Fachbereich Erziehungswissenschaft und Psychologie

der Freien Universität Berlin

Zur Wissenschaftlichkeit von Psychotherapie

Eine Untersuchung der Verfahrensregeln zur wissenschaftlichen Anerkennung von psychotherapeutischen Verfahren und Methoden anhand von Studien zu psychodynamischen Kurz- und Langzeittherapien

Dissertation

zur Erlangung des akademischen Grades

Doktorin der Philosophie (Dr. phil.)

vorgelegt von

Dipl.-Psych. Melanie Ratzek

Berlin, 2014

Erstgutachter:	Prof. Dr. Dieter Kleiber (Freie Universität Berlin)
Zweitgutachterin:	Prof. Dr. Anna Auckenthaler (Freie Universität Berlin)
Tag der Disputation:	18.12.2014

Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig verfasst und dabei keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Diese Arbeit wurde nicht an anderer Stelle als Abschlussarbeit eingereicht.

Berlin, Oktober 2014

Danksagung

Ein großer Dank geht an Prof. Dr. Dieter Kleiber, der mich über die ganzen letzten Jahre mit Zuversicht und wachem Interesse und mit einem unnachahmlichen Humor begleitet hat. Ebenso danke ich Prof. Dr. Anna Auckenthaler für die vielen ermutigenden Worte in der Anfangsphase der Arbeit und für die Genauigkeit und Klarheit ihrer Rückmeldungen.

Danken möchte ich außerdem Dr. Dr. Burkhard Gusy für die vielen Stunden, in denen wir gemeinsam nachdachten. Ihm habe ich entscheidende Impulse für die vorliegende Arbeit zu verdanken.

Unersetzbar wurde mir die Zusammenarbeit mit Dipl.-Psych. Luisa von Hauenschild. Sie ist und bleibt die beste Kollegin, die ich mir denken kann!

Und auch an drei weitere Projektmitarbeiterinnen und -mitarbeiter geht ein großer Dank: An Dipl.-Psych. Matthias Mohse, an Dipl.-Psych. Uta Czech und an Laura Diedrich (geb. Schweikert) M.Sc., die sich an den Kodierungen der Studien zuverlässig und mit großer Ausdauer beteiligten.

Für seine exzellenten Ratschläge in manchen methodischen Fragen möchte ich Prof. Dr. Michael Eid danken. Gleiches gilt für Dr. Johannes Zimmermann und Dr. Joachim Bretz, von deren Methodenkenntnis ich profitieren durfte.

Außerdem danke ich Dipl.-Psych. Franziska Wörfel und Dipl.-Psych. Sabine Stark, damals noch studentische Hilfskräfte am Institut für Public Health, die mir bei der Beschaffung der Studienpublikationen eine unverzichtbare Hilfe waren. Marian Jäger danke ich für seine Unterstützung beim aufwändigen Einscannen und bei der grafischen Optimierung aller Studienkodierbögen.

Mein Dank geht auch an Dipl.-Psych. Timo Harfst. Ihm habe ich viele gute und aufschlussreiche Ratschläge für die Praxis der Studienkodierung zu verdanken.

Ein riesiger Dank geht an die Wissenschaftliche Summer School der Deutschen Psychoanalytischen Gesellschaft (DPG) unter der Leitung von Prof. Dr. Cord Benecke, Prof. Dr. Hermann Staats, Prof. Dr. Dr. Dorothea Huber und Prof. Dr. Rainer Krause. Die Summer School wurde mir für die Entwicklung und Weiterführung der Arbeit eine Art geistiges Zuhause, das mich ermutigte und antrieb. Vor allem Cord Benecke möchte ich an dieser Stelle danken, der durch den richtigen Kommentar an richtiger Stelle immer wieder meine Gedanken in eine gute Richtung zu lenken vermochte. Durch meine Zusammenarbeit mit Dorothea Huber habe ich in den vergangenen Jahren mehr und mehr die Praxis der Psychotherapieforschung kennenlernen dürfen, was mir den Blick für die vorliegende Arbeit geschärft hat. Der DPG danke ich außerdem für die Drittmittelunterstützung meines Promotionsprojekts.

Ebenso danke ich der Hochschulleitung der International Psychoanalytic University Berlin, das sind Prof. Dr. Martin Teising, Prof. Dr. Lilli Gast und Dr. Rainer Kleinholz, dafür, dass ich die universitäre Infrastruktur zur Fertigstellung bis hin zum Druck der Arbeit nutzen durfte.

Meinem Vater Hans, meinem Großvater Walter und meiner Tante Heike danke ich für die aufmunternden Gesten und Worte in den letzten Jahren.

Ralph, dir danke ich für deinen klugen Geist, mit dem du meine Arbeit Korrektur gelesen hast, für deine Scharfsinnigkeit und dein unkonventionelles, schönes Denken. Und für deine geduldige Begleitung im Schachspiel gegen mich selbst. Antje, ohne deine kreativen Korrekturvorschläge und ohne deine liebevolle Fürsorge, deinen charmanten Witz und unsere sehr enge Freundschaft wären die letzten Monate wahrlich eintönig geworden. Carolina und Magali, einen lieben Dank fürs Redigieren des Abstracts. Sonja und Christiane, Euch danke ich für die aufgeweckten Abende, die für einige Stunden alle Sorgen vergessen ließen. Jens, dir danke ich für alles.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AMQ	Allgemeine methodische Qualität
AP	Analytische Psychotherapie
APA	American Psychological Association
APS	Wirksamkeit Analytischer Psychotherapie und Kognitiver Verhaltenstherapie bei Angst- plus Persönlichkeitsstörung (Studie)
ÄZQ	Ärztliches Zentrum für Qualität in der Medizin
AVM	Arbeitsgemeinschaft für Verhaltensmodifikation
BDI	Beck Depression Inventory
BL	Beschwerdeliste
CA	Clusteranalyse
CGI	Clinical Global Impression
CGI-S	Clinical Global Impression - Severity
CHMP	Committee for Medicinal Products for Human Use
CONTAN	Committee On Test Affairs Netherland
CPPS	Comparative Psychotherapy Process Scale
DFT	Deutsche Fachgesellschaft für tiefenpsychologisch fundierte Psychotherapie
DGPPN	Deutsche Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde
DGPs	Deutsche Gesellschaft für Psychologie
DGPT	Deutsche Gesellschaft für Psychoanalyse, Psychotherapie, Psychosomatik und Tiefenpsychologie
DGVT	Deutschen Gesellschaft für Verhaltenstherapie

DIN	Deutsche Industrienorm
DPG	Deutsche Psychoanalytische Gesellschaft
DSM	Diagnostic and Statistical Manual of Mental Disorders
ECRS-ATR	Ereignis-Context-Relation-Schwere Schema – Adverse Treatment Reaction
EFPA	European Federation of Psychologists Associations
EG	Experimentalgruppe
EPDS	Edinburgh Postnatal Depression Scale
EST	Empirically Supported Treatments
EbM	Evidenzbasierten Medizin
EMDR	Eye-Movement-Desensitization and Reprocessing
EV	Externe Validität
FLZ	Fragebogen zur Lebenszufriedenheit
FPI	Freiburger Persönlichkeitsinventar
GAD	Generalized Anxiety Disorder
GAS	Goal Attainment Scale
GBB	Gießener Beschwerdebogen
GHQ-12	General Health Questionnaire
GRADE	Grades of Recommendation, Assessment, Development, and Evaluation
GT	Gießen Test
HAM-D	Hamilton Depression Scale
HARS	Hamilton Anxiety Rating Scale
ICD	International Statistical Classification of Diseases and Related Health Problems
i.d.R	in der Regel

IG	Interventionsgruppe
IIP	Inventory of Interpersonal Problems
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
IRES	Indikatoren des Reha-Status
ITT	Intent-To-Treat
IV	Interne Validität
KG	Kontrollgruppe
LAC	Langzeittherapien bei chronischen Depressionen
M	Mittelwert
MeSH	Medical Subject Headings
MPS	Münchener Psychotherapiestudie
NCBI	National Center for Biotechnology Information
NICE	National Institute for Health and Care Excellence
NV	Normalverteilung
OPD	Operationalisierte Psychodynamische Diagnostik
PGWBI	Psychological General Well-Being Index
PQS	Psychotherapy-Process Q-Sort
Psa	Psychoanalyse
PsychThG	Psychotherapeutengesetz
QAP	Quasi-Alternativprogramm
RCI	Reliable Change Index
RCT	Randomized Controlled Trial
SAS-SR	Social Adjustment Scale - Self-Report
SCID	Structured Clinical Interview for DSM Disorders
SCL-90-Anx	Symptom Checklist Anxiety Scale

SCL-90-R	Symptom Checklist-90 Revised
SD	Standardabweichung
STPP	Short-term Psychodynamic Psychotherapy
TAU	Treatment-As-Usual
TBS-TK	Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen
TP	Tiefenpsychologisch fundierte Psychotherapie
VEV	Veränderungsfragebogen des Erlebens und Verhaltens
VG	Vergleichsgruppe
WAI	Work Ability Index
WBP	Wissenschaftlicher Beirat Psychotherapie
WL	Warteliste

Kurzfassung

Hintergrund

Die wissenschaftliche Anerkennung von Psychotherapie durch den Wissenschaftlichen Beirat Psychotherapie nach § 11 PsychThG (WBP) basiert maßgeblich auf der Begutachtung psychotherapeutischer Verfahren und Methoden im Hinblick auf ihre empirische Evidenz zur Wirksamkeit. Ziel der Begutachtung ist es zu ermitteln, inwieweit Wirksamkeitsnachweise für ein bestimmtes psychotherapeutisches Verfahren oder eine bestimmte Methode mittels wissenschaftlicher Methoden und Forschungsstrategien erbracht worden sind. Zu diesem Zweck hat der WBP im Jahr 2007 ein Methodenpapier veröffentlicht, in dem er das Prozedere der wissenschaftlichen Anerkennung transparent macht (WBP, 2007, 2010). Einen zentralen Bestandteil dieses Papiers bildet ein ausführlicher Kriterienkatalog, mit Hilfe dessen Wirksamkeitsstudien auf den drei Dimensionen der allgemeinen methodischen Qualität sowie der internen und externen Validität bewertet werden.

Fragestellung

Das Dissertationsprojekt nimmt seinen Ausgangspunkt bei den genannten Bewertungskriterien und betrachtet die damit formulierten Regeln zur wissenschaftlichen Anerkennung von Verfahren und Methoden in Anlehnung an Leichsenring (2008) als ein neu konzipiertes diagnostisches Instrument, das sich nun in seiner Anwendung zu bewähren hat. Es wurde der Frage nachgegangen, wie sich die vom WBP vorgenommene Operationalisierung von interner Validität in Form interner Validitätskriterien auf die Bewertung von existierenden Wirksamkeitsstudien zu psychodynamischen Langzeitbehandlungen auswirkt. Die Frage gewann insofern an Relevanz, als der WBP mit Veröffentlichung des Methodenpapiers die wissenschaftliche Anerkennung der psychodynamischen Psychotherapie auf psychodynamische Langzeit-

therapien (> 100 Stunden) ausweitete (WBP, 2008a). Darüber hinaus werden insbesondere Langzeitbehandlungen als ein Gegenstand diskutiert, der sich, im Gegensatz zu Kurzzeitbehandlungen, der RCT-Methodologie und den damit verbundenen Strategien zur Sicherung interner Validität weitestgehend entzieht. Es war daher davon auszugehen, dass sich interne Validitätskriterien, die sich strikt an die RCT-Methodologie anlehnen, in der Bewertung benachteiligend auf Langzeittherapiestudien auswirken würden. Die Fragestellung der Arbeit bezieht sich damit auf die Untersuchung der Gegenstandsadäquatheit der internen Validitätskriterien im Hinblick auf Studien zu psychodynamischen Langzeitbehandlungen (> 100 Stunden).

Methode

Mittels einer erschöpfenden Studienrecherche wurden für den Zeitraum von 1999 bis 2009 alle verfügbaren Wirksamkeitsstudien zur psychodynamischen Psychotherapie (Erwachsenenalter) in den Anwendungsbereichen der affektiven Störungen sowie der „gemischten Störungen“ zusammengestellt. Diese Studien wurden sodann mit Hilfe des WBP-Kriterienkatalogs kodiert. Um einen ersten Eindruck darüber zu erhalten, welche Kriterien sich potentiell benachteiligend im Hinblick auf Langzeittherapiestudien auswirken könnten, erfolgte ein systematischer Vergleich der Studien zu Langzeittherapien und zu Kurzzeittherapien in ihrem Abschneiden auf den einzelnen Kriterienstufen (1 = „Anforderung erfüllt“ bis 3 = „Anforderung nicht erfüllt“). Für diese Verteilungsvergleiche wurden die Kriterien der internen Validität sowie zentrale Kriterien der methodischen Qualität herangezogen. Kriterien, bei denen anteilig mehr Langzeittherapiestudien „schlecht“ bewertet wurden als Kurzzeittherapiestudien, wurden darauf folgend einer feinanalytischen Untersuchung unterzogen. In dieser wurden die Gründe für ein gefundenes Bewertungsgefälle durch Rückgriff und erneute Sichtung der Studien eingehend eruiert. Vor dem Hintergrund allgemeiner Erkenntnisse aus

der Evaluationsforschung sowie der empirischen Psychotherapieforschung wurden diese Gründe reflektiert. Dieser Schritt diente der Einschätzung, ob die Gründe tatsächlich mit dem Gegenstand "Langzeittherapie" zusammenhängen oder als unabhängig davon betrachtet werden müssen. Nur so konnte beurteilt werden, ob von einem Kriterium tatsächlich eine benachteiligende Wirkung ausgeht oder das Bewertungsgefälle den Studien selbst anzulasten ist.

Ergebnisse

Auf beiden Dimensionen, der methodischen Qualitätsdimension und der internen Validitätsdimension, zeigte sich nur bei vergleichsweise wenigen Kriterien ein bedeutendes Bewertungsgefälle zuungunsten der kodierten Langzeittherapiestudien im Vergleich zu Kurzzeittherapiestudien. Durch die an diese Vergleiche sich anschließenden Feinanalysen konnte darüber hinaus bei einem Teil dieser Kriterien der Verdacht einer benachteiligenden Wirkung hinsichtlich Langzeittherapiestudien ausgeräumt werden. Lediglich drei der 12 internen Validitätskriterien konnte auf Basis der hier durchgeführten Untersuchung eine tatsächlich benachteiligende Wirkung attestiert werden. Ferner zeigte sich jedoch, dass dem Kriterienkatalog solche Kriterien fehlen, die die interne Validität von Studien ohne Kontroll- bzw. Vergleichsgruppe sowie von komparativen Studien (Vergleiche zwischen verschiedenen Alternativtherapien) beurteilen.

Schlussfolgerungen

Sowohl die Dimension des WBP-Kriterienkatalogs zur allgemeinen methodischen Qualität als auch die zur internen Validität sind für die Bewertung der Evidenz von psychodynamischer Psychotherapie als durchaus angemessen anzusehen. Auch im Hinblick auf Untersuchungen zu Langzeitbehandlungen (> 100 Stunden) stellen die meisten Kriterien keine unzumutbaren Hürden dar. Ausgehend vom Konzept eines komplementären – statt eines inversen – Verhält-

nisses von interner und externer Validität werden Techniken zur internen Validitätssicherung benannt, die sich insbesondere für Studiendesigns ohne Kontroll-/Vergleichsgruppen und für komparative Studien eignen. Diese Techniken zu berücksichtigen und in Form von Kriterien dem Kriterienkatalog hinzuzufügen, könnte u.a. einen positiven Anreiz für die Planung und Durchführung künftiger Wirksamkeitsstudien setzen.

Abstract

Background

The scientific recognition of psychotherapy by the *German Advisory Board for Psychotherapy* is mainly based on the evaluation of the empirical evidence of psychotherapeutic treatments. The aim of the scientific recognition of psychotherapy is to examine to which extend efficacy and effectiveness studies are based on scientific methods and research strategies. For this purpose, the *German Advisory Board* published the *new guidelines for scientific evaluation and scientific recognition of psychotherapy* (WBP, 2007, 2010). A key component of these guidelines is a catalogue of criteria, based on three dimensions, which evaluates the general scientific quality, the internal and the external validity of efficacy and effectiveness studies.

Purpose

The aim of this study is to investigate the criteria of the internal validity dimension and of the general scientific quality dimension concerning their appropriateness to evaluate studies of long-term psychoanalytic treatments. In this regard, the criteria catalogue is viewed as an assessment measure, which has to prove itself in its practical application (cf. Leichsenring, 2008). This approach is relevant in the context of the modification of the statement concerning the scientific status of psychodynamic and psychoanalytic psychotherapy: In 2008 the *Scientific Advisory Board* extended the scientific recognition of psychodynamic/psychoanalytic psychotherapy by also including psychoanalytic long-term treatments (> 100 sessions) (WBP, 2008a). This modification is justified by the development and publication of the *new guidelines* mentioned above. Concurrently, long-term treatments as research object, in contrast to short-term treatments, are considered difficult to evaluate by principles of the RCT-methodology. Therefore, it is to assume that criteria of the internal validity di-

mension, which are based solely on the principles of the RCT-methodology, would be neither applicable nor appropriate to long-term psychotherapy studies. In that case the criteria would cause a discriminatory evaluation of these studies compared to short-term psychotherapy studies.

Method

First, an exhaustive review of the literature was done to collect all empirical evidence available for efficacy and effectiveness of psychodynamic and psychoanalytic psychotherapy published between 1999 and 2009. The collected studies had to include adult clients with mood disorders or mixed samples of clients. In the second step, the evaluation of these studies was conducted by using the criteria catalogue. To get an overall idea about the criteria, which might not be appropriate to evaluate long-term psychotherapy studies, long-term psychotherapy studies and short-term psychotherapy studies were systematically compared on each criterion (1 = "sufficient" to 3 = "insufficient"). When distributions showed more long-term psychotherapy studies with insufficient results in comparison to short-term psychotherapy studies, this was regarded as a preliminary indication of a potential discriminatory impact of the specified criterion. For those criteria a more refined analysis was performed to determine the exact reasons responsible for the unequal distributions regarding the long-term studies. This refined analysis contained both: 1) an intensive review of the long-term psychotherapy studies with poor results and 2) a review of the literature on evaluation research in general and psychotherapy research in particular. This procedure helps to identify whether "long-term psychotherapy" as research object caused the poor results on the criteria – in that case, the found criteria had to be regarded as discriminatory criteria concerning long-term psychotherapy studies. Otherwise, if it could be shown that the reasons responsible for the poor results do not

plausibly interrelate with "long-term psychotherapy" as research object, the criteria were not regarded as discriminatory concerning long-term psychotherapy studies.

Findings

Both, the analyzed scientific quality criteria and the internal validity criteria reveal relatively few unequal distributions of the studies to the disadvantage of the long-term psychotherapy studies which required subsequent refined analyses. Moreover, the conducted refined analyses showed that just few of the criteria effectively have discriminatory impact. In total, three of the 12 criteria of the internal validity dimension could be disclosed as discriminatory concerning long-term psychotherapy studies. Further along it could be shown that the criteria catalogue lacks adequate criteria measuring internal validity of studies without control condition (one-group pretest posttest design) and also of comparative outcome studies comparing two (or more) alternative psychotherapy conditions.

Conclusions

The *Scientific Advisory Board* has been mainly successful in developing criteria that are adequate to measure the scientific quality and the internal validity of outcome studies investigating the efficacy or effectiveness of psychodynamic and psychoanalytic psychotherapy. This is also true for psychoanalytic long-term psychotherapy (> 100 sessions). In this study a concept that emphasizes the complementary relationship of the internal and the external validity rather than an inverse relationship is assumed. Deriving from this concept, strategies are discussed that are suitable to assure the internal validity in both, one-group pretest posttest designs and comparative outcome studies. The integration of internal validity criteria concerning the last mentioned study designs could be a positive incentive for future implementations of efficacy and effectiveness studies.

1 Einleitung

Die Psychotherapie sieht sich seit geraumer Zeit der Verpflichtung gegenüber, ihre Wirksamkeit in Form robuster wissenschaftlicher Befunde gegenüber der interessierten Öffentlichkeit sowie den Solidargemeinschaften zu belegen. Dies gilt auch für den Bereich der analytischen Langzeittherapie (> 100 Behandlungsstunden). Diese nimmt derzeit zwar einen verhältnismäßig geringen Anteil der psychotherapeutischen Versorgung in Deutschland ein: So konnte die Kassenärztliche Bundesvereinigung anhand einer groß angelegten Kohortenstudie zeigen, dass ein Patientenanteil von lediglich 2.4% der insgesamt 385.885 einbezogenen Personen eine analytische Psychotherapie im Erhebungsjahr 2009 begonnen hat (vgl. Multmeier, 2014). Ferner konnte zu einem weiteren Erhebungszeitpunkt im Jahr 2012 gezeigt werden, dass mit einer medianen Therapielänge von 89 Sitzungen sowie 39 Sitzungen für das untere und 160 Sitzungen für das obere Quartil ein Großteil der analytischen Behandlungen ein Kontingent von höchstens 160 Sitzungen umfasst. Die reguläre Behandlungshöchstdauer von 300 Stunden in der analytischen Psychotherapie wird hingegen nur noch von etwas mehr als 1% der analytischen Psychotherapiepatienten ausgeschöpft. Dennoch stellt die analytische Langzeittherapie im Vergleich zu den kürzeren Behandlungsformen die zeitaufwändigste und – bezogen auf den reinen Behandlungsaufwand – die kostenintensivste Therapieform dar (vgl. Multmeier, 2014). Aus diesem Grund sieht sich die analytische Langzeittherapie einem besonderen Rechtfertigungsdruck ausgesetzt, dem sie seit einiger Zeit verstärkt mit der Durchführung von Wirksamkeitsstudien zu begegnen versucht. Ein Grund für die sich erst seit wenigen Jahren verdichtende Evidenzlage zur Wirksamkeit analytischer Langzeittherapie ist sicherlich darin zu sehen, dass sich in der psychoanalytischen „community“ zunächst ein gewisser Widerstand formierte. Dieser richtete sich insbesondere gegen eine blinde Übernahme

der Evidenzbasierungsmethoden der Medizin und der Pharmazie auf den Gegenstand der analytischen Langzeittherapie.

In Deutschland erfuhr die Debatte um psychotherapieadäquate Evidenzbasierungsmethoden, die längst nicht nur innerhalb der psychoanalytischen Provenienz und im Hinblick auf Langzeitbehandlungen geführt wurde, im Jahr 1999 einen Aufwind. In diesem Jahr trat das Psychotherapeutengesetz in Kraft und es bildete sich ein Gutachtergremium, der Wissenschaftliche Beirat Psychotherapie (WBP), für den fortan die empirische Evidenz zur Wirksamkeit psychotherapeutischer Verfahren und Methoden von besonderem Interesse war. Seitdem besteht seine Aufgabe darin, die Evidenz zur Wirksamkeit von Psychotherapieverfahren zwecks berufsrechtlicher und wissenschaftlicher Anerkennung zu begutachten und für „gut" oder eben „unzulänglich" zu befinden (dazu ausführlich in Kap 1.1). Insofern kommt dem WBP die verantwortungsvolle Aufgabe zu, diese Begutachtung unter Berücksichtigung der besonderen Erfordernisse des psychotherapeutischen Gegenstands vorzunehmen. Das gilt insbesondere für die analytische Langzeittherapie – dem „Heiligen Gral der psychoanalytischen Kliniker" (Fonagy, 2009, deutsche Übersetzung S. 13) –, deren Evidenzbasierung vielen Forschern[1] und auch dem WBP (2008a) zufolge ein besonderes Unterfangen darstellt. Denn gegenüber allen anderen therapeutischen Verfahren, die in der Regel von kürzerer Dauer sind, entzöge sich die Beforschung der Langzeittherapie – so der Tenor – weitestgehend der Methodologie der Evidenzbasierten Medizin (EbM) (vgl. Kap. 1.2.1).

Der Frage, inwieweit der WBP die ihm zugetragene Aufgabe der Begutachtung der empirischen Evidenz von Psychotherapie mittels gegenstandsangemessener Kriterien zu begegnen vermag und inwieweit ihm dies selbst im Hinblick auf den schwierigen Bereich der analytischen Langzeitbehandlungen gelingt, soll in dieser Arbeit nachgegangen werden. Da-

[1] Der besseren Lesbarkeit zuliebe wird auf die gleichzeitige Verwendung weiblicher und männlicher Sprachformen verzichtet. Sämtliche Personenbezeichnungen gelten jedoch für alle Geschlechter.

bei sollen in Abgrenzung zu bisherigen, eher theoretisch geführten Debatten zu diesem Themenbereich explizit Wirksamkeitsstudien herangezogen werden, die im Zeitraum von 1999 bis 2009 publiziert wurden. Hierbei wird sich auf die psychodynamischen Verfahren begrenzt. Dadurch soll ein empiriebasierter Eindruck darüber geliefert werden, wie tiefenpsychologisch fundierte und analytische Psychotherapie tatsächlich hinsichtlich ihrer Wirksamkeit beforscht werden und ob der WBP diese Form der Evidenz in adäquater Form „einzufangen“ und abzubilden weiß.

Die Darlegung des theoretischen Hintergrunds sowie die Herleitung der Fragestellung der Arbeit erfolgt in zwei Kapiteln: Zunächst wird der WBP und sein Aufgabenbereich der berufsrechtlichen in Abgrenzung zur sozialrechtlichen Anerkennung durch den Gemeinsamen Bundesausschuss genauer dargestellt (Kap. 1.1). Im Anschluss folgt eine Beschreibung der aktuellen Verfahrensregeln des WBP (Version 2.8; 2010), in denen er sein Vorgehen im Prozedere der berufsrechtlichen bzw. wissenschaftlichen Anerkennung detailliert niedergelegt hat (Kap. 1.2). Hier wird es u.a. um eine Annäherung an zentrale Begriffe wie die der *internen Validität* und der *externen Validität* von Untersuchungen gehen.

Die Entwicklungsgeschichte der Evidenzbasierung psychotherapeutischer Verfahren, die den Verfahrensregeln des WBP vorausging, wird in Kapitel 1.2.1 näher beschrieben. Dabei wird auch ein eingehender Blick in den amerikanischen Sprachraum und auf die Erstellung von Listen sog. *empirically supported treatments* (ESTs) geworfen. Zudem wird die national sowie international geführte Debatte über psychotherapieadäquate Evidenzbasierungsmethoden umrissen. Im darauf folgenden Kapitel (Kap. 1.2.2) wird schließlich der aus drei Dimensionen bestehende WBP-Kriterienkatalog – der zentrale Untersuchungsgegenstand der vorliegenden Arbeit – vorgestellt, indem jedes einzelne der 44 Kriterien kurz charakterisiert

und im Hinblick auf seine Dimensionszugehörigkeit eingeordnet wird. Ein abschließendes Fazit leitet sodann über zu Kapitel 2, in dem die genaue Fragestellung formuliert wird.

1.1 Der Wissenschaftliche Beirat Psychotherapie nach § 11 PsychThG und seine Funktion in der berufsrechtlichen Anerkennung von Psychotherapie

Durch das Psychotherapeutengesetz wurde der neue Berufsstand des Psychologischen Psychotherapeuten und des Kinder- und Jugendlichenpsychotherapeuten in die kassenärztlichen Vereinigungen integriert und damit die psychotherapeutische in die kassenärztliche Versorgung. Das Psychotherapeutengesetzt legt in § 2 die Approbation zum Psychologischen Psychotherapeuten fest, die die Voraussetzung für die Berufsausübung bildet; diese kann nur durch Ausbildung an staatlich anerkannten Ausbildungsinstituten erlangt werden. Die staatliche Anerkennung solcher Institute hängt (nach § 6 des PsychThG) wiederum u. a. davon ab, ob an ihnen vertiefte Ausbildungsgänge in *wissenschaftlich anerkannten* Verfahren erfolgen:

> Einrichtungen sind als Ausbildungsstätten nach Absatz 1 anzuerkennen, wenn in ihnen
> 1. Patienten, die an psychischen Störungen mit Krankheitswert leiden, nach wissenschaftlich anerkannten psychotherapeutischen Verfahren stationär oder ambulant behandelt werden . . . . (PsychThG § 6 Absatz 2)

Dem WBP obliegt die im Zitat erwähnte wissenschaftliche Anerkennung von psychotherapeutischen Verfahren (vgl. § 11 des PsychThG). Die wissenschaftliche Anerkennung durch den WBP beruht in erster Linie auf der Sichtung und Begutachtung der empirischen Evidenz (Wirksamkeitsstudien) einzelner psychotherapeutischer Verfahren. Auf die genaue Vorgehensweise dieses Prozederes wird in Kapitel 1.2 detailliert eingegangen.

Die staatliche Anerkennung von Ausbildungsstätten erfolgt über die zuständigen Landesbehörden, die sich in gutachterlichen Fragen bzgl. der wissenschaftlichen Anerkennung

von Verfahren an den WBP wenden und von diesem beraten werden. Ebenso wenden sich psychotherapeutische Fachverbände an den WBP, um die berufsrechtliche Anerkennung bestimmter psychotherapeutischer Verfahren oder Methoden zu beantragen, um sie sodann als Vertiefungsgebiet im Rahmen der Ausbildung mit Approbationsmöglichkeit anbieten zu können (zur Abgrenzung zwischen *Verfahren* und *Methoden* vgl. Kap. 1.2). Da es sich bei dieser Berater- und Gutachtertätigkeit um die Beantwortung von Fragen handelt, die eine psychotherapeutische, methodische und evaluatorische Expertise verlangt, setzt sich der WBP hauptsächlich aus habilitierten Psychologen und Ärzten zusammen. Dabei entsenden sowohl die Bundespsychotherapeutenkammer als auch die Bundeärztekammer jeweils sechs Vertreter, so dass der Beirat sich paritätisch aus sechs Psychologischen Psychotherapeuten und Kinder- und Jugendlichenpsychotherapeuten einerseits und sechs ärztlichen Vertretern zusammensetzt. Letztere stammen aus den Bereichen „Psychiatrie und Psychotherapie", „Psychosomatische Medizin und Psychotherapie" sowie „Kinder- und Jugendlichenpsychiatrie und –psychotherapie" (vgl. Auckenthaler, 2012, S. 105).

Grundsätzlich eröffnet die Approbation in einem durch den WBP wissenschaftlich anerkannten Verfahren nur dann den Weg in die kassenärztliche Versorgung, wenn es sich bei dem Verfahren auch um ein sozialrechtlich anerkanntes Verfahren handelt. Zu diesem zählen zum aktuellen Zeitpunkt nur die psychoanalytisch begründeten Verfahren[2] (tiefenpsychologisch fundierte und analytische Psychotherapie) sowie die Verhaltenstherapie. Die sozialrechtliche Anerkennung fällt in den Verantwortungsbereich des Gemeinsamen Bundesausschusses. Dessen Unterausschuss Psychotherapie sowie das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) prüfen die psychotherapeutischen Maßnahmen,

[2] In der vorliegenden Arbeit werden die Bezeichnungen „psychoanalytisch begründete Verfahren" (vgl. Gemeinsamer Bundesausschuss, 2013) und „psychodynamische Psychotherapie" (vgl. WBP, 2005) synonym verwendet.

die in den Leistungskatalog der gesetzlichen Krankenkassen aufgenommen werden (oder in ihm bestehen bleiben) sollen im Hinblick auf ihren indikationsbezogenen Nutzen, ihre medizinische Notwendigkeit und auf ihre Wirtschaftlichkeit. Damit kommt ein Kriterium ins Spiel, das bei der Beurteilung von Verfahren durch den WBP irrelevant ist: die Wirtschaftlichkeit. So ist es durchaus möglich, dass ein Verfahren zwar berufsrechtlich, jedoch nicht sozialrechtlich anerkannt wird, wie es seit bereits 12 Jahren bei der Gesprächspsychotherapie und seit 2008 auch bei der systemischen Therapie der Fall ist. Auf diese Weise existiert eine Art Verklammerung zwischen dem WBP und dem Gemeinsamen Bundesausschuss, die in den Psychotherapierichtlinien folgendermaßen zum Ausdruck kommt:

> Über die in § 13 genannten Verfahren hinaus können als Psychotherapie gemäß Abschnitt A der Richtlinie in der vertragsärztlichen Versorgung andere Verfahren Anwendung finden, wenn nachgewiesen ist, dass sie die nachstehenden Voraussetzungen nach Nummer1 bis 3 erfüllen:
> 1. [1]Feststellung durch den wissenschaftlichen Beirat gemäß § 11 des Psychotherapeutengesetzes, dass das Verfahren als wissenschaftlich anerkannt für eine vertiefte Ausbildung zur Psychologischen Psychotherapeutin oder zum Psychologischen Psychotherapeuten oder zur Kinder-und Jugendlichenpsychotherapeutin oder zum Kinder-und Jugendlichenpsychotherapeuten angesehen werden kann. (Gemeinsamer Bundesausschuss, § 17 Absatz 1)

In § 17 Absatz 3 heißt es zwar, dass in begründeten Ausnahmefällen von der Voraussetzung der wissenschaftlichen Anerkennung durch den WBP abgewichen werden kann, in der Regel fungiert der WBP und seine Stellungnahme zu einem psychotherapeutischen Verfahren jedoch als eine notwendige Bedingung für die sozialrechtliche Anerkennung. Dies gilt umso mehr, seitdem das Methodenpapier des WBP (2010) die Grundlage der berufsrechtlichen Anerkennung bildet (vgl. Kap. 1.2.2): Das Methodenpapier und insbesondere der darin enthaltene Kriterienkatalog zur standardisierten Begutachtung von Wirksamkeitsstudien wurde in enger Zusammenarbeit mit dem Gemeinsamen Bundeausschuss entwickelt. Damit greifen

beide Gremien, der WBP und der Gemeinsame Bundesausschuss, was die Wirksamkeits- bzw. Nutzenbewertung psychotherapeutischer Maßnahmen betrifft, auf denselben Maßstab in Form der Bewertungskriterien zurück – auch, wenn es nach Aussage von Schulte „dabei im Detail Unterschiede gibt“ (Deutsches Ärzteblatt, 2008a, S. 390).

Der Fokus der vorliegenden Arbeit liegt allein auf dem Prozedere der berufsrechtlichen respektive wissenschaftlichen Anerkennung durch den WBP. Bis zum heutigen Zeitpunkt (08/2014) wurden folgende psychotherapeutische Verfahren durch den WBP wissenschaftlich anerkannt: Verhaltenstherapie (Kinder- und Jugendalter sowie Erwachsenenalter; WBP, 2004a), psychodynamische Psychotherapie (Erwachsenenalter; WBP, 2005), Gesprächspsychotherapie (Erwachsenenalter; WBP, 2002a) und systemische Therapie (Kinder- und Jugendalter sowie Erwachsenenalter; WBP, 2009a). Als psychotherapeutische Methode wurde bislang die Eye-Movement-Desensitization and Reprocessing Methode (EMDR-Methode) für den Indikationsbereich der posttraumatischen Belastungsstörung bei Erwachsenen wissenschaftlich anerkannt (WBP, 2006a); die wissenschaftliche Anerkennung der neuropsychologischen Therapie als *Verfahren* wurde im Jahr 2008 korrigiert – seitdem ist diese Behandlungsform nunmehr als *Methode* für den Anwendungsbereich der hirnorganischen Störungen (Erwachsenenalter) anerkannt (WBP, 2008b). Die interpersonelle Psychotherapie ist für die Anwendungsbereiche der affektiven Störungen und der Essstörungen (Erwachsenenalter) als Methode wissenschaftlich anerkannt (WBP, 2006b), die Hypnotherapie für den Indikationsbereich der psychischen und sozialen Faktoren bei somatischen Krankheiten sowie für Abhängigkeit und Missbrauch (Erwachsenenalter) (WBP, 2006c). Zu den beiden Verfahren „Verhaltenstherapie“ und „psychodynamische Psychotherapie“ gab der WBP lediglich *gutachterliche Stellungnahmen* (statt vollwertige *Gutachten*) ab, da diese Verfahren aufgrund ihres Status’ als Richtlinienverfahren keiner Prüfung durch den WBP mehr bedurften.

Insgesamt erfolgten die aufgezählten Anerkennungsverfahren nach Regularien, die heute keine Geltung mehr beanspruchen. Diese Regularien umfassen sog. Mindestanforderungen für die Bewertung von Wirksamkeitsstudien (WBP, 2004b), zwei getrennte Auflistungen von Anwendungsbereichen (Indikationsbereiche) für den Kinder- und Jugendbereich sowie für den Erwachsenenbereich (WBP, 2000a/b; 2002b), denen zu entnehmen ist, welche Indikationsbereiche durch Wirksamkeitsstudien abgedeckt werden sollten. Außerdem existierte seit 1999 ein Leitfaden für die Erstellung von Gutachtenanträgen, dem zu entnehmen ist, welche Materialien beim WBP zunächst eingereicht werden müssen, bevor er gutachterlich tätig wird (vgl. WBP, 1999).

Die genannten Papiere wurden mit Veröffentlichung der neuen Verfahrensregeln des WBP (Version 2.6) im Jahre 2007 außer Kraft gesetzt (vgl. Deutsches Ärzteblatt, 2008b). Gleichsam erfolgte die genannte Korrektur in Bezug auf die neuropsychologische Therapie u.a. aufgrund der neuen Verfahrensregeln des WBP. Eine weitere Korrektur anlässlich der neuen Verfahrensregeln wurde für die psychodynamischen Verfahren vorgenommen, indem die Anerkennung nunmehr auch auf psychodynamische Langzeittherapien mit einer Dauer über 100 Stunden ausgeweitet wurde (WBP, 2008a) – eine Korrektur, auf die noch näher einzugehen sein wird (vgl. Kap. 2).

Nach aktuellem Kenntnisstand liegen jedoch noch keine Gutachten seitens des WBP vor, die auf der Basis einer systematischen Anwendung der neuen Verfahrensregeln erstellt wurden, jedoch werden die derzeit (08/2014) laufenden Gutachtenverfahren zur wissenschaftlichen Anerkennung der EMDR-Methode für das Kinder- und Jugendalter sowie der humanistischen Psychotherapie unter Anwendung der neuen Verfahrensregeln erfolgen. Mittlerweile wurden die Verfahrensregeln mehrfach modifiziert (vgl. Version 2.7; WBP, 2009b), die aktuelle Ver-

sion (2.8) wurde im Jahr 2010 veröffentlicht und stellt folglich die Fassung dar, die gegenwärtig zur Anwendung kommt. Im nun folgenden Kapitel werden die Verfahrensregeln ausführlicher dargestellt.

1.2 Das Methodenpapier des Wissenschaftlichen Beirats Psychotherapie

Was bedeutet "wissenschaftliche Anerkennung" von Psychotherapie? Der WBP definiert dies folgendermaßen:

> . . . der Wissenschaftliche Beirat Psychotherapie [geht] davon aus, dass **die wissenschaftliche Anerkennung eines Psychotherapieverfahrens dann festzustellen ist, wenn es sich aus wissenschaftlicher Sicht um ein Psychotherapieverfahren handelt, dessen Durchführung in der Praxis zur Heilung oder Linderung von Störungen mit Krankheitswert führt.** (Hervorhebungen im Original, 2010, S. 5)

Als Basis der wissenschaftlichen Anerkennung fungieren in erster Linie Untersuchungen, denen die Wirksamkeit psychotherapeutischer Maßnahmen zu entnehmen sind. Dabei unterliegen die sog. „psychotherapeutischen Maßnahmen“ einer Art Dreiteilung: Der WBP unterscheidet psychotherapeutische Techniken, Methoden und Verfahren, wobei Verfahren die übergeordnete Kategorie darstellen. Als Verfahren sind bspw. die Verhaltenstherapie, die psychodynamische Psychotherapie, die Gesprächspsychotherapie oder die systemische Therapie zu bezeichnen. Diesen ist gemeinsam, dass sie über umfassende Theorien zur Ätiologie und Kuration von psychischen Störungen verfügen und die daraus ableitbaren Behandlungsstrategien sich auf eine Bandbreite von unterschiedlichen Störungsbildern anwenden lassen. So hält ein psychotherapeutisches Verfahren Konzepte zur Indikationsstellung bereit sowie zur Behandlungsplanung und zur therapeutischen Beziehungsgestaltung (vgl. WBP, 2010). Eine psychotherapeutische Methode erfährt ihre definitorische Engführung dadurch, dass sie

sich konzeptionell und in ihren Indikationen nicht mehr auf ein ganzes Spektrum an psychischen Störungen zu beziehen hat, sondern allein auf eine oder ggf. mehrere Störungen. Als Beispiel sei die EMDR-Methode anzuführen (s.o.), die mentalisierungsbasierte Psychotherapie für Borderline-Persönlichkeitsstörungen nach Bateman und Fonagy (2008), die übertragungsfokussierte Psychotherapie für strukturelle Störungen nach Clarkin, Yeomans und Kernberg (2006) oder die *brief supportive-expressive psychodynamic psychotherapy for gerealized anxiety disorders* nach Crits-Christoph, Wolf-Palacio, Ficher und Rudick (1995). Letzteren Methoden ist die Wurzel in den psychoanalytisch begründeten Verfahren gemeinsam, so dass diese in einem wissenschaftlichen Anerkennungsprozedere diesen zugerechnet werden würden. Psychotherapeutische Techniken bezeichnen konkrete Vorgehensweisen, die im Rahmen von Verfahren oder Methoden zur Anwendung kommen, der WBP (2010) führt beispielhaft die Übertragungsdeutung oder die Reizkonfrontation in vivo an. Weitere Techniken wären psychoedukative Interventionen, Achtsamkeitsübungen, szenische Darstellungen in der Psychodramatherapie, das Klarifizieren von Affekten in der analytischen Therapie, Imaginationsübungen etc..

Die wissenschaftliche Anerkennung durch den WBP bezieht sich allein auf empirische Untersuchungen, denen die Wirksamkeit von *Methoden* oder *Verfahren* zu entnehmen ist. Dazu führt der WBP in seinen Verfahrensregeln (2010) *vier grundlegende Kriterien* auf, die näher definieren, wie solche empirischen Wirksamkeitsstudien gestaltet sein sollten. Da für die vorliegende Arbeit allein gruppenstatistische Wirksamkeitsuntersuchungen und keine Einzelfalluntersuchungen, die der WBP seit 2010 ebenfalls zwecks wissenschaftlicher Anerkennung berücksichtigt, von Belang sind, wird im Folgenden dementsprechend nur auf diejenigen Studienanforderungen Bezug genommen, die sich auf Gruppenuntersuchungen beziehen.

Mit dem übergeordneten *Kriterium 1* fordert der WBP die Durchführung von Wirksamkeitsuntersuchungen an Personen, „die unter einer Störung mit Krankheitswert leiden, und

[bei denen] der beobachtete therapeutische Effekt . . . eine Heilung oder Linderung dieser Störung dar[stellt]“ (WBP, 2010, S. 6). Dieses Kriterium leitet sich unmittelbar aus dem Psychotherapeutengesetz ab, in dem die Ausübung von Psychotherapie sich ausschließlich auf krankheitswertige Störungen und nicht etwa auf Gegenstände, wie „die Aufarbeitung und Überwindung sozialer Konflikte“ (PsychThG § 1 Absatz 3) beziehen darf. Zudem sollen die zu begutachtenden Untersuchungen Aufschluss darüber geben, ob und inwieweit das Ziel der Heilung oder Linderung einer Störung erreicht wurde; das bedeutet, die Untersuchungen sollten derart angelegt sein, dass Aussagen zu Veränderungen im Verlauf einer Behandlung im Hinblick auf solche Zielkriterien (Outcomemaße) gemacht werden können, an denen sich Linderung oder Heilung einer Störung auch tatsächlich ablesen lassen (bspw. Symptomschwere, Strukturniveau, interpersonelles Verhalten, Arbeitsfähigkeit, Lebensqualität, Inanspruchnahmeverhalten).

Kriterium 2 verlangt die objektive und replizierbare Feststellung von Behandlungseffekten. Demnach sollten in Bezug auf die Studienpatientengruppe objektive, reliable und valide Diagnosen gestellt sowie ebensolche Zielkriterien erhoben worden sein. Diese Gütemerkmale wurden vom WBP systematisch in Form eines umfassenden Kriterienkatalogs zusammengestellt, der im Anhang des Methodenpapiers zu finden ist (vgl. Anhang A in dieser Arbeit). Dieser Kriterienkatalog basiert auf insgesamt drei Dimensionen, die zuvor genannten Gütemerkmale gehören der Dimension der *allgemeinen methodischen Qualität* an und werden in Kapitel 1.2.2 noch näher beschrieben.

Kriterium 3 bezieht sich auf die zweite Dimension des erwähnten Kriterienkatalogs – der Dimension der *internen Validität*. Unter der internen Validität einer Untersuchung wird gemeinhin Folgendes verstanden:

> We use the term *internal validity* to refer to inferences about whether observed covariation between A and B reflects a causal relationship from A to B in the form in which the variables were manipulated or measured. To support such an inference, the researcher must show that A preceded B in time, that A covaries with B . . . and that no other explanations for the relationship are plausible. (Hervorhebungen im Original, Shadish, Cook & Campbell, 2002, S. 53)

Das Hauptbemühen in der Konzipierung einer intern validen Untersuchung ist daher das Antizipieren und infolgedessen die Kontrolle systematischer Störvariablen. Systematische Störvariablen sind solche, die mit der unabhängigen Variablen kovariieren und sich auf die abhängige Variable auswirken. Ist dies der Fall, könnte eine plausible Erklärung für einen beobachteten Effekt – alternativ zur unabhängigen Variablen – eben eine solche systematische Störvariable sein. Mit dem übergeordneten *Kriterium 3* des Methodenpapiers wird daher gefordert, dass der beobachtete Effekt mit hoher Wahrscheinlichkeit auf die psychotherapeutische Maßnahme und nicht auf systematische Störvariablen (etwa Spontanremission o.ä.) zurückzuführen ist. Um diese zu kontrollieren, werden in den Untersuchungen Vergleichsgruppen gefordert, in denen Interventionen realisiert werden, denen die zu überprüfende Therapie sich als überlegen erweisen soll. Hier kommen Vergleichsbedingungen wie unbehandelte Kontrollgruppen oder Wartelisten, *Treatment-As-Usual* (TAU) oder sog. Placebobehandlungen in Frage (vgl. Kap. 3.1.1). Alternativ kann auch eine Vergleichsgruppe mit einem bereits etablierten Verfahren eingesetzt werden, also einem Verfahren, das seine Wirksamkeit bereits unter Beweis gestellt hat. Hier sollte die zu überprüfende Behandlung sich nicht als bedeutsam unterlegen erweisen (vgl. Kap. 3.2.1).

Auf weitere Strategien der internen Validitätssicherung, die laut WBP entsprechend des Kriterienkatalogs erfüllt sein sollten, wird in Kapitel 1.2.2 eingegangen.

Das letzte grundlegende Kriterium, das vom WBP gefordert wird, *Kriterium 4*, bezieht sich nun auf die *externe Validität*, die gleichsam die dritte Dimension des umfassenden Kriterienkatalogs bildet. Unter externer Validität wird im Allgemeinen Folgendes verstanden:

> External validity concerns inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes. (Shadish et al., 2002, S. 83)

Um beobachtete Effekte innerhalb einer Studie bspw. von den untersuchten Studienpatienten auf eine intendierte Zielpopulation übertragen bzw. generalisieren zu können, gilt es zum einen abzuschätzen, inwieweit sich die Studienpatientengruppe von der intendierten Zielpopulation unterscheidet. Zum anderen gilt es daraufhin einzuschätzen, ob antizipierte Unterschiede zwischen diesen beiden Gruppen sich auch in unterschiedlichen Effekten einer psychotherapeutischen Maßnahme abzeichnen würden. Windeler (2008) bezeichnet diesen Umstand als "Effektmodifikation" und meint damit: „Es geht bei der Frage der Übertragbarkeit nicht darum, ob Patienten in Studien „anders“ sind als Patienten in der späteren Praxis (das ist sicher), sondern es geht darum, ob die Therapieeffekte anders sind“ (S. 255). Psychotherapeutische Interventionen stellen nun nach Schulte einen besonderen Gegenstand dar: „Die Berücksichtigung der externen Validität ist für Psychotherapieverfahren wichtiger als für die Überprüfung eines Medikaments, bei dem in der Regel der Unterschied zwischen den experimentellen Untersuchungsbedingungen und der alltäglichen Praxis nicht so gravierend ist“ (Deutsches Ärzteblatt, 2008a, S. 389). Es ist demnach davon auszugehen, dass im Rahmen von Psychotherapiestudien mehr Mühe auf die Herstellung von externer Validität verwendet werden muss, als es bspw. in rein pharmakologischen Studien der Fall ist. Zur Bewertung der externen Validität einer Studie siedelt der WBP diese in seinen Verfahrensregeln auf zweierlei Ebenen an, die mit Leichsenring und Kollegen (2011) folgendermaßen definiert werden können: Zu unterscheiden ist zwischen der externen Validität als *representativeness* einerseits und als *transfe-*

rability andererseits (vgl. auch Borkovec & Castonguay, 1998). Nach ersterer Definition fragt die externe Validität danach, ob die betreffenden Interventionen in einer Untersuchung derart durchgeführt wurden, dass sie als repräsentativ für die klinische Praxis betrachtet werden können, so dass der beobachtete Effekt auf diese übertragbar angenommen werden kann. Mit der zweiten Definition fragt die externe Validität nunmehr danach, ob die durchgeführte Behandlung in die klinische Praxis transferierbar ist:

> . . . that is, practitioners should be able to relatively easily apply the respective treatment to their patients, that is, for example, after a defined amount of training. Furthermore, the treatment should be applicable to those patients the therapists in clinical practice usually treat without the necessity to perform greater modifications of the treatment. (Leichsenring et al., 2011, S. 315)

Der WBP prägt dafür den treffenden Begriff "Praxistransfer". Der Praxistransfer setzt nicht erzwungenermaßen voraus, dass die durchgeführte Behandlung in der Studie auch tatsächlich als repräsentativ für die Praxis angesehen werden muss, sie muss jedoch in die klinische Praxis integrierbar und übertragbar sein. Wie der WBP mit Hilfe des Kriterienkatalogs den beiden Definitionen der externen Validität gerecht wird, wird ebenfalls in Kapitel 1.2.2 näher beleuchtet.

Die Begutachtung einer einzelnen Studie endet mit einer Ergebnisbewertung. Hier wird überprüft, ob die in den Untersuchungen ausgewiesenen Effekte für die Wirksamkeit des geprüften Verfahrens sprechen. Dafür sieht der WBP in seinen Verfahrensregeln folgende Optionen vor:

- Die in einer Studie für das spezifizierte Verfahren gezeigten Effekte zwischen Prä- und Postzeitpunkt sollten auf den primären Outcomemaßen signifikant größer sein als bei einer Kontrollgruppe (unbehandelte Gruppe/Warteliste, TAU oder Placebobehand-

lungen) *oder* mindestens äquivalent zu einer bereits etablierten Vergleichsbedingung. Bei letzterem Vergleich müssen zudem ausreichend große Stichproben dafür sorgen, dass auch kleine Effekte mit einer angemessenen statistischen Power zu entdecken sind

oder

- das in einer Studie gesetzte Behandlungsziel einer Heilung oder Besserung einer Störung sollte bei signifikant mehr Patienten erreicht worden sein als bei einer Kontrollgruppe *oder* bei einer annähernd gleich großen Anzahl an Patienten bei einem Vergleich mit einem etablierten Treatment (bei angemessen großen Stichproben)

oder

- bei einer Studie ohne Vergleichsbedingungen sollten die Effekte zwischen Prä- und Postmessung eine signifikante Veränderung in Richtung einer Besserung anzeigen, zudem sollten sich die gezeigten Veränderungen als klinisch bedeutsam ausweisen.

Der bis hier beschriebene Teil skizziert den wissenschaftlichen Begutachtungsprozess auf der "Studienebene", auf der empirische Wirksamkeitsuntersuchungen im Hinblick auf ihre methodische Qualität und Validität sowie auf ihre Ergebnisse bewertet werden. Auf den Bewertungsprozess der Gütemerkmale einer Studie mittels des Kriterienkatalogs des WBP, unter dessen Zuhilfenahme die Begutachtung standardisiert erfolgt, wird in Kapitel 1.2.2 noch detailliert eingegangen.

Die nächsthöhere Ebene soll hier als "Anwendungsbereichsebene" bezeichnet werden. Diese umfasst die Forderung einer Mindestanzahl an wissenschaftlich adäquaten Studien pro Anwendungs- bzw. Indikationsbereich, um als psychotherapeutisches Verfahren anerkannt zu

werden[3]. Für diese Forderung gilt entsprechend der Verfahrensregeln des WBP folgende Regel: Es müssen pro Anwendungsbereich mindestens insgesamt drei Studien vorliegen, die durch den WBP-Kriterienkatalog als methodisch adäquat einzustufen sind (Dimension der allgemeinen methodischen Qualität). Zudem müssen davon mindestens zwei Studien als hinreichend intern valide (Dimension der internen Validität) und mindestens zwei Studien als hinreichend extern valide (Dimension der externen Validität) einzustufen sein. Dies impliziert, dass mindestens eine Studie sowohl auf der internen als auch auf der externen Validitätsdimension mit hinreichend gutem Ergebnis abschneiden muss. Andernfalls ist dieses Mindestkriterium zu ersetzen durch vier methodisch adäquate Studien, davon zwei intern valide und zwei extern valide Studien. Zusätzlich muss geringstenfalls eine der drei bzw. vier Studien über eine Katamnesemessung verfügen, die einen Katamnesezeitraum von mindestens 6 Monaten umfasst. Zudem sollten die Studien auf keine erheblich schädigenden Effekte hinweisen (Nebenwirkungen, negative Folgewirkungen etc.); ist dies bei zwei oder mehr Studien der Fall, so kann die wissenschaftliche Anerkennung für den betreffenden Indikationsbereich nicht ausgesprochen werden. Das bedeutet, liegen bspw. für den Anwendungsbereich affektive Störungen drei bzw. vier gruppenstatistische Studien vor, die den eben genannten Mindestkriterien entsprechen und deren Ergebnisse (Effekte) für eine hinreichende Wirksamkeit des Verfahrens sprechen, dann gilt ein psychotherapeutisches Verfahren für diesen Indikationsbereich als wissenschaftlich anerkannt.

Jedoch muss ein Verfahren, um als *Verfahren* die wissenschaftliche Anerkennung zu erlangen und infolge zur vertieften Ausbildung empfohlen zu werden, seine empirische Evidenz in

[3] Es wird sich hier und im Weiteren lediglich auf die wissenschaftliche Anerkennung von *Verfahren* bezogen. Die Mindestforderungen, die für die wissenschaftliche Anerkennung von psychotherapeutischen *Methoden* erfüllt sein müssen, sind den Verfahrensregeln des WBP (2010) zu entnehmen.

mehreren Anwendungsbereichen unter Beweis gestellt haben. Diese Forderung betrifft nun die Ebene, die im Folgenden als "Verfahrensebene" bezeichnet werden soll: Zusammen mit dem Gemeinsamen Bundesausschuss hat der WBP im Jahr 2007 das sog. *Schwellenkriterium* beschlossen (vgl. Gemeinsamer Bundesausschuss, 2007), demzufolge für den Erwachsenenbereich[4] nur noch solche Psychotherapieverfahren als *Verfahren* anerkannt werden sollen, die zum einen über methodisch adäquate und valide Wirksamkeitsnachweise im Indikationsbereich der affektiven Störungen (F3) einschließlich der reaktiven Bindungsstörung mit Beginn in der Kindheit und Jugend (F94.1) und der psychischen oder Verhaltensstörungen im Wochenbett (F53) verfügen (vgl. WBP, 2010). Im Weiteren muss ein Verfahren über hinreichende Evidenz im Indikationsbereich der Angst-und Zwangsstörungen (F40 – F42; F93 und F94.0) verfügen.

Zusätzlich müssen entweder ausreichend viele methodisch adäquate und valide Untersuchungen aus *einem* der folgenden Indikationsbereiche vorliegen:

3. Somatoforme Störungen und dissoziative Störungen (Konversionsstörungen (F44 - F48)
4. Abhängigkeiten und Missbrauch (F1, F55)
5. Persönlichkeitsstörungen und Verhaltensstörungen (F6)

   (WBP, 2010, S. 27).

Alternativ können auch Studien aus *zwei* der folgenden Anwendungsbereiche vorgelegt werden:

6. Anpassungsstörungen und Belastungsstörungen (F43)

[4] Da die vorliegende Arbeit sich auf die wissenschaftliche Anerkennung von Verfahren für das Erwachsenenalter beschränkt, bleibt hier und im Folgenden der Bereich der Psychotherapie für das Kinder- und Jugendalter unberücksichtigt.

7. Essstörungen (F50)
8. Nicht-organische Schlafstörungen (F51)
9. Sexuelle Funktionsstörungen (F52)
10. Psychische und soziale Faktoren bei somatischen Erkrankungen (F54)
11. Schizophrenie, schizotype und wahnhafte Störungen (F2)
12. Organische, einschließlich symptomatischer psychischer Störungen F0)
13. Psychische und soziale Faktoren bei Intelligenzminderung (F7) und tiefgreifende Entwicklungsstörungen (F84)
14. Hyperkinetische Störungen (F90) und Störungen des Sozialverhaltens (F91, F94.2-F94.9)
18. Ticstörungen und Stereotypien (F95 und F98.4)

(WBP, 2010, S. 27).

Damit muss ein Verfahren, um als solches wissenschaftlich anerkannt und zur vertieften Ausbildung empfohlen zu werden, über methodisch adäquate und valide Studien in mindestens drei bzw. vier Anwendungsbereichen verfügen, wobei die beiden Bereiche der affektiven Störungen sowie der Angst- und Zwangsstörungen als obligatorisch zu betrachten sind. Diese „Schwelle“ begründen der Gemeinsame Bundesausschuss sowie der WBP mit der besonderen Versorgungsrelevanz dieser beiden Störungen (vgl. Gemeinsamer Bundesausschuss, 2007). Das Konzept der Versorgungsrelevanz ist dabei als ein mehrdimensionales zu verstehen, das sich aus der bevölkerungsepidemiologischen Dimension einerseits und der versorgungsepidemiologischen Dimension andererseits zusammensetzt. Dietmar Schulte, seinerzeit Vorsitzender des WBP, fasst diese beiden Dimensionen in folgende einfache Worte: „Wie weit verbreitet sind Störungen [bevölkerungsepidemiologische Dimension], und inwieweit kommen

Patienten mit diesen Störungen tatsächlich in psychotherapeutische Behandlung [versorgungsepidemiologische Dimension]?“ (Deutsches Ärzteblatt, 2008a, S. 389).

Der WBP sieht in Abstimmung mit dem Gemeinsamen Bundesausschuss neben den genannten Anwendungsbereichen eine weitere Patientengruppe vor, denen im wissenschaftlichen Anerkennungsprozedere gewissermaßen eine Art Sonderstellung zukommt:

> Anstelle der wissenschaftlichen Anerkennung in einem der Anwendungsbereiche . . . kann im Einzelfall ein Nachweis der Wirksamkeit durch eine entsprechende Anzahl von Studien zu „gemischten Störungen“ . . . anerkannt werden, sofern diese Studien einzeln und zusammen die unter II.5 genannten Kriterien erfüllen und sofern eine Zuordnung der einzelnen Studien zu einem der 18 Anwendungsbereiche der Psychotherapie nicht möglich ist und wenn der durch die einzelnen Studien geführte Wirksamkeitsnachweis nicht überwiegend auf Behandlungseffekte bei Störungen aus solchen Anwendungsbereichen zurückzuführen ist, für die bereits ein indikationsspezifischer Wirksamkeitsnachweis erbracht worden ist. (WBP, 2010, S. 27)

Unter der Patientengruppe mit „gemischten Störungen“ versteht der WBP „Patienten mit komplexen Störungen, die unter mehreren Diagnosen nach der *International Statistical Classification of Diseases and Related Health Problems* (ICD-Diagnosen) abgebildet werden, und/oder diagnostisch gemischte Patientengruppen“ (2010, S. 10). Ihre Sonderstellung erfährt diese Patientengruppe bzw. dieser Studientypus zum einen durch eine Fußnote, in der der WBP betont, dass die gemischten Störungen *keinen* weiteren Anwendungsbereich darstellen (S. 10) und zum anderen durch folgenden Zusatz:

> Eine Berücksichtigung von Wirksamkeitsnachweisen für ein Psychotherapieverfahren durch Studien zu gemischten Störungen bei der Empfehlung für die Ausbildung zum Psychologischen Psychotherapeuten bedarf einer umfassenden Abwägung im Einzelfall, inwieweit dieser Wirksamkeitsnachweis in seiner

> Bedeutung einem Wirksamkeitsnachweis in einem der Anwendungsbereiche 6 bis 14 oder in dem Anwendungsbereich 18 gleichkommt. (WBP, 2010, S. 27)

Damit unterliegt die Integration dieser Studien einer Art Entwicklung, die ihren Ausgangspunkt in vorherigen Versionen des Methodenpapiers nimmt und bei dem eben zitierten Zusatz aus der aktuellen Version des Methodenpapiers (Version 2.8) endet. So heißt es bspw. noch in Version 2.6:

> Die Kategorie „Gemischte Störungen“ . . . wird wie ein weiterer, neunzehnter Anwendungsbereich gewertet, der die wissenschaftliche Anerkennung eines Psychotherapieverfahrens in einem der Anwendungsbereiche 6 bis 14 oder in dem Anwendungsbereich 18 ersetzen kann. (WBP, 2007, S. 28)

In einem 2008 im Psychotherapeutenjournal veröffentlichten Interview mit dem damaligen Vorsitzenden des WBP, Dietmar Schulte, sowie mit dessen Stellvertreter, Gerd Rudolf, wird von letzterem zudem auf die Wichtigkeit der Kategorie der gemischten Störungen hingewiesen:

> Rudolf: Der wissenschaftliche Umgang mit komplexen, komorbiden Störungen ist zweifellos schwierig. In der Logik der Verhaltenstherapie lassen sich hier durch sorgfältige Diagnostik durchaus Hauptdiagnosen definieren, die dann für die Indikation, Behandlung und Ergebniseinschätzung handlungsleitend sind. Aus psychodynamischer Sicht, die sich um die Beschreibung einer ganzheitlichen Persönlichkeitsentwicklung bemüht, erscheint die Heraushebung einer Hauptdiagnose und darauf zugeschnittener Behandlungen eher fraglich. Als ein Schritt zur besseren Berücksichtigung dieser für Praxisstudien bedeutsamen Patientengruppe hat der Beirat jenseits der achtzehn diagnostischen Anwendungsbereiche die Kategorie „gemischte Störungen“ eingerichtet. (Psychotherapeutenjournal, 2008, S. 115)

Bereits in der darauf folgenden Fassung der Verfahrensregeln (Versionen 2.7) aus dem Jahr 2009 wird die gleichberechtigte Bewertung der gemischten Störungen als weiterer Anwendungsbereich geschmälert, indem die grundlegende Abwägung über solche Studien im Einzelfall hervorgehoben wird. Nach welchen Gesichtspunkten diese Abwägung genau erfolgt, führen die Verfahrensregeln bis zur aktuell gültigen Fassung (Version 2.8) nicht näher aus.

Über die Begutachtung der verfahrensspezifischen empirischen Evidenz mit Hilfe des bereits eingeführten Kriterienkatalogs hinaus, berücksichtigt das WBP-Gutachtergremium sowohl Forschungsbefunde zur Wirkweise (Prozessforschung) des betreffenden Verfahrens als auch zu seiner Anwendung und Verbreitung in der Praxis. In einem Antrag auf wissenschaftliche Anerkennung eines Verfahrens müssen ferner Angaben zur theoretischen Einbettung des Verfahrens gemacht werden, außerdem zur Behandlungsplanung und zur therapeutischen Beziehungsgestaltung sowie zu Kontraindikationen und zu potentiellen unerwünschten Wirkungen. Abschließend sollte die Aus- und Weiterbildung in diesem Verfahren näher beschrieben werden. Beispielhaft sollen an dieser Stelle die Zusammenstellungen nach den genannten Gesichtspunkten für die Richtlinienverfahren angeführt werden: Für die Verhaltenstherapie wurde im Auftrag der Deutschen Gesellschaft für Verhaltenstherapie (DGVT) sowie der Arbeitsgemeinschaft für Verhaltensmodifikation (AVM) von Kröner-Herwig im Jahre 2004 eine solche Zusammenstellung vorgelegt. Diese Sammlung enthält eine umfassende Auflistung anwendungsbereichsbezogener empirischer Wirksamkeitsstudien. Für die psychodynamischen Verfahren legten Brandl und Kollegen (2004) in Zusammenarbeit mit der Deutsche Gesellschaft für Psychoanalyse, Psychotherapie, Psychosomatik und Tiefenpsychologie e.V. (DGPT) sowie in Verbindung mit weiteren psychoanalytischen Fachgesellschaften ebenfalls einen umfassenden Überblick samt Auflistung der empirischen Evidenz vor. Ergänzt wurde diese Auflistung durch eine gesonderte Zusammenstellung von Wirksamkeitsuntersuchungen,

die durch die Deutsche Fachgesellschaft für tiefenpsychologisch fundierte Psychotherapie e.V. (DFT) in Auftrag gegeben wurde (vgl. Richter, Loew, Calatzis & Krause, 2002). Allerdings erfolgte die wissenschaftliche Anerkennung der genannten Verfahren, wie in Kapitel 1.1 bereits ausgeführt, noch unter Anwendung der Vorgängerpapiere des Methodenpapiers (vgl. WBP, 1999, 2000a/b, 2002b, 2004b). Welche Entwicklungen dem Methodenpapier vorausgingen und welcher Kritik sich der WBP ausgesetzt sah, bevor er dieser mit den neuen Verfahrensregeln begegnete, ist Gegenstand des folgenden Kapitels (1.2.1).

1.2.1 Vorlauf zum Methodenpapier des Wissenschaftlichen Beirats Psychotherapie

Das Methodenpapier des WBP kann als ein Resultat eines langjährigen Diskurses betrachtet werden, in dem sich sowohl im deutschen als auch im internationalen Sprachraum über das Thema der evidenzbasierten Psychotherapie ereifert und auseinandergesetzt wurde. Ausgangspunkt der Debatten bildet das Konzept der EbM, die Sackett und Kollegen (1996) folgendermaßen definieren:

> Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research. (Sackett, Rosenberg, Gray, Haynes & Richardson, 1996, S. 71)

Demnach geht es um ein Zusammenführen von empirischer Evidenz (*external clinical evidence from systematic research*) und der klinischen Expertise, also dem Erfahrungswissen klinisch tätiger Praktiker (*individual clinical expertise*). Im Forschungskontext wurde sich vor allem auf den ersten Aspekt konzentriert, nämlich auf die Sammlung und Integration empirischer Befunde diverser medizinischer Interventionsmaßnahmen und – wohl das Herzstück der EbM – auf die kritische Bewertung der empirischen Evidenz. Als handlungsleitend können

die von unterschiedlichen Netzwerken, etwa die Cochrane Collaboration, und von sog. Task Forces (z.B. Canadian Task Force on the Periodic Health Examination, US Preventive Services Task Force) verwendeten und weiterentwickelten Evidenzhierachien betrachtet werden. Mit Hilfe solcher Taxonomien wurde es möglich, die Sicherheit von Empfehlungen (in Form von Behandlungsleitlinien) einschätzen zu können. Tabelle 1 enthält eine Fassung der allgemein verbreiteten Evidenzhierachie (vgl. auch Auckenthaler, 2012), auf die sich u.a. das Deutsche Cochrane Zentrum, das Ärztliche Zentrum für Qualität in der Medizin (ÄZQ), das Deutsche Netzwerk Evidenzbasierte Medizin e.V. (EbM Netzwerk) sowie das National Institute for Health and Care Excellence (NICE) berufen.

Tabelle 1: Evidenzhierarchie

Ia	Evidenz aus einer Metaanalyse von mindestens drei randomisierten kontrollierten Studien (randomized controlled trials, RCTs).
Ib	Evidenz aus mindestens einer randomisiert kontrollierten Studie oder einer Metaanalyse von weniger als drei RCTs.
IIa	Evidenz aus zumindest einer methodisch guten, quasi-experimentellen deskriptiven Studie.
IIb	wenigstens eine hochwertige Studie eines anderen Typs quasi-experimenteller Studien.
III	Evidenz aus methodisch guten, nichtexperimentellen Beobachtungsstudien, wie z. B. Vergleichsstudien, Korrelationsstudien und Fallstudien.
IV	Evidenz aus Berichten von Expertenkomitees oder Expertenmeinung und/oder klinische Erfahrung anerkannter Autoritäten.

aus: S 3-Leitlinie/Nationale VersorgungsLeitlinie Unipolare Depression – Langfassung Version 1.3, Deutsche Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde (DGPPN) et al., 2012.

Die oberste Maxime der in Tabelle 1 verzeichneten Systematik liegt in der Zusammenstellung von solchen Studien, denen ein hohes Vertrauen im Hinblick auf kausale Schlussfolgerungen entgegengebracht werden kann. Dazu sind Studien notwendig, die durch interne Validitätssicherungsstrategien eben diesen kausalen Schluss rechtfertigen, nämlich dass allein die unabhängige Variable (z.B. ein Medikament) einen Einfluss auf die abhängige Variable (z.B. eine Erkrankung) hat. Alternativerklärungen sollten nahezu ausgeschlossen bzw. als unplausibel betrachtet werden können (vgl. Kap. 1.2). Randomisierte kontrollierte Studien (*randomized*

controlled trials [RCTs]) werden für diesen Zweck als besonders geeignet betrachtet und gelten als Goldstandard der Evidenzbasierung (u.a. Sackett et al., 1996). Metaanalysen, die auf dieser Art Studien beruhen, stellen aus diesem Grunde die höchste Evidenzstufe dar (Stufe Ia; vgl. Tabelle 1). Auf typische Charakteristika von RCTs – neben dem, dass die Zuweisung der Probanden zu den einzelnen Treatmentarmen per Zufall erfolgt – soll weiter unten, wenn es um den spezifisch psychotherapeutischen Kontext geht, näher eingegangen werden. Evidenzstufe II bezieht sich auf die Evidenz, die durch quasiexperimentelle Untersuchungen dargestellt wird. Da eine zentrale interne Validitätssicherungstechnik in diesen Untersuchungen nicht umgesetzt wird, die Randomisierung zwecks Kontrolle bekannter und unbekannter personsgebundener Störvariablen, genießt dieser Studientyp einen niedrigeren Status in der Evidenzhierachie. Studien, wie Korrelationsstudien, werden der Evidenzstufe III zugeordnet; wird eine Evidenz allein durch Expertenmeinungen und Erfahrungswissen abgebildet, so stellt diese die letzte Evidenzstufe dar (Stufe IV). Basieren klinische Handlungsweisen allein auf Expertenmeinungen und klinischer Erfahrung, so werden diese oftmals polemisch als *eminenzbasiert* statt evidenzbasiert bezeichnet (u.a. Köbberling & Wehner, 2009; Zurhorst, 2003).

Einen ganz ähnlichen Ansatz in Bezug auf die Hierarchisierung von Evidenz verfolgt eine noch verhältnismäßig junge Arbeitsgruppe (Gründungsjahr: 2000), die *Grades of Recommendation, Assessment, Development, and Evaluation Working Group* (GRADE Working Group). Die Gruppe entwickelte ein System zur Bewertung empirischer Evidenz sowie ein Reglement zur Verknüpfung von Evidenzbewertung und den daraus folgenden Handlungsempfehlungen (vgl. Brozek et al., 2009). Im medizinischen Kontext erfreut sich das GRADE System einer immer größeren Verbreitung, so beruft sich bspw. die Cochrane Collaboration auf den GRADE Ansatz (vgl. Higgins & Green, 2011). Da diesem Ansatz weitge-

hend dieselbe Hierarchisierung von empirischer Evidenz in Form einer Betonung von RCTs zugrunde liegt, soll auf das GRADE System nicht detaillierter eingegangen werden.

Den genannten Systemen, die sich aus der Bewegung der EbM heraus entwickelt haben, sind eine Favorisierung von RCTs und eine damit einhergehende Betonung der internen Validität gemeinsam. An die externe Validität werden zumindest in den genannten und am weitesten verbreiteten Systemen keine besonderen Forderungen gestellt. Wie in Kapitel 1.2 bereits zitiert, erklärt Schulte diese Vernachlässigung damit, dass bspw. für die Nutzenbewertung eines Medikaments die Berücksichtigung dieses Gütemerkmals weniger zentral sei, da der Unterschied zwischen den Untersuchungsbedingungen in einem RCT und der klinischen Routinepraxis nicht so grundlegend sei (Deutsches Ärzteblatt, 2008a). In Bezug auf den pharmakologischen Kontext besteht hinreichender Konsens über Schultes Haltung, das Blatt wendet sich jedoch im Hinblick auf psychosoziale Interventionen, wie Psychotherapie, in der RCTs längst nicht mehr ein Alleinstellungsmerkmal bilden. Darauf wird in den folgenden Abschnitten näher eingegangen.

Inspiriert durch die Vorgehensweisen in der EbM formierte sich zu Beginn der 90er Jahre innerhalb der American Psychological Association (APA) Division 12 eine Task Force – die *Task Force on Promotion and Dissemination of Psychological Procedures* – die für die Entwicklung von Grundsätzen zur Begutachtung von empirischer Evidenz (Wirksamkeitsstudien) zuständig war. Mit Hilfe dieses Regelwerks sollten im Weiteren Listen erstellt werden, denen *empirically validated treatments* bzw. *empirically supported treatments*[5] (ESTs) entnommen

[5] Im Laufe der Jahre wurde die Bezeichnung *empirically supported treatments* (ESTs) immer populärer und löste die Bezeichnung *empirically validated therapies* ab. Diese Änderung der Begriffswahl geht auf eine Kritik an dem Begriff "*validated*" zurück, der zu sehr nahelege, dass es sich bei dem Prozess der Evidenzuntersuchung

werden konnten (vgl. Chambless, 1996; Chambless & Hollon, 1998; Chambless & Ollendick, 2001; Chambless et al., 1996; Chambless et al., 1998; Task Force on Promotion and Dissemination of Psychological Procedures, 1995). Kern des Unternehmens bildet eine Taxonomie, die die Einordnung von psychotherapeutischen Behandlungen in eine der drei Gruppen „*well-established treatments*“, „*probably efficous treatments*“ oder „*experimental treatments*“ vorsieht (vgl. Abbildung 1).

Well-established treatments

I. At least two good between-group design experiments must demonstrate efficacy in one or more of the following ways:
   A. Superiority to pill or psychotherapy placebo, or to other treatment
   B. Equivalence to already established treatment with adequate sample sizes

OR

II. A large series of single-case design experiments must demonstrate efficacy with
   A. Use of good experimental design and
   B. Comparison of intervention to another treatment

III. Experiments must be conducted with treatment manuals or equivalent clear description of treatment

IV. Characteristics of samples must be specified

V. Effects must be demonstrated by at least two different investigators or teams

Probably efficacious treatments

I. Two experiments must show that the treatment is superior to waiting-list control group

OR

II. One or more experiments must meet well-established criteria IA or IB, III, and IV above but V is not met

OR

III. A small series of single-case design experiments must meet well-established-treatment criteria

Experimental treatments

Treatment not yet tested in trials meeting task force criteria for methodology

aus: Chambless & Ollendick, 2001, S. 689.

Abbildung 1: Kriterien zur Bestimmung von empirically validated treatments nach der Task Force on Promotion and Dissemination of Psychological Procedures – APA Division 12

In den Kriterien aus Abbildung 1 spiegelt sich die Betonung experimenteller Untersuchungen wider, die bereits in den Evidenzhierarchien der EbM (vgl. Tabelle 1) zu finden war. Ferner wird von qualitativ hochwertigen Studien gefordert, dass die untersuchten Behandlungen ma-

zu einer bestimmten Behandlung irgendwann um einen unwiderruflich abgeschlossenen Prozess handelt. Aus der Bezeichnung "*supported*" hingegen ginge – so die Kritik – eher hervor, dass es sich bei der Evidenzbasierung einer Behandlung immer um einen in Entwicklung befindlichen Prozess handele, der strenggenommen nie als abgeschlossen betrachtet werden könne (vgl. Chambless, 1996; Chambless & Hollon, 1998).

nualisiert erfolgen und die Studienpatienten nach diagnostischen Kriterien spezifiziert werden (bspw. durch Klassifikationen nach dem *Diagnostic and Statistical Manual of Mental Disorders* [DSM]; vgl. Chambless & Hollon, 1998). Um als Behandlung tatsächlich das Prädikat *well-established treatment* zu erhalten, müssen die Ergebnisse außerdem von unabhängigen Forschergruppen repliziert werden. Im Hinblick auf das Studiendesign wird gefordert, dass als Vergleichsgruppen bestenfalls solche Bedingungen herangezogen werden sollen, durch die die sog. *unspezifischen Wirkfaktoren*, die allen Behandlungen gemeinsam sind, kontrolliert werden. Diese Forderung wird in der Regel von Placebo-Kontrollgruppen erfüllt. Solche Designs sollen dementsprechend Aufschluss über die *spezifische Wirkung* einer Behandlung geben, man spricht auch von der *Nettowirkung* einer Behandlung (Hager & Hasselhorn, 2000; dazu ausführlich: Kap. 3.1.1). Alternativ kann sich eine psychotherapeutische Behandlung auch dann als *well-established treatment* etablieren, wenn sie sich in einer Untersuchung mit ausreichender statistischer Power in ihrer Wirksamkeit als gleichgut im Vergleich zu einer bereits überprüften Behandlung erweist (dazu ausführlich: Kap. 3.2.3). Werden hingegen ausschließlich Vergleiche mit Wartlisten-Kontrollgruppen durchgeführt, in der sich nicht die spezifische Wirkung des zu überprüfenden psychotherapeutischen Treatments zeigen kann, so kann sich eine so überprüfte Behandlung allenfalls als *probably efficous treatment* etablieren. Gleiches trifft auf Behandlungen zu, die zwar mittels hochwertiger Untersuchungen entsprechend der *well-established treatment*-Kategorie überprüft wurden, für die jedoch das Kriterium zweier unabhängiger Forschergruppen nicht zutrifft. Als *experimental treatments* werden schließlich solche psychotherapeutischen Verfahren oder Methoden bezeichnet, die noch in keinen Studien, die den Anforderungen der Task Force entsprechen, ihre Wirksamkeit nachgewiesen haben.

Im Jahr 1995 wurde eine erste exemplarische Aufstellung derjenigen psychotherapeutischen Verfahren und Methoden herausgegeben, deren empirische Evidenz den genannten

Anforderungen der Task Force genügten (vgl. Task Force on Promotion and Dissemination of Psychological Procedures, 1995), es folgte ein Update dieser Auflistung (vgl. Chambless et al., 1996) und kurze Zeit später eine weitere Aktualisierung (vgl. Chambless et al., 1998). Parallel zu den Aktivitäten der Task Force machte sich eine Welle der Kritik breit, die sich hauptsächlich auf einen Aspekt bezog, der in folgender Stellungnahme der Task Force treffend zum Ausdruck kommt:

> However, controlled outcome studies, or a large series of single case designs, are likely to remain the source of most policy decisions and clinical recommendations. Moreover, we believe establishing efficacy in contrast to a waiting-list control group is not sufficient. Relying on such evidence would leave psychologist at a serious disadvantage vis a vis psychiatrists who can point to numerous double-blind placebo trials to support the validity of their interventions. (Task Force on Promotion and Dissemination of Psychological Procedures, 1995, S. 4)

Motivation zu den o.g. Kriterien (vgl. Abbildung 1) war es demnach, einen Korpus der Evidenz zu erschaffen, der sich mit den Qualitätsstandards pharmakologischer Studien vergleichen ließ. Eben diese Übernahme der Qualitätsstandards der EbM bot genügend Anlass zur Kritik, auf die noch näher einzugehen sein wird. An dieser Stelle soll zunächst lediglich ein Hauptstrang der Kritik in Form eines weiteren Zitats zusammengefasst werden:

> Die Task Force der Division 12 hatte von der pharmakologischen Forschung nicht nur die Bewertung der Evidenzstufen übernommen, sie hatte damit (zumindest implizit) auch alle jene psychotherapeutischen Verfahren um ihre Chance auf einen Platz in der Gesundheitsversorgung gebracht, die von ihrem Selbstverständnis her kaum die Evidenzstufe I erreichen können: weil sie zu lange dauern, als dass sie im Rahmen experimenteller Designs untersucht werden könnten, und/oder weil sie auf ein individualisiertes statt auf ein manualisiertes Vorgehen setzen. (Auckenthaler, 2012, S. 236)

In Deutschland war im Rahmen der wissenschaftlichen Anerkennung von psychotherapeutischen Verfahren und Methoden durch den WBP eine ganz ähnliche Entwicklung zu beobachten, sowohl, was die Wahl der Kriterien zur Qualitätsbemessung der empirischen Evidenz als auch, was die darauf folgende Kritik betrifft. Diese entfachte sich als erstes infolge des im Jahre 2000 veröffentlichten Gutachtens samt Minderheitenvotum zur Gesprächspsychotherapie (WBP, 2000c). Resultat der Begutachtung durch den WBP war es, dass die Gesprächspsychotherapie nicht als Verfahren für die vertiefte Psychotherapieausbildung empfohlen wurde, „da dieses Therapieverfahren nicht für die Mindestzahl von fünf der zwölf Anwendungsbereiche der Psychotherapie des Wissenschaftlichen Beirates Psychotherapie . . . beziehungsweise für mindestens vier der klassischen Anwendungsbereiche als wissenschaftlich anerkannt gelten kann“ (WBP, 2000c, S. 63). Die Regelung im Hinblick auf die Anwendungsbereiche, für die mindestens Wirksamkeitsstudien hätten vorliegen müssen, ist einem bereits genannten Papier aus dem Jahr 2000 zu entnehmen (WBP, 2000b). Nach welchen methodischen Kriterien die vom WBP gesichteten Studien im Einzelnen bewertet wurden, bleibt leider undurchsichtig, da über diese keine Veröffentlichung existiert. Allein dem Leitfaden für die Antragstellung eines Gutachtens (WBP, 1999) ist im Hinblick auf Qualitätsanforderung an Wirksamkeitsstudien zumindest Folgendes zu entnehmen:

> Als Wirksamkeitsnachweise können verschiedene Arten von Untersuchungen angeführt werden (z. B. kontrollierte Gruppenstudien, ggf. auch kontrollierte Einzelfallstudien, Metaanalysen). Erforderlich sind multimodale Erfolgsnachweise (nicht nur Beurteilungen der Therapeuten) bei den relevanten Patientenpopulationen. Weiterhin Angaben zur Dauerhaftigkeit der Therapieeffekte auf der Grundlage von Katamnesen und zur Frage, inwieweit die festgestellten Wirkungen tatsächlich auf das jeweilige Verfahren zurückzuführen sind. (S. 1015)

In einer kurzen Bekanntmachung betont der WBP zudem, dass vor allem solche Studien besonderes Gewicht erfahren, die eine Übertragung der Wirksamkeitsnachweise auf die klini-

sche Versorgungspraxis erlauben (WBP, 2000d). Solche Studien werden in der Regel als *effectiveness*-Studien bezeichnet und sind von den sog. *efficacy*-Studien abzugrenzen, denen ein randomisiert-kontrolliertes Design zugrunde liegt, die eine Übertragung auf die klinische Praxis oft erschweren (u.a. Auckenthaler, 2012; Fydrich & Schneider, 2007). Mit diesem Zusatz der Betonung von *effectiveness*-Studien begegnete der WBP bereits zu Beginn seiner Amtszeit einem Hauptkritikpunkt, der schon in Reaktion auf die oben beschriebenen Aktivitäten der Task Force der APA Division 12 geübt wurde. Diese Kritik bezog sich eben auf die beschränkte Generalisierbarkeit von Ergebnissen aus RCTs und auf die trotz dieser Beschränkung persistierende Dominanz der RCT-Methodologie im Rahmen der Listung und Verbreitung von ESTs. Trotz der angekündigten Gewichtung und Integration generalisierbarer Studien seitens des WBPs rief das Gutachten zur Gesprächspsychotherapie diverse Widerstände hervor. Aus dem Minderheitenvotum von Richter, das dem Gutachten des WBP zur Gesprächspsychotherapie anhängt, geht Folgendes hervor:

> In der Stellungnahme der Berichterstatter . . . werden zur Beurteilung ausschließlich kontrollierte Wirksamkeitsstudien (im Sinne von efficacy) herangezogen. Dies steht nicht im Einklang mit den Kriterien des Wissenschaftlichen Beirats, wonach bei der Beurteilung der wissenschaftlichen Anerkennung auch andere Studien (zum Beispiel Einzelfallstudien, insbesondere aber effectiveness-Studien) berücksichtigt werden können. (WBP, 2000c, S. 63)

Die genaue Rekonstruktion dessen, welche internen Entscheidungsschritte den WBP im Einzelnen dazu bewogen haben, nur kontrollierte Studien – und keine *effectiveness*-Studien – in sein Gutachten einfließen zu lassen und welche Studien er aus welchen Gründen unberücksichtigt ließ oder aber als methodisch unzureichend bewertet hat, hätte den Rahmen der Arbeit gesprengt. Festzuhalten bleibt jedoch, dass sich der WBP fortan der Kritik ausgesetzt sah, zu sehr der RCT-Methodologie anzuhängen und damit einem wissenschaftlichen Anerkennungsprocedere zu folgen, das dem Gegenstand (Psychotherapie) nicht angemessen sei (u.a. Au-

ckenthaler, 2000a; Petersen, 2003; Revenstorf, 2005; Tschuschke, 2005; Zepf & Hartmann, 2002; Zurhorst, 2003). An dieser Kritik änderte sich auch nach Veröffentlichung der im Jahre 2004 veröffentlichten, mithin maßgeblichen Mindestanforderungen (WBP, 2004b; vgl. Abbildung 2) nichts, vielmehr gaben sie erneut Anlass zur Beanstandung.

Für Studien, die nach dem 1. Januar 1990 publiziert wurden, gelten die folgenden Mindestanforderungen[6]:
Jedes der im Folgenden genannten fünf Kriterien muss mindestens als „ausreichend erfüllt" beurteilt werden:

1. Stichprobe
   1.1. Kennzeichnung der Stichprobe durch Klinische Diagnosen (DSM oder ICD) oder klinisch relevante Syndrome (Falls abweichend: Es muss sich um Personen mit Störungen von Krankheitswert handeln.)
   1.2. Objektive und reliable Diagnosestellung (in der Regel durch [teil-]standardisierte Interviews)
   1.3. Angabe weiterer deskriptiver Daten zur Beurteilung der Repräsentativität der Stichprobe
   1.4. Rekrutierung der Stichprobe muss Rückschluss auf Grundgesamtheit ermöglichen
   1.5. Bei Studien mit heterogener Klientel müssen die Ergebnisse jeweils für Patienten mit bestimmten Störungen getrennt beurteilt werden können.

2. Behandlung
   2.1. Operationale Definition der zu untersuchenden Behandlung und der Kontrollbedingungen (in der Regel durch Manuale)
   2.2. Festlegung eines operationalisierbaren Behandlungsziels
   2.3. Maßnahmen zur Abschätzung der Treatment-Integrität (Manualtreue)

3. Design der Studie
   3.1. Kontrollgruppendesign
   3.2. Randomisierung (oder Parallelisierung) der Untersuchungsgruppen

4. Outcome-Messung
   4.1. inhaltlich: (mehrere) Instrumente, die Rückschluss auf Heilung oder Besserung der behandelten Störung erlauben (Symptomatik, Schweregrad, Störungsfolgen wie Beeinträchtigungen, Leiden, Inanspruchnahme medizinischer Dienste)
   4.2. Kriterium: Grad der Veränderung (Prä-Post-Vergleich, Effektstärke) und/oder Grad der Zielerreichung (klinische Signifikanz; retrospektive Erfolgsbeurteilung)

5. Ergebnis
   5.1. Effektivität: Die Experimentalgruppe muss unbehandelten Kontrollgruppen deutlich überlegen sein beziehungsweise mindestens vergleichbare Effekte haben wie bereits hinreichend empirisch überprüfte Behandlungen von Kontrollgruppen.

Mindestens eine Studie zu einem der vom WBP definierten Störungsbereiche muss eine Katamnese-Untersuchung mindestens sechs Monate nach Therapieabschluss einschließen. Der Therapieerfolg muss auch noch mindestens sechs Monate nach Therapieende nachweisbar sein.
aus: WBP, 2004a.

Abbildung 2: Die Mindestkriterien des WBP

Es ist unschwer zu erkennen, dass es Parallelen zu den Kriterien aus Abbildung 1 gibt, die sich auf die Evidenzanforderungen der *well-established treatments* beziehen (Kennzeichnung

[6] Es wurden im selben Papier außerdem Mindestanforderungen für Studien, die vor 1990 publiziert wurden, formuliert. Diese entsprechen einem weniger strengen Maßstab und sind weit weniger ausführlich, so dass sich hier nur auf die in Abbildung 2 verzeichneten Kriterien bezogen wird.

der Stichprobe durch klinische Diagnosen, Manualisierung, Kontrollgruppendesign). Zudem fehlen Kriterien, die sich zur Bewertung der Generalisierbarkeit einzelner Studienergebnisse eignen – ein Gesichtspunkt, dem der WBP bereits zu diesem Zeitpunkt besonderes Gewicht verlieh (vgl. WBP 2000d).

Die Kritik, die sowohl im angloamerikanischen Sprachraum als auch in Deutschland – dort in teils extrem polemischer Art und Weise – an der weitgehenden Übernahme der Vorgehensweise der EbM bzw. der Pharmaforschung erhoben wurde, ist mannigfaltig und bezieht sich längst nicht nur auf die eingeschränkte Generalisierbarkeit von Ergebnissen aus kontrolliert-randomisierten Untersuchungen. Vielmehr geht es neben der eingeschränkten Generalisierbarkeit von RCTs noch um etwas Grundlegenderes: um eine zumindest partielle Unvereinbarkeit der RCT-Methodologie mit dem Gegenstand "Psychotherapie" (u.a. Westen, Novotny & Thompson-Brenner, 2004). Bevor im letzten Teil dieses Kapitels auf diese beiden Argumentationsstränge (eingeschränkte Generalisierbarkeit und Unvereinbarkeit) näher eingegangen wird, sollen die wichtigsten Charakteristika von *efficacy*- und *effectiveness*-Studien kurz umrissen werden.

Oftmals werden *efficacy*-Studien mit dem Begriff „Wirksamkeitsstudien“ gleichgesetzt (vgl. Auckenthaler, 2012), dieser Gleichsetzung soll in dieser Arbeit jedoch nicht gefolgt werden (vgl. auch 3.1.1). Vielmehr werden die Begriffe „Wirksamkeitsstudie“ oder „Outcomestudie“ als Oberbegriffe für beide Arten von Studien, den *efficacy*- und den *effectiveness*-Studien, betrachtet.

Eine prägnante Charakterisierung von *Efficacy* –Studien liefert Lambert (2013a):

> Efficacy studies emphasize the internal validity of experimental design through a variety of means, including (a) controlling the types of patients included in the study (e.g., limiting the number of clients with comorbid disorders), (b) using manuals to standardize treatment delivery, (c) training therapists prior to the study, monitoring therapist adherence to the treatment during the study, and supervising therapists to ensure they do not deviate from the treatment protocol, (d) managing the "dose" of treatment through analyses that include only patients who have received a specific amount of treatment, and (e) random assignment of clients to treatments, and (f) the use of blinding procedures for raters. These and other strategies are used to enhance the investigator's ability to make causal inferences based on the findings. (S. 192)

Ferner schreibt der Einsatz von Manualen in den meisten Fällen ein störungsspezifisches Vorgehen in der Behandlung und damit diagnosehomogene Patientengruppen vor (vgl. Westen et al., 2004); die Behandlungsumfänge (Sitzungsanzahl) werden in der Regel a priori festgelegt, zudem werden in RCTs Kontrollgruppendesigns in Form von Vergleichen mit Wartelistengruppen, Placebobehandlungen oder TAU favorisiert (u.a. Leichsenring, 2004a/b). Wie der letzte Satz im obigen Zitat nahelegt, dienen diese Strategien in erster Linie der internen Validitätssicherung (vgl. Kap. 1.2). In RCTs geht es damit um die maximale Kontrolle von Störvariablen und bestenfalls um die Überprüfung der *spezifischen Wirkung* eines psychotherapeutischen Verfahrens oder einer psychotherapeutischen Methode, sofern entsprechende Kontrollgruppen eingesetzt werden (vor allem Placebo). Durch Strategien, wie der Manualisierung und der Adherence-Kontrolle sollen zum einen Therapeutenvariablen kontrolliert werden (vgl. Beutler et al., 2004; Fydrich & Schneider, 2007; Westen et al., 2004), zum anderen soll sichergestellt werden, dass die als wirksam angenommenen *spezifischen* Techniken der Behandlung auch tatsächlich *lege artis* umgesetzt werden. (u.a. Chambless, 1996). Die Untersuchung möglichst monomorbider Störungsbilder bzw. „reiner" Samples dient dem Ausschluss persongebundener Störfaktoren, ferner wird durch eine Homogenisierung der

Stichproben eine Varianzminimierung erreicht, wodurch Effekte besonders gut sichtbar werden[7].

Auch für die Charakterisierung von *effectiveness*-Studien liefert Lambert (2013a) eine treffende Umschreibung:

> Typically, clients are not as carefully preselected, treatment dose is less controlled, and therapist adherence is neither monitored nor modified to a pure-form, manually determined treatment. Therapists tend to be those working in the settings and may or may not receive the same level of prestudy training as that of the efficacy study. (S. 192)

Diese Untersuchungen finden unter den möglichst unveränderten Bedingungen der psychotherapeutisch-klinischen Praxis statt, die Untersuchungsstichproben sind in der Regel keine durch strenge Ein-/Ausschlusskriterien homogenisierte Gruppen und die Patienten wählen die Behandlungsoptionen selbst. Wegen des zuletzt genannten Aspekts werden *effectiveness*-Studien auch oftmals mit quasiexperimentellen Studien gleichgesetzt oder aufgrund der Nähe zur realen Behandlungspraxis auch als „naturalistische Studien" bezeichnet. Dementsprechend erhalten die Therapeuten kein zusätzliches Training in der durchzuführenden Behandlung und der Sitzungsumfang ist nicht a priori festgelegt.

Effectiveness-Studien wird aufgrund der genannten Umstände in der Regel eine hohe externe Validität attestiert, im Gegensatz dazu gelten RCTs, wenn methodisch adäquat durchgeführt, als intern valide Untersuchungen (u.a. Lambert, 2013a). Tabelle 2 fasst die zentralen Charakteristika der beiden Studienarten noch einmal zusammen (vgl. auch Benecke, 2014a).

[7] Die Homogenisierung der Stichproben und damit erreichte Varianzminimierung sichert strenggenommen weniger die interne Validität, sondern die statistische Validität (vgl. Shadish et al., 2002).

Tabelle 2: Überblick über zentrale Charakteristika von *efficacy*- und *effectiveness*-Studien

***efficacy*-Studien (RCTs)**	***effectiveness*-Studien (naturalistische Studien)**
Frage nach spezifischer Wirksamkeit eines Treatments, d.h. möglichst unter Kontrolle allgemeiner/unspezifischer Wirkfaktoren	Frage nach Wirksamkeit unter realen Bedingungen
randomisierte Zuteilung der Patienten	Patienten teilen sich selbst den Behandlungsalternativen zu
Kontrollgruppendesigns: Vergleichsbedingungen unter Angleichung der Randbedingungen und ohne die spezifische Maßnahme (entweder Wartelistengruppen oder Placebo, TAU, aktive Kontrollgruppe)	Vergleiche meist mit alternativen Behandlungen (komparative Studien)
diagnosehomogene Behandlungsgruppen unter Ausschluss komorbider Störungen und subklinischer Symptomausprägungen	Einschluss komorbider Störungen
störungsspezifische, weitgehend standardisierte Behandlungen durch Manuale, Adherence-Kontrollen (inklusive Rückmeldung an Therapeuten)	Behandlung wie in klinischer Praxis, d.h. ohne Manuale, Adherence-Kontrollen ohne Rückmeldung an den Therapeuten
Therapeutentraining	kein explizites Therapeutentraining
a priori festgelegte Sitzungsanzahl von eher kürzeren Behandlungen	keine festgelegte Sitzungsanzahl
hohe interne Validität	hohe externe Validität

Anmerkung: TAU: Treatment-As-Usual.

Wie bereits angekündigt, soll das Kapitel schließen mit einer Beleuchtung kritischer Stimmen zu den aufgezeigten Methoden der Evidenzbasierung, die die APA Division 12 im Rahmen ihrer EST-Aktivitäten und die der WBP im Rahmen seines Vorgehens in der wissenschaftlichen Anerkennung von psychotherapeutischen Behandlungen nutzte. Es geht also um kritische Stellungnahmen zu und Auseinandersetzungen mit dem, was in der Debatte gemeinhin als "RCT-Paradigma" bezeichnet wird (u.a. Tschuschke, 2005).

Schulte wurde in den letzten Abschnitten bereits in verschiedenen Zusammenhängen mit seinem Hinweis zitiert, dass die Berücksichtigung der externen Validität im Rahmen der Wirk-

samkeitsüberprüfung von Psychotherapie weit wichtiger sei als bspw. in der Wirksamkeitsüberprüfung eines Medikaments (vgl. Deutsches Ärzteblatt, 2008a). Er begründet dies mit der grundlegenden Unterschiedlichkeit zwischen dem, wie Psychotherapie unter den experimentellen Bedingungen eines RCTs, und dem, wie sie in der klinischen Praxis stattfindet. Bestärkt wird Schulte durch zahlreiche Autoren, die auf dieselbe Unterschiedlichkeit und daraus erwachsende beschränkte Generalisierbarkeit von RCTs hinweisen: So wird oftmals auf die in RCTs vorgenommene störungsspezifische Manualisierung der Behandlungen hingewiesen, „wodurch die angewendeten Behandlungsformen nicht denen der Versorgungspraxis entsprechen“ (Leichsenring, 2004a, S. 211). Jenseits des Generalisierungsmoments wird das manualisierte und störungsspezifische Vorgehen in RCTs im Hinblick darauf kritisiert, dass zentrale und für das Gelingen einer Behandlung bedeutende Aspekte unterlaufen und in den Hintergrund gedrängt würden:

> Die Reduktion von Psychotherapie auf die Behandlung von Störungen bedeutet, daß biographische und soziale Hintergründe der Klienten vernachlässigt oder sogar völlig ausgeklammert bleiben und daß die subjektive Bedeutung einer Störung bzw. eines Problems unbeachtet bleibt . . . . Damit wird nicht nur das Konzept des „Klienten“ aufgegeben . . .; auch die Person des Therapeuten rückt in den Hintergrund, wenn es nur noch darum geht, nach den einzelnen Störungen spezifizierte Behandlungen zu „verabreichen“. (Auckenthaler, 2000b, S. 215)

Beutler et al. (2004) geben zudem zu bedenken, dass die Manualisierung in RCTs das intendierte Ziel, die Therapeutenvariable und ihren Einfluss auf das Outcome möglichst zu eliminieren (vgl. Kap. 1.2.2), nicht erreiche. Ferner legen Befunde aus der Prozess-Outcomeforschung nahe, dass neben den jeweils spezifischen Techniken einzelner psychotherapeutischer Verfahren und Methoden den ansatzübergreifenden Wirkfaktoren eine bedeutende Rolle zukommt (u.a. Orlinsky, Rønnestad & Willutzki, 2004; Wampold, 2001). Durch die Fokussierung auf Manuale und auf das damit beabsichtigte Ziel, vor allem die spezifischen

Techniken der unterschiedlichen Therapieansätze prominent zu machen, würde, so die Kritik, die Erkenntnis um die ansatzübergreifenden Faktoren und deren Impact auf den Therapieerfolg vernachlässigt. In dieser Vernachlässigung respektive in der Dominanz von Manualen in RCTs wird nach Wampold (2001) das Verständnis von Psychotherapie in Anlehnung an das sog. "medizinische Modell" sichtbar:

> It is straightforward to understand how the treatment manual is embedded in the medical model. The typical components of the manual – which include defining the target disorder, problem, or complaint; providing a theoretical basis for the disorder, problem, or complaint, as well as the change mechanism; specifying the therapeutic actions that are consistent with the theory; and the belief that the specific ingredients lead to efficacy – are identical to the components of the medical model. (S. 17)

Das medizinische Modell begreift Psychotherapie demnach in erster Linie als etwas, das durch die korrekte Umsetzung von (schul-)spezifischen Faktoren seine Wirksamkeit entfaltet. Im Gegensatz dazu betont das sog. "kontextuelle Modell" – wie der Name schon nahelegt – kontextuelle Faktoren, wie eine vertrauensvolle Beziehung zwischen Therapeut und Patient, positive Erwartungen an die Behandlung seitens des Patienten sowie des Therapeuten, die beidseitige Überzeugung von der gewählten Behandlungsform und deren Erklärungsansatz bzgl. des psychischen Leids und dessen Heilung etc.. Erst durch diese kontextuellen Faktoren können spezifische Techniken ihre Wirkung entfalten. Das kontextuelle Modell negiert also keineswegs die Anwendung bestimmter Techniken und spricht diesen auch nicht grundsätzlich ihre Wirkung ab, jedoch betrachtet es die genannten kontextuellen Faktoren als Bedingungen, die erfüllt sein müssen, damit spezifische Interventionen tatsächlich wirken. Darüber hinaus löst das kontextuelle Modell die Therapieschulengrenzen auf, indem es den Therapeuten nicht festlegt auf die Techniken eines bestimmten Ansatzes: Viel wichtiger ist es dem kontextuellen Modell zufolge, dass Therapeut und Patient sich auf einen ihnen plausibel erscheinenden (Erklärungs-) Ansatz einigen, der in der Behandlung und unter der aktiven Mit-

arbeit des Patienten sodann verfolgt wird (vgl. Auckenthaler, 2012; Wampold, 2001). Eine Psychotherapieforschung bzw. eine Wirksamkeitsevaluation, die in Anlehnung an das medizinische Modell in erster Linie auf die isolierte Überprüfung manualisierter, (schul-) spezifischer Techniken setzt, kann daher dem kontextuellen Modell nicht genügen, da sie zentrale Aspekte von psychotherapeutischen Behandlungen gerade außer Acht lässt. Dies spräche für eine verstärkte Beforschung eben *dieser* Aspekte – z.B. der Therapeuten- und Patientenvariablen – in ihrem Impact auf das Outcome (u.a. Tschuschke, Crameri, Koemeda, Schulthess, von Wyl & Weber, 2009).

Kriz' (2007) Kritik wiederum setzt an den unterschiedlichen Therapieschulen an, indem er aufzeigt, dass das modulare/standardisierte und störungsspezifische Vorgehen eher dem behavioralen Ansatz entspricht, während die psychodynamischen Verfahren und die Gesprächspsychotherapie eher einen verfahrensspezifischen Fokus verfolgten. Bei letzteren Verfahren geht es Kriz zufolge um die Entfaltung von Behandlungsprinzipien und weniger um die manualtreue Anwendung von Behandlungsmethoden (S. 256). Zwischen der RCT-Methodologie, die dem manualisierten/störungsgruppenspezifischen Fokus folgt, und den Verfahren, die eher dem verfahrensspezifischen Fokus entsprächen, würde aus diesem Grunde eine Inkompatibilität bestehen, deren Konsequenzen naheliegen: Ein allein auf der Experimentallogik von RCTs beruhendes Bewertungssystem für die empirische Evidenz unterschiedlicher Psychotherapieverfahren wäre in erster Linie für behaviorale Methoden angemessen und würde andere therapeutische Ansätze – die störungsübergreifend konzipiert sind – in der Bewertung außen vor lassen und benachteiligen.

Die zahlreichen Standpunkte, die sich auf den störungsspezifischen Fokus und den Einsatz von Manualen in RCTs beziehen, sind hiermit ganz sicher nicht erschöpfend dargestellt. Ziel war es jedoch, die Thematik mit zumindest einer gewissen Bandbreite dieser Standpunkte zu

illustrieren und darzulegen, dass ein großer Teil der Kritik sich keineswegs allein auf die fragliche Generalisierbarkeit manualisierter Studien bezieht, sondern weit grundlegendere Argumentationsmuster betrifft.

Um die Manualisierungsdebatte vorerst abzuschließen, sollen noch zwei bedeutende Psychotherapieforscher – Horst Kächele und Peter Fonagy – zu Wort kommen. Beide stammen aus der psychoanalytischen Provenienz und beiden sind – vielleicht überraschend – befürwortende Statements im Hinblick auf Manuale zu entnehmen. So schreibt Kächele:

> The time when only cognitive–behavioral treatments were manualized and psychodynamic treatments were believed not to require these efforts is over. The impact of the evidence-based medicine movement has dramatically changed psychodynamic therapists' view of manuals. (Kächele, 2013, S. 627)

Liest sich aus diesem Zitat noch eine pragmatische Haltung im Hinblick auf die Verwendung von Manualen, so nimmt Fonagy in einem 2009 gehaltenen und ins Deutsche übersetzten Vortrag eine Haltung ein, in der sich der Sinn von Manualen im Rahmen der RCT-Methodologie widerspiegelt:

> Eine solche Spezifizierung der Zutaten des psychodynamischen Zaubertranks wird von vielen . . . als überaus wünschenswert und längst überfällig angesehen. Zwar gibt es Kritiker, die von Reduktionismus sprechen, tatsächlich aber stellen die identifizierten Kompetenzen eine Verbindung zur Evidenz her, weil sie aus den Studien über psychodynamische Kurzzeittherapie als exakt jene Elemente hervorgehen, die gegenüber der Kontrollbedingung überlegen sind. (Fonagy, 2009, deutsche Übersetzung S. 9)

Wichtig scheint in diesem Kontext jedoch die Präzisierung, die Kächele in einem Artikel mit dem Titel *Therapie-Manual: Forschungsmethode und/oder Praxisrealität?* (2010) in Bezug auf Manuale vornimmt: Er betont, dass eine starre Befolgung der psychodynamischen Manualen zu entnehmenden Regeln vollkommen kontraindiziert wäre, ginge es doch vielmehr um

„individuelles Handeln im Rahmen vorgegebener Regeln“ (S. 242). Er spitzt diese Herangehensweise noch zu, indem er schreibt, „ . . . dass in einer guten psychoanalytischen Arbeit die Ausnahme die Regel sei“ (S. 242). Es wird also offenbar, um welche Art Manuale und Manualgebrauch es sich im Rahmen der psychoanalytischen Verfahren handelt, nämlich eher um die Anwendung sog. Behandlungsprinzipien als um die Umsetzung modularisierter Schritt-für-Schritt-Handlungsanweisungen (vgl. Kap. 3.2.2).

Die Kritik an der uneingeschränkten Übernahme der Methoden der EbM auf den psychotherapeutischen Kontext bezieht sich noch auf weit mehr Aspekte, als lediglich auf den störungsspezifischen Fokus und den obligatorischen Einsatz von Manualen. So zieht der Vorgang der Randomisierung Kritik auf sich, die, neben ethischen Bedenken, die Generalisierbarkeit der Ergebnisse in Frage stellt: Die zufällige Zuteilung der Patienten zu unterschiedlichen Bedingungen unterbinde die „intrinsische Verklammerung von Patientenmerkmalen und der Therapiemethode“ (Beutel, Doering, Leichsenring & Reich, 2010, S. 51). Diese Verklammerung und die bewusste Entscheidung für die eine oder die andere Therapieform ist in der klinischen Praxis jedoch möglich, wodurch randomisierte Studien in ihrer Repräsentativität bzw. in ihrer externen Validität in Frage zu stellen sind (u.a. Benecke, 2014a; de Maat, Dekker, Schoevers & de Jonghe, 2007; Seligman, 1995). In die gleiche Richtung der eingeschränkten Generalisierbarkeit zielt die Kritik an den in RCTs meist zusammengestellten Patientengruppen, die hinsichtlich ihres Störungsbildes homogenisiert sind – was nicht zuletzt den störungsspezifisch ausgerichteten Manualen geschuldet ist – und in denen Patienten mit Mehrfachdiagnosen in der Regel ausgeschlossen werden (u.a. Auckenthaler, 2000b). Die eingeschränkte Generalisierbarkeit resultiert aus der einfachen Tatsache, die Westen et al. (2004) durch folgende Zahlen zusammenfassen: Zwischen 50% und 90% der Patienten mit einer Achse-I-Störung weisen mindestens eine weitere Achse-I- oder eine zusätzliche Achse-II-Störung auf. Diesel-

ben Autoren weisen auf ein weiteres die Generalisierbarkeit von RCTs in Frage stellendes Moment hin:

> However, the best available data from both naturalistic and community (catchment) studies suggest that between one third and one half of patients who seek mental health treatment cannot be diagnosed using the *DSM* because their problems do not fit or cross thresholds for any existing category (see Howard et al., 1996; Messer, 2001). (Westen et al., 2004, S. 634)

Damit wirkt sich allein die Tatsache, dass in RCTs nur Patienten aufgenommen werden, deren krankheitswertige Störung und Behandlungsindikation durch DSM- oder ICD-Diagnosen bestätigt wird, hingegen Patienten mit subklinischer Symptomausprägung systematisch ausgeschlossen werden, ebenfalls mindernd auf die externe Validität dieser Untersuchungen aus. Heekerens (2005) gibt darüber hinaus zu bedenken, dass nach anderen Kriterien als der Störungsgleichheit zusammengestellte Untersuchungsgruppen für die Praxis sehr viel fruchtbarer und informativer sein könnten. Er weist auf eine in Großbritannien von Guthrie et al. (1999) publizierte Studie[8] hin, in der Patienten mit unterschiedlichen Störungsbildern mittels psychodynamisch-interpersonaler Psychotherapie behandelt wurden. Gemeinsames Merkmal dieser Patienten war es, dass alle Patienten psychisch chronisch erkrankt waren und bis dato alle psychiatrischen Hilfsangebote scheiterten. In dieser Problemlage erkennt Heekerens eine Art Standardsituation, mit der er praktisch arbeitende Kliniker regelhaft konfrontiert sieht, weswegen er dieser Studie eine besondere klinische Relevanz attestiert.

Neben den bis hier diskutierten Punkten werden im Hinblick auf bestehende Unterschiede zwischen RCTs und der Praxisrealität noch weitere Aspekte angeführt, die nun nur noch kurz umrissen werden sollen: So führt Fonagy (2009) bspw. die in RCTs realisierte kür-

[8] Diese Studie ist Bestandteil des in dieser Arbeit herangezogenen Primärstudiendatensatzes (vgl. Anhang F).

zere Therapiedauer und den von vornherein begrenzten Zeitraum der Behandlung an, ferner den Ausschluss interkurrenter Behandlungen (z.B. Pharmakotherapie) sowie die spezifischen Fähigkeiten und das Engagement der Behandler (etwa durch spezielle Therapeutentrainings). Es sind diese Aspekte, die Fonagy an der Übertragbarkeit der in RCTs ermittelten Wirksamkeit einer Therapie auf die klinische Praxis zweifeln lassen und es sind dieselben Aspekte, die zahlreiche Psychotherapieforscher zu der Forderung veranlassen, die *efficacy*-Forschung durch die *effectiveness*-Forschung zu ergänzen, die eine höhere externe Validität verspricht (u.a. Benecke, 2014a; Benecke, Boothe, Frommer, Huber, Krause & Staats, 2009; Borkovec & Castonguay, 1998; Heekerens, 2005; Seligman, 1995; WBP, 2010).

Wie bereits angeklungen ist, existiert innerhalb der Debatte um angemessene Methoden zur Evidenzbasierung von Psychotherapie ein weiterer Argumentationsstrang, der sich auf kritische Überlegungen jenseits der eingeschränkten Generalisierbarkeit von RCTs stützt (u.a. Wampold, 2001, Kriz, 2007). Vor allem Westen et al. (2004) ist eine bemerkenswerte Arbeit mit dem Titel *The Empirical Status of Empirically Supported Psychotherapies: Assumptions, Findings, and Reporting in Controlled Clinical Trials* über die Inkompatibilität der RCT-Methodologie mit einem Großteil psychotherapeutischer Behandlungen gelungen. In dieser Arbeit loten die Autor/innen die Unvereinbarkeit der RCT-Methodologie mit dem Gegenstand "Psychotherapie" unter Hinzuziehung empirischer Befunde detailliert aus. Sie gehen der Frage nach, welche Annahmen der RCT-Methodologie bzw. der EST-Logik über die Funktionsweise psychotherapeutischen Intervenierens sowie über psychotherapeutische Heilungs- und Veränderungsprozesse psychischer Störungen eigentlich zugrunde liegen. Diese Annahmen werden sodann mit empirischen Befunden aus der epidemiologischen, klinisch-psychiatrischen und Psychotherapieforschung verglichen und die Diskrepanzen zu und Inkompatibilitäten mit diesen RCT-Annahmen aufgezeigt. Eine der von den Autor/innen unter-

suchten Grundannahmen innerhalb der EST-Methodologie bezieht sich bspw. darauf, dass Veränderungs- und Heilungsprozesse von psychischen Störungen innerhalb kurzer Zeit stattfinden können – Westen et al. berichten von Behandlungslängen von ESTs zwischen 6 bis 16 Sitzungen. Dieser Grundannahme stellen sie zahlreiche epidemiologische Befunde und solche aus der Psychotherapieforschung gegenüber, wovon hier nur ein kleiner Ausschnitt wiedergegeben werden soll:

> The malleability assumption is also inconsistent with data from naturalistic studies of psychotherapy, which consistently find a dose–response relationship, such that longer treatments, particularly those of 1 to 2 years and beyond, are more effective than briefer treatments (Howard, Kopta, Krause, & Orlinsky, 1986; Kopta, Howard, Lowry, & Beutler, 1994; Seligman, 1995). Of particular relevance is the finding from naturalistic samples that substantial symptom relief often occurs within 5 to 16 sessions, particularly for patients without substantial personality pathology; however, enduring "rehabilitation" requires substantially longer treatment, depending on the patient's degree and type of characterological impairment (Howard, Lueger, Maling, & Martinovich, 1993; Kopta et al., 1994). (Westen et al., 2004, S. 633)

Resümierend zeigt das Autorenteam auf, dass sich das Vorgehen entsprechend des RCT-Paradigmas im Rahmen der Evidenzbasierung psychotherapeutischer Verfahren in begrenztem Rahmen durchaus als angemessen erweist: Geht es etwa um die Erforschung der Evidenz von modularisierbaren bzw. nach einem genauen Ablaufschema strukturierten, störungsspezifischen Kurzzeitbehandlungen, mittels derer Patienten mit einem relativ eng umgrenzten Störungsbild und ohne bedeutsame strukturelle Beeinträchtigungen (z.B. eine isolierte Phobie) behandelt werden, dann kann – so die Autor/innen – die RCT-Methodologie zweifelsohne als geeignete Beforschungsform angesehen werden. Anders verhält es sich hingegen, wenn es um die Evidenzbasierung von Behandlungen komplexer, komorbider Störungskonstellationen geht, in denen strukturelle Beeinträchtigungen im Rahmen längerfristiger, prinzipiengeleiteter

Behandlungen bearbeitet werden: Dieser Forschungsgegenstand entzieht sich regelrecht den Regeln eines RCTs und damit der EST-Methodologie.

Nach dieser knappen Darstellung der Herangehensweise und der Schussfolgerungen der Gruppe um Drew Westen soll nun also folgender Gedanke für die weitere Argumentation der vorliegenden Arbeit festgehalten werden: Die RCT-Methodologie birgt neben der eingeschränkten Generalisierbarkeit noch ein weiteres Problem in sich, das sich auf eine grundlegende Inkompatibilität zwischen der EST-Logik und dem Gegenstand "Psychotherapie" bezieht. Dieses Problem wird weitestgehend sichtbar bei den Langzeitbehandlungen, die sich schwerlich mit zentralen Vorgaben eines RCTs vereinbaren lassen (u.a. Sandell, 2001). Unter Langzeittherapien werden hier in Anlehnung an den WBP (2008a) Behandlungen mit über 100 Sitzungen verstanden, was gleichsam eine Beschränkung auf die analytische Psychotherapie bedeutet (vgl. Gemeinsamer Bundesausschuss, 2013 sowie Kap. 3.2.2). De Maat, Dekker et al. (2007) widmen diesem Problem einen ganzen Artikel (Titel: *The effectiveness of long-term psychotherapy: Methodological research issues*) und diskutieren im Hinblick auf Kontrollgruppen, wie sie in RCTs gängig sind, deren Unvereinbarkeit mit Langzeittherapiestudien[9]. So steht klassischen Kontrollbedingungen wie Wartelisten- oder Placebo-Kontrollgruppen zum einen die Dauer längerfristiger Behandlungen im Wege, denn sie müssen der Dauer der Therapiebedingung angeglichen werden (vgl. Hager, 2000 sowie Kap. 3.2.1); zum anderen sei eine randomisierte Zuteilung zu diesen Bedingungen fraglich, „because few well-informed patients will participate in a study lasting at least 1 year and offering a 50% chance on no treatment, wait-list, or treatment only resembling psychotherapy“ (de Maat, Dekker et al. 2007, S. 61). Werden als TAU-Bedingung „Behandlungen bei schlecht ausgebildeten Therapeuten/Sozialarbeitern, die eine große Zahl von Patienten zu betreuen

[9] De Maat, Dekker et al. (2007) subsumieren unter Langzeittherapien bereits Behandlungen über 50 Sitzungen bzw. von der Länge eines Jahres oder mehr.

haben“ (Buchholz, 2008, S. 16) angeboten, so stellt diese Bedingung nach de Maat und Kollegen ebenfalls keine akzeptable Behandlungsoption dar, zu der sich Patienten per Zufall zuteilen lassen, vor allem dann nicht, wenn die Alternative etwa eine psychotherapeutische Langzeitbehandlung ist[10]. Auch Fonagy (2009) betrachtet die Randomisierung vor allem in Bezug auf Langzeitbehandlungen als Problem: „Sie [die Forscher] hatten große Mühe, das Problem der Randomisierung zu bewältigen, weil dies bedeutete, dass sich die Teilnehmer bereit erklären mussten, auf die präferierte Behandlung 18 Monate oder länger zu verzichten“ (S. 13). Leichsenring (2011) betont zudem, dass die bewusste Entscheidung für eine Therapie und für einen bestimmten Therapeuten sowie die Passung zwischen Therapeut und Patient vor allem im Hinblick auf (psychodynamische) Langzeitbehandlungen grundlegend sei, was durch eine Zufallszuteilung unterlaufen würde. Vor allem im Zusammenhang mit Langzeitbehandlungen kann eine geplante Randomisierung daher zur Folge haben, dass Patienten diese ablehnen, wie Leichsenring anhand einer schwedischen Studie[11] aufzeigt. Seligman (1995) macht auf eine weitere Schwierigkeit im Hinblick auf die Untersuchung langfristiger Behandlungen nach RCT-Maßstäben aufmerksam, nämlich auf die Standardisierung von Langzeittherapien durch Manuale. Westen et al. (2004) kommentieren dies mit „The longer the therapy, the more variability within experimental conditions . . .” (S. 633) und verleihen damit der Tatsache Ausdruck, dass lang andauernde psychotherapeutische Behandlungen sich dem Versuch einer annähernden Standardisierung durch Manuale, wie er in Behandlungen zwischen 6-16 Sitzungen angestrebt wird, zweifellos entziehen.

[10] Zur Problematik der eindeutigen Definition von „TAU“ vgl. Kapitel . 3.2.4.

[11] Diese Studie nach Sandell, Blomberg und Lazar (1999) ist Bestandteil des in dieser Arbeit herangezogenen Primärstudiendatensatzes (vgl. Anhang F).

Die dargestellte Inkompatibilität zwischen (analytischen) Langzeitbehandlungen und der RCT-Methodologie stellt diejenigen Forscher, die sich mit der Evidenzbasierung derselben Behandlungsform befassen, vor folgende Herausforderung: Die genannten Probleme scheinen den Gegenstandsbereich der Langzeittherapie auf den ersten Blick in das Metier der *effectiveness*-Forschung zu verbannen (keine Randomisierung, keine klassischen Kontrollgruppen, keine Standardisierung durch Manuale; vgl. Tabelle 2). Ein allzu oft bestehendes Missverständnis beruht nun auf der Auffassung, *effectiveness*-Studien würden sich allein der externen Validität verschreiben. Dieses Missverständnis wiederum beruht auf dem Postulat, die interne und die externe Validität stünden in einem disjunkten bzw. inversen Verhältnis zueinander. Dieser Auffassung soll in der vorliegenden Arbeit *nicht* gefolgt werden:

> Man kann und darf interne und externe Validität nicht gegeneinander ausspielen, sie markieren Endpunkte eines methodologischen Kontinuums . . . . Und: Die Sicherung der internen Validität ist nicht ausreichend, aber unabdingbar . . . . Interne und externe Validität stehen in einem bestimmten Spannungsverhältnis, und traditionelle Laborstudien und übliche Feldstudien weisen bedeutsame Unterschiede auf. Wenn man aber das Gebiet der Psychotherapieevaluation unter bestimmten Blickwinkeln in den Blick nimmt, zeigt sich, dass es Möglichkeiten gibt, die Spannungen zu vermindern und die Unterschiede zu verringern. (Heekerens, 2005, S. 360)

Einen ähnlichen Ansatz vertritt Leichsenring (2004a/b) mit seiner wissenschaftstheoretischen Analyse des Verhältnisses von interner zu externer Validität. Leichsenring erstellt in diesem Rahmen zwei getrennte Evidenzhierarchien – eine für die Bemessung der methodischen Qualität von *efficacy*-Studien (er nennt sie „Labor-Studien“) und eine für die Bemessung der methodischen Qualität für *effectiveness*-Studien („Feld-Studien“). In der Evidenzhierarchie für Feld-Studien wird deutlich, dass diese keineswegs allein die externe Validitätssicherung zum Ziel haben sollten; unter Heranziehung des umfassenden Werkes *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* von Shadish et al. (2002) formuliert

Leichsenring dann diverse Strategien, die sich eignen, die interne Validität in quasiexperimentellen Feld-Studien zu sichern (etwa Parallelisierung und Matching, Verwendung zusätzlicher Vergleichsgruppen, Erhebungen zu mehreren Messzeitpunkten, Vorhersage komplexer Ergebnismuster, blinde Ratings durch externe Beobachter, statistische Kontrollen etc.). Mit dieser Konzeption liefert er die Möglichkeit, die Heekerens im obigen Zitat erwähnt, um die Spannung zwischen den beiden Validitätsarten zu verringern: Wirksamkeitsuntersuchungen – ob im Labor oder im Feld – haben grundsätzlich die Möglichkeit, interne Validitätssicherungsstrategien umzusetzen. Diese werden in Abhängigkeit vom Untersuchungsgegenstand („Labor-Therapie“ oder „Feld-Therapie“) unterschiedlich sein. Dieses Konzept lässt sich ebenfalls auf Untersuchungsgegenstände wie die Langzeitbehandlungen anwenden. Es konnte gezeigt werden, dass Langzeitbehandlungen von der „community“ der Psychotherapieforschung offenkundig als Gegenstand betrachtet wird, der sich nur schwerlich mittels RCT-typischer Methoden evidenzbasieren lässt. Auch hier werden andere Strategien notwendig sein, für die Leichsenring (2004a/b) mit seiner Taxonomie unterschiedlicher Evidenzhierarchien gute Hinweise liefert.

Fazit

Es ist davon auszugehen, dass eine zu eng an die RCT-Methodologie angelehnte Begutachtungspraxis einen systematischen Bias im Hinblick auf Langzeittherapiestudien erzeugen würde, indem diese in der Bewertung weitestgehend benachteiligt würden. Da Langzeittherapien unter dem Schutz der Richtlinienverfahren stehen und somit als Kassenleistung durchgeführt werden, ist davon auszugehen, dass es dem WBP ein Anliegen ist, auch – oder vielleicht insbesondere – für diese Behandlungsform eine robuste und qualitativ hochwertige Evidenz zu fordern. Damit verpflichtet er sich gleichsam, mit dem Kriterienkatalog ein Schema vorge-

legt zu haben, das alternative interne Validitätssicherungsstrategien adäquat zu erheben und zu bewerten vermag.

Bevor die darauf ausgerichtete Fragestellung genauer eruiert wird, sollen im nun folgenden Kapitel 1.2.2 die drei Dimensionen des Kriterienkataloge beschrieben werden.

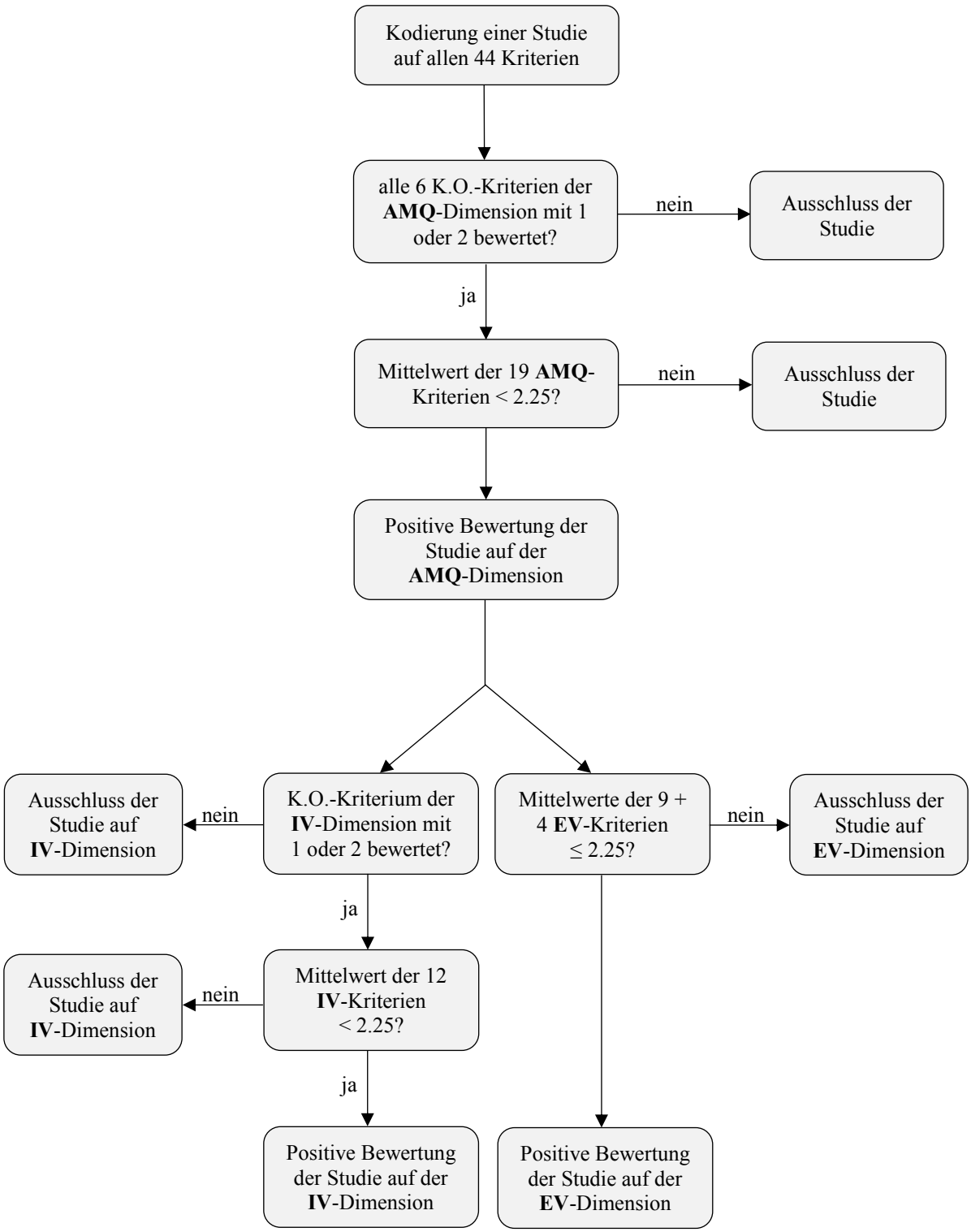
1.2.2 Die Bewertung von Studien durch den Kriterienkatalog des Wissenschaftlichen Beirats Psychotherapie

Wie bereits dargelegt, beruht die wissenschaftliche Anerkennung psychotherapeutischer Verfahren in erster Linie auf einer systematischen Begutachtung der empirischen Evidenz des betreffenden Verfahrens. Die Begutachtung von Studien erfolgt mit Hilfe des standardisierten Kriterienkatalogs des WBP, genauer: mittels der auf den drei Dimensionen der allgemeinen methodischen Qualität, der internen und der externen Validität angesiedelten 44 Kriterien (vgl. Anhang A). Bei dem Kriterienkatalog handelt es sich um ein *kriteriumsorientiertes* – in Abgrenzung zu einem *normorientierten* – Messinstrument (vgl. Klauer, 1987). Das bedeutet, es handelt es sich bei der allgemeinen methodischen Qualität und bei der internen und externen Validität um Qualitätsmerkmale von Studien, deren Ausprägung erst durch die Referenz zu einem festgelegten Standard Bedeutung erlangt (vgl. Herzberg & Frey, 2011, S. 281). Ziel der Erhebung mittels des WBP-Kriterienkatalogs ist es, das Verfehlen oder Erreichen eben dieses Standards – der *Idealnorm* – der methodischen Qualität und Validität einer Studie festzustellen (vgl. Amelang & Zielinski, 2002). Einschränkend muss hinzugefügt werden, dass dieser Maßstab allein für solche Wirksamkeitsstudien bindend ist, die nach 1990 publiziert wurden. Für ältere Studien können liberalere und weniger ausführliche Bewertungsregeln herangezogen werden, die im Folgenden jedoch außer Acht gelassen werden (vgl. WBP, 2010, S. 22).

Die allgemeine methodische Qualität wird anhand von 19 Kriterien bemessen, 12 bemessen die interne Validität, die restlichen 13 Kriterien die externe Validität. Bevor auf die

inhaltliche Gestaltung der einzelnen Dimensionen näher eingegangen wird, soll die Bewertung pro Kriterium bis hin zur Gesamtbewertung einer Studie auf der Grundlage aller 44 Einzelkriterien kurz dargelegt werden: Jedes Kriterium ist mit einem 3-stufigen Antwortformat versehen, wobei jede Stufe mit numerischen Marken sowie mit genaueren Operationalisierungen und ggf. mit Ankerbeispielen versehen ist. Die Gesamtbewertung einer Studie erfolgt – gewissermaßen entsprechend eines Entscheidungsbaumes – über mehrere Schritte, die in Abbildung 3 (S. 59) illustriert sind. Nachdem eine Studie zunächst auf allen 44 einzelnen Kriterien kodiert wurde, wird überprüft, ob die Studie auf den sechs sog. K.O.-Kriterien der allgemeinen methodischen Qualitätsdimension mit „1“ oder „2“ bewertet wurde. Diese K.O.-Kriterien sind im WBP-Kriterienkatalog mit dem Zusatz „Stufe 3 = Ausschlusskriterium“ versehen (vgl. Anhang A) und dementsprechend besonders gewichtet: Schneidet eine Studie auf einem solchen Kriterium mit „3“ ab, so gilt sie unmittelbar als methodisch unzureichend und es kommt zum Ausschluss der Studie. Andernfalls gelangt die Studie zum nächsten Entscheidungsknotenpunkt und es wird geprüft, ob der Mittelwert, der über alle 19 Kriterien der allgemeinen Qualitätsdimension gebildet wird, kleiner als der Wert 2.25 ist. Trifft dies nicht zu, so kommt es ebenfalls zu einem Ausschluss der Studie, andernfalls wird die allgemeine methodische Qualität als erfüllt betrachtet und die Studie erreicht die nächste Stufe der Bewertung – die interne und die externe Validitätsbewertung. Hierbei verfügt lediglich die interne Validitätsdimenion über ein K.O.-Kriterium, das erfüllt sein muss, um auf dieser Validitätsdimension mit einem positiven Ergebnis abzuschneiden. Erreicht eine Studie auf diesem Kriterium nicht eine „1“ oder „2“, so fällt diese Studie auf der internen Validitätsdimension automatisch durch und kann sich allenfalls noch auf der externen Validitätsdimension etablieren. Überwindet die Untersuchung hingegen das K.O.-Kriterium der internen Validitätsdimension, so kommt es auch hier zu einer Mittelwertbildung über alle 12 internen Validitätskriterien und der Studie wird eine hinreichende interne Validität attestiert, wenn der Mittel-

wert unter dem Wert 2.25 liegt. Die externe Validitätsdimension verfügt über keine Hürde im Sinne von K.O.-Kriterien, demgegenüber wird in der Gesamtbewertung der externen Validität eine zweigeteilte Dimensionsbewertung vorgenommen, indem zum einen die Kriterien C.1. bis C.9. und zum anderen die Kriterien C.10. bis C.13. zu Teildimensionen segregiert werden (vgl. Anhang A). Für beide Teildimensionen werden auch hier jeweils die Mittelwerte gebildet, die kleiner oder gleich dem Wert 2.25 sein sollten, damit einer Untersuchung hinreichende externe Validität attestiert werden kann.



Anmerkung: AMQ: Allgemeine methodische Qualität, EV: Externe Validität, IV: Interne Validität.

Abbildung 3: Entscheidungsbaum für die Studienbewertung mittels des WBP-Kriterienkatalogs

Grundsätzlich wird eine Studie nur dann als Wirksamkeitsnachweis für einen Indikationsbereich gewertet, wenn die Voraussetzung einer hinreichenden allgemeinen methodischen Qualität erfüllt ist und die Untersuchung zusätzlich auf mindestens einer der beiden Validitätsdimensionen positiv bewertet werden kann. Der WBP räumt ein: „. . ., doch es gibt auch Studien, die aufgrund ihres Untersuchungsdesigns sowohl Aussagen zur internen, als auch zur externen Validität erlauben“ (2010, S. 7). Mit diesem Zusatz wird das oftmals als invers postulierte Verhältnis von interner und externer Validität aufgeweicht, indem Studien antizipiert werden, denen trotz Umsetzung diverser interner Validitätssicherungsstrategien eine hinreichende Generalisierbarkeit der Schlussfolgerungen und ein hinreichender Praxistransfer bescheinigt werden kann (vgl. Kap. 1.2.1). Die Möglichkeit, dass Studien in der Begutachtung sowohl als hinlänglich intern als auch als extern valide bewertet werden können, eröffnet der WBP durch die dimensionale und separate Gestaltung der beiden Validitätsarten. Eine Studie gelangt somit nicht notgedrungen in eine der beiden Kategorien „experimentell“ *versus* „naturalistisch“, sondern kann gewissermaßen Aspekte der beiden Extrempole auf sich vereinigen. Die Voraussetzung für eine positive Bewertung auf der internen Validiätsdimension bildet jedoch – neben einer positiven Bewertung auf der allgemeinen methodischen Qualitätsdimension – das Erfüllen des K.O.-Kriteriums. Letzteres impliziert, dass allein Studien in Mehrgruppendesigns mit mindestens parallelisierter Gruppenzusammenstellung positiv abschneiden können (vgl. Kriterium B.8.; Anhang A). Die externe Validitätsbewertung formuliert, außer einer positiven Bewertung der allgemeinen methodischen Qualität, keine zusätzlichen Voraussetzungen in Form von K.O.-Kriterien. Daraus folgt, dass auf der externen Validitätsdimension grundsätzlich alle Studiendesigns mit einem positiven Ergebnis abschließen können.

Diese Konzeption der internen und externen Validitätsbemessung ermöglicht es, dass eine Studie sowohl interne als auch externe Validitätsanforderungen erfüllen kann. Der WBP

legt damit ein Konzept der beiden Validitätsarten vor, das mit (wissenschafts-) theoretischen Herangehensweisen an die beiden Validitätsarten durchaus übereinstimmt (vgl. Kap. 1.2.1).

Die Dimension der allgemeinen methodischen Qualität

Auf der im Hinblick auf die Kriterienanzahl umfassendsten Dimension werden methodische Themen unterschiedlicher Bereiche abgefragt, die der Einteilung im WBP-Kriterienkatalog entsprechend vorgestellt werden sollen (vgl. Anhang A). Die Gesamtbewertung der allgemeinen methodischen Qualität einer Studie reicht jedoch über die dimensionseigenen Kriterien hinaus, da für die Gesamtbewertung von den insgesamt sechs K.O.-Kriterien drei nicht auf der methodischen Qualitätsdimension angesiedelt sind. Ein K.O.-Kriterium (B.12.) stammt von der internen Validitätsdimension, zwei (C.1. und C.9.) von der externen Validitätsdimension. Diese drei Kriterien sollen daher ihrer eigentlichen Dimensionszugehörigkeit entsprechend beschrieben werden, so dass im Folgenden zunächst nur auf die 19 Kriterien der methodischen Qualität Bezug genommen wird.

Manipulation der Daten: Diese Rubrik umfasst nur ein Kriterium (K.O.-Kriterium A.1.), das als absolute Ausnahme lediglich auf zwei Kriterienstufen bewertet werden kann: Dieses Kriterium wird mit „1“ bewertet, wenn es keine Hinweise auf Ergebnismanipulationen gibt und mit „3“, wenn es begründeten Anlass zur Annahme gibt, dass die Daten/Ergebnisse der Studie manipuliert wurden. Indikatoren für Manipulationen könnten bspw. uneinheitliche Ergebnisdarstellungen oder variierende Stichprobengrößen über mehrere Publikationen hinweg sein, deren Zustandekommen nicht nachvollziehbar ist. Wird eine Studie hier mit „3“ kodiert, so kommt es zu einem Ausschluss der Studien. Aufgrund dieser drastischen Folge wird man als Gutachter dieser Studie nicht leichtfertig diese Bewertung wählen, sondern in der Regel die Autoren der Studie kontaktieren, um Ungereimtheiten aufzuklären. Darüber hinaus ist davon

auszugehen, dass tatsächliche Manipulationen derart umgesetzt werden, dass sie den Publikationen nicht zu entnehmen sind, so dass der Verdacht auf Ergebnismanipulation schwer zu begründen sein wird. Eine tatsächlich negative Bewertung auf diesem Kriterium wird daher äußerst selten und nur bei triftigen Gründen zu verantworten sein.

Patienten: In Bezug auf die Studienpatienten wird der Standardisierungsgrad der Diagnosestellung erfragt (K.O.-Kriterium A.2.) sowie die Dropoutrate zwischen der Prä- und der Postmessung (Kriterium A.3.) und zwischen Post- und Katamnesemessung (Kriterium A.4.).

In einer methodisch qualitativ angemessenen Studie – darunter wird hier und im Folgenden eine Studie verstanden, die auf einem Kriterium mit Stufe „2“ oder besser abschneidet – sollte die Diagnosestellung mindestens mittels sog. Diagnosechecklisten erfolgen oder aber das klinische Urteil sollte anderweitig nachvollziehbar sein (K.O.-Kriterium A.2.). Bestenfalls sollten die Diagnosen mittels standardisierter Verfahren erhoben worden sein (entspräche Stufe „1“).

Die Dropoutquote über alle Treatmentarme zwischen Prä- und Postzeitpunkt sollten 40% nicht übersteigen (Kriterium A.3.); für die Dropoutquoten zwischen Post- und Katamnesezeitpunkt (Kriterium A.4.) wird keine genaue Prozentangabe gemacht, sondern auf „Studien mit vergleichbaren Patientengruppen und vergleichbarem Katamnesezeitraum“ (WBP, 2010, S. 30) hingewiesen. Das bedeutet, für die Kodierung einer Studie sind entsprechend des darin realisierten Katamnesezeitraums und entsprechend der darin untersuchten Patientengruppe sog. Referenzstudien heranzuziehen und die Dropoutrate in der betreffenden Studie mit diesen Referenzen zu vergleichen (dazu detailliert: Kap. 3.2.3). Weichen die Dropoutraten in der Studie von denen in Referenzuntersuchungen deutlich nach oben hin ab, so fällt dies im Hinblick auf die methodische Qualität negativ ins Gewicht (Stufe „3“). Liegen die Dropoutra-

ten anzahlmäßig deutlich unter denen der Referenzstudien, so fällt dies folglich positiv ins Gewicht (Stufe „1“).

Für die methodische Qualität sind die zuletzt genannten Aspekte insoweit von Belang, als dass sehr hohe Dropoutquoten die Aussagekraft einer Untersuchung schmälern, da die Outcomes von Studien- und/oder Therapieabbrechern allenfalls noch geschätzt werden können (durch Methoden wie *last observation carried forward*, regressionsstatistische Schätzmethoden etc.; vgl. Mayer, 2010). Der Standardisierungsgrad diagnostischer Instrumente beeinflusst die Objektivität und Reliabilität und damit auch die Validität von Diagnosen, womit die methodische Qualität einer Untersuchung derart in Frage gestellt ist, dass der WBP dieses Kriterium (A.2.) zu einem K.O.-Kriterium erhebt: Eine Studie mit inadäquater Diagnosestellung, die keinen eindeutigen Rückschluss auf die Zusammensetzung der untersuchten Patientengruppe im Hinblick auf die zu behandelnden Störungen zulässt, fällt durch die Begutachtung der allgemeinen methodischen Qualität durch.

Studiendesign: Hinsichtlich des Studiendesigns wird die Stichprobengröße abgefragt (Kriterium A.5.) sowie der Zeitpunkt, an dem festgelegt wurde, welches Messdesign vorgenommen wird und ggf. welche Behandlungsgruppen miteinander verglichen werden (Kriterium A.6.).

In einer methodisch adäquaten Studie sollten die Stichproben pro Treatmentarm nicht kleiner als n=10 sein. Zudem sollte die Festlegung der Vergleichsgruppen und der Messzeitpunkte bestenfalls a priori – also vor Beginn der Untersuchung – erfolgt sein (Stufe „1“); ist hingegen bspw. ein Teil der Erhebungszeitpunkte erst a posteriori – also im Verlauf oder nach Abschluss der Studien – bestimmt worden, so wird eine solche Studie mit „2“ kodiert. Eine retrospektive Studie erhält hier automatisch eine Bewertung mit „3“. Retrospektive Untersuchungen gelten in der Regel als anfällig gegenüber solchen Einflussnahmen, durch die gewissermaßen Wunschergebnisse produziert werden können (z.B. Auswahl bestimmter Probanden

oder bestimmter Outcomemaße zu wiederum bestimmten Erhebungszeitpunkten aus einem bereits vorliegenden Datensatz).

Outcomemessung: Ähnlich der a priori Festlegung der Messzeitpunkte und der Vergleichsbedingungen (s.o.) fordert der Kriterienkatalog mit Kriterium A.7. die a priori Festlegung der primären und sekundären Zielkriterien. Ferner sollten die Zielkriterien mit Kriterium A.8. mittels reliabler und valider Messinstrumente erhoben werden. Ebenso wie nicht reliable Diagnosestellungen (s.o.) führt der Einsatz nicht reliabler und valider Erhebungsverfahren für Outcomemessungen zum Ausschluss der Studie (dazu ausführlich: Kap. 3.2.3). Mit Kriterium A.9. sollten in einer Untersuchung neben der gruppenstatistischen Signifikanz auch Maße der klinischen Bedeutsamkeit erhoben werden. Die klinische Bedeutsamkeit bildet dabei einen Oberbegriff für unterschiedliche Strategien, die Wirksamkeit einer Intervention auch auf *individueller* Ebene abzubilden. Auf dieses methodische Gütekriterium weist die APA in den von ihr verfassten *Criteria for Evaluating Treatment Guidelines* (2002) mit folgenden Worten hin:

> *Clinical significance.* Ideally, outcome descriptions should specify clinical significance (i.e., actual clinical benefit) in addition to reporting any statistical significance. The full range of responses to the intervention should be reported, including such outcomes as (a) functioning within normal limits, (b) much improved but not functioning within normal limits, (c) improved, (d) no change, and (e) deterioration. The mandate for a particular intervention is enhanced if it normalizes functioning. (S. 1055)

Die APA bezieht sich damit auf die Berechnung des *Reliable Change Indexes* (RCI[12]) nach Jacobson und Truax (1991) sowie auf den Vergleich des individuellen Messwertes zum Postzeitpunkt mit normgebundenen Cutoff-Werten, die eine Grenzlinie zwischen dem funktionalen und dem dysfunktionalen Bereich markieren. Die Wichtigkeit dessen, sich in Untersu-

[12] Oftmals auch mit "RC" abgekürzt.

chungen nicht allein auf gruppenstatistische Daten, wie Mittelwertvergleiche, zu verlassen, sondern diese zu kombinieren mit individuumsbezogenen Daten, betonen auch Westen et al., (2004):

> A treatment that has an enormous effect in 20% of patients can appear superior to another treatment that has a smaller but clinically meaningful impact on 90% of patients. These caveats are not meant to "demean the mean," or to devalue probability statistics, only to suggest that mean differences and their corresponding significance values and effect size estimates provide only one measure of efficacy. (S. 644)

Insoweit sollte eine Studie nach Kriterium A.9. in jedem Fall Maße der klinischen Bedeutsamkeit berichten. Ein Überblick über weitere klinische Relevanzmaße neben den hier bereits angerissenen sowie über die genaue Spezifikation der einzelnen Kriterienstufen wird in Kapitel 3.2.3 gegeben.

Mit den beiden Kriterien A.10. und A.11. fordert der WBP, dass Outcomemaße möglichst auf der Grundlage mehrerer Informationsquellen erhoben werden sollten. Neben Instrumenten zur Selbstbeurteilung (Patientenperspektive) sollte mindestens noch eine weitere Informationsquelle hinzugezogen werden (Stufe „2" auf Kriterium A.10.). Diese Fremdbeurteilung kann durch den Therapeuten oder durch unabhängige Rater erfolgen. Bestenfalls sollte die Erhebung jedoch aus drei (oder mehr) Perspektiven erhoben werden (Patient, Therapeut, externer Rater). Warum die multiperspektivische Erhebung von Bedeutung ist, fassen Hill und Lambert (2004) folgendermaßen zusammen:

> The now necessary and, to some degree, common practice of applying multiple criterion measures in research studies has made it obvious that multiple measures from different sources do not yield unitary results . . . . The lack of consensus across sources of outcome evaluation, especially when each source is presumably assessing the same phenomena, has been viewed as a threat to the validity of data. How-

> ever, outcome data provide not only evidence about changes made by the individual, but also information about the differing value orientations and motivations of the individuals providing the data (S. 113f.)

Ähnlich dem Einsatz unterschiedlicher Maße zur Abbildung von Wirksamkeit (statistische Signifikanz und klinische Bedeutsamkeit; s.o.) oder dem Einsatz unterschiedlicher Zielkriterien (symptombezogen, interpersonal, Persönlichkeitsfragebögen etc.) sorgen also auch multiperspektivische Erhebungen für ein genaueres und ganzheitlicheres Bild über die Effektivität einer psychotherapeutischen Maßnahme. Werden in einer Untersuchung Fremdeinschätzungsverfahren durch unabhängige Beurteiler eingesetzt, so sollten diese – zwecks Reliabilität der Ratings – in dem von ihnen angewandten Instrument trainiert sein (Kriterium A.11.). Bei Mehrgruppendesigns ist es zudem wichtig, dass die externen Rater blind für die Gruppenzugehörigkeit der zu beurteilenden Probanden sind, so dass evtl. *allegiance effects* ausgeschlossen werden können (vgl. Lambert & Ogles, 2004).

Kriterium A.12. bezieht sich auf die Vollständigkeit der Ergebnisdarstellung: Es sollten die Ergebnisse für alle in einer Studie angesetzten Zielkriterien zu allen realisierten Messzeitpunkten tatsächlich berichtet werden. Andernfalls könnte der Verdacht auf „Schönung“ der Daten aufkommen – etwa derart, dass zwar die Ergebnisse über alle Outcomemaße zum Postzeitpunkt dargestellt werden, jedoch aufgrund mangelnder Stabilität einzelner Outcomes nur noch vereinzelt zur Katamnesemessung.

Das letzte Kriterium (A.13) der Rubrik „Outcomemessung“ bezieht sich auf einen eher stiefmütterlich behandelten Bereich der Outcomeforschung – nämlich auf den Bereich der unerwünschten Wirkungen psychotherapeutischer Maßnahmen. Haupt und Linden (2011) berichten in einem zur *Fehlerkultur in der Psychotherapie* veröffentlichten Themenheft über Häu-

figkeiten zwischen 3-15% der Behandlungsfälle, bei denen es zu unerwünschten respektive negativen Wirkungen kommt. Sie definieren Nebenwirkungen als eben diese unerwünschten/negativen Ereignisse, die therapiebedingt auftreten. Zentral ist dabei, dass diese Nebenwirkungen als negative Wirkungen einer *korrekt* durchgeführten Behandlung auftreten können. Treten Nebenwirkungen hingegen aufgrund einer unkorrekt durchgeführten Behandlung auf, so spricht man von „Kunstfehler". Zudem können Nebenwirkungen auch bei Erreichen sonstiger erwünschter Therapieziele auftreten, was die Autoren dazu veranlasst, Nebenwirkungen als „eigenständiges Phänomen" (S. 14) zu betrachten, das es daher gesondert zu erheben gilt. Zu den zu erhebenden unerwünschten Ereignissen zählen sie bspw. unzureichende Therapieergebnisse, eine Therapieverlängerung oder das Auftreten neuer Symptome, aber auch Änderungen in der familiären Situation oder Probleme im Beruf. Dimidjian und Hollon (2010) nennen in Anlehnung an die *Food and Drug Administration* und deren Definition unerwünschter Ereignisse weitere (extrem) negative Vorkommnisse, wie Suizid, Inhaftierungen, Krankenhausaufenthalte, Unfälle, schwere somatische Erkrankungen, die im Rahmen von Psychotherapiestudien systematisch miterhoben werden sollten, und zwar „. . . even though it is unlikely that such outcomes are attributable to the intervention" (S. 26). Damit sprechen sie einen zentralen Aspekt in der Eruierung von Nebenwirkungen an, nämlich das Vorkommen und die Erhebung aversiver Ereignisse während einer Behandlung einerseits und die Feststellung der tatsächlichen Therapiebedingtheit andererseits. Erst, wenn ein Bezug zur Therapie hergestellt werden kann, ist tatsächlich von Nebenwirkungen zu sprechen. Für die Feststellung der Behandlungsbedingtheit legen Haupt und Linden (2011) ein umfassendes Schema vor – das *Ereignis-Context-Relation-Schwere Schema – adverse treatment reaction* (ECRS-ATR) (vgl. auch Linden, 2013). In diesem Schema werden typische, primär verhaltenstherapeutische Therapieprozesse als Ankerbeispiele beschrieben, die prädestiniert sind – selbst bei korrekter Umsetzung der Behandlung –Nebenwirkungen hervorzurufen. Vor dem Hinter-

grund der beschriebenen Kontexte sieht das von den Autoren konzipierte, standardisierte Klassifikationsmuster ein Wahrscheinlichkeitsrating vor: Es wird die Wahrscheinlichkeit eingeschätzt, mit der ein unerwünschtes Ereignis durch eben diesen oder jenen therapeutischen Prozess (Kontext) hervorgerufen worden sein könnte. Mit diesem System haben die Autoren eine Vorgehensweise geschaffen, mit der Nebenwirkungen tatsächlich als ein eigenständiges Phänomen, jenseits der intendierten positiven Behandlungswirkung, erfasst werden können.

Der am häufigsten zu findende und gleichzeitig der pragmatischste Zugang zum Thema „Nebenwirkungen von Psychotherapie“ ist u.a. bei Dimidjian und Hollon (2010) sowie bei Barlow (2010) zu finden, der die „more individual idiographic approaches“ (S. 13) betont. Dieser Zugang findet sich wieder in dem bereits beschriebenen Konzept der klinischen Bedeutsamkeit (s.o.), mit dem die Erhebung individueller Veränderungs- und Zielerreichungswerte gemeint ist. Barlow führt dafür für die Major Depression folgende „*5 Rs*“ an:

- *response* (Ansprechen), als Maß bedeutsamer Reduktion der Symptomschwere (bspw. Reduktion um 50%)
- *remission* (Remission), definiert als vollständige *response* und Erreichen des funktionalen Bereichs (z.B. Hamilton Depression Score [HAM-D-Score] < 7; vgl. Laux, 2008a)
- *recovery* (Heilung) bezeichnet ein Aufrechterhalten des erreichten Zustands der Remission über mehrere Monate (in der Regel 6 Monate)
- *relapse* (Rückfall) steht für einen Rückfall in der Periode zwischen *remission* und *recovery*
- *recurrence* (Wiederauftreten) steht für eine erneute depressive Episode nach Erreichen der vollständigen Heilung.

Barlow empfiehlt für die systematische Erhebung negativer Therapieeffekte die gezielte Beobachtung von auffällig langsamem Ansprechen auf die Behandlung (Einsetzen vergleichs-

weise später *response*), von geringen Remissions- oder Heilungsraten sowie von hohen Rückfall- und Wiederauftretensraten. Diese sollten sodann im Vergleich zu Daten aus Studien zum natürlichen Verlauf der betreffenden Störung[13] bewertet werden. Darüber hinaus empfehlen Dimidjian und Hollon (2010), diejenigen Subgruppen, bei denen Verschlechterungen oder keinerlei Veränderungen in den Zielkriterien sichtbar werden, genauer zu spezifizieren und deren Pendant ggf. in den alternativen Therapiebedingungen zu suchen. Treten bei diesen Patienten in den alternativen Therapiebedingungen keine vergleichbaren Verschlechterungen auf, so liegt es den Autoren zufolge nahe, dass es sich um therapiebedingte Wirkungen handeln muss[14]. Auch Barlow betont die genauere Charakterisierung dieser Patientengruppe:

> . . ., a thorough-going analysis at a more individual level of who might experience adverse effects for one reason or another and why. (2010, S. 16)

Eine qualitativ „gute" Wirksamkeitsstudie sollte daher – wenn sie Nebenwirkungen nicht derart systematisch erhebt, wie es bspw. das Klassifikationsschema ECRS-ATR nach Haupt und Linden (2011) fordert – Verschlechterungen und *Non response* in unterschiedlichen Outcomemaßen erheben und berichten. Um ein genaueres Bild darüber zu vermitteln, bei wem die Behandlung möglicherweise negative oder ausbleibende Effekte hervorrufen könnte, sollte die eruierte Patientengruppe zudem näher charakterisiert werden.

Statistische Methodik: Die letzte Rubrik der allgemeinen methodischen Qualitätsdimension bezieht sich nun zum einen auf die in den Studien verwendeten statistischen Auswertungsver-

[13] Auf den natürlichen Verlauf von Störungen wird im Zusammenhang mit den sog. störungsangemessenen Katamnesezeiträumen noch detailliert eingegangen (vgl. Kap. 3.2.3).

[14] Dieser Ansatz wirft aufgrund der verhältnismäßig geringen Baseline an Patienten mit Verschlechterungen und persistierender Symptomatik einige methodische Probleme auf, auf deren Darstellung und Diskussion an dieser Stelle jedoch verzichtet werden soll.

fahren. Zum anderen werden Qualitätsmerkmale, wie die Durchführung von *Intent-To-Treat* Analysen (ITT-Analysen), von Dropoutanalysen sowie von Poweranalysen abgefragt.

Die ersten beiden Kriterien dieser Rubrik (A.14. und A.15.) beziehen sich auf statistische Auswertungen im engeren Sinne: Danach sollte eine methodisch angemessene Studie die Anwendungsvoraussetzungen für die verwendeten inferenzstatistischen Verfahren überprüft haben (Kriterium A.14.). Außerdem wird gefordert, dass die angewandten Verfahren dem vorliegenden Rohdatenmaterial sowie der Fragestellung der Untersuchung angemessen sind und korrekt umgesetzt werden (Kriterium A.15.). Zu Letzterem zählt bspw. die α-Fehler-Korrektur bei multiplen Tests oder die korrekte Entscheidung zwischen parametrischen und nonparametrischen Prozeduren. Zu überprüfende Anwendungsvoraussetzungen (Kriterium A.14.) wären z.B. die Sphärizität[15] bei Varianzanalysen mit Messwiederholung und ggf. die Korrektur bei Verletzung der Annahme. Annahmen über die Homoskedastizität oder Normalverteilungsannahmen, bspw. bei t-Tests, sollten in methodisch angemessenen Studien vor allem bei ungleich großen Vergleichsgruppen und bei kleinen Stichproben überprüft werden.

Die drei Kriterien A.16., A.18. und A.19. beziehen sich nun auf Dropout- und ITT-Analysen: Man unterscheidet im Hinblick auf Wirksamkeitsaussagen hauptsächlich zwei Bezugsgruppen: 1. die Therapie- und Studiencompleter und 2. die zum Zeitpunkt der Randomisierung oder des Therapiebeginns eingeschlossenen Patienten (vgl. Kleist, 2009; Lambert & Ogles, 2004). Bei einer von Kendall und Kollegen (2004) postulierten Studiendropoutrate von ca. 20% liegt es nahe, dass das sog. *full analysis set* stets die größere Bezugsgruppe darstellt. Verglichen mit der Effektivitätsschätzung auf Grundlage des Completersamples stellt die Schätzungen der Effektivität auf Basis des *full analysis sets* die konservativere Schätzung dar. Die Frage, der die Auswertung über das *full analysis set* nachgeht, lautet folgendermaßen:

[15] Sphärizität bedeutet, dass die Varianzen der Differenzvariablen (Messwertdifferenzen zwischen den Messzeitpunkten) gleich sind.

„What are the effects of treatment for people seeking treatment for a certain problem?“. Im Vergleich dazu beantwortet die Auswertung über das reine Completersample folgende Frage: „What are the effects when someone completes treatment?“ (Kendall, Holmbeck & Verduin, 2004, S. 28). ITT-Analysen bezeichnen nun genau die Auswertung über das *full analysis set* und bedienen sich dabei unterschiedlicher Methoden zur Schätzung von Outcomewerten derjenigen, die die Studie vor Beendigung abgebrochen haben (vgl. u.a. Mayer, 2010).

Eine qualitativ „gute“ Studie sollte die Studien-/Therapieabbrecher zum einen genau beschreiben (Kriterium A.18.), zudem sollten Dropoutanalysen durchgeführt werden (Kriterium A.19.). Zur Beschreibung der Dropouts gehören neben allgemeinen Merkmalen der Zeitpunkt des Abbruchs sowie die Gründe für diesen Abbruch. Hiller, Bleichhardt und Schindler (2009) unterscheiden hinsichtlich der Abbruchgründe zwischen „potenziell oder sicher qualitätsrelevanten Abbrüchen“ und „nicht-qualitätsrelevanten Abbrüchen“. Zu den qualitätsrelevanten Abbrüchen zählen sie u.a. Gründe, wie fehlende Therapiemotivation seitens des Patienten, Aufsuchen einer anderen Behandlung, geringer Therapieerfolg, Ablehnung der Kostenübernahme durch Krankenkasse, aber auch, wenn keine Gründe vom Patienten angegeben werden. Nicht-qualitätsrelevante Gründe sind hingegen Mutterschaft der Therapeutin, Wohnortwechsel des Patienten etc.. ITT-Analysen (Kriterium A.16.) werden somit umso bedeutsamer, je mehr Abbrüche aus qualitätsrelevanten Gründen vorliegen, andernfalls würde der Therapieeffekt deutlich überschätzt. Daher ist die Erhebung von Abbruchgründen essentiell. Dropoutanalysen (Kriterium A.19.) beziehen sich in Mehrgruppendesigns auf den Zwischengruppenvergleich der Abbrecher im Hinblick auf prognostisch bedeutsame Faktoren (z.B. Geschlecht, Bildungsgrad, Strukturniveau und Symptomschwere zur Baseline). In einer methodisch adäquaten Studie werden diese potentiellen Zwischengruppenunterschiede überprüft und ggf. statistisch korrigiert.

Kriterium A.17. der Rubrik „Statistische Methodik“ bezieht sich allein auf Mehrgruppendesigns, in denen ein Psychotherapieverfahren mit einem bereits etablierten Treatment verglichen wird. Für diese Zwecke sollten in der Regel sog. *Non-Inferiority*-Hypothesen überprüft werden, deren Annahme (H_1) es ist, dass die zu überprüfende Therapieform im Vergleich zum etablierten Treatment mindestens gleichgut oder maximal unmaßgeblich unterlegen abschneidet (vgl. Klemmert, 2004; ausführlich: Kap. 3.2.3). Grundsätzlich geht es bei Vergleichen mit etablierten Treatments um die inferenzstatistische Absicherung minimaler Effekte, die nur dann mit einer genügend großen Power erfolgen kann, wenn dementsprechend große Stichproben zur Verfügung stehen. Eine methodisch angemessene, komparative Wirksamkeitsstudie[16] sorgt daher über hinreichend große Stichproben für eine ausreichend hohe Power (mindestens .50-.80).

Die Dimension der internen Validität

Die Dimension der internen Validität fußt auf insgesamt 12 Kriterien, die sich auf vier Rubriken aufteilen (vgl. Anhang A). Ein Kriterium (B.8.) fungiert als K.O.-Kriterium, das jedoch nicht zu einem absoluten Ausschluss der Studie führt, sondern lediglich zu einem Ausschluss der Studie auf der internen Validitätsdimension.

Patienten: Die Rubrik „Patienten“ umfasst zwei Kriterien (B.1. und B.2.). Beide beziehen sich auf die in einer Untersuchung aufgestellten Ein- und Ausschlusskriterien. Kriterium B.1. fordert die genaue Spezifizierung von Ein- und Ausschlusskriterien, Kriterium B.2. die valide Erhebung derselben. Eine Behandlung, die für eine bestimmt Patientengruppe konzipiert wur-

[16] Unter komparativen Wirksamkeitsstudien werden Untersuchungen verstanden, in denen zwei Alternativtherapien miteinander verglichen werden.

de (schwere strukturelle Störungen, Traumapatienten, unipolare Depression o.ä.), sollte ihre Wirksamkeit im Rahmen einer Untersuchung bei genau dieser Gruppe unter Beweis stellen. Dementsprechend müssen Ein- und Ausschlusskriterien dafür sorgen, dass eine solche Gruppe zwecks Untersuchung rekrutiert und zusammengestellt wird. Eine Studie, in der die Merkmale, nach denen Patienten in die Untersuchung ein- oder ausgeschlossen werden, nicht eindeutig spezifiziert und valide erhoben werden, ist anfällig gegenüber der Inklusion nicht intendierter Patientenmerkmale (z.B. nicht intendierter Diagnosegruppen, Symptomschweregrade etc.). Diese nicht intendierten Merkmale geraten somit unentdeckt in den Gesamtpool der Patienten und könnten dort – ebenso unentdeckt – ungleich verteilt in die unterschiedlichen Untersuchungsgruppen gelangen. Dies kann selbst bei randomisierter Gruppenzuweisung geschehen, denn: Randomisierung soll zwar die Wahrscheinlichkeit maximieren, dass sich alle persongebundenen Störvariablen auf die unterschiedlichen Treatmentbedingungen gleichverteilen, einen Garanten stellt sie jedoch nicht da: „Der Nachteil ist, dass der Zufall nicht immer vertrauenswürdig ist“ (Eid et al., 2010, S. 59). Vor allem bei kleinen Stichproben ist die Gleichartigkeit der Gruppen durch randomisierte Zuweisungsstrategien in Zweifel zu ziehen (vgl. Hsu, 1989).

Ungleich verteilte patientenbezogene Variablen, die durch mangelhaft spezifizierte Ein-/Ausschlusskriterien und durch nicht valide Messung der Kriterien unentdeckt in den Patientenpool gelangt sind, könnten somit eine Alternativerklärung für einen beobachteten Zwischengruppeneffekt (bei Unterschiedshypothesen) bzw. für einen nicht zustande gekommenen Zwischengruppeneffekt (bei Äquivalenzhypothesen) darstellen. Eine intern valide Studie hingegen verhindert durch genaue Spezifizierungen und valide Erhebungen der Ein-/Ausschlusskriterien die Aufnahme nicht intendierter und unentdeckter Patientenvariablen und somit mögliche Konfundierungen, die die interne Validität gefährden.

Intervention: Die zweite Rubrik der internen Validitätsdimension umfasst fünf Kriterien zum Bereich „Intervention“. Einige davon beziehen sich jeweils auf denselben Aspekt, z.B. auf die Umsetzung der Kontrollbedingung, und werden im Folgenden entsprechend zusammenhängend beschrieben.

Die beiden Kriterien B.3. und B.6. beziehen sich auf die Realisierung der psychotherapeutischen Interventionen in der Experimentalgruppe sowie, bei komparativen Studien, in der Vergleichsgruppe: Kriterium B.3. fragt danach, ob das therapeutische Vorgehen des jeweiligen Verfahrens in der Studie in einer Weise operationalisiert ist, dass es zwischen den behandelnden Therapeuten einer Treatmentgruppe vergleichbar und zudem in weiteren Untersuchungen replizierbar ist. Wird in einer Studie auf ein Manual bzw. auf manualähnliche Behandlungsrichtlinien[17] verwiesen und werden diese auch offenkundig im Rahmen der Untersuchung verwendet, so wird diese Studie auf Kriterium B.3. mit Stufe „1“ bewertet. Mit einer „3“ wird eine Untersuchung hingegen bewertet, wenn das Psychotherapieverfahren in der Publikation lediglich benannt wird und weitere Ausführungen zum Vorgehen fehlen. Die Verwendung von Manualen wird demnach vom WBP als Merkmal der internen Validitätssicherung betrachtet. Dies wird am besten durch den experimentallogischen Begriff der „bedingungsgebundenen Störvariablen“ verdeutlicht, der sich auf solche Störfaktoren bezieht, die mit den unterschiedlichen experimentellen Bedingungen (Treatments) selbst verbunden sind (u.a. Eid et al., 2010). Bezogen auf den psychotherapeutischen Kontext sind dies nach Fydrich und Schneider (2007) Therapeutenvariablen sowie Interaktionsvariablen (Therapeutenvariablen x Patientenvariablen), deren Einfluss auf das Outcome durch den Einsatz von Manualen möglichst minimiert werden soll. Da eine randomisierte Zuteilung von Therapeuten zu den unterschiedlichen Treatmentarmen in der Regel unüblich und nicht möglich ist, muss die

[17] Vgl. Kapitel 3.2.2.

Kontrolle therapeutengebundener Störvariablen mittels anderer Strategien erfolgen, indem man den Behandlungsprozess möglichst standardisiert, „so that therapy factors could be detected and extracted from the "noise" of therapist and patient factors (Nathan, Stuart, & Dolan, 2000)" (Beutler et al., 2004, S. 245). Durch Manuale und die damit angestrebte Standardisierung soll das therapeutische Vorgehen also in seiner möglichst „reinen" Form umgesetzt werden, um so die genannten Störfaktoren zu kontrollieren. Andernfalls wäre bspw. eine potentielle Alternativerklärung für einen entdeckten Unterschied zwischen zwei Treatments eben ein systematischer Unterschied im „*noise*" respektive im „Störfaktor Therapeut" und seiner Interaktion mit dem Klienten. Allerdings wird der Standardisierungsgrad in Abhängigkeit von der Länge, auf die eine Behandlung angelegt ist, sowie vom Konzept des therapeutischen Verfahrens durchaus variieren (dazu ausführlich: Kap. 3.2.2).

Kriterium B.6. bezieht sich auf die Adherence-Kontrolle bzw. auf die Kontrolle der sog. *Treatment Integrity*. Das bedeutet, es sollte im Rahmen einer Studie kontrolliert und sichergestellt werden, dass die Interventionen, die für ein Verfahren spezifisch sind, konstruktvalide umgesetzt werden, demgegenüber sollten Schlüsselkomponenten der Vergleichsbehandlung ausgeschlossen werden (vgl. Shadish et al., 2002). Dieses Gütemerkmal einer Studie hat nur bedingt etwas mit dem Einsatz von Manualen zu tun, auch wenn der WBP den Begriff *Treatment Integrity* mit *Manualtreue* gleichsetzt: Zwar ist die *Treatment Integrity* und damit die Konstruktvalidität durch den Einsatz von Manualen eher zu sichern und zudem einfacher zu kontrollieren. Jedoch kann die *Treatment Integrity* selbst in Untersuchungen, in denen unmanualisierte Behandlungen durchgeführt werden, durchaus überprüft werden. So legte die Arbeitsgruppe um Hilsenroth (2005) mit ihrer *Comparative Psychotherapy Process Scale* (CPPS) ein Adherence-Messinstrument mit der folgenden Begründung vor:

> The CPPS is not intended to replace existing manual specific instruments but rather is intended to offer a reliable alternative that has general real-world applicability to a variety of treatments. (Hilsenroth, Blagys, Ackerman, Bonge & Blais, 2005, S. 341)

Unabhängig davon also, ob in einer Studie manualisierte oder nicht manualisierte Behandlungen durchgeführt werden, gibt die Adherence-Kontrolle eine Antwort darauf, ob und inwieweit die jeweils prototypischen Techniken und Vorgehensweisen unter den jeweiligen Treatmentbedingungen umgesetzt wurden. Würde keine solche Kontrolle durchgeführt werden, so wäre nicht nur die Konstruktvalidität der Treatments in Frage gestellt, vielmehr könnte man selbst bei manualisierten Behandlungen dann nicht ausschließen, dass systematisch auch andere verfahrens*un*typische Techniken verwendet wurden. Dies könnte im Rahmen einer entdeckten Wirksamkeitsäquivalenz zwischen zwei Therapiealternativen die Frage aufwerfen, ob diese tatsächlich auf gleichwirksame, jedoch unterschiedliche prototypische Techniken und Vorgehensweisen zurückzuführen ist. Oder ob die Äquivalenz eher auf deutliche Überschneidungen der therapeutischen Prozeduren zurückgeführt werden muss (vgl. Leichsenring et al., 2011).

Eine Studie schneidet dementsprechend zufriedenstellend ab, wenn die *Treatment Integrity* mindestens durch eine Fragebogenerhebung bei den behandelnden Therapeuten belegt ist; bestenfalls erfolgt die Kontrolle hingegen videogestützt durch externe Rater.

Grundsätzlich gilt für beide Kriterien (B.3. und B.6.), dass die darin geforderten internen Validitätssicherungsstrategien nicht nur in dem Treatmentarm mit dem zu überprüfenden Verfahren umgesetzt werden sollten, sondern ebenfalls in den Vergleichsbedingungen. Andernfalls könnte ein beobachteter Effekt allein darauf zurückzuführen sein, dass in dem einen Arm ein Manual verwendet wird, in dem anderen hingegen nicht. Gleiches trifft auf die Adherence-Kontrolle zu (vgl. Leichsenring et al., 2011).

Die beiden Kriterien B.4. und B.5. der Rubrik „Intervention“ beziehen sich zum einen auf die operationale Definition der Kontrollbedingung (Kriterium B.4.) und zum anderen auf die strukturelle Äquivalenz der Kontrollbedingung im Vergleich zu dem zu überprüfenden psychotherapeutischen Verfahren (Kriterium B.5.). Mit „operationaler Definition der Kontrollbedingungen“ ist gefordert, dass dasjenige, was in diesem Treatmentarm geschieht, einerseits prospektiv festgelegt wurde und andererseits möglichst so ausführlich beschrieben wird, dass eindeutig zu beurteilen ist, *was* mit dieser Bedingung tatsächlich kontrolliert wird. Diese Beurteilung wiederum bildet die Basis für die Bewertung auf Kriterium B.5., das nach der Strukturgleichheit der Kontroll- und der Interventionsbedingung fragt. Damit ist gemeint, dass allgemeine Faktoren, wie die therapeutische Zuwendung, die Sitzungsanzahl und die Settingbedingungen in der Kontrollbedingung denen der Interventionsbedingung angeglichen werden sollten. Erst die Angleichung dieser Randbedingungen macht eine Kontrollgruppe nach Hager (2000) zu einer wirksamen Strategie der internen Validitätssicherung:

> *Ob und ggf. wie weit die interne Validität in einer konkreten Studie tatsächlich gesichert ist, hängt in erster Linie davon ab, was man mit den Vergleichsgruppen anstellt* (Hager, 1998a, 1998b). So kann selbst bei einer „Placebo-Intervention“ die interne Validität entscheidend vermindert sein, weil sich die Häufigkeit und Intensität der Kontakte (Randbedingungen) in den beiden Vergleichsgruppen (Experimental- und Vergleichsgruppe) voneinander unterscheiden . . . . (Hervorhebungen im Original, Hager, 2000, S. 183)

In den meisten Fällen verfolgt vor allem die Realisierung von Placebo-Kontrollgruppen den Zweck, Randbedingungen der o.g. Art in Angleichung an die Interventionsbedingung umzusetzen und damit zu kontrollieren. In Kapitel 3.1.1 wird auf die Wahl und Umsetzung unterschiedlicher Kontrollbedingungen sowie auf deren Bedeutung für die Interpretation der Wirksamkeit eines psychotherapeutischen Verfahrens detailliert Bezug genommen. Für hier soll

festgehalten werden: Soll in einer Studie in erster Linie über die Wirksamkeit verfahrensspezifischer Techniken eine Aussage getroffen werden, so gilt es, möglichst viele andere Quellen, von denen ebenfalls Wirkung ausgeht, zu kontrollieren. Bei diesen Wirkquellen handelt es sich um Faktoren, die allen psychotherapeutischen Interventionen gemeinsam sind (z.B. eine regelmäßig stattfindende, intensive zwischenmenschliche Zuwendung) und die daher in der Regel nicht als Technik eines bestimmten Verfahrens betrachtet werden. Je mehr eine Kontrollbedingung diese allgemeinen Faktoren kontrolliert, desto höher wird mit Kriterium B.5. die interne Validitätssicherung eingestuft.

Kriterium B.7. der internen Validitätsdimension bezieht sich auf weitere begleitende psychotherapeutische/psychiatrische Interventionen neben den eigentlich intendierten und damit auf Konfundierungen zwischen diesen unterschiedlichen Wirkquellen. Werden diese zusätzlichen Wirkquellen nicht von vornherein ausgeschlossen oder durch Randomisierung über die Treatmentarme kontrolliert, so sind sie als bedingungsgebundene Störvariablen zu betrachten, die als Alternativerklärung für einen beobachteten Effekt in Frage kommen. Um solche differentiellen Effekte aufzudecken und ggf. statistisch zu kontrollieren, muss die Inanspruchnahme zusätzlicher Interventionen zunächst dokumentiert und als möglicherweise differentiell wirkender Störfaktor analysiert werden. Daher wird von einer Studie, in der grundsätzlich zusätzliche, nicht randomisierte Interventionen zugelassen sind, gefordert, dass sie den Einfluss begleitender Interventionen dokumentiert und analysiert. Weisen die Analysen auf einen differentiellen Einfluss hin, so schneidet eine solche Studie mit „3“ ab. Wird in einer Studie dieser Störfaktor von vornherein eliminiert, so wird eine solche Untersuchung auf Kriterium B.7. mit „1“ bewertet.

Studiendesign: Die beiden ersten Kriterien B.8. und B.9. der Rubrik „Studiendesign“ befassen sich mit der Strukturgleichheit der Vergleichsgruppen im Hinblick auf potentielle Störvariablen. Kriterium B.8. stellt zudem das K.O.-Kriterium der internen Validitätsbewertung dar, d.h. Studien, in denen die Gruppen weder durch Zufallszuweisung noch durch Parallelisierung zusammengestellt werden, fallen auf der internen Validitätsdimension automatisch durch. Gleiches trifft freilich auf Untersuchungen ohne Vergleichsgruppen (Ein-Gruppen-Designs) zu. Der Grund, warum die randomisierte Gruppenzuweisung in der experimentellen Forschung stets als der „Königsweg“ der Störvariablenkontrolle bezeichnet wird, liegt darin, dass allem voran auch *unbekannte* Störfaktoren von der Randomisierung berührt werden. Die Randomisierung bietet damit den unwiderruflichen Vorteil, sowohl bekannte als auch unbekannte Störvariablen über die unterschiedlichen Bedingungen gleich zu verteilen respektive konstant zu halten: „Randomization equates groups on expectations of *every variable before treatment, whether observed or not*“ (Hervorhebungen im Original, Shadish et al., 2002, S. 251).

Doch was bedeutet „on expectations of every variable“ in diesem Zusammenhang? Es wurde bereits darauf hingewiesen, dass man dem Zufall nicht blind vertrauen sollte (vgl. Eid et al., 2010). Eben diese Vorsicht dem Zufall gegenüber steckt in dem zugrunde liegenden Prinzip von „on expectations of every variable“, dessen Bedeutung Shadish und Kollegen an folgendem Beispiel veranschaulichen:

> First, it does not mean that random assignment equates unites on *observed* pretest scores. Howard, Krause, and Orlinsky (1986) remind us that when a deck of 52 playing cards is well shuffled, some players will still be dealt a better set of cards than others. This is called the luck of the draw by card players . . . . In card games, we do not expect every player to receive equally good cards for each hand. All this is true of the randomized experiment . . . . Technically, then, random assignment equates groups on the *expectation* of group means at pretest – that is, on the mean of the distribution of all possible sample means resulting from all possible random assignments of units to conditions. (Hervorhebungen im Original, 2002, S. 250)

Es ist demnach davon auszugehen, dass selbst randomisierte Zuweisungen keine Gleichverteilungen aller Störvariablen im Sinne von „*observed* pretest scores“ garantieren – die Gleichverteilung ist zudem noch mehr in Zweifel zu ziehen, je kleiner die Stichproben sind (vgl. Hsu, 1989). Umso wichtiger ist die systematische Untersuchung der Gruppen hinsichtlich prognostisch zentraler Variablen zur Baseline – Inhalt des Kriteriums B.9.. Bevor auf Kriterium B.9. näher eingegangen wird, soll noch ein Blick auf die Stufenoperationalisierungen von B.8. geworfen werden: Mit einer „1“ werden solche Studien bewertet, die bei angemessener Stichprobengröße ($n > 30$ pro Gruppe) eine randomisierte Zuweisung der Studienpatienten zu den Treatments vornehmen; mit einer „2“ werden Studien mit randomisierter Zuweisung, jedoch kleineren Stichproben bewertet sowie Untersuchungen, in denen die Gruppenzusammenstellung nur teilweise randomisiert oder quasi-randomisiert[18] erfolgt. Außerdem werden Studien mit einer „2“ bewertet, in denen die Untersuchungsgruppen parallelisiert zustande kommen. Im Vergleich zur Randomisierung kann die Parallelisierung ausschließlich im Hinblick auf eine begrenzte Anzahl a priori *bekannter* und gemessener Störvariablen erfolgen. Erfolgt eine Gruppenzusammenstellung weder per Zufallszuweisung noch parallelisiert, so wird eine solche Studie mit „3“ bewertet und erhält eine globale Negativbewertung auf der internen Validitätsdimension.

Nach Kriterium B.9. ist eine Untersuchung dann als intern valide zu betrachten, wenn die Vergleichsgruppen im Hinblick auf prognostisch relevante und möglicherweise ungleich verteilte Faktoren untersucht und diese ggf. statistisch kontrolliert werden. Findet keine Analyse bekannter personengebundener Störvariablen zum Prämesszeitpunkt statt oder zeigt sich bei der Analyse ein bedeutsamer Unterschied zwischen den Vergleichsgruppen hinsichtlich dieser

[18] Unter einer Quasi-Randomisierung wird eine Zuteilung nach Aspekten, wie Geburtsdatum o.ä. verstanden (im Gegensatz zur computergenerierten Zufallszuordnung, die einer „echten“ Randomisierung gleichkommt).

Störfaktoren, dann wird eine solche Studie mit „3“ bewertet. Werden keine Unterschiede zwischen den Gruppen entdeckt, so wird eine solche Untersuchung auf Kriterium B.9. mit „1“ kodiert.

Die beiden internen Validitätskriterien B.10. und B.11. beziehen sich nun auf das Messdesign der zu bewertenden Studien. Kriterium B.10. fordert für eine sehr gute Bewertung (Stufe „1“) die prospektive Festlegung der Messzeitpunkte sowie, zusätzlich zu den Prä-Postmessungen, Messungen über den Therapieverlauf. Mit einer „3“ wird eine Untersuchung bewertet, die lediglich über eine Outcomemessung zum Postzeitpunkt verfügt. Inwiefern drückt sich ein Messdesign mit mehreren Messzeitpunkten in einer potentiellen Erhöhung der internen Validität aus? Die Antwort findet sich am ehesten im Konzept des *coherent pattern matching* nach Shadish et al. (2002), das Leichsenring (2004a) übersetzt mit der Vorhersage komplexer Ergebnismuster. Diesem Konzept zufolge werden hinlänglich bewährte Hypothesen über die Wirkweise einer Intervention und über die daraufhin zu erwartenden Ergebnismuster aufgestellt, die sich sodann in empirischen Studien bestätigen sollten. Je komplexere Ergebnismuster vorhergesagt werden und je mehr die empirische Evidenz diese Muster bestätigt, desto eher ist davon anzugehen, dass diese Muster ein Resultat der Interventionswirkung abbilden und Alternativerklärungen als unplausibel zu betrachten sind. Bspw. könnten auf der Basis von Befunden aus der Psychotherapieprozessforschung Hypothesen über den Symptomverlauf und, parallel dazu, über den Veränderungsverlauf interpersonaler Probleme oder des Strukturniveaus während einer längerfristigen Therapie aufgestellt werden. Diese Hypothesen würden somit den prototypischen Verlauf der genannten Indikatoren unter Einwirkung der therapeutischen Intervention markieren und vorhersagen. Um zu überprüfen, inwiefern die theoretisch vorhergesagten Ergebnismuster über den Therapieverlauf auch empirisch bestätigt werden, bedarf es mehrerer Messzeitpunkte, zu denen diejenigen Outcomes, über die Hypo-

thesen aufgestellt wurden, erhoben werden. Ein Messdesign, das eine solche Überprüfung zulässt, wird durch Kriterium B.10. gefordert.

Kriterium B.11. weitet das Zeitfenster, über das Aussagen über den Störungsverlauf gemacht werden können, auf den Zeitraum nach Beendigung der Behandlung aus – nämlich auf den Katamnesezeitraum. Damit fragt B.11. nach dem weiteren Störungsverlauf nach Abschluss einer Therapie und somit nach der langfristigen Wirkung der Behandlung. Kriterium B.11. liefert die zu überprüfende Hypothese (im Sinne des *coherent pattern matching*) im Grunde genommen mit, indem ein sog. "störungsangemessener Katamnesezeitraum" gefordert wird. Dieser Forderung liegt die empirisch fundierte Annahme zugrunde, dass für ein bestimmtes Störungsbild ein prototypischer, natürlicher Verlauf der Symptomatik beschrieben werden kann. Ein Katamnesezeitraum gilt daher genau dann als angemessen, wenn er den Zeitraum, innerhalb dessen mit hohen Rückfallraten nach Remission oder Genesung zu rechnen ist, im Rahmen einer Wirksamkeitsstudie möglichst „einfängt“ (dazu ausführlich: Kap. 3.2.3). Die Voraussage im Sinne des *coherent pattern matching* wäre dann, dass sich die Rückfallraten innerhalb eines bestimmten Zeitraums nach Beendigung der Therapie im Vergleich zu denen, die bei einem *unbehandelten* (natürlichen) Verlauf innerhalb desselben Zeitraums zu erwarten sind, deutlich vermindern. Erst dann wäre auch von einer langfristigen Wirkung der Therapie auszugehen. Komplexere Hypothesen würden o.g. Erwartungen noch für unterschiedliche Subgruppen und Outcomemaße präzisieren, um Alternativerklärungen für den beobachteten Effekt mit höherer Sicherheit ausschließen zu können. Mit einer „1“ werden Studien daher nur dann bewertet, wenn sie einen störungsangemessenen Katamnesezeitraum realisieren und zudem die Ausschöpfung der Stichprobe hoch ist. Verfügt eine Studie über keine Katamnesemessung, so schneidet sie in B.11. mit „3“ ab.

Outcomemessung: Die vierte Rubrik der internen Validitätsdimension bezieht sich auf die Messung und Darstellung der Ergebnisse und besteht aus einem Kriterium (B.12.). Da dieses Kriterium in Kapitel 3.2.3 ausführlich beschrieben wird, soll der Inhalt des Kriteriums an dieser Stelle lediglich kurz skizziert werden. Kriterium B.12. – zugleich K.O.-Kriterium der Dimension der allgemeinen methodischen Qualität – fordert die Darstellung der Outcomes mittels unterschiedlicher Indikatoren. Hierbei gibt das Kriterium vor, dass sowohl Indikatoren der *Veränderung* als auch der *Zielerreichung* berichtet werden sollten. In Anlehnung an Schulte (1993) sind Veränderungen über einen Therapieverlauf zum einen durch gruppenstatistische Prä-Postdifferenzen (ggf. plus Katamnese) festzustellen, zum anderen durch Effektstärken sowie durch Veränderungsmaße auf individueller Ebene. Letzterer Indikator wurde bereits im Zusammenhang mit der klinischen Relevanz (s.o.) eingeführt und betrifft die Abbildung der individuellen Veränderung über den RCI nach Jacobson und Truax (1991). Als Maß der Zielerreichung gilt der Vergleich individueller Post- oder Katamnesewerte mit normorientierten Cutoff-Werten. Diese Cutoff-Werte markieren den Grenzbereich zwischen dem dysfunktionalen und dem funktionalen Bereich und werden vor dem Hintergrund geeigneter Erhebungen an gesunden und dysfunktionalen Normpopulationen festgelegt. Auf die Operationalisierungen der einzelnen Kriterienstufen wird in Kapitel 3.2.3 noch detailliert eingegangen. Es soll an dieser Stelle jedoch noch der Wert dieses Kriteriums als eines der internen Validität reflektiert werden. Welche potentiellen Alternativerklärungen für einen beobachteten Effekt sollen mittels unterschiedlicher Indikatoren der Effektivität ausgeräumt werden? Es wurde bereits dargelegt, inwiefern gruppenstatistische Ergebnisse auf der Basis von Mittelwertvergleichen und individuumsbezogene Ergebnisse in ihren Aussagen über die Wirksamkeit einer Intervention differieren können: „Reliance solely on statistical significance can lead to perceiving differences (i.e., traitment gains) as potent when in fact they may be clinically insignificant“ (Kendall et al., 2004, S. 28). Dementsprechend können Verände-

rungsmaße, wie inferenzstatistisch abgesicherte Prä-Post- oder Zwischengruppendifferenzen zum Postzeitpunkt in Abhängigkeit von der Power bzw. von der Stichprobengröße durchaus zu signifikanten Ergebnisse führen – die stichprobengrößen*un*abhängigen Effektstärken könnten demgegenüber vergleichsweise klein ausfallen. Im Zusammenhang mit Kriterium A.9., das allein den Bericht klinischer Relevanzmaße in einer Studie bewertet (s.o.), wurde gezeigt, dass gänzliche *unterschiedliche* Szenarien individuumsbezogener Ergebnisse (via RCI und Cutoff-Wert) zu ein und demselben gruppenstatistischen Mittelwertunterschied führen können. Auch wäre es denkbar, dass sich die Symptomatik bei zahlreichen Patienten zwar reliabel verbessert, jedoch nahezu niemand den vorgegebenen normorientierten Cutoff-Wert überschreitet. Aus diesem Grund können unterschiedliche Ergebnismaße u.U. unterschiedliche Interpretationen über die Wirksamkeit einer Intervention nahelegen. Eine einseitige Betrachtung der Wirksamkeit bspw. allein durch inferenzstatistische Mittelwertvergleiche als Maß der Veränderung könnte somit die interne Validität beeinträchtigen und zu falschen Schlussfolgerungen über die Wirksamkeit einer Behandlung führen: Alternativ zur Behandlungswirksamkeit könnte die Größe der Stichprobe für das signifikante Ergebnis ins Gewicht fallen.

Die Dimension der externen Validität

Die externe Validitätsdimension beläuft sich auf insgesamt 13 Kriterien, wobei die ersten neun Kriterien (C.1.-C.9.) sich auf den bereits eingeführten Aspekt der *representativeness* beziehen und die letzten vier Kriterien (C.10.-C.13.) auf den Aspekt der *transferability* (vgl. Borkovec & Castonguay, 1998; Leichsenring et al., 2011). Damit bewerten die Kriterien C.1. bis C.9., inwieweit die untersuchten Patienten, die Studientherapeuten, die durchgeführten Interventionen, der in der Studie realisierte Zugang zur Behandlung sowie die Zielkriterien der Behandlung die klinische Praxis (des deutschen Gesundheitssystems) *repräsentieren*. Die-

se Kriterien lehnen sich größtenteils an diejenigen an, die auch Shadish, Navarro, Matt und Phillips (2000) in ihrer Metaanalyse zur Bemessung der Repräsentativität von Outcomestudien anlegten.

Mit den Kriterien C.10. bis C.13. wird hingegen bewertet, inwieweit die in einer Studie vorkommenden Settingbedingungen und Interventionen, die notwendigen Behandlerqualifikationen sowie die notwendige Erfassung patientenseitiger Merkmale in die Praxis tatsächlich integrierbar bzw. *transferierbar* sind. Strenggenommen wird mit dem Einbezug des Praxistransfers in die externe Validitätsbewertung über die o.g. Definition der externen Validität hinausgegangen (siehe Kap. 1.2): Diese bezog sich allein auf die Übertragbarkeit eines beobachteten Effekts auf andere Personengruppen, Orte, Situationen und Zeitpunkte. Der WBP trägt dieser Unterscheidung zwischen Generalisierbarkeit und Praxistransfer Rechnung, indem er die beiden Aspekte in der Gesamtbewertung der externen Validität getrennt voneinander verrechnet. Damit werden die *representativeness* und die *transferability* in ihrer jeweiligen Dignität unabhängig voneinander betrachtet und eingeschätzt. Eine nähere Betrachtung der Kriterien folgt nun:

Patienten: Die ersten drei Kriterien (C.1., C.2., C.3.) beziehen sich auf die Repräsentativität der untersuchten Patienten. Bei Kriterium C.1. handelt es sich um ein K.O.-Kriterium, d.h. eine Studie fällt bereits durch die Bewertung der allgemeinen methodischen Qualität durch, wenn sie auf C.1. nur mit „3“ bewertet wird. Kriterium C.1. bezieht sich auf den Krankheitswert einer psychischen Störung, dieser bemisst sich laut Operationalisierung des Kriteriums daran, ob bspw. die Kriterien einer ICD- oder DSM-Diagnose erfüllt sind. Bestehen die Untersuchungsgruppen einer Studie ausschließlich aus Patienten mit einer krankheitswertigen Störung (Stufe „1“) oder verfügen maximal 20% der Untersuchungsgruppen über subklini-

sche Symptomausprägungen (Stufe „2“), so wird eine solche Studie bzgl. der Patientenklientel als für den Versorgungskontext repräsentativ betrachtet.

Kriterium C.2. bezieht sich auf die Selektion der Patienten, die durch den in der Studie realisierten Zugang zur Behandlung zustande kommen kann: Erfolgt der Zugang, wie im Routinekontext üblich, überwiegend über den Primärzugang oder aber vermittelt durch den Haus- oder Facharzt, so wird eine solche Studie als für den Versorgungskontext repräsentativ betrachtet. Ist durch die Art der Rekrutierung der Patienten hingegen mit starken Selektionseffekten zu rechnen – etwa, indem durch Anzeigenwerbung zur Studienteilnahme und damit zur Behandlung motiviert wird – so wird eine solche Studie mit „3“ bewertet. Entsprechend dieser Operationalisierung ist davon auszugehen, dass sich unter diesen Umständen eine Patientengruppe formiert, die derjenigen Klientel unähnlich ist, die im Routinekontext eine psychotherapeutische Behandlung aufsucht.

Kriterium C.3. bezieht sich auf Selektionseffekte, die durch die in einer Studie angelegten Ausschlusskriterien zustande kommen können. Eine Patientengruppe wird laut Kriterium C.3. genau dann als selektiv angenommen, wenn Ausschlusskriterien bspw. aus Gründen der statistischen Validität dafür sorgen, dass die Untersuchungsgruppe sich systematisch von derjenigen Klientel unterscheidet, für die die Behandlung in der Praxis eigentlich konzipiert ist. Soll eine Aussage über einen Behandlungseffekt beim Störungsbild der unipolaren Depression gemacht werden und ist die Untersuchungsgruppe gleichzeitig durch den Ausschluss komorbider Störungen oder besonders schwerer Symptomausprägungen derart homogen zusammengesetzt worden, so dass Effekte auf jeden Fall sichtbar werden (statistische Validität), dann ist die externe Validität in Frage gestellt: Die Behandlungseffekte sind strenggenommen lediglich auf einen sehr speziellen Teil der Population depressiver Patienten zu generalisieren, nämlich auf denjenigen Teil mit nur leichten bis mittelschweren Symptomausprägungen und monomorbider Depression. Vor dem Hintergrund hoher Komobiditätsraten depressiver Stö-

rungen – Laux (2008a) berichtet allein für die Major Despression im Zusammenhang mit Angststörungen von einer 65%igen Komobiditätsrate – fällt der Ausschluss komorbider Störungen für diese Störungsgruppe für die Einschätzung der Selektivität der Untersuchungsgruppen stark ins Gewicht. Je mehr epidemiologisch relevante komorbide Störungen ausgeschlossen werden, mit desto größeren Selektionseffekten ist zu rechnen. Werden keine Ausschlusskriterien dieser Art formuliert und alle Patienten in die Untersuchung aufgenommen, so wird eine solche Studie mit „1“ bewertet.

Intervention: Die Rubrik „Intervention“ der externen Validitätsdimension umfasst vier Kriterien, die sich auf die Repräsentativität der in einer Studie durchgeführten Behandlung beziehen. Das erste Kriterium (C.4.) bemisst die Generalisierbarkeit der Studientherapie auf die klinische Alltagspraxis, bezogen auf das therapeutische Vorgehen und die Dauer der Behandlung. Zur Kodierung dieses Kriteriums bedarf es zunächst einer Festlegung, was als *prototypisches* Vorgehen und als ebensolche *prototypische* Dauer eines bestimmten psychotherapeutischen Vorfahrens betrachtet werden kann. Diese Festlegung und die damit einhergehenden Schwierigkeiten, vor allem in Bezug auf die psychoanalytisch begründeten Verfahren, werden in Kapitel 3.2.2 ausführlich beschrieben.

Kriterium C.5. bezieht sich auf Einflüsse eines im Rahmen einer Studie implementierten Monitoringprogramms. Werden bspw. allein zu Studienzwecken Supervisionen durchgeführt, die die Behandlungsqualität und die Adherence sichern sollen, so unterscheidet sich nach Kriterium C.5. eine solche Studienbehandlung systematisch von einer Behandlung in der klinischen Praxis. Davon zu unterscheiden sind jedoch *regulär* stattfindende Supervisionen oder Intervisionen oder aber Audio- und Videoaufzeichnungen, die zwecks Adherence-Kontrollen vorgenommen werden, von denen jedoch kein nennenswerter Einfluss auf das Therapeutenverhalten anzunehmen ist. Wird das Therapeutenverhalten hingegen durch ein

implementiertes Rückmeldesystem kontinuierlich kontrolliert und beeinflusst, so wird eine Untersuchung auf diesem Kriterium folglich mit einer „3“ bewertet; keine oder nur geringfügige Einflussnahmen werden im Sinne der externen Validität (Generalisierbarkeit) dementsprechend mit „1“ oder „2“ bewertet.

Kriterium C.6. bildet gewissermaßen das Pendant zum internen Validitätskriterium B.7., das den Ausschluss zusätzlicher, nicht randomisierter Interventionen fordert (s.o.). Mit C.6. werden Studien hingegen dann als extern valide betrachtet, wenn sie keine Einschränkungen im Hinblick auf solche begleitenden Interventionen vornehmen, die in der Routinepraxis regelhaft anzutreffen sind (meist Pharmakotherapie).

Das letzte Kriterium der Rubrik „Intervention“ (C.7.) stellt nun insoweit eine Besonderheit dar, als dass es sich in drei Subkriterien aufgliedert, die sich allesamt auf die Qualifikation der behandelnden Studientherapeuten beziehen. Mit Kriterium C.7a. wird vorausgesetzt, dass die primär in der psychotherapeutischen Routinepraxis tätigen Behandler approbierte Psychotherapeuten mit hauptberuflich klinischer Tätigkeit sind. Somit wird eine Studie, in der die Studientherapeuten hauptberuflich in der Forschung tätig sind und lediglich selten Psychotherapien durchführen hinsichtlich der externen Validität schlechter bewertet (Stufe „2“). Gleiches trifft auf Ausbildungskandidaten als Studientherapeuten zu. Handelt es sich bei dem in der Untersuchung behandelnden Personal um Mediziner oder Psychologen ohne psychotherapeutische Ausbildung, so entspricht dies nach Kriterium C.7a. in keiner Weise der psychotherapeutischen Routineversorgung im deutschen Gesundheitssystem, so dass eine solche Studie mit „3“ bewertet werden würde. Kriterium C.7b. bezieht sich auf die Breite der klinischen Tätigkeit der Studientherapeuten: Da die Ausbildung in einem psychotherapeutischen Verfahren in der Regel zur Behandlung unterschiedlicher Störungsbilder befähigt und die meisten approbierten Psychotherapeuten mit einer ebensolchen Problemheterogenität konfrontiert sein werden, sollten die Behandler in einer extern validen Studie dementsprechend

„breit aufgestellt“ sein. Schlussfolgerungen aus einer Studie, in der Therapeuten sowohl innerhalb als auch außerhalb der Studie ausschließlich Patienten mit eng umgrenzten Problemen behandeln (Schmerzpatienten, Suchtpatienten etc.), werden laut Kriterium C.7b. daher als nur eingeschränkt generalisierbar betrachtet. Kriterium C.7c. bezieht sich schließlich auf ein zu Studienzwecken durchgeführtes Therapeutentraining für die in der Studie durchzuführende Behandlung. Da derart spezifische Trainings, neben der regulären Fortbildungspflicht von Psychotherapeuten, in der klinischen Routine die Ausnahme bilden, sind intensiv trainierte Studientherapeuten nicht als repräsentativ für die in der Versorgung tätigen Therapeuten zu betrachten. Dementsprechend schneiden nur Untersuchungen, im Rahmen derer untrainierte Studientherapeuten behandeln, auf Kriterium C.7c. entsprechend zufriedenstellend ab.

Studiendesign: Die Rubrik „Studiendesign“ umfasst auf der externen Validitätsdimension ein Kriterium (C.8.), das als Pendant zum internen Validitätskriterium B.8. betrachtet werden kann: B.8. bemisst die Gruppenzuweisung und bewertet im Sinne der internen Validität eine Studie im Hinblick auf diesen Aspekt dann als zufriedenstellend (Stufe „2“), wenn die Gruppenzuweisung parallelisiert erfolgt; die randomisierte Zuweisung bei angemessener Stichprobengröße wird mit „1“ bewertet. Da die zufällige Zuteilung von Patienten nicht den Patientenzugang zu einer Behandlung in der Praxis repräsentiert, kehrt sich die Bewertung beim externen Validitätskriterium C.8. nahezu um. Mit einer „3“ werden demnach solche Studien bewertet, in denen die Patienten ihren Behandlungen randomisiert zugewiesen werden, mit „1“ werden demgegenüber Untersuchungen bewertet, in denen sich die Patienten selbst für eine Therapieform entscheiden. Im Unterschied zum internen Validitätskriterium B.8. handelt es sich bei C.8. nicht um ein K.O.-Kriterium, d.h., eine Studie fällt durch die externe Validitätsbewertung nicht automatisch durch, wenn die Gruppenzuweisung randomisiert (Stufe „3“) erfolgt.

Outcomemessung: Bei Kriterium C.9. handelt es sich um ein K.O.-Kriterium, das bedeutet, eine Studie fällt bereits im Rahmen der allgemeinen methodischen Qualitätsbewertung durch, wenn sie auf diesem Kriterium nicht mindestens Stufe „2“ erreicht. Für eine gute Bewertung (Stufe „1“ oder „2“) auf diesem Kriterium sollten in einer Untersuchung solche primären Zielkriterien angelegt werden, die als patientenrelevante Endpunkte betrachtet werden können. Darunter sind in erster Linie Zielkriterien zu verstehen, die Heilung oder Linderung von einer psychischen Störung kennzeichnen. Laut Kriterium C.9. zählen zu patientenrelevanten Endpunkten, die durch die Behandlung positiv beeinflusst werden sollten, insbesondere die Symptomatik, die Lebensqualität sowie das Inanspruchnahmeverhalten. Werden im Rahmen einer Studie hingegen ausschließlich sog. Surrogatparameter (z.B. Kontrollüberzeugung) als Zielkriterien verwendet (Stufe „3“), dann führt dies unmittelbar zum Ausschluss der Studie. Eine ähnlich strenge Haltung gegenüber der alleinigen Fokussierung auf Surrogatparameter findet sich auch im Methodenpapier (Entwurfsfassung 4.2) des IQWiG wieder. Das IQWiG begründet dies folgendermaßen:

> Surrogatendpunkte werden in der medizinischen Forschung häufig als Ersatz für patientenrelevante Endpunkte verwendet, meist um Aussagen zum patientenrelevanten (Zusatz-)Nutzen früher und einfacher zu erhalten [15, 190, 428]. Die meisten Surrogatendpunkte sind jedoch in dieser Hinsicht nicht verlässlich und können bei der Nutzenbewertung irreführend sein [100, 216, 224]. (2014, S. 40)

Demnach können Surrogatendpunkte allein dann in der Nutzenbewertung einer Intervention Beachtung finden, wenn der Einfluss des Surrogatparameters auf den patientenrelevanten Endpunkt eindeutig nachgewiesen ist. Die Hürde, die dieser Nachweis nehmen muss, ist vergleichsweise hoch, so fordert das IQWiG bspw. Metaanalysen auf der Basis randomisiert-

kontrollierter Studien für den Nachweis eines kausalen Zusammenhangs zwischen Surrogatparameter und patientenrelevantem Endpunkt.

Praxistransfer: Unter der Rubrik „Praxistransfer“ subsumiert der WBP nun diejenigen Kriterien zur externen Validität, die in Anlehnung an Leichsenring et al. (2011) die *transferability* von Studienaspekten in den klinischen Routinekontext bemessen (vgl. Kap. 1.2).

Kriterium C.10. bemisst die Übertragbarkeit der für die Intervention notwendigen Settingbedingungen auf den klinischen Routinekontext. Bedarf es für eine Behandlung bspw. besonders spezieller Ausstattungen (bestimmte Räumlichkeiten, videografische Ausstattung o.ä.), so ist der Praxistransfer in Frage gestellt und eine solche Studie schneidet mit Stufe „3“ ab. Selbiges trifft auf Studien zu, in denen für die untersuchte Behandlung eine spezielle und sehr umfangreiche Behandlerqualifikation notwendig ist (Kriterium C.11.). Mit „umfangreich“ ist hier jedoch nicht der reguläre Ausbildungsumfang zum psychologischen oder ärztlichen Psychotherapeuten gemeint, sondern deutlich darüber hinausgehende Qualifikationen, wie Zusatzausbildungen in bestimmten störungsspezifischen Methoden o.ä.. Detaillierte Informationen dazu, wie in der Bewertung dieses Kriteriums vorzugehen ist, sind Kapitel 3.2.2 zu entnehmen. Kriterium C.12. bezieht sich auf das Ausmaß zu erhebender patientenseitiger Merkmale, die für die Behandlung relevant sind. Geht die für die Behandlungsindikation notwendige Diagnostik deutlich über den regulären diagnostischen Prozess hinaus, so kann eine Studie im Hinblick auf den Praxistransfer allenfalls mit „2“ oder „3“ abschneiden. Das letzte Kriterium (C.13.) der Rubrik „Praxistransfer“ bezieht sich auf die Übertragbarkeit und Integrierbarkeit der durchgeführten Studientherapien in die Versorgungspraxis. Sind bestimmte Behandlungsmerkmale – etwa eine ungewöhnlich hohe Frequenz von täglich stattfindenden Therapiesitzungen – in der klinischen Versorgung kaum umzusetzen, so muss der Praxistrans-

fer in Zweifel gezogen und eine solche Studie dementsprechend mit Stufe „2“ oder „3“ bewertet werden.

Fazit

Mit dem Methodenpapier (Version 2.8) und insbesondere mit dem Kriterienkatalog legt der WBP ein detailliertes Regelwerk vor, anhand dessen die wissenschaftliche Anerkennungspraxis transparent wird. Durch die Integration der externen Validitätsdimension begegnet er bereits einer grundlegenden Kritik an der einseitigen Übernahme der Evidenzkriterien der EbM. Durch die dimensionale Gestaltung und die getrennte Bewertung der internen und der externen Validität eröffnet der WBP die Möglichkeit, Studien hinsichtlich beider Gütemaßstäbe zu bewerten. Er schafft damit eine differenziertere Betrachtung als die, Studien lediglich in eine der beiden Kategorien „experimentell“ oder „naturalistisch“ einzuordnen. Gleichzeitig vermeidet er dadurch, dass experimentelle Studien (*efficacy*) ausschließlich aus dem Blickwinkel der mit dieser Kategorie assoziierten Stärke – die interne Validität – betrachtet werden; umgekehrt werden *effectiveness*-Studien nicht ausschließlich im Hinblick auf die externe Validität betrachtet. Dadurch kann jede Studie im Einzelfall sowohl auf die interne Validität als auch auf die Generalisierbarkeit ihrer Ergebnisse sowie den Praxistransfer hin bewertet werden. Diese differenzierte Begutachtung entspricht konzeptionell dem wissenschaftlichen Standard, demzufolge die beiden Gütekriterien keineswegs grundsätzlich oder ausschließlich als invers zu betrachten sind (vgl. Heekerens, 2005; Leichsenring, 2004a/b; Shadish et al., 2002). Vielmehr muss davon ausgegangen werden, dass in Untersuchungen, in denen aus Gründen der externen Validität von diversen internen Validitätssicherungsstrategien, wie Randomisierung, Manualisierung oder Kontrollgruppen (Warteliste, Placebo etc.), abgesehen werden muss, die interne Validität durch alternative Strategien gesichert werden kann (vgl. Leichsenring, 2004a/b; Shadish et al., 2002). Gleiches gilt für Studien, bei denen aus ethi-

schen oder gegenstandsbezogenen Gründen von genannten Strategien abgewichen werden muss – etwa, wie es in Langzeittherapiestudien der Fall ist. Umgekehrt gilt für randomisierte Studien, dass diese nicht notgedrungen Ergebnisse liefern, denen jede Generalisierbarkeit abgesprochen werden muss (vgl. Stirman, DeRubeis, Crits-Christoph & Rothman, 2005). Vielmehr muss auch bei diesen Studien genau abgewogen werden, in Bezug auf welche Aspekte die Generalisierbarkeit sowie der Praxistransfer evtl. gefährdet sein könnte.

Der Fokus der Arbeit soll im Folgenden vertieft werden, indem sich der Konzeption und Operationalisierung der internen Validität durch den WBP unter Hinzuziehung der real vorliegenden Evidenz der psychoanalytisch begründeten Verfahren zugewandt wird. Damit wird der Frage nachgegangen, ob sich insbesondere die internen Validitätskriterien *in ihrer Anwendung* als gegenstandsangemessen in Bezug auf solche Studien erweisen, die sich der strikten RCT-Methodologie entziehen (vgl. Kap. 2).

2 Fragestellung

Im vorhergehenden Teil der Arbeit wurden Einwände gegen eine uneingeschränkte Übertragung der RCT-Methodologie auf den psychotherapeutischen Kontext dargestellt – insbesondere, was den Bereich der Langzeittherapien betrifft. Vor diesem Hintergrund soll der Frage nachgegangen werden, wie sich die vom WBP vorgenommene Operationalisierung von interner Validität in Form der internen Validitätskriterien auf deren Anwendbarkeit und auf die Bewertung von existierenden Wirksamkeitsstudien zu Langzeitbehandlungen auswirkt. In diesem Rahmen soll eruiert werden, ob und ggf. inwieweit die Anwendung der internen Validitätskriterien zu benachteiligenden Bewertungen von Langzeittherapieuntersuchungen (> 100 Sitzungen) im Vergleich zu Studien zu kürzeren Behandlungsdauern führen. Unter "benachteiligenden Bewertungen" sind entweder systematische Negativbewertungen oder aber die systematische Produktion von Missingwerten auf den einzelnen Kriterien bei der Beurteilung von Langzeittherapiestudien im Vergleich zu Kurzzeittherapiestudien zu verstehen. Da jedoch allein ein Mehr an kriterienbezogenen Negativbewertungen auf Seiten von Langzeittherapiestudien noch *nicht per se* als eine Benachteiligung betrachtet werden kann, die durch den Zuschnitt der Kriterien erzeugt wird, soll die empirische Analyse auf Basis von Studien zunächst dazu dienen, empirisch fundierte Hinweise auf *potentiell* benachteiligende Kriterien zu erlangen. Zentral wird es bei der Untersuchung daher sein, die Gründe für ein ggf. gefundenes Bewertungsgefälle zuungunsten der Langzeittherapiestudien zu reflektieren und zu diskutieren. Dies soll der Einschätzung dienen, ob die Gründe tatsächlich mit dem Gegenstand "Langzeitbehandlung" zusammenhängen oder als unabhängig davon betrachtet werden müssen.

Die Fragestellung lässt sich dahingehend zuspitzen, inwiefern insbesondere bei den internen Validitätskriterien von anwendbaren und gegenstandsangemessenen Kriterien im Hinblick auf die Begutachtung von "Langzeitbehandlung" ausgegangen werden kann. Diese Frage wird als zentral betrachtet, da es sich um einen Gegenstand handelt, der dem WBP folgend „besondere Forschungsfragen aufwirft“ (WBP, 2005, S. 74). Eben diese besonderen Forschungsfragen veranlassten den WBP dazu, vor Veröffentlichung der Version 2.6 des Methodenpapiers (2007), für diese Behandlungsform eine gesonderte Stellungnahme zu planen. In der Stellungnahme zur wissenschaftlichen Anerkennung psychodynamischer Psychotherapie blieben Langzeitbehandlungen > 100 Stunden aus diesem Grunde unberücksichtigt (vgl. WBP, 2005). Die hier beschriebene Fragestellung wird darüber hinaus dadurch virulenter, als dass der WBP mit Veröffentlichung des Methodenpapiers (2007) in einer ergänzenden Stellungnahme zu den psychodynamischen Verfahren eben diese Beschränkung auf kürzere Behandlungen (bis 100 Sitzungen) aufgehoben hat und den Bezugsrahmen der Stellungnahme von 2005 nunmehr auf Langzeittherapien (> 100 Sitzungen) ausweitete: „Vor dem Hintergrund des am 22. November 2007 in Kraft getretenen Methodenpapiers ist die auf der Grundlage der Behandlungsdauer getroffene Einschränkung der wissenschaftlichen Anerkennung nicht mehr berechtigt“ (WBP, 2008a, S. 426). Die Ausweitung des Bezugsrahmens der Stellungnahme zur wissenschaftlichen Anerkennung von psychodynamischer Psychotherapie soll in der vorliegenden Arbeit zum Anlass genommen werden, die Kriterien der aktuellen Fassung des Methodenpapiers (2010) auf ihre Gegenstandsadäquatheit eben im Hinblick auf den Gegenstand "Langzeitbehandlungen" genauer zu untersuchen. In dieser Untersuchung wird davon ausgegangen, dass den „besonderen Forschungsfragen“, die dieser Gegenstand bzgl. der Beforschung seiner Wirksamkeit aufwirft, mit einer „besonderen Begutachtungspraxis“ begegnet werden muss, die vom strikten Reglement der RCT-Methodologie abweicht. Ausgangspunkt der Untersuchung bildet daher die im ersten Teil der Arbeit bereits begründete

Annahme, dass *unterschiedlichen* Gegenständen mit *unterschiedlich* gearteten Strategien der internen Validitätssicherung begegnet werden kann – und sollte.

Inwieweit es dem WBP mit seinem Kriterienkatalog – insbesondere den internen Validitätskriterien – gelungen ist, unterschiedliche Herangehensweisen der internen Validitätssicherung mit einzubeziehen und zu berücksichtigen und damit die Prinzipien der RCT-Methodologie gegenstandsangemessen zu erweitern, soll in der vorliegenden Arbeit empirisch fundiert untersucht werden. Durch das empirische Fundament in Form von Wirksamkeitsstudien, die mittels der WBP-Kriterien bewertet werden sollen, soll zum einen die Möglichkeit eröffnet werden, eine bislang eher theoretisch geführte Debatte mit empirischen Befunden zu unterfüttern – und der Debatte damit möglicherweise eine andere Richtung zu verleihen. Zum anderen soll jedoch auch die Möglichkeit eröffnet werden, sich durch die vorzufindende empirische Evidenz gewissermaßen belehren zu lassen: Ein Fokus wird daher darauf liegen, was die empirische Evidenz zu „erzählen“ hat, und ob sie wertvolle Beiträge dergestalt leisten kann, ob und ggf. inwieweit die WBP-Kriterien einer Modifikation bedürfen, um unterschiedlichen Gegenständen (hier: allem voran den Langzeittherapiestudien) in der Begutachtung der internen Validität gerecht zu werden.

Die Beantwortung der Frage wird durch eine Art Abgleich zwischen dem, wie Kurzzeittherapien und Langzeittherapien auf ihre Wirksamkeit hin beforscht werden, d.h. zwischen Wirksamkeitsstudien zu Kurz- und Langzeitbehandlungen, und den WBP-Kriterien erfolgen. Es soll dimensions- und insbesondere kriterienbezogen anhand von Verteilungen von Wirksamkeitsstudien auf den einzelnen Kriterienstufen („optimal“, „zufriedenstellend“ und „ungenügend“) eruiert werden, ob und ggf. bei welche Kriterien Langzeittherapiestudien im Vergleich zu Kurzzeittherapiestudien sichtbar schlechter abschneiden. Daran anschließend werden ggf. die Gründe für ein solches Gefälle diskutiert (s.o.).

Für diesen Vergleich werden primär die interne/n Validitätsdimension und -kriterien herangezogen. Da ein hinreichend gutes Abschneiden auf der Dimension der allgemeinen methodischen Qualität die Voraussetzung für eine Bewertung auf der internen Validitätsdimension bildet und damit als Teil der internen Validitätsbewertung (wie auch der externen Validitätsbewertung) betrachtet werden muss, werden auch die Kriterien dieser Dimension untersucht. Um den Rahmen der Arbeit zu wahren, wird die Analyse der allgemeinen methodischen Qualitätsdimension jedoch auf die K.O.-Kriterien (vgl. Kap. 1.2.2) beschränkt werden.

Die Datenbasis bilden Wirksamkeitsstudien zu psychoanalytisch begründeten Verfahren, da hier sowohl Behandlungen von kürzerer als auch langer Dauer (> 100 Sitzungen) zu erwarten sind. Zudem wurde der zu untersuchende Studienpool auf die Anwendungsbereiche der affektiven Störungen und auf Untersuchungen an diagnoseheterogenen Störungsgruppen begrenzt. Aufgrund ihrer epidemiologischen Relevanz war zu erwarten, dass die Wirksamkeit von Behandlungen affektiver Störungen vergleichsweise gut beforscht ist, so dass in diesem Bereich mit genügend Studien zu rechnen war. Die Untersuchungen an diagnoseheterogenen Störungsgruppen wurden hinzugezogen, um dem psychoanalytischen Verständnis von der Kuration psychischer Störungen, das in erster Linie kein störungs- bzw. symptomspezifisches Verständnis ist, gerecht zu werden. Es sollten somit für die vorliegende Untersuchung gezielt auch solche Studien berücksichtigt werden, denen eben kein störungsspezifisches Konzept der Ätiopathogenese und Heilung psychischer Störungen zugrunde liegt. Damit wurde das Ziel verfolgt, der Untersuchung insbesondere *psychoanalysetypische* Evidenz als Datenbasis zugrunde zu legen.

In den folgenden Kapiteln werden die genaue Zusammenstellung der empirischen Basis sowie die Kodierung der Studien mit Hilfe des WBP-Kriterienkatalogs und den damit zusammenhängenden vorbereitenden Maßnahmen detailliert beschrieben.

3 Methodik

Die Datenbasis der Arbeit setzt sich aus Studien zur Wirksamkeit der psychodynamischen Verfahren in den Anwendungsbereichen der affektiven Störungen sowie der gemischten Störungen zusammen. Die Ein- und Ausschlusskriterien für die Zusammenstellung der Studien werden im Zusammenhang mit der Bestimmung der Studien-Zielpopulation und der Kodiereinheit "Studie" beschrieben (Kap. 3.1.1). Die Datenerhebung in Form einer erschöpfenden Studienrecherche sowie die Evaluation der Studienrecherche werden in Kapitel 3.1.2 und 3.1.3 dargestellt. Es folgt eine Illustration vorbereitender Maßnahmen, die zur Anwendung des Kriterienkatalogs für die Kodierung der Studien notwendig waren (Kap. 3.2). Dazu zählt zum einen die Entwicklung von Kodierregeln, die für nahezu jedes einzelne Kriterium definiert wurden (vgl. Kap. 3.2.1, Kap. 3.2.2 und Kap. 3.2.3). Zum anderen wurde zusätzlich zum WBP-Kriterienkatalog ein sog. Kurzkodierbogen entwickelt, mit dem die zu kodierenden Studien im Hinblick auf grundlegende Studieneigenschaften sowie methodologische Aspekte indiziert werden. Dieser Kurzkodierbogen und dessen Entwicklung werden in Kapitel 3.2.4 dargestellt.

Im Anschluss wird der Prozess der Kodierung aller Studien sowohl mit Hilfe des erwähnten Kurzkodierbogens als auch mittels des WBP-Kriterienkatalogs unter Zuhilfenahme der entwickelten Kodierregeln demonstriert (Kap. 3.3).

Das Kapitel schließt mit einer Darstellung der geplanten Datenanalyse (Kap. 3.4).

3.1 Datenerhebung

Die hier intendierte Datengrundlage in Form von Primärstudien zur Wirksamkeit der psychoanalytisch begründeten Richtlinienverfahren[19] erforderte zunächst eine intensive Studienrecherche. Dazu wurde ein Vorgehen gewählt, das in einschlägigen Werken zur Durchführung systematischer Reviews unter dem Begriff *erschöpfende* (*exhaustive*) Literaturrecherche beschrieben wird (u.a. Cooper & Hedges, 1994; Rustenbach, 2003). Als *erschöpfend* wird eine Recherche dann betrachtet, wenn keine neue, relevante Literatur mehr gefunden wird:

> In practice the stopping point may be approaching when the search has covered all the most relevant databases and bibliographies, and when further searches of databases, and scanning of bibliographies of review papers do not add to the tally of included papers. (Petticrew & Roberts, 2006, S. 100)

Nun garantiert eine erschöpfende Studienrecherche nicht notwendigerweise das „Einfangen" aller existierenden – insbesondere nicht publizierten – Forschungsbefunde (vgl. Pant, 1998; Rustenbach, 2003). Daher ist das maximal zu erreichende Ziel in Metaanalysen[20] meist auch nicht die Vollerhebung, sondern die Erhebung einer repräsentativen Studienstichprobe der intendierten Zielpopulation (Letztere umfasst in der Regel alle existierenden, relevanten Studien, auch die nicht publizierten). Diese repräsentative Stichprobe wird in diesem Fall nicht – wie es die Stichprobentheorie nahelegt – durch eine Zufallsauswahl der Studien aus der schwer greifbaren Zielpopulation realisiert, sondern durch eine anfallende Stichprobe aller zu erreichender, publizierter und nicht publizierter Studien. Rustenbach (2003) bezeichnet aus diesem Grund systematische Reviews auch als *observational studies*. Repräsentativität wird folgendermaßen zu gewährleisten versucht:

---

[19] In Ausnahmefällen auch Wirksamkeitsstudien zur Psychoanalyse (vgl. Kap. 3.1.1).

[20] Die Begriffe "systematisches Review" und "Metaanalyse" werden synonym verwendet.

> „Since there is no way of ascertaining whether the set of located studies is representative of the full set of *existing* studies on the topic, the best protection against an unrepresentative set is to locate as many of the existing studies as possible" (p. 46). (Hervorhebungen im Original, Jackson, 1978; zitiert nach White, 1994, S. 43)

Die Realisierung einer repräsentativen Studienstichprobe mittels einer erschöpfenden Studienrecherche bedeutet also, den Umfang unzugänglicher bzw. schwer zugänglicher Studien weitestgehend zu minimieren. Schwer zugänglich sind Studien in der Regel dann, wenn sie zwar durchgeführt und evtl. sogar verschriftlicht, jedoch nicht publiziert wurden. Eine Studie kann dagegen auch dadurch unentdeckt bleiben, dass sie weder in metaanalytischen Publikationen auftaucht noch in diversen Online-Literaturdatenbanken zu finden ist. So kommt es trotz der Fülle an Publikationen in einschlägigen Datenbanken, wie PubMed oder PsycINFO, durchaus vor, dass bestimmte Zeitschriften oder aber auch nur bestimmte Jahrgänge eines *Journals* (noch) nicht in den Datenbanken indiziert sind. Das kann sog. *Minor-Journals* (vgl. Rustenbach, 2003) betreffen oder aber schlicht der Tatsache geschuldet sein, dass zu Beginn eines Jahres noch nicht alle Referenzen des Vorjahres oder zum Zeitpunkt X noch nicht alle Publikationen bis zu diesem Zeitpunkt indiziert sind. Die Recherche in unterschiedlichen Datenbanken kann dabei zu einer Verminderung der Wahrscheinlichkeit unentdeckter Publikationen beitragen. Eine weitere Option, diese Wahrscheinlichkeit zu minimieren, liegt in der in bestimmten zeitlichen Abständen vorzunehmenden Aktualisierung einer digitalen Recherche für einen umgrenzten Publikationszeitraum.

Mit dem Ziel, den sog. *publication bias* möglichst gering zu halten, ist der Metaanalytiker mit der aufwändigen Aufgabe konfrontiert, sich mittels unterschiedlicher Strategien Zugang zu nicht veröffentlichter Literatur (*grey literature*) zu verschaffen (zu unterschiedlichen Strategien vgl. Rustenbach, 2003, S. 36 ff.). Ein solcher *publication bias* wird in Metaanalysen genau dann sichtbar, wenn die Desintegration nicht publizierter Befunde eine systemati-

sche Verzerrung in der Schätzung des Gesamteffekts hervorruft (in der Regel besteht die Verzerrung in einer Überschätzung des Gesamteffekts). Dieser Bias impliziert, dass die Größe von Effekten systematisch mit der Wahrscheinlichkeit, publiziert zu werden, zusammenhängt (vgl. Sutton, 2009). In systematischen Reviews ist es üblich, die durch den *publication bias* entstehende Verzerrung in der Schätzung der globalen Effektstärke mit Hilfe unterschiedlicher Techniken zu entdecken und ggf. zu korrigieren (z.B. die Methode des *Fail Safe-N*; Erstellen eines *Funnel Plot* etc., vgl. Beelmann & Bliesener, 1994; Sutton, 2009). In der vorliegenden Arbeit geht es jedoch nicht um die Integration unterschiedlicher Forschungsbefunde zum Zwecke der Gesamteffektschätzung, sondern vielmehr um die Bewertung von Studienqualität anhand des WBP-Kriterienkatalogs. Um mögliche systematische Verzerrungen in diesen Bewertungen, die durch die Publikationspraxis oder -politik hervorgerufen werden, entdecken und ggf. korrigieren zu können, müsste es Techniken der o.g. Art geben, die diese Verzerrungen aufzudecken und zu korrigieren vermögen. Der augenfällige Mangel solcher Techniken ist, neben weiteren Gründen (s.u.), ausschlaggebend dafür, dass in dieser Arbeit von vornherein eine andere Studien-Zielpopulation intendiert wurde, als es in systematischen Reviews üblich ist. Um den nicht abzuschätzenden und zu korrigierenden Bias, der durch das Nichtauffinden unpublizierter Literatur hervorgerufen würde, von vornherein zu vermeiden, wird in dieser Arbeit eine Zielpopulation zugrunde gelegt, die ausschließlich publizierte Literatur umfasst. Da zudem Dissertationsschriften nicht immer frei oder nur unter oftmals vergleichsweise hohem finanziellen und zeitlichen Aufwand zugänglich sind[21], werden diese ebenfalls als nicht zur Zielpopulation gehörend betrachtet. Jedoch sind Publikationen, die sich auf Daten aus Dissertationen stützen, selbstverständlich Teil der Zielpopulation.

[21] Zum Zeitpunkt der Studienrecherche wurde das Dissertationsprojekt noch nicht durch Drittmittel unterstützt, so dass die finanziellen Mittel auf das Budget des Arbeitsbereichs Public Health: Prävention und psychosoziale Gesundheitsforschung (Freie Universität Berlin) beschränkt waren.

Der Aussagegehalt der Arbeit beschränkt sich damit auf publizierte Wirksamkeitsuntersuchungen. Das bedeutet, dass die Untersuchung des Kriterienkatalogs auf seine Gegenstandsadäquatheit (vgl. Kap. 2). ausschließlich mittels solcher Studien vorgenommen wird, die bereits die Hürde der Publikationstauglichkeit überwunden haben.

Ein weiterer Grund für die Beschränkung auf publizierte Literatur besteht darin, dass in der vorliegenden Arbeit die zu kodierenden Studien bewusst keiner vorhergehenden Qualitätsbewertung unterzogen werden, wie es in systematischen Reviews in der Regel der Fall ist. Dort werden bspw. Mindeststandards bzgl. der Reliabilität und Validität der verwendeten Outcomemaße, des realisierten Studiendesigns oder der Dropoutanalysen aufgestellt und ggf. in Form eines Qualitätgesamtscores zusammengefasst (z.B. Timmer & Richter, 2008). Ähnliche Mindeststandards in der vorliegenden Arbeit zugrunde zu legen, hätte die Gefahr mit sich gebracht, dass diese Mindeststandards mit den Kriterien des Kriterienkatalogs konfligieren. Stattdessen sollte die „Selektionsarbeit“ vollständig dem Kriterienkatalog überlassen werden. Um dem Studienpool trotzdem eine eindeutige Grenze zu verleihen – und da Mindestqualitätsstandards aus den genannten Gründen dafür ungeeignet schienen – musste die Grenze auf einer anderen Ebene als den Qualitätsstandards gesucht werden. Aus diesem Grund wurde die Publikationshürde als eben diese Grenze gewählt. Neben dem Nachteil, dass diese Hürde keineswegs unabhängig von Qualitätsstandards ist – oder zumindest sein sollte – und diese zudem nicht explizit sind, birgt sie doch zumindest den entscheidenden Vorteil einer klar bestimmbaren Begrenzung in sich.

Resümierend kann somit festgehalten werden: In der vorliegenden Arbeit ist das mindestens zu erreichende Ziel, mit Hilfe expliziter Einschlusskriterien sowie mit Hilfe von Strategien der erschöpfenden Literaturrecherche die Repräsentativität des zu erhebenden Studienpools für die intendierte Zielpopulation zu gewährleisten (vgl. Lipsey & Wilson, 2001). Das Ideal,

nach dem gleichsam gestrebt wird, ist eine Vollerhebung aller zur Zielpopulation gehörenden Studien:

> It is unlikely that such a population contains so many studies that appreciable efficiency is gained from drawing a representative sample . . . . Typically, therefore, metaanalysts attempt to identify and retrieve every study in the defined population rather than sample from that population. (Lipsey & Wilson, 2001, S. 23)

Die Vollerhebung wird dabei u.a. von Rustenbach (2003) als Vorkehrung betrachtet, Repräsentativität zu garantieren: „Die wirksamste Maßnahme zur Gewährleistung der Repräsentativität integrierter Primärstudien ist eine Vollerhebung, *in diesem Fall ist die Studienstichprobe identisch zur Zielpopulation*“ [Hervorhebung v. Verf.] (S. 23).

Die Einschlusskriterien, die Definition der Zielpopulation als auch der Kodiereinheit "Studie" werden im folgenden Unterkapitel (Kap. 3.1.1) dargestellt.

3.1.1 Definition der Zielpopulation und der Kodiereinheit "Studie"

Die *Zielpopulation* der Studien schließt alle Wirksamkeitsuntersuchungen zu den psychoanalytisch begründeten Verfahren ein, die unter Einschluss von Probanden im Erwachsenenalter mit der Diagnose einer affektiven Störung oder aber unter Einschluss diagnoseheterogener Störungsgruppen durchgeführt und im Zeitraum zwischen 1999[22] bis 2009 veröffentlicht wurden.

[22] Als untere Datumsgrenze wurde der Zeitpunkt gewählt, der mit dem Inkrafttreten des Psychotherapeutengesetzes und der damit einhergehenden Gründung des WBP zusammenfällt. Es wird davon ausgegangen, dass die Gründung des WBP und die damit einhergehende Debatte um gegenstandsadäquate Methoden der Evidenzbasierung psychotherapeutischen Handelns einen Aufschwung an gruppenstatistischen Wirksamkeitsstudien auslöste, die eine reichhaltige Basis für die vorliegende Arbeit bieten.

Unter "Studie" werden alle Publikationen verstanden, die sich auf eine spezifische Wirksamkeitsuntersuchung beziehen und Informationen zur Untersuchungsstichprobe, Intervention und zu Outcomemmaßen sowie zur Durchführung und Auswertung der Untersuchung bereitstellen. Beziehen sich mehrere Publikationen auf ein und denselben Datensatz, so machen diese Publikationen dementsprechend *eine* Studie aus. So wurde das Eingehen doppelter Datensätze in die zusammenzustellende Studienstichprobe vermieden. Als Studie gelten sowohl Original- und Replikationsstudien als auch Reanalysen von zu einem früheren Zeitpunkt bereits veröffentlichten Untersuchungen. Für Original- und Replikationsstudien gilt, dass sie genau dann zur Zielpopulation gehören, wenn die erste Publikation der Untersuchungsergebnisse innerhalb des genannten Veröffentlichungszeitraumes fällt. Reanalysen werden nur dann als zur Zielpopulation gehörend betrachtet, wenn die Veröffentlichung(en) der Originaldaten außerhalb des genannten Publikationszeitraums publiziert wurden. So wurde ebenfalls die Integration doppelter Datensätze in die Studienstichprobe vermieden.

Um in die Zielpopulation einzugehen, muss aus der oder den Publikation(en) hervorgehen, dass es sich um Erwachsenenstichproben (ab 18 Jahren) von Klienten handelt, deren Behandlungsbedürftigkeit durch Feststellung einer psychischen Störung nach ICD oder DSM belegt wurde. Dabei wurde sich auf Studien beschränkt, die entweder diagnosehomogene Gruppen mit einer Erstdiagnose aus dem affektiven Formenkreis (F3 und F53 nach ICD-10, [vgl. WBP, 2010] bzw. 296.xx, 300.4, 301.13, 311 nach DSM-IV) einbeziehen, sowie auf Studien mit diagnoseheterogenen Gruppen („gemischte Patientengruppen“, vgl. WBP, 2010). Bei beiden Untersuchungsgruppen wurde keine Monomorbidität gefordert.

Studien qualifizieren sich jedoch ebenfalls als zur Zielpopulation gehörend, wenn die Behandlungsbedürftigkeit durch andere Phänomene, als ICD- oder DSM-Diagnosen, festgestellt wurde – etwa wenn ein Suizidversuch vorlag oder aber wenn psychische Belastungen durch schwere sexuelle Missbrauchserfahrungen eine Behandlung notwendig machten, ohne

dass Belastungs- und Anpassungsstörungen via ICD systematisch diagnostiziert wurden. Analogstudien hingegen gehören eindeutig nicht zur Zielpopulation. Selbiges gilt für Studien, in denen psychotherapeutische Interventionen als Indikation bei primär somatischen Erkrankungen (z.B. Morbus Crohn) untersucht wurden (z.B. von Wietersheim et al., 2001).

In mindestens einem Treatmentarm muss ein psychoanalytisch begründetes Verfahren realisiert worden sein, das zur Heilung oder Linderung psychischen Leids durchgeführt wurde. Zudem wurden Studien integriert, in denen Behandlungen durchgeführt wurden, die wahrscheinlich eher als klassische Psychoanalysen betrachtet werden müssten, und die nicht zum Leistungskatalog der gesetzlichen Krankenkassen gehören. Der Grund für diese Integration wird im Rahmen des folgenden Exkurses dargelegt.

Exkurs

Der Inklusion auch solcher Primärstudien, in denen klassische Psychoanalysen durchgeführt wurden, liegt folgende Taxonomie der Verfahren zugrunde: Die Unterschiedlichkeit der psychoanalytisch begründeten Richtlinienverfahren und der Psychoanalyse wird als eine dimensionale und weniger als eine kategoriale und disjunkte Anordnung dieser Verfahren verstanden, in der sich die unterschiedlichen Vorgehensweisen zwischen dem *klassisch psychoanalytischen* und dem *tiefenpsychologischen* Pol bewegen. Dabei wird unter dem Pol der klassischen Psychoanalyse Folgendes verstanden: Rüger und Reimer (2006) beschreiben die klassische Psychoanalyse als *tendenzlos* und führen weiter aus: „Diese Behandlungen werden in einem strengen Reglement durchgeführt, das durch das Liegen auf der Couch, den freien Einfall und die Traumarbeit aufseiten des Patienten sowie durch die Neutralität, Anonymität und Abstinenz aufseiten des Therapeuten gekennzeichnet ist“ (S. 4) und weiter „[sie] wird tendenzlos und ohne vorher festgelegte Begrenzung durchgeführt. Ihre Zielsetzung ist offen“ (S. 8). Plänkers (1986) beschreibt die Tendenzlosigkeit im psychoanalytischen Vorgehen

folgendermaßen: „Das Tastende, Vage, Unsichere . . . entspricht zugleich dem psychoanalytischen Prozeß: Hier werden nicht in erster Linie funktionale Beziehungen von Es, Ich und Über-Ich untersucht, sondern Analytiker und Analysand nähern sich unsicher, Suchenden auf fremdem Gebiet vergleichbar, dem Unbewußten“ (S. 695).

Der tiefenpsychologische Pol ist im Vergleich zum psychoanalytischen vielmehr durch folgende Vorgehensweisen und therapeutische Haltungen charakterisiert: Der Therapeut nimmt eine aktive Haltung ein und verlässt die streng abstinente Haltung, er interveniert zuweilen sogar stützend und ermutigend (vgl. Busch, 2010); es werden in der Behandlung inhaltliche Ziele – sog. Foki – verfolgt, der Behandler wirkt dahingehend durch explorierende Fragen lenkend auf den therapeutischen Prozess ein und initiiert eher ein symmetrisches Kommunikationsmuster (vgl. Caligor et al., 2012; Gill, 1954). Zudem konzentriert sich der Therapeut auf die positiven Aspekte in den Erzählungen des Patienten und nimmt somit implizit eine Bewertung der berichteten Inhalte vor. Werden in der Therapie die Beziehungen zu den primären Bezugspersonen wiederbelebt, werden sie eher auf Konflikte und Probleme in der gegenwärtigen äußeren Realität bezogen und weniger auf die Übertragungen auf den Therapeuten (vgl. Busch, 2010). Zentrale Themen in der Behandlung sind überwiegend Inhalte und Geschehnisse aus der Realwelt des Klienten, Träume und Phantasien des Patienten, die das zentrale Material des psychoanalytischen Pols darstellen, werden weit weniger thematisiert (vgl. Caligor et al., 2012).

Zwischen den beiden umrissenen Polen bewegen sich die unterschiedlichen psychodynamischen Verfahren, Methoden und Techniken. Die analytische Therapie kommt dabei wohl dem klassisch psychoanalytischen Pol am nächsten und bedient sich der meisten klassisch psychoanalytischen Techniken (Analysieren der infantilen Beziehungsmuster in ihrer Übertragung auf den Therapeuten, Förderung regressiver Prozesse, Arbeit am Vergangen-

heitsunbewussten, Deutung als primäre Interventionsform etc.). Das bedeutet, dass die Grenzen zwischen klassischer Psychoanalyse und analytischer Psychotherapie als fließend und durchlässig betrachtet werden müssen. Dazu Rudolf und Rüger (2006):

> So verstanden, wird sie [die analytische Psychotherapie] bisweilen mit »Psychoanalyse« gleichgesetzt. Die Bezeichnung hat sich allerdings nur im deutschsprachigen Bereich durchgesetzt und mit der Einführung der Richtlinienpsychotherapie »offiziösen« Charakter angenommen. Demnach ist analytische Psychotherapie eine Anwendungsform der Psychoanalyse mit eigenen Zielkriterien. Ihre Psychotherapiekonzepte, die Regeln der Behandlungstechnik sowie der Behandlungsprozess unterscheiden sich nicht grundsätzlich von denen der Psychoanalyse. (S. 41)

Die Autoren nehmen jedoch folgendes zentrales Unterscheidungsmerkmal vor:

> Allerdings sind die Indikationen für eine analytische Psychotherapie stets dadurch bestimmt, dass sie sich als Krankenbehandlung versteht und damit allgemeinere Zielsetzungen, wie z. B. die Förderung der Persönlichkeitsentwicklung oder z. B. die Entwicklung zur Autonomie und deren Verwirklichung, als solche für sich allein nicht zu ihren Aufgaben gehören. Verschiebt sich der Inhalt der therapeutischen Arbeit in Richtung analytischer Zielkriterien, oder ist die Behandlung von Krankheit abgeschlossen…, endet die Verpflichtung der gesetzlichen Krankenversicherung zur Übernahme der Kosten (Faber-Haarstrick, Kommentar Psychotherapie-Richtlinien, 7. Aufl.; vgl. Rüger et al. 2005, S. 41). (S. 41)

Das Ziehen einer eindeutigen Trennlinie zwischen analytischer Psychotherapie und klassischer Psychoanalyse wird zudem dadurch erschwert, dass diese Unterscheidung ein Alleinstellungsmerkmal des deutschen Versorgungssystems darstellt. Im angloamerikanischen Sprachraum existieren hingegen *psychodynamic* oder *psychoanalytic psychotherapy* einerseits und *psychoanalysis* andererseits. Übersetzt man die *psychodynamic* oder *psychoanalytic psychotherapy* in das deutsche Gesundheitssystem, so decken sich diese Verfahren eher mit den

tiefenpsychologisch fundierten Behandlungen (vgl. Rüger & Reimer, 2006), *Psychoanalysis* meint das, was im deutschen Gesundheitssystem mit klassischer Psychoanalyse gemeint ist. Ist explizit von *long-term psychoanalytic psychotherapy* oder *long-term dynamic psychotherapy* die Rede, so handelt es sich dabei in der Regel um Behandlungen, die der hiesigen analytischen Psychotherapie nahekommen. Allerdings werden bspw. in den USA bereits Therapien über 25 Sitzungen als Langzeitbehandlungen bezeichnet, was rein vom Umfang her wiederum eher der tiefenpsychologisch fundierten Behandlung nahekommt (vgl. Beutel et al., 2010).

Die Komplexität komplettiert sich außerdem durch eine uneinheitliche begriffliche Trennung der unterschiedlichen psychodynamischen Verfahren und der Psychoanalyse in der Literatur. So liest man in deutschsprachigen Publikationen oftmals die Bezeichnung „Psychoanalyse“, wenn eigentlich die kassenfinanzierte analytische Psychotherapie gemeint ist (z.B. Rief & Hofmann, 2009). Ebenso werden analytische Therapien von praktizierenden Psychoanalytikern und ihren Patienten oftmals als klassische Psychoanalysen bezeichnet, wenn sie mindestens 3x wöchentlich und im liegenden Setting durchgeführt werden. Der Gemeinsame Bundesausschuss bezeichnet die psychodynamischen Richtlinienverfahren als „psychoanalytisch begründete Verfahren“. Der WBP wiederum nutzt das Begriffspaar „Psychodynamische Psychotherapie“ und subsumiert darunter die psychodynamischen Richtlinienverfahren. Im ersten Band einer im Jahr 2010 ins Leben gerufenen Manualreihe mit dem Titel *Praxis der psychodynamischen Psychotherapie – analytische und tiefenpsychologisch fundierte Psychotherapie* von Beutel und Kollegen wird es dem WBP zunächst gleichgetan und das Begriffspaar „Psychodynamische Psychotherapie“ ebenfalls als Oberbegriff für die beiden Richtlinienverfahren genutzt. Im Verlauf des ersten Bandes verschieben die Autoren jedoch ihre Trennlinie und stellen der analytischen Psychotherapie und Psychoanalyse die psychodynamischen Verfahren im Sinne der tiefenpsychologisch fundierten Behandlung gegenüber (vgl.

Beutel et al., 2010, S. 44). Letztere Gleichsetzung von psychodynamischer Psychotherapie mit tiefenpsychologisch fundierter Behandlung hat sich in zahlreichen deutschsprachigen Publikationen durchgesetzt (u.a. Brandl et al., 2004). Insgesamt bleibt die Nomenklatur in ihrer Verwendung jedoch mannigfaltig, was nicht zuletzt den fließenden Grenzen zwischen den einzelnen Verfahren zu verdanken ist, deren theoretische Wurzeln allesamt in der psychoanalytischen Persönlichkeitstheorie sowie Krankheits- und Behandlungslehre zu finden sind (vgl. Rüger & Reimer, 2006).

Vor dem Hintergrund des vorangegangenen Exkurses ist daher anzunehmen, dass auch in empirischen Wirksamkeitsuntersuchungen nicht immer eine kategorische Trennung zwischen den konzeptionell nahe beieinander liegenden Verfahren der analytischen Psychotherapie und der klassischen Psychoanalyse zu erwarten sein wird. Aus diesem Grund wurde entschieden, auch solche Studien zur Zielpopulation zu zählen, in denen die Behandlungen als „Psychoanalyse“ deklariert werden.

Ende des Exkurses

In den Untersuchungen wird eine Mindestbehandlungsdauer von 7 Sitzungen gefordert, Behandlungen unter 7 Sitzungen werden in der Regel als Ultrakurzzeittherapie bezeichnet (vgl. Shapiro et al., 2003) und werden daher hier nicht als repräsentativ für die psychodynamische Behandlungspraxis im deutschen Sprachraum betrachtet. Studien, deren Range der Sitzungsanzahl unter die genannte Grenze fällt, in denen jedoch von einer deutlich höheren mittleren Sitzungsanzahl berichtet wird, zählen jedoch zur Zielpopulation, da der untere Range-Wert hier lediglich für die untere Grenze der Verteilung des Behandlungsumfangs steht. Trotz des in den Richtlinien festgelegten Finanzierungsumfangs von maximal 300 Sitzungen in der ana-

lytischen Psychotherapie, wurde der Behandlungsumfang der hiesigen Studien-Zielpopulation nicht begrenzt und findet ggf. seinen Niederschlag in Kriterium C.4 „Klinische Repräsentativität der Intervention hinsichtlich Vorgehen und Dauer“[23] (vgl. Kap. 3.2.2).

Aus einer Studie muss – im Mindesten durch Benennung – hervorgehen, dass es sich um eine psychodynamische Behandlung handelt; geht aus dem Text hervor, dass es sich bspw. um eine eklektische Behandlung handelt, die unterschiedliche psychotherapeutische Verfahren – u.a. psychodynamische Techniken – miteinander kombiniert, wurde diese Studie ausgeschlossen. In welcher Ausführlichkeit die psychodynamischen Behandlungen in den jeweiligen Publikationen umschrieben werden, ist wiederum „Sache des Kriterienkatalogs“. Als nicht zur Zielpopulation gehörig betrachtet werden Studien, in denen Kombinationswirkungen (psychodynamisches Verfahren kombiniert mit Psychopharmaka) oder Zusatzwirkungen psychodynamischer Behandlungen (Wirksamkeit von psychodynamischer Behandlung als Zusatzbehandlung zu Pharmakotherapie) systematisch untersucht wurden. Gleiches gilt für psychodynamische Behandlungen, die ausschließlich als Aufrechterhaltungstherapien oder aber als Beratungsangebote intendiert waren. Auszuschließen waren ebenfalls Studien, in denen psychodynamische Klinikkonzepte evaluiert wurden, womit Studien zu psychodynamischer Therapie im stationären Setting vollständig zu exkludieren waren. Die Entscheidung für den Ausschluss dieser Studien fiel allerdings erst zu einem späteren Zeitpunkt und wird in Kapitel 3.1.2 daher auch in Form eines nachträglichen Ausschlusskriteriums beschrieben.

Die psychodynamische Behandlung muss in einem face-to-face-Setting durchgeführt worden sein, Behandlungen, die via E-Mail o.ä. stattfanden, sind auszuschließen. Sowohl

[23] Hier würden (mittlere) Behandlungsumfänge einer analytischen Psychotherapie, die jedoch weit über die Grenze der Maximalbehandlungsdauer einer analytischen Behandlung (300 Sitzungen) fallen, dementsprechend mit einer „3“ bewertet werden (vgl. Ratzek & von Hauenschild, 2011a; Anhang C).

psychodynamische Individual- als auch Gruppenbehandlungen werden als zur Zielpopulation gehörig betrachtet.

Zur Zielpopulation gehörende Studien müssen psychosoziale Outcomes erheben – dies in Abgrenzung bspw. zu Outcomes in monetären Einheiten. Werden zuletzt genannte Outcomes zusätzlich zu psychosozialen Outcomes berichtet, ist dies kein Ausschlussgrund. Unter psychosozialen Outcomes sind bspw. symptombezogene Maße zu subsumieren (z.B. *Clinical Global Impression* [CGI], Guy, 1976; *Beck Depression Inventory II* [BDI-II], Beck, Steer & Brown, 1996; *Hamilton Anxiety Rating Scale* [HARS], Hamilton, 1976; *Symptom Checklist-90 Revised* [SCL-90-R], Derogatis, 1975); außerdem Maße zur Bestimmung allgemeinerer Beschwerden (z.B. Gießener Beschwerdebogen [GBB], Brähler & Scheer, 1995) der Lebenszufriedenheit (z.B. Fragebogen zur Lebenszufriedenheit [FLZ], Fahrenberg, Myrtek, Schumacher & Brähler, 2000; *Life satisfaction scale* [Indikatoren des Reha-Status: IRES], Gerdes & Jäckel, 1992) oder zum allgemeinen Wohlbefinden (z.B. *Psychological General Well-Being Index* [PGWBI], Dupuy, 1984). Zudem Maße zur Bestimmung des interpersonalen Verhaltens (z.B. *Inventory of Interpersonal Problems* [IIP], Horowitz, 1999), Persönlichkeitsfragebögen (z.B. Gießen Test [GT], Beckmann & Richter, 1972; Freiburger Persönlichkeitsinventar [FPI], Fahrenberg & Selg, 1970) sowie Therapieerfolgsmaße, wie z.B. die *Goal Attainment Scale* (GAS, Kiresuk & Sherman, 1968). Neben den genannten sind zusätzlich verfahrensspezifische (psychodynamische) Maße zu erwarten, etwa die Operationalisierte Psychodynamische Diagnostik – 2 (OPD-2, Arbeitskreis OPD, 2009), die im Gegensatz zur OPD-1 (Arbeitskreis OPD, 1996) auch zwecks Veränderungsmessung eingesetzt werden kann; oder die *Psychodynamic Functioning Scales* (PFS, Høglend et al., 2000). Psychometrische Anforderungen an die Reliabilität und Validität der in den Studien verwendeten Outcomemaße wur-

den nicht gestellt, da diese durch Kriterium A.8 „Reliable und valide Messung zumindest der primären Zielkriterien“ des Kriterienkatalogs bewertet werden.

Inwieweit eine Studie als Wirksamkeitsstudie bezeichnet werden kann respektive welche Durchführungs- und Auswertungsbeschreibungen in einer Studie enthalten sein müssen, soll einleitend durch eine genaue Definition des Begriffs "Wirksamkeit" dargelegt werden.

Der Begriff "Wirksamkeit" wird hier breit gefasst und mit Hager und Hasselhorn (2000) auf die sog. "Bruttowirkung *im weiteren Sinne*" bezogen: Unter "Bruttowirkung *im weiteren Sinne*" subsumieren die beiden Autoren alle Wirkungen, „. . . die im Gefolge einer Intervention auftreten können: die programm-, die interventionsgebundenen und die interventionsunabhängigen Wirkungen, also die Wirkungen der Intervention insgesamt“ (S. 48). Die *programmgebundene Wirkung* bezieht sich auf spezifische Veränderungen, die durch das Programm (bestimmte Therapieform) intendiert sind. Die programmgebundenen Wirkungen bspw. eines bestimmten therapeutischen Verfahrens sind daher in dem Wirkmodell dieses Therapieverfahrens enthalten (z.B. Lösung unbewusster Konflikte) und als Folge der spezifischen Techniken dieses Verfahrens zu betrachten (z.B. Arbeit mit dem Unbewussten, Übertragungsarbeit). *Interventionsgebundene Wirkungen* sind nicht mehr mit den intendierten Zielen einer Intervention verbunden, sondern stellen eher übergeordnete, von den Spezifika der Intervention unabhängige Wirkungen dar. Es sind Wirkungen, die sich allein auf die Tatsache zurückführen lassen, dass überhaupt eine Intervention stattgefunden hat. Darunter sind also unspezifische Wirkfaktoren zu verstehen, die in sozialen Interventionen immer stattfinden – etwa Aufmerksamkeit, die Tatsache, dass jemand zuhört oder selbst die Tatsache, dass man sich um einen Therapieplatz erfolgreich bemüht hat. Diese Wirkungen werden bspw. durch die Realisierung von sog. Placebobehandlungen zu kontrollieren versucht, in denen allgemeine Faktoren, jedoch keine programmgebundenen Faktoren wirksam werden sollen. Wird die

interventionsgebundene Wirkung nicht kontrolliert, kann es zur Überschätzung der programmgebundenen Wirkung kommen. Unter *interventionsunabhängiger Wirkung* verstehen Hager und Hasselhorn schließlich alle Veränderungen, die durch „zwischenzeitliches Geschehen“ (S. 47) hervorgerufen werden, die mit der Intervention im eigentlichen Sinne jedoch nichts zu tun haben (Reifung, Spontanremission etc.). Diese Wirkung lässt sich bereits durch unbehandelte Kontrollgruppen (z.B. Wartelisten) kontrollieren. Werden in einer Wirksamkeitsuntersuchung keinerlei Kontrollen durch Vergleichsgruppen umgesetzt, so wird allenfalls die bereits eingeführte "Bruttowirkung *im weiteren Sinne*" erhoben, über die spezifische programmgebundene Wirkung kann keine differenzierte Aussage getroffen werden.

Die zugrunde gelegte Definition von "Wirksamkeit *im weiteren Sinne*" ist als Rahmen der Zielpopulation zu betrachten, der alle Studien, die etwa die Untersuchung der *Nettowirkung*, d.h. der programmgebundenen Wirkung, zum Ziel haben, mit einschließt. Für die einzuschließenden Studien bedeutet das, dass eine Studie eine Beschreibung der Durchführung der summativen Ergebnisevaluation zur Feststellung der Wirksamkeit im o.g. Sinne enthalten muss (vgl. Mittag & Hager, 2000). Es muss also aus den Publikationen hervorgehen, dass eine Untersuchung stattgefunden hat, in die mindestens ein Treatmentarm eingeht, an dem wiederum mindestens Outcomemessungen nach Beendigung einer psychodynamischen Behandlung vorgenommen wurden. Darüber, ob diese Outcomemessung direkt nach Beendigung der Behandlung (Post-Messung) oder nach einer gewissen Latenzzeit (Katamnesemessung) vorgenommen wurde, werden in der Studieninklusion keine Forderungen erhoben. Diese Mindestanforderung schließt wiederum Längsschnittstudiendesigns mit ein, in denen mehrere Messzeitpunkte an der- oder denselben Stichprobe(n) vorgenommen wurden – z.B. Prä-, Post- und Katamnesemessungen. Ebenfalls schließt dies Studien mit mehr als einem Treatmentarm mit ein. Trendstudien mit mehreren Messungen zu unterschiedlichen Zeitpunkten an *unterschiedlichen* Stichproben gehören hingegen nicht zur Zielpopulation.

Die Auswertungen der Wirksamkeitshypothese sollten quantitativ erfolgen und sich auf tatsächlich beendete Behandlungen beziehen[24]. Studien, die lediglich über die Prä- und die erste Verlaufsmessung informieren, gehören nicht zur hier intendierten Zielpopulation. Welche statistischen Effektivitätsindikatoren letztlich berichtet werden (statistische Signifikanz, klinische Signifikanz etc.) bleibt hingegen der Bewertung mittels des Kriterienkatalogs überlassen.

Für die Zugehörigkeit zur Zielpopulation ist es irrelevant, ob die Untersuchungen prospektiv oder retrospektiv durchgeführt wurden: Sowohl Untersuchungen, in denen vorhandene Daten retrospektiv zum Zwecke einer Wirksamkeitsüberprüfung herangezogen wurden als auch Untersuchungen, in denen Daten (prospektiv) zum Zwecke einer Wirksamkeitsüberprüfung erhoben wurden, sind Bestandteil der Zielpopulation.

Es soll an dieser Stelle auf einen weiteren Aspekt des Zielpopulationszuschnitts hingewiesen werden, der sich allerdings erst bei der Volltextsichtung als Problem erwies und nach dem die Definition der Zielpopulation nochmals nachjustiert werden musste: Die bislang genannten Ein- und Ausschlusskriterien bilden in dieser Art einen Filter, den Prozess-Outcomestudien – neben reinen Outcomestudien – ungehindert passieren können. Wie der Name nahelegt, werden in Prozess-Outcomestudien in der Regel immer auch Outcomes respektive Ergebnisse berichtet, die Aufschluss über die Wirksamkeit eines Verfahrens geben. Im Unterschied zu reinen Outcomestudien spielen die Ergebnisse zur Wirksamkeit in Prozess-Outcomestudien nur eine nebengeordnete Rolle – sie werden eben zum Zwecke des Prozess-Outcomezusammenhangs herangezogen und aus diesem Grunde nicht (mehr) in aller Ausführlichkeit berichtet. So wird in Prozess-Outcomestudien bspw. nicht mehr das ganze Reper-

[24] In Ausnahmefällen wurden Studien auch dann aufgenommen, wenn sich eine minimale Anzahl an Patienten zum letzten Messzeitpunkt in der Endphase ihrer Behandlung befand.

toire an relevanten Darstellungen der Outcomes (in Form von Effektstärken, statistischen Signifikanzmaßen etc.) berichtet. Denkbar ist es auch, dass die Prozess-Outcomekorrespondenzen nur auf der Basis tatsächlich beendeter Therapien (*completer sample*) errechnet werden (vgl. Seybert, Huber, Ratzek, Zimmermann & Klug, 2014) und die gesamte Dropout- oder ITT-Analyse keine besondere Rolle mehr spielt. Unproblematisch ist dies, wenn Publikationen existieren, in denen zunächst und in der gebotenen Ausführlichkeit die Wirksamkeitsüberprüfungen berichtet werden (vgl. Huber, Zimmermann, Henrich & Klug, 2012) und die Prozess-Outcomezusammenhänge erst in den Folgepublikationen Beachtung finden. Als problematisch stellt es sich hingegen dar, wenn Untersuchungen von vornherein als Prozess-Outcomeanalysen angelegt waren und o.g. Aspekte, wie Ergebnisdarstellungen, Dropout- und ITT-Analysen kaum oder gar nicht berichtet werden, weil sie für die Belange einer Prozess-Outcomestudie nur noch von sekundärem Interesse sind. Da diese Studien daher keinen guten Prüfstein für die Gegenstandsadäquatheit des Kriterienkatalogs darstellen, wurde entschieden, diese Studien auszuschließen. Es wird aus diesem Grund ein weiteres, eher weiches Charakteristikum der Zielpopulation hinzugefügt: Aus den zur Zielpopulation gehörenden Studien muss mindestens die Absicht hervorgehen, die Wirksamkeit der untersuchten Behandlungen ggf. in genau demselben Ausmaß zu eruieren, wie evtl. Prozess-Outcomezusammenhänge. Werden Wirksamkeitsindikatoren nur zu Berechnungszwecken für Zusammenhangsanalysen von Prozess- und Outcomeparametern herangezogen, wird eine solche Studie nicht als der Zielpopulation zugehörig betrachtet.

Das Vorgehen in der Zusammenstellung des Primärstudienpools wird im Folgenden beschrieben.

3.1.2 Recherchestrategie für Primärstudien

Der Rechercheprozess folgte einem zweistufigen Vorgehen – auf der ersten Stufe wurde eine intensive Handsuche durchgeführt, auf der zweiten Stufe eine digitale Literaturrecherche. Um eine tatsächlich erschöpfende Zusammenstellung aller publizierten Studien zu garantieren, wurde das Augenmerk durchweg auf eine eher sensitive Recherchestrategie gelegt (hoher *recall*). Dabei wurden falsch-positive Treffer in Kauf genommen.

Der gesamte Rechercheprozess belief sich auf etwas mehr als ein Jahr (Anfang 2009 bis Anfang/Mitte 2010). Begonnen wurde im Jahr 2009 mit einer systematischen Durchsicht von zahlreichen Überblickswerken (systematische und narrative Reviews), indem diese sowohl auf potentiell relevante Primärstudien als auch hinsichtlich weiterer zitierter Reviews gesichtet wurden. Dieses Vorgehen begleitete den gesamten Rechercheprozess, so dass die Suche nach Reviews fortlaufend aktualisiert wurde.

Zunächst wurden relevante Überblicksarbeiten, die auf die Wirksamkeit psychodynamischer Verfahren fokussieren, beschafft. Dafür wurden die Reviews der für ihre metaanalytische Forschung bekannten Psychotherapieforscherinnen und -foscher, Saskia de Maat und Falk Leichsenring, herangezogen (de Maat et al., 2008; de Maat, de Jonghe, Schoevers & Dekker, 2009; de Maat, Philipzoon, Schoevers, Dekker & de Jonghe, 2007; Driessen et al., 2010; Leibing, Rabung & Leichsenring, 2005; Leichsenring, 2001, 2002, 2005, 2006, 2007, 2009a/b; Leichsenring, Hiller, Weissberger & Leibing, 2006; Leichsenring & Leibing, 2007; Leichsenring & Rabung, 2006, 2008, 2009; Leichsenring, Rabung & Leibing, 2004)[25]. Die in diesen Überblicksarbeiten relevanten Primärstudien wurden im Anschluss – wenn die Abstracts den Einschlusskriterien standhielten – im Volltext beschafft. Enthielt ein Abstract nur wenige Infomationen, wurde im Sinne eines hohen *recalls* (s.o.) die dazugehörge Studie

[25] Er werden an dieser Stelle alle Reviews der genannten Autoren zitiert, die im Verlauf des gesamten Rechercheprozesses gesichtet wurden.

ebenfalls im Volltext beschafft. Dieses Vorgehen ist der Tatsache geschuldet, dass sich Umfang und Informationsgehalt der Studienabstracts oftmals voneinander unterscheiden, so dass hier eher konservativ vorgangen wurde, um keine relevante Studie zu übersehen.

Parallel zur Abstractsichtung und Volltextbeschaffung wurden weitere relevante Reviews, auf die die o.g. Autoren hinwiesen, im Volltext beschafft.

Um einen Bias zu vermeiden, der evtl. aus einer ausschließlichen Fokussierung auf Überblickswerke primär psychodynamisch geprägter Metaanalytiker hätte resultieren können, wurde die Recherche nach Reviews in einem nächsten Schritt ausgeweitet. Hierfür wurde eine Literaturdatenbanksuche in der Datenbank PsycINFO (Datenbankanbieter: EBSCO) angeschlossen. Dabei wurde folgender Suchstring (Abbildung 4) konstruiert und bei der digitalen Suche eingesetzt:

((efficacy) OR (effectiveness) OR (therapeutic effect*) OR (practice based) OR (evidence based) OR (*therapeutic outcome*) OR (treatment outcome*) OR (clinical sampl*) OR (outpatient*) OR (inpatient*)) AND ((MR meta analysis) OR (MR literature review) OR (MR systematic review)) AND ((psychodynamic*) OR (psychoanaly*) OR (*analytical) OR (depth psycholog*) OR (short term dynamic*) OR (long* term dynamic*)) AND ((treatment*) OR (therap*) OR (psychotherap*) OR (intervention*)) AND (DT 2000-2008)

31.07.2009 (Treffer: 175)

Abbildung 4: Suchstring für Reviews (PsycINFO)

Der Suchstring setzt sich aus fünf größeren Themenblöcken zusammen, die jeweils durch den Boole'schen Operator *AND* miteinander verbunden werden. Die *AND*-Verbindung garantiert, dass alle eingegebenen und mit *AND* verbundenen Begriffe in der Freitextsuche Berücksichtigung finden. Unter *Freitextsuche* wird die Suche der Begriffe in den wichtigsten Datenfeldern, die eine Datenbank pro Publikation ausgibt, verstanden. In der Regel sind dies Titel, Autor(en), Stichwörter (*Keywords*), Abstract etc.. Diese Begriffe sind durch die Datenbank nicht standardisiert, führen zu einem hohen *recall* und unterscheiden sich damit von den sog.

controlled vocabulary terms, durch die die Suche wiederum spezifiziert wird (s.u.) (vgl. Reed & Baxter, 2009; Rustenbach, 2003; White, 2009). Durch die Trunkierung durch das Sternchen (*) werden derart abgekürzte Begriffe in all ihren Formen gefunden (z.B. findet *psychoanaly** sowohl *psychoanalysis, psychoanalyses, psychoanalytic, psychoanalyst, psychoanalytical* etc.). Das Kürzel *MR* im Suchstring ist ein datenbankspezifischer *Field Code* und steht für das Datenfeld *Methodology*, das Kürzel *DT* steht für *Date of Publication* und grenzt dementsprechend den Publikationszeitraum ein. So kann man durch unterschiedliche *Field Codes* etwa für Sprache (*LA*), Altersgruppe (z.B. *AG young adulthood, AG adulthood, AG thirties* etc.), Methodologie (z.B. *MR meta analysis, MR literature review* etc.) und Publikationszeitraum (*DT*) etc. die Suche weiter spezifizieren. Zu jedem *Field Code* findet man bei PsycINFO eine Liste von Begriffen (*controlled vocabulary terms*) (z.B. Altersgruppen: *young adulthood*, Methodologie: *literature review, clinical case study* etc.) die genau definiert und damit standardisiert sind. Mittels dieser Begriffe sind Publikationen in der Datenbank indiziert. So sind bspw. Publikationen, die sich nicht auf empirische Untersuchungen beziehen, nicht mit dem Datenfeld *Methodology* indiziert und würden durch diese Spezifizierung bei der Recherche auch ausgeschlossen werden (hohe *precision*).

Innerhalb der fünf Themenblöcke werden die Suchbegriffe mit dem Boole'schen Operator *OR* verbunden. Dieser Operator gewährleistet, dass mindestens einer der Suchbegriffe oder die mit *Field Codes* versehenen Begriffe (*controlled vocabulary terms*) in den o.g. Datenfeldern gefunden werden.

Mit Hilfe des o.g. Suchstrings wurden zum damaligen Zeitpunkt (31.07.2009) insgesamt 175 „Treffer" erzielt respektive Referenzen gefunden. Nach Sichtung von Titel und Abstract dieser Referenzen wurden insgesamt 13 Reviews im Volltext beschafft (Bloom, 2001; Boath & Henshaw, 2001; Bond, 2006; Bortolotti, Menchetti, Bellini, Montaguti & Berardi, 2008; Gibbons, Crits-Christoph & Hearon, 2008; Cuijpers, van Straten, Andersson & van Oppen, 2008;

Jones, 2004; Karel & Hinrichsen, 2000; Lewis, Dennerstein & Gibbs, 2008; Messer & Kaplan, 2004; Pinquart, Duberstein & Lyness, 2007; Pinquart & Sörensen, 2001; Scogin, Welsh, Hanson, Coates & Stump, 2005). Zudem wurde in der Cochrane Libary der Cochrane Collaboration nach weiteren relevanten Reviews recherchiert. Infolgedessen wurden insgesamt vier Reviews beschafft (Abbass, Hancock, Henderson & Kisely, 2006; Dennis & Hodnett, 2007; Dennis, Ross & Grigoriadis, 2007; Wilson, Mottram & Vassilas, 2008). Die Überblickswerke, die durch psychoanalytische Fachgesellschaften wie die DGPT und die DFT in Auftrag gegeben wurden, wurden ebenfalls im Volltext beschafft (Brandl et al., 2004; Loew, Richter, Calatzis & Krause, 2002; Richter et al., 2002). Zudem stellte Allan Abbass im Rahmen einer E-Mailkorrespondenz sein Überblickswerk *Depression Studies Pertinent to NICE Guidelines: Short-term Psychodynamic Psychotherapies (STPP)* (2008) zur Verfügung.

Im Rahmen der systematischen Durchsicht aller bisher genannten Reviews konnten weitere relevante Überblicksarbeiten gefunden und beschafft werden (Areán & Cook, 2002; Barkham et al., 2008; Churchill et al., 2001; Fonagy, 2000; Fonagy et al., 2002; Fonagy, Roth & Higgitt, 2005; Shadish et al., 2000; Shapiro et al., 2003; Westen & Morrison, 2001).

Die systematische Durchsicht der genannten narrativen und systematischen Überblicksarbeiten erfolgte sowohl im Hinblick auf die Primärstudien, die von den (systematischen) Reviews integriert wurden als auch im Hinblick auf die Literaturverzeichnisse, also auch der Studien, die ggf. aus der metaanalytischen Weiterverarbeitung ausgeschlossen wurden.

Zuletzt wurden noch Internetseiten von Forschungseinrichtungen, universitären Instituten und Universitätskliniken, von denen ein klinisch-psychodynamischer Schwerpunkt bekannt ist, gesichtet (z.B. der Abteilung für Psychosomatische Medizin und Psychotherapie des Universitätsklinikums Ulm; der Abteilung für Psychosomatische Medizin und Psychotherapie der Georg-August-Universität Göttingen; des Fachklinikums Tiefenbrunn (Rosdorf); der Kli-

nik für Psychosomatische Medizin und Psychotherapie des städtischen Klinikums München; der Abteilung für Psychosomatik und Psychotherapie des Universitätsklinikums Standort Gießen; des Sigmund-Freud-Instituts in Frankfurt). Hier wurden Literaturlisten einzelner Mitarbeiter, die in der Forschung tätig sind, sowie online gestellte Tagungsflyer und Auflistungen abgeschlossener und laufender Forschungsprojekte gesichtet.

Ergebnis dieser Handsuche waren 43 in den Reviews als Studien ausgezeichnete Primärstudien, die nach Titel- und (wenn möglich) Abstractsichtung im Volltext beschafft werden sollten[26] (vgl. Abbildung 5). Die Publikationen zu vier Studien konnten weder als Abstract gesichtet noch im Volltext beschafft werden, da sie nach Kontaktierung der Erstautoren noch nicht publiziert waren und in die o.g. Überblickswerke lediglich als Kongressbeitrag o.ä. Eingang fanden (Gerber, Fonagy, Bateman & Higgitt, 2004; Huber & Klug, 2006; Leuzinger-Bohleber & Beutel, 2009; Szecsödy et al., 1999). Von den 39 im Volltext beschafften und gesichteten Studien wurden wiederum fünf ausgeschlossen, da Kombinationswirkungen (Pharmakotherapie und Psychotherapie) untersucht wurden (Burnand et al., 2002; de Jonghe, Kool, van Aalst, Dekker & Peen, 2001; de Jonghe et al., 2004; Dekker et al., 2005; Maina, Rosso, Crespi & Bogetto, 2007). Eine weitere Studie wurde ausgeschlossen, da u.a. Kinderanalysen untersucht wurden (Erle & Goldberg, 2003). Zudem wurde in einer weiteren Studie ein psychodynamisches Beratungskonzept evaluiert und aus diesem Grund ebenfalls exkludiert (Simpson, Corney, Fitzgerald & Beecham, 2003). Insgesamt drei Studien mussten ausgeschlossen werden, da entweder aus den Publikationen klar hervorging oder aber nach Kontaktaufnahme mit den Autoren offenbar wurde, dass zahlreiche untersuchte Patienten sich zum letzten Messzeitpunkt noch in Behandlung befanden (Hartmann, 2006; Hartmann &

[26] In diese 43 Studien gingen noch Untersuchungen im stationären Setting mit ein, die erst zu einem späteren Zeitpunkt exkludiert wurden (Kap. 3.1.1).

Zepf, 2002; Freedman, Hoffenberg, Vorus & Frosch, 1999; Puschner, Kraft, Kächele & Kordy, 2007). Eine weitere Studie untersuchte unterschiedliche Psychotherapieformen (kognitive Verhaltenstherapie und psychodynamisch-interpersonelle Therapie) und berichtete keine segregierten Ergebnisse pro Verfahren (Barkham, Rees, Stiles, Hardy & Shapiro, 2002). Die Studie von Berghout und Zevalkink (2009) wurde ausgeschlossen, da es sich dabei um eine Trendstudie handelt, in der Outcomedaten von Patienten in unterschiedlichen Phasen ihrer Therapie via Querschnittuntersuchung erhoben wurden.

Nach Ausschluss aller genannten Studien blieben 27 Studien übrig. Diese 27 Studien wurden wiederum um eine Studie, auf die in einer anderen inkludierten Studie Bezug genommen wurde, erweitert (von Wietersheim, Wilke, Röser & Meder 2002, 2003), so dass das finale Ergebnis der Handsuche sich auf 28 Studien belief. Davon beziehen sich vier Studien auf Untersuchungsgruppen aus dem Kreis der affektiven Störungen und 24 Studien auf heterogene Untersuchungsgruppen.

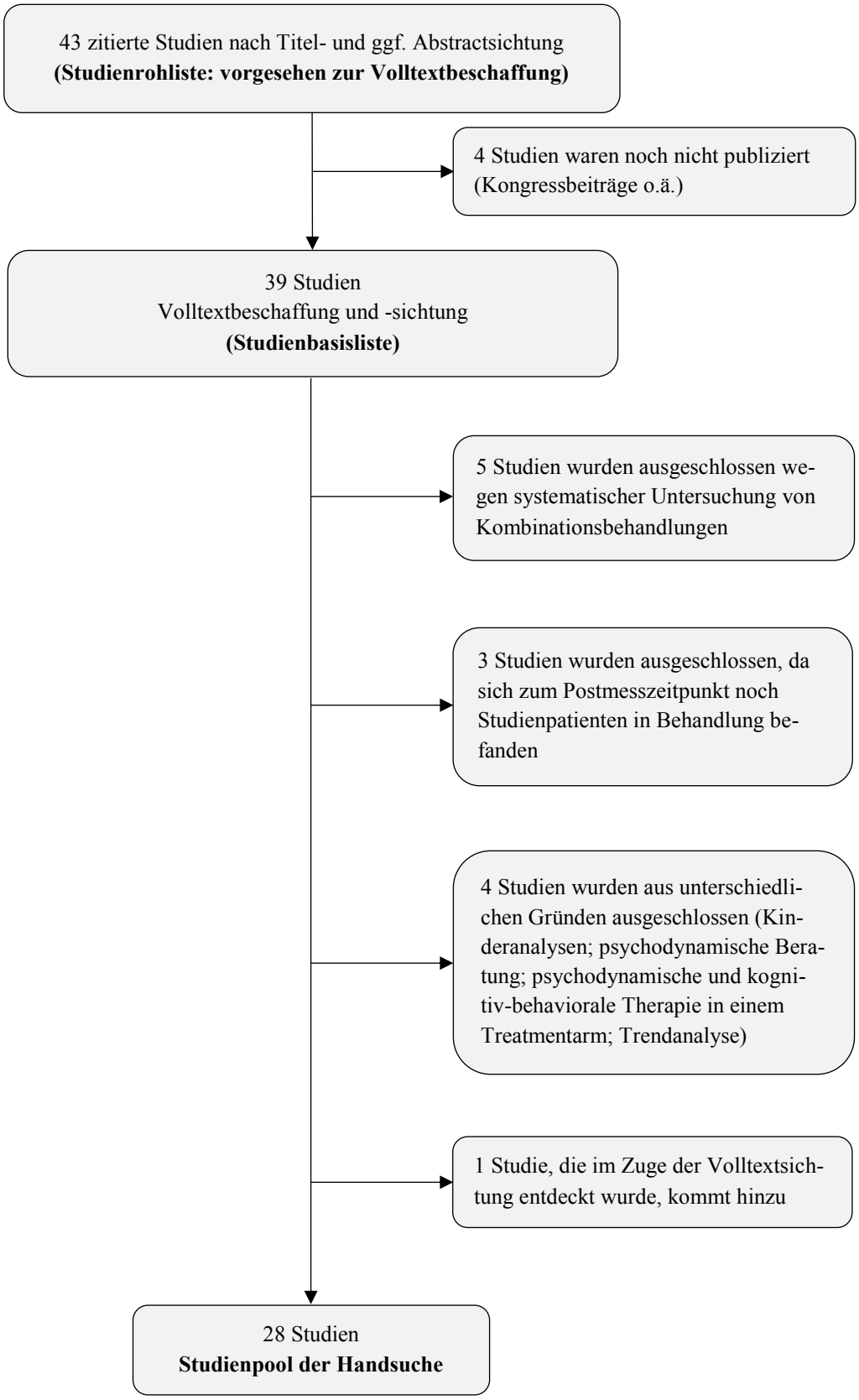


Abbildung 5: Flowdiagramm der Primärstudienselektion (Handsuche)

Im Anschluss an die beschriebene Handsuche wurde eine digitale Recherche in den Online-Datenbanken PsycINFO (EBSCO), PubMed (National Center for Biotechnology Information [NCBI]) und PSYNDEX (EBSCO) durchgeführt. Die Suche wurde zusätzlich auf die Online-Datenbank der Cochrane Library ausgeweitet.

Dafür war es zunächst notwendig, für die drei zuerst genannten Datenbanken Suchstrings zu bilden, durch die die Primärstudiensuche so sensitiv wie nötig (hoher *recall*) und so spezifisch wie möglich (hohe *precision*) gestaltet werden konnte (vgl. White, 2009). Ein ausreichender *recall* wurde über sog. Schlüsselpublikationen definiert, die durch die Suchstrings auf jeden Fall gefunden werden sollten. Die Präzision wurde durch Nutzung bereits eingeführter *Field Codes* und der Tatsache sichergestellt, dass auch hier explizite Themenblöcke gebildet und durch den Boole'schen Operator *AND* verbunden wurden.

Diese Suchtechnik erzwingt zum einen, dass sich bei der Freitextsuche mindestens ein Begriff pro Themenblock in den datenbankspezifischen Datenfeldern der indizierten Publikationen wiederfindet (hohe *precision*). Zum anderen wird durch die Verbindung von Begriffen durch den Boole'schen Operator *OR* innerhalb eines jeden Themenblocks die Chance, auf relevante Publikationen zu stoßen, erhöht (hoher *recall*).

Da sich die Online-Datenbanken PsycINFO, PSYNDEX und PubMed vor allem hinsichtlich der Indizierung durch die *controlled vocabulary terms* sowie der Literatursprachen (englisch/deutsch), voneinander unterscheiden, musste für jede einzelne Datenbank ein individueller Suchstring gebildet werden. Die Datenbank PSYNDEX wurde hauptsächlich für das Auffinden deutschsprachiger Literatur verwendet, das bedeutet, dass auch die Validierung des Suchstrings nur mittels deutschsprachiger Literatur erfolgte (s.u.). Die Suche in den Datenbanken PsycINFO und Pubmed fokussierte auf die englischsprachige Literatur. Dementspre-

chend enthalten die Suchstrings auch ausschließlich englische Suchbegriffe. Folgende Suchstrings wurden für die Suche in PsycINFO (Abbildung 6) und PubMed (Abbildung 7) erstellt:

PsycINFO:

((efficacy) OR (effectiveness) OR (psychiatric evaluation*) OR (therapeutic effect*) OR (psychotherap* evaluation*) OR (treatment evaluation*) OR (mental health program evaluation*) OR (practice based) OR (recovery) OR (symptomatic change) OR (symptom improvement) OR (outcome stud*) OR (trial therap*) OR (naturalistic stud*) OR (clinical trial*) OR (random* control*) OR (assessment*) OR (empirically supported) OR (empirically validated) OR (quasi experiment*) OR (between group design*) OR (follow up) OR (followup) OR (evidence) OR (psychotherap* outcome*) OR (*therapeutic outcome*) OR (treatment outcome*) OR (treatment result*) OR (experiment* control*) OR (clinical sampl*) OR (random sampl*) OR (outpatient*) OR (ambulatory) OR (inpatient*)) AND ((treatment*) OR (therap*) OR (psychotherap*) OR (intervention*)) AND ((MR empirical study) OR (MR field study) OR (MR longitudinal study) OR (MR quantitative study) OR (MR observational study) OR (MR treatment outcome/clinical trial) OR (MR prospective study) OR (MR retrospective study)) AND ((psychoanaly*) OR (short term *dynamic) OR (long term dynamic) OR (psychodynamic*) OR (brief dynamic) OR (*analytical) OR (depth psycholog*) OR (insight oriented psychotherap*) OR (relationship oriented therap*) OR ((interpretive psychotherap*) AND (supportive psychotherap*)) OR (nondirective psychotherap*)) AND ((affective disorder*) OR (mood disorder*) OR (depress*) OR (emotional disorder*) OR (dysthym*) OR (bipolar) OR (diagnost* heterogen*) OR (heterogen* disorder*) OR (mental disorder*) OR (psychiatric disorder*) OR (neurosis) OR (suicid*) OR (comorbid*) OR (mixed sample*) OR (naturalistic) OR (long* term) OR (catamnestic*) OR (practice based) OR (clinical effectiveness) OR (clinical evaluation*) OR (treatment effectiveness evaluation*)) AND (DT 1999-2009) AND ((AG young adulthood) OR (AG adulthood) OR (AG thirties) OR (AG middle age) OR (AG aged) OR (AG very old)) AND ((LA german) OR (LA english))

25.02.2010 (Treffer: 421)

Abbildung 6: Suchstring für Primärstudienrecherche in PsycINFO

PubMed:

("1999"[PDAT] : "2009"[PDAT]) AND (("efficacy"[TW] OR (("effect"[TW] OR "effects"[TW]) AND ("psychodynamic psychotherapy"[TW] OR "psychoanalytic psychotherapy"[TW])) OR "treatment outcome" [TW] OR "effectiveness" [TW] OR "Outcome and Process Assessment (Health Care)"[Mesh] OR "evidence-based practice" [TW] OR "Clinical Trials as Topic"[Mesh Terms] OR "psychological intervention" [TW] OR "cohort studies"[Mesh Terms] OR "prospective studies" [TW] OR "retrospective studies" [TW] OR "comparative study" [TW] OR "comparative study"[ptyp] OR "Clinical Trial"[ptyp]) OR "Evaluation Studies"[ptyp]) AND ("psychodynamic" [TW] OR "psychoanalysis" [TW] OR "psychoanalytic therapy"[Mesh:noexp] OR "insight oriented psychotherapy"[TW] OR "depth psychology" [TW] OR "short term dynamic" [TW] OR ("long term dynamic psychotherapy" [TW] AND "long-term care"[Mesh Terms]) OR "brief dynamic *therapy"[TW] OR (("psychotherapy, brief"[Mesh Terms]) AND ("psychodynamic" [TW] OR "psychoanalytic" [TW])) OR (("group therapy"[TW]) AND ("psychodynamic"[TW] OR "psychoanalytic"[TW] OR "analytic"[TW])) OR "nondirective psychotherapy"[TW] OR ("relationship-oriented"[TW] AND "psychotherapy"[TW])) AND ("mood disorders"[Mesh Terms] OR "mental disorders"[Mesh Terms] OR "long-term care/psychology"[Mesh Terms] OR "naturalistic study"[TW] OR "comorbidity" [TW] OR "suicide"[Mesh:noexp]) AND ("german" [la] OR "english" [la]) AND ("adult"[MeSH Terms]) NOT (“Case Reports”[ptyp] OR “Review”[ptyp] OR “Meta-Analysis”[ptyp])

25.02.2010 (Treffer: 268)

Abbildung 7: Suchstring für Primärstudienrecherche in PubMed

Der Suchstring für die Recherche in PsycINFO setzt sich aus acht Themenblöcken[27] zusammen, die jeweils durch den Boole’schen Operator *AND* miteinandern verbunden wurden: Der erste Block enthält eher allgemeine Begriffe, die charakterisieren, um welche Art von empirischer Studie es sich handeln soll bzw. was in welchem Kontext erhoben werden sollte (*efficacy, effectiveness*, *psychiatric evaluation*, symptomatic change, random sampl** etc.). Der zweite Block stellt sicher, dass es sich um den Untersuchungsgegenstand "Psychotherapie" handeln muss. Der dritte Block ist durch *controlled vocabulary terms* zusammengesetzt und begrenzt die Suche auf solche empirischen Studien, die mit diesen Termen standardisiert indi-

[27] Die Angaben des Publikationsraums als auch der Sprachen werden hier ebenfalls als "Themenblock" verstanden.

ziert sind (*empirical study, longitudinal study, treatment outcome/clinical trial* etc.). Der vierte Block betrifft die nähere Eingrenzung der Intervention (*short term *dynamic, insight oriented psychotherap** etc.) und wurde zum einen mit Hilfe des PsycINFO-eigenen *Thesaurus of Psychological Index Terms* zusammengestellt; zum anderen wurden die mittels Handsuche eruierten Studien auf unterschiedliche Interventionsbegriffe für den Oberbegriff „psychoanalytisch begründete Verfahren" hin abgesucht. Im vierten Themenblock ist außerdem eine *AND*-Verbindung zu finden, die hier als eine Art Bedingung fungiert: *((interpretive psychotherap*) and (supportive psychotherap*))*. Das bedeutet, dass der Fund *interpretive psychotherap** nur dann relevant ist, wenn er mit dem Term *supportive psychotherap** kombiniert auftritt (und umgekehrt). Es hat sich bei der Durchsicht der Studien der o.g. Handsuche gezeigt, dass diese Terme in relevanten Studien zusammen auftreten (z.B. Piper, McCallum, Joyce, Rosie & Ogrodniczuk, 2001).

Der fünfte Themenblock beschreibt die Untersuchungsgruppen näher. Auch hier wurde der Thesaurus eingesetzt (Oberbegriff: „affective disorder"). Vor allem für die Recherche nach den „gemischte Patientengruppen" wurden die via Handsuche eruierten Studien herangezogen und diejenigen Begriffe extrahiert, mit denen dort die Untersuchungsgruppen belegt waren (*mental disorder*, mixed sample*, psychiatric disorder**). Der sechste Block bezieht sich auf das Datenfeld des Publikationszeitraums, der siebente Block auf das Datenfeld, in dem die Altersgruppen standardisiert sind und der letzte Block auf die Publikationssprachen (deutsch oder englisch).

Der o.g. Suchstring verzeichnet das Ergebnis eines iterativen Konstruktionsprozesses, indem der Suchstring so lang optimiert wurde, bis bestimmte, zuvor festgelegte Schlüsselpublikationen gefunden wurden (vgl. Pant, 1998). Diese Optimierung gewährleistete, dass bspw. Begriffe, die thematisch an sich nicht in einen bestimmten Block gehören, dennoch in diesem Block auftraten. So findet man z.B. die Begriffsreihe *(naturalistic) or (long* term) or (catam-*

nestic)* im Themenblock, der die Untersuchungsgruppe eingrenzen soll. Es zeigte sich bei der Durchsicht der Studien aus der Handsuche, die sich auf „gemischte Untersuchungsgruppen“ beziehen, dass die o.g. Terme typische Begriffe sind, mit denen solche Studien in PsycINFO assoziiert sind. Die Trefferquote bzgl. dieser Studien wurde durch Aufnahme dieser Begriffe in eben diesem Themenblock lohnenswert erhöht, da ein großer Teil „gemischter Untersuchungsgruppen“-Studien somit gefunden werden konnte.

Schließlich wurde der Suchstring mittels 15 englischsprachiger Studien aus der Handsuche validiert. Das bedeutet, es wurde überprüft, ob mittels der digitalen Recherche diese Studien tatsächlich gefunden wurden. Dabei zeigte sich, dass insgesamt zwei Studien nicht gefunden wurden: Zum einen, weil eine gesuchte Publikation in der Datenbank (unerklärterweise) mit keinem Datenfeld für die Altersgruppen indiziert war. Zum anderen, weil der Jahrgang der betreffenden Zeitschrift, in dem die Studie erschien, nicht in PsycINFO eingespeist war. Beide Publikationen wurden jedoch im Rahmen der Suche via PubMed gefunden, die nun beschrieben wird.

Ein Spezifikum der Online-Datenbank PubMed sind die sog. *MeSH-Terms* (MeSH: Medical Subject Headings). Bei diesen handelt es sich ebenfalls um *controlled vocabulary terms*, mit denen eine Publikation in PubMed indiziert ist. Im Unterschied zu den bereits für PsycINFO eingeführten *controlled vocabulary terms*, die bestimmten Datenfeldern (*Methodology*, *Subject Headings, Classification Code, Population Group* etc.) zugeordnet sind, handelt es sich bei den MeSH-Termen um Begriffe, die keinem bestimmten Datenfeld zugeordnet sind. MeSH-Terme sind demgegenüber hierarchisch geordnet und derart miteinander verästelt, dass zu einem Unterbegriff mehrere Oberbegriffe gehören können. Ebenfalls differierend zu den *controlled vacabulary terms* bei PsycINFO werden mit MeSH-Termen sowohl Synonyme zu

diesen Termen als auch hierarchisch darunter liegende Terme gesucht (vgl. Läzer, Sonntag, Drazek, Jaeschke & Hogreve, 2010). Die Suche nach hierarchisch unter dem eingegebenen MeSH-Term liegenden Begriffen lässt sich jedoch auch unterbinden (durch den Befehl [Mesh:noexp]), was eine Spezifizierung der Suche zur Folge hat. Belegt man einen Begriff hingegen mit dem Befehl [Mesh Terms], so wird sowohl nach Synonymen als auch polyhierarchisch gesucht, was eine Ausweitung der Suche nach sich zieht (vgl. Läzer et al., 2010).

Der Befehl [TW] veranlasst, dass der damit belegte Suchstringbegriff (z.B. *"psychoanalytic psychotherapy"[TW]*) in Titel, Abstract und in den MeSH-Termen etc. gesucht wird. Das Kürzel [ptyp] bezeichnet den Publikationstyp, was dem Datenfeld *Methodology* in PsycINFO nahe kommt (z.B. *"Clinical Trial"[ptyp]) OR "Evaluation Studies"[ptyp]*). Der Befehl [PDAT] grenzt den Publikationszeitraum ein.

Der PubMed-Suchstring (Abbildung 7) setzt sich aus sechs Themenblöcken zusammen, die mit *AND* bzw. *NOT* verbunden sind. Der Boole'sche Operator *NOT* (vor dem letzten Themenblock) sorgt für den Ausschluss von Publikationen, die mit denjenigen Begriffen, auf die sich dieser Operator bezieht, indiziert sind: *NOT ("Case Reports"[ptyp] OR "Review"[ptyp] OR "Meta-Analysis"[ptyp])*. Diese Einschränkung war notwendig, da durch die Hierarchisierung der MeSH-Terme, die sich auf die Studienart beziehen (etwa *"Clinical Trials as Topic"[Mesh Terms] OR ... OR "cohort studies"[Mesh Terms]*) automatisch damit assoziierte Begriffe gesucht werden, z.B. „*Review*" etc..

Analog zum Suchstring für die Recherche in PsycINFO (Abbildung 6, S. 125) werden auch hier Terme *innerhalb* von Themenblöcken mit *AND* verbunden. Bei der Optimierung des Suchstrings entpuppte sich auch hier die Bildung von Subthemenblöcken und kleineren Einheiten als vorteilhaft.

Die Beschränkung der Altersklassen der in die Studien eingehenden Probanden wurde durch einen zusätzlichen von PubMed bereitgestellten Filter kreiert und gleicht der Altersspanne, die auch in der Recherche in PsycINFO verwendet wurde.

Auch dieser Suchstring wurde anhand von englischsprachigen Schlüsselpublikationen konstruiert und anschließend an 15 Publikationen aus der Handsuche validiert. Dabei wurden vier Studien aus folgenden Gründen nicht gefunden: Eine Studie war ohne Altersklasse indiziert; von einer weiteren Studie war der Jahrgang der Zeitschrift nicht in PubMed eingelesen; eine Studie wurde nicht gefunden, da sie als Kapitel in einem Sammelwerk erschien und das Sammelwerk in PubMed nicht indiziert war; eine letzte Studie blieb aufgrund der Einschränkung *NOT "Case Reports"[Ptyp]* unentdeckt, da sie aus unerklärten Gründen unter diesen Publikationstyp indiziert war. Drei der vier genannten Studien wurden jedoch durch die PsycINFO-Recherche gefunden.

Während PubMed und PsycINFO zum größten Teil englischsprachige Literatur indizieren, lässt sich mittels PSYNDEX sowohl nach englisch- als auch deutschsprachiger Literatur recherchieren. Im Rahmen dieser Arbeit wurde die Suche über PSYNDEX deshalb nur noch zum Auffinden deutschsprachiger Literatur verwendet. Dabei konnten mit Hilfe des folgenden Suchstrings (Abbildung 8) insgesamt 140 „Treffer“ erzielt werden:

PSYNDEX:

((wirksamkeit*) OR (wirkung*) OR (efficacy) OR (effectiveness) OR (therapieerfolgskontrolle) OR (verbesserung*) OR (outcome stud*) OR (trial therap*) OR (naturalistic stud*) OR (naturalistische studie*) OR (klinische prüfung*) OR (randomisiert* kontrolliert*) OR (assessment*) OR (follow up) OR (followup) OR (katamnese*) OR (evidence) OR (evidenz*) OR (psychotherap* outcome*) OR (treatment outcome*) OR (outpatient*) OR (ambulant*) OR (inpatient*) OR (stationär*)) AND ((treatment*) OR (behandlung) OR (therap*) OR (psychotherap*) OR (intervention*)) AND ((MR empirical study) OR (MR illustrative empirical data) OR (MR data reanalysis) OR (MR experimental study) OR (MR longitudinal empirical study) OR (MR multicenter study) OR (MR study project) OR (MR treatment program)) AND ((psychoanalyt*) OR (psychoanalysis) OR (short term dynamic) OR (((kurzpsychotherapie) OR (kurzzeittherapie*) OR (kurzzeitbehandlung*)) AND ((psychodyn*) OR (psychoanalyt*))) OR (long term dynamic) OR (((langzeitpsychotherap*) OR (langzeittherapie) OR (langzeitbehandlung*)) AND ((psychodyn*) OR (psychoanalyt*))) OR (psychodynamic*) OR (psychodynamisch*) OR (brief dynamic) OR (analytical) OR (analytisch*) OR (depth psycholog*) OR (tiefenpsycholog*)) AND ((affective disorder*) OR (affektive Störung*) OR (mood disorder*) OR (depress*) OR (emotional disorder*) OR (dysthym*) OR (bipolar) OR (mental disorder) OR (neurosis) OR (suicid*) OR (comorbid*) OR (mixed sample*) OR (representative sample*) OR (long* term) OR (katamnese*) OR (treatment effectiveness evaluation*)) AND (DT 1999-2009) AND ((LA german)) NOT ((AG childhood) OR (AG infancy) OR (AG preschool age) OR (AG school age) OR (AG adolescence))

25.02.2010 (Treffer: 140)

Abbildung 8: Suchstring für Primärstudienrecherche in PSYNDEX

Zur Vermeidung von Redundanzen, wird die Konstruktion dieses Suchstrings nur noch in aller Kürze beschrieben. Die Verwendung von Boole'schen Operatoren, Trunkierungen sowie *controlled vocabulary terms* lehnt sich an die Recherche in der Datenbank PsycINFO an, da beide Datenbanken vom selben Datenbankanbieter (EBSCO) stammen[28]. Der PSYNDEX-Suchstring setzt sich aus acht Themenblöcken zusammen. Die Besonderheit dieses Suchstrings liegt darin, dass er sich sowohl aus deutsch- als auch englischsprachigen Suchbegriffen zusammensetzt. Der Grund dafür liegt darin, dass die in PSYNDEX festgelegten *controlled*

[28] Zum Jahreswechsel 2009/10 und damit im Zeitraum der hiesigen Studienrecherche fand an der Freien Universität ein Datenbankanbieterwechsel für PSYNDEX von OVID zu EBSCO statt. Auf die Darstellung der darauf folgenden Revision des PSYNDEX-Suchstrings wird an dieser Stelle verzichtet.

vocabulary terms in den unterschiedlichen Datenfeldern sowohl in deutscher als auch englischer Sprache indiziert sind. Zudem werden Titel und Abstracts in deutschsprachigen Publikationen oftmals auch in Englisch angegeben. Die Entscheidung, dementsprechend auch englischsprachige Terme mit in den Suchstring aufzunehmen, diente allein dem Ziel, einen möglichst hohen *recall* zu erzielen und relevante Referenzen nicht unentdeckt zu lassen.

Auch diese Suche wurde anhand von Publikationen aus der Handsuche validiert, diesmal mit ausschließlich deutschsprachigen Publikationen. Dabei zeigte sich, dass von den 10 ausgewählten Publikationen insgesamt eine Studie nicht gefunden wurde. Auch hier war der gesuchte Jahrgang der Zeitschrift, in dem die Studie publiziert wurde, nicht in PSYNDEX indiziert. Da diese Studie sich auf mehrere Publikationen belief, die teils auf Englisch und teils auf Deutsch erschienen, wurde eine der englischsprachigen Publikationen jedoch im Rahmen der PubMed-Recherche gefunden.

Zusammengenommen führten die Recherchen in allen drei Online-Datenbanken zu 829 Referenzen (inklusive Dubletten), die gesammelt in das Literaturverwaltungsprogramm CITAVI (Version 2.0) importiert wurden.

Die digitale Recherche wurde an diesem Punkt durch eine Studiensuche in der Cochrane Library ergänzt. Hier wurden die Terme *(psychodyn* and depress*)* sowie der Filter zur Begrenzung des Publikationszeitraums (1999-2009) verwendet und damit ausschließlich nach Studien gesucht, in die diagnosehomogene Untersuchungsgruppen aus dem affektiven Formenkreis eingingen. Dies machte insoweit Sinn, als dass die Cochrane Collaboration der Methodologie der EbM verpflichtet ist, in der Untersuchungen von störungsspezifischen Inter-

ventionen dominieren. Mit den o.g. Termen wurde eine Recherche in Titeln, Abstracts und *Keywords* durchgeführt und zudem lediglich in der Rubrik *Clinical Trials* (heute: *Trials*) gesucht. Damit wurden 62 „Treffer" erzielt (14.03.2010), die ebenfalls in das Literaturverwaltungsprogramm CITAVI importiert wurden.

Abbildung 9 illustriert den weiteren Selektionsprozess: In die insgesamt 891 Referenzen[29] aus der digitalen Recherche (PsycINFO, PubMed, PSYNDEX und Cochrane Library) gingen insgesamt 124 Publikationen mehrfach ein, d.h. 124 Publikationen wurden in mindestens zwei Datenbanken gefunden. Aus diesem Grund konnten 135 Publikationen gelöscht werden. In einem weiteren Schritt wurden zunächst die Dissertationen extrahiert und gesichtet: Von insgesamt 60 Dissertationen wurden nach Titel- und Abstractsichtung vier Arbeiten potentiell als zur Zielpopulation gehörend indentifiziert. Drei dieser Arbeiten wurden zu einem späteren Zeitpunkt in Zeitschriften publiziert und in den Studienpool aufgenommen. Nach Abzug der 60 Dissertationen blieben 696 Publikationen übrig, von denen wiederum 35 Publikationen gelöscht werden konnten, die bereits durch die Handsuche gefunden wurden. Die übrigen 661 Publikationen wurden im Anschluss via CITAVI samt Titeln, Abstracts und Schlagwörtern ausgedruckt. Dieser Ausdruck diente als Grundlage für die anstehende Titel- und Abstractsichtung, durch die insgesamt 72 Publikationen als lohnenswert befunden wurden, im Volltext gesichtet zu werden.

Die Volltextbeschaffung erfolgte in zahlreichen Fällen über die o.g. Datenbanken, in denen Volltexte aus Zeitschriften in elektronischer Form einfach zugänglich sind. Buchkapitel oder Artikel aus Zeitschriften, zu denen kein Volltextzugang gewährleistet ist, wurden über die Fernleihe der Universitätsbibliothek (Freie Universität Berlin) beschafft.

[29] In diese 891 Studien gingen noch Untersuchungen im stationären Setting mit ein, die erst zu einem späteren Zeitpunkt exkludiert wurden.

Da sich in der Regel meist mehrere Publikationen auf eine Primärstudie beziehen, wurden die 72 Volltexte zunächst zu insgesamt 57 Studien zusammengestellt. Dabei war es zum Teil notwendig, die (Erst-) Autoren der Publikationen zu kontaktieren, da in einige Publikationen derselben Autoren und desselben Untersuchungsgegenstands oftmals ungleich große Stichproben eingingen. Damit wurde das doppelte Eingehen einer Untersuchung in den hiesigen Studienpool vermieden.

Weitere 13 von den 57 im Volltext gesichteten Studien wurden aus folgenden Gründen ausgeschlossen: In einer Studie wird ein psychodynamisches Beratungskonzept evaluiert (Archer, Forbes, Metcalfe & Winter, 2000). In insgesamt fünf Studien werden primär Prozess-Outcomezusammenhänge untersucht (Crowe & Grenyer, 2008; Götze et al., 2003; Lindgren, Barber & Sandahl, 2008; Rozmarin et al., 2008; Svartberg, Seltzer, Choi & Stiles, 2001). Eine Studie untersucht die Wirksamkeit von kognitiv-behavioraler Gruppentherapie im Vergleich zu kognitiv-behavioraler und psychodynamisch-interpersoneller Individualtherapie, wobei lediglich zur kognitiv-behavioraler Gruppentherapie ausführliche Ergebnisse berichtet werden (Kellett, Clarke & Matthews, 2007). In einer Studie werden lediglich Ergebnisse zu den ersten Verlaufsmessungen während der Therapie berichtet (Vlastelica, Jurcević & Zemunik, 2005) und in einer weiteren Studie hatten zum letzten Messzeitpunkt knapp 19% der Probanden ihre Behandlung noch nicht beendet (Lorentzen, Bøgwald & Høglend, 2002). In einer Studie entpuppte sich die im Titel erwähnte *psychodynamic group psychotherapy* als eher eklektische Intervention, bestehend aus: „(a) offering psychoeducation to participants about bipolar disorder, (b) discussing illness management issues in bipolar disorder, and (c) focusing on the quality of relationships through group discussion, feedback, active leader participation and discussion of psychodynamic and interpersonal issues" (Gonzalez & Prihoda, 2007, S. 408).

In zwei weiteren Studien wurden psychodynamische Behandlungen in Kombination mit Pharmakotherapie untersucht (Maina, Rosso & Bogetto, 2009; Van, Schoevers et al., 2008). Eine Studie beschreibt lediglich das Vorgehen einer Untersuchung, enthält jedoch keinen Ergebnisbericht (Driessen et al., 2007). Eine letzte Studie untersucht psychodynamische Kunsttherapie und wurde aus diesem Grunde ausgeschlossen (Thyme et al., 2007).

Somit blieben 43 Studien aus der digitalen Studienrecherche übrig, die als zur Zielpopulation gehörig betrachtet wurden.

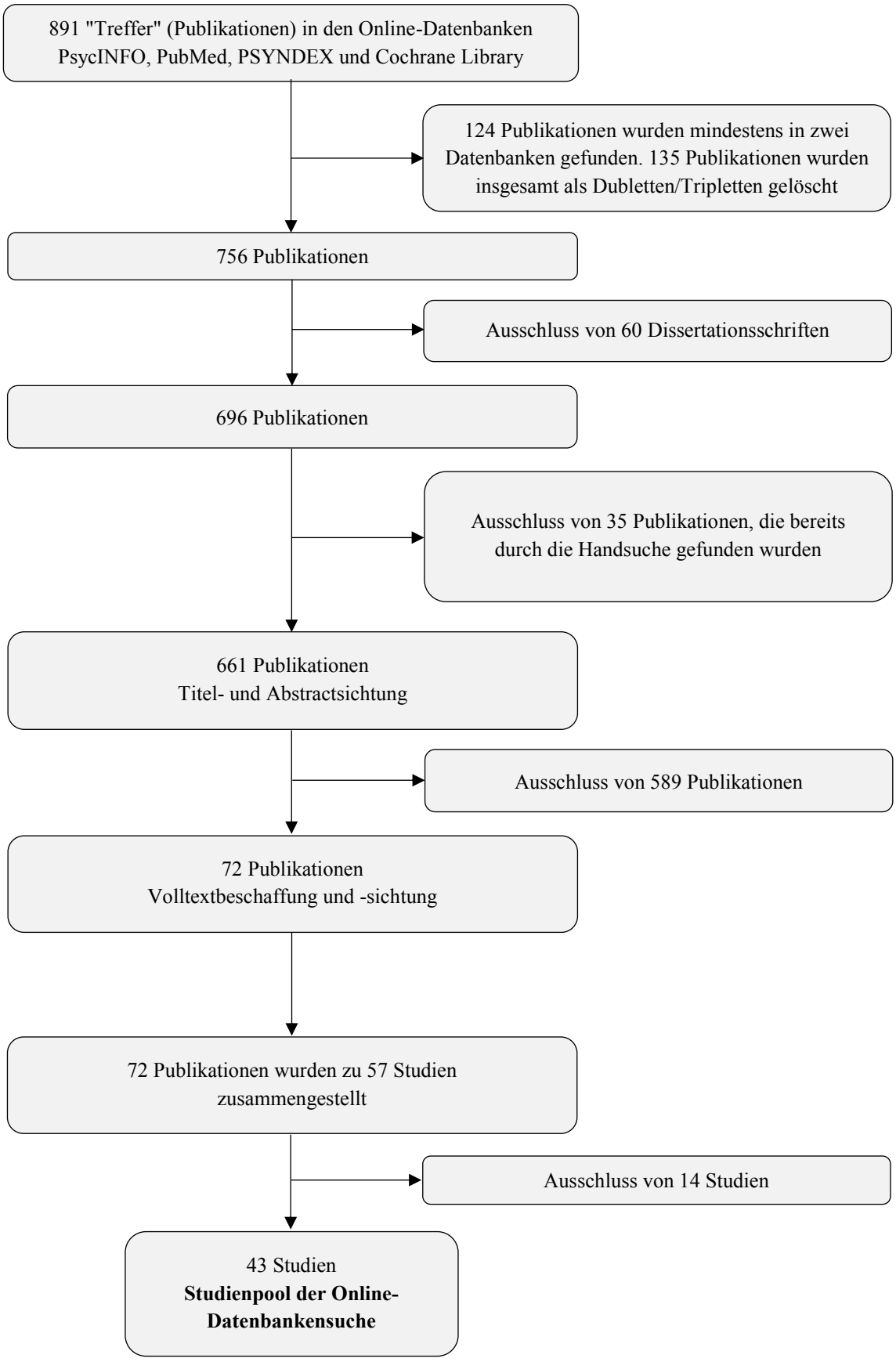


Abbildung 9: Flowdiagramm der Primärstudienselektion (Online-Datenbankensuche)

Es lagen damit insgesamt 71 Studien vor, die mittels Handsuche (28 Studien; Abbildung 5, S. 123) sowie Online-Recherche (43 Studien; Abbildung 9) extrahiert werden konnten. Im Zuge der Volltextsichtung dieser Studien stellte sich heraus, dass es sich bei den Untersuchungen, die im stationären Setting stattfanden, weniger um Evaluationen psychodynamischer Individual- oder Gruppentherapie handelte, sondern – so legt es der Gegenstand "stationäre Therapie" nahe – vielmehr um Evaluationen psychodynamischer *Klinikkonzepte.* Aus diesem Grund wurde entschieden, die Ein- bzw. Ausschlusskriterien an dieser Stelle nachzujustieren. Die Zielpopulation wurde damit modifiziert und bezieht im Weiteren nur noch Wirksamkeitsstudien nach der o.g. Definition (vgl. Kap. 3.1.1) mit ein, die in *ambulanten* Settings durchgeführt wurden. Es zeigte sich an dieser Stelle, dass sich die herangezogenen Reviews (Handsuche; Kap. 3.1.2) vorrangig auf Wirksamkeitsstudien im ambulanten Bereich beschränkten und die meisten der infolge der Modifikation der Zielpopulation auszuschließenden Studien aus der digitalen Recherche stammten. Insgesamt wurden 30 Studien ausgeschlossen: Eine Untersuchung an Patienten mit einer Störung aus dem affektiven Formenkreis und 29 Studien, die an diagnoseheterogenen Störungsgruppen durchgeführt wurden.

Damit besteht die Studienstichprobe für die Kodierung mit Hilfe des WBP-Kriterienkatalogs aus 41 Studien – 10 Studien zu affektiven Störungen und 31 Studien zu gemischten Patientengruppen (vgl. Kap. 4.1). Diese Studien mussten zum Zwecke der Kodierung, was die Anzahl an Publikationen betraf, noch aufgestockt werden: So kommt es nicht selten vor, dass Studien in ihrer Studiendesign-, Patientengruppen- oder Interventionsbeschreibung auf zuvor veröffentlichte Literatur verweisen. Diese Literatur musste, um alle notwendigen Informationen, die zur Kodierung notwendig waren, berücksichtigen zu können, nachbestellt werden. Vor allem bei reanalytischen Studien musste oftmals die Originalliteratur nachbestellt werden (z.B. die Originalliteratur der Menninger Studie *Forty-Two Lives in Treatment: A Study of*

Psychoanalysis and Psychotherapy von Wallerstein, [1986] für die Reanalyse von Blatt & Shahar [2004]). Die 41 Primärstudien belaufen sich damit auf insgesamt 77 Publikationen.

3.1.3 Evaluation der Studienrecherche

Wie in Kapitel 3.1 ausgeführt, sollte die Repräsentativität des hiesigen Studienpools für die intendierte Zielpopulation (vgl. Kap. 3.1.1) mit Hilfe von Strategien der erschöpfenden Studienrecherche sowie expliziter Einschlusskriterien garantiert (vgl. Lipsey & Wilson, 2001) und eine Vollerhebung vorgenommen werden. Das bereits beschriebene Recherchevorgehen mittels umfassender Handsuche sowie digitaler Recherche in einschlägigen Online-Literaturdatenbanken illustriert die erschöpfende Studienrecherche – jedoch lediglich zu einem bestimmten Zeitpunkt (hier: 25.02.2010 [Datum der finalen Recherche in den Online-Datenbanken PsycINFO, PubMed und PSYNDEX] und 14.03.2010 [Datum der finalen Recherche in der Cochrane Library]). In Kapitel 3.1.2 wurde bereits darauf hingewiesen, dass auch Online-Literaturdatenbanken, wie Pubmed oder PsycINFO, trotz täglicher oder wöchentlicher Updates (vgl. Reed & Baxter, 2009), nicht frei von „Lücken" sind. Vielmehr ist im Rahmen einer Recherche davon auszugehen, dass Online-Datenbanken zum Zeitpunkt X nicht die gesamte *publizierte* (hier: deutsch- und englischsprachige) Literatur enthalten, die zu einer definierten Zielpopulation als zugehörig betrachtet werden kann. Zwar erhöht sich die Wahrscheinlichkeit, wie oben gezeigt, mit zunehmender Anzahl an verwendeten Literaturdatenbanken, die „Lücken" zu schließen, gleichzeitig bleibt die Anzahl an relevanten Publikationen, die eine gemeinsame „Lücke" aller genutzten Datenbanken bilden, unbestimmt. Um sich ein Bild darüber zu verschaffen, ob und inwieweit Literatur, die zu den o.g. Zeitpunkten nicht gefunden wurde, möglicherweise im Laufe der Folgejahre nachträglich indiziert wurde, wurde am 18.10.2013 in allen vier Datenbanken (PsycINFO, PubMed, PSYNDEX, Cochrane Library) eine wiederholte Recherche mit den damaligen Suchalgorithmen durchgeführt. Die

Wiederholung erbrachte folgende Ergebnisse: In der Datenbank PsycINFO wurden nunmehr 536 „Treffer“ erzielt (zum damaligen Zeitpunkt waren es 421 „Treffer“), PubMed erbrachte 276 „Treffer“ (*versus* 268 „Treffer“ zum damaligen Zeitpunkt). PSYNDEX erbrachte 146 „Treffer“ (*versus* 140 „Treffer“) und Cochrane Library erbrachte 63 „Treffer“ (*versus* 62 „Treffer“). Ausschließlich in PubMed entstammten die neu hinzugekommenen „Treffer“ allesamt aus dem Publikationsjahr 2009, die in der Cochrane Library zusätzlich gefundene Publikation wurde ebenfalls 2009 veröffentlicht.

Daraufhin wurde untersucht, ob die neu hinzugekommenen „Treffer“ für die hiesige Zielpopulation[30] relevante Literatur enthielten. Die wiederholte Recherche in PsycINFO erbrachte zwei Publikationen, die für die Zielpopulation relevant gewesen wären (Blay et al., 2002; Lau & Kristensen, 2007)[31]. Die erste Publikation erschien in *The British Journal of Psychiatry*, die zweite in *Acta Psychiatrica Scandinavica*. Beide Zeitschriften wurden grundsätzlich im Rahmen der ersten Online-Recherche im Jahr 2010 gefunden, jedoch handelt es sich bei den beiden genannten Publikationen um Veröffentlichungen, die in den Jahren 2012 und 2013 in jeweils korrigierten Fassungen erneut von PsycINFO indiziert wurden. Dies könnte erklären, warum beide Publikationen zum Zeitpunkt des ersten Recherchedurchlaufs nicht indiziert waren und gefunden werden konnten. Es wurden in PsycINFO außerdem 19 weitere Dissertationsschriften gefunden (2010: N=49; 2013: N=68). Obwohl Dissertationsschriften ausgeschlossen wurden (vgl. Kap. 3.1), wurden die 19 neu gefundenen Arbeiten auf potentielle Relevanz überprüft. Im Rahmen der Titel- und Abstractsichtung zeigte sich, dass keine der Arbeiten von Relevanz für den in dieser Arbeit zu untersuchenden Gegenstand gewesen wäre.

[30] Exklusive des stationären Settings.

[31] Da alle damals gefundenen Publikationen gespeichert wurden, konnte leicht nachvollzogen werden, dass die beiden genannten Publikationen beim ersten Suchdurchlauf (2010) nicht durch den Suchstring gefunden wurden.

Weder die wiederholte Recherche in PSYNDEX und PubMed noch in der Cochrane Library erbrachte neue relevante Literatur. Auch hier wurden die in PSYNDEX neu hinzugekommenen vier Dissertationsschriften (2010: *N*=11; 2013: *N*=15) nochmals gesondert betrachtet und nach Titel- Abstractsichtung als irrelevant eingestuft.

Hinsichtlich der Bewertung der Repräsentativität des Primärstudienpools, kann festgehalten werden: Durch genaue Explikation und Anwendung der Studien-Einschlusskriterien wurde gewährleistet, dass jede eingeschlossene Studie in zentralen Merkmalen *proximale Ähnlichkeit* zu Prototypen der intendierten Zielpopulation aufweist (vgl. Rustenbach, 2003). Begreift man Repräsentativität hier als "Ähnlichkeit" zur intendierten Zielpopulation (vgl. Pant, 1998), so ist diese durch das Kriterium der *proximalen Ähnlichkeit* hinreichend erfüllt: Alle Studien ähneln in Bezug auf das Untersuchungsziel (Evaluation der Wirksamkeit psychodynamischer Psychotherapie), die Untersuchungsgruppen (erwachsene Probanden mit Störungen aus dem affektiven Formenkreis oder diagnoseheterogene Patientengruppen) und den Veröffentlichungszeitraum (1999-2009) der intendierten Zielpopulation.

Im Hinblick auf das Erreichen einer Vollerhebung und für die geplante Datenanalyse (vgl. Kap. 3.4) kann konstatiert werden: Der zusammengestellte Studienpool wird als Vollerhebung betrachtet, dabei stellt der Studienpool den Korpus aller publizierter Primärstudien dar, deren Auffinden zum Zeitpunkt des ersten Recherchedurchgangs (2010) technisch möglich war. Diese Festlegung ist zentral für die Auswahl der verwendeten Auswertungsmethoden, auf die in Kapitel 3.4 detailliert eingegangen wird.

3.2 Vorbereitende Maßnahmen für die Anwendung des Kriterienkatalogs

Als erste vorbereitende Maßnahme wurde ein strukturiertes Expertengespräch mit dem wissenschaftlichen Referenten der Bundespsychotherapeutenkammer, Dipl.-Psych. Timo Harfst, geführt, der u.a. an der Entwicklung des Methodenpapiers beteiligt war[32]. Hierbei wurden Fragen eines von der Verfasserin dieser Arbeit erstellten Fragenkatalogs diskutiert. So konnten zum einen allgemeine Fragen zur bisherigen Anerkennungspraxis – vor allem bezogen auf die psychodynamischen Verfahren – und zu den Vorläuferpapieren des Methodenpapiers (vgl. WBP 1999, 2002b, 2004b) sowie zur Kooperation des WBP mit dem Gemeinsamen Bundesausschuss geklärt werden. Zum anderen konnten spezifischere Fragen zum Methodenpapier (Version 2.6; WBP, 2007) geklärt werden, etwa zur Berücksichtigung von Prozessforschungsbefunden in der wissenschaftlichen Anerkennung eines Verfahrens sowie Fragen zu einzelnen Kriterien des Kriterienkatalogs. Anschließend stand Herr Harfst für weitere Anfragen zur Auslegung einzelner Kriterien durchweg zur Verfügung. Auf seine Erklärungen zu einzelnen Kriterien soll ggf. näher eingegangen werden, wenn in den folgenden Kapiteln zentrale Kriterien dargelegt werden.

Vor einem ersten Probedurchlauf, in dem die WBP-Kriterien erstmals im Form eines unabhängigen Ratings durch eine Projektmitarbeiterin (Luisa von Hauenschild[33]) sowie der Verfasserin dieser Arbeit auf einige „Probestudien“ angewendet wurden, wurden die einzelnen

[32] Dieses Gespräch fand bereits im Mai 2008 statt, soll jedoch, da es u.a. als vorbereitende Maßnahme für die Kodierung betrachtet werden kann, an dieser Stelle angeführt werden.

[33] Luisa von Hauenschild absolvierte im Jahr 2010 ein rund einjähriges Forschungspraktikum im Rahmen des in dieser Arbeit vorgestellten Dissertationsprojekts und wurde anschließend für mehrere Monate als drittmittelfinanzierte Mitarbeiterin beschäftigt. Im Jahr 2011 beschloss sie ihre Tätigkeit mit ihrer Diplomarbeit (von Hauenschild, 2011), die inhaltlich an das hiesige Dissertationsprojekt angelehnt ist. Die beschriebenen Tätigkeiten rund um die vorbereitenden Maßnahmen für die Kodierung sowie die Studienkodierung selbst fanden daher unter der Mitarbeit von Luisa von Hauenschild statt.

Kriterien in eine logisch konsistente und für das Rating handhabbare Reihenfolge gebracht. Damit wurde die Abfolge der Kriterien entsprechend ihrer Dimensionszugehörigkeit, wie der WBP sie im Anhang des Methodenpapiers vorsieht, aufgebrochen und nach Themenblöcken neu geordnet. Die Themenblöcke, denen die 44 Kriterien zugeordnet wurden, sind folgende:

- Studiendesign
- Patienten
- Eingangsdiagnostik
- Intervention
- Outcomes/Outcomediagnostik
- Dropoutanalysen
- Statistische Methodik
- Manipulation der Daten.

Anhang B ist der Kodierbogen samt vorgeschaltetem Kurzkodierbogen (vgl. Kap. 3.2.4) in der Fassung zu entnehmen, in der er in der Bewertung der Studien eingesetzt wurde.

Nach dem ersten Probedurchlauf, wurde offenbar, dass die Kriterien trotz Beschreibungen (Ankerbeispiele) der einzelnen Stufen teilweise nicht eindeutig genug zu bewerten waren, um zu übereinstimmenden Ratings zu gelangen. Dabei stellten sich Probleme auf dreierlei Ebenen:

- Zum einen zeigte sich, dass zahlreiche Kriterien nicht durchweg für alle Untersuchungsdesigns oder Studienspezifika kodierbar waren und gewissermaßen zusätzliche Kodiermöglichkeiten geschaffen werden mussten (vgl. Kap. 3.2.1).
- Zum anderen ließen einige Kriterien in ihren Definitionen der einzelnen Ratingstufen einen zu großen Interpretationsspielraum zu, der konkretisiert werden musste (vgl. Kap. 3.2.2).

- Zuletzt mussten einige Kriterienstufen ganz grundlegend spezifiziert werden, um diese Kriterien überhaupt anwenden zu können (vgl. Kap. 3.2.3).

Um diese Probleme zu lösen und zu möglichst übereinstimmenden respektive reliablen Ratings zu gelangen, wurde ein sog. Kodierregelkatalog entwickelt (Ratzek & von Hauenschild, 2001a; Anhang C). Die wichtigsten Kodierregeln werden in den folgenden drei Unterkapiteln beschrieben.

3.2.1 Kodierregeln I: Zusatzkodierregeln

In Anlehnung an die (geplante) Vorgehensweise des WBP (T. Harfst, persönl. Mitteilung, 26.08.2008)[34] wurden alle zu kodierenden Studien *vollständig* mittels des Kriterienkatalogs bewertet. Dies wurde auch dann getan, wenn es offenkundig war, dass eine Studie bspw. durch ein K.O.-Kriterium der allgemeinen methodischen Qualitätsdimension durchfiel. Auf diese Weise konnte außerdem sichergestellt werden, dass möglichst vollständige Informationen über alle Studien gewonnen werden.

Unter *Zusatzkodierregeln* werden solche Kodierungen verstanden, die sich allein auf die Vergabe von Missingwerten beziehen. So wurden für diejenigen Kriterien, für deren Kodierung bestimmte Designvoraussetzungen erfüllt sein müssen, Zusatzkodierungen derart formuliert, dass auch solche Studien bewertet werden können, die die Designvoraussetzungen eben nicht erfüllen. Zudem wurden für möglichst zahlreiche Studienspezifika unterschiedliche

[34] Da bis zum Kodierungsbeginn der Studien im Rahmen des Dissertationsprojekts (Mitte 2010) noch kein/e psychotherapeutische/s Verfahren/Methode mittels des Kriterienkatalogs durch den WBP bewertet wurde, sind die von Herrn Harfst genannten Vorgehensweisen des WBP in der Kodierung einzelner Kriterien als fachliche Einschätzungen zu betrachten, wie der WBP voraussichtlich vorgehen wird bzw. wie der WBP bislang plant, vorzugehen.

Missingwerte vergeben, um im Nachhinein die genauen Gründe für einen Missingwert nachvollziehen zu können.

Ein Beispiel:

Ein Kriterium der internen Validitätsdimension (Kriterium B.10.), mit dem u.a. die Anzahl der in einer Untersuchung realisierten Erhebungen bemessen wird, formuliert implizit die Voraussetzung, dass jede Studie (mindestens) über eine reguläre Postmessung (Messung zum Therapieende) verfügt (vgl. Abbildung 10). Studien, die über keine explizite Postmessung verfügen, sondern von Prä- und Katamnesemessungen inklusive mehrerer Messungen über den Therapieverlauf berichten, sind strenggenommen nicht eindeutig zu kodieren. Da solche Studien jedoch mit der Kodierung „3“ zu schlecht bewertet würden, wurden dementsprechende Zusatzkodierungen (Missingwerte) vergeben (siehe linkes Textfeld in Abbildung 10).

B.10.	Definition der Messzeitpunkte (Prospektive Messung; Follow-up-Messung) • Prä-Katamnesemessung inkl. mehrerer Messzeitpunkte: „4“ • Prä-Katamnesemessung: „5“ • Nur Katamnesemessung: „6“	mehrere vorab festgelegte Messzeitpunkte über den Therapieverlauf incl. Prä-Post-Messungen	(1)	Mehrere prospektive Messungen inkl. Prä-Post-Messung z.B. SCL-90-R, GBB, BL etc.
		ausschließlich Prä-Post-Messung	(2)	Nur Prä-Post-Messung, überwiegend prospektive Messungen z.B. SCL-90-R, GBB, BL etc.
		ausschließlich Post-Messung	(3)	Nur Post-Messung *oder* überwiegend retrospektive Messungen (z.B. „wie stark waren Ihre Symptome zu Beginn der Therapie? Inwieweit hat sich das Problem seither verändert?“, bzw. GAS oder VEV)

Anmerkung: BL: Beschwerdeliste, GAS: Goal Attainment Scale, GBB: Gießener Beschwerdebogen, SCL-90-R: Symptom Checklist-90 Revised, VEV: Veränderungsfragebogen des Erlebens und Verhaltens.

Abbildung 10: Kriterium B.10.: Definition der Messzeitpunkte (interne Validität)

Somit wurde versucht, möglichst für alle Studienspezifika eine Kodierung zu finden, so dass keine Informationen verloren gehen. Zudem sollte sich der erste Kodierungsdurchgang eng an

die Vorgaben des Kriterienkatalogs halten und nicht eindeutig zu kodierende Studien zunächst auch als solche behandeln bzw. durch Missings kenntlich machen. In einem Rekodierungsschritt werden einige der Missings – soweit plausibel – zu regulären Ratings auf den drei Stufen der Kriterien umgewandelt und in diesem Zuge evtl. vorzunehmende, definitorische Erweiterungen der Kriterienstufen vorgeschlagen (vgl. Kap. 4.3).

In den meisten Fällen betraf die Entwicklung von Zusatzkodierregeln zur Vergabe von Missingwerten solche Kriterien, die für die eindeutige Bewertung einer Studie irgendeine Form von Kontroll- oder Vergleichsgruppen voraussetzen. Studien ohne Kontroll- oder Vergleichsgruppen (Ein-Gruppen-Designs) produzieren somit automatisch einen Missingwert auf diesen Kriterien. Unter Ein-Gruppen-Designs sind zum einen Studien zu verstehen, in denen tatsächlich nur ein Treatmentarm eingeht. Diese werden im Weiteren als "echte Ein-Gruppen-Designs" bezeichnet. Zudem sind Studien, in denen sog. „verfahrensinterne Vergleiche“ angestellt werden, als Ein-Gruppen-Designs anzusehen: Das bedeutet, es werden die Outcomes zweier Untersuchungsgruppen miteinander verglichen, die unterschiedliche Methoden desselben Therapieverfahrens erhalten haben – etwa psychodynamische Kurzzeittherapie, in einem Treatmentarm *mit* und in dem anderen Treatmentarm *ohne* Übertragungsdeutungen (vgl. Høglend et al., 2006, 2008). Diese Designs werden in der Begutachtung mit Hilfe des Kriterienkatalogs ebenfalls als Ein-Gruppen-Designs betrachtet (T. Harfst, persönl. Mitteilung, 26.05.2008) und in der hiesigen Kodierung entsprechend als solche behandelt. Es handelt sich sozusagen um "Ein-Gruppen-Designs im erweiterten Sinne".

Diese Definition von Ein-Gruppen-Designs berücksichtigend mussten vor allem für Kriterien der internen Validitätsdimension sowie für einige wenige Kriterien der methodischen Qualitätsdimension besondere Zusatzkodierregeln formuliert werden, die Anhang C zu entnehmen sind. Exemplarisch sollen hier die Zusatzkodierregeln für drei Kriterien illustriert

werden. Die ersten beiden Kriterien entstammen der internen Validitätsdimension, das dritte der methodischen Qualitätsdimension.

Wie in Kapitel 1.2.2 dargelegt, bemessen die beiden Kriterien B.4. (Abbildung 11) und B.5. (Abbildung 12) zum einen die operationale Definition der Kontrollbedingung und zum anderen die strukturelle Äquivalenz der Kontrollbedingung im Vergleich zu dem zu überprüfenden psychotherapeutischen Verfahren. Kriterium B.4. fordert, dass das, was in der Kontrollbedingung geschieht, prospektiv festgelegt wurde und so ausführlich beschrieben wird, dass eindeutig zu beurteilen ist, was mit dieser Bedingung tatsächlich kontrolliert wird.

B.4.	Operationale Definition der Kontrollbedingungen • Eingruppendesigns (inkl. Designs mit verfahrensinternen Vergleichsgruppen): „9“ • Echte VG: „8“	Prospektive Festlegung und umfassende Beschreibung der Kontrollbedingung	(1)	**Hier sind explizit Kontrollbedingungen gemeint!** Prospektive Festlegung und umfassende Beschreibung
		Ex post facto Beschreibung der Kontrollbedingungen	(2)	Ex post facto Beschreibung (Beschreibung im Rahmen eines retrospektiven Designs)
		keine Beschreibung der Kontrollbedingung	(3)	keine Beschreibung

Anmerkung: VG: Vergleichsgruppe.

Abbildung 11: Kriterium B.4.: Operationale Definition der Kontrollbedingungen (interne Validität)

Die Beurteilung von Kriterium B.4 bildet gewissermaßen die Grundlage für die Bewertung des Kriteriums B.5., das nach der Strukturgleichheit der Kontroll- und der Interventionsbedingung fragt (vgl. Hager, 2000).

B.5.	Strukturelle Äquivalenz bei Kontrollbedingungen • Eingruppendesigns (inkl. Designs mit verfahrensinternen Vergleichsgruppen): „9“ • Echte VG: „8“	hinsichtlich des Umfangs an therapeutischer Zuwendung und der Settingbedingungen in der KG besteht Äquivalenz	(1)	**Hier sind explizit Kontrollbedingungen gemeint!**
		der Umfang der therapeutischen Zuwendung in der KG ist reduziert, die Settingbedingungen weichen von der IG ab	(2)	
		der Umfang der therapeutischen Zuwendung in der KG ist deutlich reduziert, die Settingbedingungen weichen wesentlich von der IG ab	(3)	

Anmerkung: IG: Interventionsgruppe, KG: Kontrollgruppe, VG: Vergleichsgruppe.

Abbildung 12: Kriterium B.5.: Strukturelle Äquivalenz bei Kontrollbedingungen (interne Validität)

Beide Kriterien (B.4. und B.5.) sind damit strenggenommen nicht nur für Ein-Gruppen-Designs unkodierbar, sondern zudem bei Studien mit „echten“ Vergleichsgruppen, wie sie bspw. im Rahmen komparativer Psychotherapiestudien zu finden sind. Unter „echten“ Vergleichsgruppen sind dem vergleichenden Evaluationsparadigma (vgl. Hager, 2000) zufolge Treatmentarme zu verstehen, in denen Therapieformen realisiert werden, die nicht als *Kontroll*bedingungen fungieren: Das sind zum einen „bereits etablierte Treatments“ (z.B. kognitiv-behaviorale Therapie), im Vergleich zu denen sich eine zu überprüfende Therapieform als gleich wirksam oder zumindest nicht als maßgeblich weniger wirksam erweisen sollte[35] (u.a. Lange, Bender, Ziegler, 2007, Röhmel, Hauschke, Koch & Pigeot, 2005). Zum anderen zählen dazu „verfahrensexterne Vergleichsgruppen“ mit Treatments, die zwar (noch) nicht als etabliert betrachtet werden können, jedoch ebenso, wie etablierte Treatments und anders als etwa eine Placebo-Bedingung, über ein theoretisches Behandlungskonzept verfügen (z.B. die

[35] Diese Vergleiche werden mittels sog. *Non-Inferiority*- respektive einseitigen Äquivalenzhypothesen überprüft (u.a. Committee for Medicinal Products for Human Use [CHMP], 2006; Klemmert, 2004; Rogers, Howard & Vessey, 1993; vgl. Kap. 3.2.3).

Psychodramatherapie)[36]. Im Gegensatz zu klassischen Kontrollbedingungen dürfen bei „echten“ Vergleichsgruppen die sog. Randbedingungen denen der zu überprüfenden Therapie nicht angeglichen werden:

> Um die Repräsentativität der Programme (Kazdin, 1986b, 1994) zu gewährleisten, müssen die alternativen Programme in der von ihren Autor(inn)en vorgesehenen und einer in der Praxis verbreiteten Form durchgeführt werden. Dies heißt insbesondere, dass die Randbedingungen der beiden Programme einander nicht angeglichen werden dürfen, wie dies leider oft geschieht, selbst dann nicht, wenn sie sich hinsichtlich ihrer Randbedingungen (also etwa der vorgesehenen Länge) sehr unterscheiden. Man versteht dabei die Programme . . . sozusagen als „Paket“, das für eine vergleichende Wirksamkeitsuntersuchung nicht „aufgeschnürt“ werden sollte (vgl. auch Hager et al., 1999). (Hager, 2000, S. 192)

Zu Randbedingungen sind demnach all diejenigen Aspekte zu zählen, die genuin zu einer bestimmtem Behandlungsform gehören, etwa die Sitzungsfrequenz, die Therapiedauer, das Setting etc.. Würde man die Behandlungsdauer einer kognitiv-behavioralen Individualtherapie (Maximaldauer: 80 Sitzungen) derjenigen angleichen, die in analytischen Langzeitbehandlungen möglich ist (Maximaldauer: 300 Sitzungen), würde man die Vergleichsbehandlung (kognitiv-behaviorale Therapie) derart deformieren, dass sie strenggenommen nicht mehr als „echte“ Vergleichsbehandlung angesehen werden kann: „In diesem Fall ist die Repräsentativität des Alternativprogramms in unbekanntem Maße herabgesetzt, und ich nenne die entstehende Programmvariante *„Quasi-Alternativprogramm“* (QAP)“ (Hager, 2000, S. 194).

[36] Als „etablierte Treatments“ werden all diejenigen psychotherapeutischen Verfahren betrachtet, die zum Zeitpunkt der Kodierung (Mitte 2010-Anfang 2012) vom WBP als wissenschaftlich anerkannte Therapieverfahren (für das Erwachsenenalter) eingestuft und damit für die vertiefte Ausbildung empfohlen wurden: Das sind die Verhaltenstherapie, die Gesprächspsychotherapie und die systemische Therapie.

Das bedeutet, je mehr man die Randbedingungen einer Vergleichsbehandlung denen der zu überprüfenden Therapie angleicht, desto mehr verliert die Vergleichsbehandlung ihren Status als „echte“ Vergleichsbehandlung. Dieser Umstand der Angleichung von Randbedingungen ist in Kriterium B.5. (Abbildung 12, S. 147) unter dem Begriff *struktureller Äquivalenz* subsumiert.

In beiden Kriterien (B.4. und B.5.) sind also explizit *Kontroll*bedingungen angesprochen. Auf Ein-Gruppen-Designs und vor allem auch auf komparative Psychotherapiestudien mit „echten“ Vergleichsgruppen können beide Kriterien daher nicht angewandt werden, so dass hier Zusatzkodierungen (Missings) notwendig wurden (vgl. Abbildung 11 [S. 146] und Abbildung 12 [S. 147] jeweils linkes Textfeld).

Das dritte Kriterium, das samt seiner Zusatzkodierregeln illustriert wird (Kriterium A.11.; Abbildung 13), formuliert eine Bedingung, die zunächst erfüllt sein muss, damit das Kriterium bewertet werden kann: Grundsätzlich wird mit diesem Kriterium eingeschätzt, ob Fremdbeurteilungsverfahren im Rahmen einer Studie *lege artis* eingesetzt wurden. Das bedeutet, sie sollten möglichst von unabhängigen, trainierten und verblindeten Ratern verwendet worden sein. Voraussetzung für die eindeutige Kodierung einer Studie ist, dass tatsächlich Fremdeinschätzungen seitens externer Rater vorgenommen wurden.

A.11.	Sofern Fremdeinschätzungsverfahren: externe Beurteiler (blind für die Gruppenzugehörigkeit) • Wenn keine Fremdeinschätzungsverfahren durch externe Beurteiler eingesetzt: „9“ • Eingruppendesigns (inkl. Designs mit verfahrensinternen Vergleichsgruppen): mit externen Beurteilern: „4“	validiertes Fremdeinschätzungsverfahren angewendet von trainierten, für die Gruppenbedingungen blinden externen Beurteilern	(1)	**Auf primäre und sekundäre Zielkriterien bezogen** Validiert, trainiert und blind
		validiertes Fremdeinschätzungsverfahren angewendet von trainierten, nicht-blinden externen Beurteilern	(2)	Validiert, trainiert aber nicht-blind
		validiertes Fremdeinschätzungsverfahren angewendet – Rater sind weder trainiert noch blind für die Gruppenzugehörigkeit der Patienten	(3)	Validiert aber nicht-trainiert und nicht-blind *oder* nicht-valide entsprechend A.8.

Abbildung 13: Kriterium A.11.: Verblindeter Einsatz von validen Fremdeinschätzungsverfahren (allgemeine methodische Qualität)

Zu diesem Kriterium (A.11.) wurden zwei Zusatzkodierregeln notwendig, die eine betrifft Studien, in denen gar keine Fremdeinschätzungsverfahren durch externe Rater eingesetzt wurden (vgl. Abbildung 13 linkes Textfeld). Die zweite Zusatzkodierregel betrifft wiederum Studien ohne Kontroll-/Vergleichsgruppen: Das Kriterium setzt durch den Zusatz „blind für die Gruppenzugehörigkeit“ voraus, dass es sich um Mehrgruppendesigns handelt, womit Untersuchungen in Ein-Gruppen-Designs strenggenommen nicht kodierbar sind – selbst dann nicht, wenn Fremdeinschätzungsverfahren eingesetzt wurden. Untersuchungen, die im Ein-Gruppen-Design durchgeführt wurden, jedoch externe Rater einsetzten, wurden daher mit einem separaten Missingwert belegt.

Dieses Kriterium wurde für die Rekodierung (s.o.) modifiziert, indem die einzelnen Ratingstufen derart erweitert wurden, dass die beiden hier dargelegten Missingwerte durch reguläre Ratings ersetzt werden konnten (vgl. Kap. 3.4.2).

Weitere Zusatzkodierungen zu einzelnen Kriterien sind Anhang C zu entnehmen[37]. Da diese zum größten Teil selbsterklärend sind, wird an dieser Stelle auf weitere Ausführungen verzichtet und zum nächsten Kapitel übergeleitet.

3.2.2 Kodierregeln II: Minimierung der Interpretationsspielräume

Bei der ersten Probekodierung zeigte sich, dass einige Kriterien in ihren Ratingmöglichkeiten über einen erheblichen Interpretationsspielraum verfügen, der durch Kodierregeln eingeschränkt werden musste. Einige dieser Kodierregeln werden in diesem Kapitel vorgestellt, für einen Überblick über alle entwickelten Kodierregeln sei auf Anhang C verwiesen.

Grundsätzlich wurde die Formulierung der Kodierregeln in enger Anlehnung an die Vorgaben, die das jeweilige Kriterium macht, gestaltet. Definitorische Erweiterungen wurden vermieden bzw. ausschließlich in einem Maße vorgenommen, das den grundlegenden Inhalt des Kriteriums nicht verändert. Als Beispiel einer minimalen, definitorischen Erweiterung soll als erstes das interne Validitätskriterium B.3. (Abbildung 14) angeführt werden, mit dem die operationale Definition der Therapiebedingung bemessen wird (vgl. Kap. 1.2.2).

[37] Zusatzkodierregeln zwecks Missingvergabe befinden sich in der Regel in den Textfeldern in den jeweils linken Spalten der einzelnen Kriterien.

B.3.	Operationale Definition der Interventionen (Experimental- und ggf. Kontrollgruppe)	Therapiemanual, bei dem die Interventionen so beschrieben sind, dass das therapeutische Vorgehen vergleichbar und replizierbar ist	(1)	Verweis auf Manuale/ manualähnliche Behandlungsrichtlinien, das/die in den jeweiligen Treatments auch offensichtlich eingesetzt wurden
		Therapiebeschreibung, ohne nähere Spezifikation der einzelnen Interventionen (z.B. Fehlen gegebenenfalls für das psychotherapeutische Verfahren oder die psychotherapeutische Methode notwendiger Entscheidungskriterien)	(2)	Kompakte Beschreibung der Treatments im Rahmen der Studienpublikation *oder* Wenn Manual/ manualähnliche Behandlungsrichtlinie nur bsp.haft angeführt wird
		Die Intervention ist nicht klar beschrieben, beschränkt sich auf die Benennung des Psychotherapieverfahrens bzw. der Psychotherapiemethode	(3)	Lediglich Benennung der Behandlung(en), etwa „psychodynamische Therapie“ plus ggf. verfahrenstypischer Aspekte wie „Arbeit mit unbewussten Konflikten“ o.ä. *oder* Wenn kein bekannter Manual-/Behandlungsrichtlinientitel, dann danach recherchieren. Ggf. nachsehen ob Intervention in Artikel beschrieben → 2 (sonst 3)
	Welches Manual bzw. welche manualähnlichen Behandlungsrichtlinien wurden verwendet?			-

Abbildung 14: Kriterium B.3.: Operationale Definition der Interventionen (interne Validität)

Aus den Ankerbeispielen der einzelnen Ratingstufen des Kriteriums B.3. geht hervor, dass eine Studie, um mit einer „1“ bewertet zu werden, auf die Verwendung eines Manuals verweisen muss, so dass das Treatment ggf. replizierbar ist. Allerdings ist man in Bezug auf die Verwendung von *Manualen* im Zusammenhang mit psychodynamischen Psychotherapieformen mit folgendem Problem konfrontiert:

> Es geht hier nicht darum, eine Behandlung in zeitlich genau geplanten Schritten schematisch durchzuführen. Ein solches Vorgehen kommt eher stark strukturierten Therapieformen wie der kognitiven Verhaltenstherapie entgegen, entspricht aber nicht dem Verständnis psychodynamischer Psychotherapie. Im Rahmen der psychodynamischen Psychotherapie geht es vielmehr darum, Behandlungsrichtlinien zu

> formulieren, d. h. Interventionsprinzipien, Therapieelemente, Therapieziele sowie Indikationen und Kontraindikationen zu spezifizieren. Dazu gehören auch Angaben, in welchen Phasen der Therapie und welchen Übertragungs-Gegenübertragungskonstellationen welches Vorgehen empfohlen wird. (Beutel et al., 2010, S. 82)

Aus diesem Grund wurde die zu Ratingstufe „1“ gehörende Kodierregel explizit um die eher behandlungsprinzipienbasierten Manuale (manualähnliche Behandlungsrichtlinien) erweitert (vgl. Abbildung 14 rechte Spalte). Solche Manuale sind für die psychodynamischen Verfahren in den Werken von Beutel et al. (2010) und der DGPT (2009) sowie im Überblickswerk zum Thema *Psychodynamische Psychotherapie* von Reimer und Rüger (2006) zu finden.

Ratingstufe „2“ desselben Kriteriums (B.3.) lässt Therapiebeschreibungen mit leichten Einbußen zu. Da davon auszugehen ist, dass Therapiebeschreibungen in Studienpublikationen niemals den Grad an Ausführlichkeit, Präzision und damit Replizierbarkeit erlangen werden, wie es eigens für das therapeutische Vorgehen verschriftlichte Manuale/Behandlungsrichtlinien vermögen, können Studien, in denen das Vorgehen lediglich kompakt und ohne Hinweis auf ein angewandtes Manual beschrieben wird, lediglich mit einer „2“ bewertet werden. Studienpublikationen, die über derart ausführliche Therapiebeschreibungen verfügen, dass sie einer Behandlungsrichtlinie bzw. einem Manual nahe kommen und aus denen eindeutig hervorgeht, dass entsprechend dieser Beschreibungen auch behandelt wurde, werden demgegenüber mit „1“ bewertet.

Nun existieren Studien, in denen eine kompakte Therapiebeschreibung gegeben und für das therapeutische Vorgehen *beispielhaft* auf ein oder mehrere Manuale oder manualähnliche Behandlungsrichtlinien hingewiesen wird. Auch in einem solchen Fall wird die Studie mit einer „2“ kodiert. Wird in einer Studie lediglich das Label der Therapieform benannt – etwa „psychodynamische Psychotherapie“ oder „*relational-developmental psychoanalytic*

treatment“, ohne dass nähere Beschreibungen oder Quellenverweise folgen, dann schneidet eine solche Studie mit einer „3“ ab.

Das externe Validitätskriterium C.11. (Abbildung 15) bemisst, inwieweit die in einer Studie für eine psychotherapeutische Behandlung erforderlichen Behandlerqualifikationen beschrieben und in die klinische Praxis transferierbar sind.

C.11.	Spezifikation und Herstellbarkeit der notwendigen Behandlerqualifikation (Praxistransfer)	Notwendige Behandlungsqualifikation eindeutig beschrieben und herstellbar	(1)	z.B. die reguläre Ausbildungspflicht
		Notwendige Behandlungsqualifikation eindeutig beschrieben, aber nur mit sehr großem Zeitaufwand herstellbar	(2)	Hier sind nicht die regulären Ausbildungszeiten zum PP/ÄP gemeint, sondern die über die Fortbildungspflicht eines PP/ÄP deutlich hinausgehenden
		Notwendige Behandlungsqualifikation nicht beschrieben oder praktisch nicht herstellbar	(3)	-

Anmerkung: ÄP: Ärztlicher Psychotherapeut, PP: Psychologischer Psychotherapeut.

Abbildung 15: Kriterium C.11.: Spezifikation und Herstellbarkeit der notwendigen Behandlerqualifikation (externe Validität)

In der Regel ist für die psychotherapeutische Behandlungsbefugnis eine reguläre Ausbildung zum psychologischen oder ärztlichen Psychotherapeuten erforderlich. Diese eindeutig zu beschreiben, ist der gängigen Publikationspraxis zufolge unüblich und wäre im Rahmen einer Studienpublikation eine kaum einzulösende Forderung. Aus diesem Grund wurde eine Bewertung mit „1“ auf die Forderung beschränkt, dass der Publikation eindeutig zu entnehmen sein sollte, um welche Behandlerqualifikation für welche Therapieform es sich handelt. Trotz teilweise erheblicher Ausbildungsumfänge – etwa die vertiefte Ausbildung in analytischer Psychotherapie – werden diese im hiesigen Kontext als „herstellbar“ erachtet.

Werden in einer Studie zudem umfangreiche Zusatzqualifikationen als Voraussetzung zur Teilnahme als Studientherapeut beschrieben, die ein Training über die reguläre Ausbil-

dungspflicht zum psychologischen/ärztlichen Psychotherapeuten hinaus erfordern, so wird dies im Sinne des Praxistransfers der Untersuchung mit einer „2“ bewertet. Geht aus einer Publikation überhaupt nicht hervor, über welche Qualifikation für welche Therapieform die Behandler verfügen *oder* wird eine Qualifikation beschrieben, die in der Praxis nur sehr schwer realisierbar ist, so wird die Studie mit „3“ bewertet.

Das externe Validitätskriterium C.4. (Abbildung 16) fragt nach der Repräsentativität der in einer Studie realisierten psychotherapeutischen Behandlung im Hinblick auf die klinische Praxis im deutschen Gesundheitssystem. Die zu bewertende Repräsentativität bezieht sich bei diesem Kriterium primär auf das Vorgehen und die Dauer der Behandlung.

C.4.	Klinische Repräsentativität der Intervention hinsichtlich Vorgehen und Dauer	Intervention wie in klinischer Alltagspraxis	(1)	**Bewertung entsprechend Psychotherapierichtlinien primär die Therapiedauer betreffend. Vorsicht bei Bewertung des Vorgehens, da selbst innerhalb Richtlinienverfahren uneinheitlich vorgegangen wird.** Hinsichtlich Dauer wie in Alltagspraxis des untersuchten Verfahrens (z.B. TP: mindestens 25 Sitzungen) Wenn nur Range der Dauer abweicht (Hinweis auf Ausreißer) → 1
		Intervention gegenüber klinischer Alltagspraxis teilweise verändert	(2)	Leichte Abweichungen hinsichtlich Dauer des untersuchten Verfahrens, z.B. wenn SD der Dauer abweicht
		Intervention gegenüber klinischer Alltagspraxis stark verändert	(3)	Starke Abweichungen hinsichtlich Dauer des untersuchten Verfahrens (z.B. TP: 8 Sitzungen), bzw. wenn mittlere Dauer von max. Richtliniendauer (100 TP, 300 AP) nach oben hin abweicht

Anmerkung: AP: Analytische Psychotherapie, SD: Standardabweichung, TP: Tiefenpsychologisch fundierte Psychotherapie.

Abbildung 16: Kriterium C.4.: Klinische Repräsentativität der Intervention hinsichtlich Vorgehen und Dauer (externe Validität)

Voraussetzung für eine Kodierung dieses Kriteriums ist im Hinblick auf den psychodynamischen Treatmentarm zunächst die Feststellung, um welches psychotherapeutische Verfahren (tiefenpsychologisch fundierte Psychotherapie, analytische Psychotherapie oder Psychoanalyse) es sich bei der Studientherapie handelt. Vor allem Untersuchungen, die nicht im deutschen Sprachraum durchgeführt wurden, müssen demnach ins deutsche Versorgungssystem „übersetzt" werden. Dies wiederum setzt in Bezug auf die psychoanalytisch begründeten Verfahren zunächst eine genaue Abgrenzung zwischen tiefenpsychologisch fundierter und analytischer Psychotherapie sowie Psychoanalyse voraus, die Beutel et al. (2010) folgendermaßen kommentieren:

> Dabei ist die Streuung der jeweiligen konzeptuellen Bezugsrahmen in beiden Verfahrensrichtungen [gemeint sind tiefenpsychologisch fundierte und analytische Psychotherapie] vermutlich ähnlich groß . . . . Ob sich die Verwendung der klassischen, von Greenson (1975) formulierten Vorgehensweisen Konfrontation, Klärung, Deutung und Durcharbeiten in psychodynamischen und analytischen Psychotherapien quantitativ tatsächlich unterscheidet, ist noch nicht hinreichend überprüft. Hier dürfte die Varianz innerhalb der „Richtungen" ebenfalls größer sein als zwischen ihnen. (S. 45)

Die im Zitat angesprochene, konzeptionell nicht eindeutig zu ziehende Grenze zwischen analytischer und tiefenpsychologisch fundierter Psychotherapie wird zudem dadurch erschwert, dass hinsichtlich vergleichsweise harter Fakten, wie unterschiedlichen Sitzungsfrequenzen oder bzgl. des Settings, keine unhintergehbaren Regeln bestehen. So sind zwar der Psychotherapierichtlinie (Gemeinsamer Bundesausschuss, 2013) die jeweiligen Höchstbewilligungsumfänge der Richtlinienverfahren zu entnehmen und auch, dass eine tiefenpsychologisch fundierte, eine analytische oder eine Verhaltenstherapie die Frequenz von 3 Sitzungen pro Woche nur begründet und in einem abgegrenzten Zeitraum überschreiten sollte (§ 20). Jedoch macht die Richtlinie keine darüber hinausgehenden Vorgaben bzgl. der Sitzungsfrequenz. Unter-

schiedlichen Überblickswerken zu den psychoanalytisch begründeten Verfahren ist zu entnehmen, dass die analytische Psychotherapie regelhaft mit einer Frequenz von 2-3 wöchentlichen Sitzungen (liegendes Setting) durchgeführt wird, die tiefenpsychologisch fundierte Psychotherapie dagegen mit einer Frequenz von 1-2 Sitzungen pro Woche (sitzendes Setting) (Beutel et al., 2010; DGPT, 2009, 2011). Gleichsam ist bspw. der Psychotherapierichtlinie eine Abweichung von den genannten Regeln in Form von Sonderformen der tiefenpsychologisch fundierten Psychotherapie zu entnehmen (Gemeinsamer Bundesausschuss, 2013, § 14a): Dort wird eine niederfrequente und damit längerfristig Halt gebende, tiefenpsychologisch fundierte Behandlung vorgeschlagen, in der von der o.g. Frequenz abgewichen werden kann. Hier kann es z.B. zu Sitzungen in mehrwöchigen Abständen kommen (vgl. DGPT, 2009, 2011). Auch in der analytischen Psychotherapie werden unter bestimmten Umständen (etwa bei Patienten mit schweren strukturellen Defiziten) niederfrequente Behandlungen angeraten (vgl. DGPT, 2009, 2011). Beutel et al. (2010) berichten zudem von der sog. *niederfrequenten analytischen Psychotherapie nach Hoffmann*:

> Diese auf zwei bis fünf Jahre angelegte Behandlung findet in der Regel einmal in der Woche im Sitzen „über Eck“ statt . . . . Auch kann im Liegen oder im Wechsel zwischen Sitzen und Liegen behandelt werden. Eine Grundregel wird ebenso wenig vereinbart wie eine Regelung bezüglich des Stundenausfalls. (S. 12)

Vor allem was die Frequenz sowie das Setting der Behandlungen betrifft, bestehen also offenbar begründete Ausnahmen von den ansonsten geltenden Regeln. Und selbst, was das konzeptionell begründete Vorgehen in analytischen und tiefenpsychologisch fundierten Behandlungen betrifft, scheint es zumindest in der Praxis keine klare und gebotene Grenze zwischen den beiden Verfahren zu geben (vgl. Kap. 3.1.1). Diese Ausnahmeregelungen und Grenzver-

wehungen galt es bei der Entwicklung der Kodierregeln für Kriterium C.4. (Abbildung 16, S. 155) zu beachten.

Da das Kriterium C.4. nur vor dem Hintergrund einer vorhergehenden Verfahrenseinordnung kodiert werden kann, sollen diese beiden aufeinanderfolgenden Schritte kurz dargelegt werden:

1. Mit Hilfe des in Kapitel 3.2.4 näher zu beschreibenden Kurzkodierbogens und den dazugehörigen Kodierregeln (Ratzek & von Hauenschild, 2011a/b; Anhang B und D) wurde zunächst nach folgenden Aspekten und ggf. unter Berücksichtigung möglicher Ausnahmeregelungen (s.o.) die Verfahrenseinordnung vorgenommen: Diese Aspekte umfassen den Umfang der Behandlung (Sitzungsanzahl), das Setting (sitzend *versus* liegend), die Frequenz und – unter Vorbehalt – das in der Studie verwendete Therapielabel sowie – falls vorhanden – die Interventionsbeschreibung. Da sowohl in englisch- als auch in deutschsprachigen Publikationen keine einheitliche Nomenklatur für die psychodynamischen Richtlinienverfahren plus Psychoanalyse verwendet wird, die eine eindeutige Zuordnung zur tiefenpsychologisch fundierten oder analytischen Psychotherapie in allen Fällen auf Anhieb zulässt, gilt es, die publizierte Namensgebung der Behandlung (Therapielabel) stets genauer zu prüfen (vgl. Kap. 3.1.1). In Bezug auf die Sitzungsanzahl, das Setting und die Frequenz wurden folgende Orientierungswerte angelegt: Psychodynamische Studientherapien mit einer Sitzungsanzahl bis 100 Stunden, sitzende Position und 1-2 Sitzungen/Woche, waren der tiefenpsychologisch fundierten Psychotherapie zuzuordnen; Therapien mit einer Sitzungsanzahl bis 300 Stunden, liegende Position und 2-3 Sitzungen/Woche der analytischen Psychotherapie. Bei ggf. bestehenden Abweichungen in Bezug auf diese Orientierungswerte – z.B. entsprechend o.g. Modifikationsregeln – muss vor dem Hintergrund der restlichen In-

formationen im Einzelfall entschieden werden, welchem Verfahren die publizierte Therapie am ehesten zuzuordnen ist.

Weicht die publizierte Sitzungsanzahl von 300 Stunden nach oben hin ab, und gibt es zudem Hinweise auf durchgehend hochfrequente Behandlungen (liegend, mehr als 3 Sitzungen/Woche), so war die Behandlung der Psychoanalyse zuzuordnen.

2. Die Kodierung des Kriteriums C.4. wurde hauptsächlich unter Berücksichtigung der im deutschen Versorgungssystem definitiv vorgeschriebenen Umfangsbegrenzungen der psychoanalytisch begründeten Richtlinienverfahren durchgeführt. Mit diesem Fokus wurde – im Gegensatz zu der eher variabel gehaltenen Sitzungsfrequenz (s.o.) – eine stabile Referenz geschaffen, die auch durch die Psychotherapierichtlinie (Gemeinsamer Bundesausschuss, 2013) gestützt wird. Starke Abweichungen von den vorgeschriebenen Maximaldauern wurden in Abhängigkeit von der Größe der Abweichungen auf Kriterium C.4. dementsprechend „schlechter“ bewertet. Das bedeutet, wich bspw. nur der in einer Studie angegebene *obere Range* der Sitzungsanzahl von der Maximaldauer ab, so wurde diese Überschreitung als Ausreißer betrachtet und in Kriterium C.4. mit „1“ bewertet. Wich hingegen die berichtete *mittlere* Sitzungsdauer von der vorgeschriebenen Maximaldauer ab, so wurde eine solche Studie hier mit „3“ bewertet. Gleiches gilt für publizierte Therapieumfänge von z.B. 8 Sitzungen bei einer tiefenpsychologisch fundierten Behandlung, auch diese werden mit einer „3“ kodiert, da psychodynamische Kurzzeittherapie im deutschen Versorgungssystem bis zu 25 Sitzungen umfasst und 8 Sitzungen damit als nicht repräsentativ für die ambulante, tiefenpsychologisch fundierte Behandlungspraxis zu betrachten sind.

Bezogen auf Gruppentherapiestudien wird angepasst an die maximalen Höchstdauern von 80 Doppelstunden bei der tiefenpsychologisch fundierten und 150

Doppelstunden bei der analytischen Psychotherapie (vgl. Psychotherapierichtlinie) nach demselben Schema vorgegangen.

Sitzungsfrequenzen fielen bei der Kodierung von C.4. nur dann ins Gewicht, wenn sie etwa über die oben angeführten Modifikationsmöglichkeiten hinaus gingen (z.B. 5 Sitzungen/Woche bei einer tiefenpsychologisch fundierten Psychotherapie).

Für alle in diesem Kapitel dargestellten Kriterien (sowie auch für alle anderen Kriterien) gilt: Geht aus einer publizierten Studie keinerlei Information hervor, die für die Kodierung des jeweiligen Kriteriums obligatorisch wäre, wird die Studie bei diesen Kriterien mit „3“ bewertet. Informationen, die nicht explizit beschrieben sind, jedoch eindeutig aus dem Kontext der Studie hervorgehen, gehen hingegen in die Bewertung ein.

3.2.3 Kodierregeln III: Grundlegende Spezifizierungen einzelner Kriterien

Mit einem der internen Validitätskriterien (Kriterium B.11.) wird u.a. die Länge des Katamnesezeitraums bewertet, der in einer Studie realisiert wurde (vgl. Kap. 1.2.2 und Anhang A). Dem Kriterium lässt sich die Formulierung: „zeitlich störungsangemessene Katamnese“ (WBP, 2010, S. 33) entnehmen. Ein anderes Kriterium (A.4.) bewertet die Höhe der Dropoutquote, die zwischen der Post- und der Katamnesemessung anfällt (vgl. Kap. 1.2.2 und Anhang A). Die Referenz, an der die Dropoutraten in einer Studie bemessen werden, soll folgende sein: „deutlich besser als in Studien mit vergleichbaren Patientengruppen und vergleichbarem Katamnesezeitraum“ (WBP, 2010, S. 30).

Die eben zitierten Formulierungen legen einen Vergleichsmaßstab nahe, der von den Begutachtenden zunächst eruiert werden muss. In Bezug auf den "störungsangemessenen Katam-

nesezeitraum" (Kriterium B.11.) bedeutet dies, zunächst den natürlichen Verlauf (d.h. ohne Behandlung) der betroffenen Störung zu ermitteln, um sodann eine Aussage darüber treffen zu können, innerhalb welchen Zeitraums nach symptomatischer Remission oder Genesung mit den höchsten Rückfallraten zu rechnen ist. Eben diesen Zeitraum gilt es durch eine störungsangemessene Katamnese zu umfassen (s.u.). Zur Bewertung der Höhe von Dropoutquoten zwischen der Post- und Katamnesemessung (Kriterium A.4.) wird wiederum ein Vergleichsmaßstab benötigt, der festlegt, mit welchen Dropoutquoten innerhalb dieses Zeitfensters *normalerweise* zu rechnen ist und welche Quote das normale Maß überschreitet. Dazu gilt es vergleichbare Studien heranzuziehen, die über einen ähnlichen Katamnesezeitraum verfügen, wie die zu bewertende Studie und die Dropoutquoten miteinander zu vergleichen. Dazu sollte selbstverständlich nicht nur *eine* Studie als Vergleichsmaßstab herangezogen werden, sondern eine breitere empirische Basis – soweit vorhanden (s.u.).

Neben den genannten beiden Kriterien (A.4. und B.11.) erforderten noch weitere Kriterien nähere Spezifizierungen und die Entwicklung von Vergleichsmaßstäben. Die Darstellung der in diesem Unterkapitel zu behandelnden Kodierregeln erfolgt in der Zuordnung der betreffenden Kriterien zu folgenden Themenblöcken:

- Reliabilität und Validität der primären Outcomemaße
- Bericht unterschiedlicher Indikatoren der Effektivität
- Dropoutquote/Ausschöpfungsquote und Katamnesezeitraum

Reliabilität und Validität der primären Outcomemaße

Mit dem methodischen Qualitätskriterium A.8. (Abbildung 17) wird bemessen, inwieweit die in den Studien eingesetzten Outcomemaße psychometrischen Anforderungen genügen (Reliabilität und Validität).

A.8.	Reliable und valide Messung zumindest der **primären** Zielkriterien	reliable und valide Outcome-Verfahren	(1)	In Primärstudie werden ausschließlich Outcome-Verfahren verwendet, die mit **„exzellent“**, **„gut“** oder **„zufriedenstellend“** bewertet wurden
		nur eingeschränkte Reliabilität und/oder Validität der Messverfahren	(2)	In Primärstudie werden • u.a. Outcome-Verfahren verwendet, die als **„unzureichend“** bewertet wurden, die jedoch einen Anteil von 25% der Gesamtanzahl verwendeter Outcomemaße nicht übersteigen *oder* • Outcome-Verfahren, von denen mindestens eines als **„ausreichend“** bewertet wurde
		Reliabilität und Validität der Messverfahren nicht überprüft oder Gütekriterien der Messverfahren sind unzureichend **(Stufe 3 = Ausschlusskriterium)**	(3)	In Primärstudie werden mehr als 25% Outcome-Verfahren verwendet, die als **„unzureichend“** bewertet wurden

Abbildung 17: Kriterium A.8.: Reliable und valide Messung der primären Zielkriterien (allgemeine methodische Qualität)

Um Kriterium A.8. anwenden zu können, bedarf es zum einen eines Bewertungssystems für die psychometrische Beurteilung der in den Studien verwendeten Outcomemaße. Zum anderen ist eine Operationalisierung der drei Ratingsstufen im Sinne von Cutoff-Werten erforderlich: In den meisten Studien wird mehr als nur ein primäres Outcomemaß verwendet, so dass festgelegt werden muss, wie unterschiedliche Verhältnisse psychometrisch „guter“ *versus* „unzureichender“ Outcomemaße mit Kriteriums A.8. bewertet werden sollen.

Für eine objektive psychometrische Beurteilung aller einzelnen primären Outcomemaße, die in den 41 Studien Verwendung finden, wurde eigens und in Anlehnung an einschlägige Beurteilungssysteme sowie methodische Abhandlungen ein Bewertungssystem entwickelt (Ratzek & von Hauenschild, 2011c; Anhang E). Orientiert wurde sich dazu an Systemen, wie dem niederländischen *Committee On Test Affairs Netherland* System (CONTAN System; Evers,

2001a/b), dem *Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen* (TBS-TK; Testkuratorium, 2006), dem *EFPA Review Model for the Description and Evaluation of Psychological Tests* (Bartram, Lindley & Kennedy, 2008), der *DIN 33430 für berufsbezogene Eignungsdiagnostik* (Deutsche Industrienorm 33430, 2002) sowie den *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999). Hinzugezogen wurden außerdem Publikationen namhafter Methodiker, wie Cicchetti (1994, 2001), Cicchetti und Rourke (2004), Shrout und Fleiss (1979), McGraw und Wong (1996), Fisseni (1997), Wirtz und Caspar (2002), Asendorpf und Wallbott (1979), sowie einzelne Paper aus der Reihe *EbM-Splitter* (Schwarzer, Türp & Antes, 2002a/b/c/d).

Die psychometrische Bewertung der einzelnen Outcomemaße mittels des eigens entwickelten Bewertungssystems erfolgt auf zwei Stufen: Basisbewertung (1) und Koeffizientenbewertung (2). Jede herangezogene Publikation, in der psychometrische Eigenschaften eines Selbst- oder Fremdbeurteilungsverfahren berichtet werden – Reliabilitäts- und Validitätsstudien im Original – wird derselben Bewertungsprozedur unterworfen. Um die Phase des Beurteilungsprozesses für den Rahmen dieser Dissertation angemessen zu gestalten, wurden nicht *alle* zugänglichen Reliabilitäts- und Validitätsstudien für jedes Outcomemaß gesichtet, sondern jeweils ein Mindestmaß an empirischen Befunden gefordert, das sich in der Basisbewertung (1) widerspiegelt. Dieses Mindestmaß bezieht sich auf eine Kombination aus der Anzahl unabhängiger Datensätze, der jeweilig zugrunde liegenden Stichprobengröße und der Anzahl berichteter unterschiedlicher Reliabilitäts- und Validitätsarten (vgl. Anhang E). In der Literatursuche nach Reliabilitäts- und Validitätsbefunden wurden zunächst so viele Publikationen zusammengestellt, wie möglich, um dann diejenigen Publikationen zur Bewertung psychometrischer Eigenschaften auszuwählen, die entsprechend des o.g. Mindestmaßes die robusteste

Grundlage boten[38]. Grundsätzlich wurde darauf geachtet, dass die psychometrischen Befunde sich stets auf die sprachliche Originalversion des Erhebungsinstruments beziehen. Diese Entscheidung erfolgte aus forschungspragmatischen Gründen, da zahlreiche Outcomemaße in die zu kodierenden Studien in unterschiedlichen sprachlichen Fassungen eingehen. Für jede Fassung eine eigene psychometrische Bewertung vorzunehmen, hätte den Rahmen dieser Arbeit gesprengt.

Außerdem wurden Reliabilitäts- und Validitätsnachweise gefordert, die auf der Grundlage solcher Stichproben realisiert wurden, die für die intendierte Zielpopulation des Erhebungsinstruments als repräsentativ zu betrachten sind. In der Regel waren dies klinische Zielpopulationen. Ausnahmen bildeten klinische Screening-Instrumente (etwa *General Health Questionnaire* [GHQ-12], Goldberg, 1978), bei denen neben Reliabilitäts- und Validitätsstudien aus dem klinischen Kontext auch Untersuchungen auf der Basis von Normalbevölkerungsdaten hinzugezogen wurden. Gleiches gilt für Persönlichkeitstests. Für Inventare zur Messung der allgemeinen Lebenszufriedenheit/-qualität sowie des allgemeinen Wohlbefindens, deren Anwendungskontexte entweder die Normalbevölkerung oder aber den somatischen/psychotherapeutischen Bereich betreffen, wurden ebenfalls dementsprechende Studien herangezogen. Für Instrumente zur Erhebung somatischer Beschwerden, die meist in somatischen oder psychosomatischen Kontexten eingesetzt werden, dominieren entsprechend psychometrische Untersuchungen in eben diesen Bereichen.

An diese Besonderheiten wurde die bereits formulierte Mindestanforderung (Basisbewertung) in der Weise angepasst, dass, wenn es sich um ein Instrument mit einer *heterogenen* intendierten Zielpopulation handelte, diese Heterogenität sich in den Stichproben der

[38] In der Literatursuche nach Originalstudien wurde sich zunächst an Handbüchern und Testrezensionen orientiert (z.B. Brähler, Schumacher & Strauß, 2002; Johnson, 2010; Strauß & Schumacher, 2005).

psychometrischen Untersuchungen widerspiegeln sollte. In diesen Fällen mussten demzufolge meist mehrere Studien herangezogen werden.

Die im Hinblick auf die genannten Erfordernisse realisierte Literaturrecherche pro Erhebungsinstrument brachte in der Regel zwischen 1 bis 10 Publikationen zu Reliabilitäts- und Validitätsbefunden hervor. Aus diesen wurden diejenigen ausgewählt, die entsprechend des o.g. Mindestmaßes (unabhängige Datensätze, ausreichende Stichprobengrößen, Bericht unterschiedlicher Reliabilitäts- und Validitätsarten und ggf. angemessene Breite der untersuchten Probandengruppen entsprechend der intendierten Zielpopulation) über die reichhaltigsten reliabilitäts- und validitätsrelevanten Informationen verfügten. Diese Studien bildeten in einem nächsten Schritt die Grundlage für sog. Inventarrezensionen, die für jedes einzelne Erhebungsinstrument eigens verfasst wurden. Eine Veröffentlichung der gesammelten Rezensionen über 80 Selbst- und Fremdbeurteilungsverfahren befindet sich in Vorbereitung (Ratzek & von Hauenschild, in Vorb.). Eine Rezension wurde nach einer immer selben Struktur verfasst und bezieht sich in der Regel auf bis zu sechs Reliabilitäts- und Validitätsstudien[39]. Auf Grundlage dieser Rezensionen (und ggf. unter wiederholter Hinzuziehung der Originalliteratur) wurde schließlich die Basis- und Koeffizientenbewertung vorgenommen.

In der Basisbewertung (1), die bereits für die Auswahl der heranzuziehenden Studien leitend war, wurde zunächst die Breite der empirischen Basis der Reliabilitäts- und Validitätsbefunde bewertet. Bei ausreichender Basis wurde die Koeffizientenbewertung (2) vorgenommen, d.h., die Basisbewertung fungiert als Voraussetzung für alle weiteren Bewertungsschritte.

[39] Wenn zusammenfassende Testrezensionen aus Handbüchern aussagekräftig genug waren, wurden auch diese Sekundärquellen in die Rezensionen aufgenommen.

Wie der Name es nahelegt, fokussiert die Koeffizientenbewertung auf die in den ausgewählten Studien berichteten Koeffizienten als Indikatoren für die Reliabilität und Validität des betreffenden Erhebungsinstruments. Zusammenfassend werden aus den einzelnen Koeffizientenbewertungen Mittelwerte gebildet, die zusammen mit der Basisbewertung die psychometrische Endbewertung des Instruments beschließen. Dabei kann die Koeffizientenbewertung maximal so „gut“ sein, wie die Basisbewertung. Im Extremfall bedeutet dies, dass bei extrem schlechter empirischer Basis den Koeffizienten keinerlei Beachtung mehr geschenkt wird.

In der Kodierregel für das Kriterium A.8. (Abbildung 17, S. 162) wurde versucht, eine Operationalisierung der drei Ratingstufen zu konzipieren, in der sich möglichst Äquidistanz widerspiegelt. Mit einer „1“ wird eine Studie dann bewertet, wenn die verwendeten Outcomemaße entsprechend ihrer psychometrischen Endbewertung ausschließlich mit *exzellent*, *gut* oder *zufriedenstellend* abschneiden (vgl. Anhang E). Mit einer „2“ wird eine Studien bewertet, wenn mindestens ein *ausreichendes* Erhebungsinstrumente eingesetzt wurde, oder aber, wenn auch *unzureichende* Instrumente verwendet wurden, die jedoch anteilmäßig nicht mehr als 25% aller in der betreffenden Studie verwendeten Outcomemaße ausmachen. Dementsprechend wird eine Studie mit „3“ bewertet, wenn diese 25%-Schwelle überschritten wurde.

Wurde die Reliabilität verwendeter Fremdbeurteilungsverfahren in den zu kodierenden Wirksamkeitsstudien zudem selbst überprüft, blieb diese Information nicht unberücksichtigt. Es wurde dann eruiert, ob diese Informationen sich auf die zuvor ermittelten Reliabilitätsbefunde verbessernd oder verschlechternd auswirken. Veränderungen in beide Richtungen konnten sich entsprechend auf die Bewertung in Kriterium A.8. (Abbildung 17, S. 162) auswirken.

Bericht unterschiedlicher Indikatoren der Effektivität

Die beiden im Folgenden zu behandelnden Kriterien beziehen sich auf einen ähnlichen Aspekt der Studienbewertung und werden daher bzgl. der Kodierregelentwicklung zusammenhängend illustriert. Es handelt sich zum einen um das interne Validitätskriterium B.12., mit dem bemessen wird, ob in einer Studie unterschiedliche Veränderungs- und Zielerreichungsmaße berichtet werden (statistische und klinische Signifikanzmaße sowie Effektstärken; vgl. Kap. 1.2.2). Zum anderen geht es um das methodische Qualitätskriterium A.9., das sich allein auf das Konzept der klinischen Bedeutsamkeit bezieht[40] (vgl. Kap. 1.2.2). Beide Kriterien fokussieren damit auf verschiedene Darstellungsformen der Effektivität einer Intervention.

Für die Kodierregeln des Kriteriums B.12. (Abbildung 18) wurde in erster Linie eine Publikation von Schulte (1993) herangezogen, in der detailliert die unterschiedlichen Formen der Darstellung eines Therapieerfolgs vorgestellt werden:

> Grundsätzlich können zwei Arten des Vergleichs unterschieden werden: Vergleich mit dem Ausgangszustand vor Therapiebeginn – das Resultat dieses Vergleichs kennzeichnet den Grad der *Veränderung* – und Vergleich mit einem Zielzustand oder einer Norm – das Resultat dieses Vergleichs ist der Grad der *Zielerreichung*. (S. 383)

In der Formulierung von B.12. lassen sich drei *Veränderungsdarstellungen* wiederfinden, die Messung der Differenz zwischen dem Prä- und Postzeitpunkt oder aber zwischen Gruppen im Hinblick auf ihre statistische Signifikanz (1), die Berechnung von Effektstärken (2) und die Berechnung individueller Veränderungsmessungen gemäß des RCI (3).

[40] Die Begriffe „klinische Bedeutsamkeit“ und „klinische Relevanz“ werden hier synonym verwendet.

B.12.	Erzielte Veränderungen auf den primären und sekundären Zielkriterien ggf. im Vergleich zur Kontrollgruppe (Signifikanz, Größe und Relevanz der Effekte [41])	vollständige Darstellung der erzielten Veränderungen auf den Zielkriterien inklusive der Signifikanz, Größe der Effektmaße und Ausmaß der klinisch relevanten Zielerreichung (ggf. im Vergleich zur Kontrollgruppe)	(1)	Sowohl Indikatoren der *Veränderungen* als auch der *Zielerreichungen* sind in Bezug auf die verwendeten Outcomemaße dargestellt **Achtung:** *Zielerreichung* weniger streng bewerten, wenn Cutoff angewandt
		Darstellung des Behandlungsergebnisses nur durch Veränderungs- oder Zielerreichungsmaße oder beides ist (ggf. im Vergleich zur Kontrollgruppe) bei einigen Kriterien unvollständig	(2)	Es sind nur Indikatoren der *Veränderungen* oder nur der *Zielerreichungen* in Bezug auf die verwendeten Outcomemaße dargestellt
		weitgehend unvollständige oder inadäquate Darstellung der Outcome-Kriterien (ggf. im Vergleich zur KG) **(Stufe 3 = Ausschlusskriterium)**	(3)	-

Abbildung 18: Kriterium B.12.: Veränderungs- und Zielerreichungsmessungen (interne Validität und K.O.-Kriterium der allgemeinen methodischen Qualität)

Unter dem RCI ist ein klinisches Signifikanzmaß zu verstehen, das die intraindividuelle Veränderung eines Patienten im Verlauf einer Behandlung relativiert an der Veränderung, die allein aufgrund des Messfehlers zustande kommen kann, abbildet (vgl. Jacobson & Truax, 1991; Jacobson, Follette & Revenstorf, 1984; Stieglitz, 2008). Für eine Veränderung, die als tatsächliche (reliable) Verbesserung oder Verschlechterung der Symptomatik (oder anderer Outcomemaße) interpretiert werden kann, muss die Differenz zwischen zwei Messzeitpunkten

[41] *Veränderungen:* Neben Prä-Post-Differenzen:

- Prozentuale Angabe der Probanden, die entsprechend dem „Reliable Change Index" eine signifikante Veränderung erreicht haben
- Effektstärkeberechnungen (als Indikator des Ausmaßes der Veränderung)

Zielerreichung:

- Klinische Signifikanz: Reliable Change Index plus Cutoff-Wert (Vergleich mit Normpopulation oder dysfunktionaler Population)

(vgl. Schulte, 1993; Möller, Laux & Kapfhammer, 2005).

(Prä- und Postmessung) über den zu erwartenden Standardfehler der Differenz hinausgehen. Tut sie dies nicht, so muss von einer Veränderung ausgegangen werden, die allein auf Reliabilitätseinbußen des Erhebungsinstruments zurückzuführen ist: „RC tells us whether change reflects more than the fluctuations of an imprecise measuring instrument“ (Jacobson & Truax, 1991, S. 14).

Im Gegensatz zu Effektstärken, in deren Berechnung in der Regel über die gesamte Stichprobe gemittelte Werte eingehen, wird beim RCI die *intraindividuelle* Veränderung eines jeden Patienten betrachtet. Gängige Darstellungsweisen dieses Maßes sind demzufolge prozentuale Angaben über Patienten, die sich entsprechend des RCI reliabel verändert haben. In den meisten Darstellungen wird unterschieden zwischen dem Patientenanteil, der als *improved* (verbessert), als *recovered* (genesen), *unchanged* (unverändert) oder als *deteriorated* (verschlechtert) betrachtet werden kann. Die Unterscheidung zwischen *improved* und *recovered* kommt dabei folgendermaßen zustande: Ein Patient sollte über eine reliable positive Veränderung (*improved*) hinaus einen zuvor eruierten Cutoff-Wert, in Abhängigkeit von der Polung der verwendeten Skala, über- oder unterschreiten, damit von tatsächlich erfolgter Genesung (*recovery*) ausgegangen werden kann (Jacobson, Roberts, Berns & McGlinchey, 1999). Dieser Cutoff-Wert wird in der Regel anhand von Normen festgelegt und markiert denjenigen Wert, der den dysfunktionalen vom funktionalen Bereich trennt. Da es sich dabei strenggenommen nicht mehr um ein *Veränderungsmaß* handelt, subsumiert Schulte (1993) diesen Vergleich zwischen individuellem Postmesswert und einem festgelegten Cutoff-Wert dementsprechend unter die Indikatoren der *Zielerreichung*.

Die Kodierregeln zu Kriterium B.12. (Abbildung 18) wurden folgendermaßen spezifiziert: Eine Studie, die als *Veränderungsindikatoren* neben der gruppenstatistischen Signifikanz zu-

sätzlich Effektstärken oder intraindividuelle Veränderungen gemäß RCI berechnet, und die außerdem als *Zielerreichungsindikator* einen Vergleich individueller Messwerte mit Normgruppenwerten vornimmt (Cutoff-Werte), wird mit einer „1“ bewertet. Die Zielerreichung muss in diesem Fall nicht zwingendermaßen durch die individuelle RCI-Berechnung gestützt werden. Die Berechnung des Patientenanteils, der einen normgebundenen Cutoff-Wert über- oder unterschreitet, ist in diesem Fall ausreichend[42].

Für eine Bewertung mit „1“ wurde – abweichend von den Vorgaben des Kriteriums – nicht gefordert, dass die klinisch relevante *Zielerreichung* über *alle* verwendeten Outcomemaße eruiert werden muss. Der Grund für diese etwas weniger strenge Forderung liegt darin, dass für den Vergleich individueller Postmesswerte mit Normgruppenwerten zunächst einmal solche Normen vorliegen müssen, was jedoch nicht für jedes Outcomemaß der Fall ist (vgl. Stieglitz, 2008).

Werden in einer Studie entweder Indikatoren der *Zielerreichung* oder der *Veränderung* berichtet, so schneidet eine solche Studie mit einer „2“ ab. Der Kritik an der Darstellungsform der Effektivität einer Intervention, die ausschließlich auf Signifikanztests und Effektstärken beruht, wird mit dieser Bewertung Rechnung getragen: Ohne den Einbezug eines Mindestmaßes an klinischer Signifikanzbestimmung fehlt einer Evaluationsstudie ein grundlegender Bestandteil der Ergebnisdarstellung und ihre Aussagekraft ist somit reduziert (vgl. APA, 2002; Chambless & Hollon, 1998; Kendall et al., 2004).

Eine Studie, die in ihrer Ergebnisdarstellung ausschließlich auf inferenzstatistische Auswertungen ohne Berücksichtigung von Effektstärken und Indikatoren der Zielerreichung

[42] Unter diesen Umständen ist jedoch zu beachten, dass das Über- oder Unterschreiten des Cutoff-Wertes strenggenommen nicht mit *recovery* gleichzusetzen ist, da der Nachweis einer reliablen Verbesserung fehlt. Die limitierte Aussagekraft der alleinigen Verwendung eines Cutoff-Wertes wird in Kriterium A.9. in Form einer strengeren Bewertung berücksichtigt.

setzt, wird mit einer „3“ bewertet. Eine „3“er-Bewertung kommt dabei einem Ausschluss der Studie auf der allgemeinen methodischen Qualitätsdimension gleich (vgl. Kap. 1.2.2).

Kriterium A.9. fordert nun im Speziellen den Bericht von Indikatoren, an denen sich die klinische Bedeutsamkeit ablesen lässt (vgl. Abbildung 19).

A.9.	Klinische Bedeutsamkeit der Outcome-Messung (z.B. das Konzept der klinischen Signifikanz [43])	klinische Bedeutsamkeit des Therapieeffekts (z. B. im Sinne des Konzepts der klinischen Signifikanz) ist feststellbar	(1)	• Normativer Vergleich • RCI + Cutoff-Wert • Soziale Validierung/ Subjektive Evaluation • Diagnosefreiheit **Achtung:** Hier klinische Bedeutsamkeit strenger bewerten: Wenn Cutoff <u>und</u> RCI → 1, wenn nur Cutoff <u>oder</u> RCI → 2
		klinische Bedeutsamkeit des Therapieeffekts (z. B. im Sinne des Konzepts der klinischen Signifikanz) ist nur eingeschränkt feststellbar	(2)	• Normativer Vergleich nur mittels klassischer Signifikanztests • nur RCI oder nur Cutoff-Wert
		klinische Bedeutsamkeit des (z. B. im Sinne des Konzepts der klinischen Signifikanz) Therapieeffekts ist nicht feststellbar	(3)	Die klinische Bedeutsamkeit wurde über keine der Outcomemaße berichtet

Anmerkung: RCI: Reliable Change Index.

Abbildung 19: Kriterium A.9.: Klinische Bedeutsamkeit der Outcomemessung (allgemeine methodische Qualität)

[43] Als Erhebung der klinischen Relevanz werden i.d.R. folgende Strategien betrachtet:

1. Normativer Vergleich mittels klassischer und Äquivalenzhypothesentests (vgl. Kendall, Marrs-Garcia, Nath & Sheldrick, 1999)
2. Individuelle Veränderungen (z.B. „Reliable Change Index“ als „statistically reliable improvement“ [Verbesserung]; „Cutoff-Wert“ als „clinically significance change“ [Genesung]) (vgl. Jacobson & Truax, 1991; Stieglitz, 2008)
3. Soziale Validierung/Subjektive Evaluation durch Patienten selbst, Verwandte/Freunde, Fachkräfte/externe Beobachter (vgl. Kazdin, 1994; Lambert & Ogles, 2004)
4. Kein weiteres Erfüllen der Kriterien einer psychiatrischen Diagnose (vgl. Kazdin, 2008).

Während das Begriffspaar „klinische Bedeutsamkeit“ als Oberbegriff fungiert, kennzeichnet die *klinische Signifikanz* in der Regel einen Bereich der klinischen Bedeutsamkeit, der bereits in Form des RCI und des normgebundenen Cutoff-Wertes eingeführt wurde (u.a. Kendall et al., 2004)[44]. In der einschlägigen Literatur werden darüber hinaus weitere Möglichkeiten der klinischen Bedeutsamkeitsbestimmung diskutiert: Neben der klinischen Signifikanzberechnung via RCI und Cutoff-Werten werden normative Vergleiche mittels klassischer und Äquivalenzhypothesentests (Kendall, Marrs-Garcia, Nath & Sheldrick, 1999) sowie soziale Validierungsmethoden (Hill & Lambert, 2004; Kazdin, 1994) diskutiert. Kazdin (2008) führt als weiteres klinisch relevantes Erfolgskriterium den Rückgang einer Anfangsdiagnose ein.

Bis auf den normativen Vergleich nach Kendall et al. (1999) ist allen Methoden die *individuelle* Betrachtung der Patienten gemeinsam, was als das intendierte Ziel klinischer Bedeutsamkeitsmaße angesehen werden kann (Hill & Lambert, 2004; Kazdin, 1994). Beim normativen Vergleich wird von der individuellen Betrachtung abgewichen und stattdessen eine gruppenstatistische Auswertung gefordert, in der klassische Nullhypothesentests mit Äquivalenztests kombiniert werden. Da es sich hierbei um ein zentrales Konzept handelt, soll die Logik von Äquivalenztests[45] sowie der normativen Vergleiche nach Kendall et al. im Folgenden näher beschrieben werden.

Äquivalenztests sind vor allem in pharmakologischen Äquivalenzprüfungen gängig: Ein neues, preiswerteres Präparat soll dem Effekt des etablierten Präparats nicht wesentlich nachstehen, d.h., es soll gezeigt werden, dass das neue Präparat maximal *unwesentlich* schlechter

[44] Diese Begriffsdifferenzierung wird nicht in allen Publikationen stringent genutzt (vgl. Faller & Reusch, 2004), im Englischen verwischt die Abgrenzung zudem durch die Übersetzung von „Bedeutsamkeit“ in „*significance*“. Für die vorliegende Arbeit soll jedoch die genannte Differenzierung zugrunde gelegt werden.

[45] Das Verständnis von Äquivalenztests ist zudem grundlegend für die Kodierung eines weiteren Kriteriums (A.17., vgl. Anhang C).

abschneidet, als das herkömmliche (vgl. Klemmert, 2004). Man geht demzufolge von äquivalenter Wirksamkeit beider Präparate aus, mit dem Zugeständnis einer lediglich *irrelevanten* Unterlegenheit des neuen Präparats. Solche Äquivalenzhypothesen werden in der Regel einseitig getestet, da ein sog. Irrelevanzbereich in nur eine Richtung gelegt wird. Ein Besserabschneiden des neuen Präparats ist in diesem Fall nicht von Interesse. Dies entspricht der bereits eingeführten *Non-Inferiority*-Hypothese (vgl. Kap.1.2.2). Liegt bspw. eine empirische Differenz zuungunsten des neuen Präparats vor, befindet sich samt Konfidenzintervall jedoch noch innerhalb des a priori festgelegten Irrelevanzbereichs, so würde das neue Präparat als effektäquivalent zum herkömmlichen gelten. Erst bei Überschreiten der Irrelevanzgrenze durch den empirischen Differenzwert oder aber durch eine seiner Konfidenzintervallgrenzen würde man von einer Unterlegenheit des neuen Präparats ausgehen (grafisch anschaulich dargestellt von Temple & Ellenberg, 2000, S. 461).

Das Besondere an Äquivalenztests ist, dass die Nullhypothese keinen Nulleffekt annimmt, wie dies bei klassischen Hypothesentests üblich ist:

> There is a null hypothesis asserting that the difference between two groups is at least as large as the one specified by the investigator, and there is an alternative hypothesis asserting that the difference between two groups is smaller than the specified one. As in traditional hypothesis test, the goal of the investigator is to reject the null hypothesis and accept the alternative hypothesis. (Rogers et al., 1993, S. 554)

Die Logik, die sich hinter diesem Vorgehen verbirgt, bringen Altman und Bland (1995) mit dem Titel eines Kurzaufsatzes bestechend auf den Punkt: „*Absence of evidence is not evidence of absence*“ (S. 485). In der Regel werden mittels klassischer Hypothesentests stets Unterschiede oder Zusammenhänge überprüft, indem die Alternativhypothese etwa einen spezifizierten Unterschied annimmt und gegen die Nullhypothese geprüft wird, die den Inhalt der Alternativhypothese lediglich negiert. Die Wunschhypothese ist hierbei die Alternativhypo-

these. Will man hingegen überprüfen, ob sich zwei Treatments in ihrer Wirksamkeit tatsächlich gleichen, wäre – in klassischen Hypothesentests gedacht – die Nullhypothese die Wunschhypothese. Es existieren unterschiedliche Herangehensweisen, wie mit dieser Besonderheit umgegangen werden sollte (vgl. Klemmert, 2004). Durchgesetzt hat sich v.a. in der Biostatistik die wohl robusteste Methode – der Äquivalenztest –, demzufolge die Alternativhypothese weiterhin die Wunschhypothese bleibt und die Annahme enthält, die es tatsächlich zu überprüfen gilt: Die Wirksamkeits*äquivalenz*. Würde man zur Überprüfung von Äquivalenzhypothesen traditionelle Hypothesentests anwenden und zu einem nicht-signifikanten Ergebnis gelangen, so wäre dieses Ergebnis schwierig interpretierbar. Eben diese Schwierigkeit bringt der Satz *Absence of evidence is not evidence of absence* zum Ausdruck: Inferenzlogisch ist ein nicht-signifikantes Ergebnis bei einem klassischen Unterschiedshypothesentest nicht gleichzusetzen mit der Richtigkeit des Nullhypothesenpostulats – dem Fehlen von Unterschiedlichkeit (*evidence of absence*). So schreibt Bortz (2005):

> Korrekt wäre es, wenn man nach diesem Ergebnis sagen würde, dass die H_0 mit der durchgeführten Untersuchung (vor allem bezogen auf die untersuchte Stichprobengröße; vgl. S. 125ff.) nicht verworfen werden konnte und dass im Übrigen über die Richtigkeit von H_0 und H_1 keine Aussage gemacht werden kann. (S. 118)

Denn es könnte durchaus sein: „dass der gleiche Unterschied zwischen den verglichenen Methoden in einer anderen Untersuchung mit einer größeren Stichprobe zu einem signifikanten Ergebnis führt“ (Bortz, 2005, S. 118).

Warum wählen Kendall et al. (1999) nun eine Kombination aus klassischem und Äquivalenzhypothesentest? Die Antwort liegt darin, dass klassische Nullhypothesentests nicht 1:1 in Äquivalenztests zu überführen sind. Das bedeutet, dass Äquivalenztests nicht einfach nur die

Umkehrung von klassischen Nullhypothesentests darstellen (vgl. Klemmert, 2004), denn im Unterschied zu klassischen Nullhypothesentests werden bei Äquivalenztests *Intervalle der Irrelevanz* angelegt, was bei klassischen Hypothesentests nicht der Fall ist. Daraus erwachsen unterschiedliche Aussagegehalte, deren Kombination sich Kendall et al. (1999) in ihrer Bestimmung der klinischen Bedeutsamkeit zunutze machen. Verglichen wird nach Kendalls Methode eine zu überprüfende Intervention – in der Regel der Mittelwert der Postmessungen – mit normativen Parametern, wie dem Mittelwert einer gesunden Normalpopulation. Der Vergleich erfolgt zum einen über den klassischen Hypothesentest und zum anderen über einen Äquivalenztest. Bei Letzterem ist es notwendig, ein Intervall festzulegen, innerhalb dessen eine Differenz zuungunsten der klinischen Gruppe als vernachlässigbar zu erachten ist: „The specification of the deltas [Irrelevanzbereiche] must be approached carefully because the decision affects the validity of the results of clinical equivalency testing. How close must a post-treatment mean be to a normative mean to be considered clinically equivalent?" (Kendall et al., 1999, S. 287).

Mögliche Resultate der kombinierten Anwendung beider Überprüfungsarten sind Tabelle 3 zu entnehmen.

Tabelle 3: Normativer Vergleich nach Kendall, Marrs-Garcia, Nath & Sheldrick (1999)

		Traditional Statistical Test: Significant Effect	Traditional Statistical Test: Nonsignificant Effekt
Clinical Equivalency Test	Significant	Cell I Statistically different, clinically equivalent	Cell II Clinically equivalent
	Not Significant	Cell III *Different* (not clinically equivalent)	Cell IV *Equivocal findings* (more power required)

aus: Kendall, Marrs-Garcia, Nath & Sheldrick, 1999, S. 288.

Um von tatsächlicher Gleichheit der Mittelwerte ausgehen zu können, sollte die doppelte Überprüfung zu einem Ergebnis gelangen, das Zelle II in Tabelle 3 zu entnehmen ist: Hierbei wurde die Nullhypothese des Äquivalenztests durch ein signifikantes Ergebnis zurückgewiesen, gleichzeitig führte der klassische Hypothesentest zu keinem signifikanten Ergebnis, so dass zwei Indikatoren die annähernde Äquivalenz der beiden Mittelwerte belegen.

Es leuchtet unmittelbar ein, dass es sich durch den kombinatorischen Einsatz klassischer und alternativer Hypothesentestmethoden um ein Vorgehen handelt, das zu vergleichsweise robusten Schlussfolgerungen über die klinische Bedeutsamkeit einer Intervention führen kann (vgl. Kendall et al., 2004). Eine Studie, die sich dieser Methode bedient, wird daher in der Kodierung von A.9. (Abbildung 19, S. 171) mit Stufe „1“ bewertet.

Die "soziale Validierung" bzw. "subjektive Evaluation" als Methode zur Feststellung der klinischen Bedeutsamkeit eines Effekts, beschreitet einen gänzlich anderen Weg und stellt die subjektive Bewertung des Behandlungserfolgs durch den Patienten selbst oder aber durch dessen Angehörige/Freunde in den Mittelpunkt (vgl. Kazdin, 1994; Lambert & Ogles, 2004).

Vor allem die Bewertung des Behandlungserfolgs durch Dritte begründet Kazdin (1994) folgendermaßen:

> The view of others are relevant because people in everyday life often have a critical role in identifying, defining, and responding to persons they may regard as dysfunctional or deviant . . . . (S. 56)

Und weiter:

> The opinions of others in contact with the client are important as a criterion in their own right because they often serve as a basis for seeking treatment in the first place and also reflect the evaluations the client will encounter after leaving treatment. (S. 56)

Auch, wenn Kazdin begründet auf die limitierte Aussagekraft sozialer Validierungen hinweist, da Urteile von Personen aus dem sozialen Umfeld des Patienten unterschiedlichen Verzerrungen unterliegen können, so wird dieser Form der klinischen Bedeutsamkeitsbestimmung trotzdem ein zentraler Wert beigemessen. Eine Studie, die in ihren Ergebnisdarstellungen auf subjektive Evaluationen setzt, wird aus diesem Grund mit einer „1“ bewertet. Gleiches trifft auf Studien zu, in denen sich eines weiteren klinischen Bedeutsamkeitsmaßes bedient wird, nämlich der Überprüfung des Fortbestehens der Anfangsdiagnose (Kazdin, 2008). Würde ein Patient zum Postzeitpunkt nicht mehr die Kriterien der Anfangsdiagnose erfüllen, so wäre dies als ein klinisch relevanter Effekt zu betrachten.

Die Spezifizierungen der „2“er-Bewertung setzen sich zum größten Teil aus denselben Aspekten der „1“er-Bewertung zusammen, mit dem Unterschied, dass die genannten Methoden der klinischen Bedeutsamkeitsbestimmung bei einer „2“ nicht vollständig angewandt werden (z.B. nur Anwendung des RCI oder eines normorientierten Cutoff-Werts).

Werden in einer Publikation überhaupt keine Informationen über die klinische Bedeutsamkeit der empirischen Wirksamkeitsbefunde bereitgestellt, so schneidet die Studie allenfalls mit einer „3“ ab.

Dropoutquote/Ausschöpfungsquote und Katamnesezeitraum

Der gemeinsame Nenner der drei im Folgenden zu beschreibenden Kriterien sowie der entwickelten Kodierregeln besteht darin, dass alle drei Kriterien die in den Studien berichteten Dropout- bzw. Ausschöpfungsquoten bewerten. Zwei dieser drei Kriterien beziehen sich allein auf die Abbrecherraten und bewerten gewissermaßen die Höhe der in einer Studie referierten Abbrecherraten. Das dritte Kriterium bezieht sich zudem auf ein weiteres Gütemerkmal einer Studie, nämlich auf den Katamnesezeitraum. Allen Kriterien gemeinsam ist es, dass zwecks Anwendung der Kriterien Referenzwerte eruiert werden müssen, durch die eine eindeutige Bewertung von Studien erst möglich wird.

Die zuerst darzustellenden Kodierregeln beziehen sich auf die beiden Kriterien, die sich allein auf die Abbrecherraten beziehen. Das sind zum einen das methodische Qualitätskriterium A.3., mit dem die Höhe der Abbrecherrate zwischen der Prä- und der Postmessung in einer Studie bewertet wird (Abbildung 20). Zum anderen ist dies Kriterium A.4., mit dem die Höhe der Abbrecherquote zwischen der Post- und der Katamnesemessung bewertet wird (Abbildung 21, S. 182).

A.3.	Höhe der Drop-out-Quote zu Behandlungsende (sofern nicht Erfolgskriterium)	i.d.R. Drop-out-Quote kleiner 20 %	(1)	Drop-outs werden hier als Studien- *und/oder* Therapie-Drop-outs definiert: Patienten, die die Studie abbrechen (und ggf. die Therapie weitermachen) sowie Patienten, die die Therapie abbrechen (und ggf. für Studienzwecke weiter zur Verfügung stehen) werden als Drop-outs gezählt **Achtung:** Drop-outs über alle Treatmentarme (EG, KG) berechnen Berechnung nach Flow Diagramm: $1 - \frac{Interventions/Studiencompleter}{Zugewiesen\ zur\ Intervention}$ Wenn keine Drop-outs berichtet, ohne, dass aus Publikation hervorgeht, dass es tatsächlich keine gab → 3
		i.d.R. Drop-out-Quote zwischen 20 % und 40 %	(2)	
		i.d.R. Drop-out-Quote größer 40 %	(3)	
	Vor allem in Bezug auf Studien ohne Randomisierung: Zeitpunkt, ab dem Drop-outs gezählt werden (Erstgespräch, probatorische Sitzungen, Therapiebeginn o.ä.):			-

Anmerkung: EG: Experimentalgruppe, KG: Kontrollgruppe.

Wenn bei Prä-Katamnesestudien oder reinen Katamnesestudien Post-Drop-outs und Katamnese-Drop-outs nicht separat zu eruieren sind, dann bei **A.3** alle Drop-outs zwischen Prä und Katamnese kodieren mit folgenden Richtwerten:

Affektiv < 2 Jahre ≥ 2 Jahre	 ≤ 28% ≤ 32 %	**1**
Affektiv < 2 Jahre ≥ 2 Jahre	 29% - 52% 33% - 58%	**2**
Affektiv < 2 Jahre ≥ 2 Jahre	 > 52% > 58%	**3**

Gemischt < 2 Jahre ≥ 2 Jahre	 ≤ 36% ≤ 40%	**1**
Gemischt < 2 Jahre ≥ 2 Jahre	 37% - 64% 41% - 70%	**2**
Gemischt < 2 Jahre ≥ 2 Jahre	 > 64% > 70%	**3**

Abbildung 20: Kriterium A.3.: Höhe der Drop-outquoten zu Behandlungsende (allgemeine methodische Qualität)

Die Dropoutquoten der einzelnen Bewertungsstufen des Kriteriums A.3. sind bereits mit prozentualen Intervallen spezifiziert, die festlegen, wie hoch die jeweiligen Dropoutraten zwischen der Prä- und der Postmessung maximal sein dürfen. Damit sind nicht nur die Dropoutraten in der Experimentalgruppe gemeint, sondern die Anzahl an Abbrechern aus allen Treatmentarmen zusammengenommen. Unter Dropouts sind sowohl Studien- als auch Therapieabbrecher zu subsumieren, die eine gute Studie – soweit möglich inklusive der Angabe der Abbruchgründe – segregiert aufführen sollte (vgl. Hiller et al., 2009). Von Kriterium A.3. wird dieses Gütemerkmal der Erhebung und Angabe von Abbruch*gründen* jedoch nicht tangiert, vielmehr bezieht sich dieses Kriterium ausschließlich auf die Höhe der Schwundquote: Als Kriterium der allgemeinen methodischen Qualität wird damit die Robustheit der Ergebnisse bewertet. So liegt es nahe, dass eine extrem hohe Dropoutrate den Aussagegehalt der Ergebnisse grundlegend in Frage stellt (vgl. Kendall et al., 2004).

In Anlehnung an den in der Literatur zu ITT-Analysen oft zu lesenden Satz „Once randomised, always analysed“ (Kleist, 2009, S. 450) wurde sich für die Bestimmung der Dropoutraten für Kriterium A.3. auf den Zeitraum zwischen Randomisierung und Postmessung konzentriert. Da bei Studien ohne Randomisierung kein eindeutiges Pendant zum Zeitpunkt der Randomisierung existiert (vgl. Hiller et al., 2009), wurde in der Kodierregel kein Zeitpunkt festgelegt, sondern pragmatisch diejenige Dropoutquote (Prä-Post) bewertet, die in der Publikation angegeben wird und – soweit der Publikation zu entnehmen – der Zeitpunkt notiert, ab dem die Abbruchrate in der Studie erhoben wurde (vgl. Abbildung 20).

Da bei Studien, die über keine Postmessung, dafür jedoch über eine Katamnesemessung verfügen[46], nicht derselbe Bewertungsstandard für Dropoutraten angelegt werden kann,

[46] Darunter fallen in der Regel Studien, die nicht über eine „echte“ Postmessung verfügen (damit sind individuelle Postmessungen zum Ende der Behandlung eines jeden einzelnen Patienten gemeint), sondern in denen als „Postmessungen“ deklarierte Erhebungen zu einem fixen Messzeitpunkt irgendwann nach Beendigung der The-

wie für Untersuchungen mit regulären Post-Messungen, wurden für diese Studien angemessenere Bewertungsstandards eruiert. Da das Erstellen dieser Standards jedoch zunächst die Spezifizierung des Kriteriums A.4. voraussetzte, soll an dieser Stelle zunächst auf die Kodierregeln dieses Kriteriums eingegangen werden, bevor schlussendlich auf A.3. zurückgekommen wird.

Im Gegensatz zu A.3. (Abbildung 20, S. 179), bei dem prozentuale Intervallgrenzen angegeben werden, erfordert die Kodierung von A.4. (Abbildung 21) zunächst das Aufstellen eines empirisch fundierten Referenzrahmens für Abbruchraten zwischen Post- und Katamnesemessungen, mit dem die Abbruchrate der Studie wiederum verglichen werden kann.

rapie realisiert werden. Hierbei muss im Grunde von Katamnesemessungen gesprochen werden, wobei die Zeiträume zwischen Beendigung der Therapie und Katamnesemessung individuell variieren (vgl. von Hauenschild, 2011).

<table>
<tr><td rowspan="11">A.4.</td><td rowspan="11">Höhe der Studien-Drop-outs zur Katamnese (falls Katamneseerhebung durchgeführt) (zwischen Post- und Katamnesezeitpunkt)

• Wenn keine Katamnese: „9“
• Prä-Katamnese: „4“
• Nur Katamnese: „5“</td><td colspan="2">deutlich besser als in Studien mit vergleichbaren Patientengruppen und vergleichbarem Katamnesezeitraum</td><td rowspan="3">(1)</td><td rowspan="11">Achtung: Drop-outs über <u>alle</u> Treatmentarme berechnen

Berechnung nach Flow Diagramm:

$1 - \frac{Katamnesedaten\ vorliegend\ von\ Interventions-/Studiencompletern\ zur\ Postmessung}{Interventions-/Studiencompleter\ zur\ Postmessung}$

Wenn keine Drop-outs berichtet, ohne, dass aus Publikation hervorgeht, dass es tatsächlich keine gab → 3</td></tr>
<tr><td>Affektive Störungen</td><td>Gemischte Störungsgruppen</td></tr>
<tr><td>weniger als 2 Jahre: ≤ 10%
2 Jahre oder mehr: ≤ 15%</td><td>weniger als 2 Jahre: ≤ 20%
2 Jahre oder mehr: ≤ 25%</td></tr>
<tr><td colspan="2">Drop-out-Quote vergleichbar mit Studien mit entsprechenden Patien-tengruppen und entsprechendem Katamnesezeitraum</td><td rowspan="3">(2)</td></tr>
<tr><td>Affektive Störungen</td><td>Gemischte Störungsgruppen</td></tr>
<tr><td>weniger als 2 Jahre: 11-20%
2 Jahre oder mehr: 16-30%</td><td>weniger als 2 Jahre: 21-40%
2 Jahre oder mehr: 26-50%</td></tr>
<tr><td colspan="2">deutlich schlechter als in Studien mit vergleichbaren Patientengruppen und vergleichbarem Katamnesezeitraum</td><td rowspan="3">(3)</td></tr>
<tr><td>Affektive Störungen</td><td>Gemischte Störungsgruppen</td></tr>
<tr><td>weniger als 2 Jahre: > 20%
2 Jahre oder mehr: > 30%</td><td>weniger als 2 Jahre: > 40%
2 Jahre oder mehr: > 50%</td></tr>
</table>

Abbildung 21: Kriterium A.4.: Höhe der Dropoutquoten zur Katamnesemessung (allgemeine methodische Qualität)

Da davon auszugehen ist, dass mit der Länge des Katamnesezeitraums auch die Anzahl an Abbrüchen variiert, wurde der Referenzrahmen in Anlehnung an den im Rahmen eines anderen Kriteriums bestimmten störungsangemessenen Katamnesezeitraum (s.u.) separat für unterschiedliche Zeiträume eruiert: Einmal für zu erwartende Dropoutquoten für einen Katamnesezeitraum von weniger als 2 Jahren und einmal für einen Zeitraum von mindestens 2 Jahren (vgl. Abbildung 21).

Um den Referenzrahmen zu schaffen, wurden Outcomestudien aus unterschiedlichen psychotherapeutischen Richtungen herangezogen, die über eine möglichst große Streuung an Katamnesezeiträumen verfügen. Neben der o.g. Separation (Katmnesezeitraum < 2 Jahre und Katamnesezeitraum ≥ 2 Jahre) wurde zudem hinsichtlich des Anwendungsbereichs segregiert: Für die zu kodierenden Studien zu affektiven Störungen wurden entsprechend Referenzkatamnesestudien mit Patienten derselben Störungsgruppe herangezogen; gleiches gilt für die Studien mit diagnoseheterogenen Störungsgruppen.

Bei den herangezogenen Referenzstudien mit Katamnesezeiträumen < 2 Jahren bei affektiven Störungsgruppen handelt es sich um Outcomestudien mit unterschiedlichen Vergleichsgruppen (kognitiv-behaviorale Therapie/Verhaltenstherapie, psychodynamische Psychotherapie, Entspannungsverfahren, Kombinationstherapien, *clinical management*, reine Pharmakobehandlung, Warteliste), die zusammengenommen über einen Range an Katamnesezeiträumen zwischen 3 Monaten und 2 Jahren verfügen (vgl. de Jong-Meyer, Hautzinger, Rudolf, Strauß & Frick, 1996; Gallagher-Thompson, Hanley-Peterson & Thompson, 1990; Hautzinger & Welz, 2004; McLean & Hakstian, 1979,1990; Paykel et al., 1999; Thompson, Gallagher & Breckenridge, 1987). Über diese Studien konnte ein Range der Abbruchraten (Post-Katamnese) von 5-19% eruiert werden. Da eine zu kodierende Studie laut Kriterium A.4. mit einer „2“ bewertet wird, wenn sie über *vergleichbare* Dropoutraten verfügt, wie Referenzstudien, wurde das obere Intervall der „2“er-Bewertung mit aufgerundeten 20% festgelegt.

Selbiges Vorgehen wurde für den Katamnesezeitraum ≥ 2 Jahren (affektive Störungsgruppen) gewählt: Hier zeigt sich bei einer herangezogenen Studie (McLean & Hakstian, 1979,1990) und einem Katamnesezeitraum von 27 Monaten eine Dropoutrate von 28% (Post-Katamnese), so dass die obere Intervallgrenze für eine „2“er-Bewertung auf gerundete 30% festgesetzt wurde.

Für gemischte Störungsgruppen und einen Katamnesezeitraum < 2 Jahre wurden Outcomestudien herangezogen, in denen sowohl Einzel- als auch Gruppentherapien untersucht wurden. Folgende Verfahren und Kontrollbedingungen waren Gegenstand dieser Untersuchungen: Gesprächspsychotherapie, Verhaltenstherapie, psychodynamische Psychotherapie, interaktionelle Verhaltenstherapie, systemische Therapie, *hospital treatment* sowie Wartelisten. Der Range der unterschiedlichen Katamnesezeiträume beläuft sich auf 6 bis 23 Monate, der Range der Dropoutquoten auf 0-38% (vgl. Eckert & Biermann-Ratjen, 1985; Grawe, Caspar & Ambühl, 1990; Langsley, Flomenhaft & Machotka, 1969; Langsley, Machotka & Flomenhaft, 1971; Paivio & Nieuwenhuis, 2001). Die „2"er-Bewertung (gemischte Störungen, Katamnese < 2 Jahre) wurde daher mit gerundeten 40% (obere Intervallgrenze) festgelegt.

Für gemischte Störungen und einem Katamnesezeitraum ≥ 2 Jahren wurden Outcomestudien herangezogen, in denen ebenfalls Einzel- und Gruppensettings realisiert wurden (psychodynamische Psychotherapie, Psychoanalyse)[47]. Der Range der unterschiedlichen Katamnesezeiträume beläuft sich auf dreieinhalb bis 10 Jahre, der Range der Dropoutquoten auf 13-50% (vgl. Kantrowitz, Katz & Paolitto, 1990a/b/c; Liberman et al., 1972; von Rad, Senf & Bräutigam, 1998). Die „2"er-Bewertung wurde daher mit einer oberen Intervallgrenze von 50% festgelegt (vgl. Abbildung 21, S. 182).

Ausgehend von der Mittelkategorie („2") wurde die „1"er-Bewertung des Kriteriums A.4. pragmatisch an die Festlegungen des Kriteriums A.3. (Abbildung 20, S. 179) angelehnt und als maximale Intervallgrenze jeweils die Hälfte der Prozentangabe der Mittelkategorie „2" gewählt. Der Bewertungsmaßstab der „3"er-Kategorie leitet sich ebenfalls aus der Mittelkate-

[47] In einer Studie (Liberman et al., 1972) wurde als Behandlungsform lediglich „individual psychotherapy" (S. 36) angegeben, so dass eine Verfahrenseinordnung nicht möglich war.

gorie ab: Studien, die die obere prozentuale Intervallgrenze der Mittelkategorie („2“) überschreiten, werden mit „3“ bewertet.

Es wurde oben darauf hingewiesen, dass sich aus den Kodierregeln zu Kriterium A.4. nun die prozentualen Referenzen für die Bewertung von Studien mit Prä-Katamnesemessungen (ohne spezifizierte Postmessung) bzw. von Studien mit reinen Katamnesemessungen ableiten lassen, die für Kriterium A.3. relevant sind (vgl. Abbildung 20, S. 179). Die prozentualen Intervallgrenzen wurden auch hier wieder sowohl nach Störungsgruppen (affektiv *versus* gemischt) als auch nach Katamnesezeiträumen (< 2 Jahre *versus* ≥ 2 Jahre) getrennt kalkuliert. Dafür wurden die eigens eruierten Intervallgrenzen der einzelnen Bewertungskategorien aus Kriterium A.4. mit denen aus Kriterium A.3. kombiniert und verrechnet (vgl. Abbildung 20, S. 179, unterer Teil).

Um bei Studien mit Prä-Katamnesemesszeitpunkten diese Kodierungen überhaupt vornehmen zu können, musste der Katamnesezeitraum zunächst abgeschätzt werden. Dazu wurde dann meist die mittlere Therapiedauer vom Gesamtzeitraum (Prä-Katamnese) abgezogen, um so zu einer Einschätzung des mittleren Katamnesezeitraums zu gelangen.

Da unter den zuletzt geschilderten Umständen von Prä-Katamnesemessungen lediglich Kriterium A.3. bewertet wird, wurden für Kriterium A.4. Zusatzkodierregeln (Missingwerte) entwickelt (vgl. Abbildung 21, S. 182, linkes Textfeld). Diese Zusatzkodierregeln haben jedoch ausschließlich informativen Charakter, denn für die Berechnung des Dimensionsmittelwerts der allgemeinen methodischen Qualität (vgl. Kap. 1.2.2) wurde pragmatisch entschieden, Kriterium A.4. mit demselben Wert wie Kriterium A.3. in die Berechnung eingehen zu lassen.

Bei dem dritten Kriterium, für das ebenfalls ein Referenzrahmen zur Bewertung von Studien entwickelt werden musste, handelt es sich um das interne Validitätskriterium B.11. (vgl. Abbildung 22). Dieses Kriterium fragt neben den Ausschöpfungsquoten (dem Antagonisten zur Abbruchrate) zwischen der Prä- und der Katamnesemessung danach, ob in den zu kodierenden Studien "störungsangemessene Katamnesezeiträume" realisiert wurden.

B.11.	Follow-up-Messung	zeitlich störungsangemessene Katamnese mit hoher Ausschöpfung[48] der Stichprobe • mindestens 2 Jahre (2 Jahre oder mehr) • mindestens 60%ige Ausschöpfungsquote	(1)	Es müssen beide Bedingungen erfüllt sein (Ausschöpfungsquote und Katamnesezeitraum) Die Ausschöpfungsquote bezieht sich auf alle Untersuchungsgruppen
		Katamnese mit fraglich angemessenem Zeitraum bzw. niedriger Ausschöpfung der Stichprobe • weniger als 2 Jahre • weniger als 60% Ausschöpfungsquote	(2)	Sobald eine der beiden Bedingungen nicht erfüllt ist → (2)
		keine Katamnese	(3)	Keine Katamnese

Abbildung 22: Kriterium B.11.: Follow-up-Messung (allgemeine methodische Qualität)

Die Festlegung der Ausschöpfungsquoten (Prä-Katamnese) ließ sich vergleichsweise einfach eruieren: Hierbei wurden die prozentualen Intervallgrenzen herangezogen, die für Studien mit Prä-Katamnesemessungen aufgestellt wurden. Die Orientierung erfolgte an derjenigen Prozentzahl, die für Studien mit gemischten Störungsgruppen und Katamnesen ≥ 2 Jahre für eine „1"er-Bewertung festgesetzt wurde (40%; vgl. Abbildung 20, S. 179, unterer Teil). Die geforderte Mindestausschöpfung bei Kriterium B.11. für eine Bewertung mit „1" wurde demnach

[48] Gemeint ist die Ausschöpfung der Stichprobe zwischen Prä- und Katamnesemessung.

mit 60% festgelegt. Allerdings *muss* für diese „1“er-Bewertung die Bedingung „Katamnesezeitraum ≥ 2 Jahre“ ebenfalls erfüllt sein.

Liegt die Ausschöpfungsquote unter 60% *oder* ist der Katamnesezeitraum kürzer als 2 Jahre, so wird die Studie mit „2“ bewertet.

Wie bereits erwähnt, musste der "störungsangemessene Katamnesezeitraum" zunächst in Anlehnung an den natürlichen Störungsverlauf ermittelt werden. Diese Herangehensweise, Katamnesezeiträume *störungsangemesse*n festzulegen, stellt jedoch lediglich *eine* Vorgehensweise dar. So zählt Frohburg (2004) in ihrem Überblicksartikel zu Langzeiteffekten der Gesprächspsychotherapie weitere Methoden zur Festlegung von Katamnesezeiträumen auf:

> Über die sinnvolle Länge von katamnestischen Bobachtungszeiträumen gibt es sehr unterschiedliche und kaum sachlich begründete Meinungen. So werden mindestens vier Jahre (Herzog & Deter, 1994, S. 122) oder fünf Jahre für sinnvoll gehalten (Rüger & Senf, 1994, S. 107). Oder es wird als Mindestforderung angegeben, dass der Katamnesezeitraum nicht kürzer sein sollte als der Therapiezeitraum (Rüger & Senf, 1994, S. 107). (S. 204)

Ebenso fordert Mertens (2007) in seiner Stellungnahme zum Methodenpapier[49] die Anpassung des Katamnesezeitraums an die Länge der Behandlung.

Diesen Forderungen stellen einschlägige Autoren eine Methode gegenüber, die sich mittlerweile durchzusetzen scheint (u.a. Hautzinger, 2007; Roth & Fonagy, 2005). So schreiben Boland und Keller (2009): „. . . one had to understand the natural course of the illness before investigating the effect of interventions on that course” (S. 25). Und Westen et al.

[49] Dem Veröffentlichungsdatum der Stellungnahme nach zu urteilen, müsste sie sich auf Version 2.6 der Verfahrensregeln beziehen. Da Kriterium B.11. über die unterschiedlichen Versionen hinweg nicht revidiert wurde, ist die Version der Verfahrensregeln, auf die sich Mertens’ Stellungnahme bezieht, unerheblich.

(2004) kritisieren: „A striking gap in the literature is the relative absence of follow-up studies that span the length of time during which relapse is known to be common in untreated patients for the disorder in question" (S. 650).

Der aus beiden Zitaten hervorgehenden Forderung, Katamnesezeiträume *störungsspezifisch* auszurichten, folgt der WBP in B.11.. Damit verpflichtet er sich, vor der Begutachtung einer therapeutischen Methode oder eines therapeutischen Verfahrens für unterschiedliche Störungen adäquate Katamneseintervalle festzulegen, die als Referenz für die zu begutachtenden Studien gelten. Die Festlegung erfolgt in Anlehnung an diejenigen Zeiträume nach erlangter, symptomatischer Remission oder Genesung, innerhalb derer normalerweise bzw. im natürlichen Störungsverlauf mit Rückfallraten zu rechnen ist. Eben diesen Zeitraum gilt es – so die Logik der störungsangemessenen Katamnesen – in einer Outcomestudie „einzufangen", um eine zuverlässige Aussage über den langfristigen Einfluss einer Behandlung auf den natürlichen Verlauf einer Störung machen zu können.

Für die Spezifizierung von B.11. (Abbildung 22, S. 186) bedeutete dies zunächst eine Sichtung der Literatur zum natürlichen Verlauf von affektiven Störungen. Der zentrale Parameter dieser Sichtung waren kumulative Rückfallwahrscheinlichkeiten im Zeitraum nach erreichter Remission oder Genesung (vgl. Laux, 2008a). Dabei wurden unterschiedliche Störungsbilder aus dem affektiven Formenkreis berücksichtigt (unipolar/bipolar, episodisch/chronisch, *double depression*). In einigen Studien zum Verlauf affektiver Störungen nach Remission/Genesung blieb es nicht aus, dass sog. pharmakologische Aufrechterhaltungsbehandlungen durchgeführt wurden (z.B. Gorwood, Weiller, Lemming & Katona, 2007). Auch solche Studien wurden vereinzelt berücksichtigt, obwohl sie strenggenommen nicht den rein *natürlichen*

Störungsverlauf skizzieren, der aus ethischen Gründen nicht immer einwandfrei zu eruieren ist.

In Abhängigkeit von unterschiedlichen prognostisch bedeutsamen Faktoren, wie

- der Häufigkeit vorheriger Episoden,
- dem Chronifizierungsgrad,
- dem Vorliegen einer *double depression*,
- persistierender Residualsymptomatiken nach Remission,
- bipolarer *versus* unipolarer Verlauf der affektiven Störung sowie
- dem Vorliegen komorbider Störungen,

aber auch in Abhängigkeit von den jeweiligen untersuchungsspezifischen Operationalisierungen von Remission, Genesung und Rückfall bzw. Wiedererkrankung, unterliegen die Rezidivraten bzw. die geschätzten kumulativen Rückfallwahrscheinlichkeiten starken Schwankungen. So zeigt sich in Untersuchungsgruppen mit unipolaren affektiven Störungen, in denen bspw. volle Remissionen ohne Residualsymptomatiken erreicht wurden und/oder bei denen keine oder nur wenige Episoden der Indexepisode vorausgingen, eine vergleichsweise niedrige Rückfallrate, die sich über einen langen Zeitraum nach erlangter Remission bzw. Genesung erstreckt. So berichten unterschiedliche Autoren über eine ca. 50%ige Rückfall- oder Wiedererkrankungsrate zwischen 3.5 und 7 Jahren nach Remission bzw. Genesung (vgl. Boland & Keller, 2009; Judd et al., 1998; Judd et al., 2000; Klein, Schwartz, Rose & Leader, 2000). In den meisten Fällen liegt die mediane Rückfallrate jedoch um einen Zeitraum von ca. 1 bis 2 Jahren, wobei das 2-Jahresintervall die eindeutig sicherere Basis für das „Einfangen“ einer 50%igen Rückfallwahrscheinlichkeit bildet (vgl. Boland & Keller, 2009; DGPPN, 2009; Faravelli, Ambonetti, Pallanti & Pazzagli, 1986; Gorwood et al., 2007; Judd et al., 1998; Judd et al., 2000; Judd et al., 2008; Keller, 1999; Keller, Lavori, Lewis & Klerman, 1983; Keller,

Shapiro, Lavori & Wolfe, 1982a/b; Klein et al., 2000; Laux, 2008a/b; Nierenberg et al., 2010).

Die daraus resultierende Festlegung, von Outcomestudien zu affektiven Störungen Katamnesezeiträume von mindestens 2 Jahren zu fordern, stimmt zudem mit der Einschätzung von Roth und Fonagy (2005) in Bezug auf depressive Störungen überein:

> The risk – indeed, the probability – of relapse has obvious implications for treatment trials. The effectiveness of a treatment needs to be judged not only by its capacity to manage an index episode but also by its ability to maintain remission. This poses a challenge, in part, because on the basis of figures given above, long-term follow-up of at least 2 years would be necessary to provide a conclusive result that is not confounded with the natural history of this disorder. (S. 70f.)

Selbiges postuliert Laux (2008b) über die Rückfallwahrscheinlichkeit bipolarer Störungen.

Da die Festlegung eines störungsangemessenen Katamnesezeitraums bei gemischten Störungsgruppen *per definitionem* an ihre Grenzen stößt, wurde der 2-Jahreszeitraum auch auf Studien zu gemischten Störungsgruppen übertragen, in denen der Anteil an Patienten mit affektiven Störungen in den meisten Fällen nicht unbeträchtlich ist.

Bevor mit Hilfe der dargelegten Kodierregeln mit der Kodierung begonnen werden konnte, wurde zusätzlich eine Art Kurzkodierbogen entwickelt, der im folgenden Unterkapitel 3.2.4 vorgestellt werden soll.

3.2.4 Entwicklung eines Kurzkodierbogens

Der Kurzkodierbogen (Ratzek & von Hauenschild, 2011b; Anhang B und D) wurde aus zweierlei Gründen entwickelt und den Kriterien des WBP vorangestellt: Zum einen sollten mittels

dieses Erhebungsinstruments allgemeine Angaben zu den einzelnen Studien erhoben werden, zum anderen sollten methodologische Eigenschaften der Studien festgehalten werden. Zu den allgemeinen Angaben zählen der Datenzugriff (handelt es sich um eine Originalstudie, eine Reanalyse von bereits publizierten Daten oder um eine Replikation einer Untersuchung?), der untersuchte Anwendungsbereich sowie zentrale Patientenmerkmale; außerdem der in der Studie original verwendete Verfahrenstitel und die eigens vorzunehmende Verfahrenseinordung entsprechend der Nomenklatur der Psychotherapierichtlinien (Gemeinsamer Bundesausschuss, 2013). Zusätzlich zur Verfahrenseinordnung wurde das therapeutische Setting (Gruppen- *versus* Individualtherapie), die Sitzungsfrequenz und -anzahl sowie die Behandlungsdauer erhoben, die, neben weiteren Gesichtspunkten, zur Verfahrenseinordung herangezogen wurden. Die Prozedur der Verfahrenseinordnung wurde an anderer Stelle ausführlich beschrieben (Kap. 3.2.2).

Bei der Erhebung von methodologischen Eigenschaften wurde sich vor allem auf solche Aspekte konzentriert, die typischerweise zur Charakterisierung von RCTs (*efficacy studies*) in Abgrenzung zu naturalistischen Studien (*effectiveness studies*) angeführt werden (vgl. Lambert & Ogles, 2004; Leichsenring, 2004a/b; Seligman, 1995; Westen et al., 2004). Da diese Aspekte bereits im theoretischen Teil der Arbeit diskutiert wurden (vgl. Kap. 1.2.1), wird an dieser Stelle auf eine erneute Darstellung dieser Charakteristika verzichtet und stattdessen nur auf solche Punkte eingegangen, die nicht unmittelbar selbsterklärend sind. Gleiches gilt für die Deklaration zentraler Begriffe, wie „verfahrensinterne Vergleichsgruppen“ sowie „verfahrensexterne Vergleichsgruppen“, die unter der Rubrik „Gruppenzuweisung“ im Kurzkodierbogen (Anhang B und D) wiederzufinden sind. Diese Begriffe wurden bereits in Kapitel 3.2.1 eingeführt. Näher zu spezifizieren sind daher lediglich die unter der Rubrik „Kontroll-/ Vergleichsgruppendesign“ subsumierten Arten von Kontroll- bzw. Vergleichsgruppen.

Die ersten vier Vergleichsgruppenarten stellen typische Kontrollgruppen dar, mit Hilfe derer die interventionsunabhängige Wirkung (etwa durch eine Wartelistengruppe) und die interventionsgebundene Wirkung (durch Placebo, TAU, aktive Kontrollgruppe) der Experimentalbehandlung kontrolliert werden sollen (vgl. Kap. 3.1.1). Unter Placebobehandlungen werden in der Regel unspezifische Treatments verstanden, in denen lediglich Faktoren, wie die, sich in therapeutischer Behandlung zu befinden, Aufmerksamkeit, soziale Unterstützung und Empathie zu erfahren etc. realisiert werden (Kendall et al., 2004)[50]. Placebobehandlungen und aktive Vergleichs- bzw. Kontrollgruppen werden oftmals synonym verwendet (z.B. Bandelow et al., 2013), wurden jedoch im Kurzkodierbogen separat aufgeführt („Placebo-Kontrollgruppe" *versus* „aktive Kontrollgruppe"). Mit dieser Unterscheidung soll zum Ausdruck gebracht werden, dass es zwischen einem „reinen" Placebo und Psychotherapiemethoden (im Sinne des WBP) noch therapeutische Vorgehensweisen gibt, die im gezielten Einsatz von supportiven Interventionen (Techniken) bestehen und für sich in Anspruch nehmen, bei bestimmten Problemen hilfreich zu sein (z.B. Kramer, Bernstein & Phares, 2009). Solche supportiven *Interventionen* können in unterschiedlichen Therapieverfahren zum Einsatz kommen, wenn aber von supportiven *Therapien* die Rede ist, sind damit meistens Therapieverfahren gemeint, die als Varianten der Gesprächspsychotherapie oder neueren psychodynamischen Psychotherapiemethoden betrachtet werden. Rössler (2004) verwendet das Begriffspaar „supportive *Psychotherapie*" und meint damit eine „vergleichsweise wenig ausformulierte therapeutische Methode", die „überwiegend alltagspraktischen Regeln helfenden Handelns" folgt (S. 134) und die in unterschiedlichen Psychotherapieformen vor allem in der Behandlung chronisch

[50] Das Konzept von Placebobehandlungen in der psychologischen Forschung – im Vergleich zur medizinischen Forschung – wird u.a. von Lambert und Ogles (2004) kritisch diskutiert.

psychisch gestörter Patienten Anwendung findet. Entsprechend der hier verwendeten Nomenklatur wäre dies eher den supportiven *Interventionen* zuzurechnen.

Die „TAU-Kontrollgruppe“ stellt eine weitere Kategorie dar, die Kendall et al. (2004) folgendermaßen definieren: „The use of a standard treatment (treatment-as-usual) as a control condition involves comparing new treatments with the intervention that is currently being applied for treatment of the problems and clients involved in the therapy being evaluated” (S. 21). Diese Definition legt es nahe, dass bspw. zur Evaluation einer neuartigen psychodynamischen Methode (etwa zur Behandlung von Depressionen) als TAU-Kontrollgruppe jede Behandlung in Frage käme, die bislang zur Depressionsbehandlung eingesetzt wird – dies können niederfrequent stattfindende Gespräche im Rahmen der psychiatrischen oder Hausarztversorgung sein oder aber Standardtherapie, wie psychodynamische Psychotherapie 1-3 Sitzungen/Woche oder kognitiv-behaviorale Therapie (vgl. auch Bandelow et al., 2013). Dem steht eine etwas andere Festlegung gegenüber, die primär hausärztliche Behandlung oder *clinical management* als TAU betrachtet, jedoch keine Standardtherapie (T. Harfst, persönl. Mitteilung, 27.05.2010). Wie breit TAU-Definitionen ausgelegt werden, geht aus einer randomisiert-kontrollierten, multizentrischen Studie von Tyrer et al. (2003) hervor, in der kognitiv-behaviorale Therapie mit TAU verglichen wird und TAU folgende Treatments umfasst: „. . . this varied from problem solving approaches (Nottingham), dynamic psychotherapy (South London), GP or voluntary group referral (West London and Edinburgh) or short-term counselling (Glasgow)” (S. 970).

Es kann also gezeigt werden, dass eine allgemein gültige Definition dessen, was TAU im psychotherapeutischen Kontext genau bedeutet, zum aktuellen Zeitpunkt nicht existiert. Beruft man sich auf die erste, sehr viel breitere Definition nach Kendall und Kollegen (2004), dann fungiert TAU als keine konstante Behandlungsform, sondern kann immer nur als Standardbe-

handlung zu einem gegebenen Zeitpunkt betrachtet werden (vgl. Löfholm, Brännström, Olsson & Hansson, 2013). Zudem verursacht die willkürliche Handhabung des Begriffs TAU eine erhebliche Effektvariabilität – immer in Abhängigkeit davon, ob als TAU eine sog. *bona-fide*-Therapie realisiert wird oder aber eine Behandlung, die sowohl vom Patienten als auch vom Therapeuten als inadäquat und unzureichend erlebt wird (vgl. Bandelow et al., 2013). Folglich bleibt es fraglich, in welchem Umfang man eine Überlegenheit der zu evaluierenden Therapie gegenüber TAU tatsächlich erwarten kann.

Da die Frage nach der korrekten Definition von TAU im Rahmen der Arbeit nicht geklärt werden kann, wird in der Kodierung von Studien folgendermaßen vorgegangen: Werden in der Studie explizit Begriffe wie „Routine-/Standardbehandlung“ bzw. „*standard treatment*“, „*usual care*“, „*treatment-as-usual*“ o.ä. genannt, so wird diese Vergleichsbehandlung als „TAU-Kontrollgruppe” kodiert – ungeachtet dessen, ob es sich im speziellen Fall um *bona-fide*-Therapien oder niederfrequente Gespräche mit dem Hausarzt handelt.

Im Kurzkodierbogen können als Vergleichsbedingungen in Studien, neben den vier bereits eingeführten Vergleichsgruppen, zusätzlich verfahrensexterne – also nicht psychodynamische – Vergleichsgruppen, sowie verfahrensinterne, d.h. psychodynamische Vergleichsgruppen kodiert werden. Wie bereits in Kapitel 3.2.1 dargelegt, werden in der weiteren Kodierung der Studien mittels des WBP-Kriterienkatalogs Untersuchungen, die zwei oder mehrere Formen der psychoanalytisch begründeten Verfahren (plus Psychoanalyse) miteinander vergleichen, als Ein-Gruppen-Designs gehandhabt. Um jedoch die Information, welche Varianten psychodynamischer Therapie miteinander verglichen werden, mit zu erheben, wurde die Kategorie „verfahrensinterne Vergleichsbehandlung(en)“ gebildet.

Die Kategorie „verfahrensexterne Vergleichsbehandlung(en)“ unterteilt sich in eine Kategorie „verfahrensexterne und etablierte Vergleichsbehandlung(en)“ und „verfahrensex-

terne Vergleichsbehandlung(en) kein bereits etabliertes Treatment“. Als etablierte Treatments werden, wie an anderer Stelle bereits ausgeführt, alle zum Zeitpunkt der Kodierung (Mitte 2010-Anfang 2012) vom WBP als wissenschaftlich anerkannte Therapieverfahren (für das Erwachsenenalter) angesehen (Verhaltenstherapie, Gesprächspsychotherapie und systemische Therapie). Alle anderen Psychotherapieformen, wie Psychodramatherapie etc., gelten als nicht etablierte Treatments.

Eine Vergleichsgruppe, in der ausschließlich pharmakologische Behandlungen durchgeführt werden, wäre strenggenommen ebenfalls den „verfahrensexternen Vergleichsbehandlung(en)“ zuzuordnen. Da es sich jedoch um keine psychotherapeutische Behandlung handelt, wurde die Kategorie „Vergleichsgruppe mit ausschließlich psychopharmakologischer Behandlung“ gesondert aufgeführt.

Ausschließlich "echte Ein-Gruppen-Designs" (vgl. Kap. 3.2.1) werden mit der Kategorie „keine Vergleichsgruppe“ kodiert.

3.3 Der Prozess der Kodierung

An der Kodierung der Studien mittels Kurzkodierbogen und WBP-Kriterienkatalog waren neben der Verfasserin dieser Arbeit und der Projektmitarbeiterin, Dipl.-Psych. Luisa von Hauenschild, noch drei weitere Rater (Laura Diedrich, M.Sc. in Psychologie und Dipl.-Psych. Uta Czech sowie Dipl.-Psych. Matthias Mohse) beteiligt. Mit den drei zuletzt genannten Ratern wurden zunächst die Kriterien samt Kodierregeln „trocken“ durchgegangen und alle aufkommenden Fragen geklärt[51]. Vor allem die Zusatzkodierregeln und die damit zusammenhängende Logik der "Ein-Gruppen-Designs im erweiterten Sinne" in Abgrenzung zu "echten Ein-Gruppen-Designs" (vgl. Kap. 3.2.1) bedurften einer grundsätzlichen Klärung. Zusätzlich wur-

[51] Da Luisa von Hauenschild an der Entwicklung der Kodierregeln beteiligt war und wir uns in einem ständigen Austausch darüber befanden, war ein zusätzliches „Trockentraining“ mit ihr nicht mehr notwendig.

den alle Kodierregeln beleuchtet, die nicht unmittelbar selbsterklärend sind (vgl. Kap. 3.2.2 und 3.2.3). Zudem wurde die im Kurzkodierbogen eigenständig vorzunehmende Einordnung der psychodynamischen Studientherapien entsprechend der Nomenklatur der Psychotherapierichtlinien (Gemeinsamer Bundesausschuss, 2013) genau erläutert.

Im Anschluss wurden mit allen vier Ratern jeweils zwei Studien probeweise kodiert, kriterienweise besprochen und letzte Unklarkeiten ausgeräumt. Erst nach Abschluss dieses Trainings wurde mit der regulären Kodierung der 41 Studien begonnen. Hierbei wurde jede Studie jeweils von einem der Rater und der Verfasserin der Arbeit unabhängig voneinander kodiert[52]. Bei Fragen oder Unklarheiten stand die Verfasserin der Arbeit jedoch durchweg zur Verfügung. Anschließend wurden die Studie und die vorgenommenen Ratings detailliert besprochen. Pro Studie dauerte eine solche Besprechung im Schnitt 1.5 Stunden. Besonders komplexe Studien, die sich zudem über mehrere Publikationen belaufen, konnten in der Diskussion der Ratings bis zu 3 Stunden in Anspruch nehmen.

Auch, wenn die Studien jeweils von zwei Ratern kodiert wurden, stand in der Kodierungsphase weniger die Unabhängigkeit der Rater im Vordergrund, vielmehr sollte die Genauigkeit der Ratings durch den supervisorischen Austausch und die ausführlichen Besprechungen im Anschluss an die Kodierungen gesichert werden.

Wie bereits an anderer Stelle angeführt, wurde jede Studie auf *allen* Dimensionen (d.h. allgemeine methodische Qualität, interne und externe Validität) und Kriterien kodiert. Das bedeutet, dass bspw. Studien mit einem Ein-Gruppen-Design auf der internen Validitätsdimension weiterkodiert wurden, obwohl sie strenggenommen durch das Kriterium B.8. („Gruppenzuweisung“) nicht mehr auf der internen Validitätsdimension bestehen konnten. Gleiches gilt für

[52] Jeder der vier Rater kodierte zwischen 6 bis 15 Studien und wurde aus Drittmitteln für diese Tätigkeit vergütet.

Studien, die durch K.O.-Kriterien der methodischen Qualitätsdimension durchfielen. Auf diese Weise sollte sichergestellt werden, dass alle Studien auf allen Dimensionen und Kriterien tatsächlich vergleichbar sind.

Der Zeitraum der Kodierung belief sich auf ca. 1.5 Jahre und wurde Anfang 2012 abgeschlossen.

3.4 Geplante Datenanalyse

3.4.1 Beschreibung der Studien via Kurzkodierbogen

In einem allgemeinen Überblick über die 41 Primärstudien werden die Studien zunächst in ihren Verteilungen hinsichtlich der Variablen „Anwendungsbereich", „Datenzugriff", „Therapieverfahren", „Setting" sowie „Therapieumfang" beschrieben. Daran anschließend erfolgt die Sichtung der Verteilungen auf dem Variablen „Studiendesign", „Strategien der Gruppenzuweisung", „Messzeitpunkte" und „Katamnesezeitraum" (vgl. Kap. 4.1).

Im Anschluss folgt eine Klassifizierung der Studien unter Zuhilfenahme clusteranalytischer Verfahren – einer Two-Step Clusteranalyse sowie einer hierarchischen Clusteranalyse nach der *Ward*-Methode (vgl. Kap. 4.2). Grundsätzliches Ziel von Clusteranalysen ist es, auf Basis bestimmter Variablen, die in die Clusteranalyse aufgenommen werden, Fallgruppierungen (hier: Studiengruppierungen) zu schaffen, die innerhalb dieser Gruppen (Cluster) im Hinblick auf eben diese Variablen möglichst homogen sind. Gleichsam soll zwischen den Clustern maximale Unähnlichkeit bzw. Distanz im Hinblick auf dieselben Variablen bestehen.

Das Prozedere der Two-Step Clusteranalyse basiert, wie es der Name bereits nahelegt, auf zwei Schritten – dem Pre-Cluster-Schritt und dem Cluster-Schritt (vgl. Schendera, 2010). Im Pre-Cluster-Schritt werden zunächst alle Fälle (Studien) sequentiell abgescannt und zu

sog. Sub-Clustern vorverdichtet (u.a. Janssen & Laatz, 2007)[53]. Während des Pre-Cluster-Schrittes entsteht eine Art Clusterbaum, der aus mehreren Ebenen besteht. Jeder Fall wird durch diesen Clusterbaum „geschickt", wobei mit Hilfe eines Distanzkriteriums jeweils entschieden wird, ob der jeweilige Fall in eines der bereits bestehenden Pre-Cluster eingeordnet wird oder, ob ein neues Pre-Cluster gebildet werden muss. Im zweiten Schritt (genannt: Cluster-Schritt) werden die gebildeten Sub-Cluster mittels agglomerativer hierarchischer Methoden zu den Endclustern zusammengefasst. Dabei werden nun die Sub-Cluster wie Fälle behandelt und diejenigen Sub-Cluster fusioniert, durch die die bestehende Heterogenität innerhalb der Cluster am minimalsten erhöht wird. Ergebnis der Two-Step Clusteranalyse ist, bei entsprechender Voreinstellung in der Analysesoftware, die Ausgabe einer optimalen Clusterlösung sowie derjenigen Fälle (Studien), die vom Verfahren als Ausreißer erkannt wurden.

Die hierarchische Clusteranalyse nach der *Ward*-Methode nimmt Fusionierungen der Fälle derart vor, dass N-1 Fusionierungsschritte erfolgen, im Rahmen derer vom Ausgangspunkt (jeder Fall = ein Cluster) bis zum Endpunkt (alle Fälle bilden ein einziges Cluster) die jeweiligen Heterogenitätszuwächse innerhalb der Cluster berechnet werden. Beim Ausgangspunkt besteht, mit jeweils einem Fall pro Cluster, absolute Homogenität innerhalb der Cluster, mit jeder Fusionierung steigt dementsprechend die Heterogenität. Die Heterogenität wird im Zuge der Clusteranalyse nach dem *Ward*-Verfahren als Summe der quadrierten euklidischen Distanzen (Fehlerquadratsumme) innerhalb der Cluster berechnet.

Die Durchführung der clusteranalytischen Verfahren erfolgte mithilfe der Statistik Analysesoftware SPSS (Version 20).

[53] Das sequentielle Abscannen erfolgt entsprechend der Reihenfolge, in der die Fälle in der SPSS-Matrix (zeilenweise) eingegeben wurden. Aus diesem Grunde empfiehlt es sich, die Two-Step Clusteranalyse mindestens zwei Mal durchzuführen und die Fälle zwischen den Durchgängen einer Zufallsanordnung zu unterziehen.

3.4.2 Darstellung der Studienqualität anhand der drei Dimensionen des WBP-Kriterienkatalogs (allgemeine methodische Qualität, interne Validität, externe Validität)

In Kapitel 4.3 werden die Kodierungen aller 41 Primärstudien auf den drei Dimensionen der allgemeinen methodischen Qualität sowie der internen und der externen Validität dargestellt.

Wie in Kapitel 3.2.1 ausführlich dargestellt, wurden für einige Kriterien Zusatzkodierregeln formuliert, durch die Studien mit speziellen und damit unkodierbaren Eigenschaften mit bestimmtem Missingwerten belegt wurden. Im Rahmen einer Rekodierung der Studien wurden einige dieser Zusatzkodierregeln aufgeweicht, indem durch plausible Modifikationen der Ratingstufenoperationalisierungen versucht wurde, Missingwerte zu vermeiden und zu Ratings auf den regulären Stufen („1“ bis „3“) umzuwandeln. Die Auswirkung dieser Modifikationen auf die dimensionalen Gesamtbewertungen der Studien wird ebenfalls in Kapitel 4.3 dargestellt.

3.4.3 Untersuchung der Gegenstandsadäquatheit der Kriterien der allgemeinen methodischen Qualität und der internen Validität: Vergleich von Langzeittherapiestudien mit Studien zu Therapien kürzerer Behandlungsdauer

Da ein positives Abschneiden auf der Dimension der allgemeinen methodischen Qualität eine Voraussetzung für eine positive Bewertung auf den beiden Validitätsdimensionen bildet, wird diese Dimension entsprechend als *Teil* der internen Validitätsbewertung betrachtet. Obwohl der Fokus der Untersuchung auf der internen Validitätsdimension liegt, soll aus diesem Grund die allgemeine methodische Qualitätsbewertung in die Untersuchung miteinbezogen, jedoch lediglich auf die K.O.-Kriterien dieser Dimension beschränkt werden.

Um Aufschluss über die Gegenstandsangemessenheit der Kriterien zu erlangen (vgl. Kap. 2), wird für jedes Kriterium eruiert, wie Langzeittherapiestudien im Vergleich zu Studien zu kürzeren Behandlungsdauern abschneiden. Dazu werden die Verteilungen der Langzeittherapiestudien mit den Verteilungen der Kurzzeittherapiestudien über die unterschiedlichen Kriterienstufen („optimal“, „zufriedenstellend“, „ungenügend“) verglichen. Durch diesen Vergleich sollen vor allem solche Kriterien erkannt werden, bei denen die prozentualen Anteile an Negativbewertungen auf Seiten der Langzeittherapiestudien größer sind als selbige Anteile auf Seiten der Kurzzeittherapiestudien.

Zu Auswertungszwecken werden sowohl die Variable „Sitzungsumfang“ als auch die 3-gestuften Kriterien dichotomisiert: Für die Dichotomisierung des Sitzungsumfangs werden den Langzeittherapiestudien (über 100 Sitzungen) die Studien zu Kurzzeitbehandlungen (bis 25 Sitzungen) und zu Behandlungen moderater Länge (über 25 bis 100 Sitzungen) gegenübergestellt (vgl. Leichsenring, 2005). Auf Kriterienebene werden die beiden unteren Kategorien („optimal“ und „zufriedenstellend“) der oberen Kategorie („ungenügend“) gegenübergestellt. Zusätzlich zum Vergleich zwischen Langzeittherapiestudien (über 100 Stunden) und den Studien zu kürzeren Behandlungsdauern (unter 100 Stunden) werden Extremgruppenvergleiche angestellt: Hierbei werden die Studien zu moderaten Therapieumfängen (über 25 bis 100 Sitzungen) exkludiert und den Langzeittherapiestudien die „echten“ Kurzzeittherapiestudien (bis 25 Stunden) gegenübergestellt. Mit diesem Extremgruppenvergleich wird dem Postulat entsprochen, demzufolge das RCT-Paradigma und die damit zusammenhängenden internen Validitätssicherungsstrategien nicht nur als mit dem Gegenstand "Langzeittherapie" inkompatibel betrachtet werden. Vielmehr wird das RCT-Paradigma strenggenommen als ausschließlich auf den Gegenstand "Kurzzeittherapie" anwendbar betrachtet. So ziehen Westen et al. (2004) in ihrer kritischen Analyse des RCT-Paradigmas beispielhaft immer wieder Behandlungsumfänge von „6 to 16 sessions“ heran (S. 632). Im Management Handbuch für Psy-

chotherapeutische Praxis (2011) resümiert Leichsenring die Kritik, die gemeinhin an der RCT-Methodologie formuliert wird und schreibt: „Auch sei die Methodologie des RCT allenfalls für Kurzzeittherapien angemessen, nicht jedoch für Langzeittherapien“ (S. 5). Dieser Beschränkung der RCT-Methodologie und der entsprechenden internen Validitätssicherungsmaßnahmen auf den Bereich der Kurzzeittherapie soll durch den beschriebenen Extremgruppenvergleich entsprochen werden. Es wird damit – gewissermaßen strenger – untersucht, ob die Kriterien unter Betrachtung lediglich dieser beiden Extremgruppen zu potentiellen Benachteiligungen der Langzeittherapiestudien führen.

Als Maß der Homogenität bzw. Unterschiedlichkeit der Verteilungen wird aufgrund der vorliegenden Vollerhebung (vgl. Kap. 3.1.3) auf inferenzstatistische Verfahren verzichtet und auf das Effektstärkemaß ω (Omega) zurückgegriffen (vgl. Formel 1).

Formel 1: Effektstärkemaß ω (Omega)

$$\widehat{\omega} = \sqrt{\frac{\chi^2_{df}}{n}}$$

(aus: Eid, Gollwitzer & Schmitt, S. 292, 2010).

Dieses auf dem empirischen χ^2-Wert beruhende Effektstärkemaß eignet sich auch dann für die Bewertung von Prozentwertunterschieden, wenn die Verteilung der Prüfgröße etwa aufgrund niedriger, erwarteter Zellenbesetzungen keiner χ^2-Verteilung folgt (M. Eid, persönl. Mitteilung, 10.02.2014). So stellen Bortz, Lienert und Boehnke (2000) selbiges Effektstärkemaß – allerdings unter dem Namen Φ' (Phi) – insbesondere für sehr kleine Stichproben (z.B. *N*=15; S. 343) vor.

Für die Bewertung der Effekte wird folgende Einteilung nach Cohen (1988) herangezogen:

$\omega \approx 0.10$: kleiner Effekt

$\omega \approx 0.30$: mittlerer Effekt

$\omega \approx 0.50$: großer Effekt

(aus: Eid et al., S. 292, 2010).

Die Verteilungsvergleiche bilden sodann die Grundlage für die Überlegungen, in denen der Frage nachgegangen wird, inwieweit methodische bzw. evaluatorische Gründe dafür oder dagegen sprechen, dass Studien zu Langzeitbehandlungen im Vergleich zu kürzeren Behandlungen einer tatsächlichen Benachteiligung durch besagte Kriterien unterliegen. Dabei wird sich auf solche Kriterien beschränkt, bei denen sich Unterschiede in den Verteilungen (Prozentwerten) zugunsten der Studien zu kürzeren Behandlungsdauern in Effekten ab einer Größe von $\omega \geq 0.10$ ausdrücken. Für diese Kriterien werden die Studien einer Art „Feinanalyse" unterzogen, im Rahmen derer die Gründe für das Negativergebnis auf Seiten der Langzeittherapiestudien im Kontrast du denen zu kürzeren Therapiedauern eruiert werden.

Zudem soll die feinanalytische Untersuchung in Bezug auf die interne Validitätsdimension ausgeweitet und damit strenger gestaltet werden, indem hierbei nicht nur diejenigen Häufigkeitsverteilungen der Studien über die Kriterienstufen in Augenschein genommen werden, die aufgrund einer Ungleichverteilung zugunsten der Kurzzeittherapiestudien einen Indikator für eine potentielle Benachteiligung von Langzeittherapieuntersuchungen darstellen. Vielmehr sollen auch solche Verteilungen näher betrachtet werden, die zunächst *keinen* Zusammenhang zwischen Behandlungsdauer (Sitzungsumfang) und Ergebnis auf einem Kriterium nahelegen. Bei solchen Verteilungen muss strenggenommen von der Möglichkeit ausgegangen werden, dass Zusammenhänge verschleiert werden können, indem sich bspw. im Rahmen der Durchführung mancher Kurzzeittherapiestudie bewusst dafür entschieden wurde, diese oder jene interne Validitätssicherungsstrategie aus Gründen der externen Validität *nicht* umzusetzen, wie z.B. den kontrollierten Einsatz von Behandlungsmanualen. Gleichsam könn-

te es sich bei selbigen Strategien um solche handeln, die in Langzeittherapiestudien schlicht nicht oder nur sehr schwierig umsetzbar sind. Durch Erweiterung des Feinanalysegegenstandes auch auf solche Kriterien, die auf den ersten verteilungsanalytischen Blick noch keinen Anhaltspunkt für potentielle Benachteiligungen von Langzeittherapiestudien bieten, soll das Übersehen dennoch benachteiligender Kriterien vermieden werden.

4 Ergebnisse

4.1 Allgemeiner Überblick über die Primärstudien

Die Anhänge F und G geben einen Überblick über allgemeine und methodologische Studiencharakteristika in Anlehnung an den Kurzkodierbogen (vgl. Kap. 3.2.4)[54]. Die 41 Primärstudien belaufen sich auf 10 Studien zu affektiven Störungen[55] und 31 Studien an diagnoseheterogenen Störungsgruppen[56] (vgl. Tabelle 4). Da bei insgesamt 10 der 31 Untersuchungen an gemischten Störungsgruppen keine standardisierte ICD- oder DSM-Diagnostik beschrieben oder durchgeführt wurde, konnte bei diesen lediglich geschätzt werden, dass es sich um heterogene Diagnosegruppen handelt. So ist kein Hinweis auf die Durchführung standardisierter Diagnostik (via ICD oder DSM) in insgesamt vier Studien zu finden (4, 10, 28, 35), jedoch wurden in diesen Publikationen unterschiedliche Diagnosegruppen in Form von ICD- bzw. DSM-Klassifikationen aufgezählt. In zwei weiteren Studien (36, 37) wurden lediglich einzelne Symptomatiken aufgelistet (*anxiety, depression, interpersonal problems, self-esteem, bereavement/loss, trauma/abuse, personality problems* etc.), in einer Studie (5) wird nur von prozentualen Anteilen an Achse-I und -II-Störungen berichtet, ohne einzelne Störungsgruppen zu benennen. In einer weiteren Studie (29) konnten retrospektiv lediglich für knapp 53% der Gesamtgruppe Diagnosen eruiert werden, von denen wiederum 97% der Patienten eine Diagnose aus dem depressiven Formenkreis („Diagnosis of depression“ Paley et al., 2008, S. 161)

[54] Beiden Tabellen der Anhänge F und G sind in der ersten Spalte die jeweiligen Erstautoren, ein Studientitel sowie ein hochgestellter Ziffernindex zu entnehmen. Um die Lesbarkeit im Ergebnisteil zu erleichtern, wird bei Verweis auf Studien statt der Autoren lediglich ein Platzhalter in Form der Ziffernindizes eingefügt.

[55] Das sind die Studien mit den Ziffernindizes 1, 9, 14, 20, 21, 26, 27, 32, 39, 40.

[56] Das sind die Studien mit den Ziffernindizes 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 15, 16, 17, 18, 19, 22, 23, 24, 25, 28, 29, 30, 31, 33, 34, 35, 36, 37, 38, 41.

aufwiesen. Für eine Studie, die sich auf den Datensatz der *Penn Psychoanalytic Treatment Collection* bezieht (25), konnte, nach Kontaktaufnahme zu den Erstautoren, ermittelt werden, dass „No systematic diagnostic information exists“ (P. Crits-Christoph[57], persönl. Mitteilung, 14.09.2009). Bei einer Untersuchungsgruppe (33) handelt es sich um Männer, deren Behandlungsbedürftigkeit aufgrund psychischer Folgen sexueller Missbrauchserfahrungen festgestellt wurde, wobei weitere diagnostische Details unbekannt bleiben.

Vor allem bei den zuletzt genannten Studien ist es demnach fraglich, welche psychischen Störungen entsprechend ICD oder DSM oder welche subklinischen Symptomausprägungen tatsächlich vorlagen. Es ist jedoch davon auszugehen, dass es sich um keine störungs- oder symptomhomogenen Gruppen handelt.

Tabelle 4: Allgemeine Charakteristika der Primärstudien (*N*=41)

	Affektive Störungen (*n*=10)			**Gemischte Störungen (*n*=31)**		
Datenzugriff N (%)	Originalstudie	10	(100%)	Originalstudie	28	(90.3%)
				Reanalyse	1	(3.2%)
				Replikation	2	(6.5%)
Gesamt %			**100%**			**100%**
Psychoanalytisch begründetes Psychotherapieverfahren (plus Psa) N (%)	TP	9	(90.0%)	TP	12	(38.7%)
	nicht beurteilbar	1	(10.0%)	AP	6	(19.4%)
				Psa	1	(3.2%)
				TP und AP	7	(22.6%)
				AP und Psa	3	(9.7%)
				TP, AP und Psa	1	(3.2%)
				nicht beurteilbar	1	(3.2%)
Gesamt %			**100%**			**100%**
Setting N (%)	Individualth.	9	(90.0%)	Individualth.	25	(80.6%)
	Gruppenth.	1	(10.0%)	Gruppenth.	5	(16.1%)
				beides	1	(3.2%)
Gesamt %			**100%**			**100%**

Anmerkung: AP: Analytische Psychotherapie, Individualth.: Individualtherapie, Gruppenth.: Gruppentherapie, Psa: Psychoanalyse, TP: Tiefenpsychologisch fundierte Psychotherapie.

[57] Crits-Christoph antwortete stellvertretend für Luborsky.

Bei den 10 Studien zu affektiven Störungen handelt es sich ausschließlich um Originalstudien. Die 31 Studien zu gemischten Störungsgruppen hingegen verteilen sich auf die drei Formen des Datenzugriffs, wenn es sich auch bei dem größten Teil (90.3%) ebenfalls um Originalstudien handelt (vgl. Tabelle 4).

In den Untersuchungen zu affektiven Störungen wurden nahezu ausschließlich tiefenpsychologisch fundierte Behandlungen durchgeführt. Bei einer Studie (26) war aufgrund mangelnder Information keine eindeutige Zuordnung zu einem der psychoanalytisch begründeten Richtlinienverfahren (plus Psychoanalyse) möglich.

Bei den Untersuchungen an gemischten Störungsgruppen beläuft sich der größte Teil mit 12 Studien (38.7%) ebenfalls auf die Durchführung von tiefenpsychologisch fundierter Psychotherapie (2, 7, 12, 13, 15, 24, 28, 29, 31, 33, 34, 41), gefolgt von sechs Studien (19.4%) zu analytischer Psychotherapie (3, 6, 8, 10, 18, 22) und einer Studie, in der ausschließlich klassische Psychoanalyse durchgeführt wurde (25).

Außerdem wurden in sieben Studien (22.6%) sowohl tiefenpsychologisch fundierte als auch analytische Psychotherapien durchgeführt, davon wurden in fünf Studien beide Verfahren zusammengenommen (in einem Treatmentarm) untersucht (16, 17, 36, 37, 38), während in zwei Studien ein verfahrensinterner Vergleich[58] zwischen den beiden Verfahren angestellt (11, 19) wurde.

In drei Studien (9.7%) wurden sowohl analytische Psychotherapien als auch Psychoanalysen durchgeführt, davon erfolgte wiederum in zwei Studien ein verfahrensinterner Ver-

[58] Die Definitionen zu den Bezeichnungen *verfahrensinterne Vergleiche* und *verfahrensexterne Vergleiche* lassen sich in Kapitel 3.2.1 nachlesen.

gleich zwischen den beiden Verfahren (4, 5), während in einer Studie die beiden Verfahren in einem Treatmentarm zusammengenommen untersucht wurden (23).

Eine weitere Studie bezieht sich auf alle drei Verfahren (35), wobei ein Treatmentarm die beiden Verfahren analytische Psychotherapie und Psychoanalyse enthält und mit dem zweiten Treatmentarm (tiefenpsychologisch fundierte Psychotherapie) verglichen wurde. Für eine letzte Studie aus dem gemischten Störungspool konnte nicht eindeutig eruiert werden, welche/s psychoanalytisch begründete/n Verfahren durchgeführt wurde/n (30).

Tabelle 4 (S. 205) gibt außerdem Aufschluss über das Behandlungssetting (Individual- *versus* Gruppentherapie). Dabei zeigt sich, dass affektive Störungen im hiesigen Studienpool überwiegend in Einzelsettings behandelt wurden (90.0%) (1, 9, 14, 21, 26, 27, 32, 39, 40), lediglich in einer Studie wurde selbige Störungsgruppe mittels Gruppentherapie behandelt (20).

Ein ähnliches Verhältnis der beiden Behandlungssettings ergibt sich bei den Studien zu gemischten Störungsgruppen: Hier wurden von den 31 Studien bei insgesamt 25 Studien (80.6%) Individualtherapien durchgeführt (2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 15, 16, 17, 18, 19, 22, 23, 25, 29, 34, 35, 36, 37, 41) und bei lediglich fünf Studien (16.1%) Gruppentherapien (24, 28, 31, 33, 38). In einer Studie wurden beide Settings untersucht (30).

Ein differenzierter Blick auf die Behandlungslängen (Sitzungsumfänge), segregiert nach Anwendungsbereich, Behandlungssetting und psychoanalytisch begründetem Therapieverfahren, ist Tabelle 5 zu entnehmen. Für diese Ergebnisdarstellung wurden die in die einzelnen Studien eingehenden psychoanalytisch begründeten Behandlungen (plus Psychoanalyse) in Anlehnung an Leichsenring (2005) in die Modi Kurzzeittherapie (bis 25 Sitzungen), mittelfristige Therapie (über 25 bis 100 Sitzungen) und Langzeittherapie (über 100 Sitzungen) transformiert. Wurden in einer Studie mehrere psychoanalytisch begründete Verfahren unterschiedli-

cher Sitzungsumfänge untersucht (z.B. Studie 19), so gab die Behandlung mit der höheren Sitzungsanzahl den Ausschlag für die o.g. Einordnung.

Tabelle 5: Settings, psychoanalytisch begründete Therapieverfahren und Sitzungsumfang (*N*=41)

		Affektive Störungen (*n*=10) Sitzungsumfang *n*		**Gemischte Störungen (*n*=31) Sitzungsumfang *n***	
Individual-therapie (*n*=24)	TP	KZT	6	KZT	5
		MfT	2	MfT	3
		LZT	-	LZT	-
	AP	-		KZT	-
				MfT	-
				LZT	6
	Psa	-		KZT	-
				MfT	-
				LZT	1
	TP und AP	-		KZT	2
				MfT	-
				LZT	4
	AP und Psa	-		KZT	-
				MfT	-
				LZT	3
	TP, AP und Psa	-		KZT	-
				MfT	-
				LZT	1
	nicht beurteilbar	KZT	-	-	
		MfT	-		
		LZT	1		
Gruppen-therapie (*n*=6)	TP	KZT	-	KZT	2
		MfT	1	MfT	2
		LZT	-	LZT	-
	AP	-		-	
	Psa	-		-	
	TP und AP	-		KZT	-
				MfT	-
				LZT	1
	AP und Psa	-		-	
	TP, AP und Psa	-		-	
	nicht beurteilbar	-		-	
Individual- & Gruppen-therapie (*n*=1)	*nicht beurteilbar*	-		KZT	-
				MfT	1
				LZT	-
	Gesamt n (%)	KZT	6 (60.0%)	KZT	9 (29.0%)
		MfT	3 (30.0%)	MfT	6 (19.4%)
		LZT	1 (10.0%)	LZT	16 (51.6%)
			(100%)		**(100%)**

Anmerkung: Die umrahmte Zelle verweist auf zwei Studien (36, 37), bei denen keine eindeutige Verfahrenseinordnung möglich war. AP: Analytische Psychotherapie, KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie, Psa: Psychoanalyse, TP: Tiefenpsychologisch fundierte Psychotherapie.

Bei den Studien zu ausschließlich tiefenpsychologisch fundierter Psychotherapie lassen sich settingübergreifend (Individual- *versus* Gruppentherapie) ausschließlich kurze bis moderate Therapielängen (Sitzungsumfänge) ermitteln, bei der analytischen Psychotherapie und Psychoanalyse lassen sich ausschließlich Langzeittherapien auffinden. Bei den Studien, in die mehrere psychodynamische Verfahren eingehen (in einem Treatmentarm oder in Form von verfahrensinternen Vergleichen), sind zwei Kurzzeittherapiestudien zu tiefenpsychologisch fundierter und analytischer Psychotherapie (in einem Treatmentarm) zu finden, die in Tabelle 5 (rechte Spalte) umrahmt sind. Hierbei handelt es sich um zwei Studien (eine Originalstudie [36] und deren Replikation [37]), in denen die Interventionsbeschreibungen nahezu gänzlich fehlen und die Verfahrenseinordnung lediglich aufgrund der beiden Studien zu entnehmenden Verfahrenstitel vorzunehmen war (vgl. Anhang F): „*Psychodynamic or psychoanalytic therapy*“ und in der deutschen Übersetzung (36) „psychodynamische oder psychoanalytische Therapie“. Die explizite Aufzählung *zweier* Verfahrenstitel legt nahe, dass tatsächlich *zwei* unterschiedliche Therapieformen gemeint sind – wahrscheinlich tiefenpsychologisch fundierte und analytische Psychotherapie – trotz der auffällig niedrigen mittleren Sitzungsanzahl in beiden Studien (ca. 8 Sitzungen). Da den Publikationen, außer dem mittleren Sitzungsumfang, keine weiteren Informationen zu den Therapieverfahren zu entnehmen sind, wurde pragmatisch entschieden, diese Studien in Anlehnung an die vergebenen Verfahrenstitel beiden Verfahren (tiefenpsychologisch fundierte und analytische Psychotherapie) zuzuordnen.

Zwischen den Anwendungsbereichen spiegelt sich die größere Vielfalt hinsichtlich der unterschiedlichen psychoanalytischen Verfahren auf Seiten der gemischten Störungen (vgl. Tabelle 4) folgendermaßen im Hinblick auf die Sitzungsumfänge wider (vgl. Tabelle 5): Da auf Seiten der Studien zu affektiven Störungsgruppen – mit einer Ausnahme – tiefenpsychologisch fundiert behandelt wurde, lassen sich – mit derselben Ausnahme – primär kurz- und mittelfristige

Therapien finden (6 Kurzzeittherapiestudien : 3 Studien zu mittelfristigen Therapien : 1 Langzeittherapiestudie). Bei den Studien zu den gemischten Störungen zeigt sich, dass, mit dem erhöhten Anteil an Studien zu analytischer Psychotherapie und Psychoanalyse, die Langzeittherapie anteilmäßig überwiegt (9 Kurzzeittherapiestudien : 6 Studien zu mittelfristigen Therapien : 16 Langzeittherapiestudien). Die Anteile von Kurzzeittherapiestudien und Langzeittherapiestudien zum jeweiligen Gesamt-*N* der beiden Anwendungsbereiche (affektive Störungen: *n*=10; gemischte Störungen: *n*=31) kehren sich zwischen den Anwendungsbereichen regelrecht um: Bei den affektiven Störungsgruppen machen die Kurzzeittherapiestudien 60.0%, bei den gemischten Störungsgruppen die Langzeittherapiestudien 51.6% aus.

Tabelle 6 gibt die Verteilungen der in den Studien gewählten Studiendesigns inklusive des prospektiven *versus* retrospektiven Datenzugangs wieder.

Tabelle 6: Studiendesigns und Datenzugänge der Primärstudien (*N*=41)

	Affektive Störungen (*n*=10)		**Gemischte Störungen (*n*=31)**	
Studiendesign *n* (%)	Ein-Gruppen-Design	4 (40.0%)	Ein-Gruppen-Design	14 (45.2%)
	WL, TAU, Placebo, aktive KG	3 (30.0%)	WL, TAU, Placebo, aktive KG	6 (19.4%)
	verfahrensinterner Vergleich	1 (10.0%)	verfahrensinterner Vergleich	6 (19.4%)
	verfahrensexterner Vergleich	-	verfahrensexterner Vergleich	4 (12.9%)
	Kombi verfahrensinterner & verfahrensexterner Vergleich	-	Kombi verfahrensinterner & verfahrensexterner Vergleich	1 (3.2%)
	Kombi KG (WL) & verfahrensexterner Vergleich	1 (10.0%)	Kombi KG (WL) & verfahrensexterner Vergleich	-
	Vergleich mit Pharmaka	1 (10.0%)	Vergleich mit Pharmaka	-
	Gesamt	**100%**	**Gesamt**	**100%**
Datenzugang *n* (%)	prospektiv	10 (100%)	prospektiv	18 (58.1%)
	retrospektiv	-	retrospektiv	13 (41.9%)
	Gesamt	**100%**	**Gesamt**	**100%**

Anmerkung: KG: Kontrollgruppe, TAU: Treatment-As-Usual, WL: Warteliste.

Die durchweg prospektiven Untersuchungen an affektiven Störungsgruppen enthalten vier Studien in "echten Ein-Gruppen-Designs"[59] (40%) (4, 14, 20, 26). In drei Untersuchungen wurden Vergleiche mit klassischen Kontrollgruppen vorgenommen: Eine Studie wählte eine Warteliste-Kontrollgruppe (21), eine andere Studie TAU als Kontrollgruppe (40). In der dritten Studie (9) wurde ein Treatmentarm mit TAU, ein zweiter Treatmentarm mit einer aktiven Kontrollbedingung (*non-directive counselling*) und ein dritter Treatmentarm mit einem bereits etablierten Verfahren – kognitiv-behaviorale Therapie – realisiert. Hierbei wurden jedoch ausschließlich Vergleiche der jeweiligen Treatmentarme – so auch des psychodynamischen Treatmentarms – mit der TAU-Gruppe angestellt, so dass diese Studie hier lediglich als Kontrollgruppenvergleich eingeht.

In einer Studie (39) wurde ein verfahrensinterner Vergleich durchgeführt, der in der weiteren Kodierung mittels WBP-Kriterienkatalog als "Ein-Gruppen-Design im erweiterten Sinne" zu behandeln sein wird. Hierbei wurden zwei Treatmentarme miteinander verglichen, in einem Treatmentarm befanden sich Patienten in einer tiefenpsychologisch fundierten Behandlung (s*hort-term psychodynamic supportive psychotherapy*), der sie sich selbst zugewiesen haben. Im anderen Treatmentarm wurde die gleiche Behandlung durchgeführt, wobei die Patienten dieser (im Rahmen einer anderen Untersuchung) zufällig zugewiesen wurden (vgl. Anhang G).

In keiner Studie wurde ausschließlich ein verfahrensexterner Vergleich oder eine Kombination aus verfahrensinternen und verfahrensexternen Vergleichen unternommen. In einer Studie (27) wurde ein Vergleich mit einer klassischen Kontrollgruppe und mit einer verfahrensexternen Vergleichsgruppe angestellt: Hier wurde als klassische Kontrollgruppe eine Wartelistengruppe und als verfahrensexterner Vergleich ein nicht etabliertes Verfahren einge-

[59] Zur Definition der Begriffe "echte Ein-Gruppen-Designs" sowie "Ein-Gruppen-Designs im erweiterten Sinne" vgl. Kapitel 3.2.1 .

setzt. Lediglich eine weitere Studie (32) wählte als Vergleichsbehandlung eine pharmakologische Monotherapie.

Bei den Studien an gemischten Störungsgruppen besteht im Vergleich zu den Studien an affektiven Störungsgruppen ein ähnliches Verhältnis, was den Anteil von Studien im "echten Ein-Gruppen-Design" betrifft (gemischte Störungen: 45.2%; affektive Störungen: 40.0%) (vgl. Tabelle 6, S. 211). Um sich dem Phänomen der "echten Ein-Gruppen-Designs" ein wenig zu nähern, sollen die Behandlungsumfänge in diesen Untersuchungsdesigns genauer betrachtet werden (vgl. Tabelle 7). Von den insgesamt 18 Studien im Ein-Gruppen-Design (beide Anwendungsbereiche) werden bei 14 Studien mittelfristige oder Langzeittherapien und bei vier Studien Kurzzeittherapien untersucht. Im hiesigen Studienpool überwiegen demnach in der Kategorie der Ein-Gruppen-Designs die moderaten bis langfristigen Behandlungsumfänge.

Tabelle 7: Studiendesigns und Sitzungsumfänge in den Primärstudien (*N*=41)

	Affektive Störungen (*n*=10) Therapieumfang n		**Gemischte Störungen (*n*=31)** Therapieumfang n	
Ein-Gruppen-Design (*n*=18)	KZT	1	KZT	3
	MfT	2	MfT	4
	LZT	1	LZT	7
WL, TAU, Placebo, aktive KG (*n*=9)	KZT	2	KZT	3
	MfT	1	MfT	1
	LZT	-	LZT	2
verfahrensinterner Vergleich (*n*=7)	KZT	1	KZT	1
	MfT	-	MfT	1
	LZT	-	LZT	4
verfahrensexterner Vergleich (*n*=4)	-		KZT	2
			MfT	-
			LZT	2
Kombination aus verfahrensinternem & verfahrensexternem Vergleich (*n*=1)	-		KZT	-
			MfT	-
			LZT	1
Kombination aus KG (WL) & verfahrensexternem Vergleich (*n*=1)	KZT	1	-	
	MfT	-		
	LZT	-		
Vergleich mit Pharmaka (*n*=1)	KZT	1	-	
	MfT	-		
	LZT	-		

Anmerkung: Die durchgängig umrahmten Zellen verweisen auf Studien, in denen klassische Kontrollgruppen zum Vergleich herangezogen wurden; die gestrichelt umrahmten Zellen verweisen auf Studien, in denen verfahrensinterne Vergleiche durchgeführt wurden; die grau hinterlegten Zellen verweisen auf Studien, in denen verfahrensexterne Vergleiche durchgeführt wurden. KG: Kontrollgruppe, KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie, TAU: Treatment-As-Usual, WL: Warteliste.

Auf Seiten der gemischten Störungen sind insgesamt sechs (19.4%) der 31 Studien mit einem klassischen Kontrollgruppendesign (Warteliste, TAU, Placebo, aktive Kontrollgruppe) zu finden (vgl. Tabelle 6, S. 211). Anwendungsbereichsübergreifend sind es insgesamt 10 Studien, in denen klassische Kontrollgruppen zum Vergleich herangezogen werden (in Tabelle 7 durchgängig umrahmt). Ein genauerer Blick in die Behandlungsumfänge, für die diese Kontrollgruppen herangezogen wurden, sowie in die jeweiligen Realisationen der Kontrollgrup-

pen zeigt: Auf Seiten der affektiven Störungen ist die Kategorie „Kurzzeittherapie“ mit drei von insgesamt vier Studien am höchsten besetzt, bei den gemischten Störungen ist selbige Kategorie zwar ebenfalls mit drei von sechs Studien am höchsten besetzt, jedoch sind hier auch drei Studien im Kontrollgruppendesign zu finden, die an längerfristigen Therapien (> 25 Stunden) durchgeführt wurden. Beide Anwendungsbereiche zusammengenommen, sind also insgesamt vier Studien zu finden, die Kontrollgruppendesigns bei mittelfristigen und Langzeittherapien realisierten (3, 18, 21, 41). Stellt man die Umsetzungen der Kontrollgruppendesigns bei den Studien zu längerfristigen Behandlungen denen bei Kurzzeittherapiestudien gegenüber, dann zeigt sich Folgendes: Bei einer Studie (18) zu Langzeittherapie (> 3 Jahre) überdauerte die Warteliste nicht den gesamten Behandlungszeitraum, sondern umfasste lediglich die ersten 12 Monate. Selbiges trifft auf eine weitere Studie (21) zu mittelfristiger Therapie (80 Sitzungen, entspricht bei 2 Sitzungen/Woche ca. 1 Jahr) zu, in der die Warteliste lediglich parallel zu den ersten 6 Behandlungsmonaten verlief. Eine weitere Studie (41) zu mittelfristiger Therapie (*M* 73.4 Sitzungen, entspricht bei 1-2 Sitzung/Woche durchschnittlich 1.5 Jahre) realisierte eine Warteliste ebenfalls von 6 Monaten. Lediglich in einer Studie (3) zu Langzeittherapie (*M* 159 Sitzungen, entspricht bei 1-2 Sitzungen/Woche durchschnittlich 3 Jahre) ist es gelungen, eine unbehandelte Kontrollgruppe (*n*=10) zumindest über den mittleren Zeitraum der Langzeittherapie mitzuführen. Wilczek, Weinryb, Barber, Åsberg und Gustavsson (2004) ist zu entnehmen, wie dies gelingen konnte: Im Grunde genommen handelt es sich bei dieser Kontrollgruppe nicht um eine Warteliste, sondern um eine unbehandelte Gruppe, die sich kurz vor Therapiebeginn entschied, doch nicht an der Therapie teilzunehmen. Im Verlauf der drei Folgejahre wurde festgestellt, dass niemand in dieser Gruppe sich anderweitig einer Therapie unterzogen hatte, so dass diese Gruppe als Vergleichsgruppe betrachtet werden kann, die den natürlichen Störungsverlauf kennzeichnet. So erklärt sich der verhältnismäßig lange zeitliche Umfang, den diese Kontrollgruppe umfasst.

Betrachtet man als nächstes über beide Anwendungsbereiche die sechs Kurzzeittherapiestudien mit klassischen Kontrollgruppen (inklusive der Kombination mit verfahrensexternem Vergleich), dann stellt sich folgendes Bild dar: In drei Studien wurden Wartelisten-Kontrollgruppen herangezogen (2, 27, 34). Die Behandlungslängen umfassten dabei im Mittel maximal 19.6 Sitzungen (27), was bei wöchentlichen Sitzungen einem mittleren Zeitraum von 6 Monaten entspricht. Bei den restlichen drei Kurzzeittherapiestudien wurde TAU als Kontrollgruppe herangezogen (9, 12, 40). Als TAU fungierten Hausarztbehandlungen und häusliche Gesundheitshilfen (*health visitors*) (9) sowie *Standard treatment* nach klinischen Richtlinien des chilenischen Gesundheitsministeriums (40). In einer Studie (12) wurden als TAU reguläre psychiatrische Gespräche von 15- bis 30-minütiger Dauer angeboten. Außerdem fielen in dieser Studie unter die Definition „TAU“ ebenfalls die Optionen, in eine kognitiv-behaviorale Therapie, ein kommunales Alkoholprogramm, in eine Angstbewältigungsgruppe oder aber in eine tagesklinische oder stationäre Behandlung überwiesen zu werden. Inwiefern diese Optionen tatsächlich realisiert wurden, ist der Studie nicht zu entnehmen.

Festzuhalten gilt, dass über beide Anwendungsbereiche hinweg bei den Kurzzeittherapieuntersuchungen alle Kontrollgruppen über den gesamten Behandlungszeitraum mitgeführt wurden, was bei den Studien zu längeren Behandlungen > 25 Sitzungen nur bei einer Ausnahme (3) der Fall war.

Verfahrensinterne Vergleiche wurden auf Seiten der gemischten Störungen bei insgesamt sieben (22.6%) der 31 Studien durchgeführt, bei einer dieser Studien wurde dieser Vergleich mit einem verfahrensexternen Vergleich kombiniert (19) (vgl. Tabelle 6, S. 211). Innerhalb dieser Kategorie sind eine Kurzzeittherapiestudie (31) sowie sechs Studien zu finden, die Behandlungsumfänge von mindestens moderater Länge umfassen (4, 5, 11, 15, 19, 35) (in Ta-

belle 7 gestrichelt umrahmt). Es wurden folgende verfahrensinterne Vergleiche vorgenommen (vgl. Anhang G):

- analytische Psychotherapie (*supportive-expressive psychotherapy*) und Psychoanalyse (*psychoanalysis*) (4)
- analytische Psychotherapie (*psychodynamic long-term psychotherapy*) und Psychoanalyse (*psychoanalysis*) (5)
- tiefenpsychologisch fundierte Psychotherapie (*psychodynamic therapy*) und analytische Psychotherapie (*psychoanalytic therapy*) (11)
- tiefenpsychologisch fundierte Psychotherapie (*dynamic psychotherapy*) mit *versus* ohne Übertragungsdeutungen (15)
- tiefenpsychologisch fundierte Psychotherapie (s*hort-term psychodynamic psychotherapy*) und analytische Psychotherapie (*long-term psychodynamic psychotherapy*) (19)
- tiefenpsychologisch fundierte Psychotherapie Nr. 1 (*interpretive therapy*) und tiefenpsychologisch fundierte Psychotherapie Nr. 2 (*supportive therapy*) (31)
- analytische Psychotherapie, Psychoanalyse und tiefenpsychologisch fundierte Psychotherapie (35).

Da es sich hierbei um "Ein-Gruppen-Designs im erweiterten Sinne" handelt, werden diese in der Kodierung mittels WBP-Kriterienkatalog dementsprechend als Ein-Gruppen-Designs behandelt. Ausnahme bildet die bereits eingeführte Studie, in der ein verfahrensinterner mit einem verfahrensexternen Vergleich kombiniert wurde (19).

In der Kategorie der verfahrensexternen Vergleiche (inklusive der Kombination mit dem verfahrensinternen Vergleich [19]) befinden auf Seiten der gemischten Störungen insgesamt fünf Studien (16.1%), wovon zwei Studien an Kurzzeittherapien und drei Studien an Langzeittherapien durchgeführt wurden (in Tabelle 7 grau hinterlegt). Diese verfahrensexternen Verglei-

che teilen sich ferner auf in Vergleiche mit etablierten Psychotherapieverfahren sowie mit solchen Verfahren, die (noch) nicht als etabliert betrachtet werden können: In vier Studien wurde ein Vergleich mit einem bereits etablierten Treatment vorgenommen. Vergleichsbehandlungen waren Verhaltenstherapie (8), systemische Therapie (19) sowie kognitiv-behaviorale Therapie und Gesprächspsychotherapie (36, 37) (vgl. Anhang G). In einer Studie (38) wurde dem psychodynamischen Arm ein nicht etabliertes Treatment – Psychodramatherapie – gegenübergestellt.

Auf Seiten der gemischten Störungen werden keine Vergleiche mit pharmakologischer Monotherapie durchgeführt.

Im Vergleich zu den durchweg prospektiven Studien zum affektiven Störungsbereich, wurden bei den gemischten Störungen 13 Studien retrospektiv durchgeführt (vgl. Tabelle 6, S. 211). Von diesen 13 wurden wiederum sieben Studien (53.8%) in Ein-Gruppen-Designs realisiert. Ein Blick in diese Studien verrät, dass diese Studien sich aus den einzigen zwei reinen Katamnesestudien (17, 23) im hiesigen Studienpool zusammensetzen sowie aus vier Studien, in denen Daten ausgewertet wurden, die mit hoher Wahrscheinlichkeit im Rahmen der regulären klinischen Basisdokumentation erhoben wurden (7, 10, 24, 29). In einer weiteren Studie (25) wurden audiografierte Therapiesitzungen aus der *Penn Psychoanalytic Treatment Collection* im Hinblick auf Outcomes ausgewertet.

Ebenfalls als Ein-Gruppen-Designs erweisen sich sieben (38.9%) der 18 prospektiven Studien an gemischten Störungsgruppen (vgl. Tabelle 6, S. 211). Was das primäre Heranziehen von regulären Basisdokumentationsdaten zum Zwecke gruppenstatistischer Wirksamkeitsevaluationen betrifft, lässt sich bei den prospektiven Ein-Gruppen-Design-Studien dieses Bild jedoch nicht aufzeigen.

Bevor auf multivariatem Wege unterschiedliche Studientypen eruiert werden (vgl. Kap. 4.2), sollen noch die Gruppenzuweisungsstrategien (randomisiert *versus* selbstzugeteilt) und die innerhalb der Studien realisierten Messzeitpunkte und Katamnesezeiträume näher betrachtet werden.

Tabelle 8 gibt einen Überblick über die unterschiedlichen Zuweisungsmodi, die in vier Gruppen unterteilt wurden:

- Randomisierung
- teilweise Randomisierung und teilweise Selbstzuteilung
- nur Selbstzuteilung, jedoch unter Anwendung von Strategien, wie Parallelisierung, Stratifizierung, Matching
- Selbstzuteilung ohne eine der genannten Strategien.

Tabelle 8: Gruppenzuweisungsstrategien in den Primärstudien (*n*=23)

	Affektive Störungen (*n*=6)		**Gemischte Störungen (*n*=17)**	
Gruppenzuweisung *n* (%)	Randomisierung	5 (83.3%)	Randomisierung	6 (35.3%)
	teilweise Random./ teilweise Selbstzuteilung	1 (16.7%)	teilweise Random./ teilweise Selbstzuteilung	-
	Selbstzuteilung „mit“	-	Selbstzuteilung „mit“	1 (5.9%)
	Selbstzuteilung „ohne“	-	Selbstzuteilung „ohne“	8 (47.1%)
			echte WL	2 (11.8%)
Gesamt %		**100%**		**100%**

Anmerkung: Random.: Randomisierung, Selbstzuteilung „mit“: Selbstzuteilung unter Anwendung von Strategien, wie Parallelisierung, Stratifizierung, Matching; Selbstzuteilung „ohne“: Selbstzuteilung ohne Anwendung von Strategien, wie Parallelisierung, Stratifizierung, Matching; WL: Warteliste.

Zusätzlich zu den regulären Zuweisungsmodi musste eine weitere Kategorie aufgenommen werden, die sich „echte Warteliste“ nennt. Hiermit sind Studien gemeint, in denen Wartelisten nicht zu Studienzwecken geführt werden, sondern in denen im Rahmen natürlicher Settings Wartelisten real existieren. Solchen Listen werden Patienten in der Regel weder per Zufall

zugeteilt, noch entscheiden sich Patienten bewusst für diese Option. Aus diesem Grund wurde diese Studiengruppe, die zwei Untersuchungen enthält (2, 18), gesondert aufgeführt.

Da bei Ein-Gruppen-Designs keine Zuweisungsstrategien bestimmbar sind, wurde auf die Aufnahme der anwendungsbereichsübergreifenden 18 Ein-Gruppen-Design-Studien in die Tabelle 8 verzichtet.

Auf Seiten der sechs Studien zu affektiven Störungen nahmen insgesamt fünf Studien randomisierte Gruppenzuweisungen vor (9, 21, 27, 32, 40), in einer Studie erfolgte die Gruppenzuweisung teilweise randomisiert und teilweise via Selbstzuteilung (39). Vier der fünf randomisierten Studien untersuchten Kurzzeittherapien, eine Studie untersuchte Behandlungen von mittlerer Länge (21). Die Studie mit beiden Zuweisungsmodi (39) bezieht sich auf Kurzzeittherapien. In keiner Studie erfolgte die Gruppenbildung ausschließlich über Selbstzuteilung.

Auf Seiten der gemischten Störungen zeigt sich folgendes Bild: Von den insgesamt 17 Studien erfolgte in sechs Studien die Zuweisung randomisiert (12, 15, 19, 31, 34, 41), davon wurden in einer Studie Langzeittherapien untersucht (19), bei dem Rest Kurzzeittherapien oder Therapien mittlerer Länge. In keiner Studie erfolgte die Zuweisung mittels beider Zuweisungsmodi, dafür konnten sich im Rahmen von neun Studien die Patienten selbst zuweisen. Diese neun Studien gliedern sich nochmals in eine Studie, in der Strategien zur Gruppenangleichung eingesetzt wurden (11) und acht Studien, in denen auf solche Strategien verzichtet wurde (3, 4, 5, 8, 35, 36, 37, 38). Außerdem beziehen sich sieben der neun Studien mit Selbstzuteilung auf Langzeittherapien (3, 4, 5, 8, 11, 35, 38) und zwei Studien auf Kurzzeittherapien (36, 37).

Tabelle 9 gibt Aufschluss über die Messzeitpunkte, die in den unterschiedlichen Untersuchungen realisiert wurden.

Tabelle 9: Realisierte Messzeitpunkte in den Primärstudien (*N*=41)

	Affektive Störungen (*n*=10)		**Gemischte Störungen (*n*=31)**	
Messzeitpunkte *n* (%)	Prä-Post	5 (50.0%)	Prä-Post	13 (41.9%)
	Prä-Katamnese	-	Prä-Katamnese	5 (26.1%)
	Prä-Post-Katamnese	5 (50.0%)	Prä-Post-Katamnese	9 (29.0%)
	nur Post	-	nur Post	1 (3.2%)
	nur Katamnese	-	nur Katamnese	2 (6.5%)
	nicht beurteilbar	-	*nicht beurteilbar*	1 (3.2%)
Gesamt %		**100%**		**100%**

Mit 18 (43.9%) von 41 Studien war das Prä-Post-Design anwendungsbereichsübergreifend am häufigsten besetzt, gefolgt von 14 Studien (34.1%) mit einem Prä-Post-Katamnese-Design. Das Prä-Post-Katamnese-Design liegt auf Seiten der gemischten Störungen bei neun von 31 Studien (29%) vor, zudem sind bei diesem Anwendungsbereich noch fünf Studien (26.1%) zu finden, in denen die Outcomes zu Beginn der Behandlung erhoben wurden und dann erst wieder zu einem Zeitpunkt, zu dem die Patienten bereits seit einer gewissen Zeit ihre Therapie beendet hatten (Prä-Katamnese[60]) (3, 6, 8, 18, 19). Solche Prä-Katamnesemessungen wurden ausschließlich auf Langzeittherapiestudien angewendet, gleiches trifft auf die drei Studien mit nur einem Messzeitpunkt (nur Post- oder Katamnesemessung) zu (17, 23, 35).

Auf Seiten der affektiven Störungen kommt das Prä-Katamnese-Messdesign bei keiner Studie vor.

Ferner gibt es eine Studie, bei der die Einordnung in Bezug auf die Messzeitpunkte nicht eindeutig möglich war, da die Studie Querschnitt- mit Längsschnittdaten kombiniert (5). In die-

[60] In Kapitel 3.2.3 wurde das Phänomen der Prä-Katamnesemessungen bereits eingeführt.

ser Studie wurden Gruppen, die sich in unterschiedlichen Phasen ihrer Therapie befanden, über insgesamt drei Erhebungswellen hinweg befragt, so dass es pro Patientengruppe drei Messzeitpunkte gab (z.B. „lange vor der Therapie – unmittelbar vor Therapiebeginn – früh während der Therapie“ oder „während der Therapie – spät während der Therapie – kurz nach der Therapie“ etc.). Damit können über den gesamten Zeitraum von „lange vor der Therapie“ bis „lange nach der Therapie“ in erster Linie Trendaussagen getroffen werden.

Im Hinblick auf die in den Studien gewählten Katamnesezeiträume soll abschließend noch ein Vergleich zwischen den Messzeitpunkt-Designs (Prä-Post-Katamnese, Prä-Katamnese, nur Katamnese) vorgenommen werden (vgl. Tabelle 10).

Tabelle 10: Katamnesezeiträume in den Primärstudien (*n*=21)

		Katamnesezeiträume (Range und Mittelwert)
Prä-Post-Katamnese (*n*=14)	Range	3 Monate – 4.5 Jahre
	M	1.4 Jahre
Prä-Katamnese (*n*=5)	Range	6 Monate – 4 Jahre
	M	2.1 Jahre
nur Katamnese (*n*=2)	Range	6 Jahre – 6.5 Jahre
	M	6.3 Jahre

Anmerkung: M: Mittelwert.

Der Range der Katammesezeiträume von 3 Monaten bis zu 4.5 Jahren bei den Prä-Post-Katamnese-Designs begründet die größte Spannweite, gefolgt vom Range der Katamnesezeiträume bei den Prä-Katamnese-Designs (6 Monate bis 4 Jahre).

Die Mittelwerte zeigen, dass die längsten Katamnesezeiträume bei den reinen Katamnesestudien zu finden sind (*M* 6.3 Jahre), gefolgt von der mittleren Katamnesedauer bei den Prä-Katamnesestudien von ca. 2 Jahren. Bei den zuletzt genannten Studien überwiegt anteilig

ein Katamnesezeitraum um die 2 Jahre, bei den Prä-Post-Katamnesestudien überwiegen hingegen Katamnesezeiträume zwischen 6 Monaten und 1.5 Jahren.

Im nun folgenden Abschnitt werden die wichtigsten Ergebnisse dieses ersten allgemeinen Überblicks über die Studien noch einmal zusammengefasst.

Zusammenfassung

Es fällt auf, dass auf Seiten der Studien zu den affektiven Störungen nahezu ausschließlich tiefenpsychologisch fundierte Behandlungen untersucht wurden, was zu einer Dominanz von Kurzzeittherapiestudien (60.0%), gefolgt von Studien zu mittleren Therapieumfängen (30.0%) in diesem Anwendungsbereich geführt hat. Im Gegensatz dazu dominieren auf Seiten der gemischten Störungen die Langzeittherapiestudien (51.6%), wobei die Kurzzeittherapiestudien immerhin mit einem Anteil von 29.0% vertreten sind. Insofern ist bei den gemischten Störungsgruppen-Studien eine größere Heterogenität an untersuchten psychodynamischen Verfahren vorzufinden (tiefenpsychologische fundierte und analytische Psychotherapie sowie Psychoanalyse).

Im Hinblick auf die in den Studien realisierten Studiendesigns ist die Kategorie der „Ein-Gruppen-Designs“ in beiden Anwendungsbereichen jeweils am höchsten besetzt (40.0% bzw. 45.2%). Innerhalb dieser Kategorie dominieren, über beide Anwendungsbereiche hinweg, die Studien zu mittellangen (über 25 bis 100 Sitzungen) und langen Behandlungen (> 100 Sitzungen).

Studien in Kontrollgruppendesigns bilden auf Seiten der affektiven Störungen die zweitgrößte Studiendesignkategorie (40.0%), in der sich allein Studien zu kurzen und mittellangen Therapien befinden. Bei den gemischten Störungen machen Studien, in denen klassi-

sche Kontrollgruppen (Wartelisten, TAU etc.) umgesetzt wurden, einen weit geringeren Anteil von 19.4% aus. Hier sind sowohl Studien zu kurzen als auch zu mittellangen und langen Behandlungen zu finden. Bei genauerer Betrachtung der Kontrollgruppenrealisationen innerhalb von Studien zu mittellangen (> 25 Sitzungen) und langen Therapien (> 100 Sitzungen) (beide Anwendungsbereiche) zeigt sich bei drei der vier Studien, dass die jeweiligen Wartelistenbedingungen lediglich einen Teil des gesamten Behandlungszeitraums umfassen. Bei lediglich einer Studie (3) wird die Kontrollgruppe (unbehandelte Gruppe) zum Prä-Zeitpunkt und nach 3 Jahren wiederbefragt. Dieser 3-Jahreszeitraum entspricht zumindest in etwa der mittleren Behandlungsdauer von 3 Jahren (Range 1 bis 5.5 Jahre).

Die Kategorie der verfahrensexternen Vergleiche (auch in Kombination mit Kontrollgruppen oder verfahrensinternen Vergleichsgruppen) ist in beiden Anwendungsbereichen vergleichsweise niedrig besetzt. Auf Seiten der affektiven Störungen befindet sich lediglich eine von insgesamt 10 Studien mit einem verfahrensexternen Vergleich (plus Kontrollgruppe). Bei den gemischten Störungen trifft dies auf fünf (16.1%) der 31 Studien zu. Vor dem Hintergrund der angeführten Schwierigkeit, die sich im Hinblick auf klassische Kontrollgruppen bei Studien zu längerfristigen Behandlungen ergibt, wäre zu erwarten gewesen, dass vor allem im Rahmen von Untersuchungen zu längerfristigen Therapien vermehrt auf verfahrensexterne Vergleiche zurückgegriffen wird.

Studien im sog. Ein-Messzeitpunkt-Design (nur Post- oder Katamnesemessung) befinden sich ausgenommen auf Seiten der gemischten Störungen. Zudem wurde dieses Messdesign allein im Rahmen von Langzeittherapiestudien realisiert. Gleiches trifft auf Prä-Katamnesemessungen zu. Im Gegensatz dazu wurden bei den Studien zu kurzen (bis 25 Sit-

zungen) und mittellangen Behandlungen (über 25 bis 100 Sitzungen) allein Prä-Post- bzw. Prä-Post-Katamnesemessungen durchgeführt.

Um einen möglichst kompakten Überblick über die bereits dargestellten in Kombination mit den restlichen Variablen des Kurzkodierbogens (Anhang B und D) zu erhalten, wurde im folgenden Kapitel (Kap. 4.2) der Studienpool daraufhin untersucht, ob bestimmte Studientypen im Hinblick auf methodologische Charakteristika eruiert werden können und welche Variablen dafür ggf. ausschlaggebend sind.

4.2 Clusterstruktur der Primärstudien

Um Studientypen zu identifizieren, durch die sich die einzelnen Studien in methodologischer Hinsicht charakterisieren lassen, wurden die Studien clusteranalytisch untersucht. Mit Hilfe einer Two-Step Clusteranalyse sollte zunächst eruiert werden, durch wie viele Cluster sich die Studien bestenfalls repräsentieren lassen. Im Anschluss wurde das Ergebnis der Two-Step Clusteranalyse mittels einer hierarchischen Clusteranalyse evaluiert, indem die Clusteranzahl der Two-Step Clusteranalyse als Vorgabe für das hierarchische Verfahren genutzt wurde.

In die clusteranalytischen Auswertungen wurden folgende Variablen aus dem Kurzkodierbogen (Anhang B und D) einbezogen:

- Therapieumfang (Kurzzeittherapie bis 25 Sitzungen, Therapien moderater Länge von über 25 bis 100 Sitzungen, Langzeittherapie über 100 Sitzungen)
- prospektiver *versus* retrospektiver Datenzugang
- Strategien der Gruppenzuweisung
- a priori Festlegung der Therapiesitzungsanzahl
- Störungsspezifität der Behandlung

- Einsatz von Manualen oder Behandlungsrichtlinien
- Training der Therapeuten zwecks Studie
- Implementationskontrolle (Adherence).

Die acht Kategorien, die sich im Kurzkodierbogen auf die „Strategien der Gruppenzuweisung“ beziehen, wurden zum Zwecke der Clusteranalysen zu vier Kategorien zusammengefasst:

- Randomisierte Zuteilungen
- Selbstzuweisung
- verfahrensinterne Vergleiche (randomisiert oder via Selbstzuweisung)
- Ein-Gruppen-Designs.

Für die Durchführung der hierarchischen Clusteranalyse mussten die kategorialen Variablen zunächst via Dummy-Kodierung in binäre Variablen transformiert werden. Um die Anzahl der dabei entstehenden Variablen möglichst gering zu halten, mussten insgesamt zwei Studien exkludiert werden (21, 30): Beide Studien stellen in der Variablen „a priori Festlegung der Therapiesitzungsanzahl“ eine Art Ausreißer da, die eine Studie war in dieser Variablen nicht beurteilbar (21), in der anderen Studie wurde für einen Teil der Patienten der Therapieumfang a priori festgelegt, für den anderen Teil jedoch nicht (30). Die Variable hätte unter Aufnahme beider Studien zu drei, statt nur zu einer dummy-kodierten Variablen geführt, zudem wären zwei Dummy-Variablen als Konstanten in die Clusteranalyse eingegangen, womit eine Voraussetzung für die Clusteranalyse verletzt gewesen wäre (vgl. Backhaus, Erichson, Plinke & Weiber, 2000; Schendera, 2010).

Um in beiden clusteranalytischen Verfahren dieselbe Anzahl an Variablen aufzunehmen, bildeten die binären (dummy-kodierten) Variablen (s.o.) ebenfalls die Basis für die Two-Step Clusteranalyse. Als Voraussetzung für die Durchführung beider o.g. clusteranalytischer Verfahren gilt, dass die Variablen nicht hochkorrelierend sein sollten (> .9; vgl. Backhaus et al., 2000). Aus diesem Grund wurden die Variablen zunächst auf ihre Zusammenhänge überprüft (*Phi Koeffizient*). Dabei erwiesen sich alle Zusammenhänge innerhalb der genannten Grenze, der größte Zusammenhang beläuft sich auf φ (Phi) .68. Zudem gilt für Two-Step Clusteranalysen die Voraussetzung möglichst multinomialer Verteilungen kategorialer Variablen, d.h. binomialer Verteilungen dichotomer Variablen. Inwieweit diese Voraussetzung erfüllt ist, wurde mittels der Effektgröße γ (Gamma) eruiert (vgl. Eid et al., 2010). Dabei zeigte sich bei acht der 11 dummy-kodierten Variablen eine Effektgröße von $|\gamma| > .25$ und somit große Effekte, was einer Abweichung von der Gleichverteilung entspricht (vgl. Eid et al., 2010, S. 279). Brosius (2011) weist jedoch darauf hin, dass die Two-Step Clusteranalyse gegenüber Verletzungen dieser Voraussetzung als robust gilt und sich diese lediglich unkritisch auf die Ergebnisse auswirken. Aus diesem Grund wird die Verletzung dieser Voraussetzung in Kauf genommen.

Der Forderung, die Anzahl an Ausreißern für die hierarchische Clusteranalyse – vor allem bei Anwendung der Clustermethode *Ward* (s.u.) – möglichst gering zu halten, wird durch die vorgeschaltete Two-Step Clusteranalyse entsprochen, da durch diese Ausreißer separat ausgegeben werden, die infolgedessen zwecks hierarchischer Clusteranalyse ausgeschlossen werden können (vgl. Janssen & Laatz, 2007). Unter Ausreißern sind dabei Objekte (Studien) zu verstehen, die ein derartiges Merkmalsprofil aufweisen, dass sie mit den restlichen Objekten nicht kompatibel sind – sprich: mit diesen keine homogenen Cluster bilden können (vgl. Backhaus et al., 2000).

Im Rahmen der Two-Step Clusteranalyse wurde zur Modellierung der Ähnlichkeit/Distanz zwischen den Objekten die Log-Likelihood Distanz verwendet. Die Clusteranzahl wurde automatisch ermittelt (d.h. es wurde keine feste Anzahl vorgegeben), bei der Auswahl der Clustermethode wurde das Bayes-Informationskriterium (BIC) gewählt (vgl. Janssen & Laatz, 2007). Die acht o.g. Variablen wurden in Form von 11 dummy-kodierten Variablen ins Modell aufgenommen.

Das Ergebnis der Two-Step Clusteranalyse ergab als optimale Lösung eine 2-Clusterlösung, in der drei (7.7%) der 39 Studien als Ausreißer ausgegeben wurden (14, 15, 29). Die restlichen 36 Studien belaufen sich auf Cluster 1, bestehend aus 25 Studien (64.1%), und Cluster 2, bestehend aus 11 Studien (28.2%) (vgl. Abbildung 23). Anhang H sind die Studien samt ihren Clusterzugehörigkeiten zu entnehmen.

Die Modellgüte (Qualität der Cluster) ist mit einem Silhouettenkoeffizienten (Wertebereich -1 bis 1) mit einem Wert von .5 als zufriedenstellend zu betrachten.

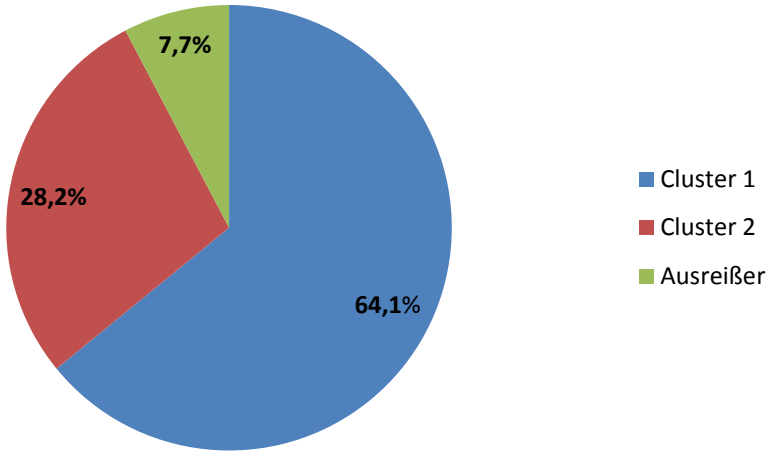


Abbildung 23: 2-Clusterlösung plus Ausreißerkategorie der Two-Step Clusteranalyse (*n*=39)

Da die Two-Step Clusteranalyse zunächst als eine erste Orientierung hinsichtlich optimaler Clusterlösungen dienen sollte, wird auf eine ausführliche Darstellung der Clusterprofile an dieser Stelle verzichtet und sich zuvor der hierarchischen Clusteranalyse zugewendet, mit Hilfe derer die Ergebnisse der Two-Step Clusteranalyse repliziert werden sollten.

Aus den unterschiedlichen Cluster-Algorithmen der hierarchischen Verfahren wurde das *Ward*-Verfahren ausgewählt, das sich für die Bildung möglichst homogener Cluster optimal eignet und von Backhaus et al. (2000) als Algorithmus bezeichnet wird, der „in den meisten Fällen *sehr gute Partitionen* findet und die Elemente "*richtig*" den Gruppen zuordnet“ (S. 366). Als Proximitätsmaß wurde die für binäre Daten geeignete quadrierte euklidische Distanz verwendet. In Anlehnung an die Clusterlösung der Two-Step Clusteranalyse wurde nun die Clusteranzahl mit zwei Clustern vorgegeben.

Da durch das *Ward*-Verfahren keine Ausreißer aufgedeckt werden und daher empfohlen wird, diese vorher zu eliminieren (vgl. Backhaus et al., 2000), wurden in die hierarchische Clusteranalyse die 41 Studien, abzüglich der drei o.g. Ausreißer-Studien und den beiden zu Beginn exkludierten Studien, aufgenommen (n=36).

In Tabelle 11 ist die Entwicklung der Fehlerquadratsumme unter Anwendung des *Ward*-Verfahrens wiedergegeben. Spalte 1 („Schritt“) ist der jeweilige Fusionierungsschritt zu entnehmen. Insgesamt wurden 35 (36-1) Fusionierungsschritte durchgeführt. Die nächste übergeordnete Spalte („Zusammengeführte Cluster“) stellt dar, welche Fälle bzw. Cluster fusioniert wurden. Bspw. wurden in 3. Schritt die beiden Studien 17 und 23 fusioniert. Bereits im 4. Schritt steht die „17“ (dritte Spalte) jedoch nicht mehr nur für „Studie 17“, sondern für das beim 3. Schritt entstandene Cluster, bestehend aus Studie 17 und 23. Der Spalte „Erstes Vorkommen des Clusters“ ist zeilenweise zu entnehmen, bei welchem Schritt ein bestimmtes

Cluster zum ersten Mal vorkommt. So bedeutet die „3“ in Zeile 4 (sechste Spalte), dass das Cluster 17 bereits in Schritt 3 einmal vorkam. Der letzten Spalte („Nächster Schritt“) ist zu entnehmen, bei welchem Schritt ein jeweils neu gebildetes Cluster erneut vorkommt. Der Spalte „Koeffizienten“ ist nun der Fehlerquadratzuwachs über die 35 Fusionierungsschritte zu entnehmen. Da am Anfang alle 36 Studien für jeweils ein „Cluster“ stehen und in insoweit keine Heterogenität innerhalb der Cluster bestehen kann, ergibt sich der Fehlerquadratzuwachs und damit der Zuwachs an Heterogenität erst mit Vergrößerung der Clusteranzahl. Beim letzten Schritt (hier: Schritt 35) werden alle Objekte zu einem Cluster zusammengeführt und man hat mit der höchsten Fehlervarianz zu rechnen.

Tabelle 11: Clusterzusammenfassung und Zunahme der Heterogenität beim *Ward*-Verfahren

Schritt	Zusammengeführte Cluster		Koeffizienten	Erstes Vorkommen des Clusters		Nächster Schritt
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	36	37	.000	0	0	15
2	35	5	.000	0	0	24
3	17	23	.000	0	0	4
4	10	17	.000	0	3	8
5	22	16	.000	0	0	6
6	6	22	.000	0	5	7
7	6	26	.000	6	0	17
8	10	25	.000	4	0	28
9	3	18	.000	0	0	10
10	38	3	.000	0	9	16
11	12	13	.500	0	0	18
12	32	40	1.000	0	0	19
13	7	24	1.500	0	0	20
14	4	11	2.000	0	0	24
15	2	36	2.667	0	1	33
16	38	8	3.417	10	0	30
17	6	19	4.217	7	0	28
18	39	12	5.050	0	11	25
19	32	28	5.883	12	0	29
20	20	7	6.717	0	13	27
21	27	9	7.717	0	0	22
22	34	27	8.717	0	21	25
23	41	33	9.717	0	0	27
24	4	35	10.967	14	2	31
25	39	34	12.300	18	22	32
26	31	1	13.800	0	0	29
27	41	20	15.467	23	20	34
28	10	6	17.778	8	17	30
29	32	31	20.944	19	26	32
30	38	10	24.468	16	28	31
31	38	4	28.980	30	24	33
32	32	39	33.768	29	25	35
33	38	2	39.205	31	15	34
34	41	38	45.775	27	33	35
35	41	32	68.139	34	32	0

In Abbildung 24 ist die Entwicklung der Fehlerquadratsumme über die 35 Fusionierungsschritte grafisch dargestellt. Der größte Heterogenitätszuwachs ist zwischen Schritt 34 und 35 zu erkennen (von 45.775 auf 68.139; vgl. Tabelle 11), was für eine 2-Clusterlösung spricht.

Grafisch ist dies am sog. *Elbow*-Kriterium – einem sichtbaren Knick im Kurvenverlauf – zu erkennen (in Abbildung 24 mit einem Pfeil versehen).

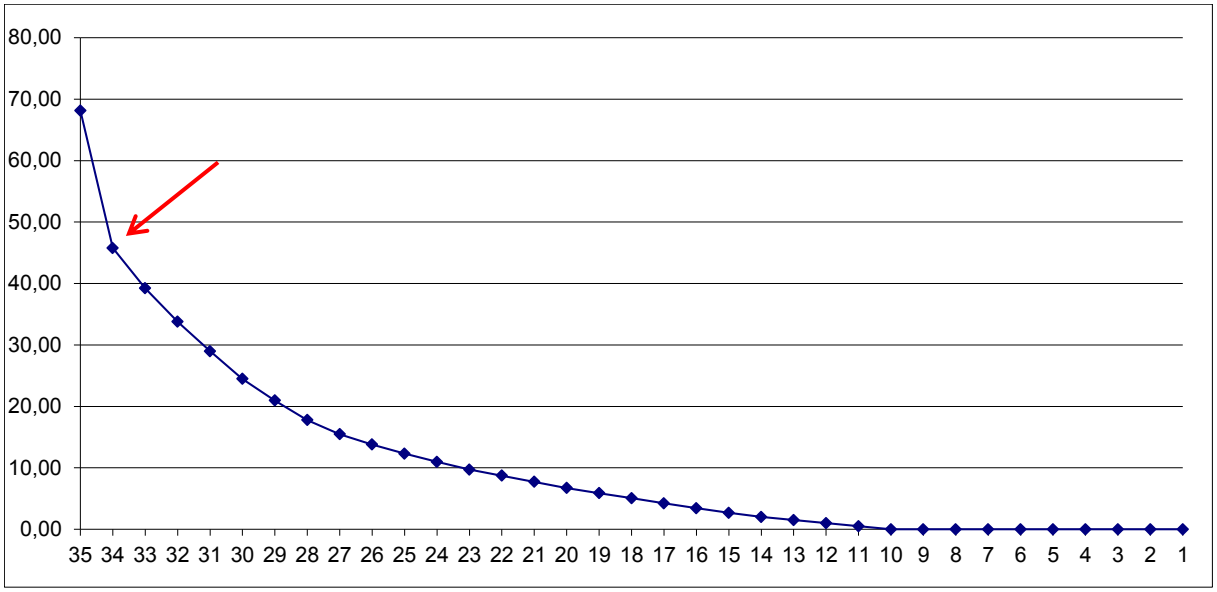


Abbildung 24: Scree-Plot: Entwicklung der Fehlerquadratsumme

Inwiefern die hierarchische Clusteranalyse die Ergebnisse der Two-Step Clusteranalyse tatsächlich repliziert, ist Tabelle 12 zu entnehmen. Es zeigte sich, dass die Cluster, was die Zuordnung der Studien in die jeweiligen Cluster betrifft, durchweg übereinstimmen. Anhang H sind die Clusterzugehörigkeiten der einzelnen Studien aus der hierarchischen Clusteranalyse zu entnehmen.

Tabelle 12: Kombination der Ergebnisse aus der Two-Step Clusteranalyse und der hierarchischen Clusteranalyse (2-Clusterlösung) (*n*=36)

		Two-Step CA		
		Cluster 1	Cluster 2	Gesamt
Hierarchische CA (2-Clusterlösung)	Cluster 1	25	0	25
	Cluster 2	0	11	11
	Gesamt	25	11	36

Anmerkung: CA: Clusteranalyse.

Backhaus et al. (2000) weisen kritisch darauf hin, dass beim Übergang von der 2- zur 1-Clusterlösung stets mit einem vergleichsweise großen Heterogenitätssprung zu rechnen ist. Dieser kritische Hinweis, zusammengenommen mit der Tatsache, dass es sich bei dem ersten Cluster um ein relativ großes Cluster handelt (*n*=25), wurde zum Anlass genommen, die Fusionierung der Studien bzw. Cluster einmal probehalber mit Schritt 33 abzuschließen und damit eine quadrierte Fehlervarianzsumme von 39.205 in Kauf zu nehmen und eine 3-Clusterlösung anzunehmen. Auch, wenn es sich hierbei – von Schritt 33 zu 34 – längst nicht mehr um einen derart deutlichen Heterogenitätszuwachs handelt (von 39.205 auf 45.775; vgl. Tabelle 11), soll eruiert werden, ob das erste Cluster sich durch diese Vorgabe inhaltlich plausibel in zwei Cluster aufteilen lässt. Tabelle 13 ist zu entnehmen, dass sich Cluster 1 (Two-Step Clusteranalyse) in zwei Cluster (Cluster 1 und 2; hierarchische Clusteranalyse 3-Clusterlösung) aufteilen lässt, bestehend aus 5 und 20 Studien.

Tabelle 13: Kombination der Ergebnisse aus der Two-Step Clusteranalyse und der hierarchischen Clusteranalyse (3-Clusterlösung) (*n*=36)

		Two-Step CA		
		Cluster 1	Cluster 2	Gesamt
Hierarchische CA (3-Clusterlösung)	Cluster 1	5	0	5
	Cluster 2	20	0	20
	Cluster 3	0	11	11
	Gesamt	25	11	36

Anmerkung: CA: Clusteranalyse.

Inwieweit die Teilung des großen Clusters in zwei kleinere Cluster tatsächlich zu inhaltlich plausiblen Separationen führt, soll in Bezug auf die clusterkonstituierenden Variablen der 2-Clusterlösung eruiert werden. Der Two-Step Clusteranalyse konnte entnommen werden, dass der Therapieumfang (Kurzzeittherapie, mittelfristige Therapie, Langzeittherapie) als zentrale clusterkonstituierende Variable fungiert. Die Verteilung dieser Variablen ist Tabelle 14 für die 2-Clusterlösung und Tabelle 15 für die 3-Clusterlösung zu entnehmen.

Tabelle 14: Therapieumfänge in den Primärstudien (2-Clusterlösung) (*n*=36)

	Cluster 1 (*n*=25)		**Cluster 2** (*n*=11)	
Therapieumfang *n* (%)	KZT	3 (12.0%)	KZT	11 (100%)
	MfT	5 (20.0%)	MfT	-
	LZT	17 (68.0%)	LZT	-

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie.

Dabei zeigt sich, dass es sich bei Cluster 1 (2-Clusterlösung; vgl. Tabelle 14) um ein noch relativ heterogenes Cluster handelt, da alle drei Therapieumfangsoptionen (Kurzzeittherapie, mittelfristige Therapie, Langzeittherapie) vorkommen. Demgegenüber entzerren Cluster 1 und 2 (3-Clusterlösung, vgl. Tabelle 15) diese Heterogenität, indem die Variablenausprägung „mittelfristige Therapielänge“ nunmehr ein eigenes Cluster (Cluster 1) bildet und Cluster 2 nur noch aus Kurzzeittherapiestudien (15.0%) und Langzeittherapiestudien (85.0%) besteht.

Tabelle 15: Therapieumfänge in den Primärstudien (3-Clusterlösung) (*n*=36)

	Cluster 1 (*n*=5)		**Cluster 2** (*n*=20)		**Cluster 3** (*n*=11)	
Therapieumfang *n* (%)	KZT	-	KZT	3 (15.0%)	KZT	11 (100%)
	MfT	5 (100%)	MfT	-	MfT	-
	LZT	-	LZT	17 (85.0%)	LZT	-

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie.

Da die 3-Clusterlösung in Bezug auf die clusterkonstituierende Variable „Therapieumfang“ und deren Kategorien „Kurzzeittherapie, mittelfristige Therapie, Langzeittherapie“ zu einem höheren Differenziertheitsgrad führt, wurde sich für die weitere Auswertung für diese Clusterlösung entschieden. Die Clusterzugehörigkeiten der einzelnen Studien im Hinblick auf die 3-Clusterlösung sind Anhang H zu entnehmen. Abbildung 25 stellt die prozentualen Anteile der drei Cluster dar und bezieht sich auf die 36 Studien, die in die hierarchische Clusteranalyse aufgenommen wurden (d.h. exklusive der Ausreißerkategorie aus der Two-Step Clusteranalyse sowie der zwei anfangs exkludierten Studien).

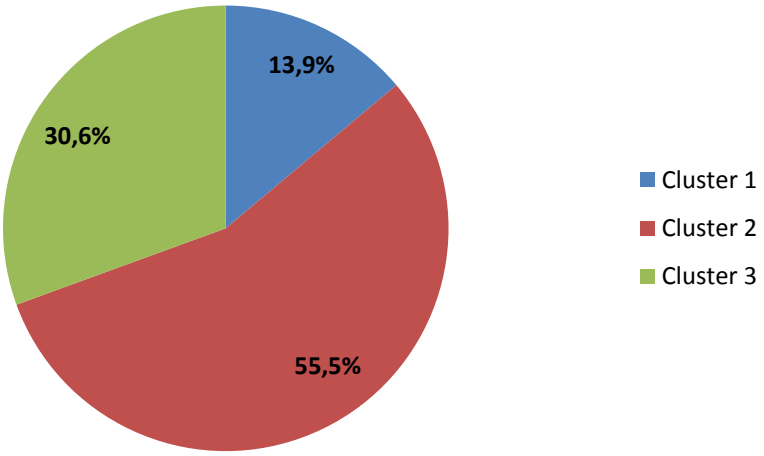


Abbildung 25: 3-Clusterlösung nach hierarchischer Clusteranalyse (*n*=36)

Die Charakterisierung der drei Cluster wird sowohl anhand der Verteilung derjenigen Variablen vorgenommen, die in die Clusteranalyse aufgenommen wurden (s.o.) als auch anhand externer Variablen, die bereits in Kapitel 4.1 dargestellt wurden. Tabelle 16 bis Tabelle 19 geben die jeweiligen Profile der drei Cluster in Form von absoluten und prozentualen Häufigkeiten der unterschiedlichen Variablenkategorien wieder.

Tabelle 16 zeigt, dass bei den fünf Studien des Cluster 1 die Studien an gemischten Störungsgruppen mit vier zu einer Studie überwiegen, in Cluster 2 verstärkt sich dieses Verhältnis mit 19 (95.0%) von insgesamt 20 Studien an gemischten Störungsgruppen. In Cluster 3 sind beide Anwendungsbereiche annähernd gleich vertreten mit einem geringfügigen Überhang an Studien zu affektiven Störungen (54.5%).

Tabelle 16: Charakterisierung der drei Cluster nach Anwendungsbereich, Therapieverfahren, Therapieumfang und Therapiesetting (*n*=36)

	Cluster 1 (*n*=5)			**Cluster 2** (*n*=20)			**Cluster 3** (*n*=11)		
Anwendungsbereich *n* (%)	AS	1	(20.0%)	AS	1	(5.0%)	AS	6	(54.5%)
	GS	4	(80.0%)	GS	19	(95.0%)	GS	5	(45.5%)
	Gesamt		**100%**	**Gesamt**		**100%**	**Gesamt**		**100%**
Psychoanalytisch begründetes Therapieverfahren (plus Psa) *n* (%)	TP	5	(100%)	TP	1	(5.0%)	TP	11	(100%)
	AP	-		AP	6	(30.0%)	AP	-	
	Psa	-		Psa	1	(5.0%)	Psa	-	
	TP und AP	-		TP und AP	7	(35.0%)	TP und AP	-	
	AP und Psa	-		AP und Psa	3	(15.0%)	AP und Psa	-	
	TP, AP und Psa	-		TP, AP und Psa	1	(5.0%)	TP, AP und Psa	-	
	nicht beurteilbar	-		*nicht beurteilbar*	1	(5.0%)	*nicht beurteilbar*	-	
	Gesamt		**100%**	**Gesamt**		**100%**	**Gesamt**		**100%**
Therapieumfang *n* (%) [61]	KZT	-		KZT	3	(15.0%)	KZT	11	(100%)
	MfT	5	(100%)	MfT	-		MfT	-	
	LZT	-		LZT	17	(85.0%)	LZT	-	
	Gesamt		**100%**	**Gesamt**		**100%**	**Gesamt**		**100%**
Therapiesetting *n* (%)	Individualth.	2	(40.0%)	Individualth.	19	(95.0%)	Individualth.	9	(81.8%)
	Gruppenth.	3	(60.0%)	Gruppenth.	1	(5.0%)	Gruppenth.	2	(18.2%)
	Gesamt		**100%**	**Gesamt**		**100%**	**Gesamt**		**100%**

Anmerkung: AP: Analytische Psychotherapie, AS: Affektive Störungen, Gruppenth.: Gruppentherapie, GS: Gemischte Störungen, Individualth.: Individualtherapie, KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie, Psa: Psychoanalyse, TP: Tiefenpsychologisch fundierte Psychotherapie.

In Bezug auf das psychoanalytisch begründete Verfahren, das in den jeweiligen Studien untersucht wurde, setzen sich Cluster 1 und 3 durchweg aus Studien zusammen, die tiefenpsychologisch fundierte Psychotherapie untersuchen, in Cluster 2 ist es hingegen lediglich eine Studie, die sich allein auf dieses Verfahren bezieht, die restlichen 19 Studien verteilen sich auf die übrigen Verfahren und Verfahrenskombinationen. Dieses Verhältnis spiegelt sich dementsprechend in der Rubrik „Therapieumfang" wider: In den beiden „Tiefenpsychologisch fundierte Psychotherapie-Clustern" finden sich ausschließlich Studien zu mittelfristigen

[61] Damit die Tabellen einen Gesamtüberblick gewähren, wurde die bereits tabellarisch dargestellte Variable „Therapieumfang" in Tabelle 16 nochmals mitaufgenommen.

Therapien (Cluster 1) sowie Kurzzeittherapiestudien (Cluster 3), wohingegen in Cluster 2 mit 17 Studien (85.0%) die Langzeittherapiestudien überwiegen. Bei den drei Kurzzeittherapiestudien in Cluster 2 handelt es sich um die eine Untersuchung an tiefenpsychologisch fundierter Psychotherapie und zwei Untersuchungen, in die sowohl tiefenpsychologisch fundierte als auch analytische Behandlungen eingehen, die mittlere Behandlungszeit jedoch lediglich um die acht Sitzungen umfasste (36, 37)[62].

Die unterschiedlichen Behandlungssettings (Individual- *versus* Gruppentherapie) verteilen sich folgendermaßen: Es befinden sich im „mittelfristigen Therapie-Cluster" (Cluster 1), im Vergleich zu den beiden anderen Clustern, mit drei Studien (60.0%) anteilig die meisten Gruppentherapiestudien; im „Langzeittherapie-Cluster" (Cluster 2) dominieren mit 19 Studien (95.0%) die Untersuchungen zu Individualtherapien. Ein ähnliches Verhältnis, jedoch in leicht abgeschwächter Form, zeigt sich mit neun Studien (81.8%) zu Individualtherapie im „Kurzzeittherapie-Cluster" (Cluster 3).

Tabelle 17 gibt einen Überblick über die drei Cluster in Hinblick auf die Variablen „Studiendesign", „prospektiver *versus* retrospektiver Datenzugang", „Gruppenzuweisung" sowie „Messzeitpunkte".

[62] Auf diese beiden Studien (Originalstudie und Replikation) wurde in Kapitel 4.1 bereits ausführlich hingewiesen.

Tabelle 17: Charakterisierung der drei Cluster nach Studiendesign, Datenzugang, Gruppenzuweisungsstrategie und Messzeitpunkte (*n*=36)

	Cluster 1 (*n*=5)		**Cluster 2** (*n*=20)		**Cluster 3** (*n*=11)	
Studiendesign *n* (%)	Ein-Gruppen-Design	4 (80.0%)	Ein-Gruppen-Design	8 (40.0%)	Ein-Gruppen-Design	3 (27.3%)
	KG-Design	1 (20.0%)	KG-Design	3 (15.0%)	KG-Design	4 (36.4%)
	verfahrensint. Vergleich	-	verfahrensint. Vergleich	4 (20.0%)	verfahrensint. Vergleich	2 (18.2%)
	verfahrensext. Vergleich	-	verfahrensext. Vergleich	4 (20.0%)	verfahrensext. Vergleich	-
	Kombi verfahrensint. & verfahrensext.	-	Kombi verfahrensint. & verfahrensext.	1 (5.0%)	Kombi verfahrensint. & verfahrensext.	-
	Kombi KG-Design & verfahrensext.	-	Kombi KG-Design & verfahrensext.	-	Kombi KG-Design & verfahrensext.	1 (9.1%)
	Vergleich mit Pharmaka	-	Vergleich mit Pharmaka	-	Vergleich mit Pharmaka	1 (9.1%)
	Gesamt	**100%**	**Gesamt**	**100%**	**Gesamt**	**100%**
Datenzugang *n* (%)	prospektiv	3 (60.0%)	prospektiv	10 (50.0%)	prospektiv	11 (100%)
	retrospektiv	2 (40.0%)	retrospektiv	10 (50.0%)	retrospektiv	-
	Gesamt	**100%**	**Gesamt**	**100%**	**Gesamt**	**100%**
Gruppenzuweisung *n* (%)	Randomisierung	1 (20.0%)	Randomisierung	1 (5.0%)	Randomisierung	6 (54.5%)
	Selbstzuteilung	-	Selbstzuteilung	7 (35.0%)	Selbstzuteilung	-
	Verfahrensint.	-	Verfahrensint.	4 (20.0%)	Verfahrensint.	2 (18.2%)
	Ein-Gruppen-Design	4 (80.0%)	Ein-Gruppen-Design	8 (40.0%)	Ein-Gruppen-Design	3 (27.3%)
	Gesamt	**100%**	**Gesamt**	**100%**	**Gesamt**	**100%**
Messzeitpunkte *n* (%)	Prä-Post	3 (60.0%)	Prä-Post	7 (35.0%)	Prä-Post	5 (45.5%)
	Prä-Katamnese	-	Prä-Katamnese	5 (25.0%)	Prä-Katamnese	-
	Prä-Post-Katamnese	2 (40.0%)	Prä-Post-Katamnese	4 (20.0%)	Prä-Post-Katamnese	6 (54.5%)
	nur Post	-	nur Post	1 (5.0%)	nur Post	-
	nur Katamnese	-	nur Katamnese	2 (10.0%)	nur Katamnese	-
	nicht beurteilbar	-	*nicht beurteilbar*	1 (5.0%)	*nicht beurteilbar*	-
	Gesamt	**100%**	**Gesamt**	**100%**	**Gesamt**	**100%**

Anmerkung: KG: Kontrollgruppe, verfahrensint.: verfahrensintern, verfahrensext.: verfahrensextern.

Cluster 1 verzeichnet mit vier von fünf Studien ein Überwiegen von Untersuchungen im Ein-Gruppen-Design. Dabei wurde in einer Studie (41) ein Kontrollgruppendesign (Warteliste) realisiert. In Cluster 2 befinden sich drei Studien (15.0%) mit einem Kontrollgruppendesign und acht Studien (40.0%) im Ein-Gruppen-Design. Außerdem wurden in vier Studien (20.0%) verfahrensinterne Vergleiche und bei vier weiteren Studien (20.0%) verfahrensexterne Vergleiche angestellt. Zudem fällt die eine Studie mit dem kombinatorischen Vergleich mit verfahrensexternen sowie verfahrensinternen Treatments ebenfalls in dieses Cluster. Cluster 3 setzt sich aus drei Studien im Ein-Gruppen-Design (27.3%) und vier Kontrollgruppendesigns (36.4%) zusammen, zudem sind zwei Studien mit einem verfahrensinternen Vergleich sowie eine Studie mit einer Kombination aus Kontrollgruppen- und verfahrensexternem Vergleich zu finden. Die einzige Studie, in der ein Vergleich mit pharmakologischer Monotherapie vorgenommen wurde, befindet sich ebenfalls in diesem „Kurzzeittherapie-Cluster".

Im zuletzt beschriebenen Cluster befinden sich ausschließlich prospektive Studien, in Cluster 1 überwiegen die prospektiven Studien leicht mit drei zu zwei retrospektiven Studien, in Cluster 2 halten sich die prospektiven und retrospektiven Studien mit jeweils 50% die Waage.

Eine in Cluster 1 befindliche Studie im Kontrollgruppendesign realisierte die Zuweisung der Patienten zur Therapie und zur Warteliste randomisiert. Ebenfalls randomisierte Zuweisungen erfolgten in sechs (von 11) Studien im „Kurzzeittherapie-Cluster" (Cluster 3). In diesen Untersuchungen wurden Kontrollgruppenvergleiche bzw. ein Vergleich mit Pharmakotherapie vorgenommen. In Cluster 2 wird in einer Studie randomisiert vorgegangen. Diese Studie (19) wurde unter die Langzeittherapiestudien subsummiert, da in einem Treatmentarm analytische Psychotherapie mit einer mittleren Sitzungsanzahl von 232 Sitzungen durchgeführt wurde (vgl. Anhang F). In sieben Studien aus Cluster 2 wurde den Patienten die Zuwei-

sung selbst überlassen, wobei es sich bei zwei Studien um Vergleiche mit „echten“ Wartelisten handelt, die hier pragmatisch den Studien mit Selbstzuweisung zugerechnet wurden.

Tabelle 17 schließt mit der Darstellung der gewählten Messzeitpunkte in den unterschiedlichen Clustern. In Cluster 1 befinden sich ausschließlich Studien mit Prä-Post- oder mit Prä-Post-Katamnesemessungen. Selbiges trifft auf Cluster 3 zu. Anders stellt es sich in Cluster 2 dar: Mit sieben Studien (35.0%) mit Prä-Postmessungen und vier Studien (20.0%) mit Prä-Post-Katamnesemessungen sind diese Messzeitpunkt-Designs im Vergleich zu den beiden anderen Clustern niedriger besetzt. Die genannten Niedrigbesetzungen in Cluster 2 werden durch fünf Studien (25.0%) in der Kategorie der Prä-Katamnesemessungen kompensiert, zudem befinden sich die zwei reinen Katamnesemessungen sowie die eine Studie mit nur einem Postmesszeitpunkt in Cluster 2. In den einzigen drei Kurzzeittherapiestudien in Cluster 2 (vgl. Tabelle 16, S. 236) wurden ausschließlich Prä-Postmessungen vorgenommen.

Tabelle 18 sind die clusterspezifischen Verteilungen der Variablen „a priori Festlegung der Sitzungsanzahl“, „Störungsspezifität der Behandlung“, „Einsatz von Manual oder Behandlungsrichtlinien“, „Therapeutentraining zwecks Studie“ sowie „Implementationskontrolle“ zu entnehmen. Im „Kurzzeittherapie-Cluster“ (Cluster 3) zeigt sich, dass drei der fünf Variablen, die in ihrer positiven Ausprägung das RCT-Paradigma repräsentieren (vgl. Kap. 1.2.1), anteilig in ihrer positiven Ausprägung überwiegen. Das sind die Variablen „a priori Festlegung der Sitzungsanzahl“, „Einsatz von Manual oder Behandlungsrichtlinien“ und „Implementationskontrolle“. In der Variable „Implementationskontrolle“ überwiegt die Anzahl an Studien mit positiver Ausprägung jedoch mit sechs zu fünf Studien nur schwach. Zudem überwiegen bei den beiden Variablen „Störungsspezifität der Behandlung“ und „Therapeutentraining zwecks Studie“ die negativen Ausprägungen.

Tabelle 18: Charakterisierung der drei Cluster nach Festlegung der Sitzungsanzahl, Störungsspezifität der Behandlung, Einsatz von Manualen, Therapeutentraining und Implementationskontrolle (*n*=36)

		Cluster 1 (*n*=5)		**Cluster 2** (*n*=20)		**Cluster 3** (*n*=11)
a priori Festlegung der Sitzungsanzahl *n* (%)	ja	3 (60.0%)	ja	-	ja	10 (90.9%)
	nein	2 (40.0%)	nein	20 (100%)	nein	1 (9.1%)
Gesamt %		**100%**		**100%**		**100%**
Störungsspezifität der Behandlung *n* (%)	ja	1 (20.0%)	ja	-	ja	4 (36.4%)
	nein	4 (80.0%)	nein	20 (100%)	nein	7 (63.6%)
Gesamt %		**100%**		**100%**		**100%**
Manual/Behandlungs-richtlinien *n* (%)	ja	-	ja	1 (5.0%)	ja	8 (72.7%)
	nein	5 (100%)	nein	19 (95.0%)	nein	3 (27.3%)
Gesamt %		**100%**		**100%**		**100%**
Therapeutentraining zwecks Studie *n* (%)	ja	-	ja	-	ja	5 (45.5%)
	nein	5 (100%)	nein	20 (100%)	nein	6 (54.5%)
Gesamt %		**100%**		**100%**		**100%**
Implementationskon-trolle *n* (%)	ja	-	ja	3 (15.0%)	ja	6 (54.5%)
	nein	5 (100%)	nein	17 (85.0%)	nein	5 (45.5%)
Gesamt %		**100%**		**100%**		**100%**

In Cluster 2 überwiegen ausnahmslos diejenigen Ausprägungen, die nicht das RCT-Paradigma, sondern eher das naturalistische Paradigma repräsentieren (vgl. Kap. 1.2.1). Selbiger Trend trifft auf das kleinste Cluster (Cluster 1) zu, mit einer Ausnahme in der Variable „a priori Festlegung der Sitzungsanzahl“, in der die positive Ausprägung mit drei zu zwei Studien überwiegt.

Die letzte clusterbezogene Betrachtung ist Tabelle 19 zu entnehmen, in der die Häufigkeitsverteilungen der Variablen „Ausschluss subklinischer Symptomausprägungen“ sowie „Ausschluss komorbider Störungen“ dargestellt sind. Letztere Variable bezieht sich jedoch ausschließlich auf *epidemiologisch relevante* komorbide Störungen, womit bspw. Störungen, wie organische Hirnschädigungen, akute Psychosen oder aktuell bestehender starker Substanzmissbrauch ausgeschlossen sind.

Tabelle 19: Charakterisierung der drei Cluster nach Ausschluss subklinischer Symptomausprägung und epidemiologisch relevanter komorbider Störungen (*n*=36)

		Cluster 1 (*n*=5)		**Cluster 2** (*n*=20)		**Cluster 3** (*n*=11)
Ausschluss subklinischer Symptomausprägung *n* (%)	ja	3 (60.0%)	ja	12 (60.0%)	ja	10 (90.9%)
	nein	2 (40.0%)	nein	4 (20.0%)	nein	1 (9.1%)
			n.b.	4 (20.0%)		
Gesamt %		**100%**		**100%**		**100%**
Ausschluss epidemiologisch relevanter komorbider Störungen *n* (%)	ja	-	ja	1 (5.0%)	ja	3 (27.3%)
	nein	5 (100%)	nein	19 (95.0%)	nein	8 (72.7%)
Gesamt %		**100%**		**100%**		**100%**

In Bezug auf den „Ausschluss subklinischer Symptomausprägungen“ besteht in allen drei Clustern Richtungsgleichheit: Das bedeutet, es überwiegt bei allen Clustern die positive Variablenausprägung, wobei diese Ausprägung im „Kurzzeittherapie-Cluster“ (Cluster 3) mit 90.9% am stärksten besetzt ist. Das bedeutet, dass clusterübergreifend bei den meisten Studien subklinische Symptomausprägungen ausgeschlossen wurden.

Beim „Ausschluss komorbider Störungen“ ist ebenfalls ein clusterübergreifender Trend zu erkennen: Es überwiegt durchweg die negative Ausprägung, das bedeutet, dass in den meisten Studien, unabhängig von der Clusterzugehörigkeit, keine epidemiologisch relevanten komorbiden Störungen ausgeschlossen wurden. Dieser Trend erreicht mit 100% und 95.0% die höchsten Ausprägungen in Cluster 1 und 2. Im „Kurzzeittherapie-Cluster“ (Cluster 3) werden bei drei Studien (27.3%) epidemiologisch relevante Störungen ausgeschlossen.

Zusammenfassend sollen im Folgenden die drei Cluster zunächst separat voneinander beschrieben werden, um im Anschluss ihre Ähnlichkeiten und Unähnlichkeiten zu eruieren.

Cluster 1 (*n*=5)

Dieses Cluster zeichnet sich allein durch Studien zu mittelfristigen Therapien (über 25 bis 100 Sitzungen) aus, in denen ausschließlich tiefenpsychologisch fundierte Behandlungen unter-

sucht wurden. In vier der fünf Studien wurden Ein-Gruppen-Designs realisiert. Die fünfte Studie weist ein randomisiertes Kontrollgruppendesign mit einer Wartelistengruppe auf, die lediglich die ersten 6 Monate der durchschnittlich knapp 2 Jahre andauernden Behandlungszeit umfasst. Zudem überwiegen mit vier Studien die Untersuchungen an gemischten Störungsgruppen. Drei der fünf Studien wurden retrospektiv durchgeführt.

In keiner der Studien wurden Manuale oder Behandlungsrichtlinien eingesetzt oder Implementationskontrollen (Adherence-Messungen) durchgeführt, gleichsam wurden in keiner Untersuchung die Therapeuten zwecks Studie trainiert. Lediglich die Sitzungsanzahl wurde bei drei der fünf Studien a priori festgelegt. In keiner der fünf Studien wurden epidemiologisch relevante komorbide Störungen ausgeschlossen; subklinische Symptomausprägungen wurden dagegen bei drei der fünf Studien exkludiert.

Cluster 2 (*n*=20)

In Cluster 2 („Langzeittherapie-Cluster“) überwiegen Untersuchungen an psychodynamischen Verfahren, die in der Regel längerfristig angelegt sind (analytische Psychotherapie, Psychoanalyse). Dieses Cluster zeichnet sich durch eine heterogene Studiendesignwahl aus, in der, neben acht Ein-Gruppen-Designs (40.0%), vor allem verfahrensinterne und verfahrensexterne Vergleichsstudien (jeweils 20.0%) vorkommen. In drei Studien (15.0%) werden unbehandelte Kontrollgruppen zum Vergleich herangezogen, davon beziehen sich zwei Studien auf Langzeittherapien. Mit einer Ausnahme beziehen sich alle Studien auf gemischte Störungsgruppen. Der Anteil prospektiver und retrospektiver Untersuchungen ist gleichgroß.

Abzüglich der insgesamt 12 verfahrensinternen Vergleichsstudien sowie Ein-Gruppen-Designs (60.0%), erfolgte bei einer Langzeittherapiestudie die Patientenzuteilung randomisiert, bei sieben Studien (35.0%) entschieden sich die Patienten selbst für eine Behandlung. In

keiner der sieben Studien wurden Strategien zur Gruppenangleichung (Parallelisierung, Matching etc.) umgesetzt.

In Cluster 2 sind im Gegensatz zu den anderen beiden Clustern fünf Langzeittherapiestudien (25.0%) zu finden, in denen Prä-Katamnesemessungen durchgeführt wurden.

In keiner der 20 Studien aus Cluster 2 wurde die Sitzungsanzahl a priori festgelegt, gleichsam wurden keine störungsspezifischen Behandlungen und keine expliziten Therapeutentrainings durchgeführt. In einer Studie (zu Kurzzeittherapie) wurde in Anlehnung an Behandlungsrichtlinien vorgegangen, in drei Studien (zu Langzeittherapien) wurden Implementationskontrollen vorgenommen. Subklinische Symptomausprägungen wurden bei 12 Studien (60.0%) ausgeschlossen, ein Ausschluss relevanter komorbider Störungen erfolgte lediglich bei einer Studie, dabei handelt es sich um die einzige Studie aus Cluster 2, in der auch randomisiert wurde.

Cluster 3 (*n*=11)

Cluster 3 („Kurzzeittherapie-Cluster“) setzt sich ausschließlich aus Studien an Kurzzeittherapien zusammen, in denen allein tiefenpsychologisch fundierte Behandlungen untersucht wurden. Die beiden Anwendungsbereiche halten sich, im Unterschied zu den beiden anderen Clustern, mit einem leichten Überhang an Studien zu affektiven Störungen (sechs Studien; 54.5%), die Waage. Es sind verhältnismäßig wenige Ein-Gruppen-Design-Studien zu finden (drei Studien; 27.3%), dafür umso mehr Studien, in denen Kontrollgruppendesigns umgesetzt wurden (fünf Studien; 45.5%), sowie eine Studie, in der psychodynamische Kurzzeitbehandlungen mit pharmakologischer Monotherapie verglichen wurden. Alle Studien wurden prospektiv durchgeführt, zudem sind in diesem Cluster mit sechs Studien anteilig die meisten Untersuchungen mit randomisierter Patientenzuweisung zu finden (54.5%). In keiner Studie erfolgte die Gruppenbildung via Selbstzuweisung.

Die Verteilungen der Studien über die Variablen „a priori Festlegung der Sitzungsanzahl“, „Verwendung von Manual/Behandlungsrichtlinien“ sowie – in abgeschwächter Form – „Implementationskontrolle“ deuten darauf hin, dass es sich eher um Studien handelt, die dem RCT-Paradigma naheliegen. Dem stehen allein die Verteilungen der Studien über die Variablen „Störungsspezifität der Behandlungen“ sowie „Therapeutentraining“ entgegen, da dort der Anteil an Studien mit negativen Ausprägungen in diesen Variablen überwiegt.

Mit knapp 91% verfügt dieses Cluster über den größten Anteil an Studien, in denen subklinische Symptomausprägungen ausgeschlossen wurden. Zudem lassen sich drei von 11 Studien (27.3%) ausmachen, in denen relevante komorbide Störungen ausgeschlossen wurden. Dabei handelt es sich um Untersuchungen an affektiven Störungsgruppen.

Ähnlichkeiten und Distanzen zwischen den drei Clustern

Vergleicht man die prozentualen Häufigkeiten der einzelnen Variablen innerhalb der drei Cluster, lassen sich vor allem hinsichtlich bestimmter Variablen Ähnlichkeiten feststellen.

Ähnlichkeiten zwischen **Cluster 1** („mittelfristige Therapie-Cluster“) und **Cluster 2** („Langzeittherapie-Cluster“) in Distanz zu Cluster 3 („Kurzzeittherapie-Cluster“):

- In den Clustern 1 und 2 kommen vergleichsweise viele Studien an gemischten Störungsgruppen vor (80.0% und 95.0%) und dementsprechend wenige Studien an affektiven Störungsgruppen; in Cluster 3 herrscht diesbezüglich annähernde Gleichverteilung.
- In beiden Clustern sind wenige Studien enthalten, die klassische Kontrollgruppen heranziehen (15.0% und 20.0%); in Cluster 3 zogen immerhin 45.5% der Studien klassische Kontrollgruppen heran.

- Der Anteil retrospektiver Studien liegt in beiden Clustern zwischen 40.0% und 50.0%, in Cluster 3 befinden sich hingegen ausschließlich prospektive Studien.
- In beiden Clustern befinden sich nur wenige Studien, in denen die Gruppenzuweisung randomisiert erfolgte (5.0% und 20.0%); in Cluster 3 trifft dies auf 54.5% der Studien zu.
- Der Anteil an Studien, in denen Manuale/Behandlungsrichtlinien eingesetzt wurden, ist in beiden Clustern mit höchstens 5.0% sehr gering; in Cluster 3 kam es bei 72.7% der Studien zum Einsatz von Manualen/Behandlungsrichtlinien.
- In keiner der Studien aus Cluster 1 und 2 wurden explizite Therapeutentrainings zu Studienzwecken durchgeführt, in Cluster 3 trifft dies auf 45.5% der Studien zu.
- In beiden Clustern wurden nur bei einer geringen Anzahl an Studien Implementationskontrollen (Adherence-Messungen) durchgeführt (0% bzw. 15.0%); in Cluster 3 fanden demgegenüber bei 54.5% der Studien Adherence-Kontrollen statt.
- Zu einem Ausschluss von subklinischen Symptomausprägungen kam es in beiden Clustern bei jeweils 60.0% der Studien; in Cluster 3 wurden bei 90.9% der Studien subklinische Symptomausprägungen ausgeschlossen.
- In beiden Clustern befinden sich kaum Studien (maximal 5.0%), in denen relevante komorbide Störungen ausgeschlossen wurden, in Cluster 3 wurden hingegen bei immerhin 27.3% der Studien relevante komorbide Störungen ausgeschlossen.

Ähnlichkeiten zwischen **Cluster 1** („mittelfristige Therapie-Cluster“) und **Cluster 3** („Kurzzeittherapie-Cluster“) in Distanz zu Cluster 2 („Langzeittherapie-Cluster“):

- Die beiden Cluster 1 und 3 setzen sich ausschließlich aus Studien zu tiefenpsychologisch fundierter Psychotherapie zusammen; in Cluster 2 befindet sich lediglich eine Studie zu tiefenpsychologisch fundierter Psychotherapie und die Mehrheit der Stu-

dien verteilt sich auf die langfristigen psychodynamischen Verfahren (analytische Psychotherapie und Psychoanalyse).

- Beide Cluster verfügen ausschließlich über Studien mit Prä-Postmessungen oder Prä-Post-Katamnesemessungen, in Cluster 2 lässt sich hingegen eine größere Variabilität an Messdesigns ausmachen, u.a. wurden bei 25.0% der Studien Prä-Katamnesemessungen realisiert.
- Der Anteil an Studien, in denen die Sitzungsanzahl a priori festgelegt wurde, liegt in Cluster 1 und 3 bei 60.0% bzw. 90.9%; demgegenüber befindet sich in Cluster 2 keine Studie mit a priori festgelegter Sitzungsanzahl.
- Beide Cluster enthalten Studien, in denen störungsspezifische Behandlungen durchgeführt wurden (20.0% und 36.4%), in Cluster 2 befindet sich hingegen keine Studie mit störungsspezifischer Behandlung.

Ähnlichkeiten zwischen **Cluster 2** („Langzeittherapie-Cluster“) und **Cluster 3** („Kurzzeittherapie-Cluster“) in Distanz zu Cluster 1 („mittelfristige Therapie-Cluster“):

- In den Clustern 2 und 3 sind verhältnismäßig viele Studien zu Individualtherapien zu finden (81.8% bzw. 95.0%), in Cluster 1 liegt der Anteil lediglich bei 40.0%.
- Die Anteile an Studien mit Ein-Gruppen-Designs sind in den Clustern 2 und 3 (27.3% bzw. 40.0%), verglichen mit Cluster 1 (80.0%), niedrig.

Damit besteht im Hinblick auf die Variablen aus dem Kurzkodierbogen (Anhang B und D) die größte Ähnlichkeit zwischen Cluster 1(„mittelfristige Therapie-Cluster“) und Cluster 2 („Langzeittherapie-Cluster“), die geringste Ähnlichkeit besteht hingegen zwischen Cluster 2 und Cluster 3 („Kurzzeittherapie-Cluster“). Cluster 1 und 2 vereinigen auf sich eher diejenigen Charakteristika, die gemeinhin dem naturalistischen Pol zugesprochen werden, Cluster 3

steht dem RCT-Paradigma näher. Jedoch sind keine der Clusterprofile als vollkommen homogen anzusehen, so sind etwa in Cluster 3 Studien zu finden, die dem RCT-Pol durchaus entgegenlaufen (keine Störungsspezifität der Behandlung, kein Therapeutentraining zwecks Studie, keine Adherence-Kontrolle, kein Ausschluss epidemiologisch relevanter komorbider Störungen und sogar Studien ohne Randomisierung bzw. in Ein-Gruppen-Designs). In Cluster 1 sind wiederum Studien zu finden, die dem rein naturalistischen Pol entgegenlaufen (a priori Festlegung der Sitzungsanzahl, Ausschluss subklinischer Symptomausprägungen). Sowohl in Cluster 1 („mittelfristige Therapie-Cluster“) als auch in Cluster 2 („Langzeittherapie-Cluster“) befinden sich – wenn auch in absoluter Minderheit – Untersuchungen, in denen die Patientenzuweisung randomisiert erfolgte. Dennoch ist eine richtungweisende Verteilung der Cluster im Hinblick auf die beiden Pole *naturalistisch* und *RCT-typisch* auszumachen.

Im Rahmen der Dummy-Kodierung wurden zwei Studien (21, 30) ausgeschlossen, da deren Einschluss zu einer unverhältnismäßigen Erhöhung der Anzahl an Dummy-Variablen geführt hätte und die neu hinzukommenden Dummy-Variablen nahezu als Konstanten in die Clusteranalysen eingegangen wären[63]. Es soll nun – nachdem die Profile der einzelnen Cluster hinreichend bekannt sind – ermittelt werden, ob und ggf. welchem Cluster die beiden anfangs exkludierten Studien zuzuordnen wären. Dabei sollen die Clusterprofile als feste Parameter dienen und die Studien jeweils in dasjenige Cluster integriert werden, mit dem sie vom Gesamtprofil her am ehesten übereinstimmen. Tabelle 20 sind die Profile der beiden Studien zu entnehmen. Dafür wurden alle Variablen herangezogen, die in die clusteranalytischen Verfah-

[63] Dieser Umstand wurde zu Beginn dieses Kapitels (4.2) beschrieben und begründet.

ren eingingen, sowie die externen Variablen, mittels derer auch die 36 Studien zuvor charakterisiert wurden.

Tabelle 20: Profile der a priori exkludierten Studien (*n*=2)

	Studie 21	**Studie 30**
Anwendungsbereich	AS	GS
Psychoanalytisch begründetes Therapieverfahren (plus Psa)	TP	*n.b.*
Therapieumfang	MfT	MfT
Therapiesetting	Individualtherapie	Individualtherapie & Gruppentherapie
Studiendesign	KG-Design	Ein-Gruppen-Design
Datenzugang	prospektiv	prospektiv
Gruppenzuweisung	Randomisierung	[Ein-Gruppen-Design]
Messzeitpunkte	Prä-Post	Prä-Post-Katamnese
a priori Festlegung der Sitzungsanzahl	*n.b.*	teils festgelegt/teils nicht festgelegt
Störungsspezifität der Behandlung	nein	nein
Manual/Behandlungsrichtlinien	nein	nein
Therapeutentraining zwecks Studie	nein	nein
Implementationskontrolle (Adherence)	nein	nein
Ausschluss subklinischer Symptomausprägung	ja	nein
Ausschluss epidemiologisch relevanter komorbider Störungen	ja	nein

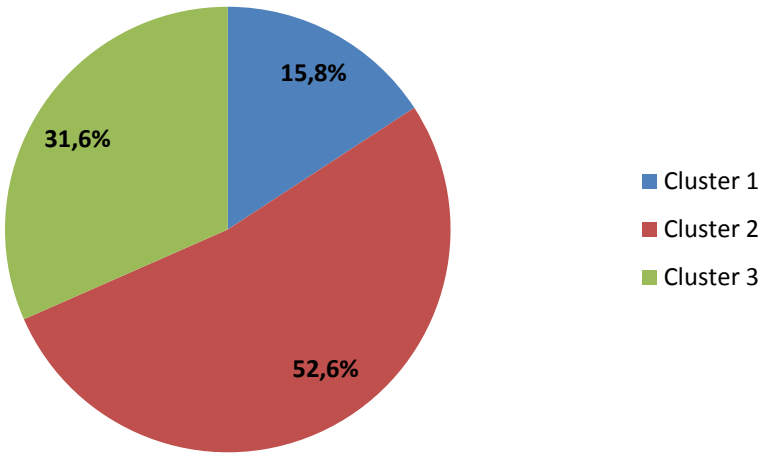
Anmerkung: AS: Affektive Störungen, KG: Kontrollgruppe, MfT: Mittelfristige Therapie, n.b.: nicht beurteilbar, Psa: Psychoanalyse, TP: Tiefenpsychologisch fundierte Psychotherapie.

Auf einen ersten, oberflächlichen Blick würde Studie 21 aufgrund ihres Profils sowohl in Cluster 1 als auch in Cluster 3 passen. Für Cluster 1 spräche, dass es sich bei dieser Studie um eine Untersuchung an Therapien von mittelfristiger Länge (über 25 bis 100 Sitzungen) handelt und die Variable „Manual/Behandlungsrichtlinien“ negativ ausgeprägt ist. Demgegen-

über entspricht die Kombination der Ausprägungen in den Variablen „Anwendungsbereich“, „Therapiesetting“, „Studiendesign“, „prospektiver *versus* retrospektiver Datenzugang“, „Gruppenzuweisung“, „Ausschluss subklinischer Symptomausprägungen“ sowie „Störungs-spezifität der Behandlung“ jedoch am ehesten dem Profil von Cluster 3.

Studie 30 würde auf den ersten Blick sowohl in Cluster 1 als auch in Cluster 2 passen, und die weitere Entscheidung zwischen den beiden Clustern fällt aufgrund ihrer Ähnlichkeit (s.o.) ungleich schwerer, als bei Studie 21. Die einzige Variable, in der sich die beiden Cluster tatsächlich grundlegend voneinander unterscheiden, ist der „Therapieumfang“. Die Tatsache, dass in Studie 30 Behandlungen von moderater Länge (über 25 bis 100 Sitzungen) untersucht wurden, lässt diese Studie – wenn auch nicht eindeutig, so doch zumindest mit einer Tendenz – Cluster 1 zuordnen. Abbildung 26 stellt dieses Kapitel abschließend die Clustergrößen nach Inklusion der beiden Studien (21, 30) dar.

Abbildung 26: 3-Clusterlösung nach hierarchischer Clusteranalyse nach Inklusion der 2 exkludierten Studien (21, 30) (*n*=38)



4.3 Ergebnisse zur Studienqualität entsprechend der drei Dimensionen des WBP-Kriterienkatalogs (allgemeine methodische Qualität, interne Validität, externe Validität)

Die Dimension der allgemeinen methodischen Qualität

Gemessen anhand des WBP-Kriterienkatalogs verfügen insgesamt 15 der 41 Studien (36.6%) über eine ausreichende allgemeine methodische Qualität, 26 Studien (63.4%) zeichnen sich hingegen durch methodische Mängel aus. Anhang J sind die Ergebnisse zur methodischen Qualität (als auch zur internen und externen Validität) pro Studie zu entnehmen.

Die Gründe, die jeweils zum Durchfallen der einzelnen Studien auf der methodischen Qualitätsdimension geführt haben, sind Tabelle 21 zu entnehmen. Hier wurden die K.O.-Kriterien herangezogen, deren Positivbewertung (mit „1“ oder „2“) als Voraussetzung für eine positive Bewertung auf der methodischen Qualitätsdimension fungiert bzw. deren Bewertung mit „3“ unmittelbar zu einer negativen Gesamtbewertung auf dieser Dimension führt. Zudem wurde im Rahmen dieser Arbeit entschieden, eine Studie dann auf der methodischen Qualitätsdimension als nicht ausreichend für eine Bewertung geeignet anzusehen, wenn mehr als 25% der 19 Kriterien nicht beurteilt werden konnten, also Missingwerte erzeugten. Als letztes enthält Tabelle 21 die Anzahl derjenigen Studien, die mit ihrem Mittelwert auf der methodischen Qualitätsdimension größer oder gleich dem Cutoff-Wert 2.25 waren.

Tabelle 21: Gründe für Negativbewertungen auf der Dimension der allgemeinen methodischen Qualität (*n*=26)

	***n* (%)**[a]
Manipulation der Daten [Kriterium A.1. mit „3“ bewertet]	-
Objektive und reliable Diagnosestellung (mittels (teil-) standardisierter Interviews) [Kriterium A.2. mit „3“ bewertet]	10 (38.5%)
Reliable und valide Messung zumindest der primären Zielkriterien [Kriterium A.8. mit „3“ bewertet]	4 (15.4%)
Erzielte Veränderungen auf den primären und sekundären Zielkriterien ggf. im Vergleich zur Kontrollgruppe (Signifikanz, Größe und Relevanz der Effekte) [Kriterium B.12. mit „3“ bewertet]	19 (73.1%)
Stichprobe von Patienten mit Störungen mit Krankheitswert [Kriterium C.1. mit „3“ bewertet]	6 (23.1%)
Primäre Zielkriterien beziehen sich auf patientenrelevante Parameter (insbesondere Schwere der Symptomatik, Leiden, Beeinträchtigung/Lebensqualität, Inanspruchnahme von Diensten des Gesundheitswesens) [Kriterium C.9. mit „3“ bewertet]	-
Anzahl der Missings > 25%	-
Durchschnitt der Kriterien A.1. – A.19. ≥ 2.25	6 (23.1)

[a] die prozentualen Häufigkeiten summieren sich nicht zu 100% auf, da einige Studien bei mehreren Kriterien mit „3“ abschnitten bzw. den Mittelwert-Cutoff überschritten.

In keiner Studie wurden beim Kodieren Hinweise auf Datenmanipulationen (Kriterium A.1.) entdeckt und in keiner Studie war der Grund für eine negative Gesamtbewertung auf der allgemeinen methodischen Qualitätsdimension die Tatsache, dass keine mehrdimensionalen, patientenrelevanten Parameter erhoben und berichtet wurden (Kriterium C.9.). Der häufigste Grund für eine Negativbewertung auf der methodischen Qualitätsdimension war ein Mangel an Veränderungs- und Zielerreichungsindikatoren im Hinblick auf die Darstellung der Outcomes (Kriterium B.12.; vgl. Kap. 3.2.3). Dies war bei insgesamt 19 Studien (73.1%) der Fall. Mit insgesamt 10 Studien (38.5%) war das Fehlen von objektiven und reliablen Diagnosestel-

lungen – mindestens anhand eines nachvollziehbaren klinischen Urteils – der zweithäufigste Grund für eine Negativbewertung von Studien auf der methodischen Qualitätsdimension (Kriterium A.2.). Bei sechs Studien (23.1%) ist das Kriterium C.1. mit „3“ bewertet worden, d.h. diese Untersuchungen bezogen sich auf Patientengruppen, von denen über 20% der Patienten über keine festgestellte Störung mit Krankheitswert entsprechend ICD oder DSM verfügten oder dies nicht eindeutig einzuschätzen war. Der Grund dafür, dass die jeweiligen Häufigkeiten an „3“er-Bewertungen bei Kriterium A.2. („Diagnosestellung-Kriterium“) und Kriterium C.1. („Krankheitswert-Kriterium“) nicht deckungsgleich sind, liegt in der einfachen Tatsache, dass in Studien durchaus detailliert die ICD- oder DSM-Diagnosen aller Studienpatienten benannt wurden, was – liberal – als Hinweis auf die Krankheitswertigkeit vorliegender Störungen betrachtet wurde; indes berichten selbige Studien von keinem diagnostischen Prozess, und es bleibt unnachvollziehbar, wie genau die berichteten ICD-/DSM-Diagnosen gestellt wurden (z.B. Studie 10 und 28).

Mit lediglich 4 Studien (15.4%), die bei Kriterium A.8. zur Bemessung der Reliabilität und Validität verwendeter Outcomemaße mit einer „3“ abschneiden, ist dieser Ausschlussgrund am wenigsten häufig besetzt. Die meisten Studien verwendeten demnach hinreichend reliable und valide Outcomemaße (vgl. Kap. 3.2.3).

Im Rahmen der Bewertung der allgemeinen methodischen Qualität musste keine Studie ausgeschlossen werden, weil sie über zu viele Missingwerte verfügt (vgl. Tabelle 21). Dafür liegen sechs Studien (23.1%) mit ihren Dimensionsmittelwerten auf oder über dem vom WBP vorgegebenen Cutoff von 2.25. Der Range dieser Mittelwerte liegt zwischen 2.25 bis 2.56. Jedoch fiel keine dieser sechs Studien allein aufgrund eines zu hohen Dimensionsmittelwerts durch. Neben der Überschreitung des Mittelwert-Cutoffs verfügen alle sechs Studien zusätz-

lich über Negativbewertungen auf mindestens einem der genannten K.O.-Kriterien aus Tabelle 21.

Obwohl, wie gezeigt werden konnte, die Anzahl der Missingwerte bei keiner Studie die eigens gesetzte 25%-Grenze überschritt, soll, wie in Kapitel 3.2.1 begründet, im Rahmen einer Rekodierung der Studien mittels plausibel modifizierter Kriterien, das Vorkommen von Missingwerten minimiert und dabei eruiert werden, inwieweit sich diese Modifikation auf die Gesamtbewertung der methodischen Qualität auswirkt. Modifikationen wurden auf dieser Dimension für das Kriterium A.11. (siehe Abbildung 13, S. 150) vorgenommen, das in Kapitel 3.2.1 ausführlich dargestellt wurde. Mittels dieses Kriteriums wird bemessen, inwieweit in einer Studie Fremdeinschätzungsverfahren durch externe, trainierte Beurteiler, die hinsichtlich der patientenseitigen Gruppenzugehörigkeit blind sind, eingesetzt wurden.

Die Modifikationen der Stufenoperationalisierungen für Kriterium A.11. sind Tabelle 22 zu entnehmen. Dazu wurden die eigens erstellten Zusatzkodierregeln (für Missingwerte) für das Kriterium A.11. in reguläre Ratingkategorien umgewandelt.

Tabelle 22: Kriterium A.11.: Rekodierung der Missingwerte

Zusatzkodierregel	**Modifikation für Rekodierung**
Missingwert „4“: Ein-Gruppen-Designs mit externen Beurteilern	„1“er-Bewertung, wenn externe Rater trainiert und Verwendung valider Fremdeinschätzungsverfahren
	„3“er-Bewertung, wenn externe Rater nicht trainiert oder 25% der durch externe Rater verwendeten Fremdeinschätzungsverfahren nicht valide
Missingwert „9“: keine Fremdeinschätzungsverfahren durch externe Beurteiler	„3“er-Bewertung, wenn keine Fremdeinschätzungsverfahren eingesetzt wurden

Demzufolge wurden Ein-Gruppen-Design-Studien im Zuge des Rekodierungsdurchgangs mit einer „1“ bewertet, wenn Fremdeinschätzungsverfahren durch externe Rater angewandt wurden, die Rater zudem in dieser Anwendung trainiert und die Verfahren als valide eingeschätzt wurden[64]. Mit einer „3“ wurden demgegenüber solche Studien kodiert, in denen die externen Rater entweder nicht trainiert waren oder aber 25% der verwendeten Fremdeinschätzungsverfahren als nicht valide eingestuft werden mussten.

Alle Studien, die im ersten Kodierungsdurchgang mit einer „9“ kodiert wurden, also die Bedingung „Fremdeinschätzungsverfahren eingesetzt“ nicht erfüllten, wurden in der Rekodierung mit „3“ bewertet. Tabelle 23 sind die Veränderungen in den absoluten und den prozentualen Besetzungshäufigkeiten der einzelnen Ratingstufen (plus Missingkategorien) zu entnehmen.

Tabelle 23: Kriterium A.11: Neuverteilung der Studien nach der Rekodierung (*N*=41)

	erste Kodierung (*N*=41) *n* (%)	**Rekodierung (*N*=41) *n* (%)**
Bewertung		
1	3 (7.3%)	10 (24.4%)
2	2 (4.9%)	2 (4.9%)
3	1 (2.4%)	29 (70.7%)
4	9 (22.0%)	-
9	26 (63.4%)	-

Von den 41 Primärstudien wurden insgesamt neun Studien (22.0%) mit dem Missingwert „4“ bewertet, da es sich um Studien in Ein-Gruppen-Designs handelte, in denen jedoch Fremdeinschätzungsverfahren durch externe Beurteilern eingesetzt wurden. Diese neun Studien verteilen sich im Rekodierungsdurchgang mit sieben Studien auf die „1“er-Kategorie und zwei Studien auf die „3“er-Kategorie. Die 26 Studien (63.4%), die im ersten Kodierungsdurchgang mit dem Missingwert „9“ belegt wurden, da in diesen Studien die Voraussetzung zur Bewer-

[64] Die Validitätseinschätzung erfolgte durch Kriterium A.8. (siehe Abbildung 17, S. 162).

tung auf diesem Kriterium nicht erfüllt war (der Einsatz von Fremdeinschätzungsverfahren), wurden mit „3“ rekodiert.

Wie Tabelle 21 (S. 252) entnommen werden kann, war die Missinganzahl bei keiner der Studien, denen eine unzureichende methodische Qualität via WBP-Kriterien attestiert wurde, ausschlaggebend für diese Negativbewertung. Dieser Umstand ändert sich selbstverständlich auch nicht durch die beschriebene Rekodierungsmaßnahme, durch die jedoch immerhin für 35 Studien die Missinganzahl vermindert werden konnte. Die Modifikation wirkt sich außerdem auf die Dimensionsmittelwerte der 35 Studien aus, jedoch führt dies bei keiner Studie zu einer Veränderung in der Gesamtbewertung auf dieser Dimension.

Die Dimension der internen Validität

Die Anwendung der internen Validitätskriterien zeichnet insgesamt drei der 15 Studien (20.0%) mit ausreichender methodischer Qualität als intern valide aus (9, 19, 27), 12 der 15 Studien (80.0%) hingegen verfügen demnach über keine hinreichende interne Validität (vgl. Anhang J). Tabelle 24 fasst die Ergebnisse zu den Gründen für Negativbewertungen auf der internen Validitätsdimension für die 12 Studien zusammen. Die Darstellung erfolgt separat für Ein-Gruppen-Designs und Mehrgruppendesigns, da, wie im ersten Teil der Arbeit ausgeführt, die interne Validitätsdimension primär für Mehrgruppendesigns konzipiert wurde (vgl. Kap. 1.2.2). Der Vollständigkeit halber sollen die Ein-Gruppen-Designs jedoch an dieser Stelle ebenfalls Erwähnung finden. Wie für die allgemeine methodische Qualität werden auch hier die formalen Voraussetzungen für ein positives Abschneiden auf der internen Validitätsdimension betrachtet – dazu gehört, dass in einer Studie die Patientenzuweisung zu den unterschiedlichen Treatmentarmen bei adäquater Stichprobengröße ($n > 30$) bestenfalls randomi-

siert, mindestens jedoch parallelisiert erfolgt sein sollte (siehe K.O.-Kriterium B.8.; Anhang C). Als eigens festgelegte Voraussetzung sollten zudem nicht mehr als 25% der 12 internen Validitätskriterien mit Missingwerten kodiert worden sein. Zuletzt soll auch hier der Mittelwert über alle 12 internen Validitätskriterien den Wert 2.24 nicht überschreiten.

Tabelle 24: Gründe für Negativbewertungen auf der Dimension der internen Validität (*n*=12)

	Ein-Gruppen-Designs (*n*=11) ***n* (%)**[a]	**Mehrgruppendesigns (*n*=1)** ***n* (%)**[a]
Randomisierte/parallelisierte vs. keine Gruppenzuweisungsstrategien [Kriterium B.8. mit „3“ bewertet]	11 (100%)	-
Anzahl der Missings > 25%	11 (100%)	-
Durchschnitt der Kriterien B.1. – B.12. ≥ 2.25	1 (9.1%)	1 (100%)

[a] die prozentualen Häufigkeiten summieren sich nicht zu 100% auf, da einige Studien sowohl bei Kriterium B.8. mit „3“ abschnitten als auch die zulässige Missinganzahl bzw. den Mittelwert-Cutoff überschritten.

Von den insgesamt 12 Studien, die negativ auf der internen Validitätsdimension abgeschnitten haben, sind 11 Studien im Ein-Gruppen-Designs durchgeführt worden – darunter drei "Ein-Gruppen-Designs im erweiterten Sinne" (verfahrensinterne Vergleiche) und acht Untersuchungen ohne jegliche Vergleichsgruppe. Lediglich eine Studie wurde in einem Mehrgruppendesign durchgeführt (Studie 34).

Auf Seiten der Ein-Gruppen-Design-Studien fallen *per definitionem* alle Studien durch das K.O.-Kriterium der internen Validitätsdimension (B.8.) durch und liegen erwartungsgemäß alle über der maximalen Missinganzahl. Eine der 11 Studien wäre zudem noch durch einen zu hohen Dimensionsmittelwert durch die interne Validitätsdimension gefallen. Die berechneten Durchschnittswerte dieser Studien sind jedoch in ihrer Aussagekraft, aufgrund der hohen Anzahl an Missings, drastischen eingeschränkt.

Die einzige Untersuchung im Mehrgruppendesign, die auf der internen Validitätsdimension durchfällt, überschreitet zwar die Hürde des K.O.-Kriteriums B.8., da es sich um eine randomisierte Studie handelt. Sie verfügt zudem nicht über zu viele Missingwerte, jedoch liegt der Durchschnittswert der internen Validitätskriterien bei dieser Studie über dem vorgegebenen Cutoff-Wert.

Analog zur allgemeinen methodischen Qualitätsdimension und dem dort modifizierten Kriterium (A.11.) soll eruiert werden, ob und ggf. wie sich eine Modifikation des Kriteriums B.10. (siehe Abbildung 10, S. 144) auf die Gesamtbewertung der internen Validität auswirkt. Mit Hilfe des Kriteriums B.10. wird bemessen, wie viele Messzeitpunkte in einer Studie realisiert wurden und ob es sich bei den verwendeten Outcomemaßen hauptsächlich um prospektive oder um retrospektive Messinstrumente handelt (vgl. Kap. 3.2.1). Die Modifikation zwecks Minimierung von Missingwerten ist Tabelle 25 zu entnehmen. Die Modifikationen beziehen sich ausschließlich auf die Bewertung der in den Studien gewählten Messzeitpunkte, die Art der Messung (retrospektiv [z.B. GAS; Kiresuk & Sherman, 1968] *versus* prospektiv [z.B. SCL-90-R; Derogatis, 1975]) bleibt von der Modifikation unberührt.

Tabelle 25: Kriterium B.10.: Rekodierung der Missingwerte

Zusatzkodierregel	**Modifikation für Rekodierung**
Missingwert „4“: Prä-Katamnesemessung inklusive mehrerer Messzeitpunkte	„2“er-Bewertung
Missingwert „5“: Prä-Katamnesemessung	„2“er-Bewertung
Missingwert „6“: nur Katamnesemessung	„3“er-Bewertung

Demzufolge wurden Studien mit Prä-Katamnesemessungen – unabhängig davon, ob noch mehrere Messungen zwischen diesen Messzeitpunkten vorgenommen wurden – im Rekodierungsdurchgang mit „2“ bewertet, reine Katamnesestudien mit „3“.

Tabelle 26 sind die Veränderungen in den absoluten und den prozentualen Besetzungshäufigkeiten der einzelnen Ratingstufen (plus Missingkategorien) zu entnehmen. Dabei bezieht sich die Darstellung ausschließlich auf diejenigen Studien, die die Voraussetzung einer ausreichenden methodischen Qualitätsbeurteilung erfüllt haben (*n*=15).

Tabelle 26: Kriterium B.10.: Neuverteilung der Studien nach der Rekodierung (*n*=15)

	erste Kodierung (*n*=15)	**Rekodierung (*n*=15)**
Bewertung		
1	3 (20.0%)	3 (20.0%)
2	10 (66.7%)	12 (80.0%)
3	-	-
4	2 (13.3%)	-
5	-	-
6	-	

Es zeigt sich, dass die Modifikation für zwei der 15 Studien zu einer Verminderung der Missinganzahl führt. Keine der dadurch erzeugten Veränderungen wirkt sich auf den jeweiligen Mittelwert der internen Validitätsdimension grundlegend in Richtung Unter- oder Überschreitung des Mittelwert-Cutoffs aus. Damit bleibt die Anzahl an Untersuchungen, denen ein ausreichendes Ausmaß an interner Validität via WBP-Kriterien attestiert werden konnte gegenüber der Anzahl an Studien, denen keine ausreichende interner Validität attestiert wurde, beim Rekodierungsdurchgang im Vergleich zur ersten Kodierung unverändert.

Die Dimension der externen Validität

Die Anwendung der externen Validitätskriterien zeichnet insgesamt 13 (86.7%) der 15 Studien mit ausreichender methodischer Qualität als extern valide aus. Zwei der 15 Studien verfügen demnach über keine hinreichende externe Validität (vgl. Anhang J). Es handelt sich dabei um zwei der drei Studien, die mittels der internen Validitätsdimension als intern valide Untersuchungen diagnostiziert wurden. Tabelle 27 fasst die Ergebnisse zu den Gründen für Negativbewertungen auf der externen Validitätsdimension für die zwei Studien zusammen.

Tabelle 27: Gründe für Negativbewertungen auf der Dimension der externen Validität (*n*=2)

	n (*n*=2)
Anzahl der Missings > 25%	-
Durchschnitt der Kriterien C.1. – C.9. > 2.25	2
Durchschnitt der Kriterien C.10. – C.13. > 2.25	-

Beide Studien erhalten eine Negativbewertung auf der externen Validitätsdimension, da der Mittelwert-Cutoff für die externen Validitätskriterien C.1. – C.9. überschritten wurde.

Betrachtet man die Kriterien-Mittelwerte der 13 Studien, die auf der externen Validitätsdimension positiv abgeschnitten haben, so zeigt sich folgendes Bild (Tabelle 28): Der Range der Mittelwerte über die externen Validitätskriterien C.1. – C.9. beträgt 1.00 bis 2.07. Die Mittelwerte über die Kriterien C.10. – C.13. produzieren überhaupt keine Streuung, da alle 13 Studien hier mit *M* 1.00 abschneiden.

Tabelle 28: Range der Mittelwerte der externen Validitätsdimension (*n*=13)

	Range der EV-Kriterien-Mittelwerte (*n*=13)
Mittelwert der Kriterien C.1. – C.9. bei positiver EV	Range 1.00 – 2.07
Mittelwert der Kriterien C.10. – C.13. bei positiver EV	-

Anmerkung: EV: Externe Validität.

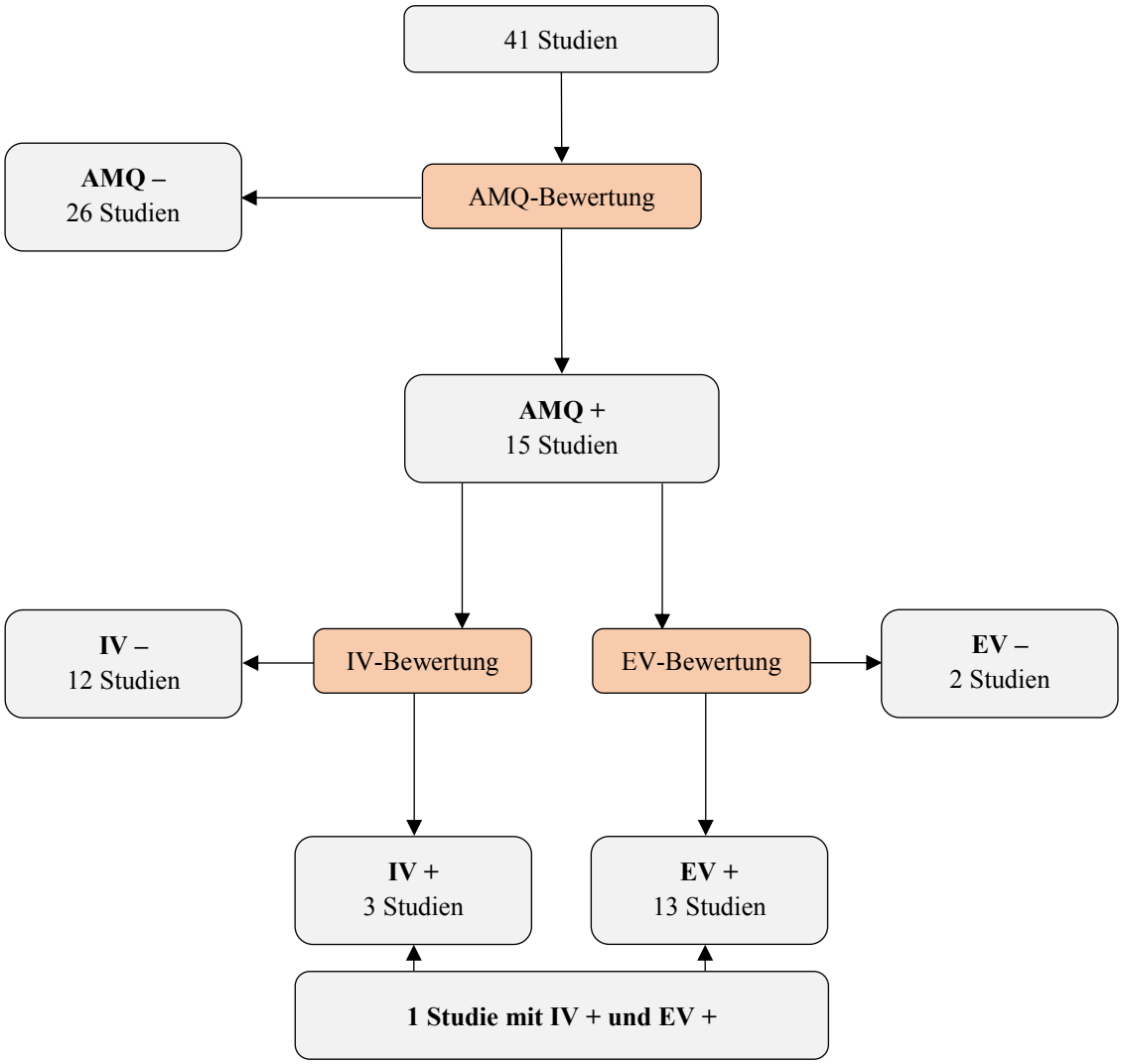
Fazit

Gemessen auf der Basis der WBP-Kriterien liegen von den 41 Primärstudien insgesamt 15 Studien mit positiv bewerteter allgemeiner methodischer Qualität vor, von denen wiederum drei Studien als intern valide und 13 Studien als extern valide beurteilt wurden. Lediglich eine Studie verfügt nach WBP-Kriterien – neben ausreichender methodischer Qualität – sowohl über ausreichende interne als auch externe Validität. An diesen Verteilungen ändert sich auch im Hinblick auf die aufgezeigten Rekodierungsdurchgänge nichts.

Damit kann für das Verhältnis der Bewertungen auf der internen Validitätsdimension im Vergleich zu den Bewertungen auf der externen Validitätsdimension Folgendes festgehalten werden: Nahezu alle Studien, die auf der internen Validitätsdimension schlecht abgeschnitten haben, schneiden auf der externen Validitätsdimension gut ab und umgekehrt. Das bedeutet, die zwei Studien, die auf der externen Validitätsdimension schlecht abschnitten, schneiden hingegen auf der internen Validitätsdimension gut ab. Diese Umkehr wird allein durch eine Studie durchkreuzt, die auf beiden Dimensionen mit gutem Ergebnis bewertet wurde. Inwieweit das sich in diesem Befund abzeichnende Konzept von einem eher inversen Verhältnis von interner und externer Validität durch die Operationalisierungen der beiden

Validitätsarten im Kriterienkatalog oder aber durch die Studien selbst hervorgerufen wird, wird in Kapitel 5.3.1 ausführlich diskutiert.

Der Gesamtverlauf der 41 Studien durch die einzelnen Bewertungsstationen (allgemeine methodische Qualität, interne und externe Validität) ist nochmals grafisch in Abbildung 27 illustriert.



Anmerkung: AMQ: Allgemeine methodische Qualität, AMQ +: positives Gesamtergebnis auf der allgemeinen methodischen Qualitätsdimension, AMQ –: negatives Gesamtergebnis auf der allgemeinen methodischen Qualitätsdimension, EV: Externe Validität, EV +: positives Gesamtergebnis auf der externen Validitätsdimension, EV –: negatives Gesamtergebnis auf der externen Validitätsdimension, IV: Interne Validität, IV +: positives Gesamtergebnis auf der internen Validitätsdimension, IV –: negatives Gesamtergebnis auf der internen Validitätsdimension.

Abbildung 27: Flowdiagramm der Primärstudien nach Anwendung des WBP-Kriterienkatalogs (*N*=41)

Im Folgenden soll außerdem ein kurzer Überblick über die Studien, die durch die Bewertungen auf den einzelnen Dimensionen in- und exkludiert wurden, gegeben werden. Dabei sollen folgende Merkmale Berücksichtigung finden:

- Anwendungsbereich
- Studiendesign
- Therapieumfang (Sitzungsanzahl).

Tabelle 29 sind die Verteilungen der zu den beiden Anwendungsbereichen gehörenden Studien über die jeweiligen Dimensionsbewertungen (positiv *versus* negativ) zu entnehmen. Von den insgesamt 10 Studien aus dem affektiven Störungsbereich schneiden vier Studien auf der allgemeinen methodischen Qualitätsdimension gut ab, die restlichen sechs Studien fallen durch die methodische Qualitätsbewertung durch. Ein ähnliches Verhältnis zeigt sich bei den gemischten Störungsgruppen-Studien: Hier schneiden 11 Studien (35.5%) gut und 20 Studien (64.5%) schlecht ab.

Tabelle 29: Ergebnisse der Primärstudien zu den affektiven und gemischten Störungsgruppen auf den drei Dimensionen des WBP-Kriterienkatalogs (*N*=41)

		Affektive Störungen *n* (%)	Gemischte Störungen *n* (%)
AMQ	+	4 (40.0%)	11 (35.5%)
	-	6 (60.0%)	20 (64.5%)
Gesamt		**10 (100%)**	**31 (100%)**
IV	+	2 (50.0%)	1 (9.1%)
	-	2 (50.0%)	10 (90.1%)
Gesamt		**4 (100%)**	**11 (100%)**
EV	+	2 (50.0%)	11 (100%)
	-	2 (50.0%)	-
Gesamt		**4 (100%)**	**11 (100%)**

Anmerkung: AMQ: Allgemeine methodische Qualität, EV: Externe Validität, IV: Interne Validität, +: positives Gesamtergebnis, –: negatives Gesamtergebnis.

Auf der internen Validitätsdimension schneiden von den vier als methodisch gut bewerteten Studien zu affektiven Störungen anteilig gleich viele gut, wie schlecht ab. Bei den zwei Studien, die hier negativ bewertet wurden, handelt es sich ausschließlich um Ein-Gruppen-Designs. Auf Seiten der gemischten Störungen zeigt sich mit 10 von 11 Studien eine vergleichsweise hohe Anzahl negativ bewerteter Studien. Ein genauerer Blick in diese 10 Studien verrät, dass davon sechs Studien in einem "echten Ein-Gruppen-Design", drei Studien in einem "Ein-Gruppen-Design im erweiterten Sinne" (verfahrensinterner Vergleich) und eine Studie im Mehrgruppendesign (mit Wartelistenvergleich) durchgeführt wurden.

Auf der externen Validitätsdimension schneiden von den vier Studien zu affektiven Störungen, die als methodisch gut bewertet wurden, zwei Studien mit gutem und zwei Studien mit schlechtem Gesamtergebnis ab: Die beiden zuvor als intern valide herausgestellten Studien werden auf der externen Validitätsdimension als *in*valide eingestuft; die als nicht intern valide klassifizierten Studien im Ein-Gruppen-Design schneiden hingegen gut auf der externen Validitätsdimension ab. Bei den gemischten Störungen verhält es sich ähnlich: Hier schneiden alle 11 Studien mit guter methodischer Qualität – darunter die 10 Studien mit negativ bewerteter interner Validität – gut auf der externen Validitätsdimension ab.

Damit schneiden auf der internen Validitätsdimension anteilig mehr Studien aus dem gemischten Störungsbereich schlecht ab, als es auf Seiten der affektiven Störungen der Fall ist. Umgekehrt schneiden auf der externen Validitätsdimension alle Studien zu gemischten Störungen gut ab, was bei den vier Studien zu affektiven Störungen bei zwei Studien der Fall ist.

Tabelle 30 vergleicht nun die unterschiedlichen Studiendesigns in ihrem Abschneiden auf den drei Dimensionen.

Tabelle 30: Ergebnisse der Primärstudien in unterschiedlichen Studiendesigns auf den drei Dimensionen des WBP-Kriterienkatalogs (*N*=41)

		Ein-Gruppen-Design *n* (%)	WL, TAU, Placebo, aktive KG *n* (%)	verfahrensinterner Vergleich *n* (%)	verfahrensexterner Vergleich *n* (%)	Kombination verfahrensinterner & verfahrensexterner Vergleich *n* (%)	Kombination KG (WL) & verfahrensexterner Vergleich *n* (%)	Vergleich mit Pharmaka *n* (%)
AMQ	+	8 (44.4%)	2 (22.2%)	3 (42.9%)	-	1 (100%)	1 (100%)	-
	-	10 (55.6%)	7 (77.8%)	4 (57.1%)	4 (100%)	-	-	1 (100%)
Gesamt		**18 (100%)**	**9 (100%)**	**7 (100%)**	**4 (100%)**	**1 (100%)**	**1 (100%)**	**1 (100%)**
IV	+	-	1 (50.0%)	-	-	1 (100%)	1 (100%)	-
	-	8 (100%)	1 (50.0%)	3 (100%)	-	-	-	-
Gesamt		**8 (100%)**	**2 (100%)**	**3 (100%)**	**-**	**1 (100%)**	**1 (100%)**	**-**
EV	+	8 (100%)	1 (50.0%)	3 (100%)	-	1 (100%)	-	-
	-	-	1 (50.0%)	-	-		1 (100%)	-
Gesamt			**2 (100%)**	**3 (100%)**	**-**	**1 (100%)**	**1 (100%)**	-

Anmerkung: AMQ: Allgemeine methodische Qualität, EV: Externe Validität, IV: Interne Validität, KG: Kontrollgruppe, TAU: Treatment-As-Usual, WL: Warteliste, +: positives Gesamtergebnis, –: negatives Gesamtergebnis.

Auf der Dimension der allgemeinen methodischen Qualität ist der Anteil von Studien im Ein-Gruppen-Design und positivem Ergebnis (44.4%) dem Anteil von Studien mit einem verfahrensinternen Vergleich und positivem Ergebnis (42.9%) sehr ähnlich. Demgegenüber schneiden anteilig vergleichsweise wenige Studien (22.2%) im Kontrollgruppendesign mit gutem Ergebnis ab und gar keine Studien mit einem verfahrensexternen Vergleich. Die zwei Studien, innerhalb derer jeweils mehrere Kontroll- bzw. Vergleichsgruppen realisiert wurden („Kombination aus verfahrensinternem und verfahrensexternem Vergleich“ sowie „Kombination aus Kontrollgruppen- und verfahrensexternem Vergleich“), schneiden beide auf der methodischen Qualitätsdimension gut ab. Hingegen fällt die eine Studie mit dem Vergleich zur pharmakologischen Monotherapie durch die methodische Qualitätsbewertung durch.

Auf der internen Validitätsdimension fallen *per definitionem* alle Studien im Ein-Gruppen-Design inklusiver derer, die in der Kodierung ebenfalls als solche behandelt werden (verfahrensinterne Vergleiche), durch. Lediglich eine Studie im Mehrgruppendesign (34) fällt ebenfalls durch die interne Validität durch. Die restlichen drei Studien im Mehrgruppendesign, davon eine im Kontrollgruppendesign (9) sowie die beiden Studien mit den kombinatorischen Vergleichen (19, 27), schneiden positiv auf der internen Validitätsdimension ab.

Auf der externen Validitätsdimension schneiden demgegenüber alle Studien in Ein-Gruppen-Designs (inklusive der verfahrensinternen Vergleiche) mit positivem Ergebnis ab. Von den zwei kombinatorischen Vergleichen schneidet nur derjenige mit den verfahrensinternen und verfahrensexternen Vergleichsgruppen gut ab (die einzige Studie [19], die sowohl auf der internen als auch auf der externen Validitätsdimension gut abschneidet). Auf Seiten der Kontrollgruppendesigns schneidet von den zwei methodisch guten Studien eine Studie gut auf der externen Validitätsdimension ab.

Tabelle 31 zeigt die Verteilungen der Kurzzeittherapiestudien (bis 25 Sitzungen) und Langzeittherapiestudien (über 100 Sitzungen) sowie der Studien zu mittelfristig langen Behandlungen (über 25 bis 100 Sitzungen) in ihrem Abschneiden auf den drei Dimensionen des Kriterienkatalogs.

Tabelle 31: Ergebnisse der Primärstudien zu unterschiedlichen Therapieumfängen (Sitzungsanzahl) auf den drei Dimensionen des WBP-Kriterienkatalogs

		KZT *n* (%)	MfT *n* (%)	LZT *n* (%)
AMQ	+	6 (40.0%)	4 (44.4%)	5 (29.4%)
	-	9 (60.0%)	5 (55.6%)	12 (70.6%)
Gesamt		**15 (100%)**	**9 (100%)**	**17 (100%)**
IV	+	2 (33.3%)	-	1 (20.0%)
	-	4 (66.7%)	4 (100%)	4 (80.0%)
Gesamt		**6 (100%)**	**4 (100%)**	**5 (100%)**
EV	+	4 (66.7%)	4 (100%)	5 (100%)
	-	2 (33.3%)	-	-
Gesamt		**6 (100%)**	**4 (100%)**	**5 (100%)**

Anmerkung: AMQ: Allgemeine methodische Qualität, EV: Externe Validität, IV: Interne Validität, KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie, +: positives Gesamtergebnis, –: negatives Gesamtergebnis.

Von den insgesamt 15 Kurzzeittherapiestudien schneiden sechs Studien (40.0%) mit guter methodischer Qualität ab, neun der 15 Studien (60.0%) schneiden mit unzureichender methodischer Qualität ab. Ein ähnliches Verhältnis zeigt sich bei den Studien zu mittelfristigen Therapiedauern: Hier schneiden vier Studien (44.4%) mit gutem und fünf Studien (55.6%) mit schlechtem Ergebnis auf der allgemeinen methodischen Qualitätsdimension ab. Die anteilmäßig höchste Anzahl methodisch unzureichender Studien ist mit 12 Studien (70.6%) auf Seiten der Langzeittherapiestudien zu finden, denen fünf Studien (29.4%) mit guter methodischer Qualität gegenüberstehen.

Von den sechs Kurzzeittherapiestudien mit guter methodischer Qualität schneiden zwei Studien (33.3%) mit guter interner Validität und vier Studien (66.7%) mit unzureichender interner Validität ab. Bei den zwei Studien mit gutem Ergebnis auf der internen Validi-

tätsdimension (9, 27) handelt es sich um Mehrgruppendesigns; von den vier Studien, denen unzureichende interne Validität attestiert wurde, handelt es sich bei einer Studie (34) um eine Untersuchung im Mehrgruppendesign, die restlichen drei Studien wurden im Ein-Gruppen-Design (1, 13) bzw. als verfahrensinterne Vergleichsstudie (31) durchgeführt. Auf der externen Validitätsdimension schneiden vier der sechs Kurzzeittherapiestudien (66.7%) mit gutem Ergebnis ab – dabei handelt es sich um dieselben Studien, die auf der internen Validitätsdimension durchfielen. Die zwei Studien (33.3%) mit unzureichender externer Validität sind dementsprechend dieselben, die auf der internen Validitätsdimension gut abschnitten.

Auf Seiten der Untersuchungen zu mittelfristigen Behandlungsdauern schneiden alle vier Studien mit guter methodischer Qualität ebenfalls positiv auf der externen Validitätsdimension ab. Hingegen schneiden sie auf der internen Validitätsdimension durchgängig mit negativem Ergebnis ab. Bei allen vier Studien handelt es sich um Untersuchungen im Ein-Gruppen-Design (7, 14, 30) bzw. um einen verfahrensinternen Vergleich (15).

Von den fünf methodisch guten Langzeittherapiestudien schneidet eine Studie mit gutem Ergebnis auf der internen Validitätsdimension ab, dabei handelt es sich um die einzige Untersuchung (19), die durch die Validitätskriterien sowohl als intern als auch extern valide beurteilt wurde. Bei den vier Studien (80.0%) mit schlechtem Ergebnis auf der internen Validitätsdimension handelt es sich um Untersuchungen im Ein-Gruppen-Design (6, 16, 22) bzw. um einen verfahrensinternen Vergleich (11). Auf der externen Validitätsdimension schneiden alle fünf Langzeittherapiestudien mit positivem Gesamtergebnis ab.

Insgesamt wurden also alle als methodisch gut bewerteten Studien zu längerfristigen Behandlungen (über 25 Sitzungen) als extern valide eingestuft. Positive Gesamtergebnisse auf der internen Validitätsdimension blieben nahezu aus, da diese Studien, mit einer Ausnahme (19), durchgängig in Ein-Gruppen-Designs (inklusive verfahrensinterne Vergleiche) durchgeführt

wurden. Auf Seiten der sechs methodisch guten Kurzzeittherapiestudien (bis 25 Sitzungen) sind zwar anteilig mehr Studien zu finden, die als intern valide eingestuft wurden, jedoch befinden sich diese Studien ebenfalls in der Minderzahl (n=2). Auch hier bilden die als extern valide eingestuften Studien die Mehrzahl (n=4), wenn auch nicht in einem ganz so starken Ausmaß, wie es bei den Studien zu längerfristigen Behandlungen (über 25 Sitzungen) der Fall ist. Von den Kurzzeittherapiestudien schneidet keine Studie auf *beiden* Validitätsdimensionen mit positivem Ergebnis ab.

Inwieweit dieser Befund allem voran im Hinblick auf Langzeittherapiestudien (> 100 Sitzungen) auf eine „zu hohe Hürde" der internen Validitätsdimension hinweist, so dass diese Studien es kaum vermögen, auf dieser Dimension positiv abzuschneiden, wird in Kapiteln 4.5 eingehend eruiert.

4.4 Analysen zur Gegenstandsadäquatheit: Dimension der allgemeinen methodischen Qualität

Anhang A sind die zu untersuchenden K.O.-Kriterien der allgemeinen methodischen Qualitätsdimension in der Reihenfolge der sich nun anschließenden Darstellung zu entnehmen. Es wird daher innerhalb der einzelnen Kriterienanalysen auf den wiederholten Hinweis auf diesen Anhang verzichtet.

Die gemeinsamen Verteilungen der (dichotomisierten) Variablen „Behandlungslänge" und der (dichotomisierten) Kriterienstufen werden zunächst kreuztabellarisch dargestellt. Den Kreuztabellen sind zudem die Verteilungen der Langzeittherapiestudien (> 100 Sitzungen) und der Studien zu kürzeren Behandlungsdauern (< 100 Sitzungen) in Form von zeilenweisen Prozentwerten zu entnehmen. Aus letzteren Werten resultiert schließlich eine zusammenfassende Grafik, der die „Richtung" evtl. bestehender Verteilungsunterschiede eindeutig zu entnehmen sein wird. Verteilungsunterschiede werden in Form der Effektgröße ω (Omega) dar-

gestellt (vgl. Kap. 3.4.3). Der Extremgruppenvergleich (Langzeit- *versus* „echte“ Kurzzeittherapiestudien [bis 25 Stunden]) wird ausschließlich in Form von Effektgrößen ausgedrückt. Sollte sich neben dem Betrag der Effektgröße beim Extremgruppenvergleich auch die „Richtung“ des Verteilungsunterschieds ändern, wird dies gesondert kommentiert.

Manipulation der Daten (Kriterium A.1.)

Das Kriterium A.1., mit dem eingeschätzt wird, ob es Grund zu der Annahme gibt, dass die einer Studie zugrunde liegenden Daten und/oder die Datenauswertungen manipuliert wurden, ist im Gegensatz zu allen anderen Kriterien lediglich 2-gestuft („optimal“ bedeutet „keine Hinweise auf Manipulation“; „ungenügend“ bedeutet „Hinweise aus Ergebnismanipulation“). Tabelle 32 ist zu entnehmen, dass in keiner der 41 Studien angenommen werden musste – etwa aufgrund uneinheitlicher oder sich widersprechender Ergebnismuster in unterschiedlichen Publikationen zu einer Studie – dass Ergebnismanipulationen vorgenommen wurden. Aufgrund der dargestellten Besetzungshäufigkeit wird auf eine zusätzliche grafische Darstellung dieses Ergebnisses verzichtet.

Tabelle 32: Kreuztabelle über „Manipulation der Daten“ (Kriterium A.1.) und Behandlungslänge (*N*=41)

		A.1.		
		optimal	ungenügend und Ausschluss	
		n (%) zeilenweise %	*n* (%) zeilenweise %	Gesamt *n* (%) zeilenweise %
Behandlungslänge	KZT/ MfT	24 (58.5%) 100%	-	24 (58.5%) 100%
	LZT	17 (41.5%) 100%	-	17 (41.5%) 100%
	Gesamt *n* (%)	41 (100%)	-	41 (100%)

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie.

Objektive und reliable Diagnosestellung (Kriterium A.2.)

Tabelle 33 illustriert die Verteilung der Kurzzeit-/mittelfristigen Therapiestudien[65] im Vergleich zur Verteilung der Langzeittherapiestudien über das sog. „Diagnosestellung-Kriterium" (Kriterium A.2.). Mit diesem wird erhoben, inwiefern die psychiatrischen Diagnosen der Studienpatienten mittels objektiver und reliabler Diagnosesysteme (SCID, Diagnosechecklisten etc.) erhoben wurden (vgl. Kap. 1.2.2).

Tabelle 33: Kreuztabelle über „Objektive und reliable Diagnosestellung" (Kriterium A.2.) und Behandlungslänge (*N*=41)

		A.2.		
		optimal / zufriedenstellend	ungenügend und Ausschluss	
		n (%) zeilenweise %	*n* (%) zeilenweise %	Gesamt *n* (%) zeilenweise %
Behandlungslänge	KZT/ MfT	19 (46.3%) 79.2%	5 (12.2%) 20.8%	24 (58.5%) 100%
	LZT	12 (29.3%) 70.6%	5 (12.2%) 29.4%	17 (41.5%) 100%
	Gesamt *n* (%)	31 (75.6%)	10 (24.4%)	41 (100%)

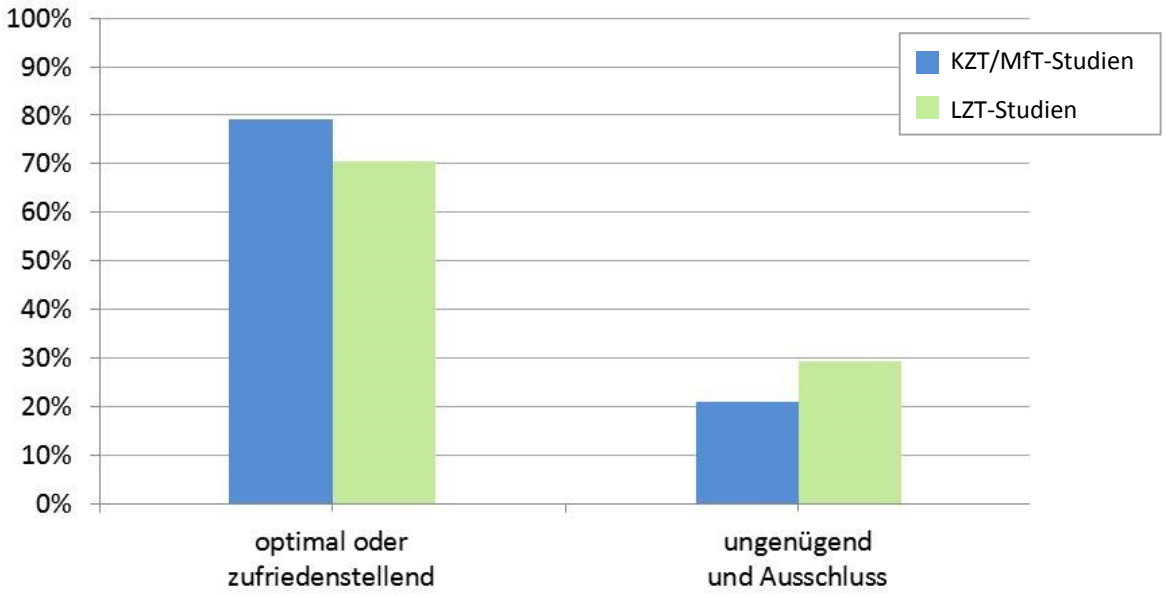
Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie.

Bei 31 der 41 Studien (75.6%) wurden die Diagnosen mindestens anhand von Diagnosechecklisten oder mittels eines nachvollziehbaren klinischen Urteils erstellt (Stufe „2"), bestenfalls jedoch auf der Basis klinischer Interviews mit Hilfe standardisierter Erhebungsverfahren (Stufe „1"). In 10 der 41 Studien (24.4%) wurden die Diagnosen entweder in für den Leser der Studie nicht nachvollziehbarer Weise gestellt (Stufe „3") – dies war dann der Fall, wenn in den jeweiligen Studienpublikationen zwar die Rede von Diagnosen war, der diagnostische Prozess jedoch im Dunkeln blieb (vgl. Kap. 4.3). Bei einem Teil der „3"er-kodierten Studien wurde die Art der Diagnosestellung hingegen durchaus benannt, musste jedoch als inadäquat

[65] Der einfacheren Lesbarkeit zuliebe wird hier und im Folgenden die kürzere Umschreibung „Kurzzeit-/mittelfristige Therapiestudien" für Studien zu kurzen und mittelfristig langen Therapien genutzt.

eingestuft werden. Bei diesen Studien wurden die Diagnosen bspw. retrospektiv und auf der Basis von Überweisungsbögen oder Krankenakten der Studienpatienten gestellt (5, 29).

Abbildung 28 stellt die prozentualen Anteilsunterschiede zwischen den Kurzzeit-/mittelfristigen Therapiestudien und den Langzeittherapiestudien auf den dichotomisierten Stufen des Kriteriums A.2. dar (vgl. zeilenweise % in Tabelle 33). Dabei wird ersichtlich, dass anteilig mehr Langzeittherapiestudien mit einem negativen Ergebnis auf diesem Kriterium abschneiden, als Kurzzeit-/mittelfristige Therapiestudien. Mit ω 0.10 (entspricht einem kleinen Effekt) ist der in Kapitel 3.4.3 festgesetzte Cutoff von ω 0.10 erreicht, was eine Feinanalyse notwendig macht. Dieser verkleinert sich beim Extremgruppenvergleich (Kurzzeittherapie *versus* Langzeittherapie; *n*=32) auf ω 0.03 .



Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie.

Abbildung 28: Verteilungsvergleich von Studien zu kürzeren Therapiedauern und Langzeittherapiestudien auf Kriterium A.2. („Objektive und reliable Diagnosestellung“) (*N*=41)

Feinanalyse zu Kriterium A.2. („Objektive und reliable Diagnosestellung“)

Ein genauerer Blick in die Langzeittherapiestudien mit negativem Ergebnis verrät, dass es sich bei allen fünf Studien um retrospektive Untersuchungen handelt. Dies machte eine Diagnosestellung, die in diesen Fällen rückwirkend hätte erfolgen müssen, da zu Beginn der Therapie nicht nachvollziehbar erstellt, schwer möglich. Gleiches trifft ebenfalls auf die Studien zu kürzeren Behandlungen und negativem Ergebnis zu, bei denen drei der fünf Studien mittels eines retrospektiven Designs durchgeführt wurden, wodurch die Diagnosestellung, wenn zu Therapiebeginn nicht nachvollziehbar erfolgt, ebenso wenig möglich gewesen sein wird. In keiner der retrospektiven Langzeittherapiestudien sind zwingende Gründe zu finden, die erklären könnten, warum diese Langezeittherapieuntersuchungen aufgrund ihres Gegenstands "Langzeittherapie" ausschließlich retrospektiv durchgeführt werden konnten. Zudem befinden sich in der Studiengruppe zur Langzeittherapie und positivem Ergebnis auf diesem Kriterium ebenfalls retrospektive Untersuchungen, gleiches trifft auf die Studiengruppe zu kürzeren Therapien zu. Das bedeutet, dass ein retrospektives Design nicht notwendigerweise impliziert, dass kein diagnostischer Prozess zur Baseline stattgefunden hat.

Reliable und valide Messung der primären Zielkriterien (Kriterium A.8.)

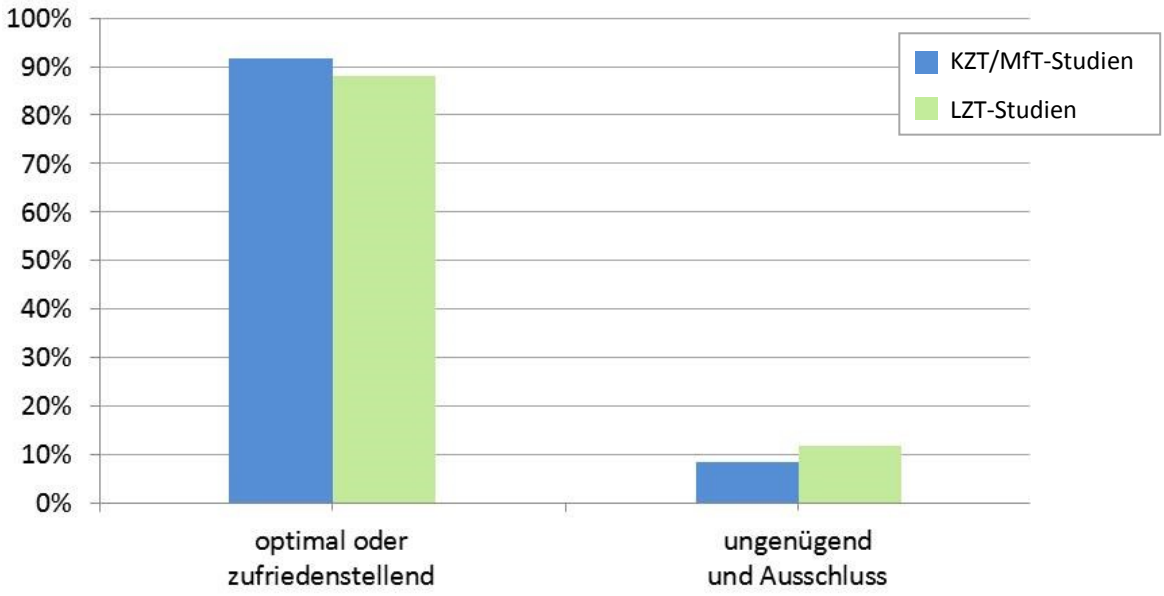
In Tabelle 34 werden die Verteilungen der Kurzzeit-/mittelfristigen Therapiestudien im Vergleich zu den Langzeittherapieuntersuchungen über die Stufen des Kriteriums A.8. dargestellt, mit dem die Reliabilität und Validität der in den Studien verwendeten Outcomemaße bewertet wird (vgl. Kap. 3.2.3). Von den insgesamt 41 Studien schneiden 37 (90.2%) mit „optimal/zufriedenstellend“ ab und eine Minderheit von vier Studien (9.8%) mit „ungenügend“. In diesen vier Studien wurden demnach, gemessen an der Gesamtanzahl der in den jeweiligen Studien verwendeten primären Outcomemaße, mehr als 25% dieser Maße als unzureichend reliabel und valide eingestuft.

Tabelle 34: Kreuztabelle über „Reliable und valide Messung der primären Zielkriterien“ (Kriterium A.8.) und Behandlungslänge (*N*=41)

		A.8.		
		optimal / zufriedenstellend	ungenügend und Ausschluss	
		n (%) zeilenweise %	*n* (%) zeilenweise %	Gesamt *n* (%) zeilenweise %
Behand-lungs-länge	KZT/ MfT	22 (53.7%) 91.7%	2 (4.9%) 8.3%	24 (58.5%) 100%
	LZT	15 (36.6%) 88.2%	2 (4.9%) 11.8%	17 (41.5%) 100%
	Gesamt *n* (%)	37 (90.2%)	4 (9.8%)	41 (100%)

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie.

Abbildung 29 illustriert, inwieweit sich die Studien zu unterschiedlichen Behandlungslängen (Kurzzeit-/mittelfristige Therapie *versus* Langzeittherapie) in ihren jeweiligen prozentualen Anteilen auf den beiden Kriterienstufen unterscheiden. Es zeigt sich, dass mit einer Effektgröße von ω 0.06 (entspricht einem sehr kleinen Effekt) ein minimaler Unterschied zugunsten der Kurzzeit-/mittelfristigen Therapiestudien besteht. Dieser Effekt vergrößert sich unter Betrachtung der Extremgruppen (Kurzzeittherapie *versus* Langzeittherapie; *n*=32) geringfügig auf eine Effektgröße von ω 0.09 . Damit schneiden wenig mehr Langzeittherapiestudien auf diesem Kriterium schlechter ab, als Studien zu kürzeren (< 100 Sitzungen) bzw. kurzen (< 25 Sitzungen) Behandlungsdauern.



Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie.

Abbildung 29: Verteilungsvergleich von Studien zu kürzeren Therapiedauern und Langzeittherapiestudien auf Kriterium A.8. („Reliable und valide Messung der primären Zielkriterien“) (*N*=41)

Veränderung und Zielerreichung auf Zielkriterien (Kriterium B.12.)

Die Verteilungen der Kurzzeit-/mittelfristigen Therapiestudien im Vergleich zu den Langzeittherapiestudien über die beiden Stufen des Kriteriums B.12. ist Tabelle 35 zu entnehmen. Mittels dieses Kriteriums wird bemessen, inwieweit in einer Studie unterschiedliche Zugänge der Wirksamkeitsabbildung realisiert wurden (Veränderungs- und Zielerreichungsmessungen mittels statistischer und klinischer Signifikanz- sowie Effektstärkenberechnungen; vgl. Kap. 3.2.3).

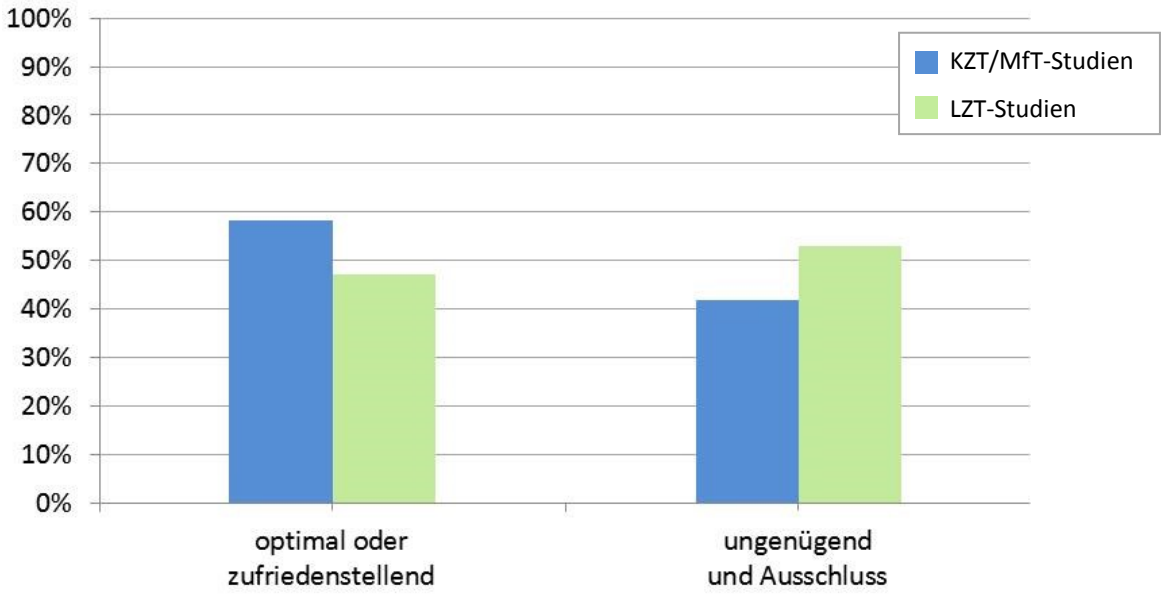
Von den 41 Studien schneiden knapp über die Hälfte der Studien (53.7%) mit positivem Ergebnis auf diesem Kriterium ab, 19 Studien (46.3%) verfügen hingegen über eine unzureichende Darstellung der Wirksamkeit. Bei dem größten Teil der 19 Studien wurden nur Signifikanzberechnungen angestellt, auf die Darstellung von Effektstärken und klinischer Signifikanzmaße hingegen vollkommen verzichtet.

Tabelle 35: Kreuztabelle über „Veränderung und Zielerreichung auf Zielkriterien“ (Kriterium B.12.) und Behandlungslänge (*N*=41)

		B.12.		
		optimal / zufriedenstellend	ungenügend und Ausschluss	
		n (%) zeilenweise %	*n* (%) zeilenweise %	Gesamt *n* (%) zeilenweise %
Behandlungslänge	KZT/ MfT	14 (34.1%) 58.3%	10 (24.4%) 41.7%	24 (58.5%) 100%
	LZT	8 (19.5%) 47.1%	9 (22.0%) 52.9%	17 (41.5%) 100%
	Gesamt *n* (%)	22 (53.7%)	19 (46.3%)	41 (100%)

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie.

Die Prozentwertunterschiede zwischen den Kurzzeit-/mittelfristigen Therapiestudien und den Langzeittherapiestudien auf den beiden Stufen des Kriteriums B.12. werden in Abbildung 30 sichtbar. Mit einer Effektgröße von ω 0.11 (entspricht einem kleinen Effekt) fällt dieser Vergleich zuungunsten der Langzeittherapiestudien aus. Dieser Effekt vergrößert sich auf ω 0.13 beim Vergleich der Extremgruppen (Kurzzeittherapie *versus* Langzeittherapie; *n*=32), bleibt jedoch insgesamt in einem Bereich, in dem man von einem „kleinen“ Effekt spricht (vgl. Eid et al., 2010). Da beide Effekte jedoch den zuvor festgelegten Cutoff-Wert von ω 0.10 überschreiten, soll dem Zustandekommen der Effekte im Folgenden feinanalytisch auf den Grund gegangen werden.



Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie.

Abbildung 30: Verteilungsvergleich von Studien zu kürzeren Therapiedauern und Langzeittherapiestudien auf Kriterium B.12. („Veränderung und Zielerreichung auf Zielkriterien“) (*N*=41)

Feinanalyse zu Kriterium B.12. („Veränderung und Zielerreichung auf Zielkriterien“)

Ein sorgfältiger Blick in die 19 Studien mit negativem Ergebnis auf diesem Kriterium zeigt, dass vor allem folgende Gründe ausschlaggebend für diese Bewertung waren: Wie oben bereits erwähnt, wurden bei den meisten Fällen über gruppenstatistische Signifikanzberechnungen hinaus weder Effektstärken noch individuumsbezogene, statistisch bedeutsame Veränderungen via RCI berichtet (vgl. Kap. 3.2.3). Zudem wurden keine normorientierten Zielerreichungsmaße (Cutoff-Werte) herangezogen, an denen sich die individuelle Veränderung von einem dysfunktionalen in einen funktionalen Bereich ablesen ließe (vgl. Stieglitz, 2008). Ein weiterer, wenn auch seltener Grund für ein negatives Ergebnis auf Kriterium B.12. war, dass in Mehrgruppendesigns keine direkten Vergleiche zwischen den untersuchten Gruppen durchgeführt wurden (2, 8, 18, 21, 38, 41), was für eine positive Bewertung auf diesem Kriterium jedoch erforderlich gewesen wäre. Zwischen den Studien zu kürzeren Behandlungsdau-

ern und den Langzeittherapiestudien sind hinsichtlich der aufgezählten Gründe keine Unterschiede feststellbar. Unterschiede bestehen allerdings dahingehend, dass die drei Studien mit nur einem Messzeitpunkt (Post- oder Katamnesemessung) und negativem Ergebnis auf Kriterium B.12. einzig auf Seiten der Langzeittherapiestudien zu finden sind. Dabei handelt es sich – neben einer Untersuchung zum Therapieerfolg rein aus der Therapeutenperspektive zu Behandlungsende (35) – bei den beiden Katamnesestudien (17, 23) um diejenigen Untersuchungen mit den längsten Katamnesezeiträumen (≥ 6 Jahre) im gesamten Studienpool (vgl. Kap. 4.1).

Alle drei Untersuchungen lassen Aussagen zu Veränderungen oder Zielerreichungen auf psychosozialen Outcomemaßen in erster Linie im Querschnitt und damit aus retrospektiver Perspektive (Patient oder Therapeut) zu. Zudem handelt es sich bei den Studien um Untersuchungen ohne Vergleichs-/Kontrollbedingung. So konnten Veränderungen über Effektstärken und – auf individueller Ebene – über RCI-Berechnungen oder das Ausmaß klinisch relevanter Zielerreichungen über normbasierte Cutoff-Werte nicht eruiert werden, da diese Wirksamkeitsindikatoren immer einen Vergleichswert benötigen (Prämessung oder Kontrollgruppenvergleich). Allein die in den beiden Katamnesestudien (17, 23) objektiv erhobenen Daten zum Inanspruchnahmeverhalten vor und nach der Therapie vermögen Aufschluss über Veränderungen in dem Sinne zu geben, wie es das Kriterium B.12. fordert.

Inwiefern die Tatsache, dass es sich bei den drei besagten Studien im Ein-Messzeitpunkt-Design um Untersuchungen zu Langzeittherapien handelt, das gewählte Messdesign rechtfertigt, berührt sehr grundlegende Überlegungen. Diese beziehen sich bspw. auf forschungsethische Vorbehalte im Hinblick auf die prospektive Untersuchung analytischer Behandlungen. Die Haltbarkeit dieser Vorbehalte wird im letzten Teil dieser Arbeit gesondert zu diskutieren sein (vgl. Kap. 5.1.6).

Stichprobe von Patienten mit Störungen mit Krankheitswert (Kriterium C.1.)

Tabelle 36 illustriert die Verteilungen der Kurzzeit-/mittelfristigen Therapiestudien im Vergleich zu den Langzeittherapiestudien über die Stufen des sog. „Krankheitswert-Kriteriums" C.1., das bemisst, inwieweit die untersuchten Studienpatienten in den jeweiligen Untersuchungen unter einer krankheitswertigen psychischen Störung leiden oder aber lediglich subklinische Symptomausprägungen aufweisen.

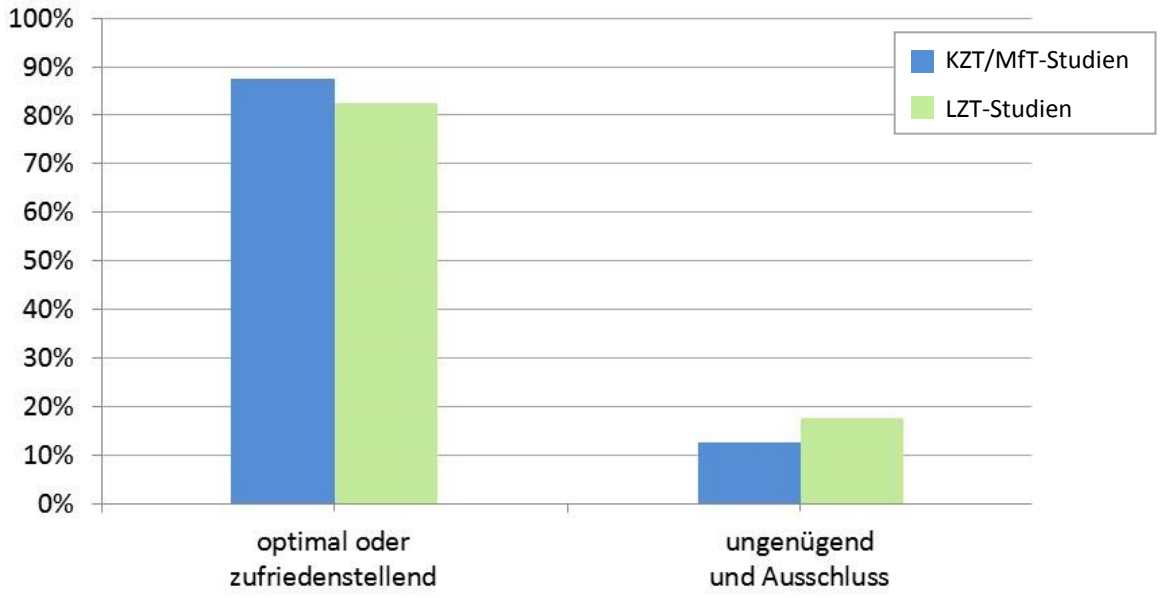
Tabelle 36: Kreuztabelle über „Stichprobe von Patienten mit Störungen mit Krankheitswert" (Kriterium C.1.) und Behandlungslänge (*N*=41)

		C.1.		
		optimal / zufriedenstellend	ungenügend und Ausschluss	
		n (%) zeilenweise %	*n* (%) zeilenweise %	Gesamt *n* (%) zeilenweise %
Behandlungslänge	KZT/ MfT	21 (51.2%) 87.5%	3 (7.3%) 12.5%	24 (58.5%) 100%
	LZT	14 (34.1%) 82.4%	3 (7.3%) 17.6%	17 (41.5%) 100%
	Gesamt *n* (%)	35 (85.4%)	6 (14.6%)	41 (100%)

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie.

Von den 41 Studien wurden 35 Studien (85.4%) mit Stufe „1" oder „2" bewertet. Das bedeutet, dass die in diesen Studien behandelten Patienten entweder alle unter einer Störung mit Krankheitswert litten (Stufe „1") oder aber die auf Stufe „2" gesetzte Grenze von 20% durch den Anteil an Patienten mit subklinischer Symptomausprägung nicht überschritten wurde. Sechs Studien (14.6%) mussten hier mit einer „3" bewertet werden, aus zwei dieser Studien (3, 29) geht eindeutig hervor, dass die 20%-Grenze durch den Anteil an Patienten mit subklinischer Symptomausprägung überschritten wird. Bei den restlichen 4 Studien (5, 25, 36, 37) konnte dies durch fehlende Angaben nicht eindeutig bestimmt werden, so dass aus diesem Grund eine „3"er-Kodierung erfolgte.

Aus Abbildung 31 wird ein lediglich kleiner Prozentwertunterschied zwischen den Kurzzeit-/mittelfristigen Therapiestudien und den Langzeittherapiestudien ersichtlich, der sich in einer Effektgröße von ω 0.07 (entspricht einem kleinen Effekt) niederschlägt. Dieser Effekt verkleinert sich auf ω 0.03 beim Extremgruppenvergleich (*n*=32) und ändert zudem „seine Richtung“: Nach Exklusion der Studien zu Behandlungen mittlerer Dauer (über 25 bis 100 Sitzungen) verändern sich die Prozentwertunterschiede dergestalt, dass nunmehr geringfügig mehr Langzeittherapiestudien auf diesem Kriterium gut abschneiden als Kurzzeittherapiestudien respektive mehr Kurzzeittherapiestudien schlecht, als Langzeittherapiestudien dies tun. Diese „Richtungsumkehr“ kommt dadurch zustande, dass alle sechs mittelfristigen Therapiestudien auf dem „Krankheitswert-Kriterium“ mit „optimal/zufriedenstellend“ abgeschnitten haben, so dass der Ausschluss im Ergebnis derart richtungsändernd ins Gewicht fällt.



Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie.

Abbildung 31: Verteilungsvergleich von Studien zu kürzeren Therapiedauern und Langzeittherapiestudien auf Kriterium C.1. („Stichprobe von Patienten mit Störungen mit Krankheitswert“) (*N*=41)

Primäre Zielkriterien beziehen sich auf patientenrelevante Parameter (Kriterium C.9.)

In Tabelle 37 werden die Verteilungen der Kurzzeit-/mittelfristigen Therapiestudien im Vergleich zu den Langzeittherapiestudien über die Stufen des Kriteriums C.9. dargestellt. Mit einer „1“ wurden Untersuchungen dann bewertet, wenn etwa neben psychosozialen Outcomemaßen (symptombezogen, interpersonal, Persönlichkeitsfragebögen etc.) auch Erhebungen zur Lebensqualität und/oder zum Inanspruchnahmeverhalten durchgeführt wurden. Mit einer „2“ wurden Studien dann bewertet, wenn ausschließlich psychosoziale Outcomemaße verwendet wurden, ohne die Lebensqualität/-zufriedenheit oder das Inanspruchnahmeverhalten mit zu erheben. Mit einer „3“ wären Studien dann bewertet worden, wenn in ihnen lediglich sog. Surrogatparameter (bspw. Kontrollüberzeugung) erhoben worden wären.

Mit diesem vergleichsweise strengen Maßstab, zumindest was eine Bewertung mit „1“ betrifft, haben 15 der 41 Studien (36.6%) mit „optimalem“ Ergebnis abgeschnitten und die restlichen 26 Studien (63.4%) mit „zufriedenstellendem“ Ergebnis. Da keine der Studien mit „ungenügend“ abgeschnitten hat, wird auf eine grafische Darstellung verzichtet.

Tabelle 37: Kreuztabelle über „Primäre Zielkriterien beziehen sich auf patientenrelevante Parameter“ (Kriterium C.9.) und Behandlungslänge (*N*=41)

		C.9.		
		optimal / zufriedenstellend	ungenügend und Ausschluss	
		n (%) zeilenweise %	*n* (%) zeilenweise %	Gesamt *n* (%) zeilenweise %
Behandlungslänge	KZT/ MfT	24 (58.5%) 100%	-	24 (58.5%) 100%
	LZT	17 (41.5%) 100%	-	17 (41.5%) 100%
	Gesamt *n* (%)	41 (100%)	-	41 (100%)

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie.

Gesamtergebnis zur allgemeinen methodischen Qualität

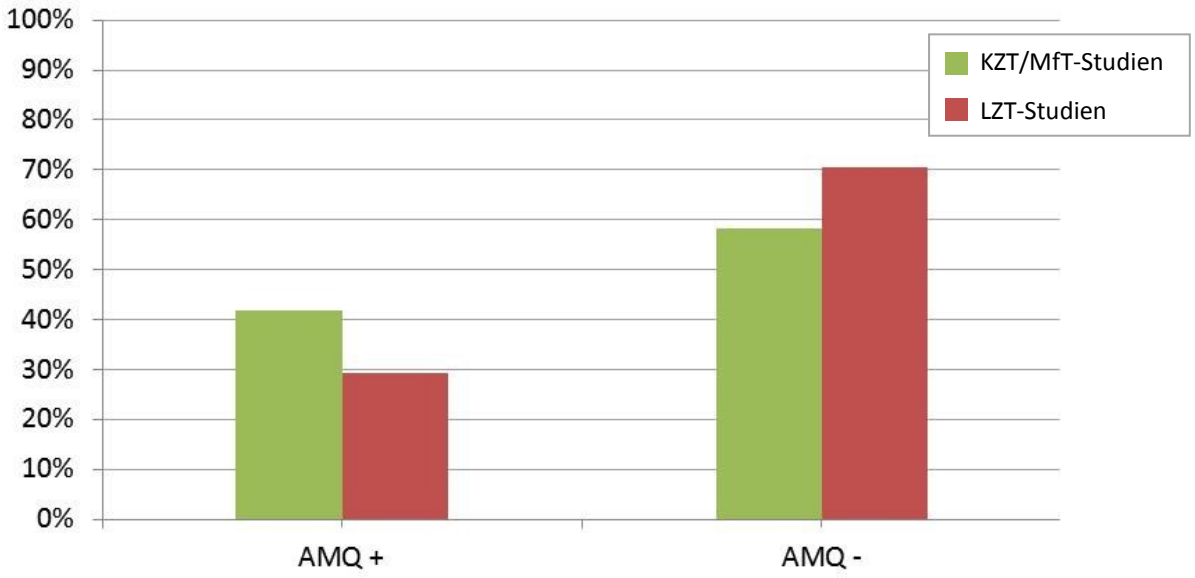
Tabelle 38 enthält schließlich die Gesamtergebnisse aller 24 Studien zu kürzeren Therapiedauern (bis 100 Sitzungen) im Vergleich zu den 17 Langzeittherapiestudien (über 100 Sitzungen) in ihrem Abschneiden auf der allgemeinen methodischen Qualitätsdimension.

Tabelle 38: Kreuztabelle über die Gesamtergebnisse der allgemeinen methodischen Qualitätsdimension und Behandlungslänge (*N*=41)

AMQ Gesamtbewertung				
		+	-	
		n (%) zeilenweise %	*n* (%) zeilenweise %	Gesamt *n* (%) zeilenweise %
Behandlungslänge	KZT/ MfT	10 (24.4%) 41.7%	14 (34.1%) 58.3%	24 (58.5%) 100%
	LZT	5 (12.2%) 29.4%	12 (29.3%) 70.6%	17 (41.5%) 100%
	Gesamt *n* (%)	15 (36.6%)	26 (63.4%)	41 (100%)

Anmerkung: AMQ: Allgemeine methodische Qualität, KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie, +: positives Gesamtergebnis auf der allgemeinen methodischen Qualitätsdimension, –: negatives Gesamtergebnis auf der allgemeinen methodischen Qualitätsdimension.

Von den insgesamt 17 Langzeittherapiestudien erhalten 12 Studien (70.6%) ein negatives Ergebnis auf dieser Dimension. Mit einem ebenfalls negativen Gesamtergebnis stehen diesen Langzeittherapiestudien 14 von 24 Studien (58.3%) auf Seiten der Kurzzeit-/mittelfristigen Therapiestudien gegenüber. Aus dem Prozentwertunterschied resultiert eine Effektgröße von ω 0.13 unter Einschluss aller 41 Studien. Nach Exklusion der neun Untersuchungen zu mittelfristig langen Behandlungen verkleinert sich dieser Effekt auf ω 0.11. Die Ergebnisse auf der methodischen Qualitätsdimension weisen auf einen Überhang an als methodisch „gut“ bewerteten Studien zu kürzeren bzw. kurzen Behandlungsdauern hin, der mit den genannten Effekten jedoch noch als geringfügig zu betrachten ist (vgl. Eid et al., 2010).



Anmerkung: AMQ: Allgemeine methodische Qualität, KZT: Kurzzeittherapie, LZT: Langzeittherapie, MfT: Mittelfristige Therapie.

Abbildung 32: Verteilungsvergleich von Studien zu kürzeren Therapiedauern und Langzeittherapiestudien auf der allgemeinen methodischen Qualitätsdimension (Gesamtergebnisse) (*N*=41)

4.5 Analysen zur Gegenstandsadäquatheit: Dimension der internen Validität

Anhang A sind die in diesem Kapitel zu untersuchenden Kriterien der internen Validität in der Reihenfolge der sich nun anschließenden Darstellung zu entnehmen. Es wird in den Erläuterungen der einzelnen Kriterien auch hier auf den wiederholten Hinweis auf diesen Anhang verzichtet.

Wie zu Beginn der Arbeit (Kap. 1.2.2) gezeigt wurde, beziehen sich die Kriterien der internen Validitätsdimension primär auf Mehrgruppendesigns. Der Vergleich der Studien zu den unterschiedlichen Therapieumfängen in ihrem Abschneiden auf den einzelnen Kriterien der internen Validität wird aus diesem Grund nur an Studien mit Mehrgruppendesigns und unter

Ausschluss von Studien in Ein-Gruppen-Designs sowie von verfahrensinternen Vergleichsuntersuchungen erfolgen. Damit stehen für diesen Vergleich von den insgesamt 15 Studien mit guter allgemeiner methodischer Qualität noch vier Studien zur Verfügung, die in Mehrgruppendesigns durchgeführt wurden (9, 19, 27, 34). Bei den restlichen 11 Untersuchungen handelt es sich demnach um "echte Ein-Gruppen-Designs" sowie um "Ein-Gruppen-Designs im erweiterten Sinne" (verfahrensinterne Vergleiche). Diese 11 ausgeschlossenen Studien setzen sich wie folgt zusammen: Von den insgesamt 10 Untersuchungen an kurzen und mittelfristigen Therapien mit guter methodischer Qualität wurden sieben Studien (70.0%) ausgeschlossen; von den insgesamt fünf Langzeittherapiestudien mit guter methodischer Qualität vier Studien (80.0%).

Die bisherige Vorgehensweise, bei der Studien zu kürzeren Therapiedauern mit Langzeittherapiestudien in ihrem Abschneiden auf den einzelnen Kriterien miteinander verglichen wurden, soll grundsätzlich beibehalten werden. Aufgrund der verhältnismäßig kleinen Anzahl an Studien (n=4), die für die folgenden Analysen noch zur Verfügung stehen, werden jedoch folgende Modifikationen vorgenommen:

- Die Dichotomisierung der Kriterienstufen wird aufgehoben, stattdessen werden die Verteilungen der vier Studien über alle drei Ratingstufen betrachtet; so ist es möglich, auch über die oberen beiden Kriterienstufen („1“ und „2“) differenziertere Aussagen zu treffen.
- Da es sich bei den vier Studien ausschließlich um Kurzzeit- und Langzeittherapiestudien handelt, fällt die Zusammenfassung der Studien zu mittelfristigen Therapiedauern mit den Kurzzeittherapiestudien weg, und es werden aufgrund dessen nur die Extremgruppen miteinander verglichen.

- Auf Effektstärkenberechnungen wird aufgrund der kleinen Stichprobengröße verzichtet, ebenso auf grafische Darstellungen.
- Im Rahmen der Beschreibung der Kreuztabellen werden die einzelnen Studien herangezogen und das Zustandekommen der einzelnen Bewertungen kommentiert; da dies der feinanalytischen Auswertung aus Kapitel 4.4 weitestgehend entspricht, wird die Separation in einen allgemeinanalytischen und einen feinanalytischen Teil damit aufgehoben.

Um die kriterienbezogenen Analysen anschaulicher zu gestalten, wird zunächst eine jeweils kurze Beschreibung der einzelnen Studien vorangestellt:

Studie 9

Titel/Autoren: *Controlled trial of short- and long-term effect of psychological treatment of post-partum depression. I. Impact on maternal mood (Cooper, Murray, Wilson & Romaniuk, 2003)*[66].

In dieser britischen Studie wurden zwei Therapieformen und eine weitere Bedingung, die in der vorliegenden Arbeit als aktive Kontrollbedingung deklariert wurde, jeweils mit TAU verglichen. Dabei handelte es sich um tiefenpsychologisch fundierte Psychotherapie (*psychodynamic therapy*) und kognitiv-behaviorale Therapie (*cognitive-behavioural therapy*), als aktive Kontrollbedingung fungierte *non-directive counselling*. Die Kontrollbedingung TAU umfasste die hausärztliche Routinebehandlung. Die Untersuchungsgruppe setzte sich aus *N*=193

[66] Diese Studie wurde im Rahmen der Studienrecherche (vgl. Kap. 3.1.2) in unterschiedlichen Reviews und Metanalysen gefunden (u.a. Bortolotti et al., 2008; Cuijpers et al., 2008; Leibing et al., 2005; Leichsenring et al., 2004) sowie im Cochrane Review zu *Short-term psychodynamic psychotherapies for common mental disorders* (Abbass et al., 2006).

Frauen mit einer postnatalen Depression zusammen, die via Randomisierung auf die vier genannten Treatments aufgeteilt wurden. Die Behandlungen erfolgten durch sechs Studientherapeuten und im Hause der Frauen. Bei allen Behandlungen handelte es sich um Kurzzeitbehandlungen von 10 wöchentlichen Sitzungen (Individualtherapie). Daten wurden zur Baseline und zum Postzeitpunkt erhoben, außerdem zu mehreren Katamnesezeitpunkten bis zu viereinhalb Jahren nach Behandlungsende. Als Outcomemaße wurden das Selbstbeurteilungsinstrument Edinburgh Postnatal Depression Scale (EPDS) sowie das Structured Clinical Interview for DSM-III-R (SCID) Depression Section (um auf evtl. Diagnosefreiheit nach Beendigung der Behandlung zu prüfen) eingesetzt. Die Behandlungen erfolgten manualisiert.

In der Bewertung mittels des WBP-Kriterienkatalogs schneidet diese Studie mit gutem Ergebnis hinsichtlich der internen Validität ab, auf der externen Validitätsdimension fällt diese Studie hingegen durch.

Studie 19

Titel/Autoren: *A randomized trial of the effect of four forms of psychotherapy on depressive and anxiety disorders: Design, methods, and results on the effectiveness of short-term psychodynamic psychotherapy and solution-focused therapy during a one-year follow-up (Knekt & Lindfors, 2004); Randomized trial on the effectiveness of long- and short-term psychodynamic psychotherapy and solution-focused therapy on psychiatric symptoms during a 3-year follow-up (Knekt, Lindfors, Härkänen et al., 2008);*

Effectiveness of short-term and long-term psychotherapy on work ability and functional capacity--A randomized clinical trial on depressive

and anxiety disorders (Knekt, Lindfors, Laaksonen, Haaramo, Järvikoski, & Raitasalo, 2008)[67].

Diese finnische Studie wird in Fachkreisen oftmals mit dem Titel „Helsinki Studie“ (bzw. *Helsinki Psychotherapy Study*) abgekürzt (u.a. Fonagy et al., 2002). In dieser Untersuchung wurden tiefenpsychologisch fundierte und analytische Psychotherapie (*short-term psychodynamic* und *long-term psychodynamic psychotherapy*) sowie systemische Therapie (*solution-focused therapy*) miteinander verglichen. Die im Titel einer Publikation erwähnte vierte Vergleichsbehandlung (s.o.) bezieht sich auf einen Treatmentarm, in dem klassische Psychoanalysen durchgeführt wurden, zu diesem wurden jedoch in den gesichteten Publikationen noch keine Ergebnisse berichtet.

Insgesamt *N*= 326 Patienten mit Angst- und/oder affektiven Störungen (plus komorbiden Störungen) wurden auf die drei Treatmentarme via Randomisierung aufgeteilt. Die Dauer der tiefenpsychologisch fundierten Psychotherapie belief sich mit im Mittel 18.5 Sitzungen auf durchschnittlich 5.7 Monate. Die Sitzungen fanden 1/Woche statt. Die systemische Therapie dauerte mit durchschnittlich 7.5 Monaten (*M* 9.8 Sitzungen) etwas länger, die Sitzungen fanden alle 2 bis 3 Wochen statt. Bei beiden Behandlungen handelt es sich damit um Kurzzeitbehandlungen. Die analytische Psychotherapie belief sich durchschnittlich auf 232 Sitzungen (*M* 31.3 Monate) und fällt damit im hiesigen Kontext unter die Langzeitbehandlungen. Einzig die systemische Therapie erfolgte manualisiert und unter Adherence-Kontrollen.

Daten wurden zu Beginn der Behandlungen erhoben sowie in Abständen von 2 bis 12 Monaten über den gesamten Therapieverlauf und den Katamnesezeitraum. Die letzte Erhe-

[67] Diese Studie wird in Lehrbüchern (z.B. Benecke, 2014a) und zahlreichen Reviews rezipiert (u.a. Abbass et al., 2006; Fonagy et al., 2002; Fonagy et al., 2005; Leichsenring, 2009a) und gilt als eine der wenigen randomisierten Studien zu psychoanalytischen Langzeitbehandlungen (vgl. Benecke, 2014a; Fonagy, 2009).

bung erfolgte 3 Jahre nach Behandlungsbeginn; zu diesem Zeitpunkt befanden sich drei der 128 Patienten im Langzeittherapiearm noch in Behandlung. Damit differieren die mittleren Katamnesezeiträume zwischen den Treatmentarmen zwischen ca. 4.5 Monaten und 2.5 Jahren. Als primäre Outcomemaße wurden das BDI, die HAM-D, die Symptom Checklist Anxiety Scale (SCL-90-Anx), die Hamilton Anxiety Rating Scale (HARS), der Work Ability Index (WAI), die Work-subscale der Social Adjustment Scale Self-Report (SAS-SR) und die Perceived Psychological Functioning Scale verwendet. Außerdem wurde der Arbeitsausfall erhoben.

In der Bewertung mit Hilfe des WBP-Kriterienkatalogs ist dies die einzige Studie, die sowohl auf der internen als auch auf der externen Validitätsdimension mit gutem Ergebnis abgeschnitten hat.

Studie 27

Titel/Autoren: *Randomized Controlled Trial Comparing Brief Dynamic and Supportive Therapy with Waiting List Condition in Minor Depressive Disorders (Maina, Forner & Bogetto, 2005)*[68].

In dieser italienischen Studie erfolgte ein Vergleich zwischen tiefenpsychologisch fundierter Psychotherapie (*brief dynamic therapy*), supportiver Psychotherapie und einer Wartlistengruppe. Die supportive Therapie wurde in diesem Fall *nicht* als aktive Kontrollbedingung, sondern als nicht etablierte Vergleichsbehandlung bewertet (vgl. Kap. 3.2.4.).
Die tiefenpsychologisch fundierte Behandlung belief sich im Mittel auf 19.6 Sitzungen und dauerte durchschnittlich 9.4 Monate. Die supportive Psychotherapie dauert mit durchschnitt-

[68] Diese Studie wurde ebenfalls im Rahmen der Studienrecherche (vgl. Kap. 3.1.2) in unterschiedlichen Reviews gefunden (u.a. Cuijpers et al., 2008; Abbass et al., 2006; Leichsenring 2009a).

lich 18.6 Sitzungen und 8.7 Monaten nur wenig kürzer. Beide Kurzzeittherapien erfolgten manualisiert. Die Wartliste wurde über 9 Monate geführt.

Insgesamt *N*=30 Patienten wurden den drei Treatmentarmen randomisiert zugeteilt. Alle 30 Patienten litten unter einer affektiven Störung (Dysthymia, Minor Depression, Anpassungsstörung mit depressiver Verstimmung). Outcomemaße wurden zu Beginn der Therapie, sowie zum Postzeitpunkt erhoben, zudem erfolgte eine Katamnesemessung 6 Monate nach Therapieende, wobei zu diesem Messzeitpunkt nur noch Daten der beiden Behandlungsgruppen erhoben wurden. Als primäres Outcomemaß fungierte die HAM-D.

In der Bewertung mittels des WBP-Kriterienkatalogs schneidet diese Studie mit gutem Ergebnis auf der internen Validitätsdimension ab, auf der externen Validitätsdimension fällt diese Studie durch.

Studie 34

Titel/Autoren: *Psychodynamic interpersonal therapy by inexperienced therapists in a naturalistic setting: a pilot study (Shaw, Margison, Guthrie & Tomenson, 2001)*[69].

Bei dieser britischen Studie handelt es sich um einen Vergleich zwischen *psychodynamic interpersonal therapy* (subsumiert unter: tiefenpsychologisch fundierter Psychotherapie) und einer Wartelistengruppe. Die Zuteilung zu den beiden Bedingungen erfolgte randomisiert. Die tiefenpsychologisch fundierte Behandlung umfasste 10-12 wöchentliche Sitzungen und fällt damit unter die Kurzzeitbehandlungen. Soweit es der Studie zu entnehmen war, verliefen die Behandlungen unter Anwendung eines eigens für die Studie modifizierten Therapiemodells

[69] Diese Studie wurde in keiner der im Rahmen der Handsuche (vgl. Kap. 3.1.2) gesichteten Reviews gefunden, sondern gelangte erst über die digitale Recherche in den Studienpool.

(Hobson's Conversational model), das für die hiesige Kodierung als manualähnliche Behandlungsrichtlinie betrachtet wurde.

Bei der Untersuchungsgruppe (*N*=54) handelt es sich um eine diagnoseheterogene Patientengruppe (detailliert: siehe Anhang F), die von in Ausbildung befindlichen Therapeuten behandelt wurde. Outcomemaße wurden zu Beginn der Therapie, sowie zum Postzeitpunkt erhoben. Als Outcomemaße fungierten das IIP und die SCL-90-R.

In der Beurteilung mit Hilfe des WBP-Kriterienkatalogs fällt diese Studie aufgrund zu vieler „3"-er Bewertungen auf der internen Validitätsdimension durch. Auf der externen Validitätsdimension schneidet sie mit positivem Ergebnis ab.

Im nun folgenden Teil werden die vier charakterisierten Studien – eingeteilt in drei Kurzzeit- und eine Langzeittherapiestudie – in ihrem Abschneiden auf den 12 Kriterien der internen Validität miteinander verglichen.

Spezifizierung der Ein- und Ausschlusskriterien (Kriterium B.1.)

Für Kriterium B.1., mit dem die eindeutige Spezifizierung der probandenbezogenen Ein- und Ausschlusskriterien bewertet wird, ergibt sich eine Verteilung der vier Studien, die kein schlechteres Abschneiden der Langzeittherapiestudie im Vergleich zu den drei Kurzzeittherapiestudien auf diesem Kriterium nahelegt (vgl. Tabelle 39).

Tabelle 39: Kreuztabelle über „Spezifizierung der Ein- und Ausschlusskriterien“ (Kriterium B.1.) und Behandlungslänge (*n*=4)

		B.1.			
		optimal *n*	zufriedenstellend *n*	ungenügend *n*	Gesamt *n*
Behandlungslänge	KZT	2	-	1	3
	LZT	1	-	-	1
	Gesamt *n*	3	-	1	4

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie.

Von den drei Kurzzeittherapieuntersuchungen schneidet eine Studie (34) mit „3“ ab. In dieser Studie wurden keine besonderen Ein-/Ausschlusskriterien aufgestellt („The criteria for placing the patients on the waiting list were not formalized and so this study reflects the practice typical of NHS service“ [Shaw et al., 2001, S. 89f.]). Jedoch ist im Weiteren die Rede von „geeigneten“ Studienpatienten: „Of the seventy patients eligible, ten were taken on by other members of staff . . .” (S. 90). Die Operationalisierung der „Eignung“ wird jedoch nicht weiter ausgeführt, so dass dies in der Kodierung von B.1. als eine uneindeutige Spezifizierung der Ein-/Ausschlusskriterien betrachtet wurde. Gleiches trifft in derselben Studie auf den Ausschluss von Patienten zu, die als dringend behandlungsbedürftig eingestuft wurden – die Feststellung der Dringlichkeit wurde nicht näher spezifiziert und lag unformalisiert im Ermessen des diagnostizierenden Klinikers. In den anderen drei Studien (9, 19, 27), darunter die eine Langzeittherapiestudie, sind die Ein-/Ausschlusskriterien hingegen eindeutig spezifiziert worden, selbst dann, wenn sie sich, wie in Studie 9, eher auf allgemeinere Faktoren statt auf spezielle Diagnosen bezogen. In dieser Studie bestand die Untersuchungsstichprobe ausschließlich aus Müttern mit einer postnatalen Depression, die Ein- und Ausschlusskriterien bezogen sich primär auf Aspekte, wie die Kinderanzahl (Einschlusskriterium: es musste das erste Kind sein) und den Ausschluss von Frauen mit einer Frühgeburt, einer Zwillingsgeburt oder einem behinderten Kind. In den beiden Studien 19 und 27 umfassten die Einschlusskrite-

rien Aspekte, wie das Alter, die Hauptdiagnosen, die bestehende Dauer der vorliegenden Störung, dem Strukturniveau sowie der Symptomschwere. Ausschlusskriterien bezogen sich ebenfalls auf bestimmte Diagnosegruppen (Psychosen, schwere Persönlichkeitsstörungen, Substanzmissbrauch, organische Hirnschädigungen etc.), auf die aktuelle Suizidneigung, die derzeitige Einnahme von Psychopharmaka sowie auf psychotherapeutische Behandlungen, die weniger als 2 Jahre zurücklagen.

Erhebung der Ein- und Ausschlusskriterien mittels valider Methoden (Kriterium B.2.)

Kriterium B.2. fragt danach, inwieweit die spezifizierten Ein- und Ausschlusskriterien (siehe B.1.) mittels valider Methoden erhoben wurden. Tabelle 40 stellt die Verteilung der vier Studien auf diesem Kriterium dar. Auch hier schneidet die Kurzzeittherapiestudie 34 mit „3“ ab, da weder die oben bereits eingeführte „Eignung“ der Patienten für die Studie noch das Ausschlusskriterium der „dringenden Behandlungsbedürftigkeit“ näher spezifiziert wurde; beide Aspekte wurden somit auch nicht mittels valider Verfahren erhoben.

Tabelle 40: Kreuztabelle über „Erhebung der Ein- und Ausschlusskriterien mittels valider Methoden“ (Kriterium B.2.) und Behandlungslänge (*n*=4)

		B.2.			
		optimal *n*	zufriedenstellend *n*	ungenügend *n*	Gesamt *n*
Behandlungslänge	KZT	2	-	1	3
	LZT	-	1	-	1
	Gesamt *n*	2	1	1	4

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie.

Bei den anderen drei Studien (9, 19, 27), wieder inklusive der Langzeittherapiestudie (19), wurden die Ein-/Ausschlusskriterien mittels valider Verfahren erhoben (HAM-D, CGI, SCID-I und -II). Einzig ein in der Langzeittherapiestudie aufgestelltes Einschlusskriterium konnte

hinsichtlich der Validität des Erhebungsinstruments nicht eingeschätzt werden: „They had to . . . be estimated in a psychodynamic assessment interview of suffering from neurosis to higher-level borderline disorder, according to Kernberg's classification of personality organization (1996)" (Knekt, Lindfors, Härkänen et al., 2008, S. 690). Mittels welcher Skala o.ä. diese Einschätzung vorgenommen wurde, bleibt unklar, aus diesem Grund schneidet diese Studien hier mit „2" („zufriedenstellend") statt mit einer „1" ab.

Operationale Definition der Interventionen (Kriterium B.3.)

Kriterium B.3. bemisst, inwieweit die operationale Definition der Intervention(en) ausführlich genug dargestellt ist, z.B. durch Verweis auf ein Therapiemanual oder aber einer detaillierten Beschreibung des therapeutischen Vorgehens. Zur Erinnerung: Mit einer „1" wurden Studien bewertet, in denen offenkundig Manuale oder Behandlungsrichtlinien angewendet wurden, mit einer „2" wurden Studien bewertet, in denen lediglich exemplarisch auf Manuale oder auf Behandlungsrichtlinien verwiesen wurde, aus denen jedoch nicht hervorgeht, dass diese auch tatsächlich angewendet wurden (vgl. Kap. 3.2.2).

Tabelle 41 stellt die Verteilung der vier Studien dar. Dabei zeigt sich, dass die Langzeittherapiestudie (19) als einzige der vier Studien mit „ungenügend" (Stufe „3") abschneidet, und dies aus folgendem Grund: Der Studienpublikation von Knekt, Lindfors, Laaksonen et al. (2008) ist zu entnehmen, dass auf Manuale für die beiden psychodynamischen Treatmentarme explizit verzichtet wurde. Beide Therapieformen (hier als tiefenpsychologisch fundierte und analytische Psychotherapie deklariert) „were conducted in accordance with clinical practice, where the interventions might be modified according to patients' needs within the psychodynamic framework" (Knekt, Lindfors, Laaksonen et al., 2008, S. 97). Im Gegensatz dazu wurde die systemische Therapie in der Vergleichsbedingung manualisiert durchgeführt (inklusive

Adherence-Kontrollen). Mit diesem Unterfangen wurde ein systematischer Unterschied zwischen den Bedingungen kreiert, der sich mindernd auf die interne Validität auswirken kann (vgl. Leichsenring et al., 2011) – daher die Bewertung mit „3“ in diesem Kriterium. In den restlichen drei Studien wurden durchweg Manuale bzw. Behandlungsrichtlinien angewendet.

Tabelle 41: Kreuztabelle über „Operationale Definition der Interventionen“ (Kriterium B.3.) und Behandlungslänge (*n*=4)

		B.3.			
		optimal *n*	zufriedenstellend *n*	ungenügend *n*	Gesamt *n*
Behandlungslänge	KZT	3	-	-	3
	LZT	-	-	1	1
	Gesamt *n*	3	-	1	4

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie.

Operationale Definition der Kontrollbedingungen (Kriterium B.4.)

Kriterium B.4. fordert die prospektive Festlegung sowie umfassende Beschreibung der Kontrollbedingung in einem Kontrollgruppendesign. Mit Kontrollbedingungen sind Wartelisten, Placebo-Kontrollgruppen, TAU oder aktive Kontrollgruppen gemeint. Tabelle 42 ist die Verteilung der vier Studien auf diesem Kriterium plus einer Missingkategorie zu entnehmen. Die Missingkategorie wurde für solche Studien gebildet, die kein klassisches Kontrollgruppendesign aufweisen (komparative Untersuchungen mit etablierten oder [noch] nicht etablierten Vergleichsbehandlungen) und die daher auf diesem Kriterium nicht bewertbar sind (vgl. Kap. 3.2.1).

Tabelle 42: Kreuztabelle über „Operationale Definition der Kontrollbedingungen“ (Kriterium B.4.) und Behandlungslänge (*n*=4)

B.4.						
		optimal *n*	zufrieden-stellend *n*	ungenügend *n*	Missing *n*	Gesamt *n*
Behand-lungs-länge	KZT	2	1	-	-	3
	LZT	-	-	-	1	1
	Gesamt *n*	2	1	-	1	4

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie.

Es zeigt sich eine Verteilung zuungunsten der Langzeittherapiestudie, die hier die Missingkategorie besetzt. Da es sich bei dieser Studie (19) um eine komparative Studie handelt, in der zwei psychodynamische Behandlungen – davon ein Arm mit Langzeitbehandlungen – mit systemischer Therapie verglichen werden, kann diese Studie auf diesem Kriterium nicht eindeutig kodiert werden. Da in den restlichen drei Kurzzeittherapiestudien klassische Kontrollgruppen in Form von TAU (9) und Wartelistengruppen (27, 34) umgesetzt wurden, sind diese Studien grundsätzlich kodierbar. Zwei der Studien (27, 34) schneiden „optimal“ ab, Studie 9 hingegen schneidet lediglich mit „2“ ab. In letzterer Studie wurde die TAU-Bedingung zwar prospektiv festgelegt, jedoch fällt die Beschreibung dieser Bedingung vergleichsweise knapp aus, so dass nicht genau eruiert werden kann, welche allgemeinen Wirkfaktoren durch diese Kontrollbedingung tatsächlich kontrolliert werden.

Strukturelle Äquivalenz bei Kontrollbedingungen (Kriterium B.5.)

Kriterium B.5. bemisst die strukturelle Äquivalenz der Kontrollbedingungen im Hinblick auf das zu überprüfende Treatment (vgl. Kap. 3.2.1). Auch hier sind mit Kontrollbedingungen ausschließlich klassische Kontrollgruppen gemeint und keine alternativen Therapiebedingungen. Die Verteilung der vier Studien wird durch Tabelle 43 dargestellt. Auch hier schneidet Studie 19 erwartungsgemäß mit einem Missingwert ab; die drei Kurzzeittherapiestudien

schneiden durchweg mit „3“ ab. Bei den beiden Studien 27 und 34 liegen Vergleiche mit Wartelisten vor, von dieser Art Kontrollbedingung geht ausschließlich die Kontrolle zeitbedingter Wirkfaktoren aus, wie Reifung oder Spontanremission. Da durch die „strukturelle Äquivalenz“ zwischen Kontroll- und Experimentalbedingung jedoch vielmehr die Angleichung der Randbedingungen gemeint ist und damit die Kontrolle unspezifischer Wirkfaktoren (vgl. Kap. 1.2.2 und 3.2.1), können Studien mit nur einer Warteliste als Kontrollbedingung hier lediglich mit Stufe „3“ abschneiden. In Studie 9, in der als Kontrollbedingung TAU fungierte, wird diese Bedingung derart knapp beschrieben, dass eine potentielle Äquivalenz der Randbedingungen nicht eingeschätzt werden konnte und aus diesem Grund mit „3“ bewertet wurde.

Tabelle 43: Kreuztabelle über „Strukturelle Äquivalenz bei Kontrollbedingungen“ (Kriterium B.5.) und Behandlungslänge (*n*=4)

		B.5.				
		optimal *n*	zufriedenstellend *n*	ungenügend *n*	Missing *n*	Gesamt *n*
Behandlungslänge	KZT	-	-	3	-	3
	LZT	-	-	-	1	1
	Gesamt *n*	-	-	3	1	4

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie.

Manualtreue, Treatment Integrity (Kriterium B.6.)

Durch Kriterium B.6. wird bemessen, inwieweit im Rahmen der jeweiligen Untersuchungen die Adherence (Manual- bzw. Therapietreue) festgestellt und kontrolliert wurde. Tabelle 44 gibt die Verteilung der vier Studien auf diesem Kriterium wieder.

Tabelle 44: Kreuztabelle über „Manualtreue, Treatment Integrity“ (Kriterium B.6.) und Behandlungslänge (*n*=4)

		B.6.			
		optimal *n*	zufriedenstellend *n*	ungenügend *n*	Gesamt *n*
Behand-lungs-länge	KZT	1	1	1	3
	LZT	-	-	1	1
	Gesamt *n*	1	1	2	4

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie.

Wie bereits zu Kriterium B.3. (Bemessung der operationalen Definition der Interventionen) beschrieben, wird in der Langzeittherapiestudie (19) ausschließlich in der Vergleichsbedingung (systemische Therapie) manualisiert vorgegangen, in den beiden psychodynamischen Treatmentarmen wurde auf Manuale verzichtet. Ebenso verhält es sich in Bezug auf die Adherence-Kontrollen, die ebenfalls ausschließlich im manualisierten Treatmentarm realisiert wurden. Somit gilt auch im Hinblick auf dieses Kriterium, dass die interne Validität allein deswegen gefährdet ist, weil die Strategie der Adherence-Kontrolle nicht einheitlich durchgeführt wird.

In zwei der drei Kurzzeittherapiestudien (9, 27) wurden Adherence-Kontrollen durchgeführt, in Studie 9 wurde diese jedoch lediglich durch Fragebögen belegt, die von den Patientinnen ausgefüllt wurden, so dass diese Studie hier mit „2“ bewertet wurde. In Studie 34 wurde hingegen auf Adherence-Kontrollen verzichtet.

Zulässigkeit, Dokumentation, Analyse des Einflusses begleitender nicht-randomisierter Interventionen (Kriterium B.7.)

Durch Kriterium B.7. wird bemessen, inwieweit begleitende Interventionen (vor allem Pharmakotherapie), zu denen keine randomisierte Zuteilung erfolgte, ausgeschlossen wurden. Tabelle 45 illustriert die Verteilung der vier Studien auf diesem Kriterium.

Tabelle 45: Kreuztabelle über „Zulässigkeit, Dokumentation, Analyse des Einflusses begleitender nicht-randomisierter Interventionen“ (Kriterium B.7.) und Behandlungslänge (*n*=4)

		B.7.			
		optimal *n*	zufriedenstellend *n*	ungenügend *n*	Gesamt *n*
Behand-lungs-länge	KZT	1	-	2	3
	LZT	-	-	1	1
	Gesamt *n*	1	-	3	4

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie.

In einer Kurzzeittherapiestudie (27) wurden jegliche begleitende und nicht randomisierte Interventionen ausgeschlossen, so dass diese Studie hier mit „1“ bewertet wurde. Die restlichen drei Studien – darunter die Langzeittherapiestudie – schnitten allesamt mit „ungenügend“ ab: In den Studien 9 und 34 fehlen jegliche Angaben zu diesem Aspekt, in der Langzeittherapiestudie (19) wurden zusätzliche Behandlungen über den gesamten Untersuchungszeitraum zugelassen und dokumentiert, jedoch im Anschluss nicht statistisch kontrolliert: „Finally, the extensive use of auxiliary treatment in the short-term therapy groups might possibly have distorted the results in favor of short-term-therapies“ (Knekt, Lindfors, Laaksonen et al., 2008, S. 104).

Gruppenzuweisung (Kriterium B.8.)

Durch Kriterium B.8. wird bemessen, ob und ggf. welcher Gruppenzuweisungsstrategien (Randomisierung, Parallelisierung etc.) sich innerhalb einer Studie bedient wird. Tabelle 46 illustriert die Verteilung der vier Studien auf diesem Kriterium.

Tabelle 46: Kreuztabelle über „Gruppenzuweisung“ (Kriterium B.8.) und Behandlungslänge (*n*=4)

		B.8.			
		optimal *n*	zufriedenstellend *n*	ungenügend *n*	Gesamt *n*
Behand-lungs-länge	KZT	1	2	-	3
	LZT	1	-	-	1
	Gesamt *n*	2	2	-	4

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie.

Alle vier Studien schneiden mit optimalem oder zufriedenstellendem Ergebnis auf diesem Kriterium ab. Die Zuteilung zu den Behandlungsgruppen erfolgte in zwei Untersuchungen (9, 19) – darunter auch die Langzeittherapiestudie – bei angemessener Stichprobengröße randomisiert. In den beiden anderen Studien (27, 34) erfolgte die Zuteilung zwar ebenfalls randomisiert, jedoch waren die Stichproben zu klein ($n < 30$ pro Gruppe), so dass die gleichverteilende Wirkung der Randomisierung im Hinblick auf bekannte und unbekannte Störfaktoren fraglich ist. Aus diesem Grund wurden die beiden zuerst genannten Studien jeweils mit Stufe „1“ bewertet, die beiden zuletzt genannten Studien (mit den zu geringen Stichprobenumfängen hingegen) mit „2“.

Vergleichbarkeit der Gruppen zur Baseline (Kriterium B.9.)

Durch Kriterium B.9. wird die Vergleichbarkeit der Gruppen zur Baseline hinsichtlich prognostisch relevanter Merkmale beurteilt. Tabelle 47 zeigt die Ergebnisse der vier Studien.

Tabelle 47: Kreuztabelle über „Vergleichbarkeit der Gruppen zur Baseline“ (Kriterium B.9.) und Behandlungslänge (*n*=4)

B.9.					
		optimal *n*	zufriedenstellend *n*	ungenügend *n*	Gesamt *n*
Behandlungslänge	KZT	1	1	1	3
	LZT	1	-	-	1
	Gesamt *n*	2	1	1	4

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie.

Die Langzeittherapiestudie (19) schneidet auf diesem Kriterium mit „1“ ab: In dieser Untersuchung wurden soziodemografische Variablen, wie das Alter, das Geschlecht, der berufliche und akademische Status sowie die Wohnsituation (alleinlebend *versus* nicht alleinlebend) erhoben; außerdem wurden klinische Faktoren erhoben, wie die vorhergehende Inanspruchnahme von Psychotherapie, Klinikaufenthalte oder die Einnahme von psychopharmakologischen Medikamenten. Zusätzlich wurden die drei Behandlungsgruppen (analytische Psychotherapie, tiefenpsychologisch fundierte Psychotherapie und systemische Therapie) hinsichtlich ihrer Zusammensetzung im Hinblick auf Diagnosegruppen und das Vorkommen von psychiatrischen Komorbiditäten verglichen. In all den genannten Variablen bestanden keine signifikanten Unterschiede zur Baseline.

Die drei Kurzzeittherapiestudien (9, 27, 34) streuen mit ihren Ergebnissen auf diesem Kriterium über die gesamte Bandbreite der drei Ratingstufen: Studie 27 schneidet mit optimalem Ergebnis ab, hierbei wurden die drei Gruppen (tiefenpsychologisch fundierte Psychotherapie, supportive Therapie und Warteliste) hinsichtlich der Merkmale Geschlecht, Alter und Bildungsgrad verglichen. Zusätzliche Vergleiche erfolgten im Hinblick auf die jeweils durchschnittlichen Ausprägungen auf der HAM-D, HARS als auch der Clinical Global Impression - Severity Scale (CGI-S). In den genannten Variablen ergaben sich keine signifikanten Unterschiede. In Studie 9 wurden ebenfalls prognostisch bedeutsame Faktoren (Alter, Bildungs-

grad, familiärer Status) erhoben und zusätzlich zwei Variablen, die den Autoren der Studie im Kontext der Behandlung postnataler Depressionen wichtig erschienen: Soziale Benachteiligung und negative Einstellung zur Mutterschaft. Im Hinblick auf die Variable „soziale Benachteiligung" zeigte sich ein Unterschied zwischen den Gruppen, der jedoch statistisch kontrolliert wurde, so dass die Untersuchung auf dem hiesigen Kriterium mit „2" bewertet wurde. Studie 34 schneidet mit „ungenügend" ab, da der Publikation keinerlei Bericht über die Überprüfung von Unterschieden zwischen den Untersuchungsgruppen zur Baseline zu entnehmen ist.

Anzahl der Messzeitpunkte und prospektive Messungen (Kriterium B.10.)

Tabelle 48 illustriert die Verteilung der vier Studien auf dem Kriterium B.10., mit dem die Anzahl der Messzeitpunkte sowie die Anwendung primär prospektiver Messungen (z.B. SCL-90-R, GBB) beurteilt wird. Für diese Darstellung wird die modifizierte Fassung des Kriteriums verwendet (vgl. Kap. 4.3).

Tabelle 48: Kreuztabelle über „Anzahl der Messzeitpunkte und prospektive Messungen" (Kriterium B.10.) und Behandlungslänge (*n*=4)

		B.10.			
		optimal *n*	zufriedenstellend *n*	ungenügend *n*	Gesamt *n*
Behandlungslänge	KZT	-	3	-	3
	LZT	-	1	-	1
	Gesamt *n*	-	4	-	4

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie.

Es zeigt sich, dass keinerlei Variabilität zwischen den Studien besteht und alle mit derselben Bewertung abschließen. In den drei Kurzzeittherapiestudien (9, 27, 34) wurden entweder Prä-Post- oder aber Prä-Post-Katamnesemessungen vorgenommen, in der Langzeittherapiestudie (19) wurden die Messzeitpunkte nicht an den individuellen Therapiephasen, sondern an der

Realzeit bemessen, d.h. es existieren fixe Messzeitpunkte mit variierenden Katamnesezeiträumen. Nach der Rekodierung der Studien mittels des modifizierten B.10.-Kriteriums wurde die Langzeittherapiestudie entsprechend mit „2“ bewertet (vorher mit einem Missingwert).

Follow-up-Messung (Kriterium B.11.)

Tabelle 49 enthält die Verteilung der Studien auf dem Kriterium B.11.. Dieses Kriterium dient der Bewertung des Katamnesezeitraums und der Ausschöpfungsquote zur Katamnesemessung (vgl. Kap. 3.2.3).

Tabelle 49: Kreuztabelle über „Follow-up-Messung“ (Kriterium B.11.) und Behandlungslänge (*n*=4)

		B.7.			
		optimal *n*	zufriedenstellend *n*	ungenügend *n* (keine Katamnese)	Gesamt *n*
Behand-lungs-länge	KZT	1	1	1	3
	LZT	-	1	-	1
	Gesamt *n*	1	2	1	4

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie.

Auch hier verteilen sich die drei Kurzzeittherapiestudien über die gesamte Bandbreite der drei Ratingstufen: Studie 9 schneidet mit ihrem Katamnesezeitraum von ca. viereinhalb Jahren mit „1“ ab, Studie 27 mit 6 Monaten mit Stufe „2“. In Studie 34 wurde keine Katamnese durchgeführt, daher die Bewertung mit „3“. Die Langzeittherapiestudie (19) wurde mit „2“ bewertet, da zwar die Katamnesezeiträume für die Kurzzeittherapien (tiefenpsychologisch fundierte und systemische Therapie) mehr als 2 Jahre umfassen, der Katamnesezeitraum des Langzeittherapiearms ist jedoch mit nur wenigen Monaten sehr kurz. Die Bewertung lehnt sich damit an den kürzesten Zeitraum an.

Veränderung und Zielerreichung auf den Zielkriterien (Kriterium B.12.)

Durch Kriterium B.12., das samt den dazugehörigen Kodierregeln in Kapitel 3.2.3 ausführlich beschrieben wurde, wird bemessen, inwieweit innerhalb der Studienpublikationen unterschiedliche Zugänge der Wirksamkeitsabbildung gewählt und dargestellt werden (statistische Signifikanz, Effekte, klinische Signifikanz via RCI und Cutoff-Werte). Aus Tabelle 50 wird ersichtlich, dass der Studienvergleich hier zugunsten der Langzeittherapiestudie (19) ausfällt. In dieser wurden Veränderungen mit Hilfe von Signifikanzberechnungen, Effektstärken und Zielerreichungen mittels Cutoff-Werten bestimmt. Die drei Kurzzeittherapiestudien schneiden allesamt mit „2“ ab: In den Studien 9 und 34 werden nur Veränderungen über Signifikanzberechnungen und Effektstärken berichtet, in Studie 27 werden zwar Veränderungs- und Zielerreichungsmaße berichtet, jedoch wird über Effektstärken nur sehr spärlich informiert (keine Zwischengruppeneffekte), so dass diese Studie hier ebenfalls mit „2“ bewertet wurde.

Tabelle 50: Kreuztabelle über „Veränderung und Zielerreichung auf Zielkriterien“ (Kriterium B.12.) und Behandlungslänge (*n*=4)

		B.12.			
		optimal *n*	zufriedenstellend *n*	ungenügend *n*	Gesamt *n*
Behandlungslänge	KZT	-	3	-	3
	LZT	1	-	-	1
	Gesamt *n*	1	3	-	4

Anmerkung: KZT: Kurzzeittherapie, LZT: Langzeittherapie.

5 Diskussion

Der erste Eindruck, den das Flowdiagramm der Studien (Abbildung 27, S. 263) nahelegt, vermittelt vom WBP-Kriterienkatalog zunächst das Bild eines recht strengen Bewertungsrasters, vor allem, was die allgemeine methodische Qualitätsbewertung als auch die interne Validitätsbewertung betrifft. Dieser erste Eindruck verflüchtigt sich in großen Teilen, wenn man sich die Gründe für die Negativbewertungen auf einzelnen Kriterien genauer ansieht. Es zeigt sich dann, dass der Maßstab, den die Kriterien etablieren, kein unrealistischer ist, der von Studien nicht erfüllt werden kann. Dies trifft sowohl auf psychodynamische Wirksamkeitsstudien im Allgemeinen also auch auf Langzeittherapiestudien (> 100 Stunden) im Besonderen zu. Zur Fundierung dieser Aussage werden die zuvor dargelegten Ergebnisse (Kap. 4) vor dem Hintergrund allgemeiner Erkenntnisse aus der Evaluationsforschung sowie unter Hinzuziehung von empirischen Befunden aus der Psychotherapieforschung reflektiert. Dazu wird vor allem auch ein Blick in die jüngere Studienlage geworfen, die durch die Begrenzung des Publikationszeitraums der hier kodierten Studien (1999-2009) nicht mehr Gegenstand der Untersuchung war.

Um den Ausgangspunkt der vorliegenden Arbeit noch einmal zu skizzieren, sollen eingangs noch mal ein ehemaliges Mitglied des WBP, Dietmar Schulte, sowie ein aktuelles Mitglied, Falk Leichsenring, zu Wort kommen. Beide waren an der Entwicklung, Veröffentlichung sowie an den vorgenommenen Revidierungen der Verfahrensregeln (WBP, 2007, 2009b, 2010) aktiv beteiligt. Das erste Zitat stammt aus einem Interview mit Dietmar Schulte, Gründungsmitglied des WBP bis 2011, zudem Vorsitzender des WBP in der Zeit von 2005 bis 2011:

> Der Beirat hat außerdem ausdrücklich festgehalten: Sollte sich herausstellen, dass die Kriterien unrealistisch sind, wird eine Überarbeitung stattfinden. (Deutsches Ärzteblatt, 2008a, S. 389)

Schulte betont damit die Vorläufigkeit der Kriterien, die revidiert werden sollen, wenn sich die Anforderungen der Kriterien als unerfüllbar erweisen. Dieser Hinweis ist auf zweierlei Art und Weisen zu verstehen: Zum einen ist es denkbar, dass ein Großteil der existierenden empirischen Evidenz dem heutigen Wissensstand an hochwertiger evaluatorischer Psychotherapiewirksamkeitsforschung *noch* nicht genügt. An dieser Stelle müsste selbst eine an der Idealnorm orientierte Begutachtung (*versus* normorientiert), wie sie der WBP mit dem Kriterienkatalog vornimmt (vgl. Kap. 1.2.2), einen Weg finden, dieses eher wissenschaftshistorisch verschuldete Defizit mit Hilfe einer Interimslösung zu überbrücken. In Teilen tut er dies bereits, indem er für ältere Studien, die vor 1990 publiziert wurden, die Möglichkeit einräumt, nach einem weniger strengen Bewertungsmaßstab begutachtet zu werden (vgl. WBP, 2010). Die zweite Verstehensweise des Zitats bezieht sich auf etwas, das der WBP in seinen Verfahrensregeln als ausnahmsweise „Kollisionen“ zwischen bestimmten Psychotherapieverfahren oder -methoden und den Forschungsstrategien, die durch die Kriterien gefordert werden, bezeichnet (vgl. WBP, 2010, S. 9). Gemeint sind Kollisionen, die nur selten vorkommen, etwa bei der Begutachtung einer Behandlungsmethode für ein sehr seltenes Störungsbild, wodurch eine Kollision mit der laut Kriterienkatalog geforderten Stichprobengröße hervorgerufen werden könnte. Für solche Fälle entpuppen sich einige der Kriterien vielleicht als *unrealistisch*, jedoch räumt der WBP der Begutachtung unter diesen Umständen die Möglichkeit ein, von den ansonsten geltenden Kriterien abzuweichen. Was aber geschieht, wenn ein bestimmter Gegenstand regelhaft mit den anzuwendenden Kriterien kollidiert?

Das zweite angekündigte Zitat setzt inhaltlich an diesem Punkt an und stammt von Falk Leichsenring, seit 2002 als Mitglied des WBP tätig:

> Im Kern handelt es sich bei dem Kriterienkatalog des Methodenpapiers um ein neu entwickeltes diagnostisches Instrument, und bei der Beurteilung von Studien und von psychotherapeutischen Verfahren/Methoden anhand dieser Kriterien um ein Ratingverfahren. Neue diagnostische Verfahren müssen sich empirisch bewähren . . . . Es muss sich auch erst noch zeigen, inwieweit der Kriterienkatalog allen Fragestellungen der Psychotherapie gerecht wird, etwa naturalistischen Studien oder Studien zur Langzeittherapie. (Leichsenring, 2008, S. 120)

Genau genommen fordert Leichsenring damit, dass der Kriterienkatalog sowohl für naturalistische als auch für Langzeittherapiestudien *realistisch* (im Sinne Schultes) sein sollte. Naturalistische und Langzeittherapiestudien stellen also keine Ausnahmefälle dar, denen mittels Ausnahmeregelungen in Bezug auf die ansonsten geltenden Kriterien begegnet werden darf.

Der WBP hat mit Veröffentlichung des Methodenpapiers (2007) die bis dahin bestehende Beschränkung der wissenschaftlichen Anerkennung der psychodynamischen Psychotherapie auf Therapielängen *bis zu* 100 Sitzungen aufgehoben (vgl. Kap. 2). Er begründete diesen Schritt mit der im Methodenpapier verankerten Unterscheidung zwischen Verfahren und Methoden. In diesem Zusammenhang wurde argumentiert, dass für unterschiedliche psychotherapeutische Methoden, die jedoch einem Verfahren angehören, keine einzelnen Wirksamkeitsnachweise zu erbringen wären. Dies gälte, so der WBP, ebenso für Variationen der Behandlungsdauer innerhalb eines Verfahrens, wie es bei der psychodynamischen Psychotherapie der Fall ist (vgl. WBP, 2008a). Gleichsam betonte der WBP in seiner Stellungnahme zur wissenschaftlichen Anerkennung psychodynamischer Psychotherapie, dass Langzeitbehandlungen besondere Forschungsfragen aufwerfen (vgl. WBP, 2005), denen offenkundig mit den Vorläuferpapieren des Methodenpapiers noch nicht angemessen begegnet wurde.

Jenseits der vom WBP abgegebenen Erklärung, mit der er die Ausweitung der wissenschaftlichen Anerkennung der psychodynamischen Psychotherapie auf Langzeitbehandlungen (> 100 Sitzungen) begründet, wurde in der vorliegenden Arbeit – und gewissermaßen Leichsenrings Zitat entsprechend – folgende Haltung eingenommen: Ein Instrument zur Messung

der methodischen Qualität und Validität von Wirksamkeitsstudien sollte der Beforschung der Wirksamkeit von Langzeitbehandlungen – trotzdem sie besondere Forschungsfragen aufwerfen mag – angemessen begegnen können. Warum ist das wichtig? Hier sollen lediglich zwei Gründe angeführt werden: Zum einen gibt es Hinweise darauf, dass vor allem Patienten mit schweren strukturellen sowie komorbiden oder chronischen Störungen längerer psychotherapeutischer Behandlungen bedürfen, als Patienten mit Störungen leichteren Schweregrades (u.a. Kopta, Howard, Lowry & Beutler, 1994; Leichsenring, Abbass, Luyten, Hilsenroth & Rabung, 2013; Perry, Banon & Ianni, 1999; Westen et al., 2004). Entsprechend der Epidemiologie komplexer Störungen oder chronischer Störungsverläufe nehmen Langzeitbehandlungen in der psychotherapeutischen Versorgung für eben diese Patientenklientel einen zentralen und aus der Versorgung nicht wegzudenkenden Stellenwert ein (vgl. Matzat, 2014; Metzner & Schlösser, 2014; Richter, 2014). Ferner legen erste Befunde zur Langzeitwirksamkeit analytischer Langzeittherapie nahe, dass diese im Vergleich zu kürzeren Behandlungen zu nachhaltigeren Erfolgen führt (vgl. Benecke, 2014a; Huber, Zimmermann et al., 2012; Leichsenring & Rabung, 2011). Diese Befundlage gewinnt nicht nur aus gesundheitsökonomischer Perspektive, sondern vor allem aus der Perspektive der Patienten besonderes Gewicht: Ein Patient, der sich in eine analytische Langzeittherapie begibt, wird bis zu dreieinhalbmal so viel Zeit und Energie in seine Behandlung investieren, als würde er sich in eine tiefenpsychologisch fundierte oder in eine Verhaltenstherapie begeben. Dieser zeitliche und psychische Mehraufwand sollte sich für einen Patienten im Mindesten durch eine langfristige Wirksamkeit rentieren. Gemessen an dieser Forderung von stabilen Therapieerfolgen sowie an der Bedarfslage bzgl. langfristiger Behandlungen für besonders komplexe Störungsbilder muss sich die analytische Langzeittherapie mittels qualitativ hochwertiger Studien ihren Platz sowohl in der Ausbildungslandschaft als auch in der psychotherapeutischen Versorgung auch künftig sichern (vgl. Benecke, 2014b; Fonagy, 2009). Aus diesen Gründen soll hier dem Vorgehen

des WBP, in der Anerkennung des wissenschaftlichen Status' von psychodynamischer Psychotherapie keinen expliziten Unterschied zwischen Kurz- und Langzeitbehandlungen vorzunehmen, nicht gefolgt werden. Stattdessen wird davon ausgegangen, dass sich die analytische Psychotherapie als Langzeitverfahren ebenso den Wirksamkeitsanforderungen stellen muss, wie es allen anderen Verfahren abverlangt wird (vgl. Benecke, 2014b; Leichsenring & Rabung, 2011). Dementsprechend ist, in Anlehnung an Leichsenrings Zitat (s.o.), auch von dem WBP-Kriterienkatalog zu fordern, dass er für die wissenschaftliche Begutachtung für eben diesen Bereich geeignet ist.

Im ersten Teil der Arbeit wurde ausführlich hergeleitet, inwiefern die RCT-Methodologie mit dem Gegenstand der Langzeitbehandlungen nur schwerlich in Übereinstimmung zu bringen ist. In dieser Herleitung kamen allem voran kritische Stimmen zu Wort, die die Inkompatibilität mit dem Bereich der Langzeittherapie hauptsächlich in internen Validtätssicherungsstrategien, wie der Randomisierung, der Manualisierung sowie der Wahl klassischer Kontrollbedingungen (Wartelistengruppen oder Placebo-Kontrollgruppen) sehen (vgl. Kap. 1). Die Fragestellung der Arbeit ging daher von der Annahme aus, dass ein Kriterienpapier, das sich in der Bewertung der internen Validität von Studien ausschließlich an die RCT-Methodologie anlehnt, einen Bias im Sinne einer systematischen Benachteiligung von Langzeittherapiestudien im Vergleich zu Kurzzeittherapiestudien erzeugen könnte (vgl. Kap. 2). Die WBP-Kriterien dahingehend zu untersuchen war das Ziel der Arbeit.

Wie sind die Ergebnisse der Anwendung der Kriterien auf die 41 Studien nun zu interpretieren? Im folgenden Teil werden dazu zunächst die Ergebnisse zur allgemeinen methodischen Studienqualität (Kap. 5.1) sowie zur internen Validität (Kap. 5.2) diskutiert. Im Anschluss erfolgt eine kritische Würdigung der Verfahrensregeln und des Kriterienkatalogs (Kap. 5.3).

Im zuletzt genannten Kapitel wird es u.a. um eine Reflektion des Konzepts von interner und externer Validität gehen, das dem WBP-Kriterienkatalog zugrunde liegt (Kap. 5.3.1). In Kapitel 5.4 wird eine Einschätzung der Generalisierbarkeit der in der Arbeit dargelegten Ergebnisse und Schlussfolgerungen vorgenommen, in Kapitel 5.5 werden schließlich die Limitationen der Arbeit beleuchtet.

5.1 Die Dimension der allgemeinen methodischen Qualität

Der erste Blick auf die rein quantitative Auslese der drei Dimensionen des Kriterienkatalogs vermittelt zunächst den Eindruck, dass es sich vor allem bei der allgemeinen Qualitätsdimension sowie bei der internen Validitätsdimension um sehr strenge Kriterien handelt, deren Anforderungen kaum realistisch zu erfüllen sind: Von den 41 Studien erlangen mit 15 Studien lediglich etwas über ein Drittel das Prädikat guter methodischer Qualität. Dieser erste Blick muss jedoch vertieft werden: Im Rahmen der Datenerhebung und der Definition der Studienzielpopulation wurde für die Arbeit bewusst entschieden, möglichst *breite* Einschlusskriterien an die Zusammenstellung des zu kodierenden Studienpools zu legen (vgl. Kap. 3.1). Der Grund für die breite Auslese war es, dass die Ein- und Ausschlusskriterien keinesfalls mit den Kriterien des WBP-Kriterienkatalogs kollidieren sollten – oder einfacher ausgedrückt: Die Arbeit der Auslese sollte allein dem Kriterienkatalog überlassen werden. Der vergleichsweise breite Einschluss der Studien erzeugte somit einen Studienpool, der keineswegs etwa mit einem Studienpool gleichzusetzen ist, der in Metaanalysen zur Wirksamkeit psychodynamischer Psychotherapie eingehen würde. Im Unterschied zu metaanalytischen Zwecken wurde für die hiesige Untersuchung keine Vorselektion durch a priori festgelegte Qualitätskriterien vorgenommen. Dies könnte zumindest in Ansätzen die verhältnismäßig hohe Anzahl an Studien erklären, die durch die methodische Qualitätsdimension aussortiert wurde; ein Vergleichsmaßstab existiert für diese Einschätzung allerdings nicht. Von zentraler Bedeutung ist

hier jedoch nicht die rein quantitative Auslese, sondern vielmehr die Gründe, die zu den zahlreichen Studienausschlüssen geführt haben. Wie in Kapitel 4.3 gezeigt werden konnte, waren für die insgesamt 26 aufgrund mangelhafter methodischer Qualität ausscheidenden Studien primär vier der sechs K.O.-Kriterien dieser Dimension verantwortlich. Es wird sich im Folgenden hauptsächlich auf diejenigen beiden K.O.-Kriterien konzentriert, die über die höchsten Ausschlussraten verfügen (vgl. Tabelle 21, S. 252).

5.1.1 Die K.O.-Kriterien der Dimension der allgemeinen methodischen Qualität mit den höchsten Ausschlussraten

Über die höchste Ausschlusskraft verfügt das K.O.-Kriterium, mit Hilfe dessen bewertet wird, inwieweit die Wirksamkeit der therapeutischen Maßnahmen sowohl über Maße der Veränderung als auch der Zielerreichung abgebildet wurde (Kriterium B.12.; vgl. Kap. 1.2.2). Allein 19 der 26 Studien wurden mindestens aufgrund schlechter Bewertungen auf diesem K.O.-Kriterium ausgeschlossen, in den meisten Fällen wurden ausschließlich statistische Prä-Post- oder Zwischengruppen-Signifikanzberechnungen angestellt und weder Effektstärken noch Zielerreichungsindikatoren (via Cutoff-Wert) berichtet. Es sei noch einmal darauf hingewiesen, dass nach den hier angelegten Kodierregeln zu diesem Kriterium eine Studie mit „2“ abschneiden kann, wenn zumindest statistische Signifikanz- mit Effektstärkeberechnungen kombiniert berichtet werden (vgl. Kap. 3.2.3). Diese Mindestanforderung an die Ergebnisdarstellung entspricht dem seit geraumer Zeit bestehenden wissenschaftlichen Standard (vgl. APA, 2010; Deutsche Gesellschaft für Psychologie [DGPs], 2007), der zudem in dieser Arbeit ausführlich begründet und mit den hier entwickelten Kodierregeln entsprechend berücksichtigt wurde (vgl. Kap. 1.2.2 und 3.2.3). Demnach sind Ergebnisse in Form von Zwischengruppen- oder Prä-Postvergleichen, die ausschließlich inferenzstatistisch abgesichert sind, vulnerabel gegenüber der Stichprobengröße und sollten daher mindestens durch Effektstärkeberechnungen ergänzt werden.

Ein weiterer Grund, der dazu führte, dass Studien aufgrund dieses K.O.-Kriteriums durch die methodische Qualitätsdimension durchfielen, besteht darin, dass im Rahmen von Studien, die als Mehrgruppendesigns angelegt waren, schlussendlich keine direkten Vergleiche (Zwischengruppenvergleiche) angestellt wurden. Stattdessen wurden lediglich Prä-Postvergleiche innerhalb der Treatmentarme angestellt und diese, wenn überhaupt, rein deskriptiv miteinander verglichen. Bei drei dieser Studien war hingegen nicht einmal ein deskriptiver Zwischengruppenvergleich zum Post- bzw. Katamnesemesszeitpunkt möglich, da in diesen Untersuchungen die jeweiligen Kontrollgruppen nicht von hinlänglicher Dauer waren (vgl. Kap. 4.1) – ein Phänomen, auf das weiter unten noch näher eingegangen wird. Da diese Studien grundsätzlich als Mehrgruppendesigns angelegt und geratet wurden, wurde diese Vorgehensweise auch konsequent bei der Bewertung von B.12. eingehalten. Alternativ hätten diese Studien aufgrund der fehlenden bzw. ausschließlich rein deskriptiven Zwischengruppenvergleiche als Studien im Ein-Gruppen-Design kodiert werden können. Stattdessen wurde sich für die vorliegende Arbeit jedoch bewusst dazu entschieden, Studien stets entsprechend derjenigen Designs zu bewerten, in denen sie angelegt wurden – in diesem Fall als Mehrgruppendesigns[70].

Insgesamt wird somit durch dieses Kriterium ein zentrales Gütemerkmal der Ergebnisdarstellung bewertet, auf dessen Forderung sowohl in grundlegenden Statistik- und Methodenlehrbüchern (u.a. Bortz & Schuster, 2010; Eid et al., 2010), in allgemeinen Werken zur Evaluationsforschung (vgl. Hager, Patry & Brezing, 2000) als auch in einschlägigen Werken zur Psychotherapieforschung (u.a. Lambert, 2013b) hingewiesen wird. Einen Grund, warum von diesem Gütemerkmal abgewichen werden sollte, hat keine der kodierten Studien gelie-

[70] Einzig bei einer Studie (38) könnte diese Vorgehensweise zu streng gewesen sein, denn aus den darauf bezogenen Publikationen geht nicht eindeutig hervor, ob die Untersuchung von zwei Psychotherapieverfahren (psychodynamische Therapie und Psychodrama) als *Vergleichs*studie angelegt war (was Zwischengruppenvergleiche notwendig gemacht hätte) oder ob lediglich zwei Therapieformen „in sich“ betrachtet werden sollten.

fert. Auch diejenigen Studien, die letztendlich auf Zwischengruppenvergleiche verzichteten, etwa, weil die Stichprobenzusammensetzungen der unterschiedlichen Behandlungsarme zu verschieden waren (Studie 8) oder die Kontrollbedingung lediglich die erste Therapiephase umfasste (Studien 18, 21 und 41), können die Forderung nach unterschiedlichen Darstellungsweisen der Wirksamkeit (statistische Signifikanz, Effekte und klinische Signifikanz) nicht in Frage stellen. Vielmehr handelt es sich hierbei um Studien, die möglicherweise gleich von vornherein als Ein-Gruppen-Designs hätten bewertet werden müssen und damit als etwas anderes, als es die Publikationen und Designs zunächst nahelegen. Dies würde nach der aktuellen Fassung des Kriterienkatalogs diesen Studien zumindest die Chance einräumen, sich als Ein-Gruppen-Design-Studien auf der methodischen Qualitätsdimension sowie auf der externen Validitätsdimension zu profilieren.

Ein weiterer potentieller Einwand, der in Bezug auf das hier diskutierte K.O.-Kriterium angeführt werden könnte, bezieht sich auf die Herleitung eines Cutoff-Werts zur Feststellung der klinisch relevanten Zielerreichung: Selbst vor dem Hintergrund, dass längst nicht alle in die Studien eingehenden Outcomemaße über Normierungsdaten verfügen, die eine valide Differenzierung zwischen der gesunden und der dysfunktionalen Population erlauben (zwecks Herleitung eines normorientierten Cutoff-Werts), ist davon auszugehen, dass es grundsätzlich möglich ist, zumindest inferenzstatistische Auswertungen *und* Effektstärkeberechnungen anzustellen und zu berichten[71]. Dadurch würde eine Studie, wie oben nochmals betont, immer noch mit zufriedenstellendem Ergebnis (Stufe „2“) auf diesem Kriterium abschneiden können und aus der weiteren Bewertung nicht ausgeschlossen werden.

[71] Auf eine Ausnahme in Form von Ein-Messzeitpunkt-Designs wird in Kapitel 5.1.6 näher eingegangen.

Über die zweistärkste Ausschlusskraft verfügt das sog. „Diagnosestellung-Kriterium“ (K.O.-Kriterium A.2.), mit dem die objektive und reliable Diagnosestellung bewertet wird. 10 der insgesamt 26 Studien, denen eine mangelhafte methodische Qualität attestiert wurde, wurden hier mit „3“ bewertet und damit ausgeschlossen, weil keine objektive Diagnosestellung – sei es auf der Basis strukturierter Interviews oder mit Hilfe von Diagnosechecklisten – feststellbar war. In (fast) keiner dieser Studien wurde der Verzicht auf Diagnosestellungen mittels klassischer Systeme, wie ICD oder DSM, bspw. damit begründet, dass anstelle dieser atheoretischen Klassifikationssysteme auf psychoanalysenähere Instrumentarien (z.B. OPD; Arbeitskreis OPD, 2009) zurückgegriffen wurde. Einzig in dem originalen Abschlussbericht zur Menninger Studie mit dem Titel *Forty-Two Lives in Treatment: A Study of Psychoanalysis and Psychotherapy* (Wallerstein, 1986), der für die Bewertung einer Reanalyse dieser umfassenden Untersuchung herangezogen wurde (Studie 4; Blatt & Shahar, 2004), lässt sich folgende Erklärung finden:

> In that sense, we did not establish diagnoses for our patients; rather, we arrived at diagnostic case formulations. Given the changes that psychoanalytic (or psychodynamic) thinking has undergone over these years, the thinking of PRP [Psychotherapy Research Project] has remained unswervingly psychoanalytic. (Wallerstein, 1986, S. 35)

Und einige Kapitel weiter:

> The other consideration that I want to emphasize is that the nature of the diagnostic thinking and conceptualization described here is what was developed and was congenial within a psychoanalytic theoretical framework as applied to the range of psychiatric pathology. It is not bound by the particular categories of DSM-II . . . . Nor of course, it is bound by the newer categories of DSM-III, with which it is far less congenial. (Wallerstein, 1986, S. 120)

Die Objektivität und Reliabilität der sehr gründlichen Fallformulierungen, die im Rahmen der Menninger Studie vorgenommen wurden, ist ganz sicher schwierig zu bestimmen; worum es hier auch vielmehr ging, ist, dass zumindest in einer Originalstudie – wenn auch nur als Zusatzmaterial für die Kodierung der Reanalyse herangezogen – der explizite Verzicht auf nicht-psychoanalysenahe Diagnosen begründet wurde. In den restlichen Studien, aus deren Publikationen kein diagnostischer Prozess erkennbar war, wurden Diagnosen teilweise retrospektiv vergeben, in den meisten Fällen fehlen in den Publikationen jedoch jedwede Angaben zur Diagnosestellung. Die oftmals gehegte Kritik an den gängigen Klassifikationssystemen (DSM und ICD) von Seiten psychoanalytischer Forscher und Kliniker, aus der heraus letztendlich psychoanalysenähere Systeme wie die OPD entwickelt wurden, ist in den hier kodierten Studien zumindest nicht in der Form wiederzuerkennen, dass auf ICD- oder DSM-Klassifikationen bewusst verzichtet und stattdessen auf psychoanalysenahe Diagnosesysteme zurückgegriffen wird. Andernfalls wären der begründete Verzicht und ein Ausweichen auf andere Diagnosesysteme (z.B. die OPD) zumindest als Hinweise darauf zu werten gewesen, dass das „Diagnosestellung-Kriterium“ (A.2.) in seiner jetzigen Fassung evtl. überdacht und dahingehend modifiziert werden müsste, dass auch Systeme jenseits der ICD- und DSM-Klassifikation zulässig werden. Eine solche Veränderung würde sich jedoch auf sehr viel mehr auswirken, als allein auf die Auslegung dieses Kriteriums (A.2.): Das gesamte Methodenpapier müsste damit von seinem Vorgehen der wissenschaftlichen Anerkennung innerhalb einzelner Anwendungsbereiche, die allesamt den ICD-Klassifikationen entsprechen, abrücken. Auch müsste der Begriff „Krankheitswert“ einer Störung neu überdacht werden, der bislang über ICD oder DSM vorgenommene Diagnosen hinreichend abgesichert ist – darauf wird weiter unten noch im Zusammenhang mit dem sog. „Krankheitswert-Kriterium“ (C.1.) zurückzukommen sein. Dennoch, die hier kodierte Studienlage legt eine solch weitreichende Modifikation nicht nahe.

5.1.2 Die Anwendbarkeit des „Fremdbeurteilung-Kriteriums“ (A.11.) vor und nach der Rekodierung

Auf der allgemeinen methodischen Qualitätsdimension fiel allem voran ein Kriterium auf, das besonders viele Missingwerte produzierte: Mit dem Kriterium A.11. wird kodiert, ob, *wenn* Fremdbeurteilungsverfahren im Rahmen der Outcomemessungen verwendet wurden, diese valide sind und von trainierten und verblindeten Beurteilern angewandt wurden. Dieses Kriterium formuliert eine Bedingung – den Einsatz von Fremdeinschätzungsverfahren – die zunächst erfüllt sein muss, um eine Studie auf diesem Kriterium kodieren zu können. Im ersten Kodierungsdurchgang wurde dieser Bedingung streng gefolgt und all diejenigen Studien, in denen keine Fremdbeurteilungen durch externe Rater vorgenommen wurden, mit einem Missingwert („9“) kodiert – insgesamt traf dies auf 26 der 41 begutachteten Studien zu. Das Kriterium formuliert darüber hinaus eine weitere Bedingung, die erst auf den zweiten Blick auffällt: Durch die Forderung einer verblindeten Fremdbeurteilung, d.h. die Beurteiler sollten blind für die Gruppenzugehörigkeit sein, setzt die Bewertung ein Mehrgruppendesign voraus. Studien im Ein-Gruppen-Design, in denen Fremdbeurteilungsverfahren eingesetzt wurden, wurden im ersten Kodierungsdurchgang daher ebenfalls mit einem Missingwert („4“) belegt – dies traf auf 9 der 41 Studien zu. In einem Rekodierungsdurchgang wurde versucht, diese Missingwerte in plausible, reguläre Ratingwerte umzuwandeln (vgl. Tabelle 22, S. 254). Für diese Umwandlung wurden eigens Rekodierungsregeln erstellt, infolge derer Ein-Gruppen-Designs, in denen Fremdeinschätzungsverfahren angewandt wurden, mit „1“ bewertet wurden, wenn diese valide waren und durch unabhängige, trainierte Rater angewandt wurden. Mit einer „3“ wurden diese Studien bewertet, wenn die externen Rater nicht trainiert oder aber über 25% der eingesetzten Fremdbeurteilungsverfahren nicht valide waren. Indem die Bedingungen der Verblindung für diese Studiendesigns gewissermaßen außer Acht gelassen wurde, konnten die genannten 9 Missingwerte umgewandelt werden, wie beschrieben. Zusätzlich

wurden alle Studien, in denen keine Fremdbeurteilungsverfahren durch externe Rater eingesetzt und die im ersten Durchgang mit einem Missingwert belegt wurden, in der Rekodierung mit „3“ bewertet.

Selbst vor dem Hintergrund der Tatsache, dass diese Rekodierungsmaßnahmen zu keinen grundlegend anderen Ergebnissen auf der methodischen Qualitätsdimension führten, konnte immerhin für 35 Studien die Missingwertanzahl reduziert werden. Da es sich um ein Kriterium der allgemeinen methodischen Qualität handelt, wäre es somit ratsam, diese Kriterien grundsätzlich für alle Studiendesigns (Mehr- und Ein-Gruppen-Designs) anwendbar zu gestalten und dahingehend keine Voraussetzungen zu formulieren.

5.1.3 Die Anwendungsbereiche und die Dimension der allgemeinen methodischen Qualität

Die Verteilungen der Gesamtergebnisse aller 41 Studien auf der methodischen Qualitätsdimension wurden im allgemeinen Ergebnisteil auf den unterschiedlichen Ausprägungen der Merkmale „Anwendungsbereich“ (affektive *versus* gemischte Störungen), „Studiendesign“ und „Therapieumfang“ miteinander verglichen (vgl. Tabelle 29, S. 264; Tabelle 30, S. 266; Tabelle 31, S. 268). Da der Therapieumfang im Zusammenhang mit der allgemeinen methodischen Qualität und der internen Validität den Fokus der Arbeit bildet, wird im Folgenden zunächst nur auf die beiden zuerst genannten Merkmale eingegangen. Der kriterienweise vorgenommene Vergleich der Studien zu unterschiedlichen Therapieumfängen im Hinblick auf die methodische Qualität sowie die interne Validität wird daher in einem extra Abschnitt diskutiert.

Der Vergleich der Studien zu affektiven Störungen (n=10) mit den Studien zu gemischten Störungen (n=31) weist mit 60.0% zu 64.5% qualitativ guter Studien lediglich einen marginalen Unterschied zugunsten der affektiven Störungsstudien auf. Dieser Befund legt nahe, dass

die methodische Qualitätsdimension, zumindest was den hiesigen Studienpool betrifft, nicht vorrangig auf die Bewertung von Untersuchungen an diagnosehomogenen Patientengruppen ausgerichtet ist. Grundsätzlich sind Untersuchungen zu beiden Anwendungsbereichen kodierbar, keine der beiden Studienarten erzeugt mehr Missingwerte als die jeweils andere und keine der beiden Studienarten schneidet im Vergleich zur anderen grundlegend schlechter ab.

5.1.4 Die Studiendesigns und die Dimension der allgemeinen methodischen Qualität

Etwas anders stellt es sich dar, wenn man die Studien in ihren unterschiedlichen Untersuchungsdesigns im Hinblick auf das Gesamtergebnis der methodischen Qualität in Augenschein nimmt, denn dort fallen die Verteilungsunterschiede zwischen den Designgruppen deutlicher aus. Es wird sich in der Diskussion lediglich auf zwei Designgruppen beschränkt, denen vergleichsweise hohe Anteile an Studien gemeinsam sind, die auf der methodischen Qualitätsdimension mit unzureichendem Gesamtergebnis abschneiden: Kontrollgruppendesigns und die verfahrensexternen Vergleiche (vgl. Tabelle 30, S. 266). Der Vergleich mit pharmakologischer Monotherapie (lediglich eine Studie mit unzureichendem Gesamtergebnis) wird hier pragmatisch den verfahrensexternen Vergleichen zugerechnet. Es zeigt sich, dass innerhalb der Gruppe der Kontrollgruppendesigns von insgesamt neun Studien sieben Untersuchungen (knapp 78%) mit unzureichendem Gesamtergebnis abgeschnitten haben; innerhalb der Gruppe der verfahrensexternen Vergleiche (plus Vergleich mit Pharmaka) schnitten alle fünf Studien unzureichend ab. Was hat im Einzelnen dazu geführt, dass innerhalb dieser Designgruppen die Raten an methodisch unzureichenden Studien so hoch ausfallen?

Die sieben Studien im Kontrollgruppendesign und schlechtem Gesamtergebnis fallen, mit einer Ausnahme, aufgrund des oben bereits diskutierten K.O.-Kriteriums B.12. (Bewertung

der Veränderungs- und Zielerreichungsindikatoren) durch. Unter diesen Studien befinden sich die drei schon eingeführten Untersuchungen (18, 21, 41), die noch einmal diskutierenswert scheinen: Bei diesen drei Studien – eine Langzeittherapiestudie (18) und zwei Untersuchungen zu Therapien moderater Länge (21, 41) – wurden zum Postzeitpunkt keine Zwischengruppenvergleiche mehr angestellt, da die Kontrollgruppen (allesamt Wartelisten) nur über die ersten 6-12 Monate mitgeführt wurden. Dieses Phänomen wurde im Ergebnisteil als eines beschrieben, das bei Studien zu längerfristigen Behandlungen auftritt (vgl. Kap. 4.1) und erklärt sich daraus, dass klassische Kontrollgruppen, etwa Wartelisten, schwierig oder gar unmöglich über den gesamten Behandlungszeitraum umzusetzen sind. So sind die in den besagten Studien gewählten Kontrollbedingungen in Form von Wartelisten strenggenommen allenfalls von illustratorischem Wert: Wenn eine Behandlung mehrere Jahre andauert und die Kontrollbedingung davon lediglich die ersten 6-12 Monate umfasst, so lässt sich zwar etwas über die „kontrollierte Wirksamkeit“ der Behandlung innerhalb dieser ersten Phase aussagen. Jedoch liegt das primäre Interesse doch auf der Wirksamkeit der gesamten Behandlung und nicht nur auf derjenigen der ersten Behandlungsphase. Eine Analogie der Zeitmessung bei einem 400-Meter-Lauf von Leichsenring und Rabung (2013) verdeutlicht das Problem: Misst man bei einem Läufer bereits nach 100 Metern seine Zeit, dann würde niemand auf die Idee kommen, von dieser Zeit auf die Gesamtzeit über die 400 Meter zu schließen. Von einer Wirksamkeitsdifferenz zwischen Interventions- und Kontrollgruppe nach der ersten Phase der Therapie auf dieselbe Differenz zum Ende der Behandlung zu schließen, wäre mindestens genauso vermessen. Wie an anderer Stelle bereits eingeräumt, hätten diese Studien für die Kodierung durchaus noch als Ein-Gruppen-Designs kodiert werden können, bei denen Zwischengruppenvergleiche *per definitionem* nicht (mit-) bewertet werden können, jedoch wurde aus ebenfalls genannten Gründen von diesem Vorgehen abgewichen.

Wie ist die methodische Qualitätsdimension – in diesem Fall nochmals das hier verantwortliche K.O.-Kriterium (B.12.) dieser Dimension – nun in Bezug auf die Bewertung von Kontrollgruppendesigns bei Studien zu längerfristigen Behandlungen einzuschätzen? Zwar sind Kontrollgruppendesigns im Allgemeinen eine kaum zu erfüllende Forderung für Langzeittherapiestudien und wenn man an Wartelisten oder Placebobehandlungen denkt, dann gilt dies bereits für Therapien von moderater Länge (vgl. de Maat, Dekker et al., 2007). Jedoch wird dieses Design durch besagtes K.O.-Kriterium auch keineswegs vorausgesetzt oder gefordert. Auch die generellen Forderungen des Methodenpapiers, das Studiendesign betreffend, legen keine unbedingte Fokussierung auf Kontrollgruppendesigns nahe (vgl. Kap. 1.2): Entweder die Therapiebedingung hat sich im Vergleich zu einer Kontrollbedingung als statistisch und klinisch bedeutsam überlegen zu erweisen, *oder* aber, sie hat sich, bei ausreichender Power, einer bereits etablierten Vergleichsbehandlung als äquivalent bzw. nicht unterlegen zu erweisen. Für Untersuchungsgegenstände, bei denen Kontrollbedingungen aus ethischen oder anderen Gründen nicht umgesetzt werden können – etwa bei längeren oder sehr langen Behandlungen oder bei akut suizidalen Patienten – ist es zwecks wissenschaftlicher Anerkennung durch den WBP durchaus möglich, die Wirksamkeit eines Verfahrens durch Vergleich mit einer etablierten Behandlung nachzuweisen. Da diesen Vergleichen die Kontrolle interventionsgebundener oder auch nur interventionsunabhängiger Wirkung[72] fehlt, sollten diese komparativen Untersuchungen zentrale andere Kriterien erfüllen, auf die in der Diskussion der internen Validitätsdimension noch näher eingegangen wird. Für die hier angestrebte Argumentation soll die Feststellung genügen, dass Kontrollgruppen, wie Wartelisten, bei Behandlungen von längerer Dauer kaum umzusetzen sind, so dass die Intention, es wie in den diskutierten Studien dennoch zu tun, lediglich sein kann, eine Art Wirksamkeits*trend* zu il-

[72] Vgl. Kapitel 3.1.1.

lustrieren. Dieser sagt jedoch vergleichsweise wenig über die programm- oder interventionsgebundene Wirkung[73] der *gesamten* Behandlung aus. Für diese Zwecke wären Vergleiche mit bereits etablierten Treatments oder aber – so verfügbar – mit dem natürlichen Störungsverlauf angemessener. Lässt man ökonomische Überlegungen einmal außen vor – komparative Studien bedeuten einen extrem hohen personellen und finanziellen Aufwand – so geben die besagten Studien im „halben Kontrollgruppendesign" keinen Anlass, in der allgemeinen methodischen Qualitätsdimension und allem voran im ausschlaggebenden K.O.-Kriterium (B.12.) eine Benachteiligung von Studien zu längerfristigen Behandlungen zu entdecken. Dies soll nicht bedeuten, dass forschungspragmatische und ökonomische Überlegungen für die Untersuchung des Kriterienkatalogs überhaupt keine Rolle spielen. Begreift man den WBP-Kriterienkatalog jedoch in erster Linie als wissenschaftliches Instrument und erst in zweiter Linie als Teil eines berufs- und gesundheitspolitischen Regulativs, dann sollten ökonomische Überlegungen für die hiesigen Zwecke zumindest nicht leitend sein.

Die zweite Kategorie an Studiendesigns, auf die im Hinblick auf ihr Abschneiden auf der allgemeinen methodischen Qualitätsdimension näher eingegangen werden soll, betrifft die verfahrensexternen Vergleiche (inklusive dem Vergleich mit Pharmakotherapie). Hiervon schneiden alle fünf Studien mit schlechter methodischer Qualität ab (vgl. Tabelle 30, S. 266). Dies liegt bei drei der Studien (8, 32, 38) an dem bereits diskutierten K.O.-Kriterium B.12., das unterschiedliche Indikatoren der Ergebnisdarstellung fordert. Bei den anderen beiden Studien handelt es sich um eine Originalstudie (36) und deren Replikation (37), bei denen sowohl das „Diagnosestellung-Kriterium" (A.2.) als auch das „Krankheitswert-Kriterium" (C.1.), jeweils zum Ausschluss der beiden Studien führten. Diese negativen Ratings resultieren dar-

[73] Vgl. Kapitel 3.1.1.

aus, dass in beiden Studien weder etwas über den diagnostischen Prozess noch über Diagnosegruppen und damit über den Krankheitswert der Störungen berichtet wird. In allen Publikationen, die für die Kodierung dieser Studien herangezogen wurden, werden die Patienten lediglich hinsichtlich ihrer Problematiken beschrieben (*anxiety, depression, interpersonal problems, self-esteem, living/welfare* etc.), ob eine systematische Störungsdiagnostik stattgefunden hat, bleibt jedoch ungewiss. Nun wurde im ersten Teil der Arbeit (Kap. 1.2.1) auf eine Kritik an der RCT-Methodologie hingewiesen, die sich auf die Rekrutierung ausschließlich solcher Studienpatienten in RCTs bezog, deren krankheitswertige Störung über Klassifikationssysteme, wie DSM[74] und ICD, festgestellt wurde. Kontrastierend dazu wurde auf Befunde hingewiesen, die nahelegen, dass zwischen ⅓ und der Hälfte der Patienten, die nach psychotherapeutischer Hilfe suchen, über keine eindeutige ICD- oder DSM-Diagnose verfügen, bei denen eine Behandlung jedoch trotzdem erforderlich scheint und auch erfolgt (vgl. Westen et al., 2004, S. 634): Diese Patienten leiden unter unterschiedlichen subklinischen Symptomausprägungen, die sich in keiner diagnostischen Kategorie des DSM oder der ICD eindeutig abbilden lassen. Das „Krankheitswert-Kriterium“ (C.1.) kommt diesem Phänomen bereits entgegen, indem für positive Bewertungen (Stufe „1“ oder „2“) entweder Patienten mit „wahrscheinlicher klinischer Störung“ gefordert werden (Traumapatienten o.ä.) oder aber, indem eine Mindestrate von 80% der Studienpatienten gefordert wird, die über Störungen mit Krankheitswert verfügen. Nun ist es durchaus möglich, dass die 80%-Rate zu hoch gegriffen ist, wenn man sie mit den Befunden, die Westen et al. (2004) berichten, vergleicht. Jedoch ist das Problem bei den beiden besagten Studien auch ganz anders gelagert: In den Studien wird nicht berichtet, dass ein diagnostischer Prozess stattgefunden hat, infolgedessen sich heraus-

[74] Davon ist jedoch das DSM-5 (American Psychiatric Association, DSM-5 Task Force, 2013) ausgenommen, zu dem Kritik laut wird, die darin vorgenommene Schwellenabsenkung würde zu einer Inflationierung von psychischen Störungen führen (vgl. Stieglitz & Hiller, 2013).

stellte, dass weit über 20% der Patienten lediglich unter einer subklinischen Störung litten. Vielmehr ist beiden Studien zu diesen Aspekten überhaupt keine Information zu entnehmen, so dass die Negativbewertung auf dem „Krankheitswert-Kriterium“ (C.1.) durchaus gerechtfertigt scheint.

Es muss in diesem Zusammenhang noch ein weiteres Problem thematisiert werden, das auf eine Art Zwangslage verweist, in der der WBP sich gewissermaßen befindet: Es ist zweifellos denkbar, dass, wie es Westen et al. (2004) postulieren, in der klinischen Praxis der Anteil an Psychotherapiepatienten mit subklinischen Symptomausprägungen nach ICD (oder DSM) weit höher liegt als bei 20%. Diese Vermutung etwa auf der Basis von Krankenkassendaten nachzuweisen, stellt ein nahezu unmögliches Unterfangen dar, denn Behandler müssen in ihren Kassenanträgen stets eine psychische Störung von Krankheitswert angeben (u.a. Auckenthaler, 2012). Diese Forderung, die mit dem sog. Bericht an den Gutachter verbunden ist, leitet sich wiederum aus der Psychotherapierichtlinie (Gemeinsamer Bundesausschuss, 2013) sowie dem Psychotherapeutengesetz ab, in denen es heißt, dass die Ausübung von Psychotherapie sich ausschließlich auf *krankheitswertige* Störungen beziehen darf. Auch die Akkreditierung von Ausbildungsstätten im Rahmen der berufsrechtlichen Anerkennung erfolgt in Abhängigkeit davon, ob an ihnen Patienten mit *krankheitswertigen* Störungen mittels wissenschaftlich anerkannter Verfahren behandelt werden (vgl. Kap. 1.1). Und auch der WBP unterliegt diesen Regularien, wodurch die mit dem „Krankheitswert-Kriterium“ (C.1.) verbundenen Forderungen verständlicher werden (zur Erinnerung: das Kriterium C.1. zählt als K.O.-Kriterium zur allgemeinen methodischen Qualitätsdimension und geht aber ansonsten in die Verrechnung der externen Validität ein): Er muss zwecks Studienbegutachtung einen Cutoff-Wert für den Anteil an Studienpatienten mit krankheitswertigen Störungen festlegen (hier: 20%), der für eine zufriedenstellende Bewertung einer Studie auf dem „Krankheitswert-Kriterium“ nicht überschritten werden darf. Nach Westen et al. (2004) wäre dieser Cutoff-

Wert nicht zu rechtfertigen, da er für Wirksamkeitsstudien etwas fordert, das sich in der Praxisrealität jedoch offenkundig ganz anders darstellt – unter der Voraussetzung, die von Westen et al. berichteten Befunde treffen auf den deutschen Versorgungskontext zu. Dem WBP könnte man folglich vorwerfen, dass er mit Kriterium C.1. einen Aspekt der RCT-Methodologie, wenn auch in abgemilderter Form, tradiert, indem er ebenfalls Studien fordert, in denen das Erfüllen von DSM- oder ICD-Diagnosen als Einschlusskriterium gilt (vgl. Kap. 1.2.1). Jedoch zielt dieser Vorwurf eigentlich auf etwas sehr viel Grundlegenderes ab als auf das Vorgehen des WBP. Die Kritik berührt vielmehr die Kluft, die zwischen der realen Versorgung von Patienten (auch subklinischer Störungsbilder) und dem gesetzlich vorgeschriebenen Behandlungsauftrag samt der damit in Verbindung stehenden Definition von „Krankheitswert“ besteht. Die Frage nach dem Zustandekommen dieser Kluft ist von grundsätzlicher Natur. Sie im Kontext der Evidenzbasierung und wissenschaftlichen Anerkennung von Psychotherapie zu klären, würde ihren grundsätzlichen Charakter verfehlen. Insofern würde auch ein an den WBP gerichteter Vorwurf im Hinblick auf seine Orientierung an ICD- oder DSM-diagnostizierten, *krankheitswertigen* Störungen und seine dementsprechende Forderung an Studien eher das Symptom eines Problems berühren als dessen Ursache.

Wie ist nun die methodische Qualitätsdimension in Bezug auf verfahrensexterne Vergleiche sowie den Vergleich mit pharmakologischer Monotherapie einzuschätzen? Es konnte bereits in den vorhergehenden Abschnitten gezeigt werden, dass das K.O-Kriterium zur Bewertung der Veränderungs- und Zielerreichungsindikatoren (B.12.) grundsätzlich für Mehrgruppendesigns anwendbar ist, *wenn* innerhalb der Studien geeignete Kontroll- bzw. Vergleichsbedingungen herangezogen werden. Das zweite hier maßgebliche K.O.-Kriterium – das sog. „Krankheitswert-Kriterium“ (C.1.) – wird von Seiten der beiden verfahrensexternen Vergleichsstudien (36, 37), die auf C.1. unzureichend abschneiden, keineswegs in Frage gestellt: Beiden Studien ist zu diesem Punkt keine verlässliche Information zu entnehmen. Auch

wurde eine möglicherweise naheliegende Kritik an der genauen Operationalisierung der einzelnen Kriterienstufen des „Krankheitswert-Kriteriums“ entkräftet, indem aufgezeigt wurde, dass diese Kritik sich vielmehr auf eine möglicherweise bestehende Kluft zwischen Versorgungsauftrag und Versorgungsbedarf bezieht und erst vermittelt darüber auf eine vielleicht unrealistische Forderung des WBP.

5.1.5 Über die Schwierigkeit, ein valides Außenkriterium zu finden: Ein Versuch der Bestimmung der Klassifikationsgüte der methodischen Qualitätsdimension

Es soll abschließend zum allgemeinen Diskussionsteil zur methodischen Qualitätsdimension ein vorsichtiger Versuch unternommen werden, die Klassifikationsfähigkeit dieser Dimension einzuschätzen. Dafür wurde folgendes Außenkriterium gewählt: Der Studienpool an 41 Studien wurde in zwei Gruppen geteilt, wovon die eine Gruppen alle Studien enthält, die im Rahmen der Studienrecherche durch die Handsuche gefunden wurden (n=22); die andere Gruppe enthält hingegen alle Studien, die durch die digitale Recherche gefunden wurden (n=19). Die erste Gruppe zeichnet sich dadurch aus, dass diese Studien bereits in Metaanalysen oder narrative Reviews eingingen und damit einer Art Qualitätscheck unterzogen wurden. Die zweite Gruppe enthält hingegen Studien, die in die damals gesichteten Reviews nicht eingegangen waren und allein durch die digitale Recherche gefunden wurden. Von letzterer Gruppe soll daher angenommen werden, dass diese Studien über keine ausreichende Qualität verfügten, um in systematische oder narrative Reviews zur Wirksamkeit psychodynamischer Psychotherapie aufgenommen zu werden. Es zeigt sich, dass bei über 68% der 41 Studien die Klassifikation der Qualitätsdimension des WBP-Kriterienkatalogs mit dem Außenkriterium "Review-Studie ja/nein" übereinstimmt, bei knapp 32% hingegen kommt die methodische Qualitätsdimension im Vergleich zum hier gewählten Außenkriterium zu gegenläufigen Einschätzungen. Es zeigt sich zudem, dass mittels der methodischen Qualitätskriterien 55% der sog. "Review-Studien" als qualitativ gut eingeschätzt werden, 45% der "Review-Studien"

werden hingegen als methodisch unzureichend eingestuft. Von denjenigen Studien, die in keines der gesichteten Reviews eingingen, werden knapp 81% methodisch unzureichend und lediglich 19% als methodisch gut eingestuft. In die Sprache der Klassifikationsgütekriterien übersetzt bedeutet das, dass die allgemeine methodische Qualitätsdimension über eine hohe Spezifität (81%), jedoch über eine vergleichsweise niedrige Sensitivität (55%) verfügt.

Dieses Ergebnis ist allerdings mit größter Vorsicht zu akzeptieren, denn die Validität des hier gewählten Außenkriteriums ist durchaus in Frage zu stellen. So wurde zwar im Rahmen der Studienrecherche eine große Bandbreite an Metaanalysen und narrativen Reviews gesichtet, jedoch ist es durchaus möglich, dass einige der hier kodierten Studien in nicht gesichtete Reviews eingingen. Dies könnte vor allem auf Studien neueren Datums zutreffen. Unter diesen Umständen hätten hier evtl. mehr Studien als "Review-Studien" bezeichnet werden müssen. Das gewichtigere Argument, das hier gewählte Außenkriterium in Frage zu stellen, besteht jedoch in folgender Tatsache: Die hier gesichteten Reviews unterscheiden sich in hohem Ausmaß im Hinblick auf die gewählten Ein- und Ausschlusskriterien und damit im Hinblick auf die Hürde der methodischen Qualitätsbemessung, die die Studien passieren mussten, um eingeschlossen zu werden. So ist es durchaus möglich, dass einige Studien in manche narrative Reviews und Überblicksartikel eher aus illustratorischen Zwecken aufgenommen wurden – etwa, um einen ganz bestimmten Aspekt einer Studie besonders hervorzuheben – und weniger, weil durch ihre besondere methodische Qualität eine Aufnahme gerechtfertigt gewesen wäre. Gerade weil das hiesige Außenkriterium auch narrative Reviews und Überblicksartikel einschloss, muss davon ausgegangen werden, dass sich möglicherweise zu viele Studien mit fraglicher methodischer Qualität in der Gruppe der "Review-Studien" befinden. So könnte sich zumindest die vergleichsweise niedrige Sensitivität der methodischen Qualitätsdimension erklären, die – sollte der zuletzt geschilderte Verdacht zutreffen – eher nach oben korrigiert werden müsste.

Insgesamt scheint die Dimension der allgemeinen methodischen Qualität einen höheren Maßstab an die Studien anzulegen, als zumindest einige der gesichteten Reviews, wobei, wie dargelegt, die Validität des Außenkriteriums fraglich ist. Ein valideres Außenkriterium wäre vielleicht dadurch zustande gekommen, indem hier allein systematische Reviews (Metaanalysen) herangezogen und narrative Reviews von vorherein als Außenkriterium ausgeschlossen worden wären. Dies hätte das Verhältnis von Sensitivität und Spezifität evtl. in eine solche Richtung verschoben, die den Bewertungsmaßstab der methodischen Qualitätsdimension weniger hoch erscheinen lässt. Vor allem jedoch vor dem Hintergrund der vorangehenden Diskussion in Bezug auf die zwei ausschlussstärksten K.O.-Kriterien dieser Dimension sowie in Bezug auf unterschiedliche Gegenstandsbereiche („Anwendungsbereiche“ und „Studiendesigns“), ist diese Dimension in Bezug auf den hiesigen Studienpool nicht als zu streng bzw. unrealistisch im Sinne Schultes (Deutsches Ärzteblatt, 2008a) zu bezeichnen.

5.1.6 Die Gegenstandsadäquatheit der Dimension der allgemeinen methodischen Qualität in Bezug auf Langzeittherapiestudien im Vergleich zu Studien zu kürzeren Therapien

Die Inkompatibilität zwischen der RCT-Methodologie und dem Gegenstand "Langzeittherapie" (> 100 Sitzungen) bezieht sich, wie in Kapitel 1.2.1 dargestellt, hauptsächlich auf die in RCTs realisierten Kontrolltechniken, die der internen Validitätssicherung dienen. Die Dimension der methodischen Qualität wurde im Hinblick auf die Gegenstandsadäquatheit nun in erster Linie als *Bestandteil* der internen Validitätsdimension betrachtet und mit untersucht, da ein hinreichend gutes Abschneiden auf dieser Dimension als obligatorisch für die Validitätsbewertung gilt. Es wurde sich hierbei ausschließlich auf die sechs K.O.-Kriterien der methodischen Qualitätsdimension konzentriert. Für die Untersuchung wurden die Kriterienstufen dichotomisiert, gleiches trifft auf die Dreiteilung Langzeittherapie (über 100 Sitzungen), The-

rapie moderater Länge (über 25 bis 100 Sitzungen) und Kurzzeittherapie (bis 25 Sitzungen) zu (vgl. Kap. 3.4.3).

Im Rahmen der Analyse wurden die Studiengruppen (Langzeittherapie *versus* kürzere Therapiedauern) im Hinblick auf ihr jeweiliges Abschneiden auf den einzelnen K.O.-Kriterien miteinander verglichen. Dies erfolgte über Verteilungsvergleiche bzw. Prozentwertvergleiche auf den dichotomisierten Stufen der Kriterien. Es wurde zudem eine Feinanalyse angeschlossen, wenn sich auf der Basis der quantitativen Verteilungsvergleiche der Verdacht einer benachteiligenden Wirkung des jeweiligen Kriteriums im Hinblick auf Langzeittherapiestudien nahelegte. Solche Verteilungsunterschiede durften lediglich als Hinweise auf *potentielle* Benachteiligungen gewertet werden, da sich aufgrund dieser Unterschiede noch keine Aussage darüber treffen ließ, ob die Gründe für einen Verteilungsunterschied tatsächlich mit dem Gegenstand "Therapieumfang" zusammenhängen. Eben diesem Zusammenhang auf die Spur zu kommen war Sinn und Zweck der Feinanalyse, in der die Studien nochmals mit dem Augenmerk auf eben diesen möglichen Zusammenhang zwischen Therapiedauer (Sitzungsanzahl) und Ergebnis auf dem jeweiligen Kriterium genau gesichtet wurden. Als Cutoff-Wert wurde ein Effekt von $\omega \geq 0.10$ gewählt, d.h. bei denjenigen Kriterien, bei denen sich Verteilungsunterschiede zu*un*gunsten der Langzeittherapiestudien in Effektgrößen, wie angezeigt, niederschlugen, erfolgte eine Feinanalyse.

Ferner wurden Extremgruppenvergleiche zwischen Langzeittherapiestudien (über 100 Sitzungen) und Kurzzeittherapiestudien (bis 25 Sitzungen) unter Ausschluss der Untersuchungen zu Behandlungen moderater Länge (über 25 bis 100 Sitzungen) angestellt. Dieser Extremgruppenvergleich liegt darin begründet, dass von zahlreichen Autoren die Inkompatibilität zwischen dem Gegenstand "Langzeittherapie" und der RCT-Methodologie stets in Abgrenzung zu „echten“ Kurzzeittherapien („6 to 16 sessions“ [Westen et al., 2004, S. 632]) und deren Kompatibilität mit der RCT-Methodologie gesehen wird (u.a. Leichsenring, 2011).

Bei zwei der sechs Kriterien war es nicht möglich, Verteilungsunterschiede zwischen den Studientypen (Langzeittherapie *versus* kürzere Therapiedauern) überhaupt festzustellen, da bei beiden Kriterien nur jeweils eine Ratingstufe besetzt war. Dabei handelt es sich zum einen um das sog. „Manipulation-Kriterium“ (A.1.), das bemisst, ob anzunehmen ist, dass die in einer Studie dargestellten Ergebnisse auf irgendeine Weise manipuliert wurden. Diese Vermutung wäre bspw. dann gerechtfertigt, wenn in mehreren Publikationen zur selben Untersuchung voneinander abweichende Stichprobenumfänge zugrunde gelegt werden, ohne dass sichtbar wird, worin die Unterschiede zwischen den unterschiedlichen Stichproben eigentlich bestehen. Auch widersprüchliche Ergebnisse über mehrere Publikationen hinweg würden einen Manipulationsverdacht nahelegen. Allerdings: Bevor man hier tatsächlich eine Negativbewertung abgibt, wird man nicht umhinkommen, die Autoren der Studie zu kontaktieren, um ungerechtfertigte Negativbewertungen möglichst zu vermeiden (vgl. Kap. 1.2.2). Im Rahmen dieser Arbeit wurden während der Studienrecherche Autorenteams dann kontaktiert, wenn aus einzelnen Publikationen nicht sicher hervorging, inwieweit diese sich auf denselben Datensatz bzw. auf sich überschneidende Datensätze beziehen. Diese Abklärung hat ausgereicht, um differierende Stichprobengrößen oder auch Outcomemaße über mehrere Veröffentlichungen hinweg korrekt einordnen zu können. Auf diesem Kriterium, das gleichzeitig das einzige Kriterium darstellt, das regulär auf nur zwei Ausprägungsstufen misst („1“ und „3“), haben alle 41 Studien mit „1“ abgeschnitten. Ein ähnliches Bewertungsmuster ist zum anderen beim dichotomisierten K.O.-Kriterium C.9. zu finden: Mittels dieses Kriteriums wird bewertet, ob es sich bei den verwendeten Outcomemaßen um patientenrelevante Parameter handelt. Hier sind alle Studien mit optimalem („1“) oder zufriedenstellendem („2“) Rating bewertet worden. Obligatorisch war für eine „1“er-Bewertung, dass unter den Outcomemaßen auch solche zur Lebensqualität oder zum Inanspruchnahmeverhalten erhoben wurden – dies traf immerhin auf

15 der 41 Studien (36.6%) zu. Die Betonung dieser beiden Ergebnismaße (Lebensqualität und Inanspruchnahmeverhalten) ist auch insoweit begrüßenswert, als dass es sich dabei fraglos um therapieschulunabhängige Zielkriterien handelt, die sich daher vor allem auch für komparative Studien eignen, in denen konzeptionell sehr verschiedene Verfahren miteinander verglichen werden – etwa analytische und Verhaltenstherapie (vgl. Hager, 2000).

Bei zwei weiteren K.O.-Kriterien fiel der Vergleich zwar zuungunsten der Langzeittherapiestudien aus, jedoch zeigten sich derart marginale Unterschiede in den Prozentwertverteilungen zwischen den Studiengruppen (Langzeit- *versus* kürzere Therapien), dass keine Feinanalysen angeschlossen wurden. Hierbei handelt es sich zum einen um das Kriterium A.8. zur Reliabilitäts- und Validitätsbemessung der in den Studien verwendeten Outcomemaße (dazu ausführlich in Kap. 3.2.3). Zum anderen handelt es sich um das bereits diskutierte K.O.-Kriterium C.1. – das sog. „Krankheitswert-Kriterium“. Bei beiden Kriterien zeigten sich Effekte unter dem gesetzten Cutoff von ω 0.10 . Dieser Cutoff-Wert wird auch nach Ausschluss der Untersuchungen zu moderaten Therapielängen (Extremgruppenvergleich) nicht überschritten. Zwar kommt es im Hinblick auf das „Krankheitswert-Kriterium“ (C.1.) zu einer „Richtungsumkehr“ der Verteilungen nach Exklusion der Studien zu mittellangen Therapien und es schneiden nunmehr mehr Langzeittherapie- als Kurzzeittherapiestudien positiv auf diesem Kriterium ab, jedoch bewegen sich diese Effekte mit ω 0.07 bzw. ω 0.03 in einem so randständigen Bereich, dass der Richtungsumkehr nicht allzu viel Bedeutung beigemessen werden sollte.

Anders stellt es sich bei den beiden folgenden K.O.-Kriterien dar: Bei beiden Kriterien handelt es sich um dieselben, die bereits im Zusammenhang mit den ausschlussstärksten Kriterien diskutiert wurden (s.o.). Sowohl das Kriterium (B.12.) zur Bemessung der Ergebnisdarstellung (Veränderungs- und Zielerreichungsmaße via statistischer Signifikanz, Effektstärken und

klinischer Signifikanzmaße) mit den höchsten Ausschlussraten als auch das zweitstärkste Ausschlusskriterium (A.2.), das sog. „Diagnosestellung-Kriterium“, erwiesen sich bei der rein verteilungsanalytischen Betrachtung als potentiell benachteiligend, was Langzeittherapiestudien betrifft. Jedoch befinden sich die Effekte mit einem Range von ω 0.10 bis ω 0.13 (beide Kriterien) immer noch innerhalb eines Bereichs, in dem man von kleinen Effekten spricht (vgl. Eid et al., 2010). Dennoch wurde der sehr streng festgelegte Cutoff-Wert von ω 0.10 erreicht bzw. überschritten und die Studienzusammensetzungen in den Zellen der Vierfeldertafeln feinanalytisch untersucht (vgl. Tabelle 33, S. 272 und Tabelle 35, S. 277).

Im Hinblick auf das „Diagnosestellung-Kriterium“ (A.2.) war zunächst festzustellen, dass alle Langzeittherapiestudien mit negativem Rating auf diesem Kriterium in einem retrospektiven Design durchgeführt wurden, was auf einen Großteil der Studien zu kürzeren Behandlungen und negativem Ergebnis ebenfalls zutraf. Wenn zur Baseline kein nachvollziehbarer diagnostischer Prozess stattgefunden hat, dann erweist es sich als schwierig bis unmöglich, diese Diagnosen retrospektiv auch nur annähernd reliabel „nachzureichen“. Dies erklärt, warum die meisten retrospektiven Studien – ungeachtet der untersuchten Therapielänge – durch dieses Kriterium ausscheiden. Jedoch lassen sich bei beiden Studiengruppen, wenn auch nur sehr wenige, retrospektive Studien finden, die auf diesem Kriterium positiv abschneiden.

In einem weiteren Schritt wurden die Studien der beiden Studiengruppen (Langzeittherapie *versus* kürzere Therapien) innerhalb der Negativbewertungskategorie nochmals gesichtet und miteinander verglichen; zudem wurden die retrospektiven Langzeittherapiestudien mit positivem Rating herangezogen und ebenfalls mit den negativ bewerteten Langzeittherapiestudien verglichen. Anhand dieser Vergleiche wurde nach expliziten oder impliziten Hinweisen gesucht, die hätten begründen können, warum besagte Langzeittherapien ausschließlich retrospektiv beforscht werden konnten. Die Frage, der in dieser Feinanalyse nachgegangen

wurde, lautete also: Lassen sich bestimmte Charakteristika feststellen, die den negativ bewerteten Langzeittherapiestudien gemeinsam sind und die einen retrospektiven Datenzugang notwendig machten?

Da die Feinanalyse keine derartigen Erkenntnisse lieferte, kann im Hinblick auf den hiesigen Datensatz geschlossen werden, dass es durch das sog. „Diagnosestellung-Kriterium“ (A.2.) zu keiner systematischen Benachteiligung der Langzeittherapiestudien im Vergleich zu Mittellang/Kurzzeittherapiestudien kam. Die Langzeittherapiestudien schnitten – genauso, wie die Untersuchungen zu kürzeren Therapiedauern – schlecht ab, weil kein diagnostischer Prozess nachvollzogen werden konnte. Dieses Defizit kann primär auf das retrospektive Design dieser Studien zurückgeführt werden. Jedoch existiert kein zwingender Zusammenhang zwischen dem Gegenstand "Langzeittherapie" und dem retrospektiven Datenzugang.

Das K.O.-Kriterium zur Veränderung und Zielerreichung (B.12.) wurde in den vorangehenden Abschnitten bereits ausführlicher diskutiert, u.a. in Bezug darauf, dass es vom evaluatorischen Standpunkt aus betrachtet wünschenswert und grundsätzlich auch möglich sein sollte, inferenzstatistisch abgesicherte Prä-Post/Katamnese-Vergleiche oder Zwischengruppenvergleiche mindestens durch Effektstärken zu bereichern; gleichsam wurde im Hinblick auf Studien in Mehrgruppendesigns herausarbeitet, dass in Abhängigkeit vom Untersuchungsgegenstand (kurze oder lange Behandlungsdauern) angemessene Kontroll- bzw. Vergleichsgruppen zu wählen sind, so dass ein Zwischengruppenvergleich tatsächlich zu Behandlungsende stattfinden kann. In diesem Zusammenhang zeigte sich, dass Wartelistengruppen in Untersuchungen zu moderaten (> 25 Sitzungen) bis umfangreichen Behandlungslängen (> 100 Sitzungen) als Kontrollbedingungen ungeeignet sind. Hinsichtlich der Gründe für etwaige Negativbewertungen auf dem Kriterium B.12. waren zwischen den beiden Studiengruppen (Langzeit- *versus* kürzere Therapielängen) keine grundlegenden Unterschiede zu entdecken, bei beiden Stu-

diengruppen wurden meist über statistische Signifikanzberechnungen hinaus keine weiteren Parameter (Effektstärken, RCI etc.) berechnet. Jedoch fiel im Rahmen der Feinanalyse ein weiterer Studientyp auf Seiten der Langzeittherapiestudien auf, der bislang in der Diskussion noch unberücksichtigt blieb: Hierbei handelt es sich um Untersuchungsdesigns mit lediglich einem Messzeitpunkt. Im hier untersuchten Datensatz waren dies eine Studie (35) mit nur einer Postmessung und zwei reine Katamnesestudien (17, 23) mit beeindruckenden Katamnesezeiträumen von bis zu 6.5 Jahren. Warum es schwierig ist, in Studien mit nur einem Messzeitpunkt tatsächlich therapiebedingte Veränderungen oder Zielerreichungen, wie es das Kriterium B.12. fordert, abzubilden, wurde bereits in Kapitel 4.4 erläutert. Shadish et al. (2002) bringen es folgendermaßen auf den Punkt:

> The absence of a pretest makes it difficult to know if a change has occurred, and the absence of a no-treatment control group makes it difficult to know what would have happened without treatment . . . . and so this design is rarely useful in this simple form. (S. 106f.)

Bei so grundlegenden Einwänden, wie sie durch das Zitat zum Ausdruck gebracht werden, muss die Frage erlaubt sein, welchem Zweck diese Studiendesigns genau dienen sollen und darüber hinaus, ob die Wahl dieses Studiendesigns mit dem Gegenstand "Langzeittherapie" zusammenhängt. Den Publikationen über die Katamnesestudie (23) von Leuzinger-Bohleber und Kollegen (2001, 2003) sind Begründungen für das gewählte Messdesign zu entnehmen, die sich folgendermaßen zusammenfassen lassen: Mit dem alleinigen Messzeitpunkt mindestens 4 Jahre und durchschnittlich 6.5 Jahre nach Beendigung der Therapie sollte sichergestellt werden, dass der analytische Behandlungsprozess durch Forschungsaktivitäten nicht gestört wurde. Zudem konnte so eine verhältnismäßig große Studienstichprobe zusammengestellt werden, die aufgrund der Querschnitterhebung frei und unabhängig von Dropoutquoten während und nach der Behandlung war. Gleichzeitig wurde damit die kostenintensive Dauer einer

prospektiven Längsschnittuntersuchung umgangen. Vor allem die zuletzt genannten Gründe sind auch in der zweiten Katamnesestudie (17) von Keller und Kollegen (2001) zu finden. Sie beziehen sich primär auf ökonomische Argumente und zweifellos sind prospektive Untersuchungen inklusive katamnestischer Erhebungen zu (besonders) langen Behandlungen äußerst kosten- und zeitintensiv. So investierte bspw. die Forschergruppe um Huber und Klug in die prospektive, komparative Münchener Psychotherapiestudie (MPS) über 12 Jahre Zeit (Prä-Post-Katamnese-Design, Katamnesezeitraum: 3 Jahre) (vgl. Huber, Henrich, Clarkin & Klug, 2013; Huber, Zimmermann et al., 2012). In dieser Outcomestudie an depressiven Patienten wurden psychodynamische Verfahren (u.a. Langzeittherapien > 100 Stunden) mit kognitiver Verhaltenstherapie verglichen. In der ebenfalls prospektiv angelegten multizentrischen Untersuchung zu Langzeittherapien (psychodynamische Verfahren und kognitive Verhaltenstherapie) bei chronischen Depressionen (LAC-Studie) von Leuzinger-Bohleber und Kollegen wurde die Rekrutierung der Patienten im Jahr 2007 begonnen und bis 2013 fortgesetzt. Der Gesamterhebungszeitraum soll sich pro Patient auf 5 Jahre belaufen, d.h. auch hier wird die Gesamtprojektlaufzeit 10 Jahre bei Weitem überdauern (vgl. Beutel et al., 2012). Gleiches trifft auf eine aktuell laufende multizentrische Studie von Benecke und Kollegen zu, in der analytische Psychotherapie mit kognitiver Verhaltenstherapie bei Angststörungen mit komorbiden Persönlichkeitsstörungen verglichen wird (APS; vgl. Benecke, Huber, Schauenburg & Staats, 2012). Diese Studie ist mit einer Gesamtlaufzeit von mindestens 10 Jahren geplant. Auch hier sollen Katamnesen bis zu 3 Jahren nach Behandlungsende stattfinden. So sehr die drei genannten sehr aufwändigen Untersuchungen die Kosten- und Zeitintensität illustrieren, so sehr zeigen sie jedoch auch, dass sie – unter der Bedingung von hinreichend finanzieller Zuwendung – prinzipiell durchführbar sind. Es wurde zudem an anderer Stelle bereits eingeräumt, warum im Zusammenhang mit der Fragestellung der vorliegenden Arbeit ökonomische Begründungen für oder gegen bestimmte Studiendesigns zumindest nicht leitend sein sollen.

Daher soll ein anderer Punkt, den Leuzinger-Bohleber und Kollegen in einer ihrer Publikationen zur o.g. Katamnesestudie herausheben, näher beleuchtet werden:

> Katamnestische Untersuchungen haben unbestritten – verglichen mit prospektiven Untersuchungen – ihre Nachteile. Sie verfügen allerdings über einen großen Vorteil: Sie üben in keiner Weise einen Einfluß auf die Intimität einer therapeutischen Situation aus und beeinflussen den psychotherapeutischen Prozeß durch ihre Forschung nicht, im Gegensatz zu prospektiven Untersuchungen, sie sich immer auf das laufende Therapiegeschehen auswirken. (Leuzinger-Bohleber, Stuhr, Rüger & Beutel, 2001, S. 198)

Auch, wenn sich der im Zitat benannte Vorbehalt gegen die prospektive Beforschung von Psychotherapie nicht explizit auf die psychoanalytisch begründeten Verfahren und insbesondere auf die analytische Langzeittherapie beschränkt, so legen unterschiedliche Arbeiten jedoch einen solchen Vorbehalt vor allem von Seiten der psychoanalytischen „community“ in Bezug auf psychodynamische Behandlungen nahe (vgl. Benecke, 2014b; Protz, Kächele & Taubner, 2012). So sind der Interviewstudie zum Thema *Ambivalenz mit der Therapieforschung* von Protz et al. (2012), neben durchaus forschungszugeneigten Aussagen, auch Aussagen von praktizierenden Psychoanalytikern zu entnehmen, die die Autoren zu folgenden übergeordneten Statements zusammenfassen:

- Ich möchte die Patienten und den analytischen Raum schützen . . .
- Standardisierte Messinstrumente greifen in die psychoanalytische Behandlung ein . . .
- Ich finde Audioaufzeichnungen im analytischen Prozess zu intrusiv . . .
- Der analytische Raum ist einzigartig und durch Forschung wird etwas geopfert . . . (S. 81f.)

Benecke (2014b) ist zum selben Thema ein beeindruckend klarsichtiger Artikel gelungen, in dem er aufzeigt, welche Missverständnisse der oftmals geäußerten Kritik an empirischer

(quantitativer) Forschung seitens der psychoanalytischen „community“ zugrunde liegen. Seine Argumentation soll an dieser Stelle nicht weiter ausgeführt werden, stattdessen soll sein resümierender Standpunkt zu o.g. Forschungsvorbehalten aufgegriffen werden:

> Der Verweis auf den dann möglicherweise nicht mehr so geschützten Rahmen oder auf die Forscher als „Dritte im Bunde“ ist natürlich wichtig und bedarf der kritischen Diskussion. Grundsätzlich ist das auf diese Weise Beforschtwerden allerdings meist weniger für die Patienten ein Problem als vielmehr für die Therapeuten. (Das lehrt mich zumindest meine mittlerweile langjährige Erfahrung in diesem Feld.) Und die patientenseitigen Fantasien über „die Forscher“ können meiner Erfahrung nach, analog beispielsweise der Bearbeitung von Fantasien bezüglich der Berichte an die Gutachter bei Verlängerungsanträgen, gut und auch gewinnbringend analytisch bearbeitet werden. (S. 65f.)

Inwieweit das Beforschtwerden vor allem für die psychodynamischen Behandlungen und weniger für andere therapeutische Behandlungen eine Gefahr „als intrusives Drittes“ bedeutet, kann hier letztlich nicht befriedigend beantwortet werden. Auch die Frage, inwieweit Therapien selbst unter naturalistischen Forschungsbedingungen anders verlaufen, als ohne die forschende Beobachtung von außen, muss an dieser Stelle unbeantwortet bleiben. Grundsätzlich wird jedoch die Tatsache, ob das Beforschtwerden zu einer Gefahr für die therapeutische Arbeit und den analytischen/therapeutischen Raum wird, davon abhängen, wie sich die Beforschung tatsächlich gestaltet und ob zentrale Bedingungen seitens der Forschenden erfüllt werden: Zum einen muss der Datenschutz und der gewissenhafte Umgang mit den erhobenen Daten explizit und ausnahmslos eingehalten werden; bei komparativen Studien, in denen die Zuteilung der Patienten zu den unterschiedlichen Behandlungsformen per Zufall erfolgt, muss sichergestellt werden, dass für jeden einzelnen Patienten grundsätzlich beide Verfahren indiziert sind (vgl. Huber, Henrich, Gastner & Klug, 2012). Letzteres gilt natürlich auch für Kontrollgruppendesigns, in denen jedoch keine Indikation für die Kontrollbedingung eingeräumt, sondern vielmehr ethische Bedenken gegenüber der Kontrollbedingung ausgeräumt werden

müssen. Den für die Studie in Frage kommenden Patienten muss zugesichert werden, dass ihnen bei Ablehnung der Studienteilnahme keinerlei Nachteile erwachsen (etwa Behandlungsangebot zu einem späteren Zeitpunkt als unter Studienteilnahme). Nach Protz et al. (2012) kann es sich zudem positiv auswirken, wenn für die behandelnden Studientherapeuten weitgehende Transparenz herrscht (soweit wie möglich), wenn sie in den Forschungsprozess und in Absprachen bzgl. des forschenden Vorgehens einbezogen werden und wenn den Forschenden die besonders sensible therapeutische/psychoanalytische Situation bekannt ist und sich dies in einem dementsprechend umsichtigen und verantwortungsvollen Umgang mit den Patienten niederschlägt. Zudem sollten behandlungsbezogene Entscheidungen dem Therapeuten und dem Patienten obliegen, ohne dass eine sog. Forscherinstanz von außen eingreift. Unter diesen Bedingungen, die für Kurzzeittherapie- und Langzeittherapiestudien gleichermaßen gelten, könnte es gelingen, negative Einflüsse auf die Intimität der therapeutischen Situation, wie sie Leuzinger-Bohleber et al. (2001) postulieren, im Rahmen prospektiver Beforschungen zu minimieren. Da in den meisten Bewilligungsverfahren von Drittmittelprojekten mittlerweile ein positives Votum einer unabhängigen Ethikkommission obligatorisch ist, soll hier (wenn auch vorsichtig) angenommen werden, dass ein Großteil der o.g. Bedingungen mittlerweile zu den ethischen Standards der psychotherapeutischen Evaluationsforschung gehört. Dies könnte die von Leuzinger-Bohleber et al. (2001) vorgebrachte Begründung für ein rein katamnestisches Design zwecks Wahrung der therapeutischen Intimität möglicherweise in ihrer Bedeutung schmälern – wenn auch sicherlich nicht gänzlich widerlegen. Psychotherapieforschung, ob zur Wirkweise oder zur Wirksamkeit oder ob zu kurzen oder langen Behandlungsdauern, bleibt ein Unterfangen, in dem zahlreiche ethische Regeln zu beachten sind und die bei Nichtbeachtung u.U. großen Schaden anrichten können. Dennoch kann festgehalten werden:

> Patienten bzw. deren Solidargemeinschaften (die Krankenkassen) haben ein natürliches Interesse daran, zu erfahren, mit welcher *Wahrscheinlichkeit* ein Patient mit einer gegebenen psychischen Problematik

> *ungefähr* zu welchem der oben zitierten Rückblicke [Therapieerfolg oder -misserfolg] kommt. Wer würde sich beispielsweise einer komplizierten Operation am offenen Herzen unterziehen, wenn die Ärzte, die diese Operationen durchführen, sich weigerten, die Ergebnisse und möglichen Komplikationen dieser Operationsmethode untersuchen zu lassen, indem sie auf ihre besondere Wissenskultur verweisen? Eine solche Operationsmethode hätte wohl zu Recht nichts im Regelleistungskatalog der gesetzlichen Krankenkassen zu suchen. (Benecke, 2014b, S. 62)

Nimmt man als praktizierender Psychotherapeut und als Psychotherapieforscher also die Verpflichtung ernst, die mit der berufsrechtlichen und sozialrechtlichen Anerkennung von Psychotherapie einhergeht, für das, was Psychotherapie bewirken kann, die möglichst beste Evidenz *unter Wahrung aller legitimen Forderungen des Gegenstands* vorzulegen, dann muss an diesem Punkt eingeräumt werden, dass Ein-Messzeitpunkt-Designs keine Grundlage für die beste Evidenz bieten. Selbst im Hinblick auf den Gegenstand "Langzeittherapie" kann nach den erfolgten Ausführungen nicht angenommen werden, dass dieser es besonders erfordert, primär retrospektiv bzw. rein katamnestisch beforscht zu werden. In Bezug auf das hier fokussierte K.O.-Kriterium zur Veränderungs- und Zielerreichungsdarstellung (B.12.), das dafür sorgte, dass die drei Studien mit ausschließlicher Post- bzw. katamnestischer Querschnittmessungen als methodisch unzureichend bewertet wurden, kann somit festgehalten werden, dass von B.12. keine tatsächlich benachteiligende Wirkung hinsichtlich Langzeittherapiestudien ausgeht. Die drei besagten Langzeittherapieuntersuchungen schneiden aufgrund ihres Designs (Ein-Messzeitpunkt-Design) negativ auf diesem Kriterium ab, jedoch besteht zwischen diesem Design und dem Gegenstand "Langzeittherapie" kein zwingender Zusammenhang.

5.1.7 Zusammenfassende Bewertung der Dimension der allgemeinen methodischen Qualität

Der erste Eindruck, die allgemeine methodische Qualitätsdimension könnte einen besonders hohen Maßstab an die zu begutachtende empirische Evidenz gelegt haben, konnte auf mehrer-

lei Ebenen entkräftet werden: Zum einen konnte das augenscheinlich hohe Aufkommen an Studien mit einem negativen Gesamtergebnis auf dieser Dimension durch die Tatsache relativiert werden, dass für die Studienrecherche ein vorwiegend weitmaschiges Netz an Einschlusskriterien angelegt wurde. Somit gelangten Untersuchungen in den Studienpool, die bspw. in keine der gesichteten Metaanalysen eingingen, da in diesen die zu integrierenden Studien in der Regel einer Qualitätskontrolle unterzogen werden. Zusätzlich müssen auch diejenigen Studien, die durch die Handsuche in Sekundärstudien und Überblickswerken gefunden wurden, nochmals unterteilt werden in solche, die in systematische Reviews (Metaanalysen) eingingen und solche, die in narrative Reviews oder auch nur in Überblicksartikeln rezitiert wurden. Vor diesem Hintergrund erscheint auch die hier ermittelte und verhältnismäßig niedrige Sensitivität der methodischen Qualitätsdimension in einem anderen Licht: Es muss aufgrund des gewählten Außenkriteriums und der darin zugrunde gelegten Operationalisierung von *tatsächlich* qualitativ guten Studien ("Review-Studien") unbedingt beachtet werden, dass unter „Reviews“ eben auch narrative Reviews und Überblicksartikel fallen. Somit ist davon auszugehen, dass die Variabilität der Studienqualität innerhalb der Gruppe der "Review-Studien" besonders hoch ist, was die vergleichsweise niedrige Sensitivität erklären und darüber hinaus rechtfertigen könnte.

Eine übermäßige Strenge der methodischen Qualitätsdimension konnte ferner durch eingehende Untersuchungen einzelner Kriterien ausgeräumt werden. In Bezug auf das ausschlussstärkste K.O.-Kriterium (B.12.: Bewertung der Veränderungs- und Zielerreichungsmaße) konnte aufgezeigt werden, dass eine optimale Bewertung (Stufe „1“) zwar einen hohen Anspruch an die in den Studien verwendeten Outcomemaße stellt, der darin besteht, dass sie zwecks Herleitung eines Cutoff-Werts über Normierungsdaten verfügen müssen, die zwischen der gesunden und dysfunktionalen Population zu trennen vermögen. Gleichsam konnte jedoch auch gezeigt werden, dass es Untersuchungen – mit Ausnahme etwa der Ein-

Messzeitpunkt-Designs sowie der „halben Kontrollgruppendesigns“ – gelingen kann, dieses K.O.-Kriterium mühelos mit zumindest zufriedenstellendem Rating (Stufe „2“) zu bestehen und somit nicht aus der weiteren Bewertung auszuscheiden. Sowohl in Bezug auf Langzeittherapiestudien im retrospektiven Ein-Messzeitpunkt-Design als auch bzgl. Studien zu längeren Behandlungen in „halben Kontrollgruppendesigns“ wurde deren Aussagekraft im Hinblick auf Wirksamkeitsaussagen kritisch reflektiert. Auf Basis dieser Reflektion und unter Verweis auf komparative Langzeittherapiestudien neueren Datums, in denen prospektive Prä-Post-Katamnesemessungen realisiert wurden bzw. werden, kann auch der Verdacht ausgeräumt werden, das K.O.-Kriterium B.12. könnte Langzeittherapiestudien explizit benachteiligen.

Auch in Bezug auf das K.O.-Kriterium A.2., das sog. „Diagnosestellung-Kriterium“ mit den zweithöchsten Ausschlussraten, konnten nach intensiver Sichtung der Studien mit negativen Ratings auf diesem Kriterium kaum Hinweise darauf gefunden werden, dass es sich bei den Anforderungen dieses Kriteriums um inadäquate handelt. Einzige Ausnahme bildet der Originalbericht zur Menninger Studie (Wallerstein, 1986), der als Zusatzliteratur für die Bewertung einer Reanalyse (Studie 4) herangezogen wurde. In diesem wird explizit auf die Inkompatibilität gängiger Klassifikationssysteme (ICD und DSM) und psychoanalytischer Diagnostik hingewiesen und damit eine Kritik vorweggenommen, die 10 Jahre später zur Entwicklung der OPD (Arbeitskreis OPD, 1996) führte. Nun muss an dieser Stelle eingeräumt werden, dass auch die OPD, wenn sie auch vorrangig auf psychoanalytische Konstrukte fokussiert (Konflikte, Struktur etc.), eine eigene Diagnostikachse für ICD-Klassifikationen vorsieht. Das bedeutet, psychoanalysenahe Diagnostik via OPD schließt die atheoretische Klassifikation via ICD keineswegs aus. Doch muss noch ein anderer Aspekt berücksichtigt werden: Im Zusammenhang mit dem sog. „Krankheitswert-Kriterium“ (C.1.) wurde eingehend diskutiert, inwieweit sich in diesem Kriterium der Versorgungsauftrag psychologischer und ärztli-

cher Psychotherapeuten widerspiegelt, der die Behandlung von Klienten mit *krankheitswertigen* psychischen Störungen vorschreibt. Der Krankheitswert einer psychischen Störung wurde in dieser Arbeit in Anlehnung an den WBP mit ICD- oder DSM-Diagnosen gleichgesetzt: „In diesem Sinne besteht im DSM-IV und in der ICD-10 eine Tendenz, den Begriff der Störung implizit mit dem der Krankheitswertigkeit zu verbinden“ (WBP, 2008c). Was bedeutet dies für Wirksamkeitsstudien? Strenggenommen bedeutet dies, dass sich Wirksamkeitsstudien dann für die berufsrechtliche (und sozialrechtliche) Anerkennung als irrelevant erweisen, wenn sie sich auf keine Patienten mit krankheitswertiger oder zumindest „wahrscheinlicher klinischer Störung“ (WBP, 2010, S. 34) beziehen. Die damit verbundenen Forderungen im „Krankheitswert-Kriterium“ (C.1.) und auch im „Diagnosestellung-Kriterium“ (A.2.) (beides K.O.-Kriterien der methodischen Qualitätsdimension) resultieren daher – neben methodischen Gründen – aus gesundheitspolitischen bzw. gesetzlichen Gegebenheiten heraus. Mögliche Kritik an der „Philosophie“ der beiden genannten Kriterien ist daher auf einer anderen Ebene zu diskutieren, als auf der Ebene des Methodenpapiers.

Auch die auf den systematischen Vergleich zwischen Kurz- und Langzeittherapiestudien folgende Feinanalyse im Hinblick auf das „Diagnosestellung-Kriterium“ (A.2.) konnte keine stichhaltigen Indizien hervorbringen, die auf eine tatsächlich benachteiligende Wirkung dieses Kriteriums in Richtung Langzeittherapiestudien hinweisen.

Insgesamt sollen daher die wirksamsten Kriterien der allgemeinen methodischen Qualitätsdimension – die K.O.-Kriterien (vgl. Tabelle 21, S. 252) – im Hinblick auf die hier zugrunde gelegten Studien als angemessen und im Sinne Schultes (Deutsches Ärzteblatt, 2008a) als realistisch zu erfüllende Kriterien bezeichnet werden. Den beiden feinanalytisch untersuchten K.O.-Kriterien (gleichsam die ausschlussstärksten Kriterien) konnte keine benachteiligende Wirkung auf Langzeittherapiestudien nachgewiesen werden. Inwiefern dies auf jedes einzelne

der restlichen Kriterien der methodischen Qualitätsdimension zutrifft, bleibt eine weiterhin zu untersuchende Forschungsfrage. Die Einschätzung, inwieweit der hier gezogene Schluss auch auf andere Studienpopulationen zu übertragen ist, soll zu einem späteren Zeitpunkt vorgenommen werden.

5.2 Die Dimension der internen Validität

Von den 15 Studien mit positiver Bewertung auf der allgemeinen methodischen Qualitätsdimension wurden nur drei Studien als intern valide bewertet. Einer einzigen Studie wird hinreichende interne *und* externe Validität attestiert. Die interne Validitätsdimension samt ihres K.O.-Kriteriums (B.8.), demzufolge nur randomisierte oder parallelisierte Studien auf dieser Dimension bestehen können, legt augenscheinlich besonders hohe Maßstäbe an Wirksamkeitsstudien an. Jedoch relativiert sich dieser Eindruck nach einem vertiefenden Blick in die Studienlage der 15 als allgemein methodisch gut beurteilten Studien, denn diesen liegen folgende Studiendesigns zugrunde:

Die 15 Studien setzen sich zusammen aus

- acht "echten Ein-Gruppen-Designs"
- drei "Ein-Gruppen-Designs im erweiterten Sinne"
- vier Mehrgruppendesigns.

Wie in Kapitel 1.2.2 dargestellt wurde, ist die interne Validitätsdimension für die Bewertung von Untersuchungen im *Mehrgruppendesign* vorgesehen, so dass sowohl die acht Studien in "echten Ein-Gruppen-Designs" als auch die drei Studien in "Ein-Gruppen-Designs im erwei-

terten Sinne"[75] automatisch auf dieser Dimension durchfallen. Es lohnt ein genauerer Blick in die drei Studien, in denen zwar mehrere Treatmentarme miteinander verglichen werden, die jedoch in der Begutachtung durch den WBP-Kriterienkatalog wie Ein-Gruppen-Designs behandelt werden ("Ein-Gruppen-Designs im erweiterten Sinne"). Dabei handelt es sich bei einer Studie (11) um einen Vergleich zwischen tiefenpsychologisch fundierter und analytischer Psychotherapie und bei einer weiteren Studie (31) um einen Vergleich zwischen zwei Behandlungsformen, die in Deutschland beide unter die tiefenpsychologisch fundierte Therapie subsumiert werden würden (*time-limited short-term group therapy: interpretive* versus *supportive therapy*). Bei der dritten Studie (15) handelt es sich schließlich um einen Vergleich von psychodynamischer Therapie *mit* Übertragungsdeutungen *versus* psychodynamischer Therapie *ohne* Übertragungsdeutungen (beide Treatments wurden eingeordnet unter die tiefenpsychologisch fundierte Therapie) und damit um eine sog. *Dismantling*-Studie. Unter *Dismantling*-Studien sind Untersuchungen zu verstehen, in denen eine bestimmte technische Komponente, die an sich fester Bestandteil einer Therapieform ist, bei einer Vergleichsgruppe entfernt wird, um bspw. die inkrementelle Wirksamkeit eben dieser Komponente zu bemessen (vgl. Hager, 2000). Weitere Möglichkeiten bestehen darin, einer Therapieform explizit neue, verfahrensfremde Bestandteile hinzuzufügen (=*Constructive*-Designs) oder aber bestimmte Elemente einer Therapieform über mehrere Bedingungen hinweg zu variieren (=*Parametric*-Designs) (vgl. Kazdin, 1994). Ziel dieser Designs ist es, durch Erhebung möglicher Wirksamkeitsunterschiede zwischen den Treatmentarmen mehr darüber zu erfahren, ob die variierten Komponenten (Techniken) einer bestimmten Therapieform die tatsächlich wirksamen sind: „These designs allow increasingly specific cause-and-effect conclusions and thus markedly enhance our basic knowledge ("this causes that") about therapeutic change“ (Borkovec &

[75] Zu den Bezeichnungen der "echten Ein-Gruppen-Designs" und der "Ein-Gruppen-Designs im erweiterten Sinne" vgl. Kapitel 3.2.1.

Castonguay, 1998, S. 137). Nun handelt es sich bei der eben erwähnten *Dismantling*-Studie um eine prospektive, randomisierte, manualisierte Untersuchung mit Prä- und Postmessung sowie einer 3-Jahres-Katamnese, die auf der internen Validitätsdimension *als Mehrgruppen-design* an sich sehr gut abgeschnitten hätte (vgl. Høglend et al., 2006, 2008). Wieso aber ist eine solche methodisch gute, *eigentlich* intern valide *Dismantling*-Studie für die wissenschaftliche Anerkennung nicht in ihrer gesamten Informationsbreite von Interesse, sondern geht allenfalls als Studie im Ein-Gruppen-Design in die Begutachtung ein? Die Antwort könnte folgendermaßen lauten: Der WBP interessiert sich – zumindest, was seine systematische Bewertung der Evidenz und allem voran die interne Validitätsbewertung betrifft – nicht für alle Fragen der Psychotherapieforschung und bei genauerer Betrachtung zeigt sich, dass er sich eher für einen nur sehr umgrenzten Teil davon interessiert. Um dies zu verdeutlichen, muss an dieser Stelle der Diskussion statt einer Betrachtung allein auf der Ebene der internen Validitätsdimension eine Charakterisierung und Bewertung des Kriterienkatalogs auf allgemeinerer Ebene vorweg genommen werden. Der Grund für diese Vorwegnahme besteht darin, dass sich die Frage danach, für welche empirischen Befunde der Psychotherapieforschung sich der WBP maßgeblich interessiert, zwar partiell auf der Ebene der internen Validitätsdimension beantworten lässt. Gleichzeitig spannt diese Frage jedoch einen sehr viel weiteren Bogen, der wiederum eine grundlegende Kritik am WBP berührt, die im Rahmen dieser Arbeit nicht unkommentiert bleiben soll. In Bezug auf die interne Validitätsbewertung interessiert sich der WBP für solche Forschungsbemühen, die Aussagen zur sog. *isolierten Wirksamkeit* oder aber zur *vergleichenden Wirksamkeit* eines Verfahrens oder einer Methode erlauben. Die isolierte Wirksamkeit fragt Hager (2000) zufolge danach, ob eine Intervention überhaupt wirkt, und diese Frage ist am besten zu beantworten, indem man eine Intervention mit Kontrollbedin-

gungen (z.B. Placebo, TAU) vergleicht (u.a. Borkovec & Castonguay, 1998)[76]. Eine vergleichende Wirksamkeitsuntersuchung dagegen sagt lediglich etwas über die *relative* Wirksamkeit zwischen zwei oder mehreren psychotherapeutischen Verfahren aus (z.B. zwischen analytischer und Verhaltenstherapie), da komparative Untersuchungen nicht über dasselbe Ausmaß an Kontrolle verfügen, wie es klassische Kontrollgruppendesigns tun. Wird jedoch als Vergleichsbedingung ein etabliertes Verfahren gewählt, das in seiner Wirksamkeit bereits als hinlänglich bewährt angesehen werden kann, so kann man sich, unter bestimmten Umständen, auf die weiter unten noch näher eingegangen wird, zumindest einer Aussage über die Wirksamkeit des zu überprüfenden Verfahrens annähern.

Nun sind *Dismantling*-Studien als eine Art Sonderform vergleichender Wirksamkeitsuntersuchungen zu betrachten, da sie ebenfalls zwei (oder mehrere) alternative Therapiebedingungen miteinander vergleichen – jedoch handelt es sich hier um einen verfahrensinternen Vergleich und nicht um einen verfahrensexternen Vergleich mit einem etablierten Verfahren. Aus diesem Grund gehen diese Studien lediglich als Ein-Gruppen-Designs in die Kodierung ein, was sich am deutlichsten in der Bewertung der internen Validität niederschlägt: Diese Untersuchungen fallen als Ein-Gruppen-Design-Studien auf der internen Validitätsdimension automatisch durch. Das bedeutet, dass solche Studien nicht im Hinblick auf die interne Validität bewertet werden, *die für die Fragestellung und Schlussfolgerungen dieser Studien zentral ist*. Die o.g. *Dismantling*-Studie würde unter diesen Umständen auf dieser Dimension sehr gut abschneiden. Stattdessen werden diese Studien vor dem Hintergrund der Frage nach der isolierten oder vergleichenden Wirksamkeit auf ihre interne Validität hin bewertet.

[76] Vergleiche mit Wartlisten-Bedingungen erfüllen nach Hager (2000) nicht die Anforderung einer isolierten Wirksamkeitsüberprüfung. Isolierte Wirksamkeitsüberprüfungen fokussieren die Nettowirkung eines Psychotherapieverfahrens, Wartelisten erlauben jedoch lediglich Aussagen zur Bruttowirkung eines Verfahrens (vgl. Kap. 3.1.1). Im Gegensatz dazu lässt der WBP jedoch Vergleiche mit Wartelisten-Bedingungen durchaus zu.

Nun mag man den WBP dafür kritisieren, dass er im Dienste seines primären Interesses an der isolierten oder vergleichenden Wirksamkeit eines Verfahrens den Informationsverlust in Kauf nimmt, der mit der Gleichsetzung von verfahrensinternen Vergleichsstudien und Studien in Ein-Gruppen-Designs einhergeht. Und hier spannt sich der Bogen von der internen Validitätsdimension, hin zu einer grundlegenderen Kritik am allgemeinen Vorgehen des WBP. Denn in der Tat erntet der WBP fundamentale Kritik dafür, dass er zentrale Befunde aus der Psychotherapieforschung wenig registriert bzw. nicht systematisch sichtet. So konzentriert sich eine Kritik an den Verfahrensregeln und dem Kriterienkatalog bspw. darauf, dass der WBP sich damit nicht für Fragestellungen zur Wirkweise eines Verfahrens oder einer Methode interessiert oder aber für Ergebnisse aus qualitativen Einzel- oder Gruppenstudien (vgl. Kriz, 2008). Man könnte vor allem die zuletzt genannte Kritik als Kritik am naturwissenschaftlichen Paradigma interpretieren, in dem quantitative Methoden die prominente Rolle spielen (vgl. Benecke, 2014b). Im Hinblick auf diese Kritik stellt Benecke aber zu Recht in Frage:

> Es wird aber so gut wie nie gefragt, warum das [die Prominenz quantitativer Wirksamkeitsnachweise] so ist. Warum werden von der (wissenschaftlichen) Öffentlichkeit, von Patientenvertretern und von Solidargemeinschaften naturwissenschaftliche Belege verlangt, wenn es darum geht, zu beurteilen, ob es sich bei der Psychoanalyse um eine hinreichend gute, weil für die meisten Patienten wirksame Behandlungsmethode handelt? Die Antwort ist ganz einfach: Weil die naturwissenschaftliche Strategie zur Ermittlung und zur Validierung *dieses* speziellen Wissens die angemessenere ist. (2014b, S. 63)

Es soll daher festgehalten werden: Für Fragen nach der (isolierten oder vergleichenden) Wirksamkeit von Verfahren sind Befunde auf der Basis quantitativer Gruppenstudien essentiell und diese können weder durch Befunde aus der Prozessforschung ersetzt oder mittels ausschließlich qualitativer Methoden erbracht werden. Die Konsequenz ist, dass im Rahmen der

wissenschaftlichen Anerkennung zahlreiche Befunde aus der Psychotherapieforschung somit unbeachtet bleiben oder zumindest keiner systematischen Begutachtung unterzogen werden. Und die Konsequenz ist zudem, dass Befunde aus quantitativen Gruppenstudien im Falle verfahrensinterner Vergleichsstudien (z.B. *Dismantling*-Studien) nicht entsprechend ihrer eigentlichen Intention und Forschungsfrage begutachtet werden. Vor dem Hintergrund, dass das gesamte Begutachtungsprozedere letztlich der *wissenschaftlichen Anerkennung* von Psychotherapieverfahren/-methoden dient, darf allerdings in Frage gestellt werden, ob der alleinige Fokus auf empirische Befunde zur Wirksamkeit eines Verfahrens für das Prädikat „wissenschaftlich" ausreichend ist. Oder ob für diese Auszeichnung nicht tatsächlich auch Befunde zur Wirkweise zu fordern wären, denn: „Prozessanalysen ohne Ergebnis-Relevanz erscheinen ebenso unbefriedigend wie Erfolgsstudien, welche die Frage, was denn eigentlich zum Erfolg geführt hat, nicht differenzierter als mit dem Hinweis auf die Einhaltung eines Manuals beantworten können" (Caspar & Jacobi, 2007, S. 398).

Die Frage danach, welche Befunde für eine *wissenschaftliche Anerkennung* tatsächlich essentiell sind oder obligatorisch sein sollten, soll hier nicht abschließend beantwortet werden, denn letztlich bewegen sich das Thema und die Fragestellung der vorliegenden Arbeit auch innerhalb der Grenzen der quantitativen Wirksamkeitsstudien. Vielmehr sollte mittels dieses Exkurses beleuchtet werden, warum sich der WBP vorrangig für bestimmte Studiendesigns (zur isolierten oder vergleichenden Wirksamkeit) interessiert und welche Konsequenzen daraus hervorgehen (erinnere: *Dismantling*-Studien). Ferner sollte deutlich gemacht werden, dass die kritische Forderung, der WBP möge eine sehr viel breitere Befundlage (Prozessforschung, qualitative Einzelfallforschung etc.) zwecks wissenschaftlicher Anerkennung heranziehen, als eine *zusätzliche* (und möglicherweise gerechtfertigte) Forderung betrachtet werden muss. Die Prozessforschung oder die qualitative Einzelfallforschung kann hingegen nicht als

alternative Forderung betrachtet werden, denn sie kann fehlende Nachweise zur isolierten oder vergleichenden Wirksamkeit keinesfalls kompensieren.

Nach diesen Ausführungen soll der Fokus der Diskussion nun wieder auf die interne Validitätsdimension gelegt werden.

Neben den eingangs (Kap. 5.2) genannten vier Studien im Mehrgruppendesign und den bereits diskutierten drei verfahrensinternen Vergleichsstudien, setzen sich die insgesamt 15 Studien mit guter methodischer Qualität zudem aus acht Untersuchungen im "echten Ein-Gruppen-Design" zusammen. Wie in Kapitel 5.1.6 im Zusammenhang mit Studien mit lediglich einem Messzeitpunkt sollen auch zu Studien ohne jedwede Vergleichsbedingung noch einmal Shadish et al. (2002) zu Wort kommen:

> However, social scientists in field settings will rarely be able to construct confident causal knowledge with the simple pretest-posttest design unless the outcomes are particularly well behaved and the interval between pretest and posttest is short. Persons considering this design should consider adding even more design elements[77]. (S. 110)

Selbst Kurzzeitbehandlungen werden die im Zitat genannte Forderung nach möglichst kurzen Intervallen zwischen Prä- und Postmessungen nicht erfüllen können, so dass sich in Bezug auf den vorliegenden Studienpool die Frage stellt, warum verhältnismäßig viele Untersuchungen lediglich in Ein-Gruppen-Designs realisiert werden. Von einem ethischen und forschungspragmatischen Standpunkt aus betrachtet, wäre zu erwarten gewesen, dass es sich bei diesen Studien vorrangig um Untersuchungen an Stichproben mit besonders seltenen Störungsbildern handelt, so dass eine Vergleichs- oder Kontrollgruppe schlichtweg kaum zustande zu

[77] Unter *design elements* sind bspw. Kontrollbedingungen, mehrere Prämessungen etc. gemeint.

bringen war; oder aber, dass es sich um Störungen handelt, für die nur eine einzige Behandlungsindikation vorlag und jede Vergleichsbehandlung (bspw. Verhaltenstherapie) unzulässig und jede Form von Kontrollbedingung ethisch nicht vertretbar gewesen wäre. Keine der acht Untersuchungen im Ein-Gruppen-Design und guter methodischer Qualität bestätigt jedoch diese Erwartung. Vielmehr werden entweder ökonomische Gründe angeführt, die die Wahl des zugrunde gelegten Designs rechtfertigen sollen, im Ausblick einiger Studien erfolgt dann der Hinweis, dass die erbrachten Befunde in kontrollierten Designs repliziert werden sollten. Oder aber es wird genau andersherum argumentiert, indem auf Befunde verwiesen wird, die bereits zur Genüge aufgezeigt haben, dass Therapie besser wirkt als keine Behandlung und es nunmehr darum ginge, diese Befunde mittels explizit naturalistischer Untersuchungen zu replizieren. Letzterem Argument liegt ein zu Beginn dieser Arbeit ausführlich beleuchtetes Missverständnis zugrunde, nämlich das Missverständnis, naturalistische, extern valide Studien hätten sich um Strategien der internen Validitätssicherung nicht zu bemühen (vgl. Kap. 1.2.1). Wichtig ist in diesem Zusammenhang, dass der WBP in seinen Verfahrensregeln explizit Studien in Ein-Gruppen-Designs akzeptiert (vgl. Kap. 1.2), jedoch mit folgender Einschränkung:

> Sofern bei einer Studie keine Kontrollgruppen vorliegen (*gegebenenfalls bei methodisch adäquaten Studien mit hoher externer Validität* [Hervorhebung v. Verf.]), können ihre Ergebnisse als Wirksamkeitsnachweis gelten, wenn der Therapieeffekt sowohl durch eine signifikante Prä-Post-Veränderung als auch die klinische Bedeutsamkeit der erreichten Veränderung nachgewiesen ist. (WBP, 2010, S. 21)

Diese Regelung spielt solchen Studien in Ein-Gruppen-Designs direkt in die Hände, die keinerlei Strategien zur internen Validitätssicherung umsetzen. Doch wird diese Beschränkung auf die externe Validität auch solchen Studien in Ein-Gruppen-Designs gerecht, in denen interne Validitätssicherungsstrategien umgesetzt werden? Im diesem Kontext kann vor allem auf eine Langzeittherapiestudie (22) – die sog. Göttinger Psychotherapiestudie – unter den

acht Ein-Gruppen-Design-Studien verwiesen werden, der folgendes Vorgehen zu entnehmen ist: „However, the problem of a control condition can be solved tentatively by assessing how much change occurs in patients who are in need of psychotherapeutic treatment, but have not received it" (Leichsenring, Biskup, Kreische & Staats, 2005, S. 440f.). In einer Arbeit mit dem Titel *Change norms: A complementary approach to the issue of control groups in psychotherapy outcome research* stellen Leichsenring und Rabung (2006) das der genannten Langzeittherapiestudie zugrunde liegende Konzept vor. Der Grundgedanke der beiden Autoren besteht darin, der von Shadish et al. (2002) in Bezug auf Studien ohne Kontroll-/Vergleichsbedingungen genannten Forderung von weiteren Designelementen nachzukommen, indem sie einen *externen Vergleichsmaßstab* kreieren: „Epidemiological studies are required to gain data of both short-term and long-term changes in untreated patients with psychiatric disorders" (Leichsenring & Rabung, 2006, S. 606). Systematisch erhobene Daten zum natürlichen Verlauf psychischer Störungen sollen demnach als Kontrollbedingung für Untersuchungen in Ein-Gruppen-Designs fungieren können. Die Autoren geben allerdings im Weiteren zu bedenken:

> At present, data like this are not yet available. Therefore, in a first and preliminary attempt, we chose another approach to obtain such data of reference: to date in psychotherapy research, many RCTs have included control groups that provide data about the changes occurring in patients who did not receive any specific treatment. (S. 606)

Mittels systematischer Sichtung von insgesamt 26 RCTs zu psychodynamischer Psychotherapie eruieren die Autoren einen mittleren Gesamteffekt für Kontrollbedingungen (Wartelisten und TAU), der bei 0.12 (SD 0.19, Konfidenzintervall 95%: 0.05 – 0.28) liegt. Dieser Gesamteffekt wird in der o.g. Langzeittherapiestudie (Göttinger Psychotherapiestudie) zum Vergleich herangezogen und lässt die Autoren zu folgendem Schluss kommen: „These results speak

clearly against the conclusion that the effects that we found were due to spontaneous remission, regression to the mean or similar effects“ (Leichsenring et al., 2005, S. 448).

Fraglos sind beide Vorgehensweisen, die Wahl eines Vergleichsmaßstabs in Form natürlicher Verlaufsstudien als auch ein Vergleichsmaßstab in Form von realen Kontrollgruppen aus ehemals durchgeführten RCTs, mit Limitationen versehen. So weisen de Maat, Dekker et al. (2007) darauf hin, dass es zwischen den Patienten aus Studien zum natürlichen Störungsverlauf und solchen in (Studien-) Therapien einen zentralen Unterschied gibt, der im „help-seeking behavior“ (S. 64) besteht. Den Kontrollgruppen aus anderen RCTs hängt hingegen folgender Nachteil an: So berichten Leichsenring und Rabung (2006) zwar von nicht signifikanten Zusammenhängen zwischen den Effekten der RCT-Kontrollbedingungen und den zeitlichen Umfängen dieser Bedingungen und schließen daraus, dass Kontrollbedingungen aus RCTs sich daher als externe Vergleichsmaßstäbe für Therapiestudien zu ganz unterschiedlichen Behandlungslängen eignen. Am Ende relativieren sie diesen Schluss jedoch mit dem Hinweis, dass die von ihnen gesichteten Kontrollbedingungen aus insgesamt 26 RCTs allesamt aus RCTs zu kurzen bis moderaten Behandlungslängen stammen – die Gründe dafür sind hinlänglich bekannt. Somit sind die von ihnen berechneten Zusammenhänge zwischen den Effekten und Dauern der Kontrollbedingungen auch auf diese Längen zu beschränken. Dies könnte für Langzeittherapiestudien wiederum eher Studien zum natürlichen Störungsverlauf als externen Vergleichsmaßstab nahelegen, die durchaus über mehrere Jahre durchgeführt werden könnten, wenn auch mit erheblichem Aufwand.

Insgesamt ist die Idee einer Entwicklung einer sog. *Veränderungsnorm*, gegen die Ergebnisse aus Untersuchungen ohne Kontroll- oder Vergleichsbedingung getestet werden können, durchaus begrüßenswert, wenn auch eine solche Vergleichsnorm ganz sicher nicht dieselbe interne Validitätssicherungskraft besitzt, wie ein randomisiertes oder parallelisiertes Kontrollgruppendesign (vgl. auch Leichsenring & Rabung, 2006). Aus diesem Grund soll

eine weitere Strategie vorgestellt werden, die einen ganz anderen Weg der internen Validitätssicherung beschreitet, als die meisten bisher vorgestellten und die für Studien in Ein-Gruppen-Designs gut geeignet scheint. Es handelt sich um das in Kapitel 1.2.2 bereits eingeführte Konzept des *coherent pattern matching* nach Shadish et al. (2002). Dieses Konzept beruht darauf, dass eine zu überprüfende Behandlung in ein nomologisches Netz an hinlänglich bewährten Hypothesen eingebettet wird, die sich im Rahmen einer empirischen Wirksamkeitsstudie sodann bestätigen sollten. Praktisch bedeutet das, dass bestimmte Ergebnismuster vorhergesagt werden, die sich in den Daten der Studie wiederfinden lassen sollten: „In work with a single treatment group and no control, much depends on theory predicting different results for different but conceptually related outcomes . . . .“ (Shadish et al., 2002, S. 486). Je komplexer die Hypothesen und die dadurch vorhergesagten Wirk- und Ergebnismuster sind und je mehr in einer Wirksamkeitsstudie eben diese Muster empirisch bestätigt werden, desto geringer ist die Wahrscheinlichkeit, dass etwas anderes als die Behandlung für diese (komplexen) Ergebnismuster verantwortlich ist. Die Wahrscheinlichkeit, dass etwas anderes als die Behandlung selbst ein solches Ergebnismuster zustande bringt, sinkt also mit zunehmender Komplexität der Vorhersagemuster. Die bisherigen Erkenntnisse aus der Outcomeforschung, aber auch aus der Psychotherapieprozessforschung könnten für die Strategie des *coherent pattern matching* ein wertvolles Fundament bieten, aus dem sich hinlänglich abgesicherte Hypothesen ableiten ließen.

Auch vor dem Hintergrund, dass sich in keiner der acht hier diskutierten Studien der Strategie des *coherent pattern matching* bedient wurde, und dass lediglich eine Studie (22) auf eine externe Vergleichsnorm als Kontrollgruppensubstitut zurückgreift, kann doch festgehalten werden, dass sich Studien in Ein-Gruppen-Designs, die sich solcher Strategien bedienen würden, essentiell von solchen Studien (im Ein-Gruppen-Design) unterscheiden, in denen keinerlei interne Validitätssicherungsstrategien umgesetzt werden. Gleichzeitig muss jedoch

festgestellt werden, dass sich dieser Unterschied in der Bewertung mittels der WBP-Kriterien zur internen Validität in keiner Weise niederschlagen würde. Durch das K.O.-Kriterium der internen Validitätsdimension (B.8.), das als Einfallstor für Mehrgruppendesigns fungiert, werden Studien ohne Kontroll-/Vergleichsbedingung nicht mehr differenziert betrachtet, sondern, was die Bewertung der internen Validität betrifft, allesamt gleichbehandelt. An dieser Stelle schlägt sich in der Bewertung der Studien – ähnlich der Gleichbehandlung von verfahrensinternen Vergleichsstudien (z.B. *Dismantling*-Studien) und "echten Ein-Gruppen-Designs" – nieder, dass es nicht um die Bewertung der internen Validität einer Studie *per se* geht. Der WBP setzt hier eindeutig eine Schwelle, über die Studien mit Hilfe eines mindestens parallelisierten Mehrgruppendesigns gelangen müssen, um auf der internen Validitätsdimension überhaupt bestehen zu können. Dieses Vorgehen ist sicher der Tatsache geschuldet, dass es primär um die Selektion solcher Studien geht, die für eindeutige respektive valide Antworten auf die Frage nach der Wirksamkeit eines Verfahrens sorgen. Aus dieser Perspektive scheint es zunächst korrekt, dass der WBP für die wissenschaftliche Anerkennung eines Verfahrens in einem Anwendungsbereich diese nicht etwa durch drei bzw. vier Studien in Ein-Gruppen-Designs abdecken lässt[78], in denen bspw. o.g. interne Validitätssicherungsstrategien umgesetzt werden. Aus einer anderen Perspektive heraus, die in Kapitel 1.2.1 hinlänglich begründet wurde, erscheint dieses Vorgehen jedoch fragwürdig: Es würde näher liegen, von *allen* Studien, die der WBP zwecks wissenschaftlicher Anerkennung zulässt – und dazu gehören eben auch die Studien im Ein-Gruppen-Design – eine Umsetzung von Strategien zur internen Validitätssicherung zu fordern. Mit diesem Vorgehen könnte der WBP sich immer noch auf den Standpunkt stellen, dass Studien in Ein-Gruppen-Designs selbst unter Verwendung der Vorhersage komplexer Ergebnismuster (*coherent pattern matching*) nicht denselben

[78] Die Forderung von drei bzw. vier Studien pro Anwendungsbereich wurde in Kapitel 1.2 dargelegt.

Status an intern validem Wirksamkeitsnachweis erlangen, wie Designs mit Kontroll-/Vergleichsgruppen, denn: „But pattern matching is no panacea that permits confident claims of no bias such as random assignment permits when it is successfully implemented“ (Shadish et al., 2002, S. 486). Gleichsam würde der WBP jedoch erreichen, dass Studien in Ein-Gruppen-Designs sich eben nicht allein durch eine hohe externe Validität zu profilieren haben. Vor dem Hintergrund recht breiter Anwendungsbereiche (z.B. affektive Störungen), in denen psychotherapeutische Verfahren und Methoden ihre Wirksamkeit nachzuweisen haben, gepaart mit einer relativ überschaubaren Anzahl an Studien, die pro Anwendungsbereich gefordert werden (drei bzw. vier), wäre es empfehlenswert, den Maßstab im Hinblick auf Studien ohne Vergleichs-/Kontrollgruppe an dieser Stelle tatsächlich etwas höher zu schrauben. Dies würde bedeuten, Studien in Ein-Gruppen-Designs explizit in die Bewertung der internen Validität mit aufzunehmen. Das würde aber gleichzeitig bedeuten, dass andere interne Validitätskriterien erst noch entwickelt werden müssten, denn diejenigen der aktuellen Fassung eigenen sich vorrangig für Mehrgruppendesigns (vgl. Kap. 1.2.2).

Noch ein anderer Aspekt könnte für eine solche Modifikation sprechen: Es ist davon auszugehen, dass in der Planung von Wirksamkeitsstudien oftmals die WBP-Kriterien als eine Art Leitfaden herangezogen werden (zumindest von deutschsprachigen Forscherteams). Damit kommt der Kriterienkatalog einer Art Richtschnur gleich, wie Wirksamkeitsstudien mit möglichst hohem Aussagegehalt idealerweise durchzuführen sind. Umso wichtiger wäre die Signalwirkung, dass gegenstandsangemessene Strategien der internen Validitätssicherung – hier in Bezug auf den Gegenstand unterschiedlicher Untersuchungsdesigns und auf den Gegenstand naturalistischer Studien – immer umzusetzen sind. Dies stünde im Einklang damit, dass der WBP Studien ohne Kontroll-/Vergleichsgruppe explizit akzeptiert und gleichsam damit, dass sich interne und externe Validität nicht grundsätzlich ausschließen (vgl. Heekerens, 2005; Leichsenring, 2004a/b; Shadish et al., 2002).

5.2.1 Die Anwendbarkeit des „Messdesign-Kriteriums“ (B.10.) vor und nach der Rekodierung

Das Kriterium B.10. bemisst, wie viele Messzeitpunkte in einer Untersuchung realisiert wurden und ob es sich bei den verwendeten Outcomemaßen hauptsächlich um prospektive oder um retrospektive Messinstrumente handelt (vgl. Kap. 3.2.1). Eine Rekodierung wurde notwendig, da etwa Studien mit Prä- und Katamnesemessung (ohne spezifische Postmessung) und ggf. mehreren Messzeitpunkten über den Therapieverlauf nicht eindeutig zu kodieren waren. Es konnte gezeigt werden, dass die Modifikation der Stufenoperationalisierungen sich lediglich für zwei der insgesamt 15 Studien mit guter allgemeiner methodischer Qualität als relevant erwies (vgl. Tabelle 26, S. 259). Trotz dieses vergleichsweise geringen Impacts der Rekodierung im Hinblick auf den hiesigen Studienpool, könnte die vorgeschlagene Modifikation von Vorteil sein: Vor allem Wirksamkeitsstudien neueren Datums sind bestrebt, neben psychosozialen Outcomemaßen vermehrt auch Aspekte der Effizienz, d.h. zu Kosten-Nutzen-Aspekten, abzubilden. Dies trifft bspw. auf die Studie zu, der im hier kodierten Studienpool als einzige sowohl hohe interne als auch externe Validität attestiert wurde (Studie 19). Ferner trifft dies z.B. auf die APS von Benecke et al. (2012) zu. Kosten-Nutzen-Analysen werden in der Regel auf der Basis leicht veränderter Messdesigns durchgeführt, hierbei richten sich die Messzeitpunkte nicht nach den individuellen Therapiephasen (z.B. vor Behandlungsbeginn, unmittelbar nach sowie 3 Jahre nach individuellem Behandlungsende), sondern nach der Realzeit ab Behandlungsbeginn (vgl. auch LAC-Studie; Beutel et al., 2012). Das bedeutet, es werden zu Studienbeginn feste Messzeitpunkte ab Therapiebeginn festgelegt, zu denen die Studienpatienten „bemessen“ werden, egal, in welcher Phase ihrer Behandlung sie sich gerade befinden. Allein dieses Messdesign lässt es zu, die Kosten und den Nutzen in eine sinnvolle Relation zueinander zu bringen. Nun fällt die Begutachtung der Effizienz eines psychotherapeutischen Verfahrens nicht in den Aufgabenbereich des WBP, sondern in den des Gemein-

samen Bundesausschusses, doch ist ein solches Messdesign noch aus einem anderen Grund, als der Effizienzbewertung, interessant, denn mit ihm wird etwas erhoben, das gewissermaßen die Patientenperspektive widerspiegelt: Mit welchem Besserungsverlauf hat ein („mittlerer“) Patient in Therapieform A verglichen zu Therapieform B mit welchem Zeit- und Energieaufwand innerhalb eines bestimmten Zeitraums (ab Therapiebeginn) zu rechnen?

Setzen sich diese Messdesigns vermehrt durch, dann würde sich eine erweiternde Modifizierung des Kriteriums B.10. um die Möglichkeit, solche Messdesigns eindeutig kodieren zu können, durchaus lohnen.

Noch ein anderer Aspekt zum sog. „Messdesign-Kriterium“ (B.10.) soll kurz erwähnt werden: Es konnte einige Abschnitte zuvor gezeigt werden, dass Studien mit nur einem Messzeitpunkt (Post- oder Katamnesemessung) keine hinreichenden Aussagen zu Veränderungen oder Zielerreichungen machen können, wie es das darauf bezogene K.O.-Kriterium B.12. fordert. Damit scheiden solche Studien im Querschnittdesign auf der allgemeinen methodischen Qualitätsdimension aus, und es wird fraglich, warum man durch Stufe „3“ des Kriteriums B.10. („ausschließlich Postmessung“) etwas Gegenteiliges signalisieren sollte. Auch dahingehend wäre eine Modifikation des Kriteriums B.10. sinnvoll.

5.2.2 Die Gegenstandsadäquatheit der Dimension der internen Validität in Bezug auf Langzeittherapiestudien im Vergleich zu Kurzzeittherapiestudien

Da die Dimension der internen Validität aus den bekannten Gründen ausschließlich an Studien in Mehrgruppendesigns untersucht werden sollte, von denen lediglich vier Studien zur Verfügung standen, wurde das Vorgehen der Analyse leicht modifiziert. So wurde die Dichotomisierung der drei Kriterienstufen aufgehoben und auf Effektstärkeberechnungen verzichtet. Ebenso wurde auf Feinanalysen verzichtet, da die Auswertung durchweg unter Hinzuziehung der einzelnen Studien erfolgte.

Bei drei der insgesamt 12 internen Validitätskriterien schnitt die eine Langzeittherapiestudie (19) mit negativem Ergebnis (Stufe „3“) ab. Das sind folgende Kriterien:

- „Manual-Kriterium“ (B.3.)
- „Adherence-Kriterium“ (B.6.)
- „Kriterium zur Zulässigkeit nicht randomisierter Interventionen“ (B.7.).

Bei zwei weiteren Kriterien schnitt die Langzeittherapiestudie im Gegensatz zu allen anderen Studien mit einem Missingwert ab, dazu zählen folgende Kriterien:

- „Kriterium zur operationalen Definition der Kontrollbedingungen“ (B.4.)
- „Kriterium zur strukturellen Äquivalenz bei Kontrollbedingungen“ (B.5.).

In Bezug auf die ersten drei genannten Kriterien stellt sich die Frage, ob der Grund für das schlechte Abschneiden der Langzeittherapiestudie im Vergleich zu den Kurzzeittherapiestudien (die hier meist besser abschnitten) tatsächlich mit dem Gegenstand "Langzeittherapie" zu tun hat. Die gleiche Frage stellt sich im Hinblick auf die beiden zuletzt genannten Kriterien, nur dass sich hier die Frage darauf richtet, inwieweit der durch die Langzeittherapiestudie produzierte Missingwert auf diesen Kriterien durch den Gegenstand "Langzeittherapie" erklärt werden kann.

Für die beiden Kriterien, die sich auf den Einsatz eines Manuals (B.3.) bzw. auf die Bemessung der Adherence (B.6.) beziehen, konnte gezeigt werden, dass sich im Rahmen der Langzeittherapiestudie (19) explizit aus Gründen der externen Validität dazu entschieden wurde, auf beide Studienelemente zu verzichten. Nun ist diese Begründung im Hinblick auf den Manualeinsatz durchaus plausibel, jedoch leuchtet sie in Bezug auf die Adherence-Kontrolle nicht unmittelbar ein: Adherence-Kontrollen müssen weder notwendigerweise in Form eines

Monitorings durchgeführt werden, mit dem Ziel die therapeutischen Interventionen durch aktives Eingreifen in den Behandlungsprozess verfahrenstypisch zu optimieren. Noch setzen Adherence-Kontrollen unbedingt einen Manualeinsatz voraus. Vielmehr können Adherence-Kontrollen auch durchgeführt werden, indem von externen Beobachtern auf der Basis audiografierter oder videografierter Therapiesitzungen beurteilt wird, ob das, was in der Behandlung geschieht, im Einklang mit dem Behandlungskonzept steht, ohne dass es zu korrigierenden Rückmeldungen an die Therapeuten kommt (vgl. Leichsenring et al., 2011). Somit bleibt der naturalistische Charakter der Behandlung trotz Adherence-Kontrollen gewahrt. Für Adherence-Kontrollen, die nicht auf der Basis von Therapiemanualen durchgeführt werden, existieren mittlerweile Skalen, wie die bereits eingeführte CPPS (Hilsenroth et al., 2005), die bspw. in der ebenfalls schon erwähnten LAC-Studie zur Adherence-Messung angewandt wird (vgl. Leuzinger-Bohleber et al., 2010). Ferner eignet sich das *Psychotherapy-Process Q-Sort* (PQS; vgl. Ablon & Jones, 1998; Jones, 2000) für die Bemessung der Adherence (vgl. Huber et al., 2013). Mit dem PQS kann beurteilt werden, ob in einer Therapie z.B. prototypisch verhaltenstherapeutische oder prototypisch psychodynamische Internventionen stattgefunden haben (vgl. Benecke et al., 2012). Der Vorteil der genannten Messinstrumente liegt also darin, dass sie in naturalistischen bzw. unmanualisierten Studien eingesetzt werden können. Allein der „Inhalt", der gemessen wird, ist bei unmanualisierten Therapiestudien ein anderer, da gewissermaßen nicht die Manualtreue, sondern vielmehr die Therapiekonzepttreue gemessen wird. Vor diesem Hintergrund ist die Begründung, die in der Langzeittherapiestudie (19) in Bezug auf die bewusst unterlassene Adherence-Messung vorgebracht wird, nicht in Gänze nachvollziehbar.

Lässt man diese Unplausibilität einmal außer Acht, so wird für die hiesige Argumentation jedoch ein ganz anderer Aspekt zentral, nämlich der, dass die Begründung, *warum* auf beide Elemente – Manuale und Adherence-Messung – verzichtet wurde, dem Ansinnen folgt,

die Therapien möglichst nahe an der klinischen Praxis durchführen zu lassen („Accordingly, no manuals were used and no adherence monitoring was organized" [Knekt, Lindfors, Laaksonen et al., 2008, S. 97]). Diese Begründung weist auf keinen Zusammenhang mit der untersuchten Behandlungslänge hin und könnte in einer Untersuchung zu Kurzzeitbehandlungen ebenso herangezogen werden – und faktisch wurde es das in der hier als Langzeittherapiestudie bezeichneten Untersuchung auch, denn neben dem psychodynamischen Langzeittherapiearm wurde noch ein weiterer Arm mit psychodynamischer Kurzzeittherapie durchgeführt.

Nun liegt es auf der Hand, dass zahlreiche Manuale für psychodynamische Kurzzeitbehandlungen existieren und nur vergleichsweise wenige zu Langzeitbehandlungen (vgl. Beutel et al., 2010). Caligor (2005) vergleicht in ihrem Artikel mit dem Titel *Treatment manuals for long-term psychodynamic psychotherapy and psychoanalysis* insgesamt drei Manuale, die sich explizit für analytische Langzeitbehandlungen über mehrere Jahre eignen, davon beziehen sich allerdings zwei primär auf Behandlungen von Borderline-Störungen (*transference focused psychotherapy* nach Otto F. Kernberg und *mentalization-based treatment* nach Peter Fonagy). In der LAC-Studie wird der psychodynamische Treatmentarm nach dem *Travistock-Manual der psychoanalytischen Psychotherapie unter besonderer Berücksichtigung der chronischen Depression* durchgeführt (vgl. Leuzinger-Bohleber et al., 2010; Taylor, 2010). In der derzeit realisierten Therapievergleichsstudie (APS) von Benecke und Kollegen wurde von den Untersuchungsleitern eigens ein Manual verfasst (vgl. Benecke et al., 2012; Benecke, 2014a). Es bleibt also festzuhalten, dass man auf eine relativ geringe Anzahl an Manualen zurückgeworfen sein wird, möchte man psychodynamische Langzeittherapien auf ihre Wirksamkeit hin untersuchen und entscheidet sich, dies manualisiert bzw. behandlungsprinzipiengeleitet zu tun. Dennoch ist auch in Bezug auf psychodynamische Langzeitbehandlungen ein Trend zu beobachten, den Heekerens (2005) folgendermaßen auf den Punkt bringt:

> Auch wenn die Forderung nach voll strukturierten Therapiemanualen in bestimmten Fällen un- oder gar widersinnig erscheint, so muss bei der Evaluation einer Therapie doch sichergestellt sein, dass auch wirklich „drin“ ist, was „drauf“ steht . . . . Das wird auch von Evaluationsforschern aus dem psychoanalytischen Lager (Rad et al. 2001) gefordert. (S. 363)

In Übereinstimmung mit diesem Zitat und vor dem Hintergrund des beschriebenen Trends (APS, LAC-Studie) kann also ebenfalls festgehalten werden, dass der Einsatz von Manualen oder Behandlungsrichtlinien vor allem in komparativen Studien ein sinnvolles Gütekriterium darstellt, das unabhängig von der Therapielänge umsetzbar ist. Der einzige Nachteil, der für Langzeittherapiestudiestudien besteht, ist die Tatsache, dass nur verhältnismäßig wenige Manuale für besonders lange Therapien existieren. Insofern mag vom „Manual-Kriterium“ (B.3.) eine benachteiligende Wirkung im Hinblick auf Langzeittherapiestudien ausgehen, die jedoch nicht konzeptioneller und grundsätzlicher Art ist, sondern allein auf einen Mangel an Langzeittherapiemanualen zurückzuführen ist. Beim „Adherence-Kriterium“ (B.6.) ist hingegen von keiner benachteiligenden Wirkung bzgl. Langzeittherapiestudien auszugehen, es konnte gezeigt werden, dass Adherence-Kontrollen in Kurz- wie in Langzeittherapiestudien auch dann durchzuführen sind, wenn keine Manuale oder Behandlungsrichtlinien angewandt wurden.

Es muss an dieser Stelle noch ein weiterer grundlegender Punkt diskutiert werden, der gegen die Manualisierung bzw. Manualisierbarkeit von Langzeitbehandlungen oftmals angeführt wird. Dieser bezieht sich auf den Grad an Standardisierung – ein Ziel der Manualisierung – der mit zunehmender Behandlungslänge nur abnehmen kann (vgl. Westen et al., 2004). Koss und Shiang (1994) gehen sogar so weit, eine annähernde Standardisierung von psychotherapeutischen Behandlungen grundsätzlich in Frage zu stellen: „At the same time, it is foolish to

believe that the use of manuals alone will “standardize” a therapy. The actual delivery of therapy is dependent on the contributions and interactions that take place between the two people” (S. 692). Und Beutler et al. (2004) schreiben: „Unfortunately standardizing the treatment has not eliminated the influence of the individual therapist on outcomes . . . .” (S. 245). Und so wundert es auch nicht, dass Heekerens (1998, 2005) in diesem Zusammenhang das durch Manuale und Adherence-Kontrollen eingelöste Gütemerkmal unter den Begriff der "Variablenvalidität" subsumiert, statt unter den der internen Validität. Auch Leichsenring et al. (2011) diskutieren in ihrem Artikel zum Thema Therapietreue die Adherence-Kontrollen als Element der "Konstruktvalidität". Beide Begriffe – Variablenvalidität und Konstruktvalidität – werden synonym benutzt (vgl. Westermann, 2000) und beschreiben das Ausmaß, in dem die in einer Untersuchung vorgenommenen Operationalisierungen und Realisierungen der unabhängigen und abhängigen Variablen tatsächlich die hypothetischen Konstrukte repräsentieren, die von Interesse sind (vgl. Shadish et al., 2002). Manuale und Adherence-Kontrollen wären demnach als Mittel zu betrachten, die die Konstruktvalidität der unabhängigen Variablen – der psychotherapeutischen Behandlungen – gewährleisten. In dem eben erwähnten Artikel von Leichsenring und Kollegen (2011) wendet sich das Autorenteam der interessanten Frage zu, inwieweit sich das Studienelement der Adherence-Messung und damit zusammenhängend auch der Manualeinsatz aus unterschiedlichen Perspektiven in den Dienst ganz unterschiedlicher Validitätsarten stellen lässt (interne und externe Validität, Konstruktvalidität und statistische Validität). Im Zusammenhang mit der internen Validität verweisen sie noch auf ein anderes Ziel, das mit Manualen und Adherence-Messungen verfolgt wird, als das meist rezitierte Ziel der Standardisierung der Behandlungen (vgl. auch Kap. 1.2.2):

> Lack of treatment differences in a comparative outcome study may also be due to failures to ensure sufficient adherence, competence or differentiation of treatments. If, for example, there is considerable

> overlap in therapeutic procedures, no differences in efficacy may be found between treatments. (Leichsenring et al., 2011, S. 314)

Durch Manuale oder durch das Befolgen expliziter Behandlungsprinzipien sowie durch Adherence-Kontrollen wird also zumindest annähernd gesichert, dass die zu vergleichenden Behandlungen – etwa analytische und Verhaltenstherapie – sich tatsächlich in zentralen Techniken voneinander unterscheiden. Zeigt sich dann in einer Studie, dass beide Verfahren auf bestimmten Outcomemaßen annähernd gleich gut abschneiden, dann ist zumindest die Alternativerklärung für diese Effektgleichheit ausgeräumt, dass die Verfahren sich zu einem großen Teil gleicher Techniken bedienen. Gleichsam würde z.B. bei einem beobachtbaren Effektunterschied zugunsten der Verhaltenstherapie dieser nicht auf die Alternativerklärung zurückzuführen sein, dass bspw. die analytische Therapie gar keine „echte" analytische Therapie war, weil zentrale Techniken dieses Verfahrens gar nicht umgesetzt wurden o.ä..

Insoweit kann festgehalten werden, dass der WBP mit seiner Entscheidung, die beiden Kriterien („Manual-Kriterium" und „Adherence-Kriterium") der internen Validität zuzurechnen, plausibel gehandelt hat, auch wenn das oft proklamierte Ziel der Standardisierung durch Manuale offenbar verfehlt zu sein scheint – und dies nicht allein bei Langzeittherapien (vgl. Beutler et al., 2004). Es konnte gezeigt werden, dass noch andere Gründe dafür sprechen, im Dienste der internen Validität sicherzustellen, „dass auch wirklich „drin" ist, was „drauf" steht" (s.o.).

Auf Kriterium B.7., dem sog. „Kriterium zur Zulässigkeit nicht randomisierter Interventionen", zeigte sich, dass insgesamt drei Studien, darunter die Langzeittherapiestudie – lediglich mit Stufe „3" abschneiden. Eine Kurzzeittherapiestudie (27) wurde hingegen mit „1" bewer-

tet. Die Frage, die es demnach zu beantworten gilt, lautet, ob es zwingende Gründe in Bezug auf die Langzeittherapiestudie gibt, die ein besseres Ergebnis auf diesem Kriterium grundsätzlich verhindern und die im Gegensatz dazu auf die beiden Kurzzeittherapiestudien nicht zutreffen. Ein Blick in die Studienpublikationen verriet, dass in der Langzeittherapiestudie die Ungleichheit, die zwischen den Gruppen im Hinblick auf die Inanspruchnahme weiterer Interventionen bestand, zwar dokumentiert und auch kritisch diskutiert wurde, jedoch wurden keine statistischen Korrekturen vorgenommen, was schlussendlich zu der Bewertung mit „3“ führte. Wären hingegen statistische Korrekturen vorgenommen worden, so hätte die Studie zumindest nach den hier verwendetet Kodierregeln (vgl. Anhang C) mit „2“ bewertet werden können. Allein eine Bewertung mit „1“ wäre ausgeschlossen gewesen, da dafür jedwede zusätzlichen Interventionen hätten eliminiert werden müssen, wie in einer der Kurzzeittherapiestudien (27) geschehen. Möglicherweise ist hier das Potential einer Benachteiligung zu sehen, da es für Behandlungen umso schwerer wird, weitere Inanspruchnahmen zielführend auszuschließen, je länger sie andauern. Diese Benachteiligung beschränkt sich jedoch allein auf die beiden ersten Stufen des Kriteriums (Stufe „1“ und „2“). Zu der Bewertung mit „3“ hat hingegen nicht die Tatsache geführt, dass es sich bei besagter Studie um eine Untersuchung zu Langzeitbehandlungen handelt, dieser Makel hätte bei allen drei Studien mit einer „3“-er-Bewertung gleichermaßen vermieden werden können.

Als letztes sind die beiden Kriterien mit den vorgesehenen Missingwerten für Studien in Mehrgruppendesigns jedoch ohne Kontrollgruppen, sondern mit psychotherapeutischen Vergleichsgruppen zu besprechen (vgl. Kap. 3.2.1). Hierbei zeigte sich sowohl beim „Kriterium zur operationalen Definition der Kontrollbedingungen“ (B.4.) als auch beim „Kriterium zur strukturellen Äquivalenz bei Kontrollbedingungen“ (B.5.), dass die Langzeittherapiestudie als einzige auf der Missingkategorie landet, da es sich um eine komparative Psychotherapiestudie

ohne Kontrollgruppe handelt. Wie bereits im ersten Teil der Diskussion (Kap. 5.1) erörtert, ist die Umsetzung von Kontrollgruppen, wie Wartelisten, Placebo-Kontrollen oder auch TAU, bei Untersuchungen zu längeren Therapiedauern mit besonderen Schwierigkeiten behaftet. Als einen potentiellen Ausweg haben Leichsenring und Rabung (2006) zwar den Vergleich mit natürlichen Störungsverläufen angeraten, jedoch muss die Umsetzung dieses Vorhabens zum jetzigen Zeitpunkt für viele Störungsbereiche noch als „Zukunftsmusik" betrachtet werden, da die epidemiologische Forschung zu unbehandelten Störungsbildern, samt systematischer Erhebungen mittels zentraler Outcomemaße, noch vergleichsweise spärlich ist (u.a. de Maat, Dekker et al., 2007, S. 63). Damit bleibt die Untersuchung zu Langzeitbehandlungen in Mehrgruppendesigns zunächst auf komparative Vergleichsstudien begrenzt[79], die, wie in Kapitel 3.2.1 ausführlich dargelegt, die in Kriterium B.5. geforderte strukturelle Äquivalenz der Vergleichsgruppen nicht erfüllen können bzw. sollten. Dies hätte sich in der hiesigen Langzeittherapiestudie (19) dahingehend ausgewirkt, dass die analytische Behandlung in Sitzungsfrequenz und Dauer dem systemischen Therapiearm hätte angeglichen werden müssen oder umgekehrt. Somit hätte die jeweils strukturell angeglichene Vergleichsbedingung nicht mehr als „echte" Vergleichsgruppe fungiert, sondern nur noch als etwas, das Hager (2000) als QAP (Quasi-Alternativprogramm) bezeichnet (vgl. Kap. 3.2.1).

Was sagt dies nun über die beiden hier diskutierten Kriterien B.4. und B.5. aus? Zunächst ist festzuhalten, dass, unabhängig von der untersuchten Therapielänge, komparative Studien auf diesen beiden Kriterien nicht zu bewerten sind und erzwungenermaßen Missinwerte erzeugen. Im Gegensatz zu Kurzzeittherapiestudien existiert für Untersuchungen zu längeren Behandlungen kaum die Möglichkeit, Kontrollgruppen heranzuziehen, so dass vor diesem Hinter-

[79] de Maat, Dekker et al. (2007) berichten zwar von der Umsetzbarkeit von TAU-Kontrollbedingungen bei psychotherapeutischen Langzeitbehandlungen, allerdings beschränken sie diese allein auf den Anwendungsbereich der Borderline-Störung (vgl. Kap. 5.4).

grund eine Benachteiligung von Langzeittherapiestudien gesehen werden kann. Diese führt dazu, dass Studien zum Gegenstand "Langzeittherapie" mit hoher Wahrscheinlichkeit als „nicht beurteilbar“ (Missingwert) kodiert werden müssen. Gleichzeitig werden diejenigen Studienelemente, die für die interne Validität bei komparativen Studien zentral sind, durch den WBP-Kriterienkatalog nicht abgefragt. Diese Elemente wären jedoch gerade im Hinblick darauf, dass eben Langzeitbehandlungen primär im Rahmen komparativer Studien beforscht werden können, nicht zu vernachlässigen: Es wurde bereits weiter oben darauf hingewiesen, dass aus komparativen Studien strenggenommen nur auf die relative Wirksamkeit, jedoch nicht auf die isolierte Wirksamkeit geschlossen werden kann (siehe Kap. 5.2 und vgl. Hager, 2000). Dieser Umstand ist unbefriedigend und wird nicht allein dadurch verbessert, dass als Vergleichsbedingung ein Verfahren gewählt wird, das in seiner Wirksamkeit bislang als hinlänglich bewährt gilt. Um sicher zu stellen, dass die Vergleichsbedingung in der komparativen Studie tatsächlich ihr gesamtes Wirksamkeitspotential ausschöpft und einen angemessenen Vergleichsmaßstab darstellt, sollte in einer komparativen Studie die zum Vergleich herangezogene Behandlung unter ähnlichen Bedingungen durchgeführt werden, wie in vormaligen Untersuchungen, in denen eben diese Vergleichsbehandlung sich klassischen Kontrollbedingungen als überlegen erwies (vgl. Hager, 2000; Temple & Ellenberg, 2000). Die Ähnlichkeit bezieht sich bspw. auf die untersuchte Patientenklientel oder die Dauer der Behandlung (vgl. Kleist, 2006). Wie schon des Öfteren im Rahmen dieser Arbeit betont, ist es in komparativen Studien darüber hinaus zentral, dass beide (oder alle) psychotherapeutischen Behandlungen der unterschiedlichen Treatmentarme so durchgeführt werden, wie es dem jeweiligen Therapiekonzept entspricht. Das bedeutet, dass eben *keine* strukturelle Angleichung zwischen den Treatmentarmen stattfinden sollte, wie von Kriterium B.5. gefordert. Demgegenüber sollte das Merkmal „therapeutische Berufserfahrung“ über alle Bedingungen konstant gehalten werden, um evtl. Effektunterschiede nicht auf unterschiedliche therapeutische Kompetenzen zu-

rückführen zu können (vgl. Hager, 2000). Das Durchführen von Psychotherapievergleichsstudien sollte zudem möglichst innerhalb eines Forscherverbundes stattfinden, der sich aus Anhängern *beider* bzw. *aller* zu vergleichenden Therapieformen zusammensetzt – so geschehen bspw. in der Vergleichsstudie „SOPHO-NET“ (Leichsenring et al., 2009). Diese bewusste Zusammensetzung dient der Kontrolle eines möglichen *researcher allegiance bias*' (vgl. Caspar & Jacobi, 2007; Munder, Brütsch, Leonhart, Gerger & Barth, 2013).

Eher der allgemeinen methodischen Qualität als der internen Validität zuzurechnen, sollten die Stichproben in komparativen Untersuchungen von adäquater Größe sein, um kleine Effektunterschiede mit angemessener Power entdecken zu können – etwas, das der WBP mit einem seiner methodischen Qualitätskriterien (A.17.) bereits explizit erhebt. Eng damit zusammen hängt zudem, dass die inferenzstatistische Analyse der Daten in komparativen Studien im Grunde genommen nach der Logik der *Non-Inferiority*-Hypothesentests erfolgen müsste, statt nach klassischen Hypothesentests. In Kapitel 3.2.3 wurde mit einem kurzen Zitat dieser Sachverhalt pointiert: „*Absence of evidence* [of a difference] *is not evidence of absence* [of a difference]“ (Kleist, 2006, S. 815). In keiner der hier kodierten komparativen Studien wurden *Non-Inferiority*- oder Äquivalenzhypothesentests durchgeführt, sondern ausschließlich klassische Nullhypothesentests. Hella Klemmert (2004) hat über die Frage, warum sich dieses wohl eher in der Biostatistik ansässige Prinzip der *Non-Inferiority*- oder Äquivalenzhypothesentests in der psychologischen/psychotherapeutischen Forschung nicht durchzusetzen scheint, eine bemerkenswerte Arbeit verfasst. Um den Rahmen der vorliegenden Arbeit nicht zu sprengen, soll es bei dem Verweis auf diese Arbeit belassen und der Frage an dieser Stelle nicht weiter nachgegangen werden.

5.2.3 Zusammenfassende Bewertung der Dimension der internen Validität

In der Auseinandersetzung mit der internen Validitätsdimension und der Beschränkung der Dimension auf die Begutachtung von Studien in „echten“ Mehrgruppendesigns (in Abgrenzung etwa zu *Dismantling*-Studien o.ä.) konnte zunächst zweierlei aufgezeigt werden: Zum einen, dass der WBP sich in seiner systematischen Bewertung der internen Validität von Studien für diejenigen Befunde interessiert, die Aussagen zur isolierten und vergleichenden Wirksamkeit von psychotherapeutischen Verfahren und Methoden zulassen. Schlüsse, die etwa aus *Dismantling*-Studien oder sonstigen verfahrensinternen Vergleichen oder auch aus reinen Prozessanalysen sowie der qualitativen Einzelfallforschung gezogen werden können, sind für die systematische Begutachtung nicht von Interesse. Geht es in der wissenschaftlichen Anerkennung von Verfahren/Methoden vorrangig um die Feststellung der (isolierten oder vergleichenden) Wirksamkeit und weniger z.B. um die Wirkweise einer Therapieform, so handelt der WBP mit seiner primären Forderung von quantitativen Gruppenstudien sehr richtig und die an ihn gerichtete Kritik muss in diesem Punkt als haltlos zurückgewiesen werden. Allein die Frage bleibt offen, ob das vorrangige Interesse an der Wirksamkeit für das Prädikat „wissenschaftlich“ ausreichend ist oder ob es dazu nicht tatsächlich einer Ausweitung auf die systematische Begutachtung der Forschung zur Wirkweise bedarf.

Zum anderen konnte auf Basis der kodierten Studien aufgezeigt werden, dass die Dimension der internen Validität in einem Aspekt der Auffassung eines inversen Verhältnisses von interner und externer Validität in die Hände spielt: 11 der 15 Studien mit guter methodischer Qualität fielen aufgrund ihres Ein-Gruppen-Designs automatisch durch die interne Validitätsbewertung durch. Vor dem Hintergrund, dass der WBP diese Studienart grundsätzlich für seine systematische Begutachtung mittels des Kriterienkatalogs zulässt, ist es nicht verständlich, warum er von diesen Studien nicht explizit mehr fordert – nämlich die Umsetzung solcher Strategien der internen Validitätssicherung, die für eben dieses Studiendesign ange-

messen und möglich sind. Vermutlich haben pragmatische Gründe zu der Entscheidung geführt, Studien in Ein-Gruppen-Designs lediglich als extern valide Studien zuzulassen, denn eine Bewertung der internen Validität würde die Entwicklung anderer Kriterien erfordern, als die jetzigen. Diese pragmatische Entscheidung des WBP steht jedoch nicht im Einklang mit dem wissenschaftlichen Standard, demzufolge die interne Validität stets als grundlegendes Gütemerkmal einer Studie betrachtet wird. Auch, um den wissenschaftlichen Status naturalistischer Studien – ob in Mehr- oder Ein-Gruppen-Designs – zu stärken, wäre es daher ratsam, von ihnen die maximale Umsetzung interner Validitätssicherungsstrategien zu fordern. Diese Forderung sollte sich dementsprechend in den WBP-Kriterien niederschlagen. Für die wissenschaftliche Anerkennung pro Anwendungsbereich könnte die bisherige Forderung von ein bis zwei intern und ein bis zwei extern validen Studien (vgl. Kap. 1.2) etwa derart revidiert werden, dass diejenigen Studien mit einem Fokus auf der externen Validität (nicht randomisiert, unmanualisiert, keine engen Ein-/Ausschlusskriterien etc.) nur dann als ausschlaggebend betrachtet werden, wenn *auch* Strategien zur internen Validitätssicherung sinnvoll umgesetzt wurden. Die Präferenz könnte problemlos auch weiterhin auf Studien in Mehrgruppendesigns liegen, die sich bspw. allein durch *mehrere* Studien in Ein-Gruppen-Designs ersetzen ließen.

Die Modifikation des sog „Messdesign-Kriteriums“ (B.10.) hatte, was die Anzahl der Studien betrifft, die dadurch kodierbar wurden, keinen beachtlichen Impact – zumindest wenn man nur diejenigen Studien mit ausreichend guter methodischer Qualität in Betracht zieht. Dennoch konnte aufgezeigt werden, dass sich die vorgenommene Modifikation vor allem für Messdesigns, die in neueren Wirksamkeitsstudien umgesetzt werden, als vorteilig erweisen würde. Gerade auch im Hinblick auf Untersuchungen zur Wirksamkeit analytischer Langzeitbehandlungen ist davon auszugehen, dass diese vermehrt auch Effizienzanalysen integrieren und daran angelehnt ihr Messdesign ausrichten werden (vgl. APS; Benecke et al., 2012).

Die Diskussion der vergleichenden Untersuchung von Kurz- und Langzeitbehandlungen, für die nur noch eine vergleichsweise geringe Anzahl an Studien zur Verfügung stand (n=4), hat aufgezeigt, dass sowohl der Manualeinsatz bzw. der Einsatz von Behandlungsrichtlinien als auch Adherence-Kontrollen sinnvolle Studienelemente zwecks interner Validitätssicherung darstellen. In Bezug auf die Adherence-Kontrollen kann davon ausgegangen werden, dass diese unabhängig vom Behandlungsumfang umsetzbar sind. Der Manualeinsatz hingegen stellt sich für Langzeittherapiestudien aufgrund eines Mangels an Langzeittherapiemanualen weit schwieriger dar. Neuere Studien wie die APS (Benecke et al., 2012) oder die LAC-Studie (Leuzinger-Bohleber et al., 2010) zeigen jedoch, dass sich der Einsatz von Manualen in Langzeittherapien nicht grundsätzlich verbietet. Daher muss festgehalten werden, dass die benachteiligende Wirkung des „Manual-Kriteriums" allein auf die geringe Anzahl an verfügbaren Manualen für Langzeitbehandlungen zurückzuführen ist und nicht etwa darauf, dass manualisierte Langzeitbehandlungen *per se* ein Tabu darstellen.

Grundsätzlich muss für die Anwendbarkeit des „Manual-Kriteriums" eine Definition von „Manualen" zugrunde gelegt werden, die sich nicht allein auf strikt modularisierte Schritt-für-Schritt-Handlungsanleitungen beschränkt. Eine solche Engführung des Begriffs „Manual" würde sich auf Studien zu psychodynamischen Behandlungen – ob Kurz- oder Langzeitbehandlungen – nicht anwenden lassen (vgl. Beutel et al., 2010 und Kap. 3.2.2).

In Bezug auf das „Kriterium zur Zulässigkeit nicht randomisierter Interventionen" (B.7.) konnte allenfalls eine Benachteiligung von Langzeittherapiestudien ausgemacht werden, die sich allein auf die ersten beiden Ratingstufen („1" und „2") dieses Kriteriums beschränkt. Hingegen konnte im Hinblick auf die beiden Kriterien B.4. („Kriterium zur operationalen Definition der Kontrollbedingungen") und B.5. („Kriterium zur strukturellen Äquivalenz bei

Kontrollbedingungen“) eine grundsätzlichere Benachteiligung von Langzeittherapiestudien festgestellt werden. Diese Benachteiligung wird unter Berücksichtigung des für Langzeittherapiestudien vorrangig geeigneten Untersuchungsdesigns – komparative Designs – sichtbar, in denen sich die strukturelle Angleichung zwischen den Treatmentarmen konzeptionell verbietet. Hingegen werden Gütemerkmale, die für komparative Studien ausschlaggebend sind, durch die WBP-Kriterien nur in Form eines Kriteriums (A.17: „Power-Kriterium“) abgefragt. Hier wäre es ratsam, weitere Kriterien zu entwickeln, die explizit die Bewertung interner Validitätssicherungsstrategien bei komparativen Studien zum Ziel haben.

5.3 Eine Kritische Würdigung der Verfahrensregeln und insbesondere des Kriterienkatalogs

Es wird in den nun folgenden Abschnitten eine kritische Würdigung der Verfahrensregeln und des Kriterienkatalogs vorgenommen. Dabei wird zunächst das dem Kriterienkatalog zugrunde liegende Konzept von interner und externer Validität vor dem Hintergrund der hier kodierten Studien diskutiert (Kap. 5.3.1). Danach wird sich dem Anwendungsbereich der „gemischten Störungen“ zugewandt und dessen Stellung in der wissenschaftlichen Anerkennung reflektiert (Kap. 5.3.2). Zuletzt erfolgt der Versuch einer Bewertung des Kodierprozesses im Rahmen des Promotionsprojekts sowie eine Beurteilung der Anwendbarkeit der Kriterien (Kap. 5.3.3).

5.3.1 Reflektion des Konzepts der internen und der externen Validität im WBP-Kriterienkatalog vor dem Hintergrund der kodierten Studienlage

Zu Beginn der Arbeit wurde hergeleitet, dass der WBP mit der dimensionalen Gestaltung der beiden Validitätsdimensionen dem wissenschaftlichen Standard grundsätzlich zu entsprechen scheint, demzufolge die beiden Gütemerkmale keineswegs und ausschließlich als invers, sondern eher als komplementär zu betrachten sind (vgl. Heekerens, 2005; Leichsenring, 2004a/b; Shadish et al., 2002). Das Ergebnis der einzelnen Studienkodierungen, demzufolge 14 Studien

entweder als intern *oder* als extern valide eingestuft wurden, während lediglich eine einzige Studie auf beiden Validitätsdimensionen positiv abschnitt, widerspricht hingegen dem Konzept eines komplementären Verhältnisses der beiden Validitätsarten (vgl. Abbildung 27, S. 263). Hingegen schienen die Ergebnisse der Clusteranalyse das komplementäre Verhältnis der beiden Validitätsarten auf der Basis von 36 Studien zunächst zu bestätigen: Alle drei Cluster vereinigen, wenn auch in unterschiedlichem Ausmaß, sowohl naturalistische als auch RCT-typische Eigenschaften auf sich (vgl. Kap. 4.2). Dies könnte als Hinweis darauf gewertet werden, dass Studien interne und externe Validitätsstrategien miteinander kombinieren. Die Frage, inwieweit diese Bewertung tatsächlich zutrifft und wie diese mit dem Endergebnis der Studienkodierung (Abbildung 27, S. 263) in Übereinstimmung zu bringen ist, bildet den Fokus dieses Kapitels. Es wird also ermittelt, durch was das inverse Verhältnis der beiden Validitätsarten, das sich im Gesamtbild der Studienkodierungen abzeichnet, primär hervorgerufen wird – durch die Studien selbst oder durch den Zuschnitt der Validitätsdimensionen im WBP-Kriterienkatalog.

Der Einfachheit halber werden die beiden Cluster, in denen sich überwiegend Untersuchungen zu längerfristigen Therapien befinden (Cluster 1 und Cluster 2) und die sich im Rahmen der Clustercharakterisierungen als hinreichend ähnlich entpuppten, hier zu einem Hauptcluster zusammengefasst und dem anderen Hauptcluster (dem sog. „Kurzzeittherapie-Cluster“ = Cluster 3) gegenübergestellt. Da die beiden zusammengefassten Cluster sich primär durch naturalistische Elemente hervortaten, wird in Folge vom sog. „naturalistischen Hauptcluster“ gesprochen.

Es bildete sich ab, dass keines der beiden Hauptcluster („Kurzzeittherapie-Cluster“ und „naturalistisches Hauptcluster“) in Gänze als homogen zu bezeichnen ist, was RCT-

typische oder naturalistische Eigenschaften betrifft[80]. Beim „Kurzzeittherapie-Cluster", das die meisten RCT-typischen Eigenschaften auf sich vereinigte, zeigte sich, dass einige Studien durchaus typische Aspekte naturalistischer Studien umsetzen (z.B. keine Randomisierung, keine expliziten Therapeutentrainings, keine störungsspezifischen Behandlungen, kein Ausschluss epidemiologisch relevanter komorbider Störungen; seltener: Verzicht auf Manuale). Gleichsam waren auf Seiten des „naturalistischen Hauptclusters" durchaus RCT-typische Aspekte zu entdecken (z.B. Randomisierung, Kontrollgruppendesigns, a priori festgelegte Sitzungsanzahl, Adherence-Kontrollen; seltener: Manualeinsatz). Einschränkend muss jedoch in Bezug auf die methodologische Heterogenität des „naturalistischen Hauptclusters" Folgendes berücksichtigt werden: Zum einen wurden in die Clusteranalyse auch solche Variablen aufgenommen, die zwar als RCT-*typisch* betrachtet werden können, jedoch keine unmittelbaren internen Validitätssicherungsmaßnahmen darstellen – bspw. die Variablen „a priori Festlegung der Sitzungsanzahl" oder „Störungsspezifität der Behandlung". Zwar treten diese Merkmale vermehrt im Zusammenhang mit der Variable „Manualeinsatz" auf und bilden zusammen mit dieser primär in Mehrgruppendesigns eine Strategie der internen Validitätssicherung (vgl. Kap. 5.2.2). Jedoch können sie durchaus auch genuine Merkmale einer unmanualisiert (naturalistisch) durchgeführten Behandlungsmethode darstellen, z.B. dann, wenn ein „störungsspezifisch" ausgerichtetes und zeitlich befristetes Gruppentherapieprogramm für Patienten mit sexuellen Missbrauchserfahrungen (plus Ängsten und Depressionen) in seiner Wirksamkeit beforscht wird (vgl. Studie 33). Weder von der hiesigen „Störungsspezifität" noch von der a priori festgelegten Zeitbegrenzung der Behandlung kann hier eine tatsächlich intern validitätssichernde Wirkung angenommen werden. Zum anderen stellen einige der in die Clusteranalyse aufgenommenen Variablen(-ausprägungen) zwar offenkundig Techniken

[80] Vgl. Tabelle 17, S. 238; Tabelle 18, S. 241; Tabelle 19, S. 242.

der internen Validitätssicherung dar, jedoch wurden sie in manchen Studien des „naturalistischen Hauptclusters“ nicht *lege artis* umgesetzt, so dass das validitätssichernde Potential dieser Techniken bei diesen Studien fraglich ist. Die Rede ist hier von in Untersuchungen zu längerfristigen Behandlungen herangezogenen Kontrollbedingungen, die jedoch, was den Zeitraum betrifft, nur einen Bruchteil der zu kontrollierenden psychotherapeutischen Behandlungen umfassten. Wie in Kapitel 5.1.4 ausgeführt wurde, sind diese Kontrollgruppenvergleiche allenfalls in der Lage, einen Wirksamkeits*trend* zu illustrieren. Das bedeutet, dass von ihnen vergleichsweise wenig interne Validitätskontrolle ausgeht.

Damit existieren also Studien im „naturalistischen Hauptcluster“, die zwar Elemente umsetzen, die als RCT-*typisch* betrachtet werden können, die jedoch in den eben genannten Fällen keinen unbedingten Beitrag zur internen Validitätssicherung leisten. Gleichzeitig prägen diese Studien aber eben auch das oben angenommene Bild der methodologischen Heterogenität innerhalb der Cluster mit, indem sie naturalistische mit RCT-typischen Elementen kombinieren. Diese methodologische Heterogenität muss nun also zumindest auf Seiten des „naturalistischen Hauptclusters“ relativiert werden.

Wie sieht es im „Kurzzeittherapie-Cluster“ aus? In diesem Cluster befinden sich Untersuchungen ohne Vergleichs- oder Kontrollgruppen (gemeint sich hier Studien in "echten Ein-Gruppen-Designs"), die es allein deswegen in dieses Cluster „geschafft“ haben, da in diesen Studien diverse RCT-typische Elemente umgesetzt wurden – z.B. manualisierte Behandlungen, Adherence-Kontrollen, a priori festgelegte Sitzungsumfänge, störungsspezifische Therapien und für die Studie durchgeführte Therapeutentrainings. Wie an anderer Stelle bereits ausgeführt, bedarf es zur internen Validitätssicherung in Untersuchungen ohne Kontroll-/Vergleichsgruppe jedoch anderer bzw. zusätzlicher Strategien (z.B. externe Vergleichsnorm, *coherent pattern matching*; vgl. Leichsenring, 2004a/b; Leichsenring & Rabung, 2006; Sha-

dish et al., 2002). Folglich muss im Hinblick auf diese Studien festgehalten werden, dass sie es zwar aufgrund der Umsetzung diverser RCT-typischer Strategien in besagtes Cluster „geschafft“ haben; dieser Umstand ist jedoch auch hier nicht damit gleichzusetzen, dass dadurch automatisch auch designadäquate Strategien zur internen Validitätssicherung wirkungsvoll umgesetzt wurden. Betrachtet man die Tatsache, dass in diesen Studien *per definitionem* keine Randomisierung stattgefunden hat, als naturalistisches Element, so können diese Studien dennoch nicht als Untersuchungen betrachtet werden, in denen die Kombination interner und externer Validitätsstrategien tatsächlich als geglückt betrachtet werden kann.

Insofern muss der oben geschilderte Eindruck bestehender methodologischer Heterogenität, gemäß derer in Studien regelhaft interne *und* externe Validitätselemente miteinander kombiniert werden, revidiert werden: Durch die Wahl der clusterkonstituierenden Variablen und durch die isolierte Betrachtung dieser Variablen wird die genannte methodologische Heterogenität offenkundig überschätzt. Ein realistischeres Bild der hiesigen Studienlage sowohl im Hinblick auf Strategien der internen Validitätssicherung als auch auf Bemühungen, in Untersuchungen *beide* Validitätsarten hinreichend zu berücksichtigen, erhält man erst nach Sichtung weiterer Studienmerkmale (z.B. des Studiendesigns [Ein- oder Mehrgruppendesign] oder der tatsächlichen Umsetzung einer Kontrollbedingung). Nur so kann abgewogen werden, inwiefern RCT-typische Elemente tatsächlich der internen Validität dienlich waren. Tut man dies, dann relativiert sich der anfangs vermutete, clusteranalytisch vermittelte Eindruck, im vorliegenden Datensatz befänden sich mehrerlei Studien, in denen interne Validitätssicherungstechniken unter Wahrung der externen Validität realisiert wurden.

Die allgemeine methodische Qualitätsbewertung und damit einhergehende Reduzierung des Datensatzes auf 12 methodisch adäquate Studien[81] wirkt sich folgendermaßen auf den clusteranalytischen Ursprungsdatensatz (*n*=36) aus: Aus dem „naturalistischen Hauptcluster" scheiden all diejenigen Studien aus, die zunächst vermeintlich als Untersuchungen hervortraten, in denen interne *und* externe Validitätselemente sinnvoll miteinander kombiniert wurden (s.o.). Darunter befindet sich auch die einzige manualisierte Studie in diesem Hauptcluster. Somit zeichnet sich das „naturalistische Hauptcluster" (*n*=6) nunmehr durch Studien in Ein-Gruppen-Designs (inklusive verfahrensinterne Vergleiche) aus – die einzige Ausnahme bildet die randomisierte Helsinki Studie mit einer Kombination aus verfahrensinternem und verfahrensexternem Vergleich (19).

Das „Kurzzeittherapie-Cluster" (*n*=6) setzt sich nach der methodischen Qualitätsbewertung paritätisch aus Studien in Ein-Gruppen-Designs (inklusive verfahrensinterne Vergleiche) sowie randomisierten Studien mit Kontrollgruppen bzw. verfahrensexternen Vergleichsgruppen zusammen. In all diesen Studien erfolgten die Therapien manualisiert, in den meisten wurden zudem Adherence-Kontrollen durchgeführt.

Insgesamt ist das „naturalistische Hauptcluster" im Vergleich zu seiner ursprünglichen Zusammensetzung, d.h. nach Abzug aller methodisch unzureichenden Studien, insoweit homogener geworden, als dass nahezu alle Studien ausschieden, in denen RCT-typische Eigenschaften (a priori festgelegte Sitzungsanzahl, störungsspezifische Behandlungen, Kontrollgruppendesigns, Einsatz von Manualen) realisiert wurden. Gleichzeitig blieben diesem Hauptcluster auch nach der methodischen Qualitätskontrolle noch folgende zwei Studien mit

[81] Achtung: Es wurden 15 der insgesamt 41 Studien als methodisch adäquat bewertet. Da hier jedoch nur die 36 Studien, die von der Clusteranalyse klassifiziert wurden, zugrunde gelegt werden, vermindert sich die Anzahl methodisch adäquater Studien auf 12.

internen Validitätssicherungsmaßnahme erhalten: Zum einen die bereits erwähnte, randomisierte und gleichzeitig größtenteils unmanualisierte Helsinki Studie (19) sowie die an anderer Stelle schon diskutierte, ebenfalls unmanualisierte Göttinger Psychotherapiestudie (22) im Ein-Gruppen-Design, in der eine externe Vergleichsnorm herangezogen wurde (vgl. Kap. 5.2). Letzterer hängt allerdings der entscheidende Nachteil an, dass es sich bei der externen Vergleichsnorm um keine Befunde zum natürlichen Störungsverlauf, sondern um Vergleichsdaten aus Kontrollbedingungen aus anderen RCTs handelt. Auf dieses Problem wurde in Kapitel 5.2 hinreichend hingewiesen, so dass diese Studie an dieser Stelle zwar würdigend erwähnt, jedoch nicht als Studie im Ein-Gruppen-Design mit tatsächlich gelungener interner Validitätssicherung angeführt werden soll.

Auch im „Kurzzeittherapie-Cluster“ ist nach der methodischen Qualitätsbewertung eine Homogenisierung dahingehend zu verzeichnen, dass bspw. alle Studien ausschieden, in denen die Behandlungen unmanualisiert erfolgten, zudem ein Großteil der Studien, in denen keine Adherence-Messungen vorgenommen wurden. Gleichzeitig bleiben dem Cluster, neben randomisierten, manualisierten Kontrollgruppendesigns, aber auch Studien in "echten Ein-Gruppen-Designs" erhalten, die aufgrund der *per definitionem* unterlassenen Zufallszuweisung der Patienten zu ihren Behandlungen zumindest in diesem Punkt als naturalistisch bezeichnet werden können. Diesen Studien hängt jedoch nach wie vor die oben beschriebene Schwäche an, dass sie zwar RCT-*typische* Elemente umsetzen, die jedoch vergleichsweise wenig zur internen Validitätssicherung beitragen.

Vor dem Hintergrund des eben Beschriebenen erscheint nun das Bild, das sich im Hinblick auf die Ergebnisse der kodierten Studien auf den beiden Validitätsdimensionen des Kriterienkatalogs abzeichnet, viel weniger verwunderlich (vgl. Abbildung 27, S. 263): In beiden Hauptclustern sind zusammengenommen lediglich vier Studien zu finden, die aktiv etwas für

die interne Validitätssicherung tun. Das sind aus dem „naturalistischen Hauptcluster“ die Helsinki Studie (19), aus dem „Kurzzeittherapie-Cluster“ sind dies drei randomisierte Kontrollgruppenstudien (9, 27, 34)[82].

Bezieht man nun alle als methodisch adäquat bewerteten Studien (*n*=15) mit ein, dann stehen diesen vier Studien insgesamt 11 Studien gegenüber, in denen vergleichsweise wenig für die interne Validität getan wird[83]. Der weitere Verlauf der 15 Studien im Bewertungsraster der beiden Validitätsarten ist hinlänglich bekannt (Abbildung 27, S. 263) und soll daher nicht mehr im Einzelnen nachgezeichnet werden. Im Großen und Ganzen verteilen sich die 15 Studien derart auf die positiven Ergebniskategorien der beiden Validitätsdimensionen, wie es aufgrund der letzten Abschnitte zu erwarten war. Drei der vier besagten Studien schneiden positiv auf der internen Validitätsdimension ab, das trifft auf keine der 11 Studien zu, in denen relativ wenig für die interne Validitätssicherung unternommen wurde.

Fazit

Die Frage, ob der Kriterienkatalog für das in den Studienbewertungen sich abzeichnende Verhältnis von interner und externer Validität verantwortlich ist oder ob dies den Studien selbst angelastet werden muss, ist vor dem Hintergrund der zuletzt dargelegten Sachverhalte folgendermaßen zu beantworten: Legt man die in der Arbeit eruierte Clusterstruktur der Stu-

[82] An der Anzahl von vier Studien würde sich im Übrigen auch dann nichts ändern, berücksichtigte man die Studien mit, die aus der Clusteranalyse a priori ausgeschlossen (*n*=2) bzw. durch die Clusteranalyse zu Ausreißern deklariert (*n*=3) wurden. Drei dieser fünf Studien wurden als methodisch adäquat eingestuft. Bei diesen Studien (14, 15, 30) handelt es sich ausschließlich um Untersuchungen in Ein-Gruppen-Designs bzw. um verfahrensinterne Vergleiche, in denen weder externe Vergleichsnormen oder andere designadäquate interne Validitätssicherungsstrategien herangezogen wurden, um etwaige Alternativerklärungen für den beobachteten (Prä-Post-) Effekt sinnvoll auszuschließen.

[83] Hier spielen die Studien mit verfahrensinternen Vergleichen eine Sonderrolle und müssten strenggenommen aus der hiesigen Argumentation ausgeschlossen werden.

dien zugrunde, dann enthüllt sich, wenn auch erst auf den zweiten Blick, eine Studienstichprobe, in der Studien mit zentralen naturalistischen Elementen (nicht randomisierte und nicht manualisierte Studien des „naturalistischen Hauptclusters“) zwar durchaus RCT-typische Elemente umsetzen (u.a. Kontrollbedingungen), jedoch konnte gezeigt werden, dass diese aus unterschiedlichen Gründen vergleichsweise wenig zur internen Validitätssicherung beitragen. Gleiches trifft auf nicht randomisierte Studien im „Kurzzeittherapie-Cluster“ zu, hier sind Untersuchungen in "echten Ein-Gruppen-Designs" gemeint, in denen zahlreiche RCT-typische Elemente umgesetzt werden, die jedoch auch hier relativ wenig zur internen Validitätssicherung beitragen. Hier wäre es also an den Studien bzw. den Durchführenden dieser Studien selbst, sich zur Auflage zu machen, Wirksamkeitsuntersuchungen immer unter Umsetzung adäquater interner Validitätssicherungsstrategien zu planen und zu realisieren.

Vor dem Hintergrund der rein zahlenmäßig recht dominanten Studiengruppe der "echten Ein-Gruppen-Designs", von denen kaum eine adäquate interne Validitätssicherungsstrategien umsetzt, könnte man sich beim vorliegenden Datensatz verleiten lassen, eine Ergänzung der internen Validitätskriterien um solche, die sich zur Bemessung der internen Validität bei eben diesen Studien eignen, für nahezu entbehrlich zu erachten. Löst man sich hingegen vom hiesigen Datensatz bzw. richtet seinen Blick auf die Göttinger Psychotherapiestudie (22), die zumindest einen richtungweisenden und vielversprechenden Vorschlag zur internen Validitätssicherung für Studien in Ein-Gruppen-Designs unterbreitet, und betrachtet man den Kriterienkatalog obendrein als Richtschnur, an der sich Forschende in der Planung und Realisierung von Wirksamkeitsstudien orientieren können sollen, dann erscheint eine Erweiterung der internen Validitätsdimension durchaus sinnvoll.

Es kann also festgehalten werden, dass das inverse Verhältnis der beiden Validitätsarten, das sich nach der Kodierung der Studien in Abbildung 27 (S. 263) widerspiegelt, eher durch die

Studien selbst und nicht durch den Zuschnitt der beiden Validitätsdimensionen hervorgerufen wird: Von den Studien, die auf der externen Validität gut abschneiden, lassen sich kaum solche ausmachen, die in wirkungsvoller Weise etwas zwecks interner Validitätssicherung unternehmen. Dagegen zeigt vor allem die Helsinki Studie (19), dass interne und externe Validitätselemente sinnvoll miteinander zu kombinieren sind. Zieht man zudem die an anderer Stelle schon erwähnte LAC-Studie sowie die MPS heran – beides Studien jüngeren bzw. aktuellen Datums – in denen die jeweiligen Zufallszuweisungen der Patienten[84] mit unterschiedlichen naturalistischen Elementen kombiniert werden (z.B. unmanualisierte Behandlungen in der MPS [vgl. Huber, Zimmermann et al., 2012]; Vergleich des Einflusses von Randomisierung *versus* Präferierung[85] in der LAC-Studie [vgl. Leuzinger-Bohleber et al., 2010]), dann scheint sich ein Trend abzuzeichnen, vermehrt randomisierte Studien unter expliziter Berücksichtigung externer Validitätsaspekte durchzuführen. Wünschenswert wären ebenfalls nicht randomisierte Studien, in denen sich aktiv um interne Validitätssicherungsstrategien bemüht wird. Es ist davon auszugehen, dass sich das komplementäre Verhältnis der beiden Validitätsarten im Bewertungsprofil durch den Kriterienkatalog dann auch abzeichnen würde – zumindest für Studien in Mehrgruppendesigns.

5.3.2 Der Anwendungsbereich der „gemischten Störungen“

Unter den Anwendungsbereich der gemischten Störungen wurden in dieser Arbeit solche Studien subsumiert, in denen psychotherapeutische Behandlungen an diagnoseheterogenen Stichproben untersucht wurden. Damit wurde gewissermaßen von derjenigen Definition abgewichen, die der WBP für gemischte Störungen zugrunde legt, indem er auch Gruppen von Patienten mit komplexen Störungen (im Sinne von Mehrfachdiagnosen) dazu zählt. Insofern

[84] In der MPS erfolgte die Randomisierung nur zwischen zwei der insgesamt drei Behandlungsarme.

[85] Unter „Präferierung“ ist die patientenseitige Selbstzuteilung zu einer der Behandlungen zu verstehen.

würde man zu Patienten mit komplexen Störungen z.B. auch solche mit einer Hauptdiagnose aus dem affektiven Formenkreis plus weiteren Diagnosen zählen. Von den insgesamt 10 der hier kodierten Studien zu affektiven Störungen würden vier Studien (1, 14, 26, 40) unter diesen Umständen in die Studiengruppe der gemischten Störungen gelangen, da sich in diesen Studien in mehr oder weniger hohen Anteilen (43-100%) Patienten mit komorbiden Störungen befinden (Persönlichkeitsstörungen und posttraumatische Belastungsstörungen). Ferner zeigt sich, dass lediglich in einer einzigen Studie (32) aus dem gesamten Studienpool tatsächlich alle Patienten mit komorbiden Störungen systematisch ausgeschlossen wurden. Die oftmals gehegte Kritik, in RCTs zu psychotherapeutischen Behandlungen würden vorsätzlich ausschließlich monomorbide Patientengruppen untersucht werden, kann auf Basis des hier kodierten Studienpools somit nicht bestätigt werden (vgl. auch Stirman et al., 2005). Darüber hinaus legen die hier kodierten Studien den Schluss nahe, dass eine derart strenge Auslegung der o.g. Definition von gemischten Störungen, der zufolge Studien, die sich auf Patienten mit komplexen Störungen im Sinne von Mehrfachdiagnosen beziehen, automatisch zur Studiengruppe der gemischten Störungen zählen, nur schwerlich aufrechterhalten werden kann. Es ist viel eher davon auszugehen, dass der WBP solche Studien, in denen die Hauptdiagnosen aller eingeschlossenen Patienten ein und demselben Anwendungsbereich angehören, auch eben diesem Anwendungsbereich zuordnet. Dies wird er aller Voraussicht nach auch unabhängig davon tun, wie hoch der Anteil an Patienten mit zusätzlichen, komorbiden Störungen in den Untersuchungsgruppen ist.

Folgt man, wie in dieser Arbeit geschehen, der zuletzt genannten Zuordnungsregel, so ist der Anteil mit insgesamt 31 Studien an diagnoseheterogenen Untersuchungsgruppen immer noch vergleichsweise hoch. Inwieweit es sich bei diesem Phänomen um ein primär bei Studien zu den psychodynamischen Verfahren auftauchendes handelt, kann hier nicht abschließend beantwortet werden. Dazu müsste ein Vergleich mit der Evidenz anderer Verfah-

ren angestellt werden. Dem Gutachten des WBP zur systemischen Therapie (2009a) ist allerdings zu entnehmen, dass auch für dieses Verfahren Studien zu gemischten Störungen eingereicht wurden. Gleiches trifft auf die Gesprächspsychotherapie und die für dieses Verfahren eingereichten Studien zu (vgl. Bundespsychotherapeutenkammer, 2007). In dem an anderer Stelle bereits eingeführten Überblickswerk von Kröner-Herwig (2004) über die Evidenz zur Wirksamkeit der Verhaltenstherapie ist hingegen kein Anwendungsbereich zu finden, der sich explizit auf gemischte Störungsgruppen bezieht. Zieht man das in Kapitel 1.2 angeführte Zitat aus einem Interview mit Gerd Rudolf zu den gemischten Störungen nochmals heran, dann scheint dies nahezulegen, dass Untersuchungen an diagnoseheterogenen Patientengruppen innerhalb der psychodynamischen Verfahren zumindest nicht untypisch und sogar begründet sind: „Aus psychodynamischer Sicht . . . erscheint die Heraushebung einer Hauptdiagnose und darauf zugeschnittener Behandlungen eher fraglich. [Daher] hat der Beirat jenseits der achtzehn diagnostischen Anwendungsbereiche die Kategorie „gemischte Störungen“ eingerichtet“ (Psychotherapeutenjournal, 2008, S. 115). Dieses Zitat bezieht sich noch auf eine Fassung der Verfahrensregeln (Version 2.6.; WBP 2007), nach der gemischte Störungen quasi wie ein neunzehnter Anwendungsbereich behandelt und dementsprechend gleichberechtigt zu anderen Anwendungsbereichen[86] betrachtet wurden. Dieser gleichberechtigte Status wurde spätestens mit der aktuellen Fassung der Verfahrensregeln (Version 2.8; WBP, 2010) zurückgenommen, indem ausdrücklich darauf hingewiesen wird, dass Studien zu gemischten Störungen nur nach umfassender Abwägung und im Einzelfall herangezogen und wie ein neunzehnter Anwendungsbereich behandelt werden. Die Studienlage der vorliegenden Arbeit soll zum Anlass genommen werden, die Richtigkeit der Entscheidung des WBP in Bezug auf den Umgang mit Studien zu gemischten Störungen in Frage zu stellen, denn folgende Studien

[86] Gemischte Störungen konnten nur bestimmte Anwendungsbereiche ersetzen (vgl. Kap. 1.2; WBP, 2007, S. 28).

wären somit nicht in die engere Auswahl der Begutachtung durch den WBP gekommen: Die Helsinki Studie (19) sowie die Frankfurt-Hamburg-Studie (8), die sich auf Patienten mit Angst- oder depressiven Störungen beziehen. Gleiches trifft auf Studien zu, in denen den Studienpatienten eine bestimmte Problemlage gemeinsam ist (z.B. sexuelle Missbrauchserfahrungen und damit zusammenhängende Ängste und Depressionen [Studie 33], therapieresistente, schwere, chronische Störungen [Studie 12], die sog. „komplizierte Trauer“ [Studie 31]). Studie 12, die sich auf Patienten bezieht, die bislang in ihrer Vergangenheit an allen Behandlungsangeboten scheiterten, wird etwa von Heekerens (2005) als von hoher klinischer Relevanz beschrieben, insofern er auch die Frage aufwirft: „Ist bei Evaluationsmaßnahmen für Zwecke der Praxis eine Zusammenfassung von Klienten/Patienten nach einem anderen Kriterium als dem der Störungsgleichheit denn nicht viel fruchtbarer?“ (S. 364).

An dieser Stelle muss es dem WBP überlassen werden, ob er sich im Rahmen weiterer Überarbeitungen der Verfahrensregeln berufen fühlt, den Studien zu gemischten Störungen wieder einen zentraleren Stellenwert einzuräumen, als er es derzeit tut.

5.3.3 Die Anwendbarkeit des WBP-Kriterienkatalogs

Nach einem ersten Probedurchlauf, in dem von der Projektmitarbeiterin Luisa von Hauenschild und der Verfasserin der Arbeit einige Studien unabhängig voneinander kodiert wurden, zeigte sich, dass trotz der Ankerbeispiele, mit denen die einzelnen Kriterien ausgestattet sind, einige Kriterienstufen über einen derart breiten Interpretationsspielraum verfügen, dass Kodierregeln festgelegt werden mussten, um diesen einzugrenzen (vgl. Kap. 3.2). Dazu gehören bspw. das K.O.-Kriterium zur Bemessung der in den Studien berichteten Zielerreichungs- und Veränderungsindikatoren (statistische und klinische Signifikanz, Effektstärken; Kriterium

B.12.) oder das externe Validitätskriterium C.4. zur Beurteilung der Repräsentativität der Behandlung im Hinblick auf die Versorgungspraxis im deutschen Gesundheitssystem.

Noch ein weiterer und vielleicht noch grundlegenderer Anlass machte Kodierregeln notwendig, nämlich die Tatsache, dass einige der Kriterien ohne weitere Ausdifferenzierung strenggenommen gar nicht anwendbar gewesen wären. Dazu gehören vor allem diejenigen Kriterien, deren Anwendung zunächst die Erstellung eines „störungsgruppenspezifischen Maßstabs“ nötig machte: So wäre ohne die vorherige Bestimmung eines störungsangemessenen Katamnesezeitraums für den Anwendungsbereich der affektiven Störungen das interne Validitätskriterium B.11. nicht kodierbar gewesen. Gleiches trifft auf die beiden methodischen Qualitätskriterien zur Bemessung der Höhe der Dropoutraten zu (Kriterien A.3. und A.4.). Ebenfalls nicht ohne geleistete Vorarbeiten ist das Kriterium A.8. anzuwenden, mit dem beurteilt wird, inwiefern die in den Studien verwendeten Outcomemaße psychometrischen Anforderungen genügen.

Um also eine möglichst reliable Kodierung der Studien zu ermöglichen, wurde ein ausführlicher Kodierregelkatalog (vgl. Anhang C) unentbehrlich. Damit die hier angestellten Kodierungen und darauf beruhenden Ergebnisse der Studienkodierungen zudem transparent und nachvollziehbar sind, wurden einige der Kodierregeln in dieser Arbeit ausführlich dargestellt (vgl. Kap. 3.2). Damit erweist sich der Kriterienkatalog (ohne Kodierregeln) nicht ausnahmslos als eingängige und eindeutig zu interpretierende Richtschnur für die Planung und Durchführung von Wirksamkeitsstudien, darüber hinaus ist von einer ungeschulten Kodierung von Studien abzuraten. Ferner hat es sich trotz Kodierregeln bei allen Studien, vor allem aber bei komplexeren Studien (z.B. 4, 5, 6, 21, 23), die sich über mehrere Publikationen erstrecken, als vorteilig erwiesen, diese von zwei Ratern kodieren zu lassen und ggf. auftretende Diskrepanzen zu diskutieren und auszuräumen.

5.4 Generalisierbarkeit der Befunde

In diesem Unterkapitel soll reflektiert werden, inwieweit die dargestellten und diskutierten Befunde zu den methodischen Qualitätskriterien und den internen Validitätskriterien in ihrer Auswirkung auf Langzeittherapiestudien generalisierbar sind. Es wird also der Frage nachgegangen, ob sich diejenigen Kriterien, denen auf Basis der hiesigen Studienpopulation eine benachteiligende Wirkung in Bezug auf Langzeittherapiestudien attestiert wurde, ebenso benachteiligend auf Langzeittherapiestudien aus *anderen* Anwendungsbereichen auswirken würden – z.B. bei Studien zu Persönlichkeitsstörungen oder Angststörungen. Ebenso wird der Frage nachgegangen, ob diejenigen Kriterien, denen *keine* benachteiligende Wirkung in Bezug auf Langzeittherapiestudien attestiert wurde, sich gleichermaßen nicht-benachteiligend auf Langzeittherapiestudien aus *anderen* Anwendungsbereichen auswirken würden.

Nun wurden Verallgemeinerungen im Hinblick auf vereinzelte Kriterien gewissermaßen schon im Rahmen der Feinanalysen und der Reflektion der Ergebnisse vorgenommen. Das war vor allem dann der Fall, wenn auf der Hand lag, dass eine benachteiligende Wirkung eines Kriteriums sich nicht nur auf die hier kodierten Langzeittherapiestudien beschränken ließ. So wird der Umstand, dass Kontrollgruppen, wie Wartelisten, sich schlecht über mehrere Jahre realisieren lassen, nicht nur auf die hier kodierte Langzeittherapiestudie (Helsinki Studie) zutreffen, die für die Untersuchung der internen Validitätskriterien noch zur Verfügung stand. Vielmehr wird dies auch auf andere Langzeittherapiestudien zutreffen. Dennoch soll in diesem Kapitel nochmals genauer abgeschätzt werden, welche der auf Basis des kodierten Studiendatensatzes gezogenen Schlussfolgerungen über einzelne Kriterien sich tatsächlich auf andere Anwendungsbereiche ausweiten bzw. übertragen ließen.

Im Hinblick auf die beiden methodischen Qualitätskriterien, denen feinanalytisch auf den Grund gegangen wurde – das waren das sog. „Diagnosestellung-Kriterium“ (A.2.) und das

Kriterium zur Bewertung von Veränderungs- und Zielerreichungsmaßen (B.12.) – wurde festgestellt, dass die Langzeittherapiestudien *nicht* aufgrund ihres Gegenstands "Langzeittherapie" schlechter auf diesen Kriterien abschnitten. Daraus war zu schließen, dass von beiden Kriterien in Bezug auf den hiesigen Studiendatensatz keine benachteiligende Wirkung auf Langzeittherapiestudien ausgeht. Wie plausibel ist nun die Annahme, dass diese eben nicht benachteiligende Wirkung der beiden Kriterien auch auf Langzeittherapiestudien etwa zu somatoformen Störungen oder zu Angst- und Zwangsstörungen zu übertragen ist?

In Bezug auf das „Diagnosestellung-Kriterium“ (A.2.) ist diese Annahme durchaus plausibel: Es ist nicht davon auszugehen, dass langzeittherapeutisch zu behandelnde Studienpatienten im Vergleich zu kurzzeittherapeutisch behandelten Patienten in Abhängigkeit von ihrer Störungsgruppe mal mehr und mal weniger objektiv und reliabel diagnostiziert werden können. Ferner kann zwischen dem Grund, der im hier kodierten Studiendatensatz auf Seiten der Langzeittherapiestudien zu schlechten Bewertungen auf dem „Diagnosestellung-Kriterium“ führte – der retrospektive Datenzugang nämlich – und dem Gegenstand "Langzeittherapie" auch in Bezug auf *andere* Anwendungsbereiche kein plausibler Zusammenhang ausgemacht werden.

Im Hinblick auf die Bewertung von Veränderungs- und Zielerreichungsmaßen (Kriterium B.12.) konnte ebenfalls keine tatsächlich benachteiligende Wirkung auf Langzeittherapiestudien festgestellt werden. Auch diese Schlussfolgerung sollte auf andere Anwendungsbereiche übertragbar sein, zumindest ist ein Störungsbild schwer vorstellbar, in dem Patienten in Langzeitbehandlungen partout nur zu einem Messzeitpunkt befragt werden können, was eine Veränderungs- und Zielerreichungsbestimmung in der Tat schwierig machen würde (vgl. Kap. 5.1.6).

In Bezug auf die beiden feinanalytisch untersuchten K.O.-Kriterien der allgemeinen methodischen Qualitätsdimension kann also davon ausgegangen werden, dass die darauf bezogenen und in dieser Arbeit hergeleiteten Schlussfolgerungen durchaus auch auf andere Anwendungsbereiche generalisierbar sind.

Wie steht es um die internen Validitätskriterien und die darauf bezogenen Schlüsse in ihrer Übertragung auf andere Anwendungsbereiche? Im Hinblick auf das „Manual-Kriterium" (B.3.) wurde festgestellt, dass, allein vermittelt über die verhältnismäßig geringe Anzahl verfügbarer Manuale für Langzeitbehandlungen, eine benachteiligende Wirkung durch dieses Kriterium im Hinblick auf Langzeittherapiestudien angenommen werden kann. Es wurde jedoch auch festgestellt, dass diese Benachteiligung keineswegs grundlegender Art ist, etwa in der Form, dass Manuale/Behandlungsrichtlinien und Langzeitbehandlungen ein *per se* nicht zu vereinigendes Gegensatzpaar darstellen. Und dennoch haben es Langzeittherapiestudien tendenziell schwerer, auf dem „Manual-Kriterium" positiv bewertet zu werden, als es für psychodynamische Kurzzeittherapiestudien der Fall ist, für die immerhin zahlreiche Manuale vorliegen (vgl. Beutel et al. 2010). Dieser Schluss kann zweifelsohne auch auf andere Anwendungsbereiche übertragen werden, da Manuale für Langzeittherapien grundsätzlich in absoluter Minderheit vorliegen – nicht nur für die Anwendungsbereiche der hier untersuchten affektiven und gemischten Störungen.

In Bezug auf das „Adherence-Kriterium" (B.6.) konnte gezeigt werden, dass Skalen existieren, wie die CPPS nach Hilsenroth et al. (2005) sowie das PQS Verfahren (Ablon & Jones, 1998; Jones, 2000), die die Adherence auch bei unmanualisierten Langzeittherapiestudien zu messen vermögen. Insoweit ist die Adherence-Messung nicht unmittelbar gekoppelt an den Einsatz von Manualen und es wurde geschlossen, dass nicht zuletzt aus diesem Grund keine benachteiligende Wirkung von diesem Kriterium im Hinblick auf Langzeittherapiestu-

dien ausgeht. Ferner kann nicht angenommen werden, dass sich die Adherence in Langzeittherapien weniger gut bemessen lässt als in Behandlungen von kürzerer Dauer. Könnte es jedoch sein, dass sich dies in der Behandlung anderer Störungsbilder anders darstellt? Es kann kaum davon ausgegangen werden, dass Langzeittherapien bestimmter Störungsgruppen jede Erhebung der Manual- oder Therapiekonzepttreue begründet verbieten würden.

Im Weiteren wurde das „Kriterium zur Zulässigkeit nicht randomisierter Interventionen" (B.7.) diskutiert, dem infolgedessen ebenfalls eine benachteiligende Wirkung auf Langzeittherapiestudien abgesprochen wurde. Grundsätzlich ist anzunehmen, dass dies auch für Studien gilt, die innerhalb anderer Anwendungsbereiche durchgeführt werden. Zumindest erscheint kein Grund plausibel, warum etwa in Studien zu Langzeitbehandlungen von Persönlichkeits- oder somatoformen Störungen die Dokumentation von zusätzlichem Inanspruchnahmeverhalten sowie ggf. die statistische Kontrolle desselben nicht möglich sein sollte.

Beim „Kriterium zur operationalen Definition der Kontrollbedingungen" (B.4.) sowie beim „Kriterium zur strukturellen Äquivalenz bei Kontrollbedingungen" (B.5.) schnitt die Langzeittherapiestudie im Gegensatz zu den drei Kurzzeittherapiestudien jeweils mit einem Missingwert ab. Daher war hier nicht die Frage, ob von diesen Kriterien evtl. eine benachteiligende Wirkung derart ausgeht, dass Langzeittherapiestudien systematisch „schlechter" bewertet werden. Vielmehr richtete sich die Frage auf die *prinzipielle* Bewertbarkeit von Langzeittherapiestudien. Beide Kriterien wurden als benachteiligend im Hinblick auf Langzeittherapiestudien eingestuft und zwar vor dem Hintergrund, dass in Langzeittherapiestudien keine klassischen Kontrollbedingungen, wie Wartelisten, Placebo-Kontrollen oder auch TAU, realisiert werden können, die jedoch die Voraussetzung für die Bewertbarkeit einer Studie auf den beiden Kriterien B.4. und B.5. bilden. Diese Schlussfolgerung muss jedoch weiter begründet

– und vielleicht begrenzt – werden: TAU wurde hier verstanden als das, was Buchholz (2008) drastisch mit „Behandlungen bei schlecht ausgebildeten Therapeuten/Sozialarbeitern, die eine große Zahl von Patienten zu betreuen haben“ (S. 16) beschreibt. Und de Maat, Dekker et al. (2007) ergänzen: „Without exaggeration, it may be said that TAU mostly consists of minimal therapy, hardly rising above the first three options mentioned previously [no treatment, wait list, placebo treatment]“ (S. 61). Nun werden aber bspw. in unterschiedlichen Wirksamkeitsstudien zu langfristigen Behandlungen von Patienten mit Borderline Persönlichkeitsstörungen durchaus Vergleiche mit TAU realisiert (vgl. Bateman & Fonagy, 2009; auf weitere Studien verweisen de Matt, Dekker et al., 2007). Wie ist das möglich? De Maat, Dekker et al. (2007) finden folgende Antwort:

> This design is feasible because there is an acceptable TAU for BPD [borderline personality disorder] patients, namely treatment at a psychiatric outpatient clinic. . . . To our knowledge, no RCTs compare LTP [long-term psychotherapy] and TAU in non-BPD-patients, not because there is no LTP for these patients but because there is no acceptable TAU for well-informed patients. (S. 61)

Das bedeutet, die o.g. Schlussfolgerung, die besagte, dass Langzeittherapiestudien durch die Kriterien B.4. und B.5. eine Benachteiligung im Sinne der Nicht-Bewertbarkeit erfahren, muss strenggenommen auf Studien solcher Störungsgruppen beschränkt werden, innerhalb derer keine akzeptablen TAU, sondern lediglich TAU in der Lesart von Buchholz (2008) zur Verfügung stehen. Ausgenommen sind davon Studien für die Störungsgruppe der Borderline Persönlichkeitsstörung, für die sog. *Structured Clinical Management* Programme existieren, die als TAU bezeichnet werden können. Diese fußen in der schon erwähnten Studie von Bateman und Fonagy (2009) bspw. auf einem supportiven Beratungskonzept, das *Case Management*, juristische Beratungen und problemorientierte psychotherapeutische Interventionen miteinschließt (vgl. auch DGPT, 2011). Diese Behandlung ist weit entfernt von dem, was

Buchholz unter TAU subsumiert, jedoch scheinen diese Programme, glaubt man de Maat, Dekker et al. (2007), Patienten mit Borderline Persönlichkeitsstörungen vorbehalten zu sein. Damit ist zu erwarten, dass sich die Kriterien zur „operationalen Definition der Kontrollbedingungen“ (B.4.) sowie zur „strukturellen Äquivalenz bei Kontrollbedingungen“ (B.5.) auf Langzeittherapiestudien genau dann nicht benachteiligend auswirken, wenn akzeptable Kontrollbedingungen zur Verfügung stehen. Dies kann zum jetzigen Zeitpunkt allein für die Gruppe der Borderline Persönlichkeitsstörungen angenommen werden.

Zusammenfassend kann also festgestellt werden, dass die Schlussfolgerungen zu denjenigen Kriterien der allgemeinen methodischen Qualität und der internen Validität, die feinanalytisch untersucht bzw. vor dem Hintergrund der allgemeinen Evaluationsforschung reflektiert wurden, mit einer Ausnahme durchaus störungsgruppenübergeifend Geltung beanspruchen können. Im Hinblick auf die genannte Ausnahme (der Realisierung von TAU-Bedingungen in Langzeittherapiestudien) muss davon ausgegangen werden, dass die Schlussfolgerungen gewissermaßen von temporärem Charakter sein könnten: Würde der Fall eintreten, dass künftig für mehrere Störungsbilder breitflächig sog. strukturierte Behandlungsprogramme (*Structured Clinical Management* Programme) eingeführt und diese zudem zur generellen Behandlungspraxis – TAU – würden, dann böten sich diese durchaus als Kontrollbedingungen im Rahmen von Langzeittherapiestudien an.

5.5 Limitationen

Durch das in der Arbeit gewählte zweistufige Vorgehen, in dem die Verteilungsvergleiche (Kurzzeit- *versus* Langzeittherapiestudien) auf den einzelnen Kriterien den Ausgangspunkt für die zweite Stufe, die Feinanalysen, bildeten, könnten bestimmte Kriterien in ihrer benachteiligenden Wirkung durchaus unentdeckt geblieben sein. Dies trifft jedoch in erster Linie auf

die methodische Qualitätsdimension zu, auf der allein ein Effekt $\omega \geq 0.10$ zuungunsten der Langzeittherapiestudien zum Anlass genommen wurde, eine *potentielle* Benachteiligung von Langzeittherapiestudien zu vermuten und diesem Befund im Weiteren feinanalytisch auf den Grund zu gehen. Es ist anzunehmen, dass die Verteilungen der Studien auf den Kriterien noch von ganz anderen Faktoren beeinflusst werden, als von der Behandlungslänge der Studientherapien allein – so z.B. von der Verteilung bestimmter methodologischer Präferenzen der Forschenden selbst oder auch einfach von der Verteilung der Forschungskompetenz der Studiendurchführenden. Diese Faktoren können eben auch dazu geführt haben, dass sich kein Verteilungsunterschied zwischen den Studiengruppen (Kurzzeit- *versus* Langzeittherapie) zeigt, obwohl sich bei näherem Hinsehen im Rahmen der Feinanalyse vielleicht durchaus hätte zeigen können, dass eine benachteiligende Wirkung von dem einen oder anderen Kriterium ausgeht. Eine weitere Limitation die methodische Qualitätsdimension betreffend, ist gewiss darin zu sehen, dass lediglich die K.O.-Kriterien untersucht wurden und nicht alle 19 Kriterien. Insofern besteht hier Nachholbedarf in der Untersuchung der restlichen Kriterien.

Im Hinblick auf die interne Validitätsdimension wurde insofern strenger vorgegangen, als dass sich durchweg denjenigen Kriterien detaillierter zugewandt wurde, bei denen die eine Langzeittherapiestudie mit negativem Ergebnis (Stufe „3“) abschnitt – und zwar unabhängig davon, ob dies für die Kurzzeittherapiestudien gleichermaßen galt. Damit sollte die Möglichkeit ausgelotet werden, ob einem Kriterium, obwohl die Studienverteilungen dies nicht nahelegten, dennoch eine benachteiligende Wirkung in Richtung Langzeittherapiestudien zugeschrieben werden könnte (vgl. Kap. 3.4.3). Insofern ist die Wahrscheinlichkeit, dass auf der internen Validitätsdimension Kriterien mit benachteiligender Wirkung unentdeckt blieben im Vergleich zur methodischen Qualitätsdimension als geringer einzuschätzen.

In Bezug auf die Einteilung der Studien in Langzeittherapiestudien (über 100 Sitzungen), Studien zu Behandlungen moderater Länge (über 25 bis 100 Sitzungen) sowie Kurzzeittherapiestudien (bis 25 Sitzungen) muss berücksichtigt werden, dass vor allem die Abgrenzung zwischen Langzeittherapien und Therapien moderater Länge unscharf ist: Eine 80 Stunden umfassende Behandlung mit wöchentlichen Therapiesitzungen (hier eingestuft als Therapie von moderater Länge) kann dem zeitlichen Umfang (ca. 2 Jahre) einer 160 Stunden umfassenden Langzeittherapie mit zwei Therapiesitzungen pro Woche durchaus gleichen. Diese Unschärfe ist gewissermaßen den teilweise spärlichen Informationen der Studien selbst zu verdanken, denen zwar zumeist die „Dosis“ (Sitzungsanzahl), jedoch nicht immer die Frequenz (Anzahl der Sitzungen pro Woche) zu entnehmen war. Strenggenommen müssten an dieser Stelle die Autoren der Studien kontaktiert werden, um auf Basis der Sitzungsfrequenzen die realen Therapielängen abschätzen zu können. Dadurch wäre es möglich, die o.g. Dreiteilung nicht nur auf der Basis einer Dimension („Dosis“), sondern auf der Basis zweier Dimensionen („Dosis“ und reale Therapielänge) vorzunehmen, wodurch die Dreiteilung möglicherweise an Trennschärfe gewinnen würde.

Eine weitere Limitation der Arbeit bezieht sich auf den Kodierprozess der Studien selbst: Hier wurden keine Reliabilitätsüberprüfungen in Form von Interraterkonkordanzen vorgenommen, so dass eine Aussage zur Reliabilität des im Rahmen der Arbeit entwickelten Kodierregelkatalogs nicht getroffen werden kann. Eine solche Reliabilitätsüberprüfung hat an sich *vor* der regulären Anwendung eines Messinstruments zu erfolgen, insofern besteht auch hier weiterer Forschungsbedarf.

Vor allem in Bezug auf die externe Validitätsdimension und die darauf bezogenen Kodierregeln ist durchaus in Frage zu stellen, ob die hier als Richtschnur herangezogenen Psychotherapierichtlinien (Gemeinsamer Bundesausschuss, 2013) etwa für die Beurteilung,

inwieweit psychodynamische Studienbehandlungen hinsichtlich ihrer Dauer die klinische Praxis im deutschen Gesundheitssystem repräsentieren, tatsächlich angemessen sind. Auch die Frage danach, ob bestimmte, durch die Studien nahegelegte Selektionseffekte, z.B. durch die Art der Rekrutierung der Stichprobe, sich tatsächlich mindernd auf die externe Validität ausgewirkt haben, kann hier nicht abschließend beantwortet werden (erinnere: „Es geht bei der Frage der Übertragbarkeit nicht darum, ob Patienten in Studien „anders“ sind als Patienten in der späteren Praxis (das ist sicher), sondern es geht darum, ob die Therapieeffekte anders sind“ [Windeler, 2008, S. 255]). Demzufolge wäre eine systematische Untersuchung der externen Validitätskriterien sowie der dazugehörigen Kodierregeln wünschenswert.

6 Zusammenfassung und Ausblick

In dieser Arbeit wurde der Frage nachgegangen, inwieweit der Kriterienkatalog des Methodenpapiers (Version 2.8; WBP, 2010) dem Gegenstand psychodynamischer Langzeittherapie und seiner Beforschung im Hinblick auf seine Wirksamkeit gerecht wird. Diese Frage gewann u.a. dadurch an Relevanz, als der WBP in einer 2008 veröffentlichten Ergänzung der Stellungnahme zur psychodynamischen Psychotherapie den Geltungsbereich der wissenschaftlichen Anerkennung, bis dato begrenzt auf psychodynamische Psychotherapie vom Umfang von bis zu 100 Sitzungen, auf Langzeitbehandlungen (> 100 Sitzungen) ausweitet. Begründet wird diese Ausweitung mit der Veröffentlichung der neuen Verfahrensregeln (Version 2.6; 2007), die, so der WBP (2008a), nicht mehr dazu berechtigten, die wissenschaftliche Anerkennung auf psychodynamische Behandlungen von bestimmter Dauer zu beschränken. Zurück geht diese Entscheidung auf die im Methodenpapier vorgenommene Trennung von psychotherapeutischen *Methoden* und *Verfahren* und der damit zusammenhängenden Regelung, der zufolge die wissenschaftliche Anerkennung von psychotherapeutischen Verfahren keineswegs über die flächendeckende Anerkennung aller dem Verfahren zuordenbaren Methoden zu erfolgen hat. Indem der WBP diese Regelung auch auf unterschiedliche Variationen von Behandlungslängen innerhalb eines Verfahrens überträgt, schafft er die Voraussetzung, die wissenschaftliche Anerkennung der psychodynamischen Verfahren nunmehr auf psychodynamische Langzeittherapien ausweiten zu können. In der vorliegenden Arbeit wird dieser Schritt, zusammen mit dem 2005 vom WBP geäußerten Hinweis, bei Langzeitbehandlungen (> 100 Stunden) handele es sich um einen Gegenstand, der „besondere Forschungsfragen“ (S. 74) aufwerfe, zum Anlass genommen, der Frage nachzugehen, inwieweit der Kriterienkatalog diesen besonderen Forschungsfragen eigentlich gerecht wird. Diese Frage wurde umso virulenter, als mit den neuen Verfahrensregeln des WBP und vor allem mit dem darin enthaltenen Kriterienkatalog erstmals ein Regelwerk vorgelegt wurde, das sich aktiv um eine psy-

chotherapieangemessene Begutachtung der empirischen Evidenz bemüht. Dies erreichte der WBP durch die dimensionale Gestaltung des Kriterienkatalogs, basierend auf den drei Dimensionen der allgemeinen methodischen Qualität, der internen und der externen Validität (vgl. Kap. 1.2.2).

Die Untersuchung ist damit im größeren Themenbereich der Evidenzbasierung in der Psychotherapie anzusiedeln, dem seit den 90er Jahren zunächst im Zuge der EST-Aktivitäten im angloamerikanischen Sprachraum und spätestens mit Inkrafttreten des Psychotherapeutengesetzes (1999) auch im deutschsprachigen Raum vermehrte Aufmerksamkeit zuteilwurde. Im Gegensatz zu den bislang eher theoretisch geführten Debatten über die gegenstandsadäquate Beforschung und Evidenzbasierung von Psychotherapie, wurde in dieser Arbeit die empirische Evidenz in Form von Wirksamkeitsstudien systematisch hinzugezogen.

Der Arbeit wurde folgender Standpunkt zugrunde gelegt, der gewissermaßen einen Ausgangspunkt der Untersuchung bildet: Die beiden Gütemerkmale der internen und der externen Validität bilden kein Gegensatzpaar, in dem sich die beiden Validitätsarten ausschließlich invers zueinander verhalten (vgl. Heekerens, 2005). Vielmehr kann angenommen werden, dass in Untersuchungen, in denen bspw. aus Gründen der externen Validität von bestimmten internen Validitätssicherungsstrategien, wie Randomisierung, Manualisierung oder Kontrollgruppen (Wartelisten, Placebo etc.) abgesehen werden muss, die interne Validität durch alternative Strategien gesichert werden kann (vgl. Leichsenring, 2004a/b; Shadish et al., 2002). Dies trifft gleichsam auf die in dieser Arbeit fokussierte Studiengruppe der Langzeittherapiestudien zu, bei denen – so legten es zahlreiche kritische Stimmen nahe (vgl. Kap. 1.2.1) – ebenfalls davon auszugehen war, dass typische RCT-Strategien, wie der Manualeinsatz, randomisierte Patientenzuweisung und klassische Kontrollgruppen sich nur schwerlich umsetzen lassen. Insoweit stellten Langzeittherapiestudien einen willkommenen Gegenstand, man

könnte auch sagen, einen gelungenen Prüfstein dar, an dem sich bemessen ließ, inwieweit insbesondere die interne Validitätsdimension des WBP-Kriterienkatalogs der reinen RCT-Methodologie entwachsen ist. Es war zu erwarten, dass eine sich zu eng an die RCT-Methodologie anlehnende Begutachtungspraxis und zudem eine Begutachtungspraxis, die alternative interne Validitätssicherungsstrategien (vgl. Leichsenring, 2004a/b) unberücksichtigt lässt, einen systematischen Bias im Hinblick auf Langzeittherapiestudien erzeugen würde – resultierend aus einer benachteiligenden Bewertung der Qualität bzw. der internen Validität eben dieser Studien. In diesem Fall wäre die Gegenstandsadäquatheit der internen Validitätskriterien in Bezug auf Langzeittherapiestudien als eingeschränkt zu bezeichnen.

Als Datengrundlage für die Untersuchung wurden Wirksamkeitsstudien zur psychodynamischen Psychotherapie aus den Anwendungsbereichen der affektiven Störungen sowie der „gemischten Störungen“, hier als diagnoseheterogene Störungsgruppen verstanden, gewählt. Um der Untersuchung einen Studiendatensatz zugrunde zu legen, der beide Anwendungsbereiche für den Publikationszeitraum 1999-2009 gut repräsentiert, wurde eine erschöpfende Studienrecherche durchgeführt und eine Vollerhebung angestrebt (vgl. Kap. 3.1). Das bedeutet, es sollten möglichst alle in diesem Zeitraum in englischer oder deutscher Sprache publizierten Wirksamkeitsstudien zur psychodynamischen Psychotherapie zu den genannten Anwendungsbereichen (ausschließlich Erwachsenenalter) miteinbezogen werden. Mittels intensiver Handsuche sowie digitaler Recherche in den Online-Datenbanken PsycINFO, PSYNDEX und PubMed sowie in der Cochrane Library konnten insgesamt 41 Studien eruiert werden, die den in Kapitel 3.1 näher beschriebenen Ein- und Ausschlusskriterien genügten. Anhand einer nachträglichen Evaluierung der erfolgten Recherche, konnte gezeigt werden, dass der zusammengestellte Primärstudiendatensatz als Vollerhebung der zum Zeitpunkt der Recherche technisch zugänglichen Studien betrachtet werden kann (vgl. Kap. 3.1.3).

Zur Beantwortung der o.g. Fragestellung wurden alle 41 Studien mit Hilfe des WBP-Kriterienkatalogs von der Verfasserin der Arbeit und einem weiteren Rater kodiert. Insgesamt waren inklusive der Verfasserin der Arbeit fünf Rater an den Kodierungen der Studien beteiligt. Zum Zwecke der Kodierung und zur Erhöhung der Reliabilität der Kodierungen wurden vorher sog. Kodierregeln erstellt, mittels derer die einzelnen Kriterienstufen (Stufen „1“ bis „3“) genauer definiert wurden (vgl. Anhang C). Zwar bietet der WBP-Kriterienkatalog bereits stufenweise Erklärungen und Ankerbeispiele an, jedoch erwiesen sich diese in einigen Fällen als unzureichend für eine reliable Kodierung. So ließen die Ankerbeispiele bzw. Operationalisierungen einiger Kriterien(-stufen) einen noch zu großen Interpretationsspielraum in ihrer Anwendung auf Studien zu; ferner erwiesen sich einige Kriterien ohne weitere Ausdifferenzierung als gar nicht anwendbar (vgl. Kap. 3.2.2 und 3.2.3). Letzteres traf vor allem auf die methodischen Qualitätskriterien zur Bemessung der Dropoutquoten (Kriterien A.3. und A.4.) zu, außerdem auf das Qualitätskriterium zur Bemessung der Reliabilität und Validität der in den Studien verwendeten primären Outcomemaße (Kriterium A.8.) sowie auf Kriterium A.9., mit dem bewertet wird, ob in einer Studie etwas zur klinischen Bedeutsamkeit der patientenseitigen Veränderungen berichtet wird. Auf der internen Validitätsdimension musste das Kriterium B.11. weiter ausdifferenziert werden, indem zunächst der sog. „störungsangemessene Katamnesezeitraum“ für affektive Störungen eruiert wurde; gleiches galt für die Stufen des Kriteriums B.12. zur Bemessung der berichteten Veränderungs- und Zielerreichungsmaße, die ebenfalls einer genaueren Operationalisierung bedurften.

Um Aufschluss über allgemeine und methodologische Merkmale der Studien zu erhalten, wurde darüber hinaus ein Kurzkodierbogen entwickelt, mit dem ebenfalls alle 41 Studien kodiert wurden (vgl. Kap. 3.2.4; Anhang B und D).

Die Ergebnisse der Arbeit lassen sich in drei größere Blöcke aufteilen: Der erste Block umfasst einen allgemeinen Überblick über die Studien, der auf den mittels Kurzkodierbogen erhobenen Informationen basiert. Die im Kurzkodierbogen enthaltenen Variablen gingen auch in die sich anschließende Clusteranalyse ein. Der zweite Block bezieht sich auf die Ergebnisse der Studienkodierungen mit Hilfe des WBP-Kriterienkatalogs. Hier konnte anhand eines Flowdiagramms (Abbildung 27, S. 263) der Verlauf (der „Flow“) der Studien durch das Raster der Bewertungen auf den drei Dimensionen des Kriterienkatalogs nachvollzogen werden. Zudem wurden die ausschlussstärksten K.O.-Kriterien der allgemeinen methodischen Qualitätsdimension identifiziert. Der dritte Block umfasst schließlich die Untersuchung der Gegenstandsadäquatheit der internen Validitätsdimension sowie einzelner Kriterien der methodischen Qualitätsdimension. In den folgenden Abschnitten sollen die zentralen Ergebnisse samt den daraus gezogenen Schlussfolgerungen nochmals zusammengefasst werden.

Die Ermittlung der Clusterstruktur, wurde mittels solcher clusterkonstituierenden Variablen vorgenommen, die, je nach Ausprägung, eher typisch naturalistische Elemente oder aber typische RCT-Elemente einer Studie darstellen (vgl. Kap. 4.2). Diese Variablen entstammten dem Kurzkodierbogen und umfassten Merkmale, wie den „Therapieumfang“, den „Strategien der Gruppenzuweisung“ (z.B. randomisiert), der „a priori Festlegung der Therapiesitzungsanzahl“, der „Störungsspezifität der Behandlung“, den „Manualeinsatz“ und weitere. Mittels einer Two-Step Clusteranalyse wurde zunächst eruiert, durch wie viele Cluster sich die Studien bestmöglich repräsentieren lassen. Im Anschluss wurde das Ergebnis der Two-Step Clusteranalyse mittels einer hierarchischen Clusteranalyse nach der *Ward*-Methode evaluiert. Hierbei diente die Clusteranzahl der Two-Step Clusteranalyse als Vorgabe für das hierarchische Verfahren. Ziel der Clusteranalyse war es, mögliche Studientypen ausfindig zu machen, die sich hinsichtlich „RCT-typischer“ und „naturalistischer“ Merkmale charakterisieren las-

sen. Insgesamt konnten drei Cluster ermittelt werden, die hauptsächlich durch die Variable „Therapieumfang“ (Kurzzeittherapie [bis 25 Sitzungen], Therapien moderater Länge [über 25 bis 100 Sitzungen], Langzeittherapie [über 100 Sitzungen]) konstituiert wurden. Von den drei Clustern ähnelten sich vor allem zwei Cluster, in denen primär Studien zu Behandlungen mittlerer Dauer sowie Langzeittherapiestudien enthalten waren. Beide Cluster zeichneten sich durch vergleichsweise wenige RCT-typische, sondern eher durch naturalistische Merkmale aus. Das dritte Cluster setzte sich ausschließlich aus Kurzzeittherapiestudien zusammen und vereinigte eher RCT-typische Merkmale auf sich. Jedoch trat keines der Cluster durch eine vollkommene Homogenität, was genannte Merkmale betrifft, hervor. Vielmehr beinhaltete das „RCT-typische Cluster“ ebenfalls Studien, die naturalistische Elemente enthielten (z.B. keine randomisierte Patientenzuweisung), und die beiden eher „naturalistischen Cluster“ enthielten ebenfalls Studien, in denen etwa RCT-typische Aspekte umgesetzt wurden (z.B. störungsspezifische Behandlungen oder a priori festgelegte Sitzungsumfänge).

Die eruierte Clusterstruktur, zusammengenommen mit den Studienkodierungen via WBP-Kriterien, konnte ferner Aufschluss darüber geben, inwiefern sich das komplementäre Konzept von interner und externer Validität (im Gegensatz zum Konzept vom inversen Verhältnis der beiden Validitätsarten) in den Studien niederschlägt. Die beschriebene methodologische Heterogenität innerhalb der Cluster schien zunächst im Gegensatz zu dem Bild zu stehen, das sich aufgrund der Studienkodierungen via WBP-Kriterien ergab: Hier schnitt allein eine einzige Studie (die sog. Helsinki Studie; Studie 19) mit positivem Ergebnis auf der internen *und* der externen Validität ab. Demgegenüber legte die Heterogenität innerhalb der Cluster auf den ersten Blick nahe, dass vermehrt Studien darum bemüht sind, beide Validitätsarten auf sich zu vereinigen. Dieser Eindruck musste jedoch revidiert werden (vgl. Kap. 5.3.1): Die Wahl der clusterkonstituierenden Variablen sowie die isolierte Betrachtung dieser Variablen ohne detaillierten Rückbezug auf die Studien selbst, führte zu einer Überschätzung der me-

thodologischen Heterogenität innerhalb der Cluster bzw. zu einer Überschätzung dessen, dass innerhalb der Studien tatsächlich interne und externe Validitätsmerkmale wirkungsvoll miteinander kombiniert werden. Zieht man nämlich weitere Studiencharakteristika hinzu und erhält so einen Eindruck davon, welche Studien in welcher Form RCT-typische Elemente realisiert haben, dann zeigt sich: In den zwei naturalistischen Clustern setzen zwar durchaus einige Studien diverse RCT-typische Elemente um, jedoch leisten diese im Rahmen der jeweiligen Studien keinen unbedingten Beitrag zur internen Validitätssicherung. Gleiches gilt für Studien aus dem Cluster, das vermehrt RCT-typische Elemente auf sich vereinigte. Hier fanden sich bspw. Studien in Ein-Gruppen-Designs, die es zwar durch Umsetzung zahlreicher RCT-typischer Strategien in dieses Cluster „schafften", jedoch ist die Umsetzung dieser Strategien nicht damit zu verwechseln, dass tatsächlich wirkungsvoll etwas für die interne Validität in diesen Studiendesigns getan wurde.

Vor dem beschriebenen Hintergrund war es, nach weiterer Reduzierung des Datensatzes durch die methodische Qualitätsbewertung (Kriterienkatalog), wenig überraschend, dass lediglich einer einzigen Studie auf Basis der Kodierung der internen und externen Validität attestiert wurde, beide Validitätsarten auf sich vereinigt und das oft proklamierte inverse Verhältnis der beiden Validitätsarten unterlaufen zu haben. Insgesamt verfügt der hier kodierte Studiendatensatz kaum über Untersuchungen, die dieser Studie (Helsinki Studie; Studie 19) in diesem Bestreben ähneln. Auch naturalistische Studien (ohne Randomisierung und Manualisierung), in denen etwa durch Techniken, wie der Vorhersage komplexer Ergebnismuster (*coherent pattern matching*; Leichsenring, 2004a/b; Shadish et al., 2002), der Hinzuziehung natürlicher Störungsverlaufsdaten (de Maat, Dekker et al., 2007) oder durch Adherence-Kontrollen *trotz* unmanualisierter Behandlungen, alternative Strategien der internen Validitätssicherung umgesetzt werden, bilden nahezu eine Leerstelle im hiesigen Studiendatensatz. Allein die Göttinger Psychotherapiestudie (22), eine naturalistische Studie im Ein-Gruppen-

Design, zieht explizit externe Vergleichsdaten heran, die als Kontrollbedingung betrachtet werden könnten. Die herangezogene Kontrollbedingung – aggregierte Daten aus Wartlisten- und TAU-Bedingungen ehemals durchgeführter RCTs – ist jedoch mit Problemen behaftet (vgl. Kap. 5.2), demzufolge die Studie allenfalls gelungene Hinweise liefert, wie interne Validitätssicherungsmaßnahmen in Ein-Gruppen-Design-Studien künftig umgesetzt werden könnten (vgl. Leichsenring und Rabung, 2006).

Es war daher zu schließen, dass im Hinblick auf den hier zugrunde gelegten Studiendatensatz das inverse Verhältnis der beiden Validitätsarten, das sich in den Kodierungsergebnissen der Studien widerspiegelt (vgl. Abbildung 27, S. 263), eher durch die Studien selbst aufrechterhalten, denn durch den WBP-Kriterienkatalog tradiert wird. Gleichzeitig war jedoch durch einen Blick in jüngere Studien aus dem deutschen Sprachraum, etwa der LAC-Studie (Leuzinger-Bohleber et al., 2010) oder der MPS (Huber, Zimmermann et al., 2012), ein Trend zu verzeichnen, demzufolge vermehrt randomisierte Studien unter expliziter Berücksichtigung der externen Validität durchgeführt werden. Wünschenswert wären ferner naturalistische Studien (ohne Randomisierung und Manualisierung), in denen sich aktiv um interne Validitätssicherungsstrategien bemüht wird.

Der zweite große Ergebnisblock bezog sich auf den Verlauf (den „Flow“) der 41 Studien durch das Bewertungsraster des Kriterienkatalogs (vgl. Kap. 4.3, insbesondere Abbildung 27, S. 263). Auf ein wichtiges Teilergebnis des Bewertungsprozederes wurde im letzten Abschnitt bereits durch den Hinweis auf die einzige Studie (19) vorgegriffen, die mit gutem Ergebnis sowohl auf der internen als auch auf der externen Validität abgeschnitten hat. Zunächst zeigte sich jedoch im Rahmen der methodischen Qualitätsbewertung der Studien, die die Voraussetzung für eine positive Bewertung auf den beiden Validitätsdimensionen bildet, eine vergleichsweise hohe Anzahl an Studien mit negativem Ergebnis auf dieser Dimension: Ins-

gesamt 26 (63.4%) der 41 Studien fielen durch dieses Bewertungsraster durch. Die Vermutung, die methodische Qualitätsbewertung via WBP-Kriterien könnte einen allzu hohen Maßstab an die Studien anlegen, konnte zum einen dadurch relativiert werden, dass nochmals explizit auf das sehr weitmaschige Netz an Einschlusskriterien hingewiesen wurde, das in der Studienrecherche Anwendung fand. Somit konnten Untersuchungen in den Studienpool gelangen, die bspw. in keine der gesichteten Metaanalysen eingingen, in denen die zu integrierenden Studien in der Regel einer Qualitätskontrolle unterzogen werden. Zum anderen konnte eine übermäßige Strenge der methodischen Qualitätsbewertung via WBP-Kriterien durch eine eingehende Analyse der Gründe der Negativbewertungen entkräftet werden. Vor allem zwei K.O.-Kriterien dieser Dimension taten sich als besonders ausschlussstark hervor: Zum einen das sog. „Diagnosestellung-Kriterium“ (A.2.) und zum anderen das Kriterium, mit dessen Hilfe bewertet wird, inwieweit die Wirksamkeit der therapeutischen Maßnahmen sowohl über Maße der Veränderung als auch der Zielerreichung abgebildet wurde (Kriterium B.12.). In Bezug auf das ausschlussstärkste K.O.-Kriterium B.12 wurde gezeigt, dass eine optimale Bewertung (Stufe „1“) zwar einen durchaus hohen, jedoch keineswegs unerreichbaren Anspruch an Studien stellt. Ferner konnte gezeigt werden, dass es Untersuchungen gelingen kann, dieses Kriterium mühelos mit zumindest zufriedenstellendem Rating (Stufe „2“) zu bestehen und somit nicht aus der weiteren Bewertung auszuscheiden. Stufe „2“ fordert etwas, das genauso auch von gängigen Publikationsrichtlinien (APA, 2010; DGPs, 2007) gefordert wird und entspricht gewissermaßen dem regulären wissenschaftlichen Standard. Allein für Untersuchungen im Ein-Messzeitpunkt-Design (nur Post- oder Katamnesemesung) sowie für Untersuchungen im „halben Kontrollgruppendesign“ (Kontrollbedingungen erfassen nur einen Bruchteil des zeitlichen Umfangs der Therapiebedingung) war die Erfüllung dieses Kriteriums nicht möglich. Bei diesen Studien handelte es sich durchgängig um Untersuchungen an längerfristigen Behandlungen (> 25 Sitzungen). Die Aussagekraft beider Studiendesigns wurde im Hin-

blick auf Wirksamkeitsaussagen kritisch reflektiert. Dadurch konnte, u.a. durch Hinzuziehung und Sichtung einiger jüngerer Outcomestudien, wie der APS (Benecke et al., 2012), der MPS (Huber, Zimmermann et al., 2012) oder der LAC-Studie (Beutel et al., 2012; Leuzinger-Bohleber et al., 2010), aufgezeigt werden, dass Untersuchungen zu länger- bis langfristigen Therapien in aussagekräftigen Mehrgruppendesigns – ohne Kontrollgruppen – durchaus realisierbar sind. Insofern konnte der Verdacht ausgeräumt werden, das K.O.-Kriterium B.12. lege einen zu hohen Maßstab an die Studien an und würde sich darüber hinaus auf Studien zu längerfristigen Behandlungen benachteiligend auswirken.

Auch in Bezug auf das K.O.-Kriterium A.2. („Diagnosestellung-Kriterium“) mit den zweithöchsten Ausschlussquoten konnte nach intensiver Sichtung der Studien mit negativen Ratings auf diesem Kriterium der Verdacht ausgeräumt werden, dass dieses Kriterium inadäquate Forderungen stellt. Im Zusammenhang mit dem „Diagnosestellung-Kriterium“ und einem möglichen Abweichen von gängigen Klassifikationssystemen, wie ICD oder DSM, wurde auch das sog. „Krankheitswert-Kriterium“ (C.1.) eingehend diskutiert. Es wurde aufgezeigt, dass sich in diesem Kriterium (C.1.) der Versorgungsauftrag psychologischer und ärztlicher Psychotherapeuten widerspiegelt, der die Behandlung von Klienten mit *krankheitswertigen* psychischen Störungen vorschreibt. Der Krankheitswert einer psychischen Störung wurde in dieser Arbeit in Anlehnung an den WBP (2008c) mit ICD- oder DSM-Diagnosen gleichgesetzt. Insofern sind das „Krankheitswert-Kriterium“ (C.1.) und auch das „Diagnosestellung-Kriterium“ (A.2.) (beides K.O.-Kriterien der methodischen Qualitätsdimension), über ihre jeweiligen methodischen Implikationen hinaus, auch als Produkte der gesundheitspolitischen bzw. gesetzlichen Gegebenheiten zu betrachten. Daraus wurde gefolgert, dass Kritik an der „Philosophie“ dieser beiden Kriterien daher auf einer anderen Ebene zu diskutieren wäre, als auf der Ebene des Methodenpapiers.

Die Kodierung der 15 methodisch adäquaten Studien auf der internen Validitätsdimension war allein für vier Untersuchungen zielführend, da es sich bei den restlichen 11 Studien um Untersuchungen im Ein-Gruppen-Design bzw. um verfahrensinterne Vergleichsstudien handelte, die in der Kodierung jedoch wie Ein-Gruppen-Designs behandelt werden (vgl. Kap. 3.2.1). In der ausführlichen Vorstellung der internen Validitätskriterien (vgl. Kap. 1.2.2) wurde dargelegt, dass sich die Kriterien der internen Validität vorrangig zur Kodierung von Mehrgruppendesignstudien eignen. Diese Regelung, so wurde es in dieser Arbeit herausgearbeitet, spielt der Auffassung eines inversen Verhältnisses von interner und externer Validität gewissermaßen direkt die Hände: Vor dem Hintergrund, dass der WBP Studien in Ein-Gruppen-Designs grundsätzlich zwecks systematischer Begutachtung zulässt, bleibt es unverständlich, warum er von diesen Studien nicht explizit mehr fordert – nämlich die Umsetzung interner Validitätssicherungsstrategien, die für eben dieses Studiendesign adäquat und möglich sind (vgl. Leichsenring, 2004a/b; Shadish et al., 2002). So könnten Kriterien, mit denen Techniken gefordert werden, wie die Vorhersage komplexer Ergebnismuster (*coherent pattern matching*) oder aber der Vergleich mit natürlichen Störungsverlaufsdaten, sich durchaus positiv auf künftige Studien ohne Kontroll-/Vergleichsgruppe auswirken und würden den wissenschaftlichen Status dieser Studien stärken. Für die wissenschaftliche Anerkennung pro Anwendungsbereich könnte die bisherige Forderung im Hinblick auf extern valide Studien demzufolge revidiert werden, indem naturalistische Studien (im Ein-Gruppen-Design) nur noch dann als ausschlaggebend betrachtet werden, wenn auch Strategien zur internen Validitätssicherung sinnvoll umgesetzt wurden.

Der dritte Block der im Rahmen der Arbeit präsentierten Ergebnisse bezieht sich auf das Herzstück der Arbeit, die Untersuchung der Gegenstandsadäquatheit der internen Validitätskriterien sowie der allgemeinen methodischen Qualitätskriterien. Letztere wurde als Teil der

internen Validitätsbewertung betrachtet, da eine hinreichend gute methodische Studienqualität eine Voraussetzung für die Validitätsbewertungen darstellt. Insoweit spielt die Bewertung der methodischen Qualität für die interne Validitätsbewertung eine zentrale Rolle, jedoch wurde sich hier nur auf die K.O.-Kriterien konzentriert.

Die Untersuchung der Gegenstandsadäquatheit der Kriterien erfolgte mittels systematischer Verteilungsvergleiche von Langzeit- und Kurzzeittherapiestudien. Verglichen wurde jeweils, mit welchen prozentualen Anteilen die beiden Studiengruppen auf den dichotomisierten Stufen der einzelnen Kriterien abgeschnitten haben. Dabei wurde in Anlehnung an die im ersten Teil der Arbeit vorgestellte RCT-Methodologie und deren Inkompatibilität mit dem Gegenstand "Langzeittherapie" (vgl. Kap. 1.2.1) davon ausgegangen, dass Studien zu Therapien von kürzerer Dauer eher den Anforderungen der RCT-Methodologie zu entsprechen vermögen, als dies auf Langzeittherapiestudien (> 100 Stunden) zutrifft. Dementsprechend war zu erwarten, dass eine ggf. zu eng an die RCT-Methodologie angelehnte Begutachtungspraxis einen systematischen Bias im Hinblick auf Langzeittherapiestudien erzeugen würde, der in einer benachteiligenden Bewertung dieser Studien bestünde. Die Verteilungsvergleiche fungierten jedoch nur als ein erster Schritt im Aufspüren *tatsächlich* benachteiligend wirkender Kriterien, denn diese Vergleiche konnten allenfalls Hinweise auf *potentielle* Benachteiligungen liefern: Wenn sich empirisch, also auf Basis der Verteilungsvergleiche, etwa auf einem Kriterium zeigte, dass anteilig mehr Langzeit- als Kurzzeittherapiestudien schlecht abgeschnitten haben, so ließ sich aufgrund dieses Verteilungsunterschieds noch keine Aussage darüber treffen, ob die Gründe für diesen Verteilungsunterschied tatsächlich mit dem Gegenstand "Therapieumfang" zusammenhängen. Eben diesem Zusammenhang auf die Spur zu kommen war Ziel des zweiten Schrittes, der in einer Feinanalyse bestand. In der Feinanalyse wurden die Studien nochmals unter besonderer Berücksichtigung eben dieses möglichen Zusammenhangs zwischen Therapiedauer und Ergebnis eingehend gesichtet.

Für die empirischen Verteilungsvergleiche auf den K.O.-Kriterien der methodischen Qualitätsdimension wurden zunächst alle 41 Studien, separiert in 24 Studien zu kürzeren Behandlungsdauern (< 100 Stunden) und 17 Langzeittherapiestudien (> 100 Stunden), integriert. Daran schloss sich ein Extremgruppenvergleich an, in dem den 17 Langzeittherapiestudien nur noch die „echten“ Kurzzeittherapiestudien bis zu 25 Sitzungen (n=15) gegenübergestellt wurden. Der Extremgruppenvergleich ging auf das Postulat zurück, demzufolge die RCT-Methodologie und die damit zusammenhängenden internen Validitätssicherungsstrategien strenggenommen als ausschließlich mit dem Gegenstand "Kurzzeittherapie" (< 25 Sitzungen) kompatibel erachtet wurden (Leichsenring, 2011; Westen et al., 2004). Mit diesem Extremgruppenvergleich wurde noch einmal strenger untersucht, ob die Kriterien unter Betrachtung der beiden Extremgruppen zu potentiellen Benachteiligungen der Langzeittherapiestudien führen. Für alle Verteilungsvergleiche wurde ein Cutoff-Wert von $\omega \geq 0.10$ angelegt (vgl. Eid et al., 2010). Das bedeutet, Verteilungsvergleiche, in denen anteilig mehr Langzeittherapiestudien mit negativem Kriteriumsergebnis sichtbar wurden und sich dieses Gefälle in einem Effekt von $\omega \geq 0.10$ ausdrückte, wurden infolgedessen feinanalytisch weiteruntersucht.

Insgesamt wurden für zwei K.O.-Kriterien der methodischen Qualitätsdimension Feinanalysen notwendig. Dabei handelte es sich wiederholt um das „Diagnosestellung-Kriterium“ (A.2.) sowie das Kriterien B.12. zur Bemessung der patientenseitigen Veränderungen und Zielerreichungen. Beide Kriterien fielen bereits als die beiden ausschlussstärksten K.O.-Kriterien der methodischen Qualitätsdimension auf (s.o.). Im Rahmen der Feinanalysen, in denen vor allem die Langzeittherapiestudien mit negativen Ergebnissen auf besagten Kriterien nochmals intensiv gesichtet wurden, konnten keine zwingenden Zusammenhänge zwischen Kriteriumsergebnis und Therapielänge identifiziert werden. Somit schnitten die Langzeittherapiestudien aus Gründen auf den beiden Kriterien schlecht ab, die nicht in einen plausiblen Zusammenhang mit dem Gegenstand "Langzeittherapie" gebracht werden konnten,

weshalb eine *tatsächliche* Benachteiligung von Langzeittherapiestudien durch diese beiden Kriterien ausgeschlossen wurde.

In Bezug auf die interne Validitätsdimension musste das Vorgehen aufgrund der geringen Studienanzahl (n=4), die für die geplanten Verteilungsvergleiche noch übrig blieben (s.o.), leicht modifiziert werden. So wurden keine Effektstärken mehr berechnet und auch die Extremgruppenvergleiche wurden hinfällig, da sich die vier Studien allein aus drei Kurzzeittherapiestudien (< 25 Stunden) und einer Langzeittherapiestudie (> 100 Stunden) zusammensetzten. Zudem wurden nicht nur solche Kriterien der internen Validitätsdimension eingehender untersucht, bei denen sich ein Bewertungsgefälle zuungunsten der Langzeittherapie zeigte. Vielmehr wurden auch diejenigen Kriterien näher in Augenschein genommen, bei denen die Langzeittherapiestudie negativ abschnitt – unabhängig vom Ergebnis der Kurzzeittherapiestudien. Damit wurde das Ziel verfolgt, möglicherweise verdeckte Zusammenhänge nicht zu übersehen. Dies hätte bspw. dann der Fall sein können, wenn sich in den Kurzzeittherapiestudien explizit aus Gründen der externen Validität dazu entschieden wurde, keine Manuale einzusetzen, was zu einem schlechten Ergebnis auf dem sog. „Manual-Kriterium“ (B.3.) geführt hätte, während sich im Gegenzug dazu der Manualeinsatz in Langzeittherapiestudien als *per se* schwieriger gestaltet (vgl. Kap. 3.4.3).

Bei drei der insgesamt 12 internen Validitätskriterien schnitt die Langzeittherapiestudie (19) mit negativem Ergebnis (Stufe „3“) ab. Dazu gehörten das „Manual-Kriterium“ (B.3.), das „Adherence-Kriterium“ (B.6.) sowie das „Kriterium zur Zulässigkeit nicht randomisierter Interventionen“ (B.7.). Bei zwei weiteren Kriterien schnitt die Langzeittherapiestudie im Gegensatz zu allen anderen Studien mit einem Missingwert ab, dazu zählten das „Kriterium zur operationalen Definition der Kontrollbedingungen“ (B.4.) und das „Kriterium zur strukturellen Äquivalenz bei Kontrollbedingungen“ (B.5.).

In Bezug auf das „Manual-Kriterium“ (B.3.) und das „Adherence-Kriterium“ (B.6.) konnte aufgezeigt werden, dass sowohl der Einsatz von Manualen bzw. von Behandlungsrichtlinien als auch die Durchführung von Adherence-Kontrollen sinnvolle interne Validitätssicherungsstrategien darstellen (vgl. Kap. 5.2.2). Im Hinblick auf Langzeittherapiestudien stellt sich der Manualeinsatz aufgrund einer nur geringen Verfügbarkeit von Langzeittherapiemanualen schwierig dar, jedoch besteht keine prinzipielle Inkompatibilität zwischen langfristigen Behandlungen und dem Einsatz von Manualen. Insbesondere jüngere Studien, wie die LAC-Studie (Leuzinger-Bohleber et al., 2010) oder die APS (Benecke et al., 2012) zeigen, dass einem Einsatz von Manualen in Langzeittherapien nichts Grundsätzliches entgegensteht. Demzufolge kann dem „Manual-Kriterium“ (B.3.) sicherlich eine benachteiligende Wirkung unterstellt werden, jedoch ist in jedem Fall zu betonen, dass diese allein auf die geringe Verfügbarkeit von Manualen für Langzeitbehandlungen zurückzuführen ist. Manualisierte Langzeitbehandlungen stellen nicht *per se* eine Unvereinbarkeit dar. Zentral ist für die Anwendbarkeit des „Manual-Kriteriums“, dass in der zugrunde gelegten Definition von „Manualen“ auch weniger strukturierte oder modularisierte Behandlungsrichtlinien inbegriffen sind. Andernfalls wäre dieses Kriterium nicht nur auf Langzeittherapiestudien, sondern auf Studien zu psychodynamischen Behandlungen generell nicht anwendbar (vgl. Beutel et al., 2010 und Kap. 3.2.2).

Vor dem geschilderten Hintergrund manualisierter Behandlungen und dem damit verbundenen Problem im Hinblick auf Langzeittherapiestudien könnte die Adherence-Kontrolle (Kriterium B.6.) einen zentraleren Stellenwert bekommen: Adherence-Kontrollen können sowohl unabhängig vom Behandlungsumfang als auch unabhängig vom Manualeinsatz durchgeführt werden – etwa mittels Instrumenten, wie der CPPS (Hilsenroth et al., 2005) oder dem PQS (Ablon & Jones, 1998; Jones, 2000). Insofern wäre zu erwägen, ob eine Gewich-

tung des „Adherence-Kriteriums“ für solche Studien sinnvoll wäre, in denen keine Manuale eingesetzt werden.

In Bezug auf das „Kriterium zur Zulässigkeit nicht randomisierter Interventionen“ (B.7.) konnte allenfalls eine marginale Benachteiligung von Langzeittherapiestudien ausgemacht werden, die sich allein auf die ersten beiden Ratingstufen (Stufen „1“ und „2“) dieses Kriteriums beschränkt.

Bei den beiden Kriterien B.4. („Kriterium zur operationalen Definition der Kontrollbedingungen“) und B.5. („Kriterium zur strukturellen Äquivalenz bei Kontrollbedingungen“) konnte wiederum eine grundsätzlichere Benachteiligung von Langzeittherapiestudien festgestellt werden. Diese besteht allerdings weniger darin, dass Langzeittherapiestudien hier systematisch „schlechter“ abschneiden, sondern vielmehr darin, dass diese beiden Kriterien auf Langzeittherapiestudien – ausgenommen solcher zu Borderline-Störungen – nicht anwendbar sind: Beide Kriterien setzen Kontrollgruppen (Wartelisten, Placebo, TAU) voraus, die in Langzeittherapiestudien in den meisten Fällen nicht realisierbar sind – zumindest dann nicht, wenn keine langfristig angelegten *Structured Clinical Management* Programme als TAU existieren, wie dies etwa für den Anwendungsbereich der Persönlichkeitsstörungen (speziell Borderline-Störungen) aufgezeigt werden konnte (vgl. de Maat, Dekker et al., 2007 und Kap. 5.4). Zur Untersuchung von Langzeittherapien eignen sich daher eher komparative Studiendesigns, d.h. Vergleiche mit bereits etablierten Therapieformen. Jedoch werden Gütemaßstäbe, wie sie explizit für komparative Studien ausschlaggebend sind, vom WBP nur in Form eines Kriteriums (A.17.: „Power-Kriterium“) abgefragt. Hier wäre es ratsam, weitere Kriterien zu entwickeln, die explizit die Bewertung interner Validitätssicherungsstrategien bei komparativen Studien zum Ziel haben. Darunter fallen etwa Strategien zur Vermeidung des sog. *researcher allegiance bias* (vgl. Caspar & Jacobi, 2007; Munder et al., 2013), die Konstanthaltung der Variable „therapeutische Berufserfahrung“ über die Therapiebedingungen sowie die For-

derung, dass zwischen den Therapiebedingungen eben *keine* strukturelle Angleichung der Randbedingungen stattfindet (vgl. Hager, 2000).

Abschließend wurden noch einige allgemeinere Empfehlungen ausgesprochen, die sich zum einen auf den Anwendungsbereich der „gemischten Störungen“ (hier verstanden als diagnoseheterogene Störungsgruppen) und zum anderen auf die generelle Anwendbarkeit des WBP-Kriterienkatalogs beziehen. Vor dem Hintergrund des doch verhältnismäßig großen Anteils an Untersuchungen an diagnoseheterogenen Störungsgruppen (n=31) in dieser Arbeit, wurde die Richtigkeit der Entscheidung des WBP (2010), diesen Anwendungsbereich allenfalls in Ausnahmefällen und nach umfassender Abwägung als gleichberechtigt zu anderen Anwendungsbereichen zu betrachten, in Zweifel gezogen. Demgegenüber wurde Rudolf gefolgt, der die Wichtigkeit der gemischten Störungen als gleichberechtigten Anwendungsbereich im Rahmen der wissenschaftlichen Anerkennung damit begründet, dass via ICD oder DSM gestellte Hauptdiagnosen im psychodynamischen Kontext nur wenig indikationsleitend seien (Psychotherapeutenjournal, 2008). Insofern ist die Zusammenstellung von Studienpatientengruppen nach ICD- oder DSM-Diagnosen im Rahmen psychodynamischer Wirksamkeitsstudien auch weniger gängig, verglichen bspw. mit Studien zur Verhaltenstherapie. Dies spiegelt sich u.a. in diversen Metaanalysen und Überblickswerken zur Wirksamkeit psychodynamischer Psychotherapie wider (z.B. DGPT, 2011; de Maat et al., 2009; Leichsenring, & Rabung, 2009). Auch, wenn ein Trend dahingehend zu verzeichnen ist, dass psychodynamische Studien an diagnosehomogenen Patientengruppen sich mehren (vgl. Benecke et al., 2012; Huber, Zimmermann et al., 2012; Leuzinger-Bohleber et al., 2010), scheint der hohe Anteil an Untersuchungen an gemischten Störungsgruppen sowohl in dieser Arbeit als auch in genannten Reviews das, was Rudolf postuliert, zu bestätigen. Vor diesem Hintergrund wurde es als ratsam

erachtet, den Anwendungsbereich der gemischten Störungen wieder aus seiner Außenseiterstellung herauszuholen, in die er vom WBP (2009b) gewissermaßen verbannt wurde.

Zur Anwendbarkeit des WBP-Kriterienkatalogs: Die Kriterien wurden in dieser Arbeit immer auch als Richtschnur für die Planung und Durchführung von Wirksamkeitsstudien betrachtet. Als solche könnten sie durchaus geeignet sein, allerdings sind sie nicht immer eindeutig zu interpretieren. Bei einigen Kriterien ergeben sich trotz Operationalisierung der einzelnen Ratingstufen immer noch enorme Interpretationsspielräume, die spätestens in der Anwendung der Kriterien auf Studien zu unreliablen Beurteilungen führen können. Zur Erhöhung der Reliabilität der Beurteilungen wurden in dieser Arbeit umfangreiche Kodierregeln entwickelt (Anhang C), durch die die Kodierungen deutlich vereinfacht wurden. Eine Untersuchung der Kodierregeln in ihrem Beitrag zur Reliabilität der Beurteilungen steht noch aus.

Insgesamt konnte der erste Eindruck einer übermäßigen Strenge insbesondere der methodischen Qualitätskriterien sowie der internen Validitätskriterien, der durch den „Flow“ (den Verlauf) der Studien durch das Bewertungsraster des WBP-Kriterienkatalogs nahegelegt wird (vgl. Abbildung 27, S. 263), in großen Teilen relativiert werden. Die gilt sowohl für die Bewertung psychodynamischer Wirksamkeitsstudien im Allgemeinen als auch für die Bewertung von Langzeittherapiestudien (> 100 Stunden) im Besonderen. Trotz der im Rahmen von Kapitel 5.4 hergeleiteten Annahme, der zufolge die in dieser Arbeit gezogenen Schlussfolgerungen über einzelne Kriterien nur mit wenigen Ausnahmen auch auf andere Anwendungsbereiche zu generalisieren sind, sollten sowohl Studien aus anderen Anwendungsbereichen als auch aus anderen Publikationszeiträumen zur weiteren Untersuchung der Kriterien herangezogen werden. Vor allem eine Ausweitung des Publikationszeitraums über das Jahr 2009 hinaus, könnte den hier gewonnenen Eindruck von größtenteils adäquaten Kriterien, jedoch auch

von einigen aufgezeigten Schwächen bekräftigen. Erste Hinweise dazu lieferte bereits ein Blick in die neuere Studienlage (vgl. Kap. 5).

Generell – nun in Richtung der Studien gesprochen – sind für die Zukunft Studien wünschenswert, die das komplementäre Verhältnis von interner und externer Validität ernst nehmen und umzusetzen versuchen. Um diesen Prozess voranzutreiben, könnte eine Modifikation und Ausweitung der WBP-Kriterien, wie in dieser Arbeit exemplarisch aufgezeigt, durchaus einen Anreiz setzen.

7 Literatur[87]

*Abbass, A. A. (2002). Office based research in Intensive Short-term Dynamic Psychotherapy (ISTDP): Data from the First 6 Years of Practice. *Ad Hoc Bulletin of Short-term Dynamic Psychotherapy, 6* (2), 5-14.

*Abbass, A. A. (2006). Intensive Short-Term Dynamic Psychotherapy of treatment-resistant depression: a pilot study. *Depression and Anxiety, 23* (7), 449-452.

Abbass, A. A. (2008). *Depression Studies Pertinent to NICE Guidelines: Short-term Psychodynamic Psychotherapies*. Verfügbar unter: http://www.istdp.ca/docs/Depression%20Studies%20Pertinent%20to%20NICE%20Guidelines.pdf [15.9.2014].

Abbass, A. A., Hancock, J. K., Henderson, J. & Kisely, S. R. (2006). Short-term psychodynamic psychotherapies for common mental disorders. *Cochrane Database of Systematic Reviews* Issue 4.

Ablon, J. S. & Jones, E. E. (1998). How expert clinicians' prototypes of an ideal treatment correlate with outcome in psychodynamic and cognitive-behavior therapy. *Psychotherapy Research, 8* (1), 71-83.

Altman, D. G. & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal, 19* (311), 485.

[87] Die mit einem * versehenen Literaturangaben kennzeichnen die Publikationen, die im Rahmen dieser Arbeit für die Studienkodierungen herangezogen wurden.

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention* (4. Aufl.). Heidelberg: Springer Medizin Verlag.

Amelang, M. & Zielinski, W. (2002). *Psychologische Diagnostik und Intervention* (3. Aufl.). Berlin: Springer.

American Educational Research Association, American Psychological Association (APA) & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychiatric Association (Hrsg.) (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5. Aufl.). Washington, DC: American Psychiatric Association.

American Psychological Association (APA) (2002). Criteria for evaluating treatment guidelines. *American Psychologist, 57* (12), 1052-1059.

American Psychological Association (APA) (2010). *Publication Manual* (6. Aufl.). Washington, DC: American Psychological Association.

American Psychological Association (APA) Presidential Task Force on Evidence-Based Practice (2006). Evidence-Based Practice in Psychology. *American Psychologist, 61* (4), 271-285.

American Psychological Association (APA) Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63* (9), 839-851.

Arbeitskreis OPD (Hrsg.) (1996). *Operationalisierte Psychodynamische Diagnostik: Grundlagen und Manual* (1. Aufl.). Bern: Huber.

Arbeitskreis OPD (Hrsg.) (2009). *Operationalisierte Psychodynamische Diagnostik OPD-2: Das Manual für Diagnostik und Therapieplanung* (2. Aufl.). Bern: Huber.

Archer, R., Forbes, Y. & Metcalfe, C. & Winter. D. (2000). An investigation of the effectiveness of a voluntary sector psychodynamic counselling service. *The British Journal of Medical Psychology, 73* (3), 401-412.

Areán, P. A. & Cook, B. L. (2002). Psychotherapy and combined psychotherapy/pharmacotherapy for late life depression. *Biological Psychiatry, 52* (3), 293-303.

Asendorpf, J. & Wallbott, H. G. (1979). Maße der Beobachterübereinstimmung: Ein systematischer Vergleich. *Zeitschrift für Sozialpsychologie, 10* (3), 243-252.

Auckenthaler, A. (2000a). Labor-Wirksamkeitsnachweise als Grundlage versorgungspolitisch relevanter Empfehlungen? Anmerkungen zur Begutachtungspraxis des Wissenschaftlichen Beirats Psychotherapie. *Gesprächspsychotherapie und Personzentrierte Beratung, 31* (1), 10-12.

Auckenthaler, A. (2000b). Die Manualisierung der Psychotherapie: Ziele und Implikationen. In M. Hermer (Hrsg.), *Forum für Verhaltenstherapie und psychosoziale Praxis: Bd. Band 43. Psychotherapeutische Perspektiven am Beginn des 21. Jahrhunderts* (S. 213–223). Tübingen: DGVT Deutsche Gesellschaft für Verhaltenstherapie.

Auckenthaler, A. (Hrsg.) (2012). *Klinische Psychologie und Psychotherapie: Grundlagen, Praxis, Kontext.* Kurzlehrbuch. Stuttgart [u.a.]: Thieme.

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2000). *Multivariate Analysemethoden*. Berlin, Heidelberg: Springer.

Bandelow, B., Wiltink, J., Alpers, G., Benecke, C., Deckert, J., Eckhardt-Henn, A., Ehrig, C., Engel, E., Falkai, P., Geiser, F., Gerlach, A. L., Harfst, T., Hau, S., Joraschky, P., Kellner, M., Köllner, V., Kopp, I., Langs, G., Lichte, T., Liebeck, H., Matzat, J., Reitt, M., Rüddel,

H. P., Rudolf, S., Schick, G., Schweiger, U., Simon, R., Springer, A., Staats, H., Ströhle, A., Ströhm, W., Waldherr, B., Watzke, B., Wedekind, D., Zottl, C., Zwanzger, P., Beutel M. E. *Deutsche S3-Leitlinie Behandlung von Angststörungen*. Verfügbar unter: www.awmf.org/leitlinien.html [9.1.2014].

*Barber, J. P., Wilczek, A., Gustavsson, J. P., Weinryb, R. M. & Åsberg, M. (2004). Change after Long-Term Psychoanalytic Psychotherapy. *Journal of the American Psychoanalytic Association, 52* (4), 1163-1184.

Barkham, M., Rees, A., Stiles, W. B., Hardy, G. E. & Shapiro, D. A. (2002). Dose-effect relations for psychotherapy of mild depression: A Quasi-experimental comparison of effects of 2, 8, and 16 sessions. *Psychotherapy Research, 12* (4), 463-474.

Barkham, M., Stiles, W. B., Connell, J., Twigg, E., Leach, C., Lucock, M., Mellor-Clark, J., Bower, P., King, M., Shapiro, D. A., Hardy, G. E., Greenberg, L. & Angus, L. (2008). Effects of psychological therapies in randomized trials and practice-based studies. *British Journal of Clinical Psychology, 47* (4), 397-415.

Barlow, D. H. (2010). Negative effects from psychological treatments: A perspective. *American Psychologist, 65* (1), 13-20.

Bartram, D., Lindley, P. & Kennedy, N. (2008). *EFPA Review Model for the Description and Evaluation of Psychological Tests: Test review form and notes for reviewers*. Verfügbar unter: http://europsyche.eu/download/9044bd41c7953b956876e06c797f8c9f [12.1.2010].

Bateman, A. W. & Fonagy, P. (2008). *Psychotherapie der Borderline-Persönlichkeitsstörung. Ein mentalisierungsgestütztes Behandlungskonzept*. Gießen: Psychosozial Verlag.

Bateman, A. & Fonagy, P. (2009). Randomized controlled trial of outpatient mentalization-based treatment versus structured clinical management for borderline personality disorder. *The American Journal of Psychiatry, 166* (12), 1355-1364.

Beck, A. T., Steer, A. T. & Brown, G. (1996). *Manual for the Beck Depression Inventory – II*. San Antonio: Psychological Corporation.

Beckmann, D. & Richter H.-E. (1975). *Gießen-Test (GT). Ein Test für Individual- und Gruppendiagnostik.* Bern, Stuttgart, Wien: Huber.

Beelmann, A. & Bliesener, T. (1994). Aktuelle Probleme und Strategien der Metaanalyse. *Psychologische Rundschau, 45* (4), 211-233.

Benecke, C. (2014a). *Klinische Psychologie und Psychotherapie: Ein integratives Lehrbuch* (1. Aufl.). Stuttgart: Kohlhammer.

Benecke, C. (2014b). Die Bedeutung empirischer Forschung für die Psychoanalyse. *Forum der Psychoanalyse, 30* (1), 55-67.

Benecke, C., Boothe, B., Frommer, J., Huber, D., Krause, R. & Staats, H. (2009). Geliebtes Feindbild „klassische Langzeitpsychoanalyse“: Kommentar zu Rief und Hofmann „Die Psychoanalyse soll gerettet werden. Mit allen Mitteln?“. *Der Nervenarzt, 80* (11), 1350-1355.

Benecke, C., Huber, D., Schauenburg, H. & Staats, H. (2012, 02. Juni). *Vorstellung der APS-Studie: Wirksamkeit Analytischer Psychotherapie und Kognitiver Verhaltenstherapie bei Angst- plus Persönlichkeitsstörung*. Berlin.

Berghout, C. C. & Zevalkink, J. (2009). Clinical significance of long-term psychoanalytic treatment. *Bulletin of the Menninger Clinic, 73* (1), 7-33.

Beutel, M. E., Doering, S., Leichsenring, F. & Reich, G. (2010). *Psychodynamische Psychotherapie: Störungsorientierung und Manualisierung in der therapeutischen Praxis*. Göttingen: Hogrefe.

Beutel, M. E., Leuzinger-Bohleber, M., Rüger, B., Bahrke, U., Negele, A., Haselbacher, A., Fiedler, G., Keller, W. & Hautzinger, M. (2012). Psychoanalytic and cognitive-behavior therapy of chronic depression: study protocol for a randomized controlled trial. *Trials, 13,* 117.

*Beutel, M. E. & Rasting, M. (2001). Langzeittherapien aus der Rückschau ehemaliger Patienten. In U. Stuhr, M. Leuzinger-Bohleber & M. E. Beutel (Hrsg.), *Langzeit-Psychotherapie. Perspektiven für Therapeuten und Wissenschaftler* (S. 187–200). Stuttgart: Kohlhammer.

Beutler, L. E., Malik, M., Alimohamed, S., Harwood, T. M., Talebi, H., Noble, S. & Wong, E. (2004). Therapist Variables. In M. J. Lambert (Hrsg.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (5th ed.) (S. 227–306). New York: John Wiley & Sons Inc.

*Blatt, S. J. & Shahar, G. (2004). Psychoanalysis with whom, for what, and how? Comparisons with Psychotherapy. *Journal of the American Psychoanalytic Association* (52), 393-447.

Blay, S. L., Vel Fucks, J. S., Barruzi, M., Di Pietro, M. C., Gastal, F. L., Neto, A. M., De Souza, M. P., Glausiusz, L. R. U. & Dewey, M. (2002). Effectiveness of time-limited psychotherapy for minor psychiatric disorders: Randomised controlled trial evaluating immediate v. long-term effects. *The British Journal of Psychiatry, 180* (5), 416-422.

*Blomberg, J., Lazar, A. & Sandell, R. (2001). Long-term outcome of long-term psychoanalytically oriented therapies: First findings of the Stockholm outcome of psychotherapy and psychoanalysis study. *Psychotherapy Research, 11* (4), 361-382.

Bloom, B. L. (2001). Planned short-term psychotherapy for depression: Recent controlled outcome studies. *Brief Treatment and Crisis Intervention, 1* (2), 169-189.

Boath, E. & Henshaw, C. (2001). The treatment of postnatal depression: A comprehensive literature review. *Journal of Reproductive and Infant Psychology, 19* (3), 215-248.

Boland, R. J. & Keller, M. B. (2009). Course and outcome of depression. In I. H. Gotlib & C. L. Hammen (Hrsg.), *Handbook of depression* (2nd ed.) (S. 23–43). New York, NY US: Guilford Press.

Bond, M. (2006). Psychodynamic psychotherapy in the treatment of mood disorders. *Current Opinion in Psychiatry, 19* (1), 40-43.

*Bond, M. & Perry, J. C. (2004). Long-term changes in defense styles with psychodynamic psychotherapy for depressive, anxiety, and personality disorders. *The American Journal of Psychiatry, 161* (9), 1665-1671.

*Bond, M. & Perry, J. C. (2006). Psychotropic medication use, personality disorder and improvement in long-term dynamic psychotherapy. *Journal of Nervous and Mental Disease, 194* (1), 21-26.

Borkovec, T. D. & Costonguay, L. G. (1998). What is the scientific meaning of empirically supported therapy? *Journal of Consulting and Clinical Psychology, 66* (1), 136-142.

Bortolotti, B., Menchetti, M., Bellini, F., Berardi, D. & Montaguti, M. B. (2008). Psychological interventions for major depression in primary care: A meta-analytic review of randomized controlled trials. *General Hospital Psychiatry, 30* (4), 293-302.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler* (6. Aufl.). Berlin, Heidelberg, New York: Springer.

Bortz, J., Lienert, G. A. & Boehnke, K. (Hrsg.) (2000). *Verteilungsfreie Methoden in der Biostatistik.* Berlin, Heidelberg: Springer.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. Berlin, Heidelberg: Springer.

*Bradshaw, W., Roseborough, D., Pahwa, R. & Jordan, J. (2009). Evaluation of psychodynamic psychotherapy in a community mental health center. *Journal of the American Academy of Psychoanalysis & Dynamic Psychiatry, 37* (4), 665-682.

Brähler, E. & Scheer, J. W. (1995). *Der Giessener Beschwerdebogen. Handbuch* (2. Aufl.). Bern: Huber.

Brähler, E., Schumacher, J. & Strauß, B. (Hrsg.) (2002). *Diagnostische Verfahren in der Psychotherapie.* Diagnostik für Klinik und Praxis, Bd. 1. Göttingen: Hogrefe.

Brandl, Y., Bruns, G., Gerlach, A., Hau, S., Janssen, P. L., Kächele, H., Leichsenring, F., Leuzinger-Bohleber, M., Mertens, W., Rudolf, G., Schlösser, A.-M., Springer, A., Stuhr, U. & Windaus, E. (2004). Psychoanalytische Therapie: Eine Stellungnahme für die wissenschaftliche Öffentlichkeit und für den Wissenschaftlichen Beirat Psychotherapie. *Forum der Psychoanalyse, 20* (1), 13-125.

*Brockmann, J., Schlüter, T. & Eckert, J. (2001). Die Frankfurt-Hamburg Langzeit-Psychotherapiestudie - Ergebnisse der Untersuchung psychoanalytisch orientierter und verhaltenstherapeutischer Langzeit-Psychotherapien in der Praxis niedergelassener Psychotherapeuten. In U. Stuhr, M. Leuzinger-Bohleber & M. E. Beutel (Hrsg.), *Langzeit-Psychotherapie. Perspektiven für Therapeuten und Wissenschaftler* (S. 271–276). Stuttgart: Kohlhammer.

*Brockmann, J., Schlüter, T., Brodbeck, D. & Eckert, J. (2002). Die Effekte psychoanalytisch orientierter und verhaltenstherapeutischer Langzeittherapien. *Psychotherapeut, 47* (6), 347-355.

*Brockmann, J., Schlüter, T. & Eckert, J. (2006). Langzeitwirkungen psychoanalytischer und verhaltenstherapeutischer Langzeitpsychotherapien. *Psychotherapeut, 51* (1), 15-25.

Brosius, F. (2011). *SPSS 19* (1. Aufl.). Heidelberg, München, Landsberg, Frechen, Hamburg: mitp.

Brozek, J. L., Akl, E. A., Alonso-Coello, P., Lang, D., Jaeschke, R., Williams, J. W., Phillips, B., Lelgemann, M., Lethaby, A., Bousquet, J., Guyatt, G. H. & Schünemann, H. J. (2009). Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions. *Allergy, 64* (5), 669-677.

Buchholz, M. B. (2008). Was ist eine methodisch adäquate Wirksamkeitsstudie? Zum Stand einer Kontroverse. *PTT: Persönlichkeitsstörungen Theorie und Therapie, 12* (1), 12-22.

Bundesministerium der Justiz und für Verbraucherschutz (1998, 16. Juni). Gesetz über die Berufe des Psychologischen Psychotherapeuten und des Kinder- und Jugendlichenpsychotherapeuten: Psychotherapeutengesetz - PsychThG.

Bundespsychotherapeutenkammer (2007). *Ergänzung der Stellungnahme der BPtK nach § 91 Abs. 8a SGB V vom 30.10.2006 zur Gesprächspsychotherapie unter Einbezug des vollständigen Berichts zur Nutzenbewertung der Gesprächspsychotherapie bei Erwachsenen vom 17.07.2006,* Bundespsychotherapeutenkammer. Verfügbar unter: http://www.bptk.de/uploads/media/20071105_stn_bptk_gpt.pdf [25.9.2014].

Burnand, Y., Andreoli, A., Kolatte, E., Venturini, A. & Rosset, N. (2002). Psychodynamic psychotherapy and clomipramine in the treatment of major depression. *Psychiatric Services, 53* (5), 585-590.

Busch, F. (2010). Distinguishing psychoanalysis from psychotherapy. *The International Journal of Psychoanalysis, 91* (1), 23-34.

Caligor, E. (2005). Treatment manuals for long-term psychodynamic psychotherapy and psychoanalysis. *Clinical Neuroscience Research, 4* (5-6), 387-398.

Caligor, E., Hilsenroth, M. J., Devlin, M., Rutherford, B. R., Terry, M. & Roose, S. P. (2012). Will patients accept randomization to psychoanalysis? A feasibility study. *Journal of the American Psychoanalytic Association, 60* (2), 337-360.

Caspar, F. & Jacobi, F. (2007). Psychotherapieforschung. In W. Hiller, E. Leibing, F. Leichsenring & S. Sulz (Hrsg.), *Bd. 1: Wissenschaftliche Grundlagen der Psychotherapie. Lehrbuch der Psychotherapie* (4 Aufl.) (S. 395–410). München: CIP-Medien.

Chambless, D. L. (1996). In defense of dissemination of empirically supported psychological interventions. *Clinical Psychology: Science and Practice, 3* (3), 230-235.

Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L. E., Calhoun, K. S., Crits-Christoph, P., Daiuto, A., DeRubeis, R., Detweiler, J., Haaga, D. A. F., Bennett Johnson, S., McCurry, S., Mueser, K. T., Pope, K. S., Sanderson, W. C., Shoham, V., Stickle, T., Williams, D. A. & Woody, S. R. (1998). Update on Empirically Validtated Therapies, II. *The Clinical Psychologist, 51* (1), 3-16.

Chambless, D. L. & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66* (1), 7-18.

Chambless, D. L. & Ollendick, T. H. (2001). Empirically Supported Psychological Interventions: Controversies and Evidence. *Annual Review of Psychology, 52* (1), 685-716.

Chambless, D. L., Sanderson, W. C., Shoham, V., Bennett Johnson, S., Pope, K. S., Crits-Christoph, P., Baker, M., Johnson, B., Woody, S. R., Sue, S., Beutler, L., Williams, D. A., & McCurry, S. (1996). An Update on Empirically Validated Therapies. *The Clinical Psychologist, 49,* 5-18.

Churchill, R., Hunot, V., Corney, R., Knapp, M., McGuire, H., Tylee, A. & Wessely, S. (2001). A systematic review of controlled trials of the effectiveness and cost-effectiveness of brief psychological treatments for depression. *Health Technology Assessment, 5* (35), 1-173.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6* (4), 284-290.

Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology, 23* (5), 695-700.

Cicchetti, D. V. & Rourke, B. P. (2004). *Methodological and biostatistical foundations of clinical neuropsychology and medical and health disciplines* (2nd ed.). Hove, England: Psychology Press/Taylor & Francis (UK).

Clarkin, J. F., Yeomans, F. E. & Kemberg, O. F. (2006). *Psychotherapie der Borderline-Persönlichkeit. Manual zur psychodynamischen Therapie.* Stuttgart: Schattauer.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Committee for Medicinal Products for Human Use (CHMP) (2006). Guideline on the choice of the non-inferiority margin. *Statistics in Medicine, 25* (10), 1628-1638.

Cooper, H. & Hedges, L. V. (Hrsg.) (1994). *The Handbook of Research Synthesis.* New York, NY, US: Russell Sage Foundation.

*Cooper, P. J., Murray, L., Wilson, A. & Romaniuk, H. (2003). Controlled trial of the short- and long-term effect of psychological treatment of post-partum depression. 1. Impact on maternal mood. *The British Journal of Psychiatry, 182* (5), 412-419.

Crits-Christoph, P., Wolf-Palacio, D., Ficher, M. & Rudick, D. (1995). Brief supportive expressive psychodynamic therapy for generalized anxiety disorder. In J. P. Barber & P. Crits-Christoph (Hrsg.), *Dynamic therapies for psychiatric disorders (Axis 1)* (S. 43-83). New York: Basic Books.

Crowe, T. P. & Grenyer, B. F. S. (2008). Is therapist alliance or whole group cohesion more influential in group psychotherapy outcomes? *Clinical Psychology and Psychotherapy, 15* (4), 239-246.

Cuijpers, P., van Straten, A., Andersson, G. & van Oppen, P. (2008). Psychotherapy for depression in adults: A meta-analysis of comparative outcome studies. *Journal of Consulting and Clinical Psychology, 76* (6), 909-922.

Dekker, J., Molenaar, P. J., Kool, S., van Aalst, G., Peen, J. & Jonghe, F. de. (2005). Dose-effect relations in time-limited combined psycho-pharmacological treatment for depression. *Psychological Medicine, 35* (1), 47-58.

Dennis, C. L. & Hodnett, E. D. (2007). Psychosocial and psychological interventions for treating postpartum depression. *Cochrane Database of Systematic Reviews*, Issue 4.

Dennis, C. L., Ross, L. E. & Grigoriadis, S. (2007). Psychosocial and psychological interventions for treating antenatal depression. *Cochrane Database of Systematic Reviews*, Issue 3.

Derogatis, L. R. (1975). *The SCL-90-R*. Baltimore: Clinical Psychometric Research.

Deutsche Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde (DGPPN), Bundesärztekammer (BÄK), Kassenärztliche Bundesvereinigung (KBV), Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF), Arzneimittelkommission der deutschen Ärzteschaft (AkdÄ), Bundespsychotherapeutenkammer (beratend) (BPtK), Bundesverband der Angehörigen psychisch Kranker (BApK), Deutsche Arbeitsgemeinschaft Selbsthilfegruppen (DAGSHG), Deutsche Gesellschaft für Allgemeinmedi-

zin und Familienmedizin (DEGAM), Deutsche Gesellschaft für Psychosomatische Medizin und Ärztliche Psychotherapie (DGPM), Deutsche Gesellschaft für Psychologie (DGPs), Deutsche Gesellschaft für Rehabilitationswissenschaften (DGRW) (Hrsg.) (2009). *S3-Leitlinie/Nationale VersorgungsLeitlinie Unipolare Depression - Langfassung Version 1.3, Januar 2012, basierend auf der Fassung von November 2009*. Verfügbar unter: http://www.versorgungsleitlinien.de/themen/depression/pdf/s3_nvl_depression_lang.pdf [30.8.2010].

Deutsche Gesellschaft für Psychoanalyse, Psychotherapie, Psychosomatik und Tiefenpsychologie (DGPT) e.V. (2009). *Stellungnahme zur Prüfung der Richtlinienverfahren gemäß §§ 13 – 15 der Psychotherapie-Richtlinie für die psychoanalytisch begründeten Verfahren,* DGPT. Verfügbar unter: http://www.dgpt.de/dokumente/DGPT%20Stllungnahme%20zur%20Pruefung%20der%20Richtlinienverfahren%202009.pdf [30.8.2010].

Deutsche Gesellschaft für Psychoanalyse, Psychotherapie, Psychosomatik und Tiefenpsychologie (DGPT) e.V. (2011). Stellungnahme zur Prüfung der Richtlinienverfahren gemäß §§ 13-15 der Psychotherapie-Richtlinie für die psychoanalytisch gegründeten Verfahren. *Forum der Psychoanalyse, 27* (Supplement 1), S3.

Deutsche Gesellschaft für Psychologie (DGPs) (2007). *Richtlinien zur Manuskriptgestaltung* (3. Aufl.). Göttingen: Hogrefe.

Deutsche Industrienorm (DIN) 33430 (2002). *DIN 33430 Berufsbezogene Eignungsdiagnostik: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin.

Deutsches Ärzteblatt (2008a). Interview mit Prof. Dr. Dietmar Schulte, alternierender Vorsitzender des WBP - „Die Hürde wird etwas höher“: Der Wissenschaftliche Beirat Psycho-

therapie hat neue Verfahrensregeln zur Beurteilung der wissenschaftlichen Anerkennung von Methoden und Verfahren vorgelegt. Ein Gespräch über die Änderungen und ihre Auswirkungen. *Deutsches Ärzteblatt, 105* (8), 388-390.

Deutsches Ärzteblatt (2008b). Bekanntmachungen: Methodenpapier des Wissenschaftlichen Beirats Psychotherapie. *Deutsches Ärzteblatt, 105* (26), A-1464.

Dimidjian, S. & Hollon, S. D. (2010). How would we know if psychotherapy were harmful? *American Psychologist, 65* (1), 21-33.

Driessen, E., Cuijpers, P., de Maat, S. C. M., Abbass, A. A., Jonghe, F. de & Dekker, J. (2010). The efficacy of short-term psychodynamic psychotherapy for depression: A meta-analysis. *Clinical Psychology Review, 30* (1), 25-36.

Driessen, E., Van, H. L., Schoevers, R. A., Cuijpers, P., van Aalst, G., Don, F. J., Hendriksen, M., Kool, S., Molenaar, P. J., Peen, J. & Dekker, J. (2007). Cognitive Behavioral Therapy versus Short Psychodynamic Supportive Psychotherapy in the outpatient treatment of depression: A randomized controlled trial. *BMC Psychiatry, 7,* 58.

Dupuy, H. J. (1984). The Psychological General Well-being (PGWB) Index. In N. K. Wenger, M. E. Mattson, C. D. Furburg & J. Elinson (Hrsg.), *Assessment of Quality of Life in Clinical Trials of Cardiovascular Therapies* (S. 184–188). New York: Le Jacq Publishing.

Eckert, J. & Biermann-Ratjen, E.-M. (1985). *Stationäre Gruppenpsychotherapie: Prozesse, Effekte, Vergleiche*. Berlin: Springer.

Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden* (1. Aufl.). Weinheim [u.a.]: Beltz.

Erle, J. B. & Goldberg, D. A. (2003). The course of 253 analyses from selection to outcome. *Journal of the American Psychoanalytic Association, 51* (1), 257-292.

Evers, A. (2001a). Improving Test Quality in the Netherlands: Results of 18 Years of Test Ratings. *International Journal of Testing, 1* (2), 137.

Evers, A. (2001b). The Revised Dutch Rating System for Test Quality. *International Journal of Testing, 1* (2), 155-182.

Fahrenberg, J., Myrtek, M., Schumacher, J. & Brähler, E. (2000). *Fragebogen zur Lebenszufriedenheit*. Göttingen: Hogrefe.

Fahrenberg, J. & Selg, H. (1970). *Das Freiburger Persönlichkeitsinventar FPI: Handanweisung* (1. Aufl.). Göttingen: Hogrefe.

Faller, H. & Reusch, A. (2004). Das experimentelle Design bei der Evaluation von Patientenschulungen. *Praxis Klinische Verhaltensmedizin und Rehabilitation, 17* (65), 13-18.

Faravelli, C., Ambonetti, A., Pallanti, S. & Pazzagli, A. (1986). Depressive relapses and incomplete recovery from index episode. *Am J Psychiatry, 143* (7), 888-891.

Fisseni, H.-J. (1997). *Lehrbuch der psychologischen Diagnostik: Mit Hinweisen zur Intervention*. Göttingen: Hogrefe.

Fonagy, P. (2009, 02. Oktober). *Veränderungen der klinischen Praxis: wissenschaftlich oder pragmatisch begründet? Vortrag auf der Jahrestagung der DGPT*. Aus dem Englischen übersetzt von: Elisabeth Vorspohl, Frankfurt/Main. Berlin.

Fonagy, P. (2000). The outcome of psychoanalysis: The hope of a future. *The Psychologist, 13* (12), 620-623.

Fonagy, P., Jones, E. E., Kächele, H., Krause, R., Clarkin, J. F., Perron, R., Gerber, A. J. & Allison, E. (Hrsg.) (2002). *An Open Door Review of Outcome Studies in Psychoanalysis* (2nd ed.). London: International Psychoanalytical Association.

Fonagy, P., Roth, A. & Higgitt, A. (2005). Psychodynamic psychotherapies: Evidence-based practice and clinical wisdom. *Bulletin of the Menninger Clinic, 69* (1), 1-58.

Freedman, N., Hoffenberg, J. D., Vorus, N. & Frosch, A. (1999). The effectiveness of psychoanalytic psychotherapy: The role of treatment duration, frequency of sessions, and the therapeutic relationship. *Journal of the American Psychoanalytic Association, 47* (3), 741-772.

Frohburg, I. (2004). Katamnesen zur Gesprächspsychotherapie: Überblicksarbeit. *Zeitschrift für Klinische Psychologie und Psychotherapie: Forschung und Praxis, 33* (3), 196-208.

Fydrich, T. & Schneider, S. (2007). Evidenzbasierte Psychotherapie. *Psychotherapeut, 52* (1), 55-68.

Gallagher-Thompson, D., Hanley-Peterson, P. & Thompson, L. W. (1990). Maintenance of gains versus relapse following brief psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 58* (3), 371-374.

Gemeinsamer Bundesausschuss (2007). *Tragende Gründe zum Beschluss über eine Änderung der Psychotherapie-Richtlinien: Einführung eines Schwellenkriteriums*. Verfügbar unter: https://www.g-ba.de/downloads/40-268-492/2007-12-20-Psycho-Schwellenkriterium_TrG.pdf [12.9.2014].

Gemeinsamer Bundesausschuss (2013). Richtlinie des Gemeinsamen Bundesausschusses über die Durchführung der Psychotherapie (Psychotherapie-Richtlinien). Verfügbar unter: https://www.g-ba.de/downloads/62-492-713/PT-RL_2013-04-18.pdf [11.10.2014].

Gerber, A. J., Fonagy, P., Bateman, A. & Higgitt, A. (2004). Structural and symptomatic change in psychoanalysis and psychodynamic psychotherapy of young adults: A quantitative study of process and outcome. *Journal of the American Psychoanalytic Association, 52* (4), 1235-1236.

Gerdes, N. & Jäckel, W. H. (1992). "Indikatoren des Reha-Status (IRES)" - Ein Patientenfragebogen zur Beurteilung von Rehabilitationsbedürftigkeit und -erfolg. *Die Rehabilitation, 31* (2), 73-79.

Gibbons, M. B. C., Crits-Christoph, P. & Hearon, B. (2008). The empirical status of psychodynamic therapies. *Annual Review of Clinical Psychology, 4,* 93-108.

Gill, M. M. (1954). Psychoanalysis and exploratory psychotherapy. *Journal of the American Psychoanalytic Association, 2,* 771-797.

Goldberg, D. P. (1978). *Manual of the General Health Questionnaire*. Windsor: NFER-NELSON Publishing.

Gonzalez, J. M. & Prihoda, T. J. (2007). A Case Study of Psychodynamic Group Psychotherapy for Bipolar Disorder. *American Journal of Psychotherapy, 61* (4), 405-422.

*Gordon, R. M. (2001). MMPI/MMPI-2 changes in long-term psychoanalytic psychotherapy. *Issues in Psychoanalytic Psychology, 23* (1-2), 59-79.

Gorwood, P., Weiller, E., Lemming, O. & Katona, C. (2007). Escitalopram prevents relapse in older patients with major depressive disorder. *The American Journal of Geriatric Psychiatry, 15* (7), 581-593.

Götze, P., Eckert, J., Nilsson, B., Biermann-Ratjen, E. M., Jährig, C., Kamp-Kowerk, M., Mohr, M., Niedermeyer, U., Papenhausen, R., Preuss, W. & Thomasius, R. (2003). Fokaltherapie. Was trägt zum Therapieerfolg bei? *Psychotherapeut, 48* (2), 122-128.

*Grande, T., Dilg, R., Jakobsen, T., Krawietz, B., Oberbracht, C., Stennes, M., Rudolf, G., Stehle, S., Langer, M. & Keller, W. (2006). Differential effects of two forms of psychoanalytic therapy: Results of the Heidelberg-Berlin study. *Psychotherapy Research, 16* (4), 470-485.

*Grande, T., Dilg, R., Jakobsen, T., Keller, W., Krawietz, B., Langer, M., Oberbracht, C., Stehle, S., Stennes, M. & Rudolf, G. (2009). Structural change as a predictor of long-term follow-up outcome. *Psychotherapy Research, 19* (3), 344-357.

*Grant, J. & Sandell, R. (2004). Close family or mere neighbours? Some empirical data on the differences between psychoanalysis and psychotherapy. In P. Richardson, H. Kächele & C. Renlund (Hrsg.), *Research on psychoanalytic psychotherapy with adults* (S. 81–108). London: Karnac.

Grawe, K., Caspar, F. & Ambühl, H. (1990). Die Berner Therapievergleichsstudie: Fragestellung und Versuchsplan. *Zeitschrift für Klinische Psychologie, 19* (4), 294-315.

*Guthrie, E., Moorey, J., Barker, H., Margison, F. & McGrath, G. (1998). Brief psychodynamic-interpersonal therapy for patients with severe psychiatric illness which is unresponsive to treatment. *British Journal of Psychotherapy, 15* (2), 155-166.

*Guthrie, E., Moorey, J., Margison, F., Barker, H., Palmer, S., McGrath, G., Tomenson, B. & Creed, F. (1999). Cost-effectiveness of brief psychodynamic-interpersonal therapy in high utilizers of psychiatric services. *Archives of General Psychiatry, 56* (6), 519-526.

Guy, W. (1976). *Clinical Global Impression. ECDEU Assessment Manual for Psychopharmacology*. Rockville, MD: U.S. Department of Health, Education, and Welfare.

Hager, W. (2000). Wirksamkeits- und Wirksamkeitsunterschiedshypothesen, Evaluationsparadigmen, Vergleichsgruppen und Kontrolle. In W. Hager, J.-L. Patry & H. Brezing (Hrsg.), *Handbuch Evaluation psychologischer Interventionsmaßnahmen. Standards und Kriterien* (S. 180–201). Bern: Huber.

Hager, W. & Hasselhorn, M. (2000). Psychologische Interventionsmaßnahmen: Was sollen sie bewirken können? In W. Hager, J.-L. Patry & H. Brezing (Hrsg.), *Handbuch Evaluati-*

on psychologischer Interventionsmaßnahmen. Standards und Kriterien (S. 41–85). Bern: Huber.

Hager, W., Patry, J.-L. & Brezing, H. (Hrsg.) (2000). *Handbuch Evaluation psychologischer Interventionsmaßnahmen. Standards und Kriterien.* Bern: Huber.

Hamilton, M. (1976). HAMA – Hamilton Anxiety Scale. In W. Guy (Hrsg.), *ECDEU Assessment Manual for Psychopharmacology, revised* (S. 193–198). Rockville, Maryland: NIMH.

Hartmann, S. (2006). *Die Behandlung psychischer Störungen. Wirksamkeit und Zufriedenheit aus Sicht des Patienten. Eine Replikation der "Consumer Reports Study" für Deutschland.* Gießen: Psychosozial Verlag.

Hartmann, S. & Zepf, S. (2002). Effektivität von Psychotherapie. Ein Vergleich verschiedener psychotherapeutischer Verfahren. *Forum der Psychoanalyse, 18* (2), 176-196.

Hauenschild, L. von. (2011). *Ist Gleichheit gerecht? Prüfung von Studien zu psychodynamischer Langzeitbehandlung auf der Grundlage des Methodenpapiers des WBPs*. Diplomarbeit, Humboldt-Universität zu Berlin.

Haupt, M.-L. & Linden, M. (2011). Nebenwirkungen und Nebenwirkungserfassung in der Psychotherapie - Das ECRS-ATR-Schema. *Psychotherapie und Sozialwissenschaft, 13* (2), 9-27.

Hautzinger, M. (2007). Psychotherapieforschung. In C. Reimer, J. Eckert, M. Hautzinger & E. Wilke (Hrsg.), *Psychotherapie: Ein Lehrbuch für Ärzte und Psychologen* (3 Aufl.) (S. 62–73). Heidelberg: Springer Medizin Verlag.

Hautzinger, M. & Welz, S. (2004). Kognitive Verhaltenstherapie bei Depressionen im Alter. Ergebnisse einer kontrollierten Vergleichsstudie unter ambulanten Bedingungen an De-

pressionen mittleren Schweregrads. *Zeitschrift für Gerontologie und Geriatrie, 37* (6), 427-435.

*Hecke, D., Hardt, J. & Tress, W. (2008). Zur Effektivität und klinischen Relevanz psychodynamischer Kurztherapie: Das Düsseldorfer Kurzzeittherapieprojekt (seit 1991). *Zeitschrift für Psychosomatische Medizin und Psychotherapie, 54* (2), 107-131.

Heekerens, H.-P. (1998). Evaluation von Erziehungsberatung: Forschungsstand und Hinweise zu künftiger Forschung. *Praxis der Kinderpsychologie und Kinderpsychiatrie, 47* (8), 589-606.

Heekerens, H. P. (2005). Vom Labor ins Feld. Die Psychotherapieevaluation geht neue Wege. *Psychotherapeut, 50* (5), 357-366.

Herzberg, P. Y. & Frey, A. (2011). Kriteriumsorientierte Diagnostik. In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Enzyklopädie der Psychologie, Themenbereich B, Methodologie und Methoden, Serie II, Psychologische Diagnostik, Bd. 2. Methoden der psychologischen Diagnostik* (S. 281–324). Göttingen: Hogrefe.

Higgins, J. P. T. & Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions: The Cochrane Collaboration* (Version 5.1.0). Verfügbar unter: www.cochrane-handbook.org [11.1.2012]

Hill, C. E. & Lambert, M. J. (2004). Methodological Issues in Studying Psychotherapy Processes and Outcomes. In M. J. Lambert (Hrsg.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (5th ed.) (S. 84–135). New York: John Wiley & Sons Inc.

Hiller, W., Bleichhardt, G. & Schindler, A. (2009). Evaluation von Psychotherapien aus der Perspektive von Qualitätssicherung und Qualitätsmanagement. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie, 57* (1), 7-22.

*Hilsenroth, M. J., Ackerman, S. J., Blagys, M. D., Baity, M. R. & Mooney, M. A. (2003). Short-term psychodynamic psychotherapy for depression: An examination of statistical, clinically significant, and technique-specific change. *Journal of Nervous and Mental Disease, 191* (6), 349-357.

Hilsenroth, M. J., Blagys, M. D., Ackerman, S. J., Bonge, D. R. & Blais, M. A. (2005). Measuring Psychodynamic-Interpersonal and Cognitive-Behavioral Techniques: Development of the Comparative Psychotherapy Process Scale. *Psychotherapy: Theory, Research, Practice, Training, 42* (3), 340-356.

*Høglend, P., Amlo, S., Marble, A., Sørbye, Ø., Heyerdahl, O., Sjaastad, M. C. & Bøgwald, K.-P. (2006). Analysis of the patient-therapist relationship in dynamic psychotherapy: An experimental study of transference interpretations. *The American Journal of Psychiatry, 163* (10), 1739-1746.

Høglend, P., Bøgwald, K.-P., Amlo, S., Heyerdahl, O., Sørbye, Ø., Marble, A., Sjaastad, M. C. & Bentsen, H. (2000). Assessment of change in dynamic psychotherapy. *Journal of Psychotherapy Practice and Research, 9* (4), 190-199.

*Høglend, P., Bøgwald, K.-P., Amlo, S., Ulberg, R., Sørbye, Ø., Johansson, P., Heyerdahl, O., Sjaastad, M. C. & Marble, A. (2008). Transference interpretations in dynamic psychotherapy: Do they really yield sustained effects? *The American Journal of Psychiatry, 165* (6), 763-771.

Horowitz, L. M. (1999). *Manual for the Inventory of Interpersonal Problems*. San Antonio: The Psychological Corporation.

Hsu, L. M. (1989). Random sampling, randomization, and equivalence of contrasted groups in psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57* (1), 131-137.

Huber, D., Henrich, G., Clarkin, J. & Klug, G. (2013). Psychoanalytic versus psychodynamic therapy for depression: a three-year follow-up study. *Psychiatry, 76* (2), 132-149.

Huber, D., Henrich, G., Gastner, J. & Klug, G. (2012). Must all have prizes? The Munich Psychotherapy Study. In R. A. Levy, J. S. Ablon & H. Kächele (Hrsg.), *Psychodynamic psychotherapy research: Evidence-based practice and practice-based evidence* (S. 51–69). Totowa, NJ, US: Humana Press.

Huber, D. & Klug, G. (2006). Munich Psychotherapy Study (MPS): The effectiveness of psychoanalytic long-term psychotherapy for depression. In Society for Psychotherapy Research (Hrsg.), *Book of abstracts - From research to practice* (S. 154). Ulm: Ulmer Textbank.

Huber, D., Zimmermann, J., Henrich, G. & Klug, G. (2012). Comparison of cognitive behavior therapy with psychoanalytic and psychodynamic therapy for depressed patients - A three-year follow-up study. *Zeitschrift für Psychosomatische Medizin und Psychotherapie, 58* (3), 299-316.

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) (2014). *Allgemeine Methoden: Entwurf für Version 4.2 vom 18.06.2014,* Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG). Verfügbar unter: https://www.iqwig.de/download/IQWiG_Methoden_Entwurf-fuer-Version-4-2.pdf [15.9.2014].

Jacobson, N. S., Follette, W. C. & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15* (4), 336-352.

Jacobson, N. S., Roberts, L. J., Berns, S. B. & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67* (3), 300-307.

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* (1), 12-19.

*Jakobsen, T. & Mattanza, G. (2005a). Die Katamneseresultate. In G. Mattanza, I. Meier & M. Schlegel (Hrsg.), *Seele und Forschung. Ein Brückenschlag in der Psychotherapie* (S. 130–142). Basel: Karger.

*Jakobsen, T. & Mattanza, G. (2005b). Hinweise auf günstige Therapieverläufe. In G. Mattanza, I. Meier & M. Schlegel (Hrsg.), *Seele und Forschung. Ein Brückenschlag in der Psychotherapie* (S. 143–164). Basel: Karger.

Janssen, J. & Laatz, W. (2007). *Statistische Datenanalyse mit SPSS für Windows: Eine anwendungsorientierte Einführung in das Basissystem und das Modul Exakte Tests* (6. Aufl.). Berlin: Springer.

Johnson, D. L. (2010). *A compendium of psychosocial measures: Assessment of people with serious mental illnesses in the community*. New York, NY US: Springer Publishing Co.

Jones, E. E. (2000). *Therapeutic action: A guide to psychoanalytic therapy*. Lanham, MD, US: Jason Aronson.

Jones, S. (2004). Psychotherapy of bipolar disorder: a review. *Journal of Affective Disorders, 80* (2-3), 101-114.

Jonghe, F. de, Hendriksen, M., van Aalst, G., Kool, S., Peen, J., Van, R., van den Eijnden, E. & Dekker, J. (2004). Psychotherapy alone and combined with pharmacotherapy in the treatment of depression. *The British Journal of Psychiatry, 185* (1), 37-45.

Jonghe, F. de, Kool, S., van Aalst, G., Dekker, J. & Peen, J. (2001). Combining psychotherapy and antidepressants in the treatment of depression. *Journal of Affective Disorders, 64* (2), 217-229.

Jong-Meyer, R. de, Hautzinger, M., Rudolf, G. A. E., Strauß, W. & Frick, U. (1996). Die Überprüfung der Wirksamkeit einer Kombination von Antidepressiva- und Verhaltenstherapie bei endogen depressiven Patienten: Varianzanalytische Ergebnisse zu den Haupt- und Nebenkriterien des Therapieerfolgs. *Zeitschrift für Klinische Psychologie, 25* (2), 93-109.

Judd, L. L., Akiskal, H. S., Maser, J. D., Zeller, P. J., Endicott, J., Coryell, W., Paulus, M. P., Kunovac, J. L., Leon, A. C., Mueller, T. I., Rice, J. A. & Keller, M. B. (1998). Major depressive disorder: A prospective study of residual subthreshold depressive symptoms as predictor of rapid relapse. *Journal of Affective Disorders, 50* (2-3), 97-108.

Judd, L. L., Paulus, M. J., Schettler, P. J., Akiskal, H. S., Endicott, J., Leon, A. C., Maser, J. D., Mueller, T., Solomon, D. A. & Keller, M. B. (2000). Does incomplete recovery from first lifetime major depressive episode herald a chronic course of illness? *The American Journal of Psychiatry, 157* (9), 1501-1504.

*Junkert-Tress, B., Schnierda, U., Hartkamp, N., Schmitz, N. & Tress, W. (2001). Effects of short-term dynamic psychotherapy for neurotic, somatoform, and personality disorders: A prospective 1-year follow-up study. *Psychotherapy Research, 11* (2), 187-200.

*Junkert-Tress, B., Tress, W., Scheibe, G., Hartkamp, N., Maus, J., Hildenbrand, G., Schmitz, N. & Franz, M. (1999). Das Düsseldorfer Kurzzeitpsychotherapie-Projekt (DKZP). *Psychotherapie Psychosomatik Medizinische Psychologie, 49* (5), 142-152.

Kächele, H. (2010). Therapie-Manual: Forschungsmethode und/oder Praxisrealität? *Zeitschrift für Individualpsychologie, 35* (3), 239-248.

Kächele, H. (2013). Manualization as tool in psychodynamic psychotherapy research and clinical practice - commentary on six studies. *Psychoanalytic Inquiry, 33* (6), 626-630.

Kantrowitz, J. L., Katz, A. L. & Paolitto, F. (1990a). Followup of psychoanalysis five to ten years after termination: I. Stability of change. *Journal of the American Psychoanalytic Association, 38* (2), 471-496.

Kantrowitz, J. L., Katz, A. L. & Paolitto, F. (1990b). Followup of psychoanalysis five to ten years after termination: II. Development of the self-analytic function. *Journal of the American Psychoanalytic Association, 38* (3), 637-654.

Kantrowitz, J. L., Katz, A. L. & Paolitto, F. (1990c). Followup of psychoanalysis five to ten years after termination: III. The relation between the resolution of the transference and the patient-analyst match. *Journal of the American Psychoanalytic Association, 38* (3), 655-678.

Karel, M. J. & Hinrichsen, G. (2000). Treatment of depression in late life: Psychotherapeutic interventions. *Clinical Psychology Review, 20* (6), 707-729.

Kazdin, A. E. (1994). Methodology, design, and evaluation in psychotherapy research. In A. E. Bergin & S. L. Garfield (Hrsg.), *Handbook of psychotherapy and behavior change* (4th ed.) (S. 19–71). Oxford, England: John Wiley & Sons.

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63* (3), 146-159.

Keller, M. B. (1999). The long-term treatment of depression. *Journal of Clinical Psychiatry, 60* (Suppl 17), 41-45.

Keller, M. B., Lavori, P. W., Lewis, C. E. & Klerman, G. L. (1983). Predictors of relapse in major depressive disorder. *JAMA: Journal of the American Medical Association, 250* (24), 3299-3304.

Keller, M. B., Shapiro, R. W., Lavori, P. W. & Wolfe, N. (1982a). Recovery in major depressive disorder: analysis with the life table and regression models. *Archives of General Psychiatry, 39* (8), 905-910.

Keller, M. B., Shapiro, R. W., Lavori, P. W. & Wolfe, N. (1982b). Relapse in major depressive disorder: analysis with the life table. *Archives of General Psychiatry, 39* (8), 911-915.

*Keller, W., Westhoff, G., Dilg, R., Rohner, R., Studt, H. H. & Arbeitsgruppe empirische Psychotherapieforschung in der Analytischen Psychologie. (2001). Wirksamkeit und Inanspruchnahme von Krankenkassenleistungen bei Langzeitanalysen: Ergebnisse einer empirischen Follow-up-Studie zur Effektivität der (Jungianischen) Psychoanalyse und Psychotherapie. *Analytische Psychologie, 32* (125), 202-229.

Kellett, S., Clarke, S. & Matthews, L. (2007). Delivering group psychoeducational CBT in Primary Care: Comparing outcomes with individual CBT and individual psychodynamic-interpersonal psychotherapy. *British Journal of Clinical Psychology, 46* (2), 211-222.

Kendall, P. C., Holmbeck, G. & Verduin, T. (2004). Methodology, Design, and Evaluation in Psychotherapy Research. In M. J. Lambert (Hrsg.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (5th ed.) (S. 16–43). New York: John Wiley & Sons Inc.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R. & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67* (3), 285-299.

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* (6), 443-453.

*Klar, F. J. (2005). Wirksamkeit individualpsychologisch-psychoanalytischer Psychotherapie. *Zeitschrift für Individualpsychologie, 30* (1), 28-50.

Klauer, K. J. (1987). *Kriteriumsorientierte Tests: Lehrbuch der Theorie und Praxis lehrzielorientierten Messens*. Göttingen: Hogrefe.

Klein, D. N., Schwartz, J. E., Rose, S. & Leader, J. B. (2000). Five-year course and outcome of dysthymic disorder: A prospective, naturalistic follow-up study. *The American Journal of Psychiatry, 157* (6), 931-939.

Kleist, P. (2006). Zehn Anforderungen an therapeutische Äquivalenzstudien. *Schweizerisches Medizin-Forum, 6* (37), 814-819.

Kleist, P. (2009). Das Intention-to-Treat-Prinzip. *Schweizerisches Medizin-Forum, 9* (25), 450-453.

Klemmert, H. (2004). *Äquivalenz- und Effekttests in der psychologischen Forschung* (1. Aufl.). Frankfurt a. M.: Peter Lang.

*Knekt, P. & Lindfors, O. (Hrsg.) (2004). *A randomized trial of the effect of four forms of psychotherapy on depressive and anxiety disorders: Design, methods, and results on the effectiveness of short-term psychodynamic psychotherapy and solution-focused therapy during a one-year follow-up.* Studies in social security and health 77. Helsinki: Edita Prima Ltd.

*Knekt, P., Lindfors, O., Härkänen, T., Virtala, E., Marttunen, M., Renlund, C., Kaipainen, M., Laaksonen, M. A. & Välikoski, M. (2008). Randomized trial on the effectiveness of

long- and short-term psychodynamic psychotherapy and solution-focused therapy on psychiatric symptoms during a 3-year follow-up. *Psychological Medicine, 38* (5), 689-703.

*Knekt, P., Lindfors, O., Laaksonen, M. A., Haaramo, P., Järvikoski, A. & Raitasalo, R. (2008). Effectiveness of short-term and long-term psychotherapy on work ability and functional capacity--A randomized clinical trial on depressive and anxiety disorders. *Journal of Affective Disorders, 107* (1), 95-106.

Köbberling, J. & Wehner, M. (2000). Alternativen zur evidenzbasierten Medizin (EBM). *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 94,* 246-248.

Kopta, S. M., Howard, K. I., Lowry, J. L. & Beutler, L. E. (1994). Patterns of symptomatic recovery in psychotherapy. *Journal of Consulting and Clinical Psychology, 62* (5), 1009-1016.

Koss, M. P. & Shiang, J. (1994). Research on brief psychotherapy. In A. E. Bergin & S. L. Garfield (Hrsg.), *Handbook of psychotherapy and behavior change* (4th ed.) (S. 664–700). Oxford, England: John Wiley & Sons.

Kramer, G. P., Bernstein, D. A. & Phares, V. (2009). *Introduction to clinical psychology* (7th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Kriz, J. (2007). Wie lässt sich die Wirksamkeit von Verfahren X wissenschaftlich begründen? Versuch der Präzisierung einer methodologischen Hintergrunds-Kontroverse. *Psychotherapeutenjournal, 6* (3), 258-261.

Kriz, J. (2008). Vermessene Wissenschaftlichkeit: Kritische Aspekte und bedenkliche Tendenzen des Methodenpapiers. *Psychotherapeutenjournal, 7* (2), 117-119.

Kröner-Herwig, B. (2004). *Die Wirksamkeit von Verhaltenstherapie bei psychischen Störungen von Erwachsenen sowie Kindern und Jugendlichen: Expertise zur empirischen Evidenz des Psychotherapieverfahrens Verhaltenstherapie*. Tübingen: DGVT-Verlag.

*Kurzweil, S. (2008). Relational-developmental group therapy for postnatal depression. *International Journal of Group Psychotherapy, 58* (1), 17-34.

Lambert, M. J. (2013a). The Efficacy and Effectiveness of Psychotherapy. In M. J. Lambert (Hrsg.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (6th ed.) (S. 169–218). Hoboken, NJ: Wiley.

Lambert, M. J. (Hrsg.) (2013b). *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (6th ed.). Hoboken, NJ: Wiley.

Lambert, M. J. & Ogles, B. M. (2004). The Efficacy and Effectiveness of Psychotherapy. In M. J. Lambert (Hrsg.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (5th ed) (S. 139–193). New York: John Wiley & Sons Inc.

Lange, S., Bender, R. & Ziegler, A. (2007). Äquivalenzstudien und Nicht-Unterlegenheitsstudien. *Deutsche Medizinische Wochenschrift (1946), 132* (Suppl 1), e53.

Langsley, D. G., Flomenhaft, K. & Machotka, P. (1969). Followup evaluation of family crisis therapy. *American Journal of Orthopsychiatry, 39* (5), 753-759.

Langsley, D. G., Machotka, P. & Flomenhaft, K. (1971). Avoiding mental hospital admission: A follow-up study. *The American Journal of Psychiatry, 127* (10), 1391-1394.

Lau, M. & Kristensen, E. (2007). Outcome of systemic and analytic group psychotherapy for adult women with history of intrafamilial childhood sexual abuse: A randomized controlled study. *Acta Psychiatrica Scandinavica, 116* (2), 96-104.

Laux, G. (2008a). Depressive Störungen. In H.-J. Möller, G. Laux & H.-P. Kapfhammer (Hrsg.), *Psychiatrie und Psychotherapie* (3 Aufl.) (S. 400–470). Heidelberg: Springer Medizin Verlag.

Laux, G. (2008b). Bipolare affektive Störungen. In H.-J. Möller, G. Laux & H.-P. Kapfhammer (Hrsg.), *Psychiatrie und Psychotherapie* (3 Aufl.) (S. 472–498). Heidelberg: Springer Medizin Verlag.

*Lazar, A., Sandell, R. & Grant, J. (2006). Do psychoanalytic treatments have positive effects on health and health care utilization? Further Findings of the Stockholm Outcome of Psychotherapy and Psychoanalysis Project (STOPPP). *Psychotherapy Research, 16* (1), 51-66.

Läzer, K. L., Sonntag, M., Drazek, R., Jaeschke, R.-I. & Hogreve, C. (2010). *Einführung in die systematische Literaturrecherche mit den Datenbanken "PsycINFO", "Pubmed" und "PEP - Psychoanalytic Electronic Publishing" sowie in das Literaturverwaltungsprogramm "Citavi": Ein Tutorial für Studierende der Fächer Psychologie, Pädagogik, Psychoanalyse und Medizin,* Universität Kassel, Institut für Psychoanalyse. Verfügbar unter: http://www.uni-kassel.de/fb01/uploads/media/Tutorial_Literaturrecherche_30.4.2010_01.pdf [15.9.2014].

*Lehto, S., Tolmunen, T., Joensuu, M., Saarinen, P. I., Vanninen, R., Ahola, P., Tiihonen, J., Kuikka, J. & Lehtonen, J. (2006). Midbrain binding of [^{123}I]nor-β-CIT in atypical depression. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 30* (7), 1251-1255.

*Lehto, S. M., Tolmunen, T., Joensuu, M., Saarinen, P. I., Valkonen-Korhonen, M., Vanninen, R., Ahola, P., Tiihonen, J., Kuikka, J. & Lehtonen, J. (2008). Changes in midbrain serotonin transporter availability in atypically depressed subjects after oneyear of psychotherapy. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 32* (1), 229-237.

*Lehto, S. M., Tolmunen, T., Kuikka, J., Valkonen-Korhonen, M., Joensuu, M., Saarinen, P. I., Vanninen, R., Ahola, P., Tiihonen, J. & Lehtonen, J. (2008). Midbrain serotonin and

striatum dopamine transporter binding in double depression: A one-year follow-up study. *Neuroscience Letters, 441* (3), 291-295.

Leibing, E., Rabung, S. & Leichsenring, F. (2005). Ist psychodynamische Kurztherapie eine wirksame Behandlungsform bei psychischen Störungen? *Forum der Psychoanalyse, 21* (4), 371-379.

Leichsenring, F. (2001). Comparative effects of short-term psychodynamic psychotherapy and cognitive-behavioral therapy in depression: A meta-analytic approach. *Clinical Psychology Review, 21* (3), 401-419.

Leichsenring, F. (2002). Zur Wirksamkeit tiefenpsychologisch fundierter und psychodynamischer Therapie. Eine Übersicht unter Berücksichtigung von Kriterien der Evidence-Based Medicine. *Zeitschrift für Psychosomatische Medizin und Psychotherapie, 48* (2), 139-162.

Leichsenring, F. (2004a). "Empirically supported treatments": Wissenschaftstheoretische und methodische Aspekte kontrollierter vs. naturalistischer Studien. *Zeitschrift für Klinische Psychologie, Psychiatrie und Psychotherapie, 52* (3), 209-222.

Leichsenring, F. (2004b). Randomized controlled versus naturalistic studies: A new research agenda. *Bulletin of the Menninger Clinic, 68* (2), 137-151.

Leichsenring, F. (2005). Are psychodynamic and psychoanalytic therapies effective? A review of empirical data. *The International Journal of Psychoanalysis, 86* (3), 841-868.

Leichsenring, F. (2006). A review of meta-analyses of outcome studies of psychodynamic therapy. In Alliance of Psychodynamic Organizations (Hrsg.), *Psychodynamic Diagnostic Manual (PDM). A collaborative effort of the American Psychoanalytic Association, International Psychoanalytical Association, Division of Psychoanalysis (39) of the American Psychological Association, American Academy of Psychoanalysis and Dynamic Psychia-*

try, National Membership Committee on Psychoanalysis in Clinical Social Work (S. 819–837). Chicago, IL, US: Independent Publishers Group.

Leichsenring, F. (2007). Zur Frage empirisch bewährter Therapie: Befunde zur psychodynamischen Therapie. *Zeitschrift für Psychotraumatologie, Psychotherapiewissenschaft, Psychologische Medizin, 5* (2), 25-37.

Leichsenring, F. (2008). Zum Methodenpapier des Wissenschaftlichen Beirats Psychotherapie. *Psychotherapeutenjournal, 7* (2), 119-120.

Leichsenring, F. (2009a). Applications of psychodynamic psychotherapy to specific disorders: Efficacy and indications. In G. O. Gabbard (Hrsg.), *Textbook of psychotherapeutic treatments* (S. 97–132). Arlington, VA US: American Psychiatric Publishing, Inc.

Leichsenring, F. (2009b). Psychodynamic psychotherapy: A review of efficacy and effectiveness studies. In R. A. Levy & J. S. Ablon (Hrsg.), *Handbook of evidence-based psychodynamic psychotherapy: Bridging the gap between science and practice* (S. 3–27). Totowa, NJ US: Humana Press.

Leichsenring, F. (2011). Empirisch fundierten Psychotherapie: Wissenschaftstheoretische Überlegungen und methodische Alternativen. In E. Behnsen, K. Bell, D. Best, H. Gerlach, H. D. Schirmer & R. Schmid (Hrsg.), *Management Handbuch für die psychotherapeutische Praxis* (S. 1670: 1-25). Heidelberg: medhochzwei Verlag GmbH.

Leichsenring, F., Abbass, A. A., Luyten, P., Hilsenroth, M. & Rabung, S. (2013). The emerging evidence for long-term psychodynamic therapy. *Psychodynamic Psychiatry, 41* (3), 361-384.

*Leichsenring, F., Biskup, J., Kreische, R. & Staats, H. (2005). The Göttingen study of psychoanalytic therapy: First results. *The International Journal of Psychoanalysis, 86* (2), 433-455.

Leichsenring, F., Hiller, W., Weissberg, M. & Leibing, E. (2006). Cognitive-Behavioral Therapy and Psychodynamic Psychotherapy: Techniques, Efficacy, and Indications. *American Journal of Psychotherapy, 60* (3), 233-259.

*Leichsenring, F., Kreische, R., Biskup, J., Rudolf, G., Jakobsen, T. & Staats, H. (2008). Die Göttinger Psychotherapiestudie - Ergebnisse analytischer Langzeitpsychotherapie bei depressiven Störungen, Angststörungen, Zwangsstörungen, somatoformen Störungen und Persönlichkeitsstörungen. *Forum der Psychoanalyse, 24* (2), 193-204.

Leichsenring, F. & Leibing, E. (2007). Psychodynamic psychotherapy: A systematic review of techniques, indications and empirical evidence. *Psychology and Psychotherapy: Theory, Research and Practice, 80* (2), 217-228.

Leichsenring, F. & Rabung, S. (2006). Change norms: A complementary approach to the issue of control groups in psychotherapy outcome research. *Psychotherapy Research, 16* (5), 594-605.

Leichsenring, F. & Rabung, S. (2008). Effectiveness of long-term psychodynamic psychotherapy: A meta-analysis. *JAMA: Journal of the American Medical Association, 300* (13), 1551-1565.

Leichsenring, F. & Rabung, S. (2009). Zur Wirksamkeit psychodynamischer Langzeittherapie bei komplexen psychischen Störungen. *Der Nervenarzt, 80* (11), 1343-1349.

Leichsenring, F. & Rabung, S. (2011). Long-term psychodynamic psychotherapy in complex mental disorders: Update of a meta-analysis. *The British Journal of Psychiatry, 199* (1), 15-22.

Leichsenring, F. & Rabung, S. (2013). Zur Kontroverse um die Wirksamkeit psychodynamischer Therapie. *Zeitschrift für Psychosomatische Medizin und Psychotherapie, 59* (1), 13-32.

Leichsenring, F., Rabung, S. & Leibing, E. (2004). The efficacy of short-term psychodynamic psychotherapy in specific psychiatric disorders. A meta-analysis. *Archives of General Psychiatry, 61,* 1208-1216.

Leichsenring, F., Salzer, S., Beutel, M. E., Consbruch, K. von, Herpertz, S., Hiller, W., Hoyer, J., Hüsing, J., Irle, E., Joraschky, P., Konnopka, A., König, H.-H., Liz, T. de, Nolting, B., Pöhlmann, K., Ruhleder, M., Schauenburg, H., Stangier, U., Strauß, B., Subic-Wrana, C., Vormfelde, S. V., Weniger, G., Willutzki, U., Wiltink, J. & Leibing, E. (2009). SOPHO-NET - Forschungsverbund zur Psychotherapie der Sozialen Phobie. *Psychotherapie, Psychosomatik, Medizinische Psychologie, 59* (3-4), 117-123.

Leichsenring, F., Salzer, S., Hilsenroth, M. J., Leibing, E., Leweke, F. & Rabung, S. (2011). Treatment integrity: An unresolved issue in psychotherapy research. *Current Psychiatry Reviews, 7* (4), 313-321.

Leuzinger-Bohleber, M., Bahrke, U., Beutel, M. E., Deserno, H., Edinger, J., Fiedler, G., Haselbacher, A., Hautzinger, M., Kallenbach, L., Keller, W., Negele, A., Pfenning-Meerkötter, N., Prestele, H., Strecker-von Kannen, T., Stuhr, U. & Will, A. (2010). Psychoanalytische und kognitiv-verhaltenstherapeutische Langzeittherapien bei chronischer Depression: Die LAC-Depressionsstudie. *Psyche: Zeitschrift für Psychoanalyse und ihre Anwendungen, 64* (9-10), 782-832.

Leuzinger-Bohleber, M. & Beutel, M. E. (2009). *Langzeittherapie bei chronischen Depressionen (LAC)*. Verfügbar unter: http://www.sfi-frankfurt.de/forschung/forschungsfeld-2/depressionsstudie/projektbeschreibung/projektdaten.html [19.1.2009].

*Leuzinger-Bohleber, M., Stuhr, U., Rüger, B. & Beutel, M. E. (2001). Langzeitwirkungen von Psychoanalysen und Psychotherapien: Eine multiperspektivische, repräsentative Ka-

tamnesestudie. *Psyche: Zeitschrift für Psychoanalyse und ihre Anwendungen, 55* (3), 193-276.

*Leuzinger-Bohleber, M., Stuhr, U., Rüger, B. & Beutel, M. E. (2003). How to study the 'quality of psychoanalytic treatments' and their long-term effects on patients' well-being: A representative, multi-perspective follow-up study. *The International Journal of Psychoanalysis, 84* (2), 263-290.

Lewis, A. J., Dennerstein, M. & Gibbs, P. M. (2008). Short-term psychodynamic psychotherapy: Review of recent process and outcome studies. *Australian and New Zealand Journal of Psychiatry, 42* (6), 445-455.

Liberman, B. L., Frank, J. D., Hoehn-Saric, R., Stone, A. R., Imber, S. D. & Pande, S. K. (1972). Patterns of change in treated psychoneurotic patients: A five-year follow-up investigation of the systematic preparation of patients for psychotherapy. *Journal of Consulting and Clinical Psychology, 38* (1), 36-41.

Linden, M. (2013). How to define, find and classify side effects in psychotherapy: From unwanted events to adverse treatment reactions. *Clinical Psychology and Psychotherapy, 20* (4), 286-296.

Lindgren, A., Barber, J. P. & Sandahl, C. (2008). Alliance to the group-as-a-whole as a predictor of outcome in psychodynamic group therapy. *International Journal of Group Psychotherapy, 58* (2), 163-184.

*Lindgren, A., Werbart, A. & Philips, B. (2010). Long-term outcome and post-treatment effects of psychoanalytic psychotherapy with young adults. *Psychology and Psychotherapy: Theory, Research and Practice, 83* (1), 27-43.

Lipsey, M. W. & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks, CA, US: Sage Publications, Inc.

Loew, T. H., Richter, R., Calatzis, A. & Krause, S. (2002). Wirksamkeitsstudien zur Psychodynamischen Psychotherapie: Studien, die nicht berücksichtigt wurden. *PDP Psychodynamische Psychotherapie: Forum der tiefenpsychologisch fundierten Psychotherapie, 1* (2), 93-107.

Löfholm, C. A., Brännström, L., Olsson, M. & Hansson, K. (2013). Treatment-as-usual in effectiveness studies: What is it and does it matter? *International Journal of Social Welfare, 22* (1), 25-34.

Lorentzen, S., Bøgwald, K. P. & Høglend, P. (2002). Change during and after long-term analytic group psychotherapy. *International Journal of Group Psychotherapy, 52* (3), 419-429.

*Lotz, M. & Jensen, H. H. (2006). Focus in psychodynamic group therapy: An empirical study of dynamic focus, diagnoses, and symptomatic outcome. *Nordic Psychology, 58* (3), 248-261.

*Luborsky, L., Stuart, J., Friedman, S., Diguer, L., Seligman, D. A., Bucci, W., Pulver, S., Krause, E. D., Ermold, J., Davison, W. T., Woody, G. & Mergenthaler, E. (2001). The Penn Psychoanalytic Treatment Collection: A set of complete and recorded psychoanalyses as a research resource. *Journal of the American Psychoanalytic Association, 49* (1), 217-234.

*Lundblad, S. (2003). Depressed women in psychotherapy: the nature and persistence of change. *International Journal of Psychotherapy, 8* (1), 53-63.

Maat, S. de, Dekker, J., Schoevers, R. & Jonghe, F. de. (2007). The effectiveness of long-term psychotherapy: Methodological research issues. *Psychotherapy Research, 17* (1), 59-65.

Maat, S. de, Dekker, J., Schoevers, R., van Aalst, G., Gijsbers-van Wijk, C., Hendriksen, M., Kool, S., Peen, J., Van, R. & Jonghe, F. de. (2008). Short psychodynamic supportive psy-

chotherapy, antidepressants, and their combination in the treatment of major depression: A mega-analysis based on three randomized clinical trials. *Depression and Anxiety, 25* (7), 565-574.

Maat, S. de, Jonghe, F. de, Schoevers, R. & Dekker, J. (2009). The effectiveness of long–term psychoanalytic therapy: A systematic review of empirical studies. *Harvard Review of Psychiatry, 17* (1), 1-23.

Maat, S. de, Philipszoon, F., Schoevers, R., Dekker, J. & Jonghe, F. de. (2007). Costs and benefits of long-term psychoanalytic therapy: Changes in health care use and work impairment. *Harvard Review of Psychiatry, 15* (6), 289-300.

*Maina, G., Forner, F. & Bogetto, F. (2005). Randomized Controlled Trial Comparing Brief Dynamic and Supportive Therapy with Waiting List Condition in Minor Depressive Disorders. *Psychotherapy and Psychosomatics, 74* (1), 43-50.

Maina, G., Rosso, G. & Bogetto, F. (2009). Brief dynamic therapy combined with pharmacotherapy in the treatment of major depressive disorder: Long-term results. *Journal of Affective Disorders, 114* (1-3), 200-207.

Maina, G., Rosso, G., Crespi, C. & Bogetto, F. (2007). Combined brief dynamic therapy and pharmacotherapy in the treatment of major depressive disorder: A pilot study. *Psychotherapy and Psychosomatics, 76* (5), 298-305.

*Małyszczak, K., Pawłowski, T., Pyszel, K., Tichonow, M. & Kiejna, A. (2006). Assessment of changes in the severity of depressive and anxiety symptoms using a tripartite model of anxiety and depression. *Archives of Psychiatry and Psychotherapy, 8* (3), 15-21.

*Mattanza, G., Jakobsen, T. & Hurt, J. (2005). Jung'sche Psychotherapie ist effizient. In G. Mattanza, I. Meier & M. Schlegel (Hrsg.), *Seele und Forschung. Ein Brückenschlag in der Psychotherapie* (S. 38–82). Basel: Karger.

Matzat, J. (2014). Gegen eine vorschriftsmäßige Unterdosierung. *Projekt Psychotherapie. Das Magazin des Bundeverbandes der Vertragspsychotherapeuten e.V.* (2), 19.

Mayer, B. (2010). *Fehlende Werte in klinischen Verlaufsstudien - Der Umgang mit Studienabbrechern*. Dissertation, Universität Ulm.

McGraw, K. O. & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1* (1), 30-46.

McLean, P. D. & Hakstian, A. R. (1979). Clinical Depression: Comparative efficacy of outpatient treatments. *Journal of Consulting and Clinical Psychology, 47* (5), 818-836.

McLean, P. D. & Hakstian, A. R. (1990). Relative endurance of unipolar depression treatment effects: Longitudinal follow-up. *Journal of Consulting and Clinical Psychology, 58* (4), 482-488.

Mertens, W. (2007). *Stellungnahme der DGPT zum Entwurf der Neufassung der Verfahrensregeln zur Beurteilung der wissenschaftlichen Anerkennung von Methoden und Verfahren der Psychotherapie des Wissenschaftlichen Beirates Psychotherapie („Methodenpapier"),* DGPT. Verfügbar unter: http://www.dgpt.de/dokumente/Stellungnahme%20der%20DGPT%20zum%20Methodenpapier.pdf [30.8.2010].

Messer, S. B. & Kaplan, A. H. (2004). Outcomes and factors related to efficacy of brief psychodynamic therapy. In D. P. Charman (Hrsg.), *Core processes in brief psychodynamic psychotherapy: Advancing effective practice* (S. 103–118). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Metzner, K. & Schlösser, A.-M. (2014). Die Erstbegutachtung darf nicht wegfallen. *Projekt Psychotherapie. Das Magazin des Bundeverbandes der Vertragspsychotherapeuten e.V.* (2), 21.

Mittag, W. & Hager, W. (2000). Ein Rahmenkonzept zur Evaluation psychologischer Interventionsmaßnahmen. In W. Hager, J.-L. Patry & H. Brezing (Hrsg.), *Evaluation psychologischer Interventionsmaßnahmen. Standards und Kriterien* (S. 102–128). Bern: Huber.

Möller, H.-J., Laux, G. & Kapfhammer, H.-P. (Hrsg.) (2005). *Psychiatrie und Psychotherapie* (2. Aufl.). Heidelberg: Springer Medizin Verlag.

Multmeier, J. (2014). Ambulante psychotherapeutische Versorgung in Deutschland - eine Kohortenbetrachtung der KBV: Die Studie der Kassenärztlichen Bundesvereinigung in ungekürzter Fassung. *Projekt Psychotherapie. Das Magazin des Bundeverbandes der Vertragspsychotherapeuten e.V.* (2), 12-22.

Munder, T., Brütsch, O., Leonhart, R., Gerger, H. & Barth, J. (2013). Researcher allegiance in psychotherapy outcome research: an overview of reviews. *Clinical Psychology Review, 33* (4), 501-511.

Nierenberg, A. A., Husain, M. M., Trivedi, M. H., Fava, M., Warden, D., Wisniewski, S. R., Miyahara, S. & Rush, A. J. (2010). Residual symptoms after remission of major depressive disorder with citalopram and risk of relapse: A STAR*D report. *Psychological Medicine: A Journal of Research in Psychiatry and the Allied Sciences, 40* (1), 41-50.

Orlinsky, D. E., Rønnestad, H. M. & Willutzki, U. (2004). Fifty Years of Psychotherapy Process-Outcome Research: Continuity and Change. In M. J. Lambert (Hrsg.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (5th ed.) (S. 307–389). New York: John Wiley & Sons Inc.

Paivio, S. C. & Nieuwenhuis, J. A. (2001). Efficacy of Emotion Focused Therapy for Adult Survivors of Child Abuse: A Preliminary Study. *Journal of Traumatic Stress, 14* (1), 115-133.

*Paley, G., Cahill, J., Barkham, M., Shapiro, D., Jones, J., Patrick, S. & Reid, E. (2008). The effectiveness of psychodynamic-interpersonal therapy (PIT) in routine clinical practice: A benchmarking comparison. *Psychology and Psychotherapy: Theory, Research and Practice, 81* (2), 157-175.

Pant, H. A. (1998). *HIV-Infektionen bei iv Drogenkonsumenten: sozialepidemiologische Befunde zur Ätiologie durch Metaanalysen und Primärdatenanalysen*. Dissertation, Freie Universität Berlin.

Paykel, E. S., Scott, J., Teasdale, J. D., Johnson, A. L., Garland, A., Moore, R., Jenaway, A., Cornwall, P. L., Hayhurst, H., Abbott, R. & Pope, M. (1999). Prevention of relapse in residual depression by cognitive therapy: A controlled trial. *Archives of General Psychiatry, 56* (9), 829-835.

Perry, J. C., Banon, E. & Ianni, F. (1999). Effectiveness of psychotherapy for personality disorders. *The American Journal of Psychiatry, 156* (9), 1312-1321.

*Perry, J. C. & Bond, M. (2009). The sequence of recovery in long-term dynamic psychotherapy. *Journal of Nervous and Mental Disease, 197* (12), 930-937.

Petersen, P. (2003). Wissenschaftlicher Beirat Psychotherapie: Starre Forschungsprinzipien. *Deutsches Ärzteblatt PP* (2), 61.

Petticrew, M. & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Malden: Blackwell Publishing.

*Philips, B., Wennberg, P., Werbart, A. & Schubert, J. (2006). Young adults in psychoanalytic psychotherapy: Patient characteristics and therapy outcome. *Psychology and Psychotherapy: Theory, Research and Practice, 79* (1), 89-106.

*Philips, B., Wennberg, P. & Werbart, A. (2007). Ideas of cure as a predictor of premature termination, early alliance and outcome in psychoanalytic psychotherapy. *Psychology and Psychotherapy: Theory, Research and Practice, 80* (2), 229-245.

*Philips, B., Werbart, A., Wennberg, P. & Schubert, J. (2007). Young adults' ideas of cure prior to psychoanalytic psychotherapy. *Journal of Clinical Psychology, 63* (3), 213-232.

Pinquart, M., Duberstein, P. R. & Lyness, J. M. (2007). Effects of psychotherapy and other behavioral interventions on clinically depressed older adults: A meta-analysis. *Aging & Mental Health, 11* (6), 645-657.

Pinquart, M. & Sörensen, S. (2001). How effective are psychotherapeutic and other psychosocial interventions with older adults? A meta-analysis. *Journal of Mental Health and Aging, 7* (2), 207-243.

*Piper, W. E., McCallum, M., Joyce, A. S., Rosie, J. S. & Ogrodniczuk, J. S. (2001). Patient personality and time-limited group psychotherapy for complicated grief. *International Journal of Group Psychotherapy, 51* (4), 525-552.

Plänkers, T. (1986). Zum Verhältnis von Psychoanalyse und Systemtheorie. *Psyche: Zeitschrift für Psychoanalyse und ihre Anwendungen, 40* (8), 678-708.

Protz, J., Kächele, H. & Taubner, S. (2012). Die Ambivalenz mit der Therapieforschung. Beweggründe und Erfahrungen von Psychoanalytikern. *Forum der Psychoanalyse, 28* (1), 67-88.

Psychotherapeutenjournal (2008). Fragen zur Zukunft der Psychotherapie vor dem Hintergrund der neuen „Verfahrensregeln zur Beurteilung der wissenschaftlichen Anerkennung von Methoden und Verfahren der Psychotherapie“ (Methodenpapier) des Wissenschaftlichen Beirats Psychotherapie (WBP): Ein Interview des Psychotherapeutenjournals mit

Prof. Dr. Dietmar Schulte, dem Vorsitzenden des WBP und Prof. Dr. Gerd Rudolf, dem stellvertretendem Vorsitzenden des WBP. *Psychotherapeutenjournal, 7* (2), 111-116.

Puschner, B., Kraft, S., Kächele, H. & Kordy, H. (2007). Course of improvement over 2 years in psychoanalytic and psychodynamic outpatient psychotherapy. *Psychology and Psychotherapy: Theory, Research and Practice, 80* (1), 51-68.

Rad, M. von, Senf, W. & Bräutigam, W. (1998). Psychotherapie und Psychoanalyse in der Krankenversorgung: Ergebnisse des Heidelberger Katamnese-Projektes. *Psychotherapie Psychosomatik Medizinische Psychologie, 48* (3), 88-100.

Ratzek, M. & Hauenschild, L. von (2011a). *Kodierregeln 1.4: Kurzkodierbogen und Kodierbogen zum Methodenpapier des Wissenschaftlichen Beirats Psychotherapie nach § 11 PsychThG (Version 2.7 / 2.8) - Zur Bewertung von Wirksamkeitsstudien eines psychotherapeutischen Verfahrens bzw. einer psychotherapeutischen Methode.* Schriftenreihe des Instituts für Prävention und psychosoziale Gesundheitsforschung: Bd. Nr. 03/P11. Berlin: Freie Universität Berlin.

Ratzek, M. & Hauenschild, L. von (2011b). *Kurzkodierbogen zur Erhebung allgemeiner Studiencharakteristika und zur methodologischen Spezifizierung von Wirksamkeitsstudien hinsichtlich naturalistischer und experimenteller Studiendesigneigenschaften.* Schriftenreihe des Instituts für Prävention und psychosoziale Gesundheitsforschung: Bd. Nr. 02/P11. Berlin: Freie Universität Berlin.

Ratzek, M. & Hauenschild., L. von (2011c). *Kodierbogen zur Beurteilung von psychometrischen Eigenschaften (Reliabilität und Validität) diagnostischer Selbst- und Fremdbeurteilungsverfahren.* Schriftenreihe des Instituts für Prävention und psychosoziale Gesundheitsforschung: Bd. Nr. 04/P11. Berlin: Freie Universität Berlin.

Ratzek, M. & Hauenschild, L. von (in Vorb.). *Überblick über Reliabilitäts- und Validitätsbefunde von klinischen und außerklinischen Selbst- und Fremdbeurteilungsverfahren*.

Reed, J. G. & Baxter, P. M. (2009). Using reference databases. In H. Cooper, L. V. Hedges & J. C. Valentine (Hrsg.), *The Handbook of Research Synthesis and Meta-analysis* (2nd ed.) (S. 73–101). New York, NY, US: Russell Sage Foundation.

Reimer, C. & Rüger, U. (Hrsg.) (2006). *Psychodynamische Psychotherapien: Lehrbuch der tiefenpsychologisch fundierten Psychotherapieverfahren* (3. Aufl.). Heidelberg: Springer.

Revenstorf, D. (2005). Das Kuckucksei - Über das pharmakologische Modell in der Psychotherapieforschung. *Psychotherapie in Psychiatrie, Psychotherapeutischer Medizin und Klinischer Psychologie, 10* (1), 22-31.

Richter, R. (2014). Pauschale Verkürzungen der Behandlungskontingente mobilisieren keine Effizienzreserven mehr. *Projekt Psychotherapie. Das Magazin des Bundeverbandes der Vertragspsychotherapeuten e.V.* (2), 23.

Richter, R., Loew, T. H., Calatzis, A. & Krause, S. (2002). Kontrollierte Wirksamkeitsstudien zur Psychodynamischen Psychotherapie: Tiefenpsychologisch fundierte Psychotherapie. *PDP Psychodynamische Psychotherapie: Forum der tiefenpsychologisch fundierten Psychotherapie, 1* (2), 19-36.

Rief, W. & Hofmann, S. G. (2009). Die Psychoanalyse soll gerettet werden. Mit allen Mitteln? *Der Nervenarzt, 80* (5), 593-597.

Rogers, J. L., Howard, K. I. & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113* (3), 553-565.

Röhmel, J., Hauschke, D., Koch, A. & Pigeot, I. (2005). Biometrische Verfahren zum Wirksamkeitsnachweis im Zulassungsverfahren: Nicht-Unterlegenheit in klinischen Studien. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz, 48* (5), 562-571.

Rössler, W. (2004). Supportive Psychotherapie. In W. Rössler (Hrsg.), *Psychiatrische Rehabilitation* (S. 134–145). Berlin: Springer.

Roth, A. & Fonagy, P. (2005). *What works for whom: A critical review of psychotherapy research* (2nd ed.). New York, NY, US: Guilford Publications.

*Roy, C. A., Perry, C. J., Luborsky, L. & Banon, E. (2009). Changes in defensive functioning in completed psychoanalyses: the penn psychoanalytic treatment collection. *Journal of the American Psychoanalytic Association, 57* (2), 399-415.

Rozmarin, E., Muran, J. C., Safran, J., Gorman, B., Nagy, J. & Winston, A. (2008). Subjective and intersubjective analyses of the therapeutic alliance in a brief relational therapy. *American Journal of Psychotherapy, 62* (3), 313-328.

*Rudolf, G., Dilg, R., Grande, T., Jakobsen, T., Keller, W., Krawietz, B., Langer, M., Stehle, S. & Oberbracht, C. (2004). Effektivität und Effizienz psychoanalytischer Langzeittherapie: Die Praxisstudie analytische Langzeitpsychotherapie. In A. Gerlach, A.-M. Schlösser & A. Springer (Hrsg.), *Psychoanalyse des Glaubens* (S. 515–528). Gießen: Psychosozial-Verlag.

Rudolf, G. & Rüger, U. (2006) Analytische Psychotherapie. In C. Reimer & U. Rüger (Hrsg.), *Psychodynamische Psychotherapien. Lehrbuch der tiefenpsychologisch fundierten Psychotherapieverfahren* (S. 39–48). Berlin: Springer.

Rüger, U. & Reimer, C. (2006) Gemeinsame Merkmale und Charakteristika psychodynamischer Psychotherapieverfahren. In C. Reimer & U. Rüger (Hrsg.), *Psychodynamische Psychotherapien. Lehrbuch der tiefenpsychologisch fundierten Psychotherapieverfahren* (S. 1–22). Berlin: Springer.

Rustenbach, S. J. (2003). *Metaanalyse – eine anwendungsorientierte Einführung.* Bern: Huber.

Sackett, D. L., Rosenberg, W. M. C., Gray, M. J. A., Haynes, B. R. & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn´t. *British Medical Journal, 312* (7023), 71-72.

*Salminen, J. K., Karlsson, H., Hietala, J., Kajander, J., Aalto, S., Markkula, J., Rasi-Hakala, H. & Toikka, T. (2008). Short-term psychodynamic psychotherapy and fluoxetine in major depressive disorder: A randomized comparative study. *Psychotherapy and Psychosomatics, 77* (6), 351-357.

Sandell, R. (2001). Can Psychoanalysis Become Empirically Supported? *International Forum of Psychoanalysis, 10* (3/4), 184-190.

*Sandell, R., Blomberg, J. & Lazar, A. (1999). Wiederholte Langzeitkatamnesen von Langzeitpsychotherapien und Psychoanalysen. Erste Ergebnisse des 'Stockholmer Outcome of Psychotherapy (STOP) Project'. *Zeitschrift für Psychosomatische Medizin und Psychoanalyse, 45* (1), 43-56.

Schendera, C. (2008). *Clusteranalyse mit SPSS* (1. Aufl.). München: Oldenbourg, R.

*Schleussner, D. (2005). Wirksamkeit individualpsychologisch-psychoanalytischer Psychotherapie - Eine Langzeitstudie zur Psychotherapieforschung. *Zeitschrift für Individualpsychologie, 30* (1), 51-77.

Schulte, D. (1993). Wie soll Therapieerfolg gemessen werden? *Zeitschrift für Klinische Psychologie, 22* (4), 374-393.

Schwarzer G., Türp J. & Antes G. (2002a). Die Vierfeldertafel (in Diagnosestudien): Sensitivität und Spezifität. *Deutsche Zahnärztliche Zeitschrift, EbM-Splitter, 57* (6), 333-334.

Schwarzer G., Türp J. & Antes G. (2002b). Sensitivität und Spezifität: Auswirkung der Wahl des Trennpunktes. *Deutsche Zahnärztliche Zeitschrift, EbM-Splitter, 57* (8), 446-447.

Schwarzer G., Türp J. & Antes G. (2002c). Nutzen eines diagnostischen Tests in der Praxis: prädiktive Werte. *Deutsche Zahnärztliche Zeitschrift, EbM-Splitter, 57* (10), 573-575.

Schwarzer G., Türp J. & Antes G. (2002d). Wahrscheinlichkeitsverhältnis (Likelihood Ratio) - Alternative zu Sensitivität und Spezifität. *Deutsche Zahnärztliche Zeitschrift, EbM-Splitter, 57* (12), 660-661.

Scogin, F., Welsh, D., Hanson, A., Coates, A. & Stump, J. (2005). Evidence-Based Psychotherapies for Depression in Older Adults. *Clinical Psychology: Science and Practice, 12* (3), 222-237.

Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The Consumer Reports study. *American Psychologist, 50* (12), 965-974.

Seybert, C., Huber, D., Ratzek, M., Zimmermann, J. & Klug, G. (2014). How to Capture Emotional Processing in a Process-Outcome Study of Psychoanalytic, Psychodynamic, and Cognitive-Behavioral Psychotherapies. *Journal of the American Psychoanalytic Association, 62* (3), NP4.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shadish, W. R., Navarro, A. M., Matt, G. E. & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin, 126* (4), 512-529.

Shapiro, D. A., Barkham, M., Stiles, W. B., Hardy, G. E., Rees, A., Reynolds, S. & Startup, M. (2003). Time is of the essence: A selective review of the fall and rise of brief therapy research. *Psychology and Psychotherapy: Theory, Research and Practice, 76* (3), 211-235.

*Sharpe, J., Selley, C., Low, L. & Hall, Z. (2001). Group Analytic Therapy for Male Survivors of Childhood Sexual Abuse. *Group Analysis, 34* (2), 195.

*Shaw, C. M., Margison, F. R., Guthrie, E. A. & Tomenson, B. (2001). Psychodynamic interpersonal therapy by inexperienced therapists in a naturalistic setting: a pilot study. *European Journal of Psychotherapy, Counselling and Health, 4* (1), 87-101.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86* (2), 420-428.

Simpson, S., Corney, R., Fitzgerald, P. & Beecham, J. (2003). A randomized controlled trial to evaluate the effectiveness and cost-effectiveness of psychodynamic counselling for general practice patients with chronic depression. *Psychological Medicine, 33* (2), 229-239.

*Stehle, S., Dilg, R. & Keller, W. (2004). Psychotherapeutische Berufstätigkeit. Ergebnisse der DGPT-Therapeutenerhebung und Schlussfolgerungen für die Qualitätssicherung. In A. Gerlach, A. Springer & A.-M. Schlösser (Hrsg.), *Psychoanalyse des Glaubens* (S. 465–513). Gießen: Psychosozial Verlag.

Stieglitz, R.-D. (2008). *Diagnostik und Klassifikation in der Psychiatrie*. Stuttgart: Kohlhammer.

Stieglitz, R.-D. & Hiller, W. (2013). Definition und Erfassung psychischer Störungen. Bestandsaufnahme. *Psychotherapeut, 58* (3), 237-248.

*Stiles, W. B., Barkham, M., Mellor-Clark, J. & Connell, J. (2008). Effectiveness of cognitive-behavioural, person-centred, and psychodynamic therapies in UK primary-care routine practice: replication in a larger sample. *Psychological Medicine, 38* (5), 677-688.

*Stiles, W. B., Barkham, M., Twigg, E., Mellor-Clark, J. & Cooper, M. (2006). Effectiveness of cognitive-behavioural, person-centred and psychodynamic therapies as practised in UK National Health Service settings. *Psychological Medicine, 36* (4), 555-566.

*Stiles, W. B., Barkham, M., Twigg, E., Mellor-Clark, J. & Cooper, M. (2007). Wirksamkeit Personzentrierter Therapie im Vergleich zu kognitiv-behavioralen und psychodynamischen

Therapien, wie sie im Rahmen des britischen National Health Service praktiziert werden. *Person, 11* (2), 105-113.

Stirman, S. W., DeRubeis, R. J., Crits-Christoph, P. & Rothman, A. (2005). Can the Randomized Controlled Trial Literature Generalize to Nonrandomized Patients? *Journal of Consulting and Clinical Psychology, 73* (1), 127-135.

Strauß, B. & Schumacher, J. (Hrsg.) (2005). *Klinische Interviews und Ratingskalen.* Diagnostik für Klinik und Praxis, Bd. 3. Göttingen: Hogrefe.

*Stuhr, U., Leuzinger-Bohleber, M. & Beutel, M. E. (Hrsg.) (2001). *Langzeit-Psychotherapie: Perspektiven für Therapeuten und Wissenschaftler*. Stuttgart: Kohlhammer.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges & J. C. Valentine (Hrsg.), *The Handbook of Research Synthesis and Meta-analysis* (2nd ed.) (S. 435–452). New York, NY, US: Russell Sage Foundation.

Svartberg, M., Seltzer, M. H., Choi, K. & Stiles, T. C. (2001). Cognitive change before, during, and after short-term dynamic and nondirective psychotherapies: A preliminary growth modeling study. *Psychotherapy Research, 11* (2), 201-219.

Szecsödy, I., Varvin, S., Beenen, F., Stoker, J., Klockars, L. & Amadei, G. (1999). *Multicenter collaboration of research on process and outcome of psychoanalysis. Presentation of AHMOS: Paper presented at the International Psychoanalytic Association Congress*. Santiago.

Task Force on Promotion and Dissemination of Psychological Procedures – American Psychological Association (APA) (1995). Training in and dissemination of empirically validated treatments: Report and recommendations. *The Clinical Psychologist, 48* (1), 3-23.

Taylor, D. (2010). Das Tavistock-Manual der psychoanalytischen Psychotherapie - unter besonderer Berücksichtigung der chronischen Depression. *Psyche: Zeitschrift für Psychoanalyse und ihre Anwendungen, 64* (9-10), 833-861.

Temple, R. & Ellenberg, S. S. (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. *Annals of Internal Medicine, 133* (6), 455-463.

Testkuratorium (2006). TBS-TK: Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. *Report Psychologie* (31), 492-500.

Thompson, L. W., Gallagher, D. & Breckenridge, J. S. (1987). Comparative effectiveness of psychotherapies for depressed elders. *Journal of Consulting and Clinical Psychology, 55* (3), 385-390.

Thyme, K. E., Sundin, E. C., Stahlberg, G., Lindstrom, B., Eklof, H. & Wiberg, B. (2007). The outcome of short-term psychodynamic art therapy compared to short-term psychodynamic verbal therapy for depressed women. *Psychoanalytic Psychotherapy, 21* (3), 250-264.

Timmer, A. & Richter, B. (2008). Systematische Übersichtsarbeiten zu Fragen der Therapie und Prävention: Eine Einführung in Frage und Antwort Teil 2 – Was macht eine gute Übersichtsarbeit aus? *Arzneimitteltherapie, 26* (7), 252-255.

Tschuschke, V. (2005). Die Psychotherapie in Zeiten evidenzbasierter Medizin. Fehlentwicklungen und Korrekturvorschläge. *Psychotherapeutenjournal, 4* (2), 106-115.

*Tschuschke, V. & Anbeh, T. (2008). *Ambulante Gruppenpsychotherapie*. Stuttgart: Schattauer.

*Tschuschke, V., Anbeh, T. & Kiencke, P. (2007). Evaluation of Long-term Analytic Outpatient Group Therapies. *Group Analysis, 40* (1), 140-159.

Tschuschke, V., Crameri, A., Koemeda, M., Schulthess, P., Wyl, A. von & Weber, R. (2009). Psychotherapieforschung - Grundlegende Überlegungen und erste Ergebnisse der naturalistischen Psychotherapie-Studie ambulanter Behandlungen in der Schweiz (PAP-S). *Psychotherapie Forum, 17* (4), 160-176.

Tyrer, P., Thompson, S., Schmidt, U., Jones, V., Knapp, M., Davidson, K., Catalan, J., Airlie, J., Baxter, S., Byford, S., Byrne, G., Cameron, S., Caplan, R., Cooper, S., Ferguson, B., Freeman, C., Frost, S., Godley, J., Greenshields, J. & Henderson, J., Holden, N., Keech, P., Kim, L., Logan, K., Manley, C., MacLeod, A., Murphy, R., Patience, L., Ramsay, L., De Munroz, S., Scott, J., Seivewright, H., Sivakumar, K., Tata, P., Thornton, S., Ukoumunne, O. C. & Wessely, S. (2003). Randomized controlled trial of brief cognitive behaviour therapy versus treatment as usual in recurrent deliberate self-harm: the POPMACT study. *Psychological Medicine: A Journal of Research in Psychiatry and the Allied Sciences, 33* (6), 969-976.

*Van, H. L., Dekker, J., Peen, J., Abraham, R. E. & Schoevers, R. (2009). Predictive Value of Self-Reported and Observer-Rated Defense Style in Depression Treatment. *American Journal of Psychotherapy, 63* (1), 25-39.

*Van, H. L., Hendriksen, M., Schoevers, R. A., Peen, J., Abraham, R. A. & Dekker, J. (2008). Predictive value of object relations for therapeutic alliance and outcome in psychotherapy for depression: An exploratory study. *Journal of Nervous and Mental Disease, 196* (9), 655-662.

Van, H. L., Schoevers, R. A., Kool, S., Hendriksen, M., Peen, J. & Dekker, J. (2008). Does early response predict outcome in psychotherapy and combined therapy for major depression? *Journal of Affective Disorders, 105* (1), 261-265.

*Vitriol, V. G., Ballesteros, S. T., Florenzano, R. U., Weil, K. P. & Benadof, D. F. (2009). Evaluation of an outpatient intervention for women with severe depression and a history of childhood trauma. *Psychiatric Services, 60* (7), 936-942.

Vlastelica, M., Jurcević, S. & Zemunik, T. (2005). Changes of defense mechanisms and personality profile during group analytic treatment. *Collegium Antropologicum, 29* (2), 551-558.

*Wallerstein, R. S. (1986). *Forty-Two Lives in Treatment: A Study of Psychoanalysis and Psychotherapy*. New York: Guilford Press.

*Wallerstein, R. S. (1989). The Psychotherapy Research Project of the Menninger Foundation: an overview. *Journal of Consulting and Clinical Psychology, 57* (2), 195-205.

Wampold, B. E. (2001). *The great psychotherapy debate: Models, methods, and findings*. Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

Westen, D. & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An empirical examination of the status of empirically supported therapies. *Journal of Consulting and Clinical Psychology, 69* (6), 875-899.

Westen, D., Novotny, C. M. & Thompson-Brenner, H. (2004). The Empirical Status of Empirically Supported Psychotherapies: Assumptions, Findings, and Reporting in Controlled Clinical Trials. *Psychological Bulletin, 130* (4), 631-663.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik: Ein Lehrbuch zur Psychologischen Methodenlehre*. Göttingen: Hogrefe.

White, H. D. (1994). Scientific communication and literature retrieval. In H. Cooper & L. V. Hedges (Hrsg.), *The Handbook of Research Rynthesis* (S. 41–55). New York, NY, US: Russell Sage Foundation.

White, H. D. (2009). Scientific communication and literature retrieval. In H. Cooper, L. V. Hedges & J. C. Valentine (Hrsg.), *The Handbook of Research Synthesis and Meta-analysis* (2nd ed.) (S. 51–71). New York, NY, US: Russell Sage Foundation.

Wietersheim, J. von, Scheib, P., Keller, W., Osborn, W., Pritsch, M., Balck, F., Fritzsche, K., Dilg, R. & Schmelz-Schumacher, E. (2001). Die Wirksamkeit psychotherapeutischer Maßnahmen bei Morbus Crohn. Ergebnisse einer randomisierten, multizentrischen Studie. *Psychotherapie, Psychosomatik, Medizinische Psychologie, 51* (1), 2-9.

*Wietersheim, J. von, Wilke, E., Röser, M. & Meder, G. (2002). Die Effektivität der Kathathym-imaginativen Psychotherapie in einer ambulanten Längsschnittstudie. In D. Mattke, G. Hertel, S. Büsing & K. Schreiber-Willnow (Hrsg.), *Störungsspezifische Konzepte und Behandlung in der Psychosomatik* (S. 379–388). Frankfurt a. M.: VAS Verlag für Akademische Schriften.

*Wietersheim, J. von, Wilke, E., Röser, M. & Meder, G. (2003). Ergebnisse der Katathymimaginativen Psychotherapie. Die Effektivität der Katathym-imaginativen Psychotherapie in einer ambulanten Längsschnittstudie. *Psychotherapeut, 48* (3), 173-178.

*Wilczek, A., Weinryb, R. M., Barber, J. P., Åsberg, M. & Gustavsson, J. P. (2004). Change in the Core Conflictual Relationship Theme after Long-Term Dynamic Psychotherapy. *Psychotherapy Research, 14* (1), 107-125.

*Wilczek, A., Weinryb, R. M., Gustavsson, P. J., Barber, J. P., Schubert, J. & Åsberg, M. (1998). Symptoms and character traits in patients selected for long-term psychodynamic psychotherapy. *Journal of Psychotherapy Practice and Research, 7* (1), 23-34.

Wilson, K., Mottram, P. G. & Vassilas, C. (2008). Psychotherapeutic treatments for older depressed people. *Cochrane Database of Systematic Reviews* Issue 1.

Windeler, J. (2008). Externe Validität. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 102* (4), 253-259.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mitttels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

Wissenschaftlicher Beirat Psychotherapie (1999). Leitfaden für die Erstellung von Gutachten-Anträgen zu Psychotherapieverfahren. *Deutsches Ärzteblatt, 96* (15), A-1015.

Wissenschaftlicher Beirat Psychotherapie (2000a). Anwendungsbereiche von Psychotherapie bei Kindern und Jugendlichen. *Deutsches Ärzteblatt, 97* (33), 2190.

Wissenschaftlicher Beirat Psychotherapie (2000b). Anwendungsbereiche von Psychotherapie bei Erwachsenen. *Deutsches Ärzteblatt, 97* (1-2), A-59.

Wissenschaftlicher Beirat Psychotherapie (2000c). Gutachten zur Gesprächspsychotherapie als wissenschaftliches Psychotherapieverfahren. *Deutsches Ärzteblatt, 97* (1-2), 61-63.

Wissenschaftlicher Beirat Psychotherapie (2000d). Wirksamkeitskriterien angewendet. *Deutsches Ärzteblatt, 97* (13), 818.

Wissenschaftlicher Beirat Psychotherapie (2002a). Gutachten zum Nachantrag zur Gesprächspsychotherapie. *Deutsches Ärzteblatt PP* (12), 567-568.

Wissenschaftlicher Beirat Psychotherapie (2002b). Anwendungsbereiche von Psychotherapie bei Erwachsenen: (geänderte Fassung nach dem Beschluss des Beirats vom 16.09.2002). *Deutsches Ärzteblatt, 99* (46), 3132.

Wissenschaftlicher Beirat Psychotherapie (2004a). Stellungnahme des Wissenschaftlichen Beirats Psychotherapie nach § 11 PsychThG zur Verhaltenstherapie. *Deutsches Ärzteblatt, 101* (6), A-367 - A-368.

Wissenschaftlicher Beirat Psychotherapie (2004b). Mindestanforderungen für die Begutachtung von Wirksamkeitsstudien im Bereich der Psychotherapie: (geänderte Fassung nach dem Beschluss des Beirats vom 15.9.2003). *Deutsches Ärzteblatt, 101* (6), A-369.

Wissenschaftlicher Beirat Psychotherapie (2005). Stellungnahme zur Psychodynamischen Psychotherapie bei Erwachsenen. *Deutsches Ärzteblatt, 102* (1-2), A-73 - A-75.

Wissenschaftlicher Beirat Psychotherapie (2006a). Gutachten zur wissenschaftlichen Anerkennung der EMDR-Methode (Eye-Movement-Desensitization and Reprocessing) zur Behandlung der Posttraumatischen Belastungsstörung. *Deutsches Ärzteblatt PP* (10), 476-478.

Wissenschaftlicher Beirat Psychotherapie (2006b). Gutachten zur wissenschaftlichen Anerkennung der Interpersonellen Psychotherapie. *Deutsches Ärzteblatt PP* (10), 473-475.

Wissenschaftlicher Beirat Psychotherapie (2006c). Gutachten zur wissenschaftlichen Anerkennung der Hypnotherapie. *Deutsches Ärzteblatt PP* (6), 285-287.

Wissenschaftlicher Beirat Psychotherapie (2007). *Methodenpapier des Wissenschaftlichen Beirats Psychotherapie nach § 11 PsychThG. Verfahrensregeln zur Beurteilung der wissenschaftlichen Anerkennung von Methoden und Verfahren der Psychotherapie.* Version 2.6. Verfügbar unter: http://www.wbpsychotherapie.de/page.asp?his=0.1.78 [6.9.2008].

Wissenschaftlicher Beirat Psychotherapie (2008a). Ergänzung der Stellungnahme zur Psychodynamischen Psychotherapie vom 30. Juni 2008. *Deutsches Ärzteblatt, 105* (33), A-1752.

Wissenschaftlicher Beirat Psychotherapie (2008b). Ergänzung zum Gutachten des Beirats zur neuropsychologischen Therapie. *Deutsches Ärzteblatt, 105* (13), 702.

Wissenschaftlicher Beirat Psychotherapie (2008c). *Glossar zu wiederkehrenden Begriffen im Zusammenhang mit den Stellungnahmen des Wissenschaftlichen Beirats Psychotherapie*. Verfügbar unter: http://www.wbpsychotherapie.de/page.asp?his=0.1.16 [30.9.2014].

Wissenschaftlicher Beirat Psychotherapie (2009a). Gutachten zur wissenschaftlichen Anerkennung der Systemischen Therapie. *Deutsches Ärzteblatt, 106* (5), 208-211.

Wissenschaftlicher Beirat Psychotherapie (2009b). *Methodenpapier des Wissenschaftlichen Beirats Psychotherapie nach § 11 PsychThG. Verfahrensregeln zur Beurteilung der wissenschaftlichen Anerkennung von Methoden und Verfahren der Psychotherapie.* Version 2.7. Verfügbar unter: http://www.wbpsychotherapie.de/downloads/Methodenpapier2720090709.pdf [15.5.2010].

Wissenschaftlicher Beirat Psychotherapie (2010). *Methodenpapier des Wissenschaftlichen Beirats Psychotherapie nach § 11 PsychThG. Verfahrensregeln zur Beurteilung der wissenschaftlichen Anerkennung von Methoden und Verfahren der Psychotherapie.* Version 2.8. Verfügbar unter: http://www.wbpsychotherapie.de/downloads/Methodenpapier28.pdf [12.9.2014].

Zepf, S. & Hartmann, S. (2002). Wissenschaftliche Prüfung und wissenschaftliche Anerkennung psychotherapeutischer Verfahren. Einige grundsätzliche Anmerkungen zu den Prüfkriterien des wissenschaftlichen Beirats "Psychotherapie". *Psychotherapeut, 47* (5), 278-284.

Zurhorst, G. (2003). Eminenz-basierte, Evidenz-basierte oder Ökologisch-basierte Psychotherapie? *Psychotherapeutenjournal, 2* (2), 97-104.

Anhang

Anhang A: Kriterienkatalog zur Beurteilung der Studienqualität von Psychotherapiestudien. Aus: Methodenpapier 2.8 des Wissenschaftlichen Beirats Psychotherapie nach § 11 PsychThG

(aus: WBP, 2010)

A.	Kriterien zur Bewertung der allgemeinen methodischen Qualität	Operationalisierung
1.	Manipulation der Daten **(Stufe 3 = Ausschlusskriterium)**	1) keine Hinweise auf Ergebnismanipulation 3) Hinweise auf Ergebnismanipulation
	Patienten	
2.	Objektive und reliable Diagnosestellung (mittels (teil-) standardisierter Interviews) **(Stufe 3 = Ausschlusskriterium)**	1) Diagnosestellung mittels strukturiertem klinischen bzw. voll standardisiertem Interview (z.B. SKID, DIPS) 2) Diagnosestellung mittels Diagnosechecklisten oder nachvollziehbarem klinischen Urteil 3) keine adäquate Diagnosestellung
3.	Höhe der Drop-out-Quote zu Behandlungsende (sofern nicht Erfolgskriterium)	1) i.d.R. Drop-out-Quote kleiner 20 % 2) i.d.R. Drop-out-Quote zwischen 20 % und 40 % 3) i.d.R. Drop-out-Quote größer 40 %
4.	Höhe der Studien-Drop-outs zur Katamnese (falls Katamneseerhebung durchgeführt)	1) deutlich besser als in Studien mit vergleichbaren Patientengruppen und vergleichbarem Katamnesezeitraum 2) Drop-out-Quote vergleichbar mit Studien mit entsprechenden Patientengruppen und entsprechendem Katamnesezeitraum 3) deutlich schlechter als in Studien mit vergleichbaren Patientengruppen und vergleichbarem Katamnesezeitraum
	Studiendesign	
5.	Stichprobengröße pro Gruppe	1) n pro Gruppe > 30 2) n pro Gruppe 10-30 3) n pro Gruppe < 10
6.	Vergleiche der (sofern vorhanden) Behandlungsgruppen und der Messzeitpunkte a priori definiert	1) a priori Definition der Vergleiche der (sofern vorhanden) Behandlungsgruppen und Messzeitpunkte erfüllt 2) teilweise post-hoc Definition der Vergleiche 3) ausschließlich post-hoc definierte Vergleiche
	Outcome-Messung	
7.	a priori Definition der primären und sekundären Zielkriterien	1) a priori definierte primäre und gegebenenfalls sekundäre Zielkriterien 2) a priori Nennung der Zielkriterien ohne Differenzierung in primäre und sekundäre Zielkriterien 3) a posteriori Definition der Zielkriterien; keine Angaben zu Zielkriterien
8.	Reliable und valide Messung zumindest der primären Zielkriterien **(Stufe 3 = Ausschlusskriterium)**	1) reliable und valide Outcome-Verfahren 2) nur eingeschränkte Reliabilität und/oder Validität der Messverfahren 3) Reliabilität und Validität der Messverfahren nicht überprüft oder Gütekriterien der Messverfahren sind unzureichend

9.	Klinische Bedeutsamkeit der Outcome-Messung (z.B. das Konzept der klinischen Signifikanz)	1) klinische Bedeutsamkeit des Therapieeffekts (z. B. im Sinne des Konzepts der klinischen Signifikanz) ist feststellbar 2) klinische Bedeutsamkeit des Therapieeffekts (z. B. im Sinne des Konzepts der klinischen Signifikanz) ist nur eingeschränkt feststellbar 3) klinische Bedeutsamkeit des (z. B. im Sinne des Konzepts der klinischen Signifikanz) Therapieeffekts ist nicht feststellbar
10.	Multiple Informationsquellen (z.B. Patient, Therapeut, Laborwerte)	1) multidimensionale Erfassung der Zielkriterien – drei oder mehr Informationsquellen 2) zwei Informationsquellen 3) eine Informationsquelle
11.	Sofern Fremdeinschätzungsverfahren: externe Beurteiler (blind für die Gruppenzugehörigkeit)	1) validiertes Fremdeinschätzungsverfahren angewendet von trainierten, für die Gruppenbedingungen blinden externen Beurteilern 2) validiertes Fremdeinschätzungsverfahren angewendet von trainierten, nicht-blinden externen Beurteilern 3) validiertes Fremdeinschätzungsverfahren angewendet – Rater sind weder trainiert noch blind für die Gruppenzugehörigkeit der Patienten
12.	Vollständige Darstellung der Ergebnisse zu allen Outcomemaßen und zu allen relevanten Messzeitpunkten	1) für alle Outcome-Variablen berichtet 2) ausschließlich für die primären Zielkriterien berichtet 3) nicht für alle primären Zielkriterien berichtet
13.	Erfassung unerwünschter Wirkungen	1) Systematische Erfassung und Bericht von unerwünschten Ereignissen, Nebenwirkungen, Verschlechterungen 2) unsystematische Erfassung und/oder unvollständiger Bericht von unerwünschten Ereignisse, Nebenwirkungen, Verschlechterungen 3) Unerwünschte Ereignisse, Nebenwirkungen, Verschlechterungen wurden nicht erfasst oder nicht berichtet
	Statistische Methodik	
14.	Anwendungsvoraussetzungen für statistische Modelle geprüft und erfüllt	1) Anwendungsvoraussetzungen geprüft und erfüllt 2) Anwendungsvoraussetzungen geprüft und lediglich leichte Verletzungen der Voraussetzungen oder keine Prüfung der Anwendungsvoraussetzungen bei gleichzeitiger Robustheit der angewendeten statistischen Verfahren 3) deutliche Verletzungen der Anwendungsvoraussetzungen oder keine Prüfung der Anwendungsvoraussetzungen bei substanziellem Risiko für deren Verletzung
15.	Angemessenheit der statistischen Analysen (inklusive der Korrektur für multiple Tests)	1) adäquate und umfassende statistische Analysen 2) weitgehend adäquate statistische Analysen 3) unangemessene statistische Analysen (fehlende Korrektur für multiple Tests, inadäquate statistische Methoden)

16.	Intention to treat–Analysen durchgeführt	1) ITT-Analysen durchgeführt 2) Keine ITT-Analysen bei geringem Risiko für einen attrition bias 3) Keine ITT-Analysen bei deutlichem Risiko für einen attrition bias
17.	Statistische Power der Vergleiche bei Vergleich mit bewährter Treatment-Gruppe	1) adäquate statistische Power der Vergleiche 2) eingeschränkte Power der statistischen Vergleiche (.50-.80) 3) unzureichende statistische Power der Vergleiche (<.50)
18.	Vollständige Beschreibung der Drop-Outs	1) vollständige Beschreibung aller Drop-outs, inkl. der Gründe und des Zeitpunkts des Drop-outs 2) unvollständige Angaben zu Gründen oder Zeitpunkten des Drop-outs 3) keine Beschreibung der Drop-outs oder definitive Angabe der Anzahl der Drop-outs in einer Gruppe
19.	Drop-out-Analysen	1) Drop-out-Analysen unter Berücksichtigung der wichtigsten prognostischen Faktoren durchgeführt; keine signifikanten Unterschiede zwischen den Vergleichsgruppen 2) Drop-out-Analysen unter Berücksichtigung wichtigster prognostischer Faktoren durchgeführt; trotz sign. Unterschiede ist die Validität der Ergebnisse nicht wesentlich eingeschränkt 3) Drop-out-Analysen nicht oder unter Vernachlässigung relevanter prognostischer Merkmale durchgeführt; Drop-out-Analysen stellen die Validität der Ergebnisse deutlich in Frage

B.	Kriterien zur Bewertung der internen Validität	Operationalisierung
	Patienten	
1.	Spezifizierung der Einschluss- und Ausschlusskriterien	1) eindeutige Spezifizierung der Ein- und Ausschlusskriterien 2) Ein- oder Ausschlusskriterien teilweise uneindeutig beschrieben 3) Ein- und Ausschlusskriterien sind nicht eindeutig definiert
2.	Erhebung der spezifizierten Einschluss- und Ausschlusskriterien mittels valider Methoden	1) Die Ein- und Ausschlusskriterien sind sämtlich klar operationalisiert und werden mittels valider Methoden erfasst (z. B. komorbide Störungen als Ausschluss werden mittels strukturiertem klinischen Interview erfasst; Ausschlusskriterien beziehen sich auf eindeutig objektivierbare Merkmale wie Alter, Geschlecht etc.) 2) Die Validität der Erhebungen von Teilen der Ein- oder Ausschlusskriterien ist teilweise eingeschränkt (z. B. komorbide Störungen als Ausschluss werden mittels globalem klinischen Urteil eingeschätzt) und wirkt sich jedoch nur in geringem Umfang auf die Zusammensetzung der Stichprobe aus 3) Die Validität der Erhebungen von Teilen der Ein- oder Ausschlusskriterien ist deutlich eingeschränkt und wirkt sich differenziell auf die Zusammensetzung der Behandlungsgruppen aus
	Intervention	
3.	Operationale Definition der Interventionen (Experimental- und ggf. Kontrollgruppe)	1) Therapiemanual, bei dem die Interventionen so beschrieben sind, dass das therapeutische Vorgehen vergleichbar und replizierbar ist 2) Therapiebeschreibung, ohne nähere Spezifikation der einzelnen Interventionen (z.B. Fehlen gegebenenfalls für das psychotherapeutische Verfahren oder die psychotherapeutische Methode notwendiger Entscheidungskriterien) 3) Die Intervention ist nicht klar beschrieben, beschränkt sich auf die Benennung des Psychotherapieverfahrens bzw. der Psychotherapiemethode
4.	Operationale Definition der Kontrollbedingungen	1) Prospektive Festlegung und umfassende Beschreibung der Kontrollbedingung 2) Ex post facto Beschreibung der Kontrollbedingungen 3) keine Beschreibung der Kontrollbedingung
5.	Strukturelle Äquivalenz bei Kontrollbedingungen	1) hinsichtlich des Umfangs an therapeutischer Zuwendung und der Settingbedingungen in der KG besteht Äquivalenz 2) der Umfang der therapeutischen Zuwendung in der KG ist reduziert, die Settingbedingungen weichen von der IG ab 3) der Umfang der therapeutischen Zuwendung in der KG ist deutlich reduziert, die Settingbedingungen weichen wesentlich von der IG ab

6.	Manualtreue, Treatment Integrity	1) Manualtreue/Treatmentintegrität durch externe Beobachter (z.B. videogestützt) belegt 2) Manualtreue/Treatmentintegrität durch Fragebögen belegt 3) keine Maßnahmen zum Monitoring der Manualtreue oder Hinweise auf substanzielle Abweichungen
7.	Zulässigkeit, Dokumentation und Analyse des Einflusses begleitender nicht-randomisierter Interventionen (insbesondere Pharmakotherapie)	1) Ausschluss begleitender nicht-randomisierter Interventionen 2) begleitende nicht-randomisierte Interventionen sind zulässig, werden jedoch detailliert dokumentiert und die Analysen weisen auf keinen substanziellen, differenziellen Einfluss der begleitenden Interventionen hin 3) begleitende nicht-randomisierte Interventionen sind zulässig, werden jedoch nicht dokumentiert oder die Analysen weisen auf eine differenzielle Inanspruchnahme von begleitenden Interventionen und deren Einfluss auf das Behandlungsergebnis hin
	Studiendesign	
8.	Gruppenzuweisung **(obligatorisches Kriterium für interne Validität (<3))**	1) angemessene Randomisierung (inkl. Cluster-Randomisierung)bei ausreichender Stichprobengröße (n>30/Gruppe[88]), die das Gelingen der Randomisierung hinsichtlich bekannter und unbekannter (nicht erfasster) prognostisch relevanter Merkmale sicherstellt 2) Parallelisierung oder teilweise randomisiert oder quasi-randomisiert oder Stichprobengröße n<30/Gruppe 3) keine randomisierte oder parallelisierte Zuweisung
9.	Vergleichbarkeit der Gruppen zur Baseline im Hinblick auf prognostisch relevante Merkmale	1) Weder statistisch noch klinisch relevante Unterschiede zwischen den Gruppen hinsichtlich prognostisch relevanter oder potentiell konfundierender Variablen 2) Vergleichbarkeit hinsichtlich der meisten prognostisch relevanten Merkmale ist weitgehend gegeben; signifikante Unterschiede hinsichtlich relevanter prognostischer Merkmale zwischen den Gruppen werden statistisch angemessen kontrolliert 3) Keine angemessene Überprüfung der Vergleichbarkeit oder Vergleichsgruppen unterscheiden sich erheblich hinsichtlich mehrerer prognostisch relevanter Merkmale und eine angemessene statistische Kontrolle des Einflusses dieser Merkmale fehlt
10.	Definition der Messzeitpunkte (Prospektive Messung; Follow-up-Messung)	1) mehrere vorab festgelegte Messzeitpunkte über den Therapieverlauf incl. Prä-Post-Messungen 2) ausschließlich Prä-Post-Messung 3) ausschließlich Post-Messung
11.	Follow-up-Messung	1) zeitlich störungsangemessene Katamnese mit hoher Ausschöpfung der Stichprobe 2) Katamnese mit fraglich angemessenem Zeitraum bzw. niedriger Ausschöpfung der Stichprobe 3) keine Katamnese

[88] es handelt sich um Anhaltszahlen.

	Outcome-Messung	
12.	Erzielte Veränderungen auf den primären und sekundären Zielkriterien ggf. im Vergleich zur Kontrollgruppe (Signifikanz, Größe und Relevanz der Effekte) **(Stufe 3 = Ausschlusskriterium)**	1) vollständige Darstellung der erzielten Veränderungen auf den Zielkriterien inklusive der Signifikanz, Größe der Effektmaße und Ausmaß der klinisch relevanten Zielerreichung (ggf. im Vergleich zur Kontrollgruppe) 2) Darstellung des Behandlungsergebnisses nur durch Veränderungs- oder Zielerreichungsmaße oder beides ist (ggf. im Vergleich zur Kontrollgruppe) bei einigen Kriterien unvollständig 3) weitgehend unvollständige oder inadäquate Darstellung der Outcome-Kriterien (ggf. im Vergleich zur KG)

C.	Kriterien zur Beurteilung der externen Validität[89]	Operationalisierung
	Patienten	
1.	Stichprobe von Patienten mit Störungen mit Krankheitswert **(Stufe 3 = Ausschlusskriterium)**	1) ausschließlich Patienten mit Störung mit Krankheitswert (z. B. ICD-, DSM-Diagnosen) 2) Stichprobe von Patienten mit wahrscheinlicher klinischer Störung (z. B. Menschen nach Trauma) oder bis zu (maximal) 20% der Patienten mit lediglich erhöhter Symptomausprägung, z.T. subklinisch 3) Patienten ohne festgestellte Störung mit Krankheitswert
2.	Art der Rekrutierung der Stichprobe	1) Patientenzugang überwiegend durch gängige klinische Routinen (Überweisung , Primärzugang, etc.); keine Selektionseffekte aufgrund der Zugangswege 2) Patientenzugang überwiegend durch gängige klinische Routinen (Überweisung , Primärzugang, etc.); Selektionseffekte aufgrund der Zugangswege 3) Patientenzugang überwiegend über Aufforderungen der Forschergruppe (z. B. Anzeigenwerbung)
3.	Selektivität der Stichprobe aufgrund der Ausschlusskriterien	1) keine Selektionseffekte aufgrund der Ausschlusskriterien: Einschluss aller Patienten 2) mittlere Selektionseffekte aufgrund der Ausschlusskriterien (z. B. Ausschluss einiger epidemiologisch relevanter komorbider Störungen) 3) deutliche Selektionseffekte aufgrund der Ausschlusskriterien
	Intervention	
4.	Klinische Repräsentativität der Intervention hinsichtlich Vorgehen und Dauer	1) Intervention wie in klinischer Alltagspraxis 2) Intervention gegenüber klinischer Alltagspraxis teilweise verändert 3) Intervention gegenüber klinischer Alltagspraxis stark verändert
5.	Art des Therapie-Monitorings (Einfluss auf Therapeutenverhalten)	1) keine Veränderung des Therapeutenverhaltens durch Therapie-Monitoring (z.B. durch Therapie-Supervision; Ausnahmen: Audio- oder Video-Aufzeichnungen ohne Feedback an Therapeuten, nur zur späteren Auswertung sind erlaubt) 2) mittlere Veränderung des Therapeutenverhaltens durch Therapie-Monitoring (gelegentliche Rückmeldungen an Therapeuten) 3) starke Veränderung des Therapeutenverhaltens durch Therapie-Monitoring (durch kontinuierliche Rückmeldungen)

[89] Die Beurteilung der externen Validität bezieht sich auf die Vergleichbarkeit mit Verhältnissen des deutschen Gesundheitssystems.

6.	Zulässigkeit begleitender Interventionen (z. B. Pharmakotherapie)	1) keine Einschränkungen 2) begleitende in der Routine-Praxis übliche Interventionen teilweise ausgeschlossen 3) alle begleitenden in der Routine-Praxis üblichen Interventionen ausgeschlossen
7.	Qualifikation der Behandler a) Klinische Tätigkeit der Therapeuten	1) Therapeuten sind praktizierende Kliniker 2) klinische Forscher, die überwiegend Forschung betreiben und seltener auch Patienten behandeln; (Ausbildungskandidaten) 3) keine Kliniker oder Kliniker, die keine Psychotherapeuten sind
	b) Breite der Klinischen Tätigkeit der Therapeuten (Problemheterogenität)	1) Therapeut behandelt Patienten mit verschiedenen Problemen innerhalb und außerhalb der Studie 2) Therapeut behandelt überwiegend Patienten mit ähnlichen Problemen innerhalb und außerhalb der Studie 3) Therapeut behandelt nur Patienten mit ähnlichen Problemen innerhalb und außerhalb der Studie (z.B. Therapeut, der nur Schmerzpatienten in einer Schmerzklinik behandelt)
	c) Spezifisches Training der Psychotherapeuten in einer Behandlungsmethode für die Studie	1) kein spezifisches Training für die Studie (z.B. Therapeuten wenden die von ihnen üblicherweise angewendete Therapie an) 2) kurzes Training für die Studie / intensives Training nur einiger Therapeuten 3) intensives Training vor der Studie
	Studiendesign	
8.	Repräsentativität der patientenseitigen Freiheit hinsichtlich der Wahl der Intervention	1) Patienten entscheiden sich selbst für eine der angebotenen Therapieformen 2) ein Teil der Patienten (z. B. Randomisierungswillige) wird der Therapie zufällig zugewiesen 3) alle Patienten werden der Therapie (zufällig) zugewiesen
	Outcome-Messung	
9.	Primäre Zielkriterien beziehen sich auf patientenrelevante Parameter (insbesondere Schwere der Symptomatik, Leiden, Beeinträchtigung/Lebensqualität, Inanspruchnahme von Diensten des Gesundheitswesens) **(Stufe 3 = Ausschlusskriterium)**	1) Zielkriterien beziehen sich auf mehrere Dimensionen patienten- bzw. störungsrelevanter Parameter unter Einbezug von Beeinträchtigung/Lebensqualität und Inanspruchnahme von Diensten des Gesundheitswesens) 2) Zielkriterien beziehen sich nur auf eine Dimension 3) Zielkriterien beziehen sich ausschließlich auf Surrogatparameter (z.B. Kontrollüberzeugung)

	Praxistransfer	
10.	Spezifikation und Herstellbarkeit notwendiger Settingbedingungen	1) Notwendige Settingbedingungen herstellbar (z. B. Infrastruktur, Kooperation, Team) 2) Notwendige Settingbedingungen nur begrenzt herstellbar (z. B. Infrastruktur, Kooperation, Team) 3) Notwendige Settingbedingungen nicht herstellbar
11.	Spezifikation und Herstellbarkeit der notwendigen Behandlerqualifikation	1) Notwendige Behandlungsqualifikation eindeutig beschrieben und herstellbar 2) Notwendige Behandlungsqualifikation eindeutig beschrieben, aber nur mit sehr großem Zeitaufwand herstellbar 3) Notwendige Behandlungsqualifikation nicht beschrieben oder praktisch nicht herstellbar
12.	Spezifikation und Erfassbarkeit relevanter Patientenmerkmale	1) Relevante Patientenmerkmale (z.B. Alter, genetische Marker) praktisch erfassbar 2) Relevante Patientenmerkmale nur mit erheblichem Aufwand erfassbar 3) Relevante Patientenmerkmale praktisch nicht erfassbar
13.	Spezifikation und Herstellbarkeit relevanter Treatmentaspekte	1) Relevante Treatmentmerkmale (Art der Interventionen, Reihenfolge, Dauer) praktisch herstellbar (z.B. durch Manual) 2) Relevante Treatmentmerkmale nur schwer herstellbar (z.B. tägliche Behandlung, Parallelbehandlung) 3) Relevante Treatmentmerkmale praktisch nicht herstellbar

Anhang B: Kurzkodierbogen und Kriterienkatalog des Wissenschaftlichen Beirats Psychotherapie nach § 11 PsychThG

(aus: Ratzek & von Hauenschild, 2011b; WBP, 2010)

Beurteiler/in: **Datum der Beurteilung:**

Autoren:

Titel:

Publikationsjahr:

Datenzugriff:

Originalstudie	
Reanalyse	
Replikation	

Weitere Charakteristiken und methodologische Eigenschaften der Studie:

Anwendungsbereich
Affektive Störungen:
Gemischte Störungsgruppen:

Zentrale Patientenmerkmale:

Anwendungsform der psychoanalytisch begründeten Verfahren
Benennung in Studie:

Einordnung nach Psychotherapierichtlinien (plus Psychoanalyse)		**Sitzungsanzahl/Dauer und Frequenz**
tiefenpsychologisch fundiert		
analytische Psychotherapie		
Psychoanalyse		
Gehen unterschiedliche Anwendungsformen der psychoanalytisch begründeten Verfahren in <u>einen</u> Treatmentarm ein oder werden die Anwendungsformen <u>separat</u> voneinander untersucht?		

Setting:	
Einzeltherapie	
Gruppentherapie	
a priori festgelegte Sitzungsanzahl	
ja	
nein	
störungs<u>un</u>spezifisches vs. störungsspezifisches Vorgehen	
störungs<u>un</u>spezifische Anwendung der analytisch begründeten Therapie	
störungsspezifische Anwendung der analytisch begründeten Therapie	

Kontroll-/Vergleichsgruppendesign		**Sitzungsanzahl/Dauer und Frequenz**
unbehandelte Kontrollgruppe (z.B. Warteliste)		
Placebo-Kontrollgruppe (stützende Gespräche, nonspecific treatment)		
TAU-Kontrollgruppe		
Aktive Kontrollgruppe (z.B. supportive Psychotherapie)		
verfahrensexterne und etablierte Vergleichsbehandlung(en)[90]		
verfahrensexterne Vergleichsbehandlungen(en) - kein bereits etabliertes Treatment		
verfahrensinterne Vergleichsbehandlungen(n) (z.B. Vergleich zweier Methoden der analytisch begründeten Verfahren)		
Vergleichsgruppe mit ausschließlich psychopharmakologischer Behandlung		
keine Vergleichsgruppe		
Wie viele Behandlungsarme insgesamt?- Welche?		

Messzeitpunkte		**prospektives vs. retrospektives Design**	
Prä-Post		prospektiv (vor Beginn des zu evaluierenden Treatments wird das Design festgelegt: Treatmentarme bzw. zu vergleichende Gruppen; Messzeitpunkte; Outcomekriterien)	
Prä-Katamnese			
Prä-Post-Katamnese			
Post-Katamnese		retrospektiv (das/die Treatment/s sind bereits erfolgt und die Auswahl der Treatmentarme, die Bestimmung der Messzeitpunkte und die Festlegung der Outcomemaße erfolgt erst nach Beendigung [ggf. während] des/r Treatments)	
nur Post			
nur Katamnese(n)			
Katamnesezeitraum (Post-Katamnese):			

[90] Als „etabliertes Treatment“ werden Verfahren betrachtet, die zum aktuellen Zeitpunkt (01/2011) als wissenschaftlich anerkannte Verfahren gelten (Verhaltenstherapie, Gesprächspsychotherapie, Systemische Therapie).

Gruppenzuweisung		
randomisierte Zuteilung	verfahrensinterne Vergleichsgruppe/n	
	verfahrensexterne Vergleichsgruppe/n *oder* Kontrollgruppe/n (Warteliste, Placebo, TAU, aktive Kontrolle)	
teilweise randomisiert und teilweise Selbstzuteilung	verfahrensinterne Vergleichsgruppe/n	
	verfahrensexterne Vergleichsgruppe/n *oder* Kontrollgruppe/n (Warteliste, Placebo, TAU, aktive Kontrolle)	
patientenseitige Selbstzuteilung unter Anwendung von Strategien wie z.B. Parallelisierung, Stratifizierung, Matching	verfahrensinterne Vergleichsgruppe/n	
	verfahrensexterne Vergleichsgruppe/n *oder* Kontrollgruppe/n (Warteliste, Placebo, TAU, aktive Kontrolle)	
patientenseitige Selbstzuteilung ohne Anwendung von Strategien wie z.B. Parallelisierung, Stratifizierung, Matching	verfahrensinterne Vergleichsgruppe/n	
	verfahrensexterne Vergleichsgruppe/n *oder* Kontrollgruppe/n (Warteliste, Placebo, TAU, aktive Kontrolle)	
Standardisierungsgrad des Treatments		
Verwendung von Manual bzw. manualähnlichen Behandlungsrichtlinien (behandlungsprinzipienbasiert)		
nicht manualisiert bzw. keine Verwendung manualähnlicher Behandlungsrichtlinien (behandlungsprinzipienbasiert)		
explizites Therapeutentraining zwecks Studiendurchführung		
kein explizites Therapeutentraining		
Implementationskontrolle		
keine Implementationskontrolle		

Selektivität der Stichprobe		
Ausschluss subklinischer Symptomausprägungen	ja	
	nein	
Ausschluss komorbider Störungen	ja	
	nein	

Von den Autor/innen formulierte Fragestellung (Untersuchungsziel):

Studiendesign:

B.8.	Gruppenzuweisung	angemessene Randomisierung (inkl. Cluster-Randomisierung)bei ausreichender Stichprobengröße (n>30/Gruppe[91]), die das Gelingen der Randomisierung hinsichtlich bekannter und unbekannter (nicht erfasster) prognostisch relevanter Merkmale sicherstellt	(1)
		Parallelisierung oder teilweise randomisiert oder quasi-randomisiert oder Stichprobengröße n<30/Gruppe	(2)
		keine randomisierte oder parallelisierte Zuweisung **(obligatorisches Kriterium für interne Validität (<3))**	(3)
C.8.	Repräsentativität der patientenseitigen Freiheit hinsichtlich der Wahl der Intervention	Patienten entscheiden sich selbst für eine der angebotenen Therapieformen	(1)
		ein Teil der Patienten (z. B. Randomisierungswillige) wird der Therapie zufällig zugewiesen	(2)
		alle Patienten werden der Therapie (zufällig) zugewiesen	(3)
A.5.	Stichprobengröße pro Gruppe	n pro Gruppe > 30	(1)
		n pro Gruppe 10-30	(2)
		n pro Gruppe < 10	(3)
A.6.	Vergleiche der (sofern vorhanden) Behandlungsgruppen und der Messzeitpunkte a priori definiert	a priori Definition der Vergleiche der (sofern vorhanden) Behandlungsgruppen und Messzeitpunkte erfüllt	(1)
		teilweise post-hoc Definition der Vergleiche	(2)
		ausschließlich post-hoc definierte Vergleiche	(3)
A.7.	a priori Definition der primären und sekundären Zielkriterien	a priori definierte primäre und gegebenenfalls sekundäre Zielkriterien	(1)
		a priori Nennung der Zielkriterien ohne Differenzierung in primäre und sekundäre Zielkriterien	(2)
		a posteriori Definition der Zielkriterien; keine Angaben zu Zielkriterien	(3)

[91] es handelt sich um Anhaltszahlen.

B.10.	Definition der Messzeitpunkte (Prospektive Messung; Follow-up-Messung)	mehrere vorab festgelegte Messzeitpunkte über den Therapieverlauf incl. Prä-Post-Messungen	(1)
		ausschließlich Prä-Post-Messung	(2)
		ausschließlich Post-Messung	(3)
B.11.	Follow-up-Messung	zeitlich störungsangemessene Katamnese mit hoher Ausschöpfung[92] der Stichprobe - mindestens 2 Jahre (2 Jahre oder mehr) - mindestens 60%ige Ausschöpfungsquote	(1)
		Katamnese mit fraglich angemessenem Zeitraum bzw. niedriger Ausschöpfung der Stichprobe - weniger als 2 Jahre - weniger als 60% Ausschöpfungsquote	(2)
		keine Katamnese	(3)
	Katamnesenzeitraum (Zeitraum zwischen Beendigung der Therapie und letzter Katamneseerhebung):		
	Katamnesemessungen auf allen Treatmentarmen?	ja	
		nein	
	Weitere/r katamnestische/r Messzeitpunkt/e (vor letzter Katamnesemessung liegend)? Wenn ja, wann?	ja	
		nein	
	Einbeziehung longitudinaler Informationen: Wurde die Inanspruchnahme weiterer psychotherapeutischer Interventionen über das gesamte Katamneseintervall erhoben (Inanspruchnahmedaten, Arztbriefe, Krankenkassendaten)?	ja	
		nein	
	Direkte persönliche Untersuchung (vs. Telefon- oder postalische Fragebogenuntersuchung)?	ja	
		nein	
	Erhebung explizit psychodynamischer Outcomekriterien (Konfliktdynamiken, strukturelle Aspekte)?	ja	
		nein	
	Stabilitätsmessung? (d.h. differenzierte Betrachtung der Responder und Non-Responder)	ja	
		nein	
	Wird der in der Studie gewählte Katamnesezeitraum explizit begründet? Wenn ja, wie?	ja	
		nein	

[92] Gemeint ist die Ausschöpfung der Stichprobe zwischen Prä- und Katamnesemessung.

C.10.	Spezifikation und Herstellbarkeit notwendiger Settingbedingungen (Praxistransfer)	Notwendige Settingbedingungen herstellbar (z. B. Infrastruktur, Kooperation, Team)	(1)
		Notwendige Settingbedingungen nur begrenzt herstellbar (z. B. Infrastruktur, Kooperation, Team)	(2)
		Notwendige Settingbedingungen nicht herstellbar	(3)
C.11.	Spezifikation und Herstellbarkeit der notwendigen Behandlerqualifikation (Praxistransfer)	Notwendige Behandlungsqualifikation eindeutig beschrieben und herstellbar	(1)
		Notwendige Behandlungsqualifikation eindeutig beschrieben, aber nur mit sehr großem Zeitaufwand herstellbar	(2)
		Notwendige Behandlungsqualifikation nicht beschrieben oder praktisch nicht herstellbar	(3)
C.12.	Spezifikation und Erfassbarkeit relevanter Patientenmerkmale (Praxistransfer)	Relevante Patientenmerkmale (z.B. Alter, genetische Marker) praktisch erfassbar	(1)
		Relevante Patientenmerkmale nur mit erheblichem Aufwand erfassbar	(2)
		Relevante Patientenmerkmale praktisch nicht erfassbar	(3)
C.13.	Spezifikation und Herstellbarkeit relevanter Treatmentaspekte (Praxistransfer)	Relevante Treatmentmerkmale (Art der Interventionen, Reihenfolge, Dauer) praktisch herstellbar (z.B. durch Manual)	(1)
		Relevante Treatmentmerkmale nur schwer herstellbar (z.B. tägliche Behandlung, Parallelbehandlung)	(2)
		Relevante Treatmentmerkmale praktisch nicht herstellbar	(3)

Population/Patienten:

C.1.	Stichprobe von Patienten mit Störungen mit Krankheitswert	ausschließlich Patienten mit Störung mit Krankheitswert (z. B. ICD-, DSM-Diagnosen)	(1)
		Stichprobe von Patienten mit wahrscheinlicher klinischer Störung (z. B. Menschen nach Trauma) oder bis zu (maximal) 20% der Patienten mit lediglich erhöhter Symptomausprägung, z.T. subklinisch	(2)
		Patienten ohne festgestellte Störung mit Krankheitswert **(Stufe 3 = Ausschlusskriterium)**	(3)

C.2.	Art der Rekrutierung der Stichprobe	Patientenzugang überwiegend durch gängige klinische Routinen (Überweisung , Primärzugang, etc.); keine Selektionseffekte aufgrund der Zugangswege	(1)
		Patientenzugang überwiegend durch gängige klinische Routinen (Überweisung , Primärzugang, etc.); Selektionseffekte aufgrund der Zugangswege	(2)
		Patientenzugang überwiegend über Aufforderungen der Forschergruppe (z. B. Anzeigenwerbung)	(3)
B.1.	Spezifizierung der Einschluss- und Ausschlusskriterien	eindeutige Spezifizierung der Ein- und Ausschlusskriterien	(1)
		Ein- oder Ausschlusskriterien teilweise uneindeutig beschrieben	(2)
		Ein- und Ausschlusskriterien sind nicht eindeutig definiert	(3)
	Ein-/Ausschlusskriterien:		
C.3.	Selektivität der Stichprobe aufgrund der Ausschlusskriterien	keine Selektionseffekte aufgrund der Ausschlusskriterien: Einschluss aller Patienten	(1)
		mittlere Selektionseffekte aufgrund der Ausschlusskriterien (z. B. Ausschluss einiger epidemiologisch relevanter komorbider Störungen)	(2)
		deutliche Selektionseffekte aufgrund der Ausschlusskriterien	(3)

Eingangsdiagnostik:

A.2.	Objektive und reliable Diagnosestellung (mittels (teil-) standardisierter Interviews)	Diagnosestellung mittels strukturiertem klinischen bzw. voll standardisiertem Interview (z.B. SKID, DIPS)	(1)
		Diagnosestellung mittels Diagnosechecklisten oder nachvollziehbarem klinischen Urteil	(2)
		keine adäquate Diagnosestellung **(Stufe 3 = Ausschlusskriterium)**	(3)
	Diagnostische Instrumente:		

B.2.	Erhebung der spezifizierten Einschluss- und Ausschlusskriterien mittels valider Methoden	Die Ein- und Ausschlusskriterien sind sämtlich klar operationalisiert und werden mittels valider Methoden erfasst (z. B. komorbide Störungen als Ausschluss werden mittels strukturiertem klinischen Interview erfasst; Ausschlusskriterien beziehen sich auf eindeutig objektivierbare Merkmale wie Alter, Geschlecht etc.)	(1)
		Die Validität der Erhebungen von Teilen der Ein- oder Ausschlusskriterien ist teilweise eingeschränkt (z. B. komorbide Störungen als Ausschluss werden mittels globalem klinischen Urteil eingeschätzt) und wirkt sich jedoch nur in geringem Umfang auf die Zusammensetzung der Stichprobe aus	(2)
		Die Validität der Erhebungen von Teilen der Ein- oder Ausschlusskriterien ist deutlich eingeschränkt und wirkt sich differenziell auf die Zusammensetzung der Behandlungsgruppen aus	(3)
B.9.	Vergleichbarkeit der Gruppen zur Baseline im Hinblick auf prognostisch relevante Merkmale	Weder statistisch noch klinisch relevante Unterschiede zwischen den Gruppen hinsichtlich prognostisch relevanter oder potentiell konfundierender Variablen	(1)
		Vergleichbarkeit hinsichtlich der meisten prognostisch relevanten Merkmale ist weitgehend gegeben; signifikante Unterschiede hinsichtlich relevanter prognostischer Merkmale zwischen den Gruppen werden statistisch angemessen kontrolliert	(2)
		Keine angemessene Überprüfung der Vergleichbarkeit oder Vergleichsgruppen unterscheiden sich erheblich hinsichtlich mehrerer prognostisch relevanter Merkmale und eine angemessene statistische Kontrolle des Einflusses dieser Merkmale fehlt	(3)
	Hinsichtlich Prognostizität untersuchte Merkmale:		

Intervention:

B.3.	Operationale Definition der Interventionen (Experimental- und ggf. Kontrollgruppe)	Therapiemanual, bei dem die Interventionen so beschrieben sind, dass das therapeutische Vorgehen vergleichbar und replizierbar ist	(1)
		Therapiebeschreibung, ohne nähere Spezifikation der einzelnen Interventionen (z.B. Fehlen gegebenenfalls für das psychotherapeutische Verfahren oder die psychotherapeutische Methode notwendiger Entscheidungskriterien)	(2)
		Die Intervention ist nicht klar beschrieben, beschränkt sich auf die Benennung des Psychotherapieverfahrens bzw. der Psychotherapiemethode	(3)
	Welches Manual bzw. welche manualähnlichen Behandlungsrichtlinien wurden verwendet?		
B.6.	Manualtreue, Treatment Integrity	Manualtreue/Treatmentintegrität durch externe Beobachter (z.B. videogestützt) belegt	(1)
		Manualtreue/Treatmentintegrität durch Fragebögen belegt	(2)
		keine Maßnahmen zum Monitoring der Manualtreue oder Hinweise auf substanzielle Abweichungen	(3)
C.4.	Klinische Repräsentativität der Intervention hinsichtlich Vorgehen und Dauer	Intervention wie in klinischer Alltagspraxis	(1)
		Intervention gegenüber klinischer Alltagspraxis teilweise verändert	(2)
		Intervention gegenüber klinischer Alltagspraxis stark verändert	(3)
C.5.	Art des Therapie-Monitorings (Einfluss auf Therapeutenverhalten)	keine Veränderung des Therapeutenverhaltens durch Therapie-Monitoring (z.B. durch Therapie-Supervision; Ausnahmen: Audio- oder Video-Aufzeichnungen ohne Feedback an Therapeuten, nur zur späteren Auswertung sind erlaubt)	(1)
		mittlere Veränderung des Therapeutenverhaltens durch Therapie-Monitoring (gelegentliche Rückmeldungen an Therapeuten)	(2)
		starke Veränderung des Therapeutenverhaltens durch Therapie-Monitoring (durch kontinuierliche Rückmeldungen)	(3)

B.4.	Operationale Definition der Kontrollbedingungen	Prospektive Festlegung und umfassende Beschreibung der Kontrollbedingung	(1)
		Ex post facto Beschreibung der Kontrollbedingungen	(2)
		keine Beschreibung der Kontrollbedingung	(3)
B.5.	Strukturelle Äquivalenz bei Kontrollbedingungen	hinsichtlich des Umfangs an therapeutischer Zuwendung und der Settingbedingungen in der KG besteht Äquivalenz	(1)
		der Umfang der therapeutischen Zuwendung in der KG ist reduziert, die Settingbedingungen weichen von der IG ab	(2)
		der Umfang der therapeutischen Zuwendung in der KG ist deutlich reduziert, die Settingbedingungen weichen wesentlich von der IG ab	(3)
B.7.	Zulässigkeit, Dokumentation und Analyse des Einflusses begleitender nicht-randomisierter Interventionen (insbesondere Pharmakotherapie)	Ausschluss begleitender nicht-randomisierter Interventionen	(1)
		begleitende nicht-randomisierte Interventionen sind zulässig, werden jedoch detailliert dokumentiert und die Analysen weisen auf keinen substanziellen, differenziellen Einfluss der begleitenden Interventionen hin	(2)
		begleitende nicht-randomisierte Interventionen sind zulässig, werden jedoch nicht dokumentiert oder die Analysen weisen auf eine differenzielle Inanspruchnahme von begleitenden Interventionen und deren Einfluss auf das Behandlungsergebnis hin	(3)
C.6.	Zulässigkeit begleitender Interventionen (z. B. Pharmakotherapie)	keine Einschränkungen	(1)
		begleitende in der Routine-Praxis übliche Interventionen teilweise ausgeschlossen	(2)
		alle begleitenden in der Routine-Praxis üblichen Interventionen ausgeschlossen	(3)
C.7a.	Qualifikation der Behandler: Klinische Tätigkeit der Therapeuten	Therapeuten sind praktizierende Kliniker	(1)
		klinische Forscher, die überwiegend Forschung betreiben und seltener auch Patienten behandeln; (Ausbildungskandidaten)	(2)
		keine Kliniker oder Kliniker, die keine Psychotherapeuten sind	(3)

C.7b.	Qualifikation der Behandler: Breite der Klinischen Tätigkeit der Therapeuten (Problemheterogenität)	Therapeut behandelt Patienten mit verschiedenen Problemen innerhalb und außerhalb der Studie	(1)
		Therapeut behandelt überwiegend Patienten mit ähnlichen Problemen innerhalb und außerhalb der Studie	(2)
		Therapeut behandelt nur Patienten mit ähnlichen Problemen innerhalb und außerhalb der Studie (z.B. Therapeut, der nur Schmerzpatienten in einer Schmerzklinik behandelt)	(3)
C.7c.	Qualifikation der Behandler: Spezifisches Training der Psychotherapeuten in einer Behandlungsmethode für die Studie	kein spezifisches Training für die Studie (z.B. Therapeuten wenden die von ihnen üblicherweise angewendete Therapie an)	(1)
		kurzes Training für die Studie / intensives Training nur einiger Therapeuten	(2)
		intensives Training vor der Studie	(3)

Outcomes/Outcomediagnostik:

A.8.	Reliable und valide Messung zumindest der primären Zielkriterien	reliable und valide Outcome-Verfahren (in Primärstudie werden ausschließlich Outcome-Verfahren verwendet, die mit **„exzellent“**, **„gut“** oder **„zufriedenstellend“** bewertet wurden)	(1)
		nur eingeschränkte Reliabilität und/oder Validität der Messverfahren (in Primärstudie werden - u.a. Outcome-Verfahren verwendet, die als **„unzureichend“** bewertet wurden, die jedoch einen Anteil von 25% der Gesamtanzahl verwendeter Outcomemaße nicht übersteigen *oder* - Outcome-Verfahren, von denen mindestens eines als **„ausreichend“** bewertet wurde)	(2)
		Reliabilität und Validität der Messverfahren nicht überprüft oder Gütekriterien der Messverfahren sind unzureichend (in Primärstudie werden mehr als 25% Outcome-Verfahren verwendet, die als **„unzureichend“** bewertet wurden) **(Stufe 3 = Ausschlusskriterium)**	(3)

	Wurde die Reliabilität (i.d.R. die Interrater-Reliabilität) in der Studie selbst untersucht und dargestellt?	ja	
		nein	
	Verändert die in der Studie berichtete Reliabilität die Gesamtbewertung eines/mehrerer Inventars/Inventare? Wenn ja, verbessernd oder verschlechternd?	ja	
		nein	
A.10.	Multiple Informationsquellen (z.B. Patient, Therapeut, Laborwerte)	multidimensionale Erfassung der Zielkriterien – drei oder mehr Informationsquellen	(1)
		zwei Informationsquellen	(2)
		eine Informationsquelle	(3)
A.11.	Sofern Fremdeinschätzungsverfahren: externe Beurteiler (blind für die Gruppenzugehörigkeit)	validiertes Fremdeinschätzungsverfahren angewendet von trainierten, für die Gruppenbedingungen blinden externen Beurteilern	(1)
		validiertes Fremdeinschätzungsverfahren angewendet von trainierten, nicht-blinden externen Beurteilern	(2)
		validiertes Fremdeinschätzungsverfahren angewendet – Rater sind weder trainiert noch blind für die Gruppenzugehörigkeit der Patienten	(3)
C.9.	Primäre Zielkriterien beziehen sich auf patientenrelevante Parameter (insbesondere Schwere der Symptomatik, Leiden, Beeinträchtigung/Lebensqualität, Inanspruchnahme von Diensten des Gesundheitswesens)	Zielkriterien beziehen sich auf mehrere Dimensionen patienten- bzw. störungsrelevanter Parameter unter Einbezug von Beeinträchtigung/Lebensqualität und Inanspruchnahme von Diensten des Gesundheitswesens)	(1)
		Zielkriterien beziehen sich nur auf eine Dimension	(2)
		Zielkriterien beziehen sich ausschließlich auf Surrogatparameter (z.B. Kontrollüberzeugung) **(Stufe 3 = Ausschlusskriterium)**	(3)

B.12.	Erzielte Veränderungen auf den primären und sekundären Zielkriterien ggf. im Vergleich zur Kontrollgruppe (Signifikanz, Größe und Relevanz der Effekte[93])	vollständige Darstellung der erzielten Veränderungen auf den Zielkriterien inklusive der Signifikanz, Größe der Effektmaße und Ausmaß der klinisch relevanten Zielerreichung (ggf. im Vergleich zur Kontrollgruppe)	(1)
		Darstellung des Behandlungsergebnisses nur durch Veränderungs- oder Zielerreichungsmaße oder beides ist (ggf. im Vergleich zur Kontrollgruppe) bei einigen Kriterien unvollständig	(2)
		weitgehend unvollständige oder inadäquate Darstellung der Outcome-Kriterien (ggf. im Vergleich zur KG) **(Stufe 3 = Ausschlusskriterium)**	(3)
A.9.	Klinische Bedeutsamkeit der Outcome-Messung (z.B. das Konzept der klinischen Signifikanz[94])	klinische Bedeutsamkeit des Therapieeffekts (z. B. im Sinne des Konzepts der klinischen Signifikanz) ist feststellbar	(1)
		klinische Bedeutsamkeit des Therapieeffekts (z. B. im Sinne des Konzepts der klinischen Signifikanz) ist nur eingeschränkt feststellbar	(2)
		klinische Bedeutsamkeit des (z. B. im Sinne des Konzepts der klinischen Signifikanz) Therapieeffekts ist nicht feststellbar	(3)

[93] *Veränderungen:* Neben Prä-Post-Differenzen:

- Prozentuale Angabe der Probanden, die entsprechend dem „Reliable Change Index" eine signifikante Veränderung erreicht haben
- Effektstärkeberechnungen (als Indikator des Ausmaßes der Veränderung)

Zielerreichung:

- Klinische Signifikanz: Reliable Change Index plus Cutoff-Wert (Vergleich mit Normpopulation oder dysfunktionaler Population)

(vgl. Schulte, 1993; Möller, Laux & Kapfhammer, 2005).

[94] Als Erhebung der klinischen Relevanz werden i.d.R. folgende Strategien betrachtet:

1. Normativer Vergleich mittels klassischer und Äquivalenzhypothesentests (vgl. Kendall, Marrs-Garcia, Nath & Sheldrick, 1999)
2. Individuelle Veränderungen (z.B. „Reliable Change Index" als „statistically reliable improvement" [Verbesserung]; „Cutoff-Wert" als „clinically significance change" [Genesung]) (vgl. Jacobson & Truax, 1991; Stieglitz, 2008)
3. Soziale Validierung/Subjektive Evaluation durch Patienten selbst, Verwandte/Freunde, Fachkräfte/externe Beobachter (vgl. Hill & Lambert, 2004; Kazdin, 1994)
4. Kein weiteres Erfüllen der Kriterien einer psychiatrischen Diagnose (vgl. Kazdin, 2008).

A.12.	Vollständige Darstellung der Ergebnisse zu allen Outcomemaßen und zu allen relevanten Messzeitpunkten	für alle Outcome-Variablen berichtet	(1)
		ausschließlich für die primären Zielkriterien berichtet	(2)
		nicht für alle primären Zielkriterien berichtet	(3)
A.13.	Erfassung unerwünschter Wirkungen	Systematische Erfassung und Bericht von unerwünschten Ereignissen, Nebenwirkungen, Verschlechterungen	(1)
		unsystematische Erfassung und/oder unvollständiger Bericht von unerwünschten Ereignisse, Nebenwirkungen, Verschlechterungen	(2)
		Unerwünschte Ereignisse, Nebenwirkungen, Verschlechterungen wurden nicht erfasst oder nicht berichtet	(3)

Drop-out-Analysen:

A.3.	Höhe der Drop-out-Quote zu Behandlungsende (sofern nicht Erfolgskriterium)	i.d.R. Drop-out-Quote kleiner 20 %	(1)
		i.d.R. Drop-out-Quote zwischen 20 % und 40 %	(2)
		i.d.R. Drop-out-Quote größer 40 %	(3)
	Vor allem in Bezug auf Studien <u>ohne</u> Randomisierung: Zeitpunkt, ab dem Drop-outs gezählt werden (Erstgespräch, probatorische Sitzungen, Therapiebeginn o.ä.):		

<table>
<tr><td rowspan="10">A.4.</td><td rowspan="10">Höhe der Studien-Drop-outs zur Katamnese (falls Katamneseerhebung durchgeführt) (zwischen Post- und Katamnesezeitpunkt)</td><td colspan="2">deutlich besser als in Studien mit vergleichbaren Patientengruppen und vergleichbarem Katamnesezeitraum</td><td rowspan="3">(1)</td></tr>
<tr><td>Affektive Störungen</td><td>Gemischte Störungsgruppen</td></tr>
<tr><td>weniger als 2 Jahre: ≤ 10%</td><td>weniger als 2 Jahre: ≤ 20%</td></tr>
<tr><td>2 Jahre oder mehr: ≤ 15%</td><td>2 Jahre oder mehr: ≤ 25%</td></tr>
<tr><td colspan="2">Drop-out-Quote vergleichbar mit Studien mit entsprechenden Patien-tengruppen und entsprechendem Katamnesezeitraum</td><td rowspan="3">(2)</td></tr>
<tr><td>Affektive Störungen</td><td>Gemischte Störungsgruppen</td></tr>
<tr><td>weniger als 2 Jahre: 11-20%</td><td>weniger als 2 Jahre: 21-40%</td></tr>
<tr><td>2 Jahre oder mehr: 16-30%</td><td>2 Jahre oder mehr: 26-50%</td></tr>
<tr><td colspan="2">deutlich schlechter als in Studien mit vergleichbaren Patientengruppen und vergleichbarem Katamnesezeitraum</td><td rowspan="3">(3)</td></tr>
<tr><td>Affektive Störungen</td><td>Gemischte Störungsgruppen</td></tr>
<tr><td>weniger als 2 Jahre: > 20%</td><td>weniger als 2 Jahre: > 40%</td></tr>
<tr><td>2 Jahre oder mehr: > 30%</td><td>2 Jahre oder mehr: > 50%</td></tr>
<tr><td rowspan="3">A.16.</td><td rowspan="3">Intention to treat–Analysen durchgeführt</td><td colspan="2">ITT-Analysen durchgeführt</td><td>(1)</td></tr>
<tr><td colspan="2">Keine ITT-Analysen bei geringem Risiko für einen attrition bias (wenn nahezu ausschließlich Drop-outs vorliegen, von denen mit hoher Wahrscheinlichkeit keine Verzerrungen des Behandlungsergebnisses zu erwarten sind [z.B. Wohnortwechsel oder Versterben des Patienten; Ausfall des behandelnden Therapeuten[95]])</td><td>(2)</td></tr>
<tr><td colspan="2">Keine ITT-Analysen bei deutlichem Risiko für einen attrition bias (wenn nahezu ausschließlich Drop-outs vorliegen, von denen mit hoher Wahrscheinlichkeit Verzerrungen des Behandlungsergebnisses zu erwarten sind [z.B. Therapieabbruch seitens Patienten aus therapiebezogenen Gründen; Abbruch seitens Therapeuten wg. patientenseitigen Motivationsmangels; Therapiewechsel; zu geringer Therapieerfolg; ohne Angabe von Gründen])</td><td>(3)</td></tr>
</table>

[95] Vgl. Hiller, Bleichhardt & Schindler (2009).

A.18.	Vollständige Beschreibung der Drop-Outs	vollständige Beschreibung aller Drop-outs, inkl. der Gründe und des Zeitpunkts des Drop-outs	(1)
		unvollständige Angaben zu Gründen oder Zeitpunkten des Drop-outs	(2)
		keine Beschreibung der Drop-outs oder definitive Angabe der Anzahl der Drop-outs in einer Gruppe	(3)
A.19.	Drop-out-Analysen	Drop-out-Analysen unter Berücksichtigung der wichtigsten prognostischen Faktoren durchgeführt; keine signifikanten Unterschiede zwischen den Vergleichsgruppen	(1)
		Drop-out-Analysen unter Berücksichtigung wichtigster prognostischer Faktoren durchgeführt; trotz sign. Unterschiede ist die Validität der Ergebnisse nicht wesentlich eingeschränkt	(2)
		Drop-out-Analysen nicht oder unter Vernachlässigung relevanter prognostischer Merkmale durchgeführt; Drop-out-Analysen stellen die Validität der Ergebnisse deutlich in Frage	(3)

Statistische Methodik:

A.14.	Anwendungsvoraussetzungen für statistische Modelle geprüft und erfüllt	Anwendungsvoraussetzungen geprüft und erfüllt	(1)
		Anwendungsvoraussetzungen geprüft und lediglich leichte Verletzungen der Voraussetzungen oder keine Prüfung der Anwendungsvoraussetzungen bei gleichzeitiger Robustheit der angewendeten statistischen Verfahren	(2)
		deutliche Verletzungen der Anwendungsvoraussetzungen oder keine Prüfung der Anwendungsvoraussetzungen bei substanziellem Risiko für deren Verletzung	(3)
A.15.	Angemessenheit der statistischen Analysen (inklusive der Korrektur für multiple Tests)	adäquate und umfassende statistische Analysen	(1)
		weitgehend adäquate statistische Analysen	(2)
		unangemessene statistische Analysen (fehlende Korrektur für multiple Tests, inadäquate statistische Methoden)	(3)

A.17.	Statistische Power der Vergleiche bei Vergleich mit bewährter Treatment-Gruppe	adäquate statistische Power der Vergleiche Als Richtwert für einen t-Test für unabhängige Stichproben: N (pro Gruppe) >310 bzw. N (gesamt) >620 [96]	(1)
		eingeschränkte Power der statistischen Vergleiche (.50-.80) Als Richtwert für einen t-Test für unabhängige Stichproben: N (pro Gruppe): 136 bis 310 bzw. N (gesamt) 272 bis 620	(2)
		unzureichende statistische Power der Vergleiche (<.50) Als Richtwert für einen t-Test für unabhängige Stichproben: N (pro Gruppe) <136 bzw. N (gesamt) <272	(3)
	Wurde die Signifikanzüberprüfung mittels eines Intervall-Nullhypothesentests (einseitiger Äquivalenztest/Non-Inferiority) oder mittels eines nil-Nullhypothesentests durchgeführt?		

Manipulation der Daten:

A.1.	Manipulation der Daten	keine Hinweise auf Ergebnismanipulation	(1)
		-	(2)
		Hinweise auf Ergebnismanipulation **(Stufe 3 = Ausschlusskriterium)**	(3)

[96] Die kalkulierten Stichprobengrößen gelten für gleichgroße Stichproben. Für ungleichgroße Stichproben werden gesonderte Kalkulationen angestellt.

Subgruppenanalysen:

Wurden Subgruppenanalysen durchgeführt?

ja	
nein	

Welche?

Anhang C: Kodierregeln zum Kriterienkatalog des Methodenpapiers 2.8 des Wissenschaftlichen Beirats Psychotherapie nach § 11 PsychThG

(aus: Ratzek & von Hauenschild, 2011a)

Studiendesign:

B.8.	Gruppenzuweisung	angemessene Randomisierung (inkl. Cluster-Randomisierung)bei ausreichender Stichprobengröße (n>30/Gruppe[97]), die das Gelingen der Randomisierung hinsichtlich bekannter und unbekannter (nicht erfasster) prognostisch relevanter Merkmale sicherstellt	(1)	**Eingruppendesigns und verfahrensinterne Vergleichsgruppen werden hier mit „3“ geratet.** Mit „n>30/Gruppe“ sind die Größen pro Gruppe im Rahmen der randomisierten Zuteilung gemeint.
		Parallelisierung oder teilweise randomisiert oder quasi-randomisiert oder Stichprobengröße n<30/Gruppe	(2)	Quasi-randomisiert: z.B. jeder 2. aufgenommene Patient gelangt in Gruppe A.
		keine randomisierte oder parallelisierte Zuweisung **(obligatorisches Kriterium für interne Validität (<3))**	(3)	-
C.8.	Repräsentativität der patientenseitigen Freiheit hinsichtlich der Wahl der Intervention Echte Warteliste: „9“	Patienten entscheiden sich selbst für eine der angebotenen Therapieformen	(1)	**Auch bei Eingruppendesigns (inkl. Designs mit verfahrensinternen Vergleichsgruppen) zu raten!**
		ein Teil der Patienten (z. B. Randomisierungswillige) wird der Therapie zufällig zugewiesen	(2)	
		alle Patienten werden der Therapie (zufällig) zugewiesen	(3)	
A.5.	Stichprobengröße pro Gruppe	n pro Gruppe > 30	(1)	Gemeint sind die Stichprobengrößen zu Beginn (ab Randomisierung oder sonstiger Gruppenzuteilung)
		n pro Gruppe 10-30	(2)	
		n pro Gruppe < 10	(3)	

[97] es handelt sich um Anhaltszahlen.

A.6.	Vergleiche der (sofern vorhanden) Behandlungsgruppen und der Messzeitpunkte a priori definiert	a priori Definition der Vergleiche der (sofern vorhanden) Behandlungsgruppen und Messzeitpunkte erfüllt	(1)	**Bei Eingruppendesigns nur auf Messzeitpunkte bezogen!** A priori Definition der Vergleichsgruppen und Messzeitpunkte
		teilweise post-hoc Definition der Vergleiche	(2)	Teilweise post-hoc Definition der Vergleichsgruppen und Messzeitpunkte
		ausschließlich post-hoc definierte Vergleiche	(3)	Ausschließlich post-hoc Definition der Vergleichsgruppen und Messzeitpunkte
A.7.	a priori Definition der primären und sekundären Zielkriterien	a priori definierte primäre und gegebenenfalls sekundäre Zielkriterien	(1)	A priori Definition *und* Differenzierung primärer und sekundärer Zielkriterien
		a priori Nennung der Zielkriterien ohne Differenzierung in primäre und sekundäre Zielkriterien	(2)	A priori Definition aber keine Differenzierung primärer und sekundärer Zielkriterien
		a posteriori Definition der Zielkriterien; keine Angaben zu Zielkriterien	(3)	Alle Zielkriterien a posteriori definiert *oder* keine Angabe zu Zielkriterien
B.10.	Definition der Messzeitpunkte (Prospektive Messung; Follow-up-Messung) • Prä-Katamnesemessung inkl. mehrerer Messzeitpunkte: „4“ • Prä-Katamnesemessung: „5“ • Nur Katamnesemessung: „6“	mehrere vorab festgelegte Messzeitpunkte über den Therapieverlauf incl. Prä-Post-Messungen	(1)	Mehrere prospektive Messungen inkl. Prä-Post-Messung z.B. SCL-90-R, GBB, BL etc.
		ausschließlich Prä-Post-Messung	(2)	Nur Prä-Post-Messung, überwiegend prospektive Messungen z.B. SCL-90-R, GBB, BL etc.
		ausschließlich Post-Messung	(3)	Nur Post-Messung *oder* überwiegend retrospektive Messungen (z.B. „wie stark waren Ihre Symptome zu Beginn der Therapie? Inwieweit hat sich das Problem seither verändert?“, bzw. GAS oder VEV)

B.11.	Follow-up-Messung	zeitlich störungsangemessene Katamnese mit hoher Ausschöpfung[98] der Stichprobe • mindestens 2 Jahre (2 Jahre oder mehr) • mindestens 60%ige Ausschöpfungsquote	(1)	Es müssen <u>beide</u> Bedingungen erfüllt sein (Ausschöpfungsquote und Katamnesezeitraum) Die Ausschöpfungsquote bezieht sich auf alle Untersuchungsgruppen
		Katamnese mit fraglich angemessenem Zeitraum bzw. niedriger Ausschöpfung der Stichprobe • weniger als 2 Jahre • weniger als 60% Ausschöpfungsquote	(2)	Sobald <u>eine</u> der beiden Bedingungen nicht erfüllt ist → (2)
		keine Katamnese	(3)	Keine Katamnese
	Katamnesenzeitraum (Zeitraum zwischen Beendigung der Therapie und letzter Katamneseerhebung):			-
	Katamnesemessungen auf <u>allen</u> Treatmentarmen?	ja		-
		nein		
	Weitere/r katamnestische/r Messzeitpunkt/e (vor letzter Katamnesemessung liegend)? Wenn ja, wann?	ja		-
		nein		
	Einbeziehung longitudinaler Informationen: Wurde die Inanspruchnahme weiterer psychotherapeutischer Interventionen über das gesamte Katamneseintervall erhoben (Inanspruchnahmedaten, Arztbriefe, Krankenkassendaten)?	ja		Es sind in erster Linie psychotherapeutische oder psychiatrische Inanspruchnahmen gemeint
		nein		
	Direkte persönliche Untersuchung (vs. Telefon- oder postalische Fragebogenuntersuchung)?	ja		-
		nein		
	Erhebung explizit psychodynamischer Outcomekriterien (Konfliktdynamiken, strukturelle Aspekte)?	ja		-
		nein		
	Stabilitätsmessung? (d.h. differenzierte Betrachtung der Responder und Non-Responder)	ja		Sind die Responder zum Postzeitpunkt bis zur Katamnese stabil geblieben?
		nein		
	Wird der in der Studie gewählte Katamnesezeitraum explizit begründet? Wenn ja, wie?	ja		-

[98] Gemeint ist die Ausschöpfung der Stichprobe zwischen Prä- und Katamnesemessung.

C.10.	Spezifikation und Herstellbarkeit notwendiger Settingbedingungen (Praxistransfer)	Notwendige Settingbedingungen herstellbar (z. B. Infrastruktur, Kooperation, Team)	(1)	Hier werden besonders aufwändige Behandlungen, in die etwa mehrere Teammitglieder integriert werden und die einer besonderen Infrastruktur (bestimmte Räumlichkeiten, videografische Ausstattung etc.) bedürfen, mit (3) bewertet
		Notwendige Settingbedingungen nur begrenzt herstellbar (z. B. Infrastruktur, Kooperation, Team)	(2)	
		Notwendige Settingbedingungen nicht herstellbar	(3)	
C.11.	Spezifikation und Herstellbarkeit der notwendigen Behandlerqualifikation (Praxistransfer)	Notwendige Behandlungsqualifikation eindeutig beschrieben und herstellbar	(1)	z.B. die reguläre Ausbildungspflicht
		Notwendige Behandlungsqualifikation eindeutig beschrieben, aber nur mit sehr großem Zeitaufwand herstellbar	(2)	Hier sind nicht die regulären Ausbildungszeiten zum PP/ÄP gemeint, sondern die über die Fortbildungspflicht eines PP/ÄP deutlich hinausgehenden
		Notwendige Behandlungsqualifikation nicht beschrieben oder praktisch nicht herstellbar	(3)	-
C.12.	Spezifikation und Erfassbarkeit relevanter Patientenmerkmale (Praxistransfer)	Relevante Patientenmerkmale (z.B. Alter, genetische Marker) praktisch erfassbar	(1)	-
		Relevante Patientenmerkmale nur mit erheblichem Aufwand erfassbar	(2)	bspw. strukturelle Beeinträchtigungen mittels aufwändiger psychodiagnostischer Erhebungsinstrumente
		Relevante Patientenmerkmale praktisch nicht erfassbar	(3)	bspw. bildgebende Verfahren
C.13.	Spezifikation und Herstellbarkeit relevanter Treatmentaspekte (Praxistransfer)	Relevante Treatmentmerkmale (Art der Interventionen, Reihenfolge, Dauer) praktisch herstellbar (z.B. durch Manual)	(1)	Behandlung in Versorgungspraxis/-alltag integrierbar
		Relevante Treatmentmerkmale nur schwer herstellbar (z.B. tägliche Behandlung, Parallelbehandlung)	(2)	In Versorgungspraxis/-alltag schwierig integrierbar (z.B. tägliche Behandlung)
		Relevante Treatmentmerkmale praktisch nicht herstellbar	(3)	z.B. 2x tägliche Behandlung oder unnachvollziehbare eklektische Behandlung

Population/Patienten:

C.1.	Stichprobe von Patienten mit Störungen mit Krankheitswert	ausschließlich Patienten mit Störung mit Krankheitswert (z. B. ICD-, DSM-Diagnosen)	(1)	-
		Stichprobe von Patienten mit wahrscheinlicher klinischer Störung (z. B. Menschen nach Trauma) oder bis zu (maximal) 20% der Patienten mit lediglich erhöhter Symptomausprägung, z.T. subklinisch	(2)	-
		Patienten ohne festgestellte Störung mit Krankheitswert **(Stufe 3 = Ausschlusskriterium)**	(3)	-
C.2.	Art der Rekrutierung der Stichprobe	Patientenzugang überwiegend durch gängige klinische Routinen (Überweisung , Primärzugang, etc.); keine Selektionseffekte aufgrund der Zugangswege	(1)	**Es steht der "Zugang zur Behandlung" im Vordergrund:** Neben „gängige klinische Routine“: • Wenn angeschriebene Studientherapeuten eines/r bestimmten Verbandes/Fachgesellschaft (DGPT, DPG etc.) ihre Pat. zur Studie vermitteln sollen → 1 (keine Selektionseffekte i.S.v. "Zugang zur Behandlung")
		Patientenzugang überwiegend durch gängige klinische Routinen (Überweisung , Primärzugang, etc.); Selektionseffekte aufgrund der Zugangswege	(2)	• Wenn Zugang über Hochschulambulanz, bestimmte Krankenkassen etc., → 2 (Selektionseffekte i.S.v. "Zugang zur Behandlung")
		Patientenzugang überwiegend über Aufforderungen der Forschergruppe (z. B. Anzeigenwerbung)	(3)	-

B.1.	Spezifizierung der Einschluss- und Ausschlusskriterien	eindeutige Spezifizierung der Ein- und Ausschlusskriterien	(1)	-
		Ein- oder Ausschlusskriterien teilweise uneindeutig beschrieben	(2)	
		Ein- und Ausschlusskriterien sind nicht eindeutig definiert	(3)	
	Ein-/Ausschlusskriterien:			
C.3.	Selektivität der Stichprobe aufgrund der Ausschlusskriterien	keine Selektionseffekte aufgrund der Ausschlusskriterien: Einschluss aller Patienten	(1)	Einschluss aller Patienten
		mittlere Selektionseffekte aufgrund der Ausschlusskriterien (z. B. Ausschluss einiger epidemiologisch relevanter komorbider Störungen)	(2)	z.B. Ausschluss einiger epidemiologisch relevanter komorbider Störungen (z.B. bei Affektiven Störungen, Angststörungen etc.)
		deutliche Selektionseffekte aufgrund der Ausschlusskriterien	(3)	z.B. Ausschluss zahlreicher/aller komorbider Störungen

Eingangsdiagnostik:

A.2.	Objektive und reliable Diagnosestellung (mittels (teil-) standardisierter Interviews)	Diagnosestellung mittels strukturiertem klinischen bzw. voll standardisiertem Interview (z.B. SKID, DIPS)	(1)	z.B. SKID, DIPS, DIA-X, CIDI etc.
		Diagnosestellung mittels Diagnosechecklisten oder nachvollziehbarem klinischen Urteil	(2)	z.B. Diagnosechecklisten (ICD- oder DSM-Diagnosechecklisten) oder globales klinisches Urteil nach erfolgter Anamnese (meist in Arztbrief festgehalten)
		keine adäquate Diagnosestellung **(Stufe 3 = Ausschlusskriterium)**	(3)	Keine Angaben zur Diagnosestellung
	Diagnostische Instrumente:			Hier sind keine zusätzlichen Erhebungsinstrumente gemeint, die im Rahmen des Ein-Ausschlusskatalogs eingesetzt werden (z.B. BDI)

B.2.	Erhebung der spezifizierten Einschluss- und Ausschlusskriterien mittels valider Methoden	Die Ein- und Ausschlusskriterien sind sämtlich klar operationalisiert und werden mittels valider Methoden erfasst (z. B. komorbide Störungen als Ausschluss werden mittels strukturiertem klinischen Interview erfasst; Ausschlusskriterien beziehen sich auf eindeutig objektivierbare Merkmale wie Alter, Geschlecht etc.)	(1)	Komorbide Störungen mittels klinischer Interviews, wie z.B. SKID, DIPS, DIA-X, CIDI etc. z.B. Schweregradeinstufung diagnostizierter Störungen mittels dafür entwickelter mindestens „zufriedenstellend“ <u>valider</u> Inventare (z.B. BDI) *oder* nur objektivierbare Merkmale
		Die Validität der Erhebungen von Teilen der Ein- oder Ausschlusskriterien ist teilweise eingeschränkt (z. B. komorbide Störungen als Ausschluss werden mittels globalem klinischen Urteil eingeschätzt) und wirkt sich jedoch nur in geringem Umfang auf die Zusammensetzung der Stichprobe aus	(2)	z.B. globales klinisches Urteil oder Diagnosechecklisten nach anamnestischem Interview (ICD- oder DSM-Diagnosechecklisten) (meist in Arztbrief festgehalten) z.B. Schweregradeinstufung diagnostizierter Störungen mittels dafür entwickelter lediglich „ausreichend“ <u>valider</u> Inventare
		Die Validität der Erhebungen von Teilen der Ein- oder Ausschlusskriterien ist deutlich eingeschränkt und wirkt sich differenziell auf die Zusammensetzung der Behandlungsgruppen aus	(3)	keine adäquate Diagnosestellung (Diagnosestellung nicht nachvollziehbar) z.B. Schweregradeinstufung diagnostizierter Störungen mittels dafür entwickelter „unzureichend“ <u>valider</u> Inventare *oder* Wenn B.1.=3 → 3

B.9.	Vergleichbarkeit der Gruppen zur Baseline im Hinblick auf prognostisch relevante Merkmale Eingruppendesigns (inkl. Designs mit verfahrensinternen Vergleichsgruppen): „9“	Weder statistisch noch klinisch relevante Unterschiede zwischen den Gruppen hinsichtlich prognostisch relevanter oder potentiell konfundierender Variablen	(1)	-
		Vergleichbarkeit hinsichtlich der meisten prognostisch relevanten Merkmale ist weitgehend gegeben; signifikante Unterschiede hinsichtlich relevanter prognostischer Merkmale zwischen den Gruppen werden statistisch angemessen kontrolliert	(2)	-
		Keine angemessene Überprüfung der Vergleichbarkeit oder Vergleichsgruppen unterscheiden sich erheblich hinsichtlich mehrerer prognostisch relevanter Merkmale und eine angemessene statistische Kontrolle des Einflusses dieser Merkmale fehlt	(3)	-
	Hinsichtlich Prognostizität untersuchte Merkmale:			-

Intervention:

B.3.	Operationale Definition der Interventionen (Experimental- und ggf. Kontrollgruppe)	Therapiemanual, bei dem die Interventionen so beschrieben sind, dass das therapeutische Vorgehen vergleichbar und replizierbar ist	(1)	Verweis auf Manuale/ manualähnliche Behandlungsrichtlinien, das/die in den jeweiligen Treatments auch offensichtlich eingesetzt wurden [99]
		Therapiebeschreibung, ohne nähere Spezifikation der einzelnen Interventionen (z.B. Fehlen gegebenenfalls für das psychotherapeutische Verfahren oder die psychotherapeutische Methode notwendiger Entscheidungskriterien)	(2)	Kompakte Beschreibung der Treatments im Rahmen der Studienpublikation *oder* Wenn Manual/ manualähnliche Behandlungsrichtlinie nur bsp.haft angeführt wird
		Die Intervention ist nicht klar beschrieben, beschränkt sich auf die Benennung des Psychotherapieverfahrens bzw. der Psychotherapiemethode	(3)	Lediglich Benennung der Behandlung(en), etwa „psychodynamische Therapie“ plus ggf. verfahrenstypischer Aspekte wie „Arbeit mit unbewussten Konflikten“ o.ä. *oder* Wenn kein bekannter Manual-/Behandlungsrichtlinientitel, dann danach recherchieren. Ggf. nachsehen ob Intervention in Artikel beschrieben → 2 (sonst 3)
	Welches Manual bzw. welche manualähnlichen Behandlungsrichtlinien wurden verwendet?			-
B.6.	Manualtreue, Treatment Integrity	Manualtreue/ Treatmentintegrität durch externe Beobachter (z.B. videogestützt) belegt	(1)	Hier gilt die Kontrolle der Manualtreue/Treatmentintegrität auch für die ggf. realisierte Vergleichsbehandlung (komparative Studien) und ebenfalls für Studien ohne Manual
		Manualtreue / Treatmentintegrität durch Fragebögen belegt	(2)	
		keine Maßnahmen zum Monitoring der Manualtreue oder Hinweise auf substanzielle Abweichungen	(3)	

[99] Bzgl. psychoanalytisch begründeter Behandlungen vgl. Beutel, Doering, Leichsenring & Reich (2010); DGPT (2009); Reimer & Rüger (2006).

C.4.	Klinische Repräsentativität der Intervention hinsichtlich Vorgehen und Dauer	Intervention wie in klinischer Alltagspraxis	(1)	**Bewertung entsprechend Psychotherapierichtlinien primär die <u>Therapiedauer</u> betreffend. Vorsicht bei Bewertung des <u>Vorgehens</u>, da selbst innerhalb Richtlinienverfahren uneinheitlich vorgegangen wird.** Hinsichtlich Dauer wie in Alltagspraxis des untersuchten Verfahrens (z.B. TP: mindestens 25 Sitzungen) Wenn nur <u>Range</u> der Dauer abweicht (Hinweis auf Ausreißer) → 1
		Intervention gegenüber klinischer Alltagspraxis teilweise verändert	(2)	Leichte Abweichungen hinsichtlich Dauer des untersuchten Verfahrens, z.B. wenn <u>SD</u> der Dauer abweicht
		Intervention gegenüber klinischer Alltagspraxis stark verändert	(3)	Starke Abweichungen hinsichtlich Dauer des untersuchten Verfahrens (z.B. TP: 8 Sitzungen), bzw. wenn <u>mittlere</u> Dauer von max. Richtliniendauer (100 TP, 300 AP) nach oben hin abweicht
C.5.	Art des Therapie-Monitorings (Einfluss auf Therapeutenverhalten)	keine Veränderung des Therapeutenverhaltens durch Therapie-Monitoring (z.B. durch Therapie-Supervision; Ausnahmen: Audio- oder Video-Aufzeichnungen ohne Feedback an Therapeuten, nur zur späteren Auswertung sind erlaubt)	(1)	Therapiesupervision, die explizit zum Zwecke der Studie durchgeführt wird, wird als das Therapeutenverhalten beeinflussend betrachtet → (2) oder (3) Keinen Einfluss: • reguläre Supervision/ Intervision • Audio-/Videoaufzeichnungen ohne Feedback an Therapeuten
		mittlere Veränderung des Therapeutenverhaltens durch Therapie-Monitoring (gelegentliche Rückmeldungen an Therapeuten)	(2)	gelegentliche, unregelmäßige Rückmeldungen bei Bedarf
		starke Veränderung des Therapeutenverhaltens durch Therapie-Monitoring (durch kontinuierliche Rückmeldungen)	(3)	kontinuierliche, regelmäßige Rückmeldung

B.4.	Operationale Definition der Kontrollbedingungen • Eingruppendesigns (inkl. Designs mit verfahrensinternen Vergleichsgruppen): „9“ • Echte VG: „8“	Prospektive Festlegung und umfassende Beschreibung der Kontrollbedingung	(1)	**Hier sind explizit Kontrollbedingungen gemeint!** Prospektive Festlegung und umfassende Beschreibung
		Ex post facto Beschreibung der Kontrollbedingungen	(2)	Ex post facto Beschreibung (Beschreibung im Rahmen eines retrospektiven Designs)
		keine Beschreibung der Kontrollbedingung	(3)	keine Beschreibung
B.5.	Strukturelle Äquivalenz bei Kontrollbedingungen • Eingruppendesigns (inkl. Designs mit verfahrensinternen Vergleichsgruppen): „9“ • Echte VG: „8“	hinsichtlich des Umfangs an therapeutischer Zuwendung und der Settingbedingungen in der KG besteht Äquivalenz	(1)	**Hier sind explizit Kontrollbedingungen gemeint!**
		der Umfang der therapeutischen Zuwendung in der KG ist reduziert, die Settingbedingungen weichen von der IG ab	(2)	
		der Umfang der therapeutischen Zuwendung in der KG ist deutlich reduziert, die Settingbedingungen weichen wesentlich von der IG ab	(3)	
B.7.	Zulässigkeit, Dokumentation und Analyse des Einflusses begleitender nicht-randomisierter Interventionen (insbesondere Pharmakotherapie) Eingruppendesigns (inkl. Designs mit verfahrensinternen Vergleichsgruppen): „9“	Ausschluss begleitender nicht-randomisierter Interventionen	(1)	Eliminierung der Störvariablen
		begleitende nicht-randomisierte Interventionen sind zulässig, werden jedoch detailliert dokumentiert und die Analysen weisen auf keinen substanziellen, differenziellen Einfluss der begleitenden Interventionen hin	(2)	Dokumentation und Analyse und ggf. statistische Kontrollen der begleitenden nicht-randomisierten Interventionen
		begleitende nicht-randomisierte Interventionen sind zulässig, werden jedoch nicht dokumentiert oder die Analysen weisen auf eine differenzielle Inanspruchnahme von begleitenden Interventionen und deren Einfluss auf das Behandlungsergebnis hin	(3)	Keine Dokumentation *oder* Dokumentation und Analyse weisen auf differenziellen Effekt der nicht-randomisierten Begleitintervention hin und es wird keine statistische Kontrolle durchgeführt
C.6.	Zulässigkeit begleitender Interventionen (z. B. Pharmakotherapie)	keine Einschränkungen	(1)	-
		begleitende in der Routine-Praxis übliche Interventionen teilweise ausgeschlossen	(2)	-
		alle begleitenden in der Routine-Praxis üblichen Interventionen ausgeschlossen	(3)	-

C.7a.	Qualifikation der Behandler: Klinische Tätigkeit der Therapeuten	Therapeuten sind praktizierende Kliniker	(1)	Behandler: mit abgeschlossener Psychotherapieausbildung und (hauptberuflich) praktizierend
		klinische Forscher, die überwiegend Forschung betreiben und seltener auch Patienten behandeln; (Ausbildungskandidaten)	(2)	Behandler: in Psychotherapieausbildung *oder* mit abgeschlossener Psychotherapieausbildung, jedoch hauptsächlich in Forschung statt Klinik tätig
		keine Kliniker oder Kliniker, die keine Psychotherapeuten sind	(3)	Behandler: ohne Psychotherapieausbildung
C.7b.	Qualifikation der Behandler: Breite der Klinischen Tätigkeit der Therapeuten (Problemheterogenität)	Therapeut behandelt Patienten mit verschiedenen Problemen innerhalb und außerhalb der Studie	(1)	Tätigkeit in breitem Anwendungsbereich, keine Tätigkeit in nur einem abgegrenzten Bereich (Sucht, Schmerz, Essstörungen)
		Therapeut behandelt überwiegend Patienten mit ähnlichen Problemen innerhalb und außerhalb der Studie	(2)	Tätigkeit hauptsächlich in einem begrenzten Anwendungsbereich, jedoch durchaus mehrere Störungsgruppen
		Therapeut behandelt nur Patienten mit ähnlichen Problemen innerhalb und außerhalb der Studie (z.B. Therapeut, der nur Schmerzpatienten in einer Schmerzklinik behandelt; Psychotherapeuten, die außerhalb der Studie keine Pat. behandeln)	(3)	Tätigkeit in abgegrenztem Anwendungsbereich (Sucht, Schmerz, Essstörungen) oder keine therapeutische Tätigkeit außerhalb der Studie
C.7c.	Qualifikation der Behandler: Spezifisches Training der Psychotherapeuten in einer Behandlungsmethode für die Studie	kein spezifisches Training für die Studie (z.B. Therapeuten wenden die von ihnen üblicherweise angewendete Therapie an)	(1)	-
		kurzes Training für die Studie / intensives Training nur einiger Therapeuten	(2)	-
		intensives Training vor der Studie	(3)	-

Outcomes/Outcomediagnostik:

A.8.	Reliable und valide Messung zumindest der **primären** Zielkriterien	reliable und valide Outcome-Verfahren	(1)	In Primärstudie werden ausschließlich Outcome-Verfahren verwendet, die mit **„exzellent“**, **„gut“** oder **„zufriedenstellend“** bewertet wurden
		nur eingeschränkte Reliabilität und/oder Validität der Messverfahren	(2)	In Primärstudie werden • u.a. Outcome-Verfahren verwendet, die als **„unzureichend“** bewertet wurden, die jedoch einen Anteil von 25% der Gesamtanzahl verwendeter Outcomemaße nicht übersteigen *oder* • Outcome-Verfahren, von denen mindestens eines als **„ausreichend“** bewertet wurde
		Reliabilität und Validität der Messverfahren nicht überprüft oder Gütekriterien der Messverfahren sind unzureichend **(Stufe 3 = Ausschlusskriterium)**	(3)	In Primärstudie werden mehr als 25% Outcome-Verfahren verwendet, die als **„unzureichend“** bewertet wurden
	Wurde die Reliabilität (i.d.R. die Interrater-Reliabilität) in der Studie selbst untersucht und dargestellt?	ja		-
		nein		-
	Verändert die in der Studie berichtete Reliabilität die Gesamtbewertung eines/mehrerer Inventars/Inventare? Wenn ja, verbessernd oder verschlechternd?	ja		-
		nein		-
A.10.	Multiple Informationsquellen (z.B. Patient, Therapeut, Laborwerte)	multidimensionale Erfassung der Zielkriterien – drei oder mehr Informationsquellen	(1)	**Auf primäre und sekundäre Zielkriterien bezogen** z.B. Selbstbeurteilung Patient, Fremdbeurteilung Therapeut, Fremdbeurteilung durch externe Beurteiler (Kliniker oder Angehörige)
		zwei Informationsquellen	(2)	-
		eine Informationsquelle	(3)	-

A.11.	Sofern Fremdeinschätzungsverfahren: externe Beurteiler (blind für die Gruppenzugehörigkeit) • Wenn keine Fremdeinschätzungsverfahren durch externe Beurteiler eingesetzt: „9“ • Eingruppendesigns (inkl. Designs mit verfahrensinternen Vergleichsgruppen): mit externen Beurteilern: „4“	validiertes Fremdeinschätzungsverfahren angewendet von trainierten, für die Gruppenbedingungen blinden externen Beurteilern	(1)	**Auf primäre und sekundäre Zielkriterien bezogen** Validiert, trainiert und blind
		validiertes Fremdeinschätzungsverfahren angewendet von trainierten, nicht-blinden externen Beurteilern	(2)	Validiert, trainiert aber nicht-blind
		validiertes Fremdeinschätzungsverfahren angewendet – Rater sind weder trainiert noch blind für die Gruppenzugehörigkeit der Patienten	(3)	Validiert aber nicht-trainiert und nicht-blind *oder* nicht-valide entsprechend A.8.
C.9.	Primäre Zielkriterien beziehen sich auf patientenrelevante Parameter (insbesondere Schwere der Symptomatik, Leiden, Beeinträchtigung/ Lebensqualität, Inanspruchnahme von Diensten des Gesundheitswesens)	Zielkriterien beziehen sich auf mehrere Dimensionen patienten- bzw. störungsrelevanter Parameter unter Einbezug von Beeinträchtigung/Lebensqualität und Inanspruchnahme von Diensten des Gesundheitswesens)	(1)	z.B. symptom-, persönlichkeits-, strukturbezogene Outcomemaße, Outcomemaße zur Lebensqualität, Inanspruchnahmedaten
		Zielkriterien beziehen sich nur auf eine Dimension	(2)	z.B. symptom-, persönlichkeits-, strukturbezogene Outcomemaße, aber weder Outcomemaße zur Lebensqualität noch zur Inanspruchnahme
		Zielkriterien beziehen sich ausschließlich auf Surrogatparameter (z.B. Kontrollüberzeugung) **(Stufe 3 = Ausschlusskriterium)**	(3)	-

B.12.	Erzielte Veränderungen auf den primären und sekundären Zielkriterien ggf. im Vergleich zur Kontrollgruppe (Signifikanz, Größe und Relevanz der Effekte [100])	vollständige Darstellung der erzielten Veränderungen auf den Zielkriterien inklusive der Signifikanz, Größe der Effektmaße und Ausmaß der klinisch relevanten Zielerreichung (ggf. im Vergleich zur Kontrollgruppe)	(1)	Sowohl Indikatoren der *Veränderungen* als auch der *Zielerreichungen* sind in Bezug auf die verwendeten Outcomemaße dargestellt **Achtung:** *Zielerreichung* weniger streng bewerten, wenn Cutoff angewandt
		Darstellung des Behandlungsergebnisses nur durch Veränderungs- oder Zielerreichungsmaße oder beides ist (ggf. im Vergleich zur Kontrollgruppe) bei einigen Kriterien unvollständig	(2)	Es sind nur Indikatoren der *Veränderungen* oder nur der *Zielerreichungen* in Bezug auf die verwendeten Outcomemaße dargestellt
		weitgehend unvollständige oder inadäquate Darstellung der Outcome-Kriterien (ggf. im Vergleich zur KG) **(Stufe 3 = Ausschlusskriterium)**	(3)	-

[100] *Veränderungen:* Neben Prä-Post-Differenzen:

- Prozentuale Angabe der Probanden, die entsprechend dem „Reliable Change Index“ eine signifikante Veränderung erreicht haben
- Effektstärkeberechnungen (als Indikator des Ausmaßes der Veränderung)

Zielerreichung:

- Klinische Signifikanz: Reliable Change Index plus Cutoff-Wert (Vergleich mit Normpopulation oder dysfunktionaler Population)

(vgl. Schulte, 1993; Möller, Laux & Kapfhammer, 2005).

A.9.	Klinische Bedeutsamkeit der Outcome-Messung (z.B. das Konzept der klinischen Signifikanz [101])	klinische Bedeutsamkeit des Therapieeffekts (z. B. im Sinne des Konzepts der klinischen Signifikanz) ist feststellbar	(1)	• Normativer Vergleich • RCI + Cutoff-Wert • Soziale Validierung/ Subjektive Evaluation • Diagnosefreiheit **Achtung:** Hier klinische Bedeutsamkeit strenger bewerten: Wenn Cutoff und RCI → 1, wenn nur Cutoff oder RCI → 2
		klinische Bedeutsamkeit des Therapieeffekts (z. B. im Sinne des Konzepts der klinischen Signifikanz) ist nur eingeschränkt feststellbar	(2)	• Normativer Vergleich nur mittels klassischer Signifikanztests • nur RCI oder nur Cutoff-Wert
		klinische Bedeutsamkeit des (z. B. im Sinne des Konzepts der klinischen Signifikanz) Therapieeffekts ist nicht feststellbar	(3)	Die klinische Bedeutsamkeit wurde über keine der Outcomemaße berichtet
A.12.	Vollständige Darstellung der Ergebnisse zu allen Outcomemaßen und zu allen relevanten Messzeitpunkten	für alle Outcome-Variablen berichtet	(1)	Diejenigen Outcomemaße, die laut Publikation zu spezifizierten Messzeitpunkten (i.d.R. Prä-Post-Katamnese) erhoben werden, sollten dementsprechend auch in der Ergebnisdarstellung auftauchen
		ausschließlich für die primären Zielkriterien berichtet	(2)	Es werden nur die Ergebnisse der primären Outcomemaße zu den spezifizierten Messzeitpunkten (i.d.R. Prä-Post-Katamnese) berichtet
		nicht für alle primären Zielkriterien berichtet	(3)	Die Ergebnisdarstellung bezieht sich nicht mal auf alle primären Outcomemaße

[101] Als Erhebung der klinischen Relevanz werden i.d.R. folgende Strategien betrachtet:

1. Normativer Vergleich mittels klassischer und Äquivalenzhypothesentests (vgl. Kendall, Marrs-Garcia, Nath & Sheldrick, 1999)
2. Individuelle Veränderungen (z.B. „Reliable Change Index“ als „statistically reliable improvement“ [Verbesserung]; „Cutoff-Wert“ als „clinically significance change“ [Genesung]) (vgl. Jacobson & Truax, 1991; Stieglitz, 2008)
3. Soziale Validierung/Subjektive Evaluation durch Patienten selbst, Verwandte/Freunde, Fachkräfte/externe Beobachter (vgl. Kazdin, 1994; Lambert & Ogles, 2004)
4. Kein weiteres Erfüllen der Kriterien einer psychiatrischen Diagnose (vgl. Kazdin, 2008).

A.13.	Erfassung unerwünschter Wirkungen	Systematische Erfassung und Bericht von unerwünschten Ereignissen, Nebenwirkungen, Verschlechterungen	(1)	Wird bspw. der %-Anteil der Verschlechterungen, Nebenwirkungen etc. berechnet, dann sollte diese Patientengruppe näher spezifiziert werden
		unsystematische Erfassung und/oder unvollständiger Bericht von unerwünschten Ereignissen, Nebenwirkungen, Verschlechterungen	(2)	Es wird bspw. nur der %-Anteil der Verschlechterungen, Nebenwirkungen etc. angegeben, ohne nähere Spezifizierung der Patientengruppe
		Unerwünschte Ereignisse, Nebenwirkungen, Verschlechterungen wurden nicht erfasst oder nicht berichtet	(3)	-

Drop-out-Analysen:

A.3.	Höhe der Drop-out-Quote zu Behandlungsende (sofern nicht Erfolgskriterium)	i.d.R. Drop-out-Quote kleiner 20 %	(1)	Drop-outs werden hier als Studien- *und/oder* Therapie-Drop-outs definiert: Patienten, die die Studie abbrechen (und ggf. die Therapie weitermachen) sowie Patienten, die die Therapie abbrechen (und ggf. für Studienzwecke weiter zur Verfügung stehen) werden als Drop-outs gezählt **Achtung:** Drop-outs über alle Treatmentarme (EG, KG) berechnen Berechnung nach Flow Diagramm: $1 - \frac{Interventions/Studiencompleter}{Zugewiesen\ zur\ Intervention}$ Wenn keine Drop-outs berichtet, ohne, dass aus Publikation hervorgeht, dass es tatsächlich keine gab → 3
		i.d.R. Drop-out-Quote zwischen 20 % und 40 %	(2)	
		i.d.R. Drop-out-Quote größer 40 %	(3)	
	Vor allem in Bezug auf Studien ohne Randomisierung: Zeitpunkt, ab dem Drop-outs gezählt werden (Erstgespräch, probatorische Sitzungen, Therapiebeginn o.ä.):			-

Wenn bei Prä-Katamnesestudien oder reinen Katamnesestudien Post-Drop-outs und Katamnese-Drop-outs nicht separat zu eruieren sind, dann bei **A.3** alle Drop-outs zwischen Prä und Katamnese kodieren mit folgenden Richtwerten:

Affektiv < 2 Jahre ≥ 2 Jahre	 ≤ 28% ≤ 32 %	**1**
Affektiv < 2 Jahre ≥ 2 Jahre	 29% - 52% 33% - 58%	**2**
Affektiv < 2 Jahre ≥ 2 Jahre	 > 52% > 58%	**3**

Gemischt < 2 Jahre ≥ 2 Jahre	 ≤ 36% ≤ 40%	**1**
Gemischt < 2 Jahre ≥ 2 Jahre	 37% - 64% 41% - 70%	**2**
Gemischt < 2 Jahre ≥ 2 Jahre	 > 64% > 70%	**3**

<table>
<tr>
<td rowspan="10">A.4.</td>
<td rowspan="10">Höhe der Studien-Drop-outs zur Katamnese (falls Katamneseerhebung durchgeführt) (zwischen Post- und Katamnesezeitpunkt)

• Wenn keine Katamnese: „9“
• Prä-Katamnese: „4“
• Nur Katamnese: „5“</td>
<td colspan="2">deutlich besser als in Studien mit vergleichbaren Patientengruppen und vergleichbarem Katamnesezeitraum</td>
<td rowspan="4">(1)</td>
<td rowspan="10">Achtung: Drop-outs über alle Treatmentarme berechnen

Berechnung nach Flow Diagramm:

$1 - \frac{Katamnesedaten\ vorliegend\ von\ Interventions-/Studiencompletern\ zur\ Postmessung}{Interventions-/Studiencompleter\ zur\ Postmessung}$

Wenn keine Drop-outs berichtet, ohne, dass aus Publikation hervorgeht, dass es tatsächlich keine gab → 3</td>
</tr>
<tr><td>Affektive Störungen</td><td>Gemischte Störungsgruppen</td></tr>
<tr><td>weniger als 2 Jahre: ≤ 10%</td><td>weniger als 2 Jahre: ≤ 20%</td></tr>
<tr><td>2 Jahre oder mehr: ≤ 15%</td><td>2 Jahre oder mehr: ≤ 25%</td></tr>
<tr>
<td colspan="2">Drop-out-Quote vergleichbar mit Studien mit entsprechenden Patien-tengruppen und entsprechendem Katamnesezeitraum</td>
<td rowspan="4">(2)</td>
</tr>
<tr><td>Affektive Störungen</td><td>Gemischte Störungsgruppen</td></tr>
<tr><td>weniger als 2 Jahre: 11-20%</td><td>weniger als 2 Jahre: 21-40%</td></tr>
<tr><td>2 Jahre oder mehr: 16-30%</td><td>2 Jahre oder mehr: 26-50%</td></tr>
<tr>
<td colspan="2">deutlich schlechter als in Studien mit vergleichbaren Patienten-gruppen und vergleichbarem Katamnesezeitraum</td>
<td rowspan="4">(3)</td>
</tr>
<tr><td>Affektive Störungen</td><td>Gemischte Störungsgruppen</td></tr>
<tr><td>weniger als 2 Jahre: > 20%</td><td>weniger als 2 Jahre: > 40%</td></tr>
<tr><td>2 Jahre oder mehr: > 30%</td><td>2 Jahre oder mehr: > 50%</td></tr>
</table>

A.16.	Intention to treat–Analysen durchgeführt Falls keine Drop-outs: „9“	ITT-Analysen durchgeführt	(1)	-
		Keine ITT-Analysen bei geringem Risiko für einen attrition bias	(2)	Wenn nahezu ausschließlich Drop-outs vorliegen, von denen mit hoher Wahrscheinlichkeit keine Verzerrungen des Behandlungsergebnisses zu erwarten sind [z.B. Wohnortwechsel oder Versterben des Patienten; Ausfall des behandelnden Therapeuten[102]]
		Keine ITT-Analysen bei deutlichem Risiko für einen attrition bias	(3)	Wenn nahezu ausschließlich Drop-outs vorliegen, von denen mit hoher Wahrscheinlichkeit Verzerrungen des Behandlungsergebnisses zu erwarten sind [z.B. Therapieabbruch seitens Patienten aus therapiebezogenen Gründen; Abbruch seitens Therapeuten wg. patientenseitigen Motivationsmangels; Therapiewechsel; zu geringer Therapieerfolg; ohne Angabe von Gründen] Wenn keine Drop-outs berichtet, ohne, dass aus Publikation hervorgeht, dass es tatsächlich keine gab → 3
A.18.	Vollständige Beschreibung der Drop-Outs Falls keine Drop-outs: „9“	vollständige Beschreibung aller Drop-outs, inkl. der Gründe und des Zeitpunkts des Drop-outs	(1)	Beschreibung von zentralen Patientenmerkmalen und derjenigen Kriterien unter B.9. *und* Gründe *und* Zeitpunkt
		unvollständige Angaben zu Gründen oder Zeitpunkten des Drop-outs	(2)	Unvollständige Beschreibung *und/oder* unvollständige Angabe von Gründen *und/oder* unvollständige Angabe der Zeitpunkte
		keine Beschreibung der Drop-outs oder definitive Angabe der Anzahl der Drop-outs in einer Gruppe	(3)	Keine Beschreibung *und* keine Angabe der Gründe *und* Zeitpunkte *oder* Keine Angabe der Anzahl der Drop-outs pro Gruppe Wenn keine Drop-outs berichtet, ohne, dass aus Publikation hervorgeht, dass es tatsächlich keine gab → 3

[102] Vgl. Hiller, Bleichhardt & Schindler (2009).

A.19.	Drop-out-Analysen Falls keine Drop-outs: „9“	Drop-out-Analysen unter Berücksichtigung der wichtigsten prognostischen Faktoren durchgeführt; keine signifikanten Unterschiede zwischen den Vergleichsgruppen	(1)	**Wenn Drop-out-Quote sehr gering, dann Rating weniger streng durchführen!** Z.B. keine signifikanten Unterschiede zwischen den Drop-outs der Untersuchungsgruppen hinsichtlich prognostisch relevanter Faktoren oder ggf. Korrektur
		Drop-out-Analysen unter Berücksichtigung wichtigster prognostischer Faktoren durchgeführt; trotz sign. Unterschiede ist die Validität der Ergebnisse nicht wesentlich eingeschränkt	(2)	Signifikante Unterschiede zwischen den Drop-outs der Treatmentgruppen hinsichtlich prognostisch relevanter Faktoren und Einschätzung in Studie, dass nicht validitätsgefährdend bzgl. Ergebnisse
		Drop-out-Analysen nicht oder unter Vernachlässigung relevanter prognostischer Merkmale durchgeführt; Drop-out-Analysen stellen die Validität der Ergebnisse deutlich in Frage	(3)	Drop-out-Analysen hinsichtlich prognostisch irrelevanter Faktoren durchgeführt *oder* Keine Drop-out-Analysen durchgeführt Wenn keine Drop-outs berichtet, ohne, dass aus Publikation hervorgeht, dass es tatsächlich keine gab → 3

Statistische Methodik:

A.14.	Anwendungsvoraussetzungen für statistische Modelle geprüft und erfüllt	Anwendungsvoraussetzungen geprüft und erfüllt	(1)	• Wenn z.B. bei Anova mit Messwiederholung: Sphärizität mittels Mauchly-Test und ggf. Korrektur (Greenhouse-Geisser-Korrektur) • Keine Überprüfung von z.B. Normalverteilungsvoraussetzung bei hinreichend großem N pro Bedingung (**N>30 bei t-Tests, N>25 bei ANOVA**) oder keine Überprüfung der Homoskedastizität bei gleichgroßen Stichproben • Voraussetzungen (v.a. bei kleinen Stichproben) überprüft und keine Verletzungen
		Anwendungsvoraussetzungen geprüft und lediglich leichte Verletzungen der Voraussetzungen oder keine Prüfung der Anwendungsvoraussetzungen bei gleichzeitiger Robustheit der angewendeten statistischen Verfahren	(2)	Voraussetzungen (v.a. bei kleinen Stichproben) überprüft und leichte Verletzungen
		deutliche Verletzungen der Anwendungsvoraussetzungen oder keine Prüfung der Anwendungsvoraussetzungen bei substanziellem Risiko für deren Verletzung	(3)	• Eindeutige Verletzung der Sphärizität ohne Korrektur oder Verletzung sonstiger Voraussetzungen (NV, Homoskedastizität) • Keine Überprüfung der Sphärizitätsannahme • Keine Überprüfung der Voraussetzungen bei „problematischen“ (kleinen, ungleich großen) Stichproben
A.15.	Angemessenheit der statistischen Analysen (inklusive der Korrektur für multiple Tests)	adäquate und umfassende statistische Analysen	(1)	z.B. α-Fehlerkorrektur
		weitgehend adäquate statistische Analysen	(2)	-
		unangemessene statistische Analysen (fehlende Korrektur für multiple Tests, inadäquate statistische Methoden)	(3)	• keine α-Fehlerkorrektur, obwohl notwendig • Anwendung von parametrischen Verfahren, wenn nonparametrische indiziert

A.17.	Statistische Power der Vergleiche bei Vergleich mit bewährter Treatment-Gruppe Bei Eingruppendesigns (inkl. Designs mit verfahrensinternen Vergleichsgruppen), Kontrollgruppendesigns (Warteliste, Placebo, TAU, aktive Kontrolle): „9“	adäquate statistische Power der Vergleiche	(1)	**Achtung: Dieses Kriterium bezieht sich allein auf verfahrensexterne Vergleiche mit bereits bewährten Behandlungen** Als Richtwert für einen t-Test für unabhängige Stichproben: N (pro Gruppe) >310 bzw. N (gesamt) >620 [103]
		eingeschränkte Power der statistischen Vergleiche (.50-.80)	(2)	Als Richtwert für einen t-Test für unabhängige Stichproben: N (pro Gruppe): 136 bis 310 bzw. N (gesamt) 272 bis 620
		unzureichende statistische Power der Vergleiche (<.50)	(3)	Als Richtwert für einen t-Test für unabhängige Stichproben: N (pro Gruppe) <136 bzw. N (gesamt) <272
	Wurde die Signifikanzüberprüfung mittels eines Intervall-Nullhypothesentests (einseitiger Äquivalenztest/Non-Inferiority) oder mittels eines nil-Nullhypothesentests durchgeführt?			nil-Nullhypothesentest: H_0: $\mu_A=\mu_B$ vs. H_1: $\mu_A\neq\mu_B$ Non-Inferiority: H_0: $\mu_A-\mu_B \geq\delta$ vs. H_1: $\mu_A-\mu_B <\delta$

Manipulation der Daten:

A.1.	Manipulation der Daten	keine Hinweise auf Ergebnismanipulation	(1)	-
		-	(2)	-
		Hinweise auf Ergebnismanipulation **(Stufe 3 = Ausschlusskriterium)**	(3)	z.B. wenn in unterschiedlichen Publikationen zur selben Studie (selber Datensatz!) voneinander abweichende Ergebnisse berichtet werden

[103] Die kalkulierten Stichprobengrößen gelten für gleichgroße Stichproben. Für ungleichgroße Stichproben werden gesonderte Kalkulationen angestellt.

Literatur:

Beutel, M. E., Doering, S., Leichsenring, F. & Reich, G. (2010). *Psychodynamische Psychotherapie. Störungsorientierung und Manualisierung in der therapeutischen Praxis*. Göttingen: Hogrefe.

Deutsche Gesellschaft für Psychoanalyse, Psychotherapie, Psychosomatik und Tiefenpsychologie (DGPT) e.V. (2009). *Stellungnahme zur Prüfung der Richtlinienverfahren gemäß §§ 13 – 15 der Psychotherapie-Richtlinie für die psychoanalytisch begründeten Verfahren.*
Verfügbar unter:
http://www.dgpt.de/dokumente/DGPT%20Stllungnahme%20zur%20Pruefung%20der%20Richtlinienverfahren%202009.pdf. [30.08.2010].

Lambert, M. J. & Ogles, B. M. (2004). The Efficacy and Effectiveness of Psychotherapy. In A. E. Bergin & S. L. Garfield (Hrsg.), *Handbook of Psychotherapy and Behavior Change* (5. Aufl., S. 139–193). Oxford England: John Wiley & Sons.

Hiller, W., Bleichhardt, G. & Schindler, A. (2009). Evaluation von Psychotherapien aus der Perspektive von Qualitätssicherung und Qualitätsmanagement. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie, 57* (1), 7-22.

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* (1), 12-19.

Kazdin, A. E. (1994). Methodology, design, and evaluation in psychotherapy research. In: A. E. Bergin, S. L. Garfield (Hrsg.), *Handbook of psychotherapy and behavior change* (4. Aufl., S. 19–71). Oxford England: John Wiley & Sons.

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63* (3), 146-159.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R. & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, *67* (3), 285-299.

Möller, H. J., Laux, G. & Kapfhammer, H. P. (Hrsg.). (2005). *Psychiatrie und Psychotherapie* (2. Aufl.). Heidelberg: Springer Medizin Verlag.

Reimer, C. & Rüger, U. (Hrsg.). (2006). *Psychodynamische Psychotherapien: Lehrbuch der tiefenpsychologisch fundierten Psychotherapieverfahren* (3. Aufl.). Heidelberg: Springer.

Schulte, D. (1993). Wie soll Therapieerfolg gemessen werden? *Zeitschrift für Klinische Psychologie, 22* (4), 374-393.

Stieglitz, R.-D. (2008). *Diagnostik und Klassifikation in der Psychiatrie* (1. Aufl.). Stuttgart: Kohlhammer.

Wissenschaftlicher Beirat Psychotherapie nach § 11 PsychThG. (2009). *Methodenpapier des Wissenschaftlichen Beirats Psychotherapie: Verfahrensregeln zur Beurteilung der wissenschaftlichen Anerkennung von Methoden und Verfahren der Psychotherapie - Version 2.7*.
Verfügbar unter: http://www.wbpsychotherapie.de/downloads/Methodenpapier2720090709.pdf [15.5.2010].

Wissenschaftlicher Beirat Psychotherapie nach § 11 PsychThG. (2010). *Methodenpapier des Wissenschaftlichen Beirats Psychotherapie: Verfahrensregeln zur Beurteilung der wissenschaftlichen Anerkennung von Methoden und Verfahren der Psychotherapie - Version 2.8*.
Verfügbar unter: http://www.wbpsychotherapie.de/downloads/Methodenpapier28.pdf [26.11.2010].

Anhang D: Kodierregeln zum Kurzkodierbogen zur Erhebung allgemeiner Studiencharakteristika und zur methodologischen Spezifizierung von Wirksamkeitsstudien hinsichtlich naturalistischer und experimenteller Studiendesigneigenschaften

(aus: Ratzek & von Hauenschild, 2011a)

Beurteiler/in: **Datum der Beurteilung:**

Autoren:

Titel:

Publikationsjahr:

Datenzugriff:

Originalstudie	
Reanalyse	
Replikation	

Weitere Charakteristiken und methodologische Eigenschaften der Studie:

Anwendungsbereich	
Affektive Störungen:	Benennung der Diagnosen (Major Depression, Bipolare affektive Störung, Dysthymia, postnatale Depression etc. inkl. %-Angaben)
Gemischte Störungsgruppen:	Auflistung aller Störungen, die in die gemischte Störungsgruppe(n) eingehen (Hauptdiagnosen) inkl. %-Angaben (falls vorhanden)
Zentrale Patientenmerkmale:	
	z.B. - Geschlechterverteilung - Altersverteilung - Herkunft - Bildungsgrad etc.

Anwendungsform der psychoanalytisch begründeten Verfahren			
Benennung in Studie:			Benennung der Behandlung entsprechend der Benennung in der Studie (Zitat)
Einordnung nach Psychotherapierichtlinien (plus Psychoanalyse)		**Sitzungsanzahl/Dauer und Frequenz**	
tiefenpsychologisch fundiert			**Mehrfachrating möglich!** Rating in Anlehnung an einschlägige Literatur. **Orientierungswerte:** • bis zu 100 Sitzungen, 1-2/Woche, sitzend → TP • bis zu 300 Sitzungen, 2-3/Woche, liegend → AP In Abhängigkeit von den Treatmentarmen kann eine Studie durchaus bspw. tiefenpsychologisch fundierte und analytische Therapien untersuchen.
analytische Psychotherapie			
Psychoanalyse			
Gehen unterschiedliche Anwendungsformen der psychoanalytisch begründeten Verfahren in einen Treatmentarm ein oder werden die Anwendungsformen separat voneinander untersucht?			Wird bspw. innerhalb eines Treatmentarms sowohl tiefenpsychologisch fundierte als auch analytische Psychotherapie durchgeführt?
Setting:			
Einzeltherapie			-
Gruppentherapie			
a priori festgelegte Sitzungsanzahl			
ja			Ist die Sitzungsanzahl bzw. Dauer der Behandlung im Rahmen der Studie limitiert?
nein			
störungsunspezifisches vs. störungsspezifisches Vorgehen			
störungsunspezifische Anwendung der analytisch begründeten Therapie			Anwendung einer störungsspezifischen Methode des psychoanalytisch begründeten Verfahrens [104]?
störungsspezifische Anwendung der analytisch begründeten Therapie			

[104] Vgl. Beutel, Doering, Leichsenring & Reich (2010); DGPT (2009); Reimer & Rüger (2006).

Kontroll-/Vergleichsgruppendesign		**Sitzungsanzahl/Dauer und Frequenz**	
unbehandelte Kontrollgruppe (z.B. Warteliste)			**Mehrfachrating möglich!**
Placebo-Kontrollgruppe (stützende Gespräche, nonspecific treatment)			
TAU-Kontrollgruppe			
Aktive Kontrollgruppe (z.B. supportive Interventionen)			
verfahrensexterne und etablierte Vergleichsbehandlung(en)[105]			„Verfahrensextern“ bedeutet ein/e Verfahren/Methode, das nicht zu den psychoanalytisch begründeten Verfahren gehört
verfahrensexterne Vergleichsbehandlungen(en)[2] - kein bereits etabliertes Treatment			
verfahrensinterne Vergleichsbehandlungen(n) (z.B. Vergleich zweier Methoden der analytisch begründeten Verfahren)			„Verfahrensintern“ bedeutet ein/e Verfahren/Methode, das zu den psychoanalytisch begründeten Verfahren gehört
Vergleichsgruppe mit ausschließlich psychopharmakologischer Behandlung			
keine Vergleichsgruppe			
Wie viele Behandlungsarme insgesamt?- Welche?			Aufzählung aller Treatmentarme (d.h. der Intervention und der Kontroll- und/oder Vergleichsgruppen). Hier werden auch die Behandlungsarme innerhalb eines verfahrensinternen Vergleichs mitberücksichtigt und aufgezählt (so sie separat untersucht wurden).

[105] Als „etabliertes Treatment“ werden Verfahren betrachtet, die zum aktuellen Zeitpunkt (01/2011) als wissenschaftlich anerkannte Verfahren gelten (Verhaltenstherapie, Gesprächspsychotherapie, Systemische Therapie).

<table>
<tr><td colspan="3">Messzeitpunkte</td><td colspan="2">prospektives vs. retrospektives Design</td><td></td></tr>
<tr><td>Prä-Post</td><td></td><td rowspan="6">Wenn kein expliziter Post-Zeitpunkt → Prä-Katamnese</td><td rowspan="3">prospektiv
(vor Beginn des zu evaluierenden Treatments wird das Design festgelegt: Treatmentarme bzw. zu vergleichende Gruppen; Messzeitpunkte; Outcomekriterien)</td><td></td><td rowspan="3">-</td></tr>
<tr><td>Prä-Katamnese</td><td></td></tr>
<tr><td>Prä-Post-Katamnese</td><td></td></tr>
<tr><td>Post-Katamnese</td><td></td><td rowspan="3">retrospektiv
(das/die Treatment/s sind bereits erfolgt und die Auswahl der Treatmentarme, die Bestimmung der Messzeitpunkte und die Festlegung der Outcomemaße erfolgt erst nach Beendigung [ggf. während] des/r Treatments)</td><td></td><td rowspan="3">-</td></tr>
<tr><td>nur Post</td><td></td></tr>
<tr><td>nur Katamnese(n)</td><td></td></tr>
<tr><td colspan="5">Katamnesezeitraum (Post-Katamnese):</td><td>Falls mehr als eine Katamnesemessung, <u>alle</u> aufzählen</td></tr>
<tr><td colspan="6">Gruppenzuweisung</td></tr>
<tr><td rowspan="2">randomisierte Zuteilung</td><td colspan="3">verfahrensinterne Vergleichsgruppe/n</td><td></td><td rowspan="4">Hier werden auch die Behandlungsarme <u>innerhalb eines verfahrensinternen Vergleichs</u> mitberücksichtigt, d.h. die Zuweisung zu verfahrensinternen Vergleichsgruppen geratet (so sie separat untersucht wurden).

„Verfahrensinterne Vergleichsgruppen“ sind Gruppen, die mit einem Verfahren / einer Methode, das/die zu den psychoanalytisch begründeten Verfahren gehört, behandelt wurden.

„Verfahrensexterne Vergleichs-/Kontrollgruppen“ sind Gruppen, die mit <u>keinem</u> Verfahren/ <u>keiner</u> Methode, das/die zu den psychoanalytisch begründeten Verfahren gehört, behandelt wurden.</td></tr>
<tr><td colspan="3">verfahrensexterne Vergleichsgruppe/n
oder
Kontrollgruppe/n (Warteliste, Placebo, TAU, aktive Kontrolle)</td><td></td></tr>
<tr><td rowspan="2">teilweise randomisiert und teilweise Selbstzuteilung</td><td colspan="3">verfahrensinterne Vergleichsgruppe/n</td><td></td></tr>
<tr><td colspan="3">verfahrensexterne Vergleichsgruppe/n
oder
Kontrollgruppe/n (Warteliste, Placebo, TAU, aktive Kontrolle)</td><td></td></tr>
</table>

patientenseitige Selbstzuteilung unter Anwendung von Strategien wie z.B. Parallelisierung, Stratifizierung, Matching	verfahrensinterne Vergleichsgruppe/n		Parallelisierung: Balancierung der Gruppen hinsichtlich bekannter Störvariable, so dass Mittelwert und Streuung dieser Variable annähernd gleich in den Gruppen Stratifizierung: Anzahlmäßige Gleichverteilung von Störvariablen (z.B. Geschlechterverteilung in allen Gruppen gleich) Matching: Paarbildung hinsichtlich Matchingvariable, so dass in Gruppen gleichverteilt
	verfahrensexterne Vergleichsgruppe/n *oder* Kontrollgruppe/n (Warteliste, Placebo, TAU, aktive Kontrolle)		
patientenseitige Selbstzuteilung ohne Anwendung von Strategien wie z.B. Parallelisierung, Stratifizierung, Matching	verfahrensinterne Vergleichsgruppe/n		-
	verfahrensexterne Vergleichsgruppe/n *oder* Kontrollgruppe/n (Warteliste, Placebo, TAU, aktive Kontrolle)		
Standardisierungsgrad des Treatments			
Verwendung von Manual bzw. manualähnlichen Behandlungsrichtlinien (behandlungsprinzipienbasiert)			Primär bezogen auf Gruppe mit psychoanalytisch begründetem Verfahren: „Manualähnliche Behandlungsrichtlinien“ [106]
nicht manualisiert bzw. keine Verwendung manualähnlicher Behandlungsrichtlinien (behandlungsprinzipienbasiert)			
explizites Therapeutentraining zwecks Studiendurchführung			-
kein explizites Therapeutentraining			
Implementationskontrolle			Hiermit ist die Kontrolle der Behandlungsimplementierung gemeint (treatment integrity)
keine Implementationskontrolle			
Selektivität der Stichprobe			
Ausschluss subklinischer Symptomausprägungen	ja		-
	nein		
Ausschluss komorbider Störungen	ja		Ausschluss epidemiologisch relevanter Störungen
	nein		

[106] Vgl. Beutel, Doering, Leichsenring & Reich (2010); DGPT (2009); Reimer & Rüger (2006).

Von den Autor/innen formulierte Fragestellung (Untersuchungsziel):

	v.a. im Abstract und unter dem Abschnitt „aim of the study…“ zu finden

Subgruppenanalysen:

Wurden Subgruppenanalysen durchgeführt?

ja	
nein	

Welche?

Literatur:

Beutel, M. E., Doering, S., Leichsenring, F. & Reich, G. (2010). *Psychodynamische Psychotherapie. Störungsorientierung und Manualisierung in der therapeutischen Praxis*. Göttingen: Hogrefe.

Deutsche Gesellschaft für Psychoanalyse, Psychotherapie, Psychosomatik und Tiefenpsychologie (DGPT) e.V. (2009). *Stellungnahme zur Prüfung der Richtlinienverfahren gemäß §§ 13 – 15 der Psychotherapie-Richtlinie für die psychoanalytisch begründeten Verfahren.*
Verfügbar unter:
http://www.dgpt.de/dokumente/DGPT%20Stllungnahme%20zur%20Pruefung%20der%20Richtlinienverfahren%202009.pdf. [30.08.2010].

Reimer, C. & Rüger, U. (Hrsg.). (2006). *Psychodynamische Psychotherapien: Lehrbuch der tiefenpsychologisch fundierten Psychotherapieverfahren* (3. Aufl.). Heidelberg: Springer.

Anhang E: Kodierbogen zur Beurteilung von psychometrischen Eigenschaften (Reliabilität und Validität) diagnostischer Selbst- und Fremdbeurteilungsverfahren

(Ratzek & von Hauenschild, 2011c)

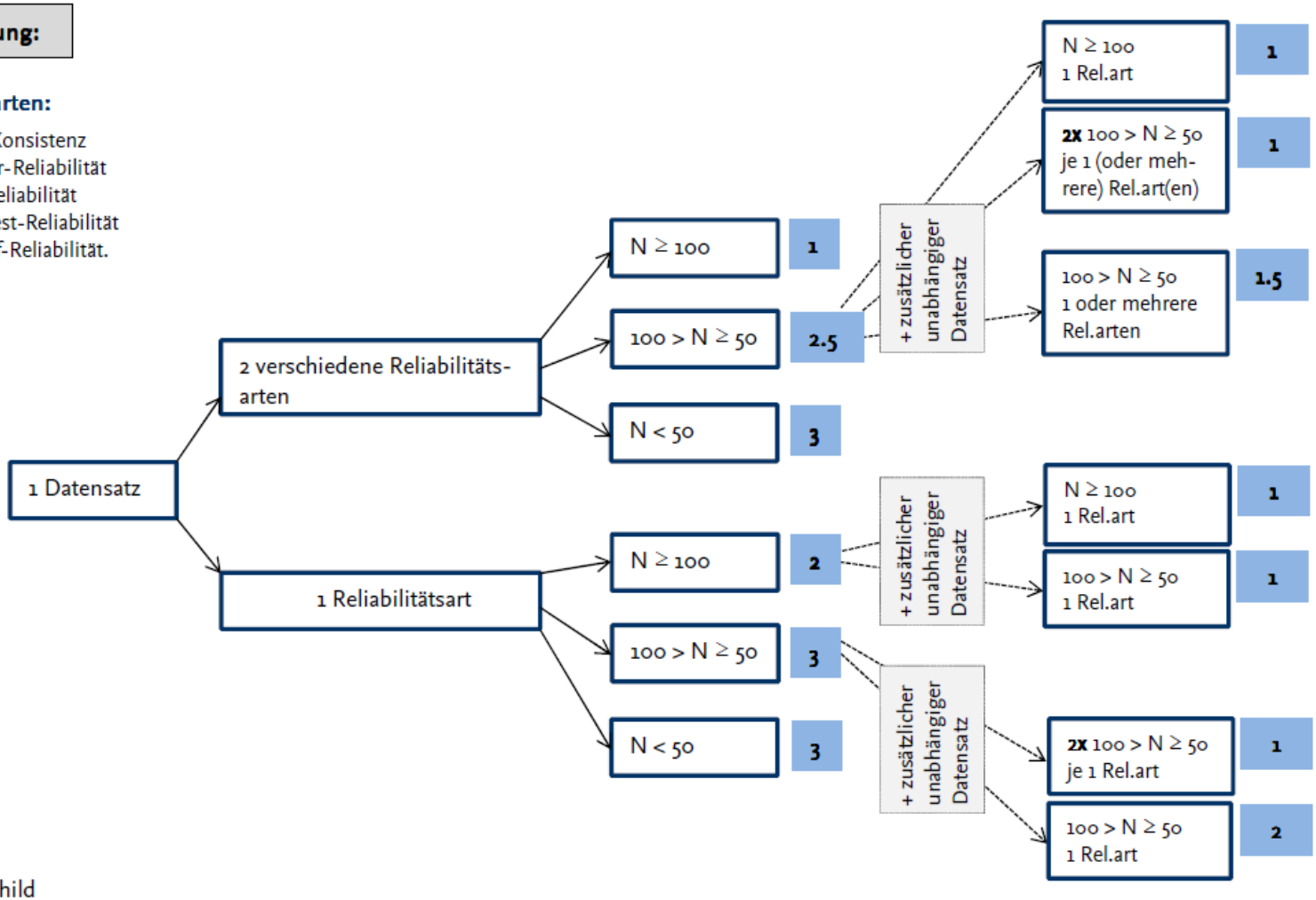
Inventar:

Entscheidungsregeln für die Reliabilitäts- und Validitätsbewertung der Outcomemaße (Kriterium A 8 des Methodenpapiers 2.7 / 2.8, WBP):

1. Basisbewertung:

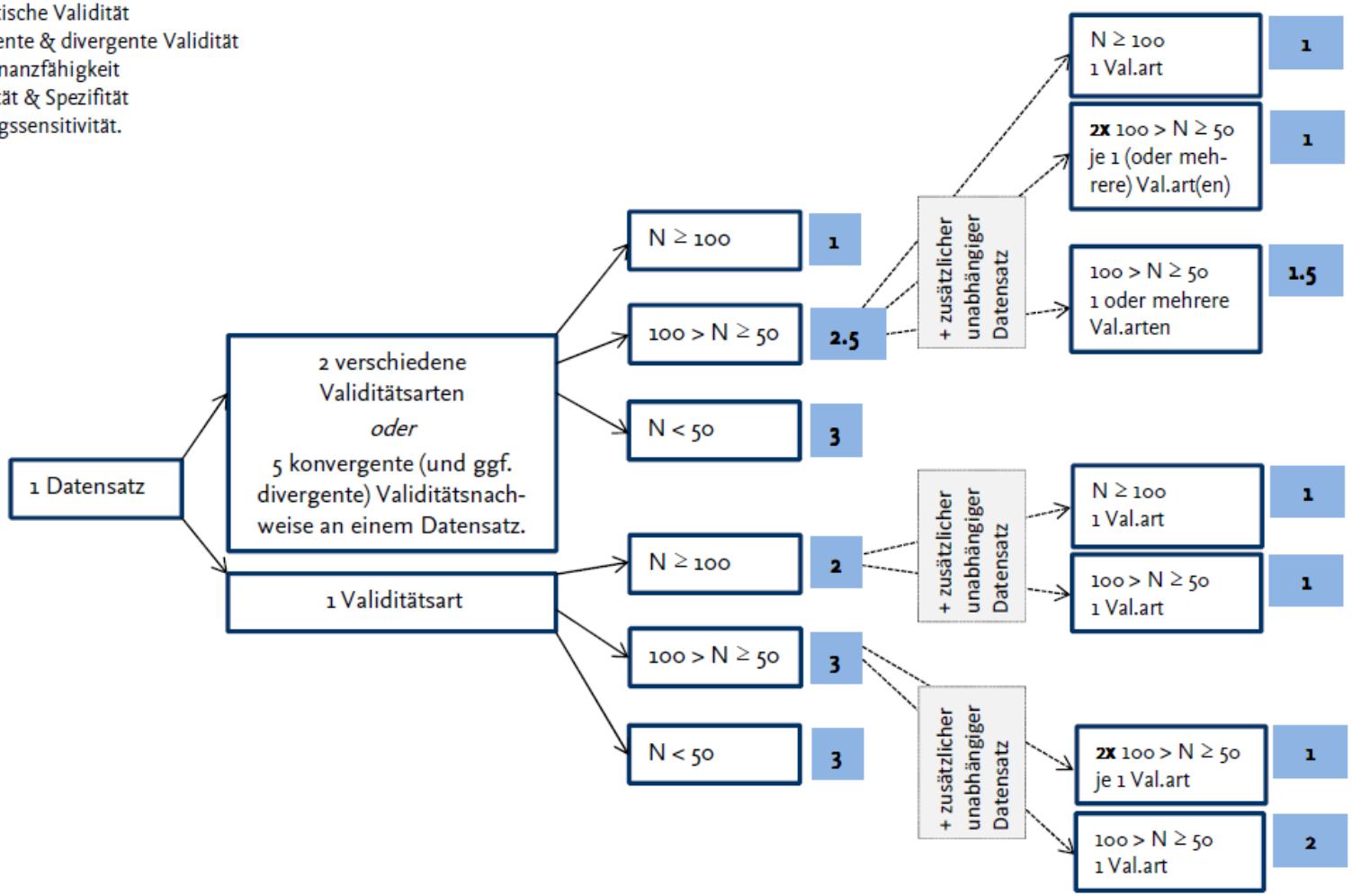
Reliabilitätsarten:

- Interne Konsistenz
- Interrater-Reliabilität
- Retest-Reliabilität
- Paralleltest-Reliabilität
- Split-half-Reliabilität.



Validitätsarten:

- Übereinstimmungsvalidität
- Prognostische Validität
- Konvergente & divergente Validität
- Diskriminanzfähigkeit
- Sensitivität & Spezifität
- Änderungssensitivität.



2. Koeffizientenbewertung:

Reliabilität		Bewertung
- Retest-Reliabilität - Interne Konsistenz (Cronbach's Alpha) - Split-half-Reliabilität - Paralleltest-Reliabilität - Interrater-Reliabilität (Konsistenz) [1]	> .70	1
	.60 - .70	2
	< .60	3
Interrater-Reliabilität (Agreement) [2] - Cohen's Kappa - ICC	> .59	1
	.40 - .59	2
	< .40	3
Interrater-Reliabilität [3] - prozentuale Übereinstimmung	> 79%	1
	70 – 79%	2
	< 70%	3

Validität		Bewertung
- Übereinstimmungsvalidität - Prognostische Validität - Konvergente & divergente Validität [4]	> .60	1
	.40 - .60	2
	< .40	3
- Diskriminanzfähigkeit - Änderungssensitivität	p≤ .01	1
	P≤ .05	2
	p> .05	3
- Sensitivität & Spezifität [5] (Diagnostische Genauigkeit)	> .79	1
	.70 - .79	2
	< .70	3

3. Zusammenfassende Koeffizientenbewertung:

Reliabilität	**Mittelwert:**
Validität	**Mittelwert:**

[1] Bewertungsmaßstab entnommen aus dem COTAN System (vgl. Evers, 2001a/b)
[2] Bewertungsmaßstab angelehnt an Cicchetti (1994, 2001) und Cicchetti u. Rourke (2004)
[3] Bewertungsmaßstab angelehnt an Cicchetti (2001) und Cicchetti u. Rourke (2004)
[4] Bewertungsmaßstab nach Fisseni (1997)
[5] Bewertungsmaßstab angelehnt an Cicchetti (2001)

4. Zusammenfassende Basis- und Koeffizientenbewertung:

Reliabilität	Koeffizientenbewertung				
Basisbewertung	1	1.1 – 1.5	1.6 - 2	2.1 – 2.5	2.6 - 3
1	1	Koeff.-bewertung	Koeff.-bewertung	Koeff.-bewertung	Koeff.-bewertung
1.5	1.5	1.5	Koeff.-bewertung	Koeff.-bewertung	Koeff.-bewertung
2	2	2	2	Koeff.-bewertung	Koeff.-bewertung
2.5	2.5	2.5	2.5	2.5	Koeff.-bewertung
3	3	3	3	3	3

Reliabilität	**Wert:**
Validität	**Wert:**

Validität	Koeffizientenbewertung				
Basisbewertung	1	1.1 – 1.5	1.6 - 2	2.1 – 2.5	2.6 - 3
1	1	Koeff.-bewertung	Koeff.-bewertung	Koeff.-bewertung	Koeff.-bewertung
1.5	1.5	1.5	Koeff.-bewertung	Koeff.-bewertung	Koeff.-bewertung
2	2	2	2	Koeff.-bewertung	Koeff.-bewertung
2.5	2.5	2.5	2.5	2.5	Koeff.-bewertung
3	3	3	3	3	3

5. Zusammenfassende Reliabilitäts- und Validitätsbewertung:

exzellent	gut	zufrieden-stellend	ausreichend	unzu-reichend
1 – 1.3	1.4 – 1.7	1.8 – 2.1	2.2 – 2.5	2.6 - 3

Literatur:

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6* (4), 284-290.

Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology, 23* (5), 695-700.

Cicchetti, D. V. & Rourke, B. P. (Hrsg.). (2004). *Methodological and biostatistical foundations of clinical neuropsychology and medical and health disciplines* (2. Aufl.). Hove, England: Psychology Press/Taylor & Francis.

Evers, A. (2001a). Improving Test Quality in the Netherlands: Results of 18 Years of Test Ratings. *International Journal of Testing, 1* (2), 137-153.

Evers, A. (2001b). The Revised Dutch Rating System for Test Quality. *International Journal of Testing, 1* (2), 155-182.

Fisseni, H.-J. (1997). Lehrbuch der psychologischen Diagnostik. Göttingen: Hogrefe.

Wissenschaftlicher Beirat Psychotherapie nach § 11 PsychThG. (2009). *Methodenpapier des Wissenschaftlichen Beirats Psychotherapie: Verfahrensregeln zur Beurteilung der wissenschaftlichen Anerkennung von Methoden und Verfahren der Psychotherapie - Version 2.7*.
Verfügbar unter: http://www.wbpsychotherapie.de/downloads/Methodenpapier2720090709.pdf [15.5.2010].

Wissenschaftlicher Beirat Psychotherapie nach § 11 PsychThG. (2010). *Methodenpapier des Wissenschaftlichen Beirats Psychotherapie: Verfahrensregeln zur Beurteilung der wissenschaftlichen Anerkennung von Methoden und Verfahren der Psychotherapie - Version 2.8*.
Verfügbar unter: http://www.wbpsychotherapie.de/downloads/Methodenpapier28.pdf [26.11.2010].

Anhang F: Allgemeine Übersicht I über alle Primärstudien (*N*=41)

Autoren & Kurztitel der Studien	Datenzugriff	Anwendungsbereich & Diagnosen	Verfahrenstitel im Original	Einordnung des psychodynamischen Verfahrens nach Psychotherapie-richtlinie	Sitzungsanzahl, Dauer & Frequenz der psychodynamischen Behandlung	Setting
Abbass (2006) **Abbass-A Studie** [1]	Originalstudie	Affektive Störungen Behandlungsresistente Depression	Intensive short-term dynamic psychotherapy	TP	*M* 13.6 Sitzungen ca. 16 Wochen	Individualtherapie
Abbass (2002) **Abbass-B Studie** [2]	Originalstudie	Gemischte Störungen Major Depression, GAD, Dysthymia, Panikstörung, posttraumatische Belastungsstörung, Agoraphobie, Essstörungen, Substanzmissbrauch, soziale Phobien	Intensive short-term dynamic psychotherapy	TP	*M* 16.9 Sitzungen	Individualtherapie
Barber et al. (2004) Wilczek et al. (2004) [Wilczek et al. (1998)] **Wilczek Studie** [3]	Originalstudie	Gemischte Störungen Mood disorders (DSM-III), anxiety disorders (DSM-III), somatoforme Störungen, Schlafstörungen, Essstörungen, sexuelle Funktionsstörungen, Persönlichkeitsstörungen	long-term psychoanalytic psychotherapy, long-term dynamic psychotherapy	AP	*M* 159 Sitzungen *M* 3 Jahre 1-2/Woche	Individualtherapie
Blatt et al. (2004) Wallerstein (1986/1989) **Reanalyse Menninger Studie** [4]	Reanalyse	Gemischte Störungen neurotic, personality disorder, latent psychotic (Blatt et al., S. 400)	Psychoanalysis und supportive-expressive psychotherapy	AP und Psa	AP: *M* 316 Sitzungen 5/Woche Psa: *M* 1017 Sitzungen 2-3/Woche	Individualtherapie

Autoren & Kurztitel der Studien	Datenzugriff	Anwendungsbereich & Diagnosen	Verfahrenstitel im Original	Einordnung des psychodynamischen Verfahrens nach Psychotherapie-richtlinie	Sitzungsanzahl, Dauer & Frequenz der psychodynamischen Behandlung	Setting
Blomberg et al. (2001) [Grant et al. (2004) Lazar et al. (2006) Sandell et al. (1999)] **Stockholm Outcome of Psychotherapy and Psychoanalysis (STOPP)** [5]	Originalstudie	Gemischte Störungen Achse-I-Diagnosen, Achse-II-Diagnosen	Psychodynamic long-term psychotherapy und psychoanalysis	AP und Psa	AP: *M* 252 Sitzungen 1-2/Woche Psa: *M* > 600 Sitzungen 3-5/Woche	Individualtherapie
Bond et al. (2004/2006) Perry et al. (2009) **Bond Studie** [6]	Originalstudie	Gemischte Störungen Angststörungen, affektive Störungen, Essstörungen, Substanzmissbrauchsstörung, Persönlichkeitsstörungen	Long-term dynamic psychotherapy	AP	*Me* 110 Sitzungen *Me* 3 Jahre 1-2/Woche	Individualtherapie
Bradshaw et al. (2009) **Bradshaw Studie** [7]	Originalstudie	Gemischte Störungen Affektive Störungen (Major Depression, nicht näher bezeichnete affektive Störung, Dysthymia), Angststörungen (GAD), posttraumatische Belastungsstörung, Anpassungsstörungen, Persönlichkeitsstörungen	Psychodynamic psychotherapy	TP	*M* 25 Sitzungen	Individualtherapie

Autoren & Kurz-titel der Studien	**Datenzugriff**	**Anwendungsbereich & Diagnosen**	**Verfahrenstitel im Origi-nal**	**Einordnung des psychodynamischen Verfahrens nach Psychotherapie-richtlinie**	**Sitzungsanzahl, Dauer & Frequenz der psychodynami-schen Behandlung**	**Setting**
Brockmann et al. (2002) Brockmann et al. (2006) **Frankfurt-Hamburg-Studie** [8]	Originalstudie	Gemischte Störungen Depressive und Angststörungen	Psychoanalytische Behand-lung, psychoanalytische Langzeittherapie	AP	*M* 209 Sitzungen 1-2/Woche	Individualthera-pie
Cooper et al. (2003) **Cooper Studie** [9]	Originalstudie	Affektive Störungen Postnatale Depression	Psychodynamic therapy	TP	10 Sitzungen 1/Woche	Individualthera-pie
Gordon (2001) **Gordon Studie** [10]	Originalstudie	Gemischte Störungen Dysthymia, anxiety disorders (DSM), Major Depression, somato-forme Störungen, Persönlichkeits-störungen	Long-term psychoanalytic psychotherapy	AP	ca. 120-240 Sitzun-gen 1-2/Woche	Individualthera-pie

Autoren & Kurztitel der Studien	Datenzugriff	Anwendungsbereich & Diagnosen	Verfahrenstitel im Original	Einordnung des psychodynamischen Verfahrens nach Psychotherapie-richtlinie	Sitzungsanzahl, Dauer & Frequenz der psychodynamischen Behandlung	Setting
Grande et al. (2006/2009) Rudolf et al. (2004) **Praxisstudie psychoanalytischer Langzeittherapie (PAL)** [11]	Originalstudie	Gemischte Störungen Depressive Episode, rezidivierende depressive Störungen, anhaltende affektive Störungen, Angststörungen, phobische Störungen, somatoforme Störungen, Zwangsstörung, sexuelle Funktionsstörungen, Reaktionen auf schwere Belastung und Anpassungsstörungen, Essstörungen, psychische und Verhaltensstörungen durch psychotrope Substanzen, Schlafstörungen, Persönlichkeitsstörungen	Psychoanalytic therapy und psychodynamic therapy	TP und AP	TP: *M* 71 Sitzungen 1/Woche AP: *M* 310 Sitzungen 3/Woche	Individualtherapie
[Guthrie et al. (1998)] Guthrie et al. (1999) **Guthrie Studie** [12]	Originalstudie	Gemischte Störungen Depressive Episode, rezidivierende depressive Störungen, Dysthmia, bipolare affektive Störung, GAD, phobische Störungen, Panikstörung, Zwangsstörung, somatoforme Störungen	Brief psychodynamic-interpersonal psychotherapy	TP	8 Sitzungen 1/Woche	Individualtherapie

Autoren & Kurztitel der Studien	Datenzugriff	Anwendungsbereich & Diagnosen	Verfahrenstitel im Original	Einordnung des psychodynamischen Verfahrens nach Psychotherapie-richtlinie	Sitzungsanzahl, Dauer & Frequenz der psychodynamischen Behandlung	Setting
Hecke et al. (2008) Junkert-Tress et al. (2001) Junkert-Tress et al. (1999) **Düsseldorfer-Kurzzeitpsychotherapie-Projekt** [13]	Originalstudie	Gemischte Störungen Somatoforme Störungen, psychische Faktoren und Verhaltenseinflüsse bei andernorts klassifizierten Krankheiten, Persönlichkeitsstörungen, depressive Episode, anhaltende affektive Störungen, phobische Störungen, Angststörungen	Psychodynamisch-interpersonelle Fokaltherapie / brief dynamic psychotherapy	TP	25 Sitzungen	Individualtherapie
Hilsenroth et al. (2003) **Hilsenroth Studie** [14]	Originalstudie	Affektive Störungen „depressive spectrum disorder“: Major Depression, nicht näher bezeichnete affektive Störungen, Dysthymia, Anpassungsstörungen mit depressiver Verstimmung	Short-term psychodynamic psychotherapy for depression	TP	*M* 30 Sitzungen *M* 7 Monate 1-2/Woche	Individualtherapie
Høglend et al. (2006) Høglend et al. (2008) **Høglend Studie** [15]	Originalstudie	Gemischte Störungen Major Depression, Dysthymia, Panikstörung, Agoraphobie, soziale Phobien, GAD, somatoforme Störungen, Anpassungsstörungen, Persönlichkeitsstörungen	Brief dynamic Psychotherapy (with or without transference interpretation)	TP	*M* 33-34 Sitzungen 1/Woche	Individualtherapie

Autoren & Kurztitel der Studien	Datenzugriff	Anwendungsbereich & Diagnosen	Verfahrenstitel im Original	Einordnung des psychodynamischen Verfahrens nach Psychotherapie-richtlinie	Sitzungsanzahl, Dauer & Frequenz der psychodynamischen Behandlung	Setting
Jakobsen et al. (2005a/b) Mattanza et al. (2005) **Schweizer Praxisstudie psychoanalytischer Langzeittherapie (Schweizer PAL)** [16]	Replikation	Gemischte Störungen Störungen durch psychotrope Substanzen, affektive Störungen, phobische Störungen, Angststörungen, Essstörungen, Persönlichkeitsstörungen	Analytische oder tiefenpsychologische Psychotherapie mit einem für die Jung'sche Psychotherapie geeigneten Ansatz	TP und AP	*M* 90 Sitzungen *M* 35 Monate 1-2/Woche	Individualtherapie
Keller et al. (2001) **Jungianische Psychoanalyse und Psychotherapie Studie** [17]	Originalstudie	Gemischte Störungen Affektive Störungen (bipolare affektive Störung, depressive Episode, rezidivierende depressive Störungen, Zyklothymia), phobische Störungen, Angststörungen, Zwangsstörung, akute Belastungsreaktion, somatoforme Störungen, Essstörungen, sexuelle Funktionsstörungen, Persönlichkeitsstörungen	Psychoanalyse und tiefenpsychologisch fundierte Psychotherapie	TP und AP	TP: *M* 78 Sitzungen *M* 2.4 Jahre AP: *M* 193 Sitzungen *M* 3 Jahre	überwiegend Individualtherapie
Klar (2005) Schleussner (2005) **Schleussner Studie** [18]	Originalstudie	Gemischte Störungen Neurotische Störungen, affektive Störungen, Persönlichkeitsstörungen	Individualpsychologisch-psychoanalytische Psychotherapie	AP	> 3 Jahre	Individualtherapie

Autoren & Kurztitel der Studien	Datenzugriff	Anwendungsbereich & Diagnosen	Verfahrenstitel im Original	Einordnung des psychodynamischen Verfahrens nach Psychotherapie-richtlinie	Sitzungsanzahl, Dauer & Frequenz der psychodynamischen Behandlung	Setting
Knekt et al. (2004) Knekt, Lindfors, Härkänen et al. (2008) Knekt, Lindfors, Laaksonen et al. (2008) **Helsinki Studie** [19]	Originalstudie	Gemischte Störungen Depressive Episode, rezidivierende depressive Episode, nicht näher bezeichnete affektive Störungen, bipolare affektive Störung, soziale Phobien, GAD, Panikstörung, spezifische Phobien, Zwangsstörung, Persönlichkeitsstörungen	Short-term psychodynamic psychotherapy und long-term psychodynamic psychotherapy	TP und AP	TP: *M* 18.5 Sitzungen 1/Woche AP: *M* 232 Sitzungen 2-3/Woche	Individualtherapie
Kurzweil (2008) **Kurzweil Studie** [20]	Originalstudie	Affektive Störungen Postnatale Depression	Long-term psychodynamic treatment group for postnatal depression; relational-developmental psychoanalytic treatment	TP	*M* 58 Sitzungen *M* 35 Monate vierzehntägig	Gruppentherapie
Lehto et al. (2006) Lehto, Tolmunen, Joensuu et al. (2008) Lehto, Tolmunen, Kuikka et al. (2008) **MAP-Psy Studie** [21]	Originalstudie	Affektive Störungen Mittelgradige depressive Episode, schwere depressive Episode, rezidivierende depressive Störung (gegenwärtig mittelgradige Episode), rezidivierende depressive Störung (gegenwärtig schwere Episode)	Psychodynamic psychotherapy	TP	80 Sitzungen 2/Woche	Individualtherapie

Autoren & Kurztitel der Studien	Datenzugriff	Anwendungsbereich & Diagnosen	Verfahrenstitel im Original	Einordnung des psychodynamischen Verfahrens nach Psychotherapie-richtlinie	Sitzungsanzahl, Dauer & Frequenz der psychodynamischen Behandlung	Setting
Leichsenring et al. (2005) Leichsenring et al. (2008) **Göttinger Psychotherapiestudie** [22]	Originalstudie	Gemischte Störungen Affektive Störungen, Angststörungen, phobische Störungen, Zwangsstörung, Anpassungsstörungen, somatoforme Störungen, Essstörungen, sexuelle Funktionsstörungen, Persönlichkeitsstörungen	Analytische Langzeitpsychotherapie, psychoanalytic therapy	AP	*M* 253 Sitzungen mehrere Sitzungen/Woche	Individualtherapie
Leuzinger-Bohleber et al. (2001) Leuzinger-Bohleber et al. (2003) Stuhr (2001) **DPV-Studie** [23]	Originalstudie	Gemischte Störungen Persönlichkeitsstörungen, affektive Störungen, neurotische Störungen, Schizophrenie	Psychoanalysen / Langzeitpsychoanalysen und psychoanalytische Langzeitpsychotherapie	AP und Psa	*M* 371 Sitzungen 1-5/Woche	Individualtherapie
Lotz et al. (2006) **Lotz Studie** [24]	Originalstudie	Gemischte Störungen Angststörungen, Depression, unspezifizierte neurotische Störungen, Persönlichkeitsstörungen	Psychoanalytic group psychotherapy	TP	39 Sitzungen	Gruppentherapie
Luborsky et al. (2001) Roy et al. (2009) **Penn Psychoanalytic Treatment Collection Studie** [25]	Originalstudie	Gemischte Störungen „No systematic diagnostic information exists“ (P. Crits-Christoph, persönl. Mitteilung, 14.09.2009)	Psychoanalyses	Psa	*M* 623 Sitzungen *Me* 3 Jahre 4/Woche	Individualtherapie

Autoren & Kurztitel der Studien	Datenzugriff	Anwendungsbereich & Diagnosen	Verfahrenstitel im Original	Einordnung des psychodynamischen Verfahrens nach Psychotherapie-richtlinie	Sitzungsanzahl, Dauer & Frequenz der psychodynamischen Behandlung	Setting
Lundblad (2003) **Lundblad Studie** [26]	Originalstudie	Affektive Störungen Major Depression mit Persönlichkeitsstörungen	Long-term psychoanalytic psychotherapy	nicht eindeutig beurteilbar	1.5-4 Jahre	Individualtherapie
Maina et al. (2005) **Maina Studie** [27]	Originalstudie	Affektive Störungen Dystymia, Minor Depression, Anpassungsstörungen mit depressiver Verstimmung	Brief dynamic therapy, focused short-term psychoanalytic psychotherapy	TP	*M* 19.6 Sitzungen 1/Woche	Individualtherapie
Małyszczak et al. (2006) **Małyszczak Studie** [28]	Originalstudie	Gemischte Störungen Neurotische, Belastungs- und somatoforme Störungen, Persönlichkeitsstörungen	Group psychodynamic therapy	TP	12 Wochen 2 Sitzungen/Tag an 5 Tagen/Woche	Gruppentherapie
Paley et al. (2008) **Paley Studie** [29]	Originalstudie	Gemischte Störungen Von den N=57 Pat. konnten retrospektiv für N=30 Pat. die Diagnosen eruiert werden: 97% hatten eine „Diagnosis of depression“ (Paley et al., 2008, S. 161)	Psychodynamic-interpersonal therapy	TP	*M* 16.9 Sitzungen	Individualtherapie

Autoren & Kurztitel der Studien	Datenzugriff	Anwendungsbereich & Diagnosen	Verfahrenstitel im Original	Einordnung des psychodynamischen Verfahrens nach Psychotherapie-richtlinie	Sitzungsanzahl, Dauer & Frequenz der psychodynamischen Behandlung	Setting
Philips et al. (2006) Philips, Wennberg et al. (2007) Philips, Werbart et al. (2007) [Lindgren et al. (2010)] **Philips Studie** [30]	Originalstudie	Gemischte Störungen Major depression, nicht näher bezeichnete Angststörung, soziale Phobien, Dysthymia, Zwangsstörung, akute Belastungsreaktion, Anpassungsstörungen mit depressiver Verstimmung, Anpassungsstörung mit Angst, Agoraphobie ohne Angabe einer Panikstörung, Störungen durch Alkohol, nicht näher bezeichnete affektive Störung, Essstörungen, GAD, posttraumatische Belastungsstörung, spezifische Phobien, psychische und Verhaltensstörungen durch psychotrope Substanzen	Psychoanalytic individual and group psychotherapy	nicht eindeutig beurteilbar	Einzeltherapie: *M* 15 Monate Gruppentherapie: *M* 14 Monate	Individualtherapie und Gruppentherapie
Piper et al. (2001) **Complicated Grief Studie** [31]	Originalstudie	Gemischte Störungen Major Depression, Dysthymia, Persönlichkeitsstörungen	time-limited short-term group therapy (interpretive and supportive), time-limited group psychotherapy for complicated grief	TP	12 Sitzungen 1/Woche	Gruppentherapie
Salminen et al. (2008) **Salminen Studie** [32]	Originalstudie	Affektive Störungen Major Depression (leichte oder mittelgradige Episode)	short-term psychodynamic psychotherapy	TP	16 Sitzungen 1/Woche	Individualtherapie

Autoren & Kurz-titel der Studien	Datenzugriff	Anwendungsbereich & Diagnosen	Verfahrenstitel im Original	Einordnung des psychodynamischen Verfahrens nach Psychotherapie-richtlinie	Sitzungsanzahl, Dauer & Frequenz der psychodynamischen Behandlung	Setting
Sharpe et al. (2001) **Sharpe Studie** [33]	Originalstudie	Gemischte Störungen Männer mit sexuellen Missbrauchserfahrungen (keine Diagnosen via ICD oder DSM)	Slow-open analytic group	TP	ca. 27-54 Sitzungen 6-12 Monate 1/Woche	Gruppentherapie
Shaw et al. (2001) **Shaw Studie** [34]	Originalstudie	Gemischte Störungen Depressive disorders (ICD-9), anxiety states (ICD-9), Abhängigkeitssyndrom, akute Belastungsreaktion, andere neurotische Störungen, Persönlichkeitsstörungen	Psychodynamic interpersonal therapy	TP	10-12 Sitzungen 1/Woche	Individualtherapie
Stehle et al. (2004) **DGPT-Therapeutenerhebung** [35]	Originalstudie	Gemischte Störungen Psychoneurosen (Depression, Dysthymia, phobische Störungen, Angststörungen, Zwangsstörung, dissoziative Störungen, andere neurotische Störungen) Persönlichkeitsstörungen, Essstörungen, funktionelle Störungen (Somatisierungsstörungen), Psychosomatosen (psychologische Faktoren und Verhaltensfaktoren bei andernorts klassifizierten Krankheiten), Psychosen, Suchterkrankungen	Psychoanalyse und tiefenpsychologisch fundierte Psychotherapie	TP, AP und Psa	*M* 249 Sitzungen 1-4/Woche	Individualtherapie

Autoren & Kurztitel der Studien	**Datenzugriff**	**Anwendungsbereich & Diagnosen**	**Verfahrenstitel im Original**	**Einordnung des psychodynamischen Verfahrens nach Psychotherapie-richtlinie**	**Sitzungsanzahl, Dauer & Frequenz der psychodynamischen Behandlung**	**Setting**
Stiles et al. (2006/2007) **Stiles I Studie** [36]	Originalstudie	Gemischte Störungen Ängste, Depressionen, interpersonale Probleme, Selbstwertprobleme etc. (keine Diagnosen via ICD oder DSM)	Psychodynamic or psychoanalytic therapy / psychodynamische oder psychoanalytische Therapie	TP und AP	*M* 8.5 Sitzungen	Individualtherapie
Stiles et al. (2008) **Stiles II Studie** [37]	Replikation	Gemischte Störungen Ängste, Depressionen, interpersonale Probleme, Selbstwertprobleme etc. (keine Diagnosen via ICD oder DSM)	Psychodynamic or psychoanalytic therapy	TP und AP	*M* 8.06 Sitzungen	Individualtherapie
Tschuschke et al. (2007) Tschuschke et al. (2008) **Projekt für ambulante Gruppentherapie-Evaluation (PAGE)** [38]	Originalstudie	Gemischte Störungen Affektive Störungen, Angststörungen, Zwangsstörung, somatoforme Störungen, Reaktion auf schwere Belastungen und Anpassungsstörungen, Verhaltensauffälligkeiten mit körperlichen Störungen und Faktoren, Abhängigkeitssyndrom, Persönlichkeitsstörungen	Analytic and psychodynamically oriented long-term outpatient groups / analytische und tiefenpsychologische Gruppenpsychotherapie	TP und AP	TP und AP: *M* 81 Sitzungen 1-2/Woche AP: *M* 101 Sitzungen	Gruppentherapie
Van, Hendriksen et al. (2008) Van et al. (2009) **Van Studie** [39]	Originalstudie	Affektive Störungen Major depression mit (*double depression*) und ohne Dysthymie	Short-term psychodynamic supportive psychotherapy	TP	16 Sitzungen Sitzungen 1-8: 1/Woche Sitzungen 9-16: vierzehntägig	Individualtherapie

Autoren & Kurztitel der Studien	Datenzugriff	Anwendungsbereich & Diagnosen	Verfahrenstitel im Original	Einordnung des psychodynamischen Verfahrens nach Psychotherapie-richtlinie	Sitzungsanzahl, Dauer & Frequenz der psychodynamischen Behandlung	Setting
Vitriol et al. (2009) **Vitriol Studie [40]**	Originalstudie	Affektive Störungen „severe depression“ (ICD-10), posttraumatische Belastungsstörung	brief dynamic psychotherapy	TP	ca. 12 Sitzungen 3 Monate 1/Woche	Individualtherapie
von Wietersheim et al. (2002/2003) **von Wietersheim Studie [41]**	Originalstudie	Gemischte Störungen Dysthymia, leichte oder mittelgradige depressive Episode, rezidivierende depressive Störungen, Angststörungen, Persönlichkeitsstörungen, Essstörungen, Zwangsstörung, somatoforme Störungen, psychische Faktoren und Verhaltenseinflüsse bei andernorts klassifizierten Krankheiten	Katathym-imaginative Psychotherapie	TP	*M* 73.4 Sitzungen 1/Woche	Individualtherapie

Anmerkung: Die grau hinterlegten Zeilen verweisen auf Studien, die auf der allgemeinen methodischen Qualitätsdimension mit positivem Gesamtergebnis abgeschnitten haben. AP: Analytische Psychotherapie, GAD: Generalized Anxiety Disorder, M: Mittelwert, Me: Median, Psa: Psychoanalyse, TP: Tiefenpsychologisch fundierte Psychotherapie.

Anhang G: Allgemeine Übersicht II über alle Primärstudien (*N*=41)

Autoren & Kurztitel der Studien	Studiendesign (Gruppenvergleich, Stichprobenumfang)	Gruppenzuweisung	Messzeitpunkte und ggf. Katamnesezeitraum	Störungsspezifische Behandlung	a priori Festlegung der Sitzungsanzahl	Verwendung von Manual bzw. Behandlungsrichtlinien	Therapeutentraining zwecks Studie	Implementationskontrolle / Adherence	Ausschluss subklinischer Symptomausprägungen	Ausschluss komorbider Störungen
Abbass (2006) **Abbass-A Studie** [1]	Ein-Gruppen-Design (TP) *N*=10 prospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: + 6 Mo	+	-	+	-	-	+	-
Abbass (2002) **Abbass-B Studie** [2]	- TP - WL *N*=166 retrospektiv	WL: weder random. noch Selbstzuteilung	Prä: + Post: + Katamnese: -	-	-	+	-	-	+	-
Barber et al. (2004) Wilczek et al. (2004) [Wilczek et al. (1998)] **Wilczek Studie** [3]	- AP - unbehandelte KG (3 Jahre nach Therapiebeginn) *N*=55 prospektiv	Selbstzuteilung (ohne Parallelisierung, Stratifizierung oder Matching)	Prä: + Post: - Katamnese: + *M* 6 Mo	-	-	-	-	-	-	-

Autoren & Kurztitel der Studien	Studiendesign (Gruppenvergleich, Stichprobenumfang)	Gruppenzuweisung	Messzeitpunkte und ggf. Katamnesezeitraum	Störungsspezifische Behandlung	a priori Festlegung der Sitzungsanzahl	Verwendung von Manual bzw. Behandlungsrichtlinien	Therapeutentraining zwecks Studie	Implementationskontrolle / Adherence	Ausschluss subklinischer Symptomausprägungen	Ausschluss komorbider Störungen
Blatt et al. (2004) Wallerstein (1986/1989) **Reanalyse Menninger Studie** [4]	verfahrensinterner Vergleich: - supportive-expressive psychotherapy (AP) - Psychoanalysis (Psa) *N*=33 retrospektiv	Selbstzuteilung (ohne Parallelisierung, Stratifizierung oder Matching)	Prä: + Post: + Katamnese: -	-	-	-	-	+	nicht beurteilbar	-
Blomberg et al. (2001) [Grant et al. (2004) Lazar et al. (2006) Sandell et al. (1999)] **Stockholm Outcome of Psychotherapy and Psychoanalysis (STOPP)** [5]	verfahrensinterner Vergleich: - Psychodynamic long-term psychotherapy (AP) - Psychoanalysis (Psa) *N*=405 retrospektiv	Selbstzuteilung (ohne Parallelisierung, Stratifizierung oder Matching)	Kombination aus Längs- und Querschnitterhebungen mit 2-Jahres-Katamnese	-	-	-	-	-	-	-

Autoren & Kurztitel der Studien	Studiendesign (Gruppenvergleich, Stichprobenumfang)	Gruppenzuweisung	Messzeitpunkte und ggf. Katamnesezeitraum	Störungsspezifische Behandlung	a priori Festlegung der Sitzungsanzahl	Verwendung von Manual bzw. Behandlungsrichtlinien	Therapeutentraining zwecks Studie	Implementationskontrolle / Adherence	Ausschluss subklinischer Symptomausprägungen	Ausschluss komorbider Störungen
Bond et al. (2004/2006) Perry et al. (2009) **Bond Studie** [6]	Ein-Gruppen-Design (AP) *N*=53 prospektiv	[Ein-Gruppen-Design]	Prä: + Post: - Katamnese: + 1.5 Jahre 2 Jahre	-	-	-	-	-	+	-
Bradshaw et al. (2009) **Bradshaw Studie** [7]	Ein-Gruppen-Design (TP) *N*=78 retrospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: -	-	-	-	-	-	-	-
Brockmann et al. (2002) Brockmann et al. (2006) **Frankfurt-Hamburg-Studie** [8]	- AP - VT *N*=62 prospektiv	Selbstzuteilung (ohne Parallelisierung, Stratifizierung oder Matching)	Prä: + Post: - Katamnese: + 4 Jahre	-	-	-	-	+	+	-
Cooper et al. (2003) **Cooper Studie** [9]	- TP - TAU - Non-directive counselling - KVT *N*=193 prospektiv	random.	Prä: + Post: + Katamnese: + 4.5 Mo 1 Jahr 4.6 Jahre	+	+	+	+	+	+	-

Autoren & Kurztitel der Studien	Studiendesign (Gruppenvergleich, Stichprobenumfang)	Gruppenzuweisung	Messzeitpunkte und ggf. Katamnesezeitraum	Störungsspezifische Behandlung	a priori Festlegung der Sitzungsanzahl	Verwendung von Manual bzw. Behandlungsrichtlinien	Therapeutentraining zwecks Studie	Implementationskontrolle / Adherence	Ausschluss subklinischer Symptomausprägungen	Ausschluss komorbider Störungen
Gordon (2001) **Gordon Studie** [10]	Ein-Gruppen-Design (AP) *N*=55 retrospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: -	-	-	-	-	-	+	-
Grande et al. (2006/2009) Rudolf et al. (2004) **Praxisstudie psychoanalytischer Langzeittherapie (PAL)** [11]	verfahrensinterner Vergleich: - Psychodynamic therapy (TP) - Psychoanalytic therapy (AP) *N*=76 prospektiv	Selbstzuteilung (mit Parallelisierung, Stratifizierung oder Matching)	Prä: + Post: + Katamnese: + 1 Jahr	-	-	-	-	+	+	-
[Guthrie et al. (1998)] Guthrie et al. (1999) **Guthrie Studie** [12]	- TP - TAU *N*=110 prospektiv	random.	Prä: + Post: + Katamnese: + 6 Mo	-	+	+	+	+	+	-

Autoren & Kurztitel der Studien	Studiendesign (Gruppenvergleich, Stichprobenumfang)	Gruppenzuweisung	Messzeitpunkte und ggf. Katamnesezeitraum	Störungsspezifische Behandlung	a priori Festlegung der Sitzungsanzahl	Verwendung von Manual bzw. Behandlungsrichtlinien	Therapeutentraining zwecks Studie	Implementationskontrolle / Adherence	Ausschluss subklinischer Symptomausprägungen	Ausschluss komorbider Störungen
Hecke et al. (2008) Junkert-Tress et al. (2001) Junkert-Tress et al. (1999) **Düsseldorfer-Kurzzeitpsychotherapie-Projekt** [13]	Ein-Gruppen-Design (TP) *N*=130 prospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: + 6 Mo 1 Jahr	-	+	+	+	+	+	-
Hilsenroth et al. (2003) **Hilsenroth Studie** [14]	Ein-Gruppen-Design (TP) *N*=27 prospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: -	+	-	-	+	+	+	-

Autoren & Kurztitel der Studien	Studiendesign (Gruppenvergleich, Stichprobenumfang)	Gruppenzuweisung	Messzeitpunkte und ggf. Katamnesezeitraum	Störungsspezifische Behandlung	a priori Festlegung der Sitzungsanzahl	Verwendung von Manual bzw. Behandlungsrichtlinien	Therapeutentraining zwecks Studie	Implementationskontrolle / Adherence	Ausschluss subklinischer Symptomausprägungen	Ausschluss komorbider Störungen
Høglend et al. (2006) Høglend et al. (2008) **Høglend Studie** [15]	verfahrensinterner Vergleich: - Dynamic psychotherapy *mit* Übertragungsdeutung (TP) - Dynamic psychotherapy *ohne* Übertragungsdeutung (TP) *N*=100 prospektiv	random.	Prä: + Post: + Katamnese: + 1 Jahr 3 Jahre	-	+	+	+	+	+	+
Jakobsen et al. (2005a/b) Mattanza et al. (2005) **Schweizer Praxisstudie psychoanalytischer Langzeittherapie (Schweizer PAL)** [16]	Ein-Gruppen-Design (TP und AP) *N*=37 prospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: + 1 Jahr 3 Jahre	-	-	-	-	-	+	-

Autoren & Kurztitel der Studien	**Studiendesign (Gruppenvergleich, Stichprobenumfang)**	**Gruppenzuweisung**	**Messzeitpunkte und ggf. Katamnesezeitraum**	**Störungsspezifische Behandlung**	**a priori Festlegung der Sitzungsanzahl**	**Verwendung von Manual bzw. Behandlungsrichtlinien**	**Therapeutentraining zwecks Studie**	**Implementationskontrolle / Adherence**	**Ausschluss subklinischer Symptomausprägungen**	**Ausschluss komorbider Störungen**
Keller et al. (2001) **Jungianische Psychoanalyse und Psychotherapie Studie** [17]	Ein-Gruppen-Design (TP und AP) *N*=152 retrospektiv	[Ein-Gruppen-Design]	Prä: - Post: - Katamnese: + 6 Jahre	-	-	-	-	-	+	-
Klar (2005) Schleussner (2005) **Schleussner Studie** [18]	- AP - WL (Prä-1Jahr) *N*=196 prospektiv	WL: weder random. noch Selbstzuteilung	Prä: + Post: - Katamnese: + *M* 2 Jahre	-	-	-	-	-	+	-
Knekt et al. (2004) Knekt, Lindfors, Härkänen et al. (2008) Knekt, Lindfors, Laaksonen et al. (2008) **Helsinki Studie** [19]	- Solution-focused therapy und verfahrensinterner Vergleich: - StPP (TP) - LtPP (AP) *N*=326 prospektiv	random.	Prä: + Post: - Katamnese: + 5 Mo bis 2.5 Jahre (*M* 1.8 Jahre) in Abhängigkeit der 3 Treatmentarme	-	-	-	-	-	+	+

Autoren & Kurztitel der Studien	Studiendesign (Gruppenvergleich, Stichprobenumfang)	Gruppenzuweisung	Messzeitpunkte und ggf. Katamnesezeitraum	Störungsspezifische Behandlung	a priori Festlegung der Sitzungsanzahl	Verwendung von Manual bzw. Behandlungsrichtlinien	Therapeutentraining zwecks Studie	Implementationskontrolle / Adherence	Ausschluss subklinischer Symptomausprägungen	Ausschluss komorbider Störungen
Kurzweil (2008) **Kurzweil Studie** [20]	Ein-Gruppen-Design (TP) *N*=31 prospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: -	-	-	-	-	-	+	-
Lehto et al. (2006) Lehto, Tolmunen, Joensuu et al. (2008) Lehto, Tolmunen, Kuikka et al. (2008) **MAP-Psy Studie** [21]	- TP - WL (6 Mo) *N*=22 prospektiv	random.	Prä: + Post: + Katamnese: -	-	k.A.	-	-	-	+	+
Leichsenring et al. (2005) Leichsenring et al. (2008) **Göttinger Psychotherapiestudie** [22]	Ein-Gruppen-Design (AP) *N*=36 prospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: + 1 Jahr	-	-	-	-	-	+	-

Autoren & Kurztitel der Studien	**Studiendesign (Gruppenvergleich, Stichprobenumfang)**	**Gruppenzuweisung**	**Messzeitpunkte und ggf. Katamnesezeitraum**	**Störungsspezifische Behandlung**	**a priori Festlegung der Sitzungsanzahl**	**Verwendung von Manual bzw. Behandlungsrichtlinien**	**Therapeutentraining zwecks Studie**	**Implementationskontrolle / Adherence**	**Ausschluss subklinischer Symptomausprägungen**	**Ausschluss komorbider Störungen**
Leuzinger-Bohleber et al. (2001) Leuzinger-Bohleber et al. (2003) Stuhr (2001) **DPV-Studie** [23]	Ein-Gruppen-Design (AP und Psa) *N*=401 retrospektiv	[Ein-Gruppen-Design]	Prä: - Post: - Katamnese: + *M* 6.5 Jahre	-	-	-	-	-	nicht beurteilbar	-
Lotz et al. (2006) **Lotz Studie** [24]	Ein-Gruppen-Design (TP) *N*=139 retrospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: -	-	+	-	-	-	+	-
Luborsky et al. (2001) Roy et al. (2009) **Penn Psychoanalytic Treatment Collection Studie** [25]	Ein-Gruppen-Design (Psa) *N*=17 retrospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: -	-	-	-	-	-	nicht beurteilbar	-

Autoren & Kurztitel der Studien	Studiendesign (Gruppenvergleich, Stichprobenumfang)	Gruppenzuweisung	Messzeitpunkte und ggf. Katamnesezeitraum	Störungsspezifische Behandlung	a priori Festlegung der Sitzungsanzahl	Verwendung von Manual bzw. Behandlungsrichtlinien	Therapeutentraining zwecks Studie	Implementationskontrolle / Adherence	Ausschluss subklinischer Symptomausprägungen	Ausschluss komorbider Störungen
Lundblad (2003) **Lundblad Studie** [26]	Ein-Gruppen-Design („Long-term psychoanalytic psychotherapy“) *N*=8 prospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: + 1 Jahr	-	-	-	-	-	+	-
Maina et al. (2005) **Maina Studie** [27]	- TP - WL - BSP *N*=30 prospektiv	random.	Prä: + Post: + Katamnese: + 6 Mo	-	+	+	-	+	+	+
Małyszczak et al. (2006) **Małyszczak Studie** [28]	Ein-Gruppen-Design (TP) *N*=77 prospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: -	-	+	-	-	-	+	-

Autoren & Kurztitel der Studien	Studiendesign (Gruppenvergleich, Stichprobenumfang)	Gruppenzuweisung	Messzeitpunkte und ggf. Katamnesezeitraum	Störungsspezifische Behandlung	a priori Festlegung der Sitzungsanzahl	Verwendung von Manual bzw. Behandlungsrichtlinien	Therapeutentraining zwecks Studie	Implementationskontrolle / Adherence	Ausschluss subklinischer Symptomausprägungen	Ausschluss komorbider Störungen
Paley et al. (2008) **Paley Studie** [29]	Ein-Gruppen-Design (TP) *N*=67 retrospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: -	-	-	+	+	-	-	-
Philips et al. (2006) Philips, Wennberg et al. (2007) Philips, Werbart et al. (2007) [Lindgren et al. (2010)] **Philips Studie** [30]	Ein-Gruppen-Design („Psychoanalytic individual and group psychotherapy") *N*=134 prospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: + 1.5 Jahre	-	teils ja / teils nein	-	-	-	-	-
Piper et al. (2001) **Complicated Grief Studie** [31]	verfahrensinterner Vergleich: - Interpretive therapy (TP) - Supportive therapy (TP) *N*=139 prospektiv	random.	Prä: + Post: + Katamnese: -	+	+	+	-	+	nicht beurteilbar	-

Autoren & Kurztitel der Studien	Studiendesign (Gruppenvergleich, Stichprobenumfang)	Gruppenzuweisung	Messzeitpunkte und ggf. Katamnesezeitraum	Störungsspezifische Behandlung	a priori Festlegung der Sitzungsanzahl	Verwendung von Manual bzw. Behandlungsrichtlinien	Therapeutentraining zwecks Studie	Implementationskontrolle / Adherence	Ausschluss subklinischer Symptomausprägungen	Ausschluss komorbider Störungen
Salminen et al. (2008) **Salminen Studie** [32]	- TP - Psychopharmakatherapie *N*=51 prospektiv	random.	Prä: + Post: + Katamnese: -	-	+	-	-	-	+	+
Sharpe et al. (2001) **Sharpe Studie** [33]	Ein-Gruppen-Design (TP) *N*=27 prospektiv	[Ein-Gruppen-Design]	Prä: + Post: + Katamnese: + 6 Mo	+	+	-	-	-	-	-
Shaw et al. (2001) **Shaw Studie** [34]	- TP - WL *N*=54 prospektiv	random.	Prä: + Post: + Katamnese: -	-	+	+	+	-	+	-

Autoren & Kurztitel der Studien	Studiendesign (Gruppenvergleich, Stichprobenumfang)	Gruppenzuweisung	Messzeitpunkte und ggf. Katamnesezeitraum	Störungsspezifische Behandlung	a priori Festlegung der Sitzungsanzahl	Verwendung von Manual bzw. Behandlungsrichtlinien	Therapeutentraining zwecks Studie	Implementationskontrolle / Adherence	Ausschluss subklinischer Symptomausprägungen	Ausschluss komorbider Störungen
Stehle et al. (2004) **DGPT-Therapeutenerhebung** [35]	verfahrensinterner Vergleich: - „Psychoanalyse“ (AP und Psa) - TP *N*=605 retrospektiv	Selbstzuteilung (ohne Parallelisierung, Stratifizierung oder Matching)	Prä: - Post: + Katamnese: -	-	-	-	-	-	nicht beurteilbar	-
Stiles et al. (2006/2007) **Stiles I Studie** [36]	- TP und AP - KVT - PZT *N*=1309 retrospektiv	Selbstzuteilung (ohne Parallelisierung, Stratifizierung oder Matching)	Prä: + Post: + Katamnese: -	-	-	-	-	-	-	-
Stiles et al. (2008) **Stiles II Studie** [37]	- TP und AP - KVT - PZT *N*=5613 retrospektiv	Selbstzuteilung (ohne Parallelisierung, Stratifizierung oder Matching)	Prä: + Post: + Katamnese: -	-	-	-	-	-	-	-

Autoren & Kurztitel der Studien	Studiendesign (Gruppenvergleich, Stichprobenumfang)	Gruppenzuweisung	Messzeitpunkte und ggf. Katamnesezeitraum	Störungsspezifische Behandlung	a priori Festlegung der Sitzungsanzahl	Verwendung von Manual bzw. Behandlungsrichtlinien	Therapeutentraining zwecks Studie	Implementationskontrolle / Adherence	Ausschluss subklinischer Symptomausprägungen	Ausschluss komorbider Störungen
Tschuschke et al. (2007) Tschuschke et al. (2008) **Projekt für ambulante Gruppentherapie-Evaluation (PAGE)** [38]	- TP und AP - Psychodramatherapie *N*=620 prospektiv	Selbstzuteilung (ohne Parallelisierung, Stratifizierung oder Matching)	Prä: + Post: + Katamnese: -	-	-	-	-	-	+	-
Van, Hendriksen et al. (2008) Van et al. (2009) **Van Studie** [39]	verfahrensinterner Vergleich: - SPSP random. (TP) - SPSP selbstzugewiesen (TP) *N*=119 prospektiv	random. und Selbstzuteilung	Prä: + Post: + Katamnese: -	-	+	+	+	+	+	-
Vitriol et al. (2009) **Vitriol Studie** [40]	- TP - TAU *N*=87 prospektiv	random.	Prä: + Post: + Katamnese: + 3 Mo	+	+	-	-	-	+	+

Autoren & Kurztitel der Studien	Studiendesign (Gruppenvergleich, Stichprobenumfang)	Gruppenzuweisung	Messzeitpunkte und ggf. Katamnesezeitraum	Störungsspezifische Behandlung	a priori Festlegung der Sitzungsanzahl	Verwendung von Manual bzw. Behandlungsrichtlinien	Therapeutentraining zwecks Studie	Implementationskontrolle / Adherence	Ausschluss subklinischer Symptomausprägungen	Ausschluss komorbider Störungen
von Wietersheim et al. (2002/2003) **von Wietersheim Studie** [41]	- TP - WL (6 Mo) *N*=140 prospektiv	random.	Prä: + Post: + Katamnese: + 1.5 Jahre	-	+	-	-	-	+	-

Anmerkung: Die grau hinterlegten Zeilen verweisen auf Studien, die auf der allgemeinen methodischen Qualitätsdimension mit positivem Gesamtergebnis abgeschnitten haben. AP: Analytische Psychotherapie, BSP: Brief Supportive Psychotherapy, k.A.: keine Angaben, KG: Kontrollgruppe, KVT: Kognitive Verhaltenstherapie, LtPP: Long-term Psychodynamic Psychotherapy, M: Mittelwert, Mo: Monate, Psa: Psychoanalyse, PZT: Personzentrierte Therapie, random.: randomisiert, SPSP: Short-term Psychodynamic Supportive Psychotherapy, StPP: Short-term Psychodynamic Psychotherapy, TAU: Treatment-As-Usual, TP: Tiefenpsychologisch fundierte Psychotherapie, VT: Verhaltenstherapie, WL: Warteliste, +: trifft zu, –: trifft nicht zu.

Anhang H: Clusterzugehörigkeit der Primärstudien (*N*=41)

Autoren & Kurztitel der Studien	Clusterlösung der Two-Step CA (2 Cluster)	Clusterlösung der hierarchischen CA (2 Cluster)	Clusterlösung der hierarchischen CA (3 Cluster)
Abbass (2006) **Abbass-A Studie [1]**	Cluster 2	Cluster 2	Cluster 3
Abbass (2002) **Abbass-B Studie [2]**	Cluster 1	Cluster 1	Cluster 2
Barber et al. (2004) Wilczek et al. (2004) [Wilczek et al. (1998)] **Wilczek Studie [3]**	Cluster 1	Cluster 1	Cluster 2
Blatt et al. (2004) Wallerstein (1986/1989) **Reanalyse Menninger Studie [4]**	Cluster 1	Cluster 1	Cluster 2
Blomberg et al. (2001) [Grant et al. (2004) Lazar et al. (2006) Sandell et al. (1999)] **Stockholm Outcome of Psychotherapy and Psychoanalysis (STOPP) [5]**	Cluster 1	Cluster 1	Cluster 2
Bond et al. (2004/2006) Perry et al. (2009) **Bond Studie [6]**	Cluster 1	Cluster 1	Cluster 2
Bradshaw et al. (2009) **Bradshaw Studie [7]**	Cluster 1	Cluster 1	Cluster 1
Brockmann et al. (2002) Brockmann et al. (2006) **Frankfurt-Hamburg-Studie [8]**	Cluster 1	Cluster 1	Cluster 2
Cooper et al. (2003) **Cooper Studie [9]**	Cluster 2	Cluster 2	Cluster 3
Gordon (2001) **Gordon Studie [10]**	Cluster 1	Cluster 1	Cluster 2

Autoren & Kurztitel der Studien	Clusterlösung der Two-Step CA (2 Cluster)	Clusterlösung der hierarchischen CA (2 Cluster)	Clusterlösung der hierarchischen CA (3 Cluster)
Grande et al. (2006/2009) Rudolf et al. (2004) **Praxisstudie psychoanalytischer Langzeittherapie (PAL)** [11]	Cluster 1	Cluster 1	Cluster 2
[Guthrie et al. (1998)] Guthrie et al. (1999) **Guthrie Studie** [12]	Cluster 2	Cluster 2	Cluster 3
Hecke et al. (2008) Junkert-Tress et al. (2001) Junkert-Tress et al. (1999) **Düsseldorfer-Kurzzeitpsychotherapie-Projekt** [13]	Cluster 2	Cluster 2	Cluster 3
Hilsenroth et al. (2003) **Hilsenroth Studie** [14]	*Ausreißer*	*Ausreißer*	*Ausreißer*
Høglend et al. (2006) Høglend et al. (2008) **Høglend Studie** [15]	*Ausreißer*	*Ausreißer*	*Ausreißer*
Jakobsen et al. (2005a/b) Mattanza et al. (2005) **Schweizer Praxisstudie psychoanalytischer Langzeittherapie (Schweizer PAL)** [16]	Cluster 1	Cluster 1	Cluster 2
Keller et al. (2001) **Jungianische Psychoanalyse und Psychotherapie Studie** [17]	Cluster 1	Cluster 1	Cluster 2
Klar (2005) Schleussner (2005) **Schleussner Studie** [18]	Cluster 1	Cluster 1	Cluster 2

Autoren & Kurztitel der Studien	Clusterlösung der Two-Step CA (2 Cluster)	Clusterlösung der hierarchischen CA (2 Cluster)	Clusterlösung der hierarchischen CA (3 Cluster)
Knekt et al. (2004) Knekt, Lindfors, Härkänen et al. (2008) Knekt, Lindfors, Laaksonen et al. (2008) **Helsinki Studie** [19]	Cluster 1	Cluster 1	Cluster 2
Kurzweil (2008) **Kurzweil Studie** [20]	Cluster 1	Cluster 1	Cluster 1
Lehto et al. (2006) Lehto, Tolmunen, Joensuu et al. (2008) Lehto, Tolmunen, Kuikka et al. (2008) **MAP-Psy Studie** [21]	*exkludiert*	*exkludiert*	*exkludiert* [Cluster 3]
Leichsenring et al. (2005) Leichsenring et al. (2008) **Göttinger Psychotherapiestudie** [22]	Cluster 1	Cluster 1	Cluster 2
Leuzinger-Bohleber et al. (2001) Leuzinger-Bohleber et al. (2003) Stuhr (2001) **DPV-Studie** [23]	Cluster 1	Cluster 1	Cluster 2
Lotz et al. (2006) **Lotz Studie** [24]	Cluster 1	Cluster 1	Cluster 1
Luborsky et al. (2001) Roy et al. (2009) **Penn Psychoanalytic Treatment Collection Studie** [25]	Cluster 1	Cluster 1	Cluster 2
Lundblad (2003) **Lundblad Studie** [26]	Cluster 1	Cluster 1	Cluster 2
Maina et al. (2005) **Maina Studie** [27]	Cluster 2	Cluster 2	Cluster 3

Autoren & Kurztitel der Studien	Clusterlösung der Two-Step CA (2 Cluster)	Clusterlösung der hierarchischen CA (2 Cluster)	Clusterlösung der hierarchischen CA (3 Cluster)
Małyszczak et al. (2006) **Małyszczak Studie [28]**	Cluster 2	Cluster 2	Cluster 3
Paley et al. (2008) **Paley Studie [29]**	*Ausreißer*	*Ausreißer*	*Ausreißer*
Philips et al. (2006) Philips, Wennberg et al. (2007) Philips, Werbart et al. (2007) [Lindgren et al. (2010)] **Philips Studie [30]**	*exkludiert*	*exkludiert*	*exkludiert* [Cluster 1]
Piper et al. (2001) **Complicated Grief Studie [31]**	Cluster 2	Cluster 2	Cluster 3
Salminen et al. (2008) **Salminen Studie [32]**	Cluster 2	Cluster 2	Cluster 3
Sharpe et al. (2001) **Sharpe Studie [33]**	Cluster 1	Cluster 1	Cluster 1
Shaw et al. (2001) **Shaw Studie [34]**	Cluster 2	Cluster 2	Cluster 3
Stehle et al. (2004) **DGPT-Therapeutenerhebung [35]**	Cluster 1	Cluster 1	Cluster 2
Stiles et al. (2006/2007) **Stiles I Studie [36]**	Cluster 1	Cluster 1	Cluster 2
Stiles et al. (2008) **Stiles II Studie [37]**	Cluster 1	Cluster 1	Cluster 2

Autoren & Kurztitel der Studien	Clusterlösung der Two-Step CA (2 Cluster)	Clusterlösung der hierarchischen CA (2 Cluster)	Clusterlösung der hierarchischen CA (3 Cluster)
Tschuschke et al. (2007) Tschuschke et al. (2008) **Projekt für ambulante Gruppentherapie-Evaluation (PAGE)** [38]	Cluster 1	Cluster 1	Cluster 2
Van, Hendriksen et al. (2008) Van et al. (2009) **Van Studie** [39]	Cluster 2	Cluster 2	Cluster 3
Vitriol et al. (2009) **Vitriol Studie** [40]	Cluster 2	Cluster 2	Cluster 3
von Wietersheim et al. (2002/2003) **von Wietersheim Studie** [41]	Cluster 1	Cluster 1	Cluster 1

Anmerkung: CA: Clusteranalyse.

Anhang J: Übersicht über alle Primärstudien (*N*=41) und deren Ergebnisse auf den Dimensionen der allgemeinen methodischen Qualität, der internen und der externen Validität

Autoren & Kurztitel der Studien	AMQ	IV	EV
Abbass (2006) **Abbass-A Studie [1]**	+	-	+
Abbass (2002) **Abbass-B Studie [2]**	-	-	-
Barber et al. (2004) Wilczek et al. (2004) [Wilczek et al. (1998)] **Wilczek Studie [3]**	-	-	-
Blatt et al. (2004) Wallerstein (1986/1989) **Reanalyse Menninger Studie [4]**	-	-	-
Blomberg et al. (2001) [Grant et al. (2004) Lazar et al. (2006) Sandell et al. (1999)] **Stockholm Outcome of Psychotherapy and Psychoanalysis (STOPP) [5]**	-	-	-
Bond et al. (2004/2006) Perry et al. (2009) **Bond Studie [6]**	+	-	+
Bradshaw et al. (2009) **Bradshaw Studie [7]**	+	-	+
Brockmann et al. (2002) Brockmann et al. (2006) **Frankfurt-Hamburg-Studie [8]**	-	-	-
Cooper et al. (2003) **Cooper Studie [9]**	+	+	-
Gordon (2001) **Gordon Studie [10]**	-	-	-

Autoren & Kurztitel der Studien	AMQ	IV	EV
Grande et al. (2006/2009)-Rudolf et al. (2004) **Praxisstudie psychoanalytischer Langzeittherapie (PAL)** [11]	+	-	+
[Guthrie et al. (1998)] Guthrie et al. (1999) **Guthrie Studie** [12]	-	-	-
Hecke et al. (2008) Junkert-Tress et al. (2001) Junkert-Tress et al. (1999) **Düsseldorfer-Kurzzeitpsychotherapie-Projekt** [13]	+	-	+
Hilsenroth et al. (2003) **Hilsenroth Studie** [14]	+	-	+
Høglend et al. (2006) Høglend et al. (2008) **Høglend Studie** [15]	+	-	+
Jakobsen et al. (2005a/b) Mattanza et al. (2005) **Schweizer Praxisstudie psychoanalytischer Langzeittherapie (Schweizer PAL)** [16]	+	-	+
Keller et al. (2001) **Jungianische Psychoanalyse und Psychotherapie Studie** [17]	-	-	-
Klar (2005) Schleussner (2005) **Schleussner Studie** [18]	-	-	-

Autoren & Kurztitel der Studien	AMQ	IV	EV
Knekt et al. (2004) Knekt, Lindfors, Härkänen et al. (2008) Knekt, Lindfors, Laaksonen et al. (2008) **Helsinki Studie [19]**	+	+	+
Kurzweil (2008) **Kurzweil Studie [20]**	-	-	-
Lehto et al. (2006) Lehto, Tolmunen, Joensuu et al. (2008) Lehto, Tolmunen, Kuikka et al. (2008) **MAP-Psy Studie [21]**	-	-	-
Leichsenring et al. (2005) Leichsenring et al. (2008) **Göttinger Psychothe-rapiestudie [22]**	+	-	+
Leuzinger-Bohleber et al. (2001) Leuzinger-Bohleber et al. (2003) Stuhr (2001) **DPV-Studie [23]**	-	-	-
Lotz et al. (2006) **Lotz Studie [24]**	-	-	-
Luborsky et al. (2001) Roy et al. (2009) **Penn Psychoanalytic Treatment Collection Studie [25]**	-	-	-
Lundblad (2003) **Lundblad Studie [26]**	-	-	-
Maina et al. (2005) **Maina Studie [27]**	+	+	-

Autoren & Kurztitel der Studien	AMQ	IV	EV
Małyszczak et al. (2006) **Małyszczak Studie [28]**	-	-	-
Paley et al. (2008) **Paley Studie [29]**	-	-	-
Philips et al. (2006) Philips, Wennberg et al. (2007) Philips, Werbart et al. (2007) [Lindgren et al. (2010)] **Philips Studie [30]**	+	-	+
Piper et al. (2001) **Complicated Grief Studie [31]**	+	-	+
Salminen et al. (2008) **Salminen Studie [32]**	-	-	-
Sharpe et al. (2001) **Sharpe Studie [33]**	-	-	-
Shaw et al. (2001) **Shaw Studie [34]**	+	-	+
Stehle et al. (2004) **DGPT-Therapeutenerhebung [35]**	-	-	-
Stiles et al. (2006/2007) **Stiles I Studie [36]**	-	-	-
Stiles et al. (2008) **Stiles II Studie [37]**	-	-	-

Autoren & Kurztitel der Studien	AMQ	IV	EV
Tschuschke et al. (2007) Tschuschke et al. (2008) **Projekt für ambulante Gruppentherapie-Evaluation (PAGE)** [38]	-	-	-
Van, Hendriksen et al. (2008) Van et al. (2009) **Van Studie** [39]	-	-	-
Vitriol et al. (2009) **Vitriol Studie** [40]	-	-	-
von Wietersheim et al. (2002/2003) **von Wietersheim Studie** [41]	-	-	-

Anmerkung: AMQ: Allgemeine methodische Qualität, EV: Externe Validität, IV: Interne Validität, +: positives Gesamtergebnis auf der betreffenden Dimension, –: negatives Gesamtergebnis auf der betreffenden Dimension.