68**	.46	-.04	.06	**.69**	.48	-.15	-.03	**.70**	.51	-.07	**.50**	.05	.26	-.03	-.05	**.65**	.43	.09	-.13	**.75**	.59
19.	Wohlbefinden	**.85**	.21	.13	.79	**.80**	.28	.10	.73	**.83**	.23	.13	.76	**.72**	.16	.29	.64	**.88**	.20	.16	.84	**.91**	.16	.13	.87
20.	Umgang mit Problemen	**.74**	.19	.07	.59	**.58**	.15	.00	.36	**.65**	.20	.16	.48	**.66**	-.03	.24	.49	**.83**	.23	.10	.75	**.82**	.21	.19	.76
21.	Selbsthilfe	**.75**	.24	.05	.62	**.70**	.28	-.02	.57	**.72**	.33	-.04	.63	**.59**	.05	.12	.37	**.77**	.28	.14	.70	**.78**	.24	.17	.70
22.	Umgang mit Enttäuschungen	**.74**	.18	.00	.58	**.62**	.20	-.09	.44	**.70**	.25	-.03	.56	**.67**	-.11	.11	.47	**.82**	.24	.04	.72	**.77**	.17	.18	.66
23.	Zurechtkommen mit Arbeit	**.73**	.10	.12	.56	**.62**	-.05	.15	.41	**.70**	.14	.04	.51	**.51**	.07	.26	.34	**.80**	.12	.18	.69	**.78**	.26	.20	.72
24.	Belastbarkeit	**.80**	.12	.15	.67	**.70**	.09	.14	.52	**.73**	.12	.11	.56	**.71**	.24	.14	.59	**.85**	.18	.18	.79	**.81**	.18	.23	.74
25.	Auskommen Mitmenschen	**.63**	.28	-.04	.47	**.49**	.33	-.18	.39	**.56**	.39	-.08	.47	**.70**	-.06	.22	.54	**.68**	.29	.07	.55	**.62**	.31	.12	.49
26.	Leben mit Einschränkungen	**.80**	.22	.07	.69	**.66**	.42	.00	.60	**.71**	.33	-.01	.61	**.81**	.03	.03	.66	**.86**	.19	.12	.79	**.82**	.21	.17	.74
27.	Ausgeglichenheit	**.79**	.20	.08	.67	**.65**	.25	.05	.49	**.71**	.22	.14	.57	**.70**	.01	.24	.54	**.85**	.24	.10	.79	**.87**	.18	.15	.81
	Varianzaufklärung in %	*43,0*	*10,7*	*8,7*	*62,4*	*35,6*	*11,5*	*10,1*	*57,2*	*40,5*	*11,6*	*8,8*	*60,9*	*31,4*	*11,6*	*9,6*	*52,6*	*48,2*	*11,7*	*8,3*	*68,2*	*47,6*	*10,9*	*8,6*	*67,1*

Tabelle 43. Faktorenanalyse von EMEK_27a, vierfaktorielle Lösung

		Gesamtstichprobe					Studie A					Studie B					Studie C					Studie D					Studie E				
		F I	F II	F III	F IV	h²	F I	F II	F III	F IV	h²	F I	F II	F III	F IV	h²	F I	F II	F III	F IV	h²	F I	F II	F III	F IV	h²	F I	F II	F III	F IV	h²
1.	Befinden zum Katamnesezeitpunkt	**.69**	.29	.12	.09	.58	**.71**	.26	.02	.09	.57	**.73**	.27	.13	.12	.64	**.44**	.31	.36	.04	.42	**.63**	.40	.10	-.01	.57	**.71**	.08	.23	-.20	.60
2.	Lebensqualität	**.73**	.36	.17	.04	.69	**.77**	.30	.18	.10	.73	**.81**	.26	.15	.05	.75	**.74**	.09	.23	.19	.64	**.77**	.32	.15	-.06	.72	**.75**	.19	.11	-.11	.62
3.	Körperliches Befinden	**.74**	.20	.13	.17	.64	**.73**	.25	.10	.12	.62	**.77**	.21	.14	.10	.67	.26	**.66**	.15	.11	.54	.51	**.62**	.14	.14	.69	**.79**	.07	.11	-.21	.69
4.	Seelisches Befinden	**.75**	.34	.17	.04	.71	**.73**	.29	.22	.10	.67	**.76**	.27	.21	.05	.69	**.64**	.21	.23	.18	.53	**.78**	.37	.09	-.06	.75	**.85**	.15	.09	-.05	.75
5.	Allgemeinbefinden	**.77**	.35	.14	.06	.74	**.80**	.31	.14	.10	.77	**.85**	.29	.11	.02	.81	**.55**	.35	.15	.04	.45	**.74**	.36	.13	.03	.69	**.87**	.17	.07	-.08	.79
6.	Leistungsfähigkeit	**.72**	.30	.09	.14	.64	**.70**	.20	.09	.18	.56	**.67**	.36	.10	.14	.61	.19	**.78**	.00	.10	.65	**.74**	.47	.09	.03	.78	**.75**	.11	.03	-.15	.60
7.	Beschwerden	**.79**	.27	.12	.14	.73	**.79**	.26	.17	.13	.73	**.81**	.19	.12	.14	.72	**.54**	.44	.18	.12	.53	**.68**	.53	.06	.08	.76	**.85**	.12	.15	-.10	.77
8.	Gesundheitszustand	**.74**	.31	.09	.12	.67	**.81**	.26	.09	.15	.75	**.65**	.29	.07	.13	.52	**.56**	.47	.08	.14	.57	**.64**	.54	.07	.07	.72	**.88**	.15	.12	-.15	.83
9.	Umgang mit Alltagsbelastungen	**.56**	.49	.15	.04	.58	.52	**.54**	.23	.08	.63	**.54**	.47	.08	.04	.51	**.68**	.17	.19	-.11	.54	**.75**	.26	.13	.03	.64	**.76**	.15	.06	.08	.62
10.	Gesundheitsbewußte Lebensführung	.10	**.31**	.14	.08	.13	.13	**.33**	.00	.16	.15	.04	**.43**	.07	-.01	.20	.17	.19	**.26**	-.18	.17	.10	.31	**.28**	.12	.20	**.28**	.00	.10	.23	.14
11.	Medikamentenkonsum	.28	-.05	.13	**.42**	.27	.18	-.04	.17	**.45**	.26	.35	-.06	.16	**.43**	.34	-.02	**.30**	-.03	.23	.15	.10	**.48**	.14	.32	.37	.20	.08	.09	**-.60**	.41
12.	Beziehungen zu Bezugspersonen	.19	.24	**.83**	.05	.79	.21	.23	**.81**	.10	.77	.20	.23	**.85**	.08	.82	.24	.14	**.80**	.02	.71	.24	.08	**.85**	-.01	.79	.28	**.83**	.01	-.07	.77
13.	Beziehung zum Partner	.16	.16	**.89**	.05	.85	.13	.07	**.89**	.11	.83	.20	.19	**.87**	.03	.84	.13	.01	**.89**	.07	.81	.18	.11	**.90**	-.02	.86	.22	**.88**	.09	.02	.84
14.	Familienleben mit Kindern	.16	.20	**.86**	.04	.81	.17	.16	**.84**	.08	.77	.15	.23	**.86**	.05	.82	.14	.12	**.87**	.01	.80	.20	.12	**.84**	.00	.77	.20	**.89**	.11	.02	.84
15.	Arbeitsfähigkeit	**.55**	.30	.12	.22	.45	**.44**	.22	.03	.37	.38	**.52**	.34	.18	.21	.47	.20	**.73**	.13	.13	.60	**.53**	.50	.09	.10	.55	**.66**	.26	.09	.00	.51
16.	Arztbesuche	.20	.08	.06	**.66**	.48	.24	.11	.12	**.65**	.50	.35	.13	.03	**.57**	.46	.01	.08	-.05	**.71**	.51	.11	.26	.08	**.59**	.43	.11	.06	**.76**	.02	.59
17.	Krankschreibungszeiten (AU)	.16	.13	.01	**.79**	.67	.13	.09	.01	**.79**	.65	.17	.23	.00	**.75**	.65	.05	.16	.11	**.83**	.72	.17	.29	.03	**.71**	.62	.27	.07	**.78**	-.05	.68
18.	Krankenhaustage	-.11	.08	-.04	**.63**	.42	-.08	.07	-.02	**.69**	.48	-.14	-.05	.02	**.66**	.45	.09	-.03	.03	**.49**	.26	.00	-.13	-.10	**.66**	.46	.08	.05	**.73**	.00	.53
19.	Wohlbefinden	**.69**	.47	.14	.06	.71	**.60**	.50	.23	.02	.66	**.74**	.38	.10	.07	.70	**.74**	.19	.15	.12	.62	**.79**	.35	.10	.04	.76	**.87**	.18	.11	.00	.80
20.	Umgang mit Problemen	.40	**.63**	.11	.10	.58	.37	**.48**	.05	.07	.38	.37	**.58**	.11	.20	.53	**.72**	-.01	.16	.04	.55	**.82**	.04	.16	.14	.72	**.80**	.20	.19	.20	.76
21.	Selbsthilfe	.32	**.71**	.11	.11	.63	.29	**.75**	.06	.11	.67	.34	**.67**	.10	.02	.58	**.60**	.08	.01	.10	.38	**.75**	-.01	.21	.21	.66	**.74**	.24	.16	.28	.70
22.	Umgang mit Enttäuschungen	.31	**.72**	.09	.05	.62	.18	**.77**	.04	.06	.63	.29	**.66**	.13	.02	.54	**.49**	.33	.02	-.20	.38	**.80**	.00	.19	.16	.70	**.73**	.15	.14	.23	.62
23.	Zurechtkommen mit Arbeit	.39	**.55**	.07	.12	.48	.30	**.52**	-.05	.18	.40	.29	**.61**	.10	.13	.49	.16	**.62**	.19	-.15	.46	**.68**	.25	.11	.13	.55	**.73**	.24	.10	.04	.60
24.	Belastbarkeit	.48	**.57**	.10	.12	.58	.36	**.55**	.15	.08	.46	.42	**.53**	.10	.22	.51	.43	**.60**	.07	-.12	.57	**.77**	.25	.16	.16	.71	**.74**	.17	.14	.10	.61
25.	Auskommen Mitmenschen	.18	**.71**	.21	-.04	.58	.10	**.64**	.26	-.11	.50	.15	**.67**	.29	-.07	.57	**.39**	.30	.18	-.25	.34	**.71**	-.07	.27	.07	.59	**.62**	.26	.03	.45	.65
26.	Leben mit Einschränkungen	.39	**.70**	.13	.07	.66	.27	**.74**	.19	.04	.66	.38	**.60**	.18	.13	.55	**.70**	.19	-.03	-.16	.55	**.83**	.12	.13	.10	.73	**.70**	.19	.18	.35	.68
27.	Ausgeglichenheit	.48	**.61**	.13	.03	.62	.40	**.58**	.14	.02	.52	.42	**.61**	.11	.09	.57	**.54**	.31	.30	-.16	.51	**.79**	.21	.16	.02	.69	**.84**	.11	.16	.16	.76
	Varianzaufklärung in %	*25,7*	*18,4*	*9,6*	*6,9*	*60,6*	*23,4*	*17,2*	*9,7*	*7,8*	*58,1*	*26,0*	*16,8*	*9,8*	*6,7*	*59,3*	*20,8*	*13,3*	*10,6*	*7,0*	*51,7*	*37,8*	*10,6*	*10,2*	*6,0*	*64,6*	*43,3*	*10,6*	*7,7*	*4,1*	*65,7*

Tabelle 44. Faktorenanalyse von EMEK_27b, vierfaktorielle Lösung

		Gesamtstichprobe					Studie A					Studie B					Studie C					Studie D					Studie E				
		FI	FII	FIII	FIV	h²	FI	FII	FIII	FIV	h²	FI	FII	FIII	FIV	h²	FI	FII	FIII	FIV	h²	FI	FII	FIII	FIV	h²	FI	FII	FIII	FIV	h²
1.	Befinden zum Katamnesezeitpunkt	**.70**	.35	.17	.10	.65	**.72**	.23	.19	.09	.61	**.75**	.29	.15	.07	.67	**.51**	.46	.27	.11	.55	**.71**	.18	.31	.03	.63	**.80**	.09	.15	.11	.68
2.	Lebensqualität	**.72**	.44	.22	.07	.76	**.75**	.35	.30	.08	.78	**.78**	.31	.23	.07	.76	**.62**	.43	.31	.10	.67	**.83**	.19	.26	.00	.78	**.82**	.14	.18	.08	.74
3.	Körperliches Befinden	**.81**	.22	.11	.14	.74	**.80**	.15	.12	.13	.70	**.83**	.19	.12	.13	.76	.16	**.77**	.08	.15	.64	**.70**	.10	.50	-.01	.75	**.82**	.06	.12	.21	.74
4.	Seelisches Befinden	**.76**	.39	.22	.03	.77	**.75**	.28	.29	.09	.73	**.79**	.25	.29	.03	.77	.32	**.73**	.25	-.07	.70	**.84**	.19	.26	-.07	.81	**.87**	.18	.07	.00	.80
5.	Allgemeinbefinden	**.79**	.38	.16	.06	.80	**.83**	.29	.22	.08	.82	**.86**	.24	.17	.06	.83	.27	**.82**	.12	-.03	.75	**.82**	.17	.24	-.01	.76	**.92**	.11	.07	.05	.86
6.	Leistungsfähigkeit	**.79**	.32	.10	.14	.76	**.82**	.20	.11	.15	.74	**.81**	.25	.09	.18	.76	.22	**.79**	.02	.19	.72	**.84**	.15	.34	-.01	.85	**.78**	.12	.19	.11	.66
7.	Beschwerden	**.82**	.30	.12	.13	.79	**.79**	.25	.22	.15	.75	**.81**	.22	.12	.14	.75	.26	**.80**	.10	.05	.73	**.78**	.09	.46	-.02	.83	**.89**	.10	.11	.11	.82
8.	Gesundheitszustand	**.80**	.31	.12	.09	.76	**.85**	.23	.14	.18	.83	**.81**	.17	.09	.10	.71	.53	**.55**	.13	.15	.62	**.76**	.10	.47	-.03	.80	**.89**	.13	.10	.17	.85
9.	Umgang mit Alltagsbelastungen	**.54**	.51	.17	.05	.58	.43	**.54**	.28	.06	.56	**.58**	.38	.15	.03	.50	**.60**	.35	.05	-.04	.49	**.79**	.19	.14	.04	.69	**.76**	.15	.04	-.11	.62
10.	Gesundheitsbewußte Lebensführung	.03	**.40**	.12	.06	.18	.05	**.42**	.07	.09	.19	.02	**.58**	.05	.09	.34	**.39**	.09	.09	-.10	.18	.16	**.28**	.23	.04	.16	**.25**	-.05	.04	-.28	.14
11.	Medikamentenkonsum	**.42**	-.10	.15	.32	.31	.25	-.08	.15	**.36**	.22	**.50**	-.17	.25	.25	.41	.13	.30	-.09	**.33**	.22	.18	.12	**.66**	.14	.50	.21	.03	.08	**.82**	.71
12.	Beziehungen zu Bezugspersonen	.18	.24	**.84**	.06	.80	.20	.24	**.82**	.11	.78	.21	.20	**.85**	.06	.80	.29	.14	**.76**	.08	.69	.24	**.88**	.02	.03	.83	.24	**.83**	.04	.06	.74
13.	Beziehung zum Partner	.16	.16	**.90**	.04	.86	.14	.13	**.89**	.13	.84	.18	.15	**.89**	.04	.84	.15	.09	**.90**	.01	.85	.20	**.91**	.08	-.03	.87	.14	**.88**	.02	.01	.79
14.	Familienleben mit Kindern	.20	.22	**.83**	.02	.79	.21	.19	**.81**	.11	.75	.19	.22	**.83**	.00	.77	.22	.15	**.83**	.00	.75	.25	**.84**	.10	-.05	.78	.30	**.86**	.01	.02	.83
15.	Arbeitsfähigkeit	**.62**	.33	.08	.21	.55	**.61**	.16	-.04	.28	.48	**.67**	.21	.13	.19	.55	**.39**	.44	.15	.36	.49	**.71**	.07	.27	.15	.61	**.72**	.21	.21	.07	.61
16.	Arztbesuche	.21	.07	.08	**.65**	.47	.24	.18	.11	**.68**	.57	.33	.14	.07	**.58**	.48	-.13	.01	.08	**.68**	.49	.13	.04	**.50**	.42	.44	.15	.18	**.72**	-.09	.59
17.	Krankschreibungszeiten (AU)	.22	.11	.05	**.84**	.77	.18	.09	.08	**.87**	.80	.31	.08	.09	**.79**	.73	.07	.14	.12	**.90**	.85	.20	.05	.41	**.73**	.75	.33	.04	**.78**	.09	.73
18.	Krankenhaustage	-.06	.09	-.06	**.74**	.57	-.01	.04	.03	**.77**	.59	-.11	.02	-.05	**.80**	.66	.03	.04	-.05	**.63**	.40	.04	-.06	-.01	**.80**	.64	.11	-.14	**.74**	.06	.59
19.	Wohlbefinden	**.68**	.54	.18	.07	.79	**.64**	.51	.22	.05	.73	**.74**	.42	.20	.04	.76	**.68**	.33	.26	.10	.65	**.86**	.19	.25	.06	.84	**.91**	.14	.10	.02	.87
20.	Umgang mit Problemen	.44	**.65**	.13	.08	.64	.40	**.46**	.07	.01	.38	**.50**	.47	.13	.16	.52	**.71**	.12	.16	.03	.54	**.85**	.22	.03	.13	.79	**.83**	.18	.18	-.19	.80
21.	Selbsthilfe	.39	**.73**	.16	.09	.71	.35	**.76**	.11	.10	.72	.47	**.64**	.23	.00	.68	**.64**	.14	.04	.11	.44	**.81**	.27	.00	.19	.76	**.79**	.22	.17	-.23	.76
22.	Umgang mit Enttäuschungen	.38	**.72**	.10	.03	.67	.25	**.76**	.01	.07	.64	.50	**.55**	.17	-.02	.58	**.69**	.13	.05	-.08	.50	**.85**	.22	-.04	.11	.79	**.78**	.15	.17	-.15	.68
23.	Zurechtkommen mit Arbeit	.49	**.57**	.05	.11	.58	.53	.35	-.10	.13	.42	**.53**	.49	.07	.04	.53	**.56**	.11	.19	.13	.38	**.81**	.11	.11	.17	.71	**.80**	.24	.18	.03	.72
24.	Belastbarkeit	.56	**.59**	.08	.12	.69	**.57**	.44	.04	.12	.53	**.59**	.47	.06	.09	.58	**.70**	.32	.08	.22	.65	**.86**	.16	.13	.16	.81	**.82**	.16	.20	-.01	.74
25.	Auskommen Mitmenschen	.25	**.70**	.20	.02	.59	.18	**.63**	.18	-.07	.47	.25	**.71**	.25	.03	.63	**.69**	.19	.19	-.07	.56	**.71**	.28	-.04	.14	.61	**.63**	.29	.11	-.16	.52
26.	Leben mit Einschränkungen	.48	**.69**	.16	.07	.74	.35	**.70**	.27	.08	.69	.50	**.59**	.24	.00	.65	**.81**	.26	-.04	.03	.73	**.87**	.17	.08	.12	.82	**.82**	.19	.16	-.15	.76
27.	Ausgeglichenheit	.51	**.63**	.15	.07	.69	**.49**	.47	.18	.04	.49	**.54**	.53	.14	.14	.61	**.70**	.20	.19	.00	.57	**.85**	.23	.13	.06	.80	**.88**	.16	.14	-.11	.82
	Varianzaufklärung in %	29,2	20,2	9,9	7,3	66,6	27,9	15,6	10,4	8,4	62,3	33,6	14,3	10,5	6,9	65,3	23,7	17,2	9,7	7,9	58,5	46,7	11,2	8,2	5,8	71,9	48,9	10,3	7,8	4,0	71,0

8.1.3 Vorhersage von EMEK_27 aus Stichprobenmerkmalen

Tabelle 45. Multiple Regression zur Vorhersage von EMEK_27a

EMEK_27a	**Studie A**	**Studie B**	**Studie C**	**Studie D**	**Studie E**
N	367	401	191	569	210
Multiple Korrelation R	**.24**	**.34**	.29	**.30**	**.44**
Geschlecht (1=m, 2=w)	.09	**.13**	.12	**.10**	**.19**
Alter in Jahren	-.04	**-.13**	-.01	-.03	-.11
Schulbildung (1=HS, 2=RS, 3=Abi)	.05	**.11**	.02	.07	.12
Familienstand verheiratet	.08	.02	-.10	.05	-.03
Erwerbstätig bei Aufnahme	-.04	-.04	.00	.02	.07
Kostenträger (0=KV, 1=RV)	.07	.03	-.08	**-.09**	**-.21**
Rentenantrag bei Aufnahme	**-.12**	**-.12**	**-.17**	**-.15**	.08
Krankheitsdauer in Jahren	-.06	**-.18**	.04	**-.11**	-.13
Somatoforme Hauptdiagnose	.07	-.04	.13	.05	.00
Nebendiagnose(n) vorhanden	-.02	-.08	.01	**-.09**	-.03
Behandlungsdauer in Tagen	**.14**	.04	.07	-.04	.01
Vorzeitige Entlassung	-.03	-.06	-.06	**-.10**	-.07

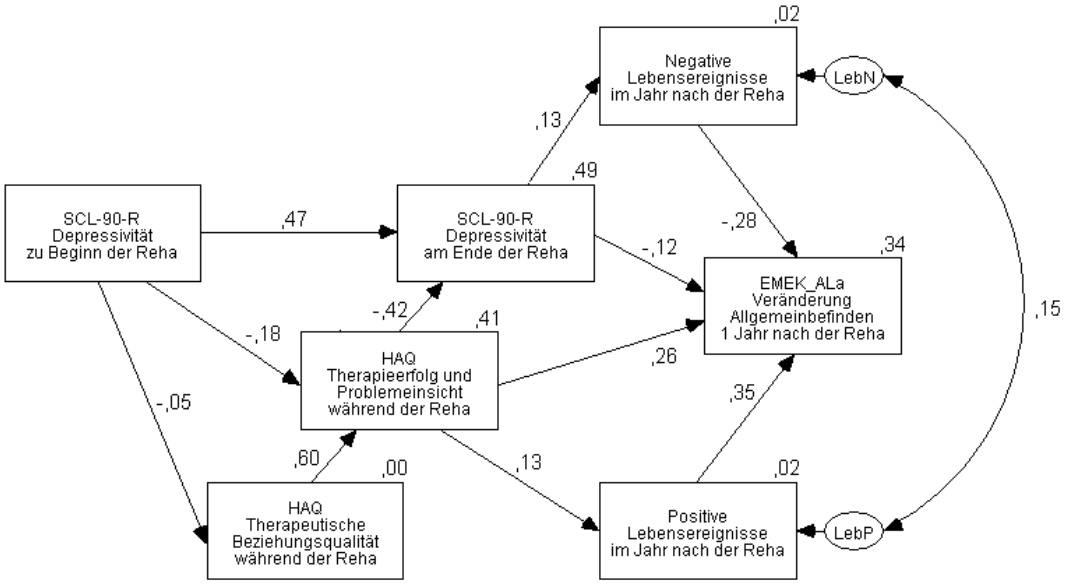
Anmerkung. Wiedergegeben sind die Beta-Gewichte (abgesehen von R). Statistisch signifikante Koeffizienten (<5%) sind durch Fettdruck hervorgehoben.

Tabelle 46. Multiple Regression zur Vorhersage von EMEK_27b

EMEK_27b	**Studie A**	**Studie B**	**Studie C**	**Studie D**	**Studie E**
N	367	401	191	569	210
Multiple Korrelation R	.21	**.33**	.25	**.32**	**.45**
Geschlecht (1=m, 2=w)	.07	.09	.10	**.10**	**.19**
Alter in Jahren	-.05	**-.13**	.04	-.04	**-.16**
Schulbildung (1=HS, 2=RS, 3=Abi)	.03	.08	.05	.08	**.14**
Familienstand verheiratet	.10	-.01	-.12	.05	-.06
Erwerbstätig bei Aufnahme	-.06	-.02	-.04	.04	.07
Kostenträger (0=KV, 1=RV)	.07	-.01	-.03	**-.12**	**-.23**
Rentenantrag bei Aufnahme	-.10	**-.15**	**-.16**	**-.14**	.08
Krankheitsdauer in Jahren	-.09	**-.18**	.04	**-.13**	-.09
Somatoforme Hauptdiagnose	.02	-.04	.09	.04	.00
Nebendiagnose(n) vorhanden	-.03	-.07	.04	**-.10**	.01
Behandlungsdauer in Tagen	.04	-.03	.04	-.07	-.08
Vorzeitige Entlassung	-.06	-.06	-.02	**-.09**	-.08

8.1.4 Vorhersage von EMEK_27 aus Prozessmerkmalen

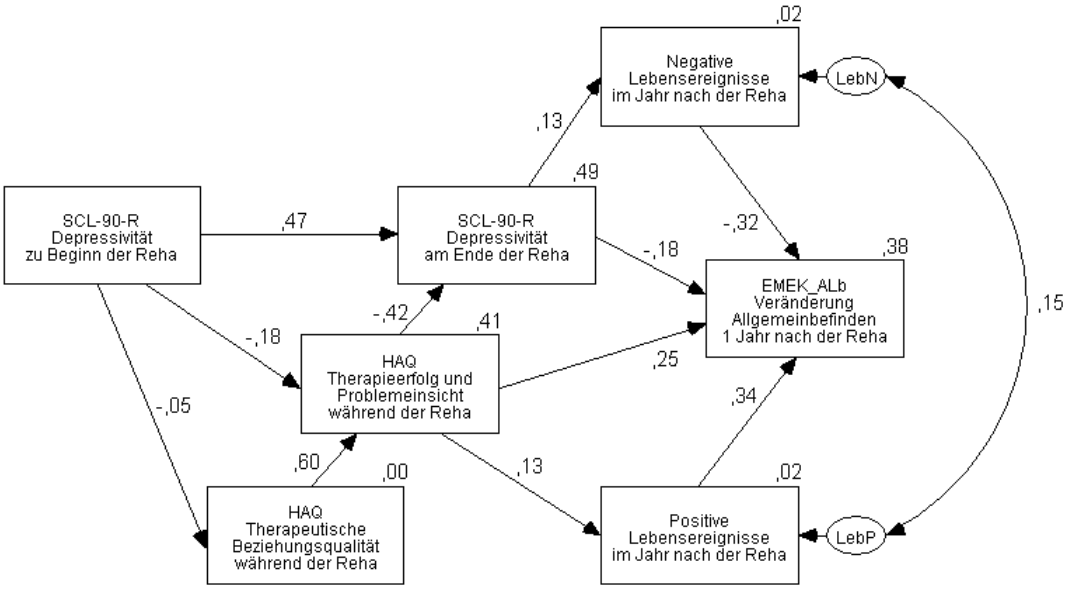
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 23,563; *df* = 9; χ^2 / *df* = 2,618; *p* = .001; R^2 = .338
SRMR = .030; *RMSEA* = .049 (*CI* 90%: .025 bis .074); *CFI* = .987

Abbildung 47. Pfadmodell zur Vorhersage von EMEK_ALa. *N* = 664 Patienten.

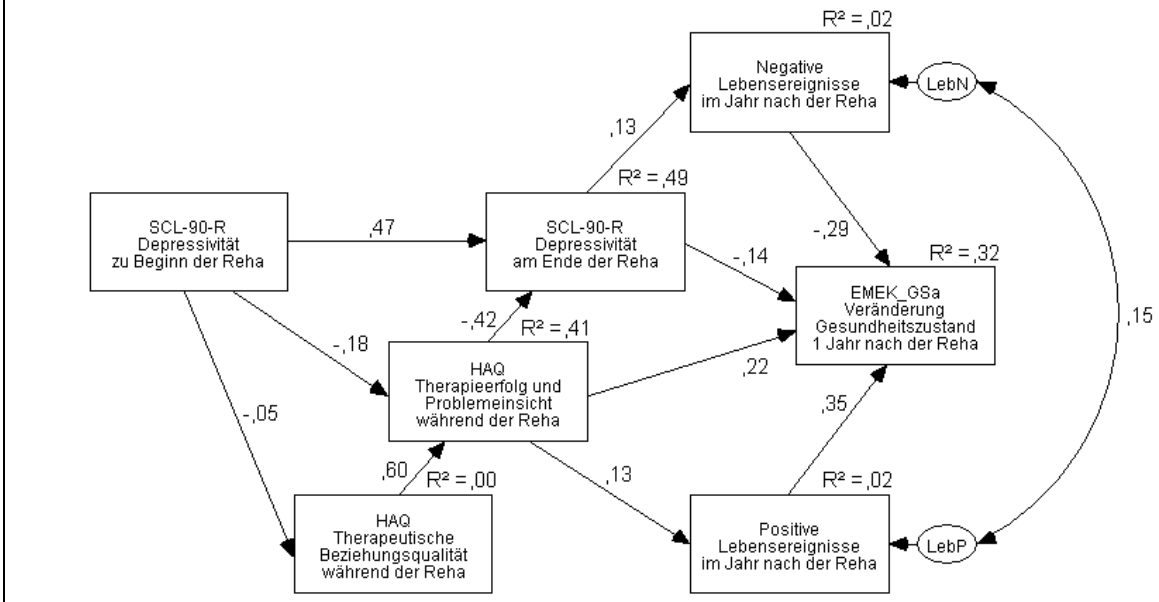
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 23,740; *df* = 9; χ^2 / *df* = 2,638; *p* = .005; R^2 = .377
SRMR = .030; *RMSEA* = .050 (*CI* 90%: .026 bis .075); *CFI* = .987

Abbildung 48. Pfadmodell zur Vorhersage von EMEK_ALb. *N* = 664 Patienten.

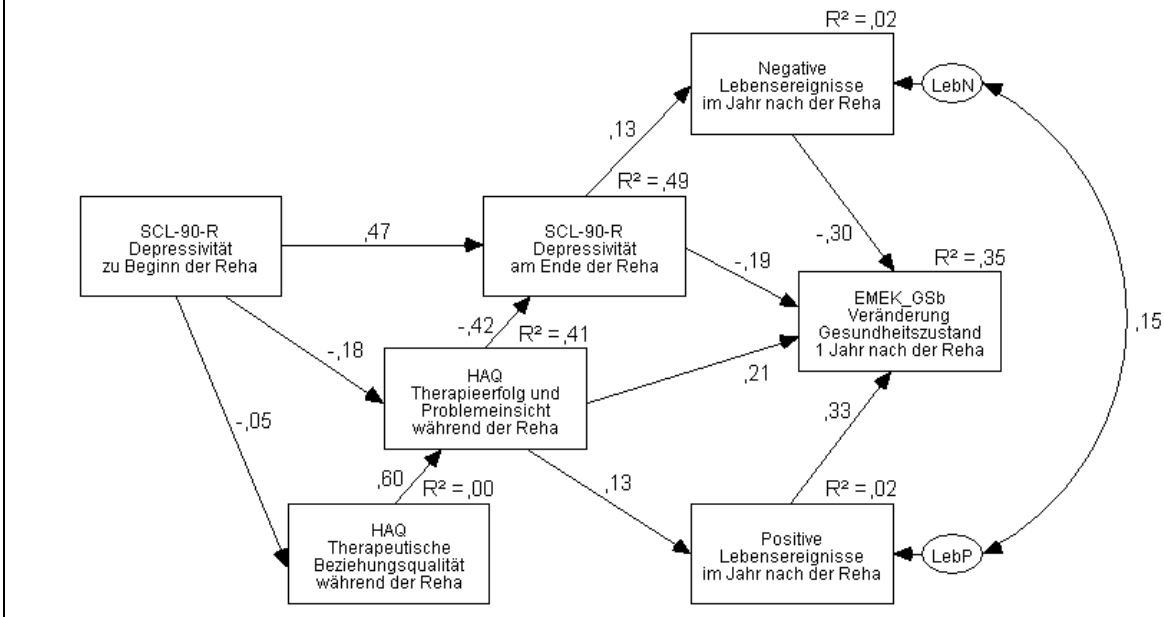
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 26,625; *df* = 9; χ^2 / *df* = 2,958; *p* = .001; R^2 = .323
SRMR = .032; *RMSEA* = .054 (*CI* 90%: .031 bis .079); *CFI* = .983

Abbildung 49. Pfadmodell zur Vorhersage von EMEK_GSa. *N* = 664 Patienten.

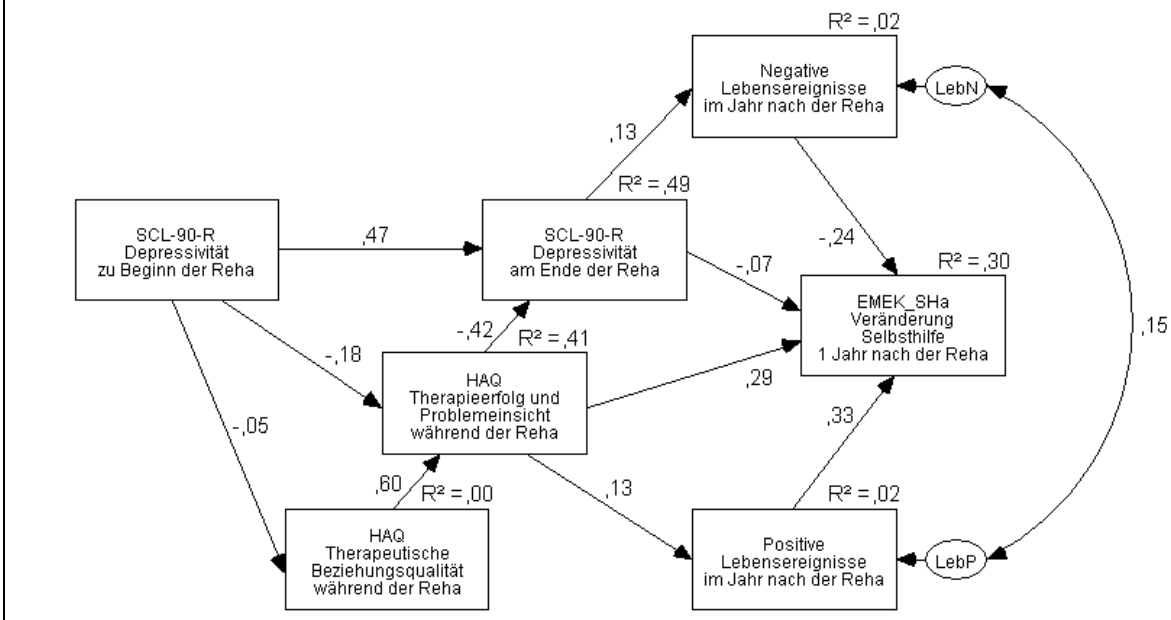
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 26,042; *df* = 9; χ^2 / *df* = 2,894; *p* = .002; R^2 = .353
SRMR = .031; *RMSEA* = .053 (*CI* 90%: .030 bis .078); *CFI* = .985

Abbildung 50. Pfadmodell zur Vorhersage von EMEK_GSb. *N* = 664 Patienten.

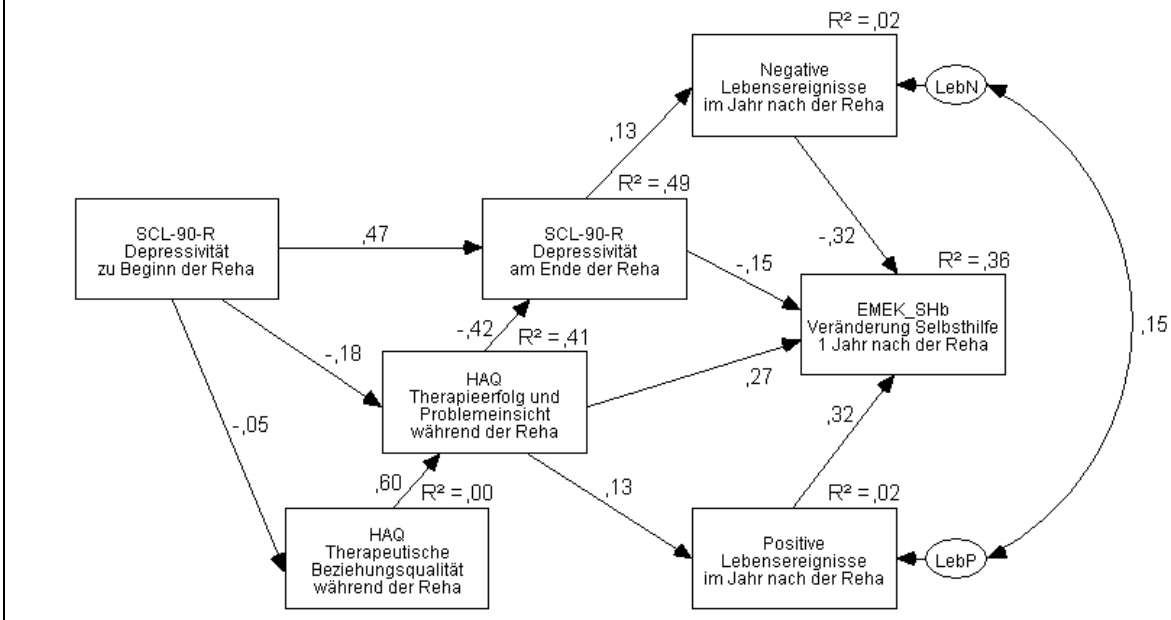
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 19,456; *df* = 9; χ^2 / *df* = 2,162; *p* = .022; R^2 = .297
SRMR = .028; *RMSEA* = .042 (*CI* 90%: .015 bis .068); *CFI* = .990

Abbildung 51. Pfadmodell zur Vorhersage von EMEK_SHa. *N* = 664 Patienten.

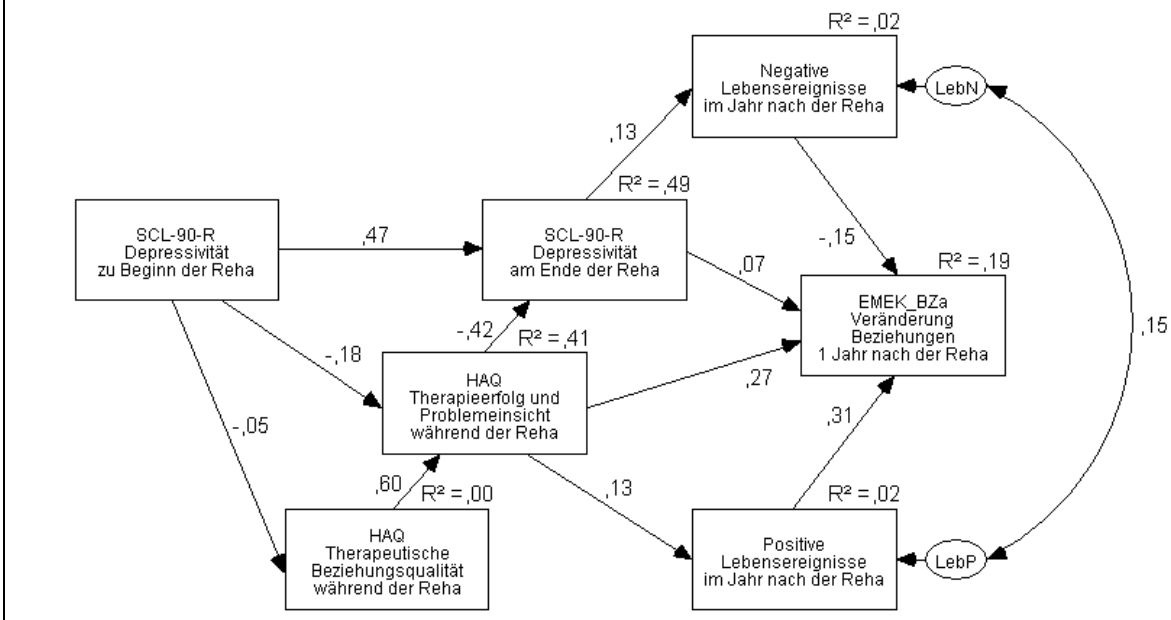
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 20,536; *df* = 9; χ^2 / *df* = 2,282; *p* = .015; R^2 = .360
SRMR = .030; *RMSEA* = .044 (*CI* 90%: .018 bis .069); *CFI* = .990

Abbildung 52. Pfadmodell zur Vorhersage von EMEK_SHb. *N* = 664 Patienten.

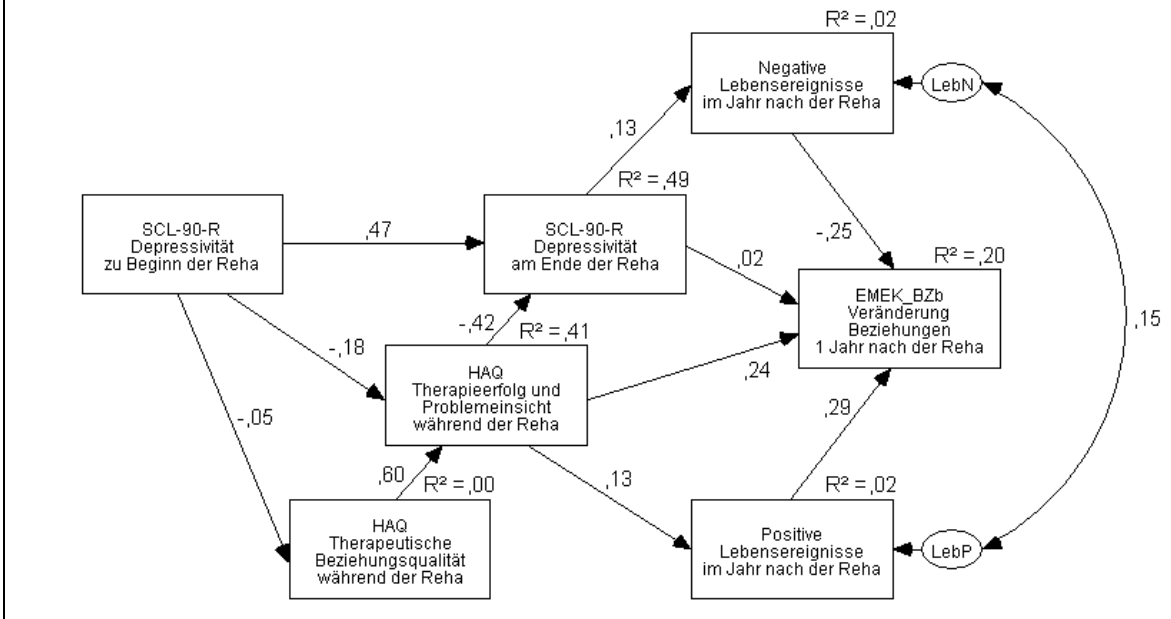
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 26,344; *df* = 9; χ^2 / *df* = 2,927; *p* = .002; R^2 = .187
SRMR = .032; *RMSEA* = .054 (*CI* 90%: .031 bis .078); *CFI* = .982

Abbildung 53. Pfadmodell zur Vorhersage von EMEK_BZa. *N* = 664 Patienten.

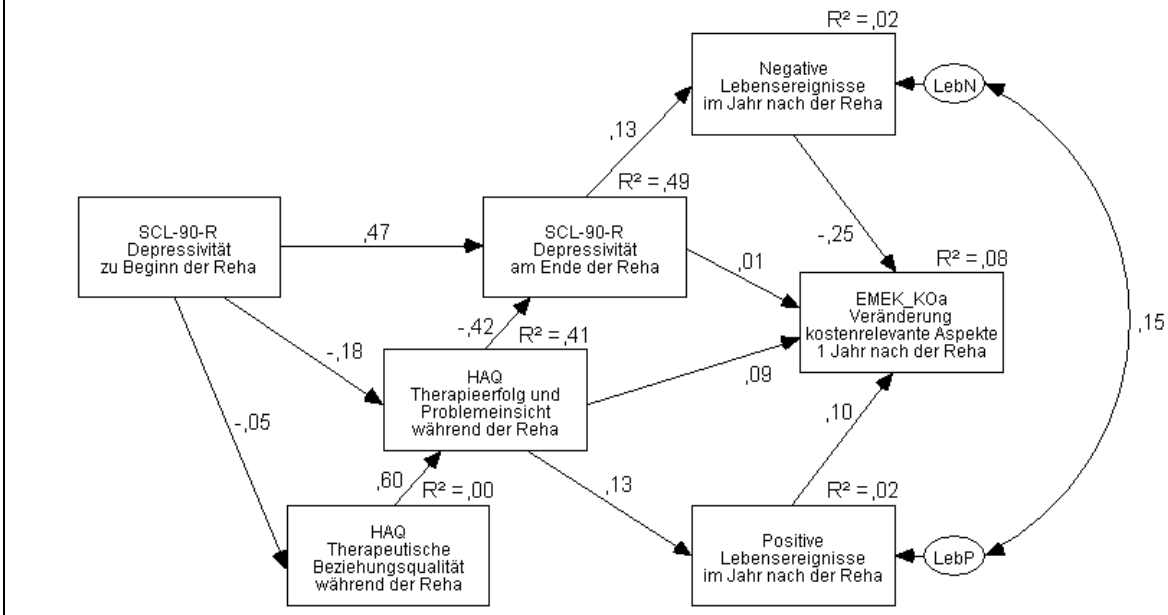
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 21,796; *df* = 9; χ^2 / *df* = 2,422; *p* = .010; R^2 = .204
SRMR = .029; *RMSEA* = .046 (*CI* 90%: .022 bis .071); *CFI* = .987

Abbildung 54. Pfadmodell zur Vorhersage von EMEK_BZb. *N* = 664 Patienten.

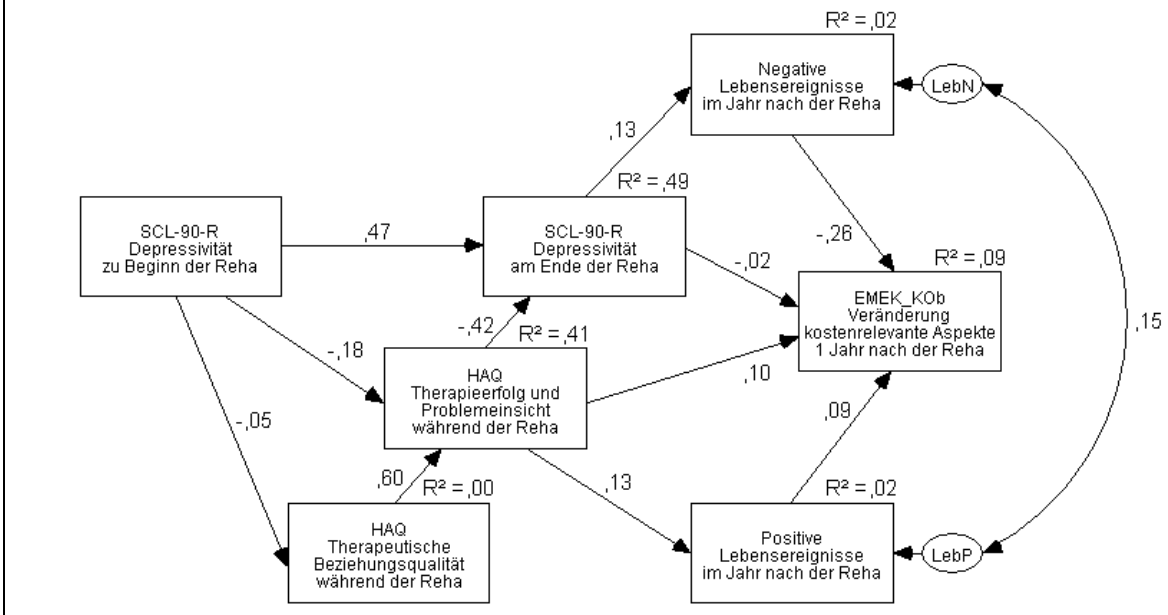
Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 26,622; *df* = 9; χ^2 / *df* = 2,958; *p* = .002; R^2 = .078
SRMR = .030; *RMSEA* = .054 (*CI* 90%: .031 bis .079); *CFI* = .980

Abbildung 55. Pfadmodell zur Vorhersage von EMEK_KOa. *N* = 664 Patienten.

Aufnahme in die Klinik	Stationäre Behandlung	Entlassung aus der Klinik	Zeitraum nach der Behandlung	1-Jahres-Katamnese



χ^2 = 24,666; *df* = 9; χ^2 / *df* = 2,741; *p* = .003; R^2 = .092
SRMR = .029; *RMSEA* = .051 (*CI* 90%: .028 bis .076); *CFI* = .983

Abbildung 56. Pfadmodell zur Vorhersage von EMEK_KOb. *N* = 664 Patienten.