PROFESSIONELLE PSYCHOTHERAPIEFORSCHUNG?

Ein Kommentar zu GRAWE, K., DONATI, R. und BERNAUER, F. (1994) „Psychotherapie im Wandel - Von der Konfession zur Profession“

Heinrich Geldschläger & Bernd Runde

„Wer die Psychologie liebt, hat oft Anlaß, sich der Psychotherapie-*Forschung* zu schämen.“ (GRAWE u. a., 1994, S. V; kursiv Gedrucktes von den Autoren dieses Beitrages hinzugefügt).

Es war dann doch keine Schamesröte, eher Unbehagen bei der Lektüre etlicher Passagen der Arbeit von GRAWE u. a., die uns als Psychologie-Liebhaber zum Verfassen dieses Diskussionsbeitrages veranlaßte.

Der Alptraum eines jeden Psychotherapieforschers besteht wohl in der Vorstellung, daß sein Werk weder erwähnt, noch kritisiert und somit auch kaum gelesen wird. Wenigstens in diesem Bereich können GRAWE u. a. beruhigt schlafen. Ihr Werk ist im ersten Jahr nach Erscheinen bereits in die dritte Auflage gegangen und hat nicht nur innerhalb der Psychotherapieforschungs-Kommune, sondern über Presse, Funk und Fernsehen auch in der Öffentlichkeit das beabsichtigte Interesse gefunden. Bei dem Buch handelt es sich einerseits um den Bericht über eine empirische Untersuchung zur Effektivität psychotherapeutischer Verfahren (v. a. Kapitel 2, 3, 4 und 5), andererseits um Schlußfolgerungen und Interpretationen der Ergebnisse dieser (und anderer) Untersuchung(en) mit erheblichen berufspolitischen Implikationen (v. a. Kapitel 1 und 6). Gerade auch wegen dieser berufspolitischen Brisanz ist das Buch mehrfach und auf verschiedenen Ebenen kritisiert worden. Wir wollen uns in diesem Beitrag im wesentlichen auf den empirischen Teil des Buches beschränken und die Vorgehensweise der Autoren kritisch würdigen. Auf andere, ebenso wichtige wie interessante Fragen, etwa nach der erkenntnistheoretischen Grundlage der Untersuchung, nach der Angemessenheit der gezogenen Schlußfolgerungen und nach deren berufspolitischen Konsequenzen, soll im Rahmen dieser Besprechung nicht eingegangen werden. Im folgenden werden wir also Fragestellung, Stichprobe und Methodik der empirischen Untersuchung von GRAWE u. a. kurz schildern und kritisch beurteilen.

Zur Fragestellung und zum Untersuchungsansatz

Das Buch von GRAWE u. a. „will aufklären über das, was ist, und das, was sein sollte“ (S. 1). Weniger pathetisch ausgedrückt soll der Frage nach der differentiellen Wirksamkeit derzeit bestehender Psychotherapie-Methoden nachgegangen werden. Die Wichtigkeit dieser Fragestellung ist allein schon vor dem Hintergrund der

kaum zu überblickenden Anzahl konkurrierender Psychotherapie-Methoden unbestritten.

Auch der Untersuchungsansatz, eine Sekundäranalyse möglichst vieler bereits vorliegender Studien durchzuführen, scheint angesichts der Vielzahl solcher Studien sinnvoll und wurde in verschiedenen Reviews und Metaanalysen bereits angewandt.

Zur Stichprobe

Was den Untersuchungsansatz von GRAWE u. a. von bereits existierenden Sekundäranalysen unterscheidet, ist der Anspruch, einen „möglichst vollständigen Überblick über die Ergebnisse aller bisher durchgeführter Therapiestudien“ (S. 50) zu geben. Um diesem Anspruch zu genügen, führten GRAWE u. a. eine umfangreiche Literatursuche nach kontrollierten Therapiestudien durch, die bis zum Jahreswechsel 1983/1984 veröffentlicht wurden. Unveröffentlichte Studien, z. B. Dissertationen, wurden nicht berücksichtigt „in der Annahme, daß die bedeutenderen Dissertationen publiziert würden“ (S. 57). Da GRAWE u. a. Aussagen über die Effektivität *psychotherapeutischer Verfahren* machen (wollen), sich diese aber nicht nur in *veröffentlichten* Effektivitätsstudien zeigt, bedeutet der Ausschluß unveröffentlichter Studien einen krassen Verstoß gegen den Anspruch, einen möglichst vollständigen Überblick zu liefern. Insbesondere unter der Annahme, daß sich die Ergebnisse unveröffentlichter Studien systematisch von denen veröffentlichter unterscheiden könnten, hätte zur Prüfung dieser Annahme zumindest eine Stichprobe unveröffentlichter Studien einbezogen werden müssen. Zur Beschaffung der sog. „grauen Literatur“ sieht das metaanalytische Methodeninventar z. B. vor, Experten mit der Bitte anzuschreiben, unveröffentlichte Forschungsarbeiten zum Themenkomplex zu nennen.

Unter dem Aspekt beschränkter Verarbeitungskapazitäten halten wir es für sehr vernünftig, daß GRAWE u. a. im Gegensatz zu anderen Überblicksarbeiten (z. B. SMITH, GLASS & MILLER, 1980) ausschließlich Studien analysierten, die klinisch relevante Stichproben umfaßten, während sogenannte Analogiestudien nicht berücksichtigt wurden.

Diese Literaturrecherche ergab etwa dreineinhalbtausend Studien, die den bisher aufgeführten Selektionskriterien entsprachen. Da „längst nicht alle ... Veröffentlichungen wirklich relevante Informationen ... enthielten“ (S. 58), wurde ein zweiter Selektionsschritt notwendig. Ausgeschlossen wurden neben Studien zu bestimmten Problembereichen (z. B. Rauchen, Übergewicht usw.) und Studien mit spezifischen Klientengruppen (Gefängnisinsassen, Kinder und Jugendliche bis 17 Jahre usw.) auch solche Studien, die bestimmten formalen Kriterien nicht genügten: Studien mit weniger als 4 Normseiten Umfang, Studien mit weniger als 4 Sitzungen Behandlungsdauer oder Studien mit weniger als 4 Patienten in einer der Behandlungsbedingungen. Die Zahl „4“ scheint hier das grundlegende Kriterium zu sein - andere

Begründungen bleiben dem Leser verschlossen! Prägnante Berichte über die Effektivität von Kurzzeittherapien würden von GRAWE u. a. also nicht berücksichtigt.

Es blieben 897 Studien, die auch diesen Selektionskriterien genügen, für die weitere Analyse.

Zur Methode

Zunächst erstellten GRAWE u. a. einen standardisierten Auswertungskatalog mit fast 1000 Einzelmerkmalen, hinsichtlich derer die Studien von den insgesamt 16 geschulten Auswertern beurteilt werden sollten. Dieser begrüßenswerten Vollständigkeit und Differenziertheit der Studienkodierung nach deskriptiven und qualitativen Aspekten steht jedoch das Problem der Urteilsgüte gegenüber. Bei nahezu 1000 Einzelmerkmalen dürfte es kaum möglich sein, eine auch nur ausreichende Beurteilergenauigkeit zu erzielen. Bei mangelhafter Urteilsqualität verliert jedoch jedes noch so differenzierte Auswertungssystem seine Aussagekraft. Daher wäre eine Überprüfung der Beurteilerübereinstimmung dringend geboten, auf die GRAWE u. a. jedoch unter Hinweis auf den zusätzlichen Aufwand verzichteten (S. 69).

Zur raschen Visualisierung der Güte der einzelnen Studien wurden aus den Urteilen zu ausgewählten Fragen des standardisierten Auswertungskatalogs acht Güteindizes gebildet und über alle Studien z-standardisiert. Diese an sich sinnvolle Informationsverdichtung weist jedoch in ihrer Umsetzung einige Mängel auf. Die acht Güteindizes sind alles andere als unabhängig, was eine Profilinterpretation unmöglich macht (Beispiel: Eine geringe Anzahl an Versuchspersonen wirkt sich negativ auf das Gütemaß „Interne Validität“ *und* „Klinische Relevanz“ aus. Interessanterweise wirkt sich auf die „Interne Validität“ ein Stichprobenumfang von *weniger als acht* negativ aus, auf die „Klinische Relevanz“ erst eine Anzahl von *weniger als zehn* - ohne daß die Autoren dies begründen!). Wozu also eine Güteprofil erstellen, wenn dies aufgrund methodischer Mängel nicht interpretierbar ist?

Weiterhin entspricht die Zuordnung einiger Fragen zu den Indizes nicht gängigem Verständnis, z. B. hat die Anzahl der Versuchspersonen keine Auswirkung auf die interne Validität. Auch einige der Güteindizes selbst sind wenig einleuchtend. Insbesondere die „Reichhaltigkeit der Ergebnisse“, die sich im wesentlichen aus der Anzahl signifikanter Effkete und Korrelationen ergibt, wird üblicherweise nicht als Gütekriterium einer Untersuchung betrachtet. Auch der Index „Vorsicht bei der Interpretation“ mutet etwas seltsam an, sollte diese sich doch aus der Gesamtheit aller Güteindizes ergeben!

Die entscheidende methodische Frage ist wohl die, wie GRAWE u. a. zu einem Urteil über die Effektivität einzelner Verfahren kommen. Hierzu wurden 10 Bereiche angenommen, in denen Veränderungen für die entsprechende Stichprobe gemessen werden konnten, z. B. Hauptsymptomatik, Befinden, Persönlichkeit, zwischenmenschlicher Bereich, etc.. Für jede Behandlungsbedingung in jeder Studie notierten GRAWE u. a., in welchen dieser Veränderungsbereiche Messungen vor-

genommen wurden, und ob Unterschiede (prä-post bzw. im Vergleich mit einer Kontroll- oder einer anderen Behandlungsgruppe) in diesen Messungen statistisch signifikant waren. Jeder Veränderungsbereich ließ sich durch mehrere Maße operationalisieren. Trat nur in einem dieser Maße ein signifikanter Unterschied auf, konstatierten GRAWE u. a. für den gesamten Veränderungsbereich einen signifikanten Unterschied (S. 94), und zwar unabhängig davon, wie die anderen Unterschiedsmessungen ausfielen. Dadurch wurden systematisch solche Studien bevorzugt, die mehrere Erfolgsmaße pro Veränderungsbereich anwendeten. Bei dem von GRAWE u. a. zugrundgelegten Signifikanzniveau von 5% ist es wahrscheinlicher, daß von 20 Unterschiedsmessungen wenigstens eine statistisch signifikant ausfällt, als daß keine Messung auf einen Unterschied hindeutet. Um bei GRAWE u. a. „gut abzuschneiden", kann also die Empfehlung gegeben werden, möglichst viele Meßinstrumente einzusetzen und darauf zu vertrauen, daß sicherlich eines dieser Instrumente einen Unterschied hervorbringt, der aber - bei einem Rest methodenkritischen Bewußtseins - keinen Unterschied mehr macht!

Für jedes von 42 Therapieverfahren (unterteilt in die fünf große Gruppen: humanistische, psychodynamische, kognitiv-behaviourale, interpersonale Therapien und Entspannungsverfahren, sowie für eklektizistische und richtungsübergreifende Therapien) wurden dann die Ergebnisse über die vorliegenden Studien kumuliert. Als Ergebnis ergab sich zu jeder Therapiemethode für jeden der zehn Veränderungsbereiche das Verhältnis von statistisch signifikanten Unterschiedsmessungen zu den insgesamt vorgenommenen Unterschiedsmessungen. Entsprechende Ergebnistabellen liegen für Prä-Post-Vergleiche und, soweit vorhanden, für Kontrollgruppenvergleiche vor (wobei zu letzteren nicht berichtet wird, ob es sich um Posttestvergleiche, um Prä-Post-Differenz-Vergleiche oder um eine Mischung beider handelt). Erfreulich ist die zusätzliche Differenzierung der Ergebnistabellen nach Verfahrensunterschieden, verschiedenen Störungsbildern und Settingvariablen, soweit dazu Daten vorliegen. Erfreulich auch, daß alle analysierten Studien in tabellarischer Form kurz beschrieben werden (bei 897 Studien machen allein diese Tabellen knapp ein Drittel des gesamten Buches aus!). Weniger erfreulich ist allerdings, daß die Anzahl der in diesen deskriptiven Tabellen beschriebenen Behandlungsbedingungen nie mit der Gesamtzahl der in den zugehörigen Ergebnistabellen berücksichtigten Bedingungen übereinstimmt. So finden sich beispielsweise für die Gestalttherapie insgesamt sieben Studien mit acht Behandlungsbedingungen. In der Ergebnistabelle werden diese jedoch auf zwei reduziert (S. 112 ff.). Als Gründe für diesen „Schwund" kommen neben fehlender Angaben auch qualitative Mängel in den Originalarbeiten in Frage. Leider explizieren GRAWE u. a. solche Ausschlußkriterien nicht. Zur Überprüfung oder der Replikation der Untersuchung von GRAWE u. a. wäre darüber hinaus eine kurze Ergebnisangabe zu jeder Einzelstudie notwendig.

Die ansonsten differenzierten und ausführlichen Ergebnisberichte können und sollten jedoch nicht über ein grundlegendes methodisches Manko der Auswertung von GRAWE u. a. hinwegtäuschen, nämlich der Problematik der statistischen Signifikanz von Unterschiedsmessungen als alleiniges Kriterium. Die Abhängigkeit

statistischer Signifikanztests von bestimmten Untersuchungsparametern wie der Stichprobengröße und der Streuung der Meßwerte und die Unterscheidung von statistischer Signifikanz und Relevanz gehört zum methodischen Einmaleins des Grundstudiums der Psychologie. Und für Studien des direkten Vergleichs der Effektivität verschiedener Therapieverfahren argumentiert GRAWE (1992, S. 143 f.) selbst ausführlich und völlig zutreffend gegen das Auszählen der Signifikanztestergebnisse (box-counting), da bei geringen Stichprobengrößen (für 83% der von GRAWE u. a. untersuchten Studien liegt die Behandlungsgruppengröße unter 30) und häufigen Varianzerweiterungen nach Psychotherapie (in 48% der von GRAWE u. a. analysierten Studien, aus denen entsprechende Daten vorliegen) die statistische Power, also die Wahrscheinlichkeit einen tatsächlichen Unterschied mit Hilfe des Signifikanztests zu 'entdecken', sehr gering (oft unter 50%) sein kann. „Wenn man dann die Signifikanzen und Nicht-Signifikanzen über viele Untersuchungen hinweg im Sinne eines 'Box-Scores' auszählt, ... dann führt dies ... zu einer 'doppelten Kumulierung des Fehlers 2. Art'"(GRAWE, 1992, S.143f.)

Genau das tun GRAWE u. a. dann für die Prä-Post- und Kontrollgruppenvergleiche zu den einzelnen Therapieverfahren. Zwar ist das Problem der statistischen Power für Prä-Post-Vergleiche nicht so schwerwiegend wie für unabhängige Kontrollgruppenvergleiche, bei wenig reliablen Meßinstrumenten (ein Thema, das GRAWE u. a. leider überhaupt nicht behandeln), sehr kleinen Stichproben und der erwähnten Streuungserweiterung scheint es uns aber immer noch so erheblich, daß die von GRAWE u. a. angewandte 'box-counting' Methode als unangemessen bezeichnet werden muß. Sie führt dazu, daß Studien mit großen und homogenen Stichproben, in denen möglichst viele Messungen (unabhängig von der Meßgüte) vorgenommen wurden, systematisch bevorzugt werden. Ansonsten wird jedoch die Stichprobengröße oder Studienqualität in keiner Weise gewichtet, ein signifikanter Unterschied aus einer qualitativ hervorragenden Untersuchung an 100 Personen wird also genau wie ein signifikanter Unterschied in einer methodisch schwachen Studie an 10 Personen gewertet.

Diese Mängel des Auswertungsverfahrens sind umso bedauerlicher, als seit Anfang der 80er Jahre (HUNTER, SCHMIDT & JACKSON, 1982) statistische Verfahren der Metaanalyse zur Berechnung mittlerer Effektstärken vorliegen, die zur Überwindung der geschilderten Probleme entwickelt wurden. Warum wandten GRAWE u. a. diese Verfahren nicht an? „Hauptsächlich deswegen, weil es uns bei der Ergebnisauswertung nicht nur auf die quantitative mittlere Wirkung ankam ..." (S. 65). Offenbar kam es den Autoren auf die quantitative mittlere Wirkung *überhaupt* nicht an, denn diese läßt sich mit dem benutzten Verfahren des 'box-counting' eben nicht ermitteln! Darüberhinaus schließt unseres Erachtens die exakte Berechnung von Effektstärken eine sorgfältige qualitative Analyse nicht aus. Als weiteres Argument gegen eine Effektstärkenberechnung führen GRAWE u. a. „bedeutsame Unterschiede in der Art der Untersuchungen zu den verschiedenen Therapiemethoden ..." (S.67) an, die bei einem Effektstärkenvergleich vernachlässigt würden. Diese Unterschiede werden bei der 'box-counting' Methode jedoch gleichermaßen vernachlässigt, während sie durch Moderatoranalysen im Rahmen einer

Metaanalyse berücksichtigt und überprüft werden können. Bleibt ein letzter Grund: „Die Berechnung von Effektstärken erfordert ein recht sophistiziertes statistisches Know-how ...“, so daß die Autoren sich außerstande sahen, „alle Auswerter (...) auf ein so hohes vergleichbares Niveau statistischer Versiertheit zu bringen“ (S. 66). Diese Einschätzung spricht wohl für sich.

Wie erschreckend ernst GRAWE u. a. diese Einschätzung meinen[1] , zeigt Kapitel 4.9, in dem GRAWE u. a. über ihren Versuch einer Metaanalyse direkter Therapievergleichsstudien berichten. In diese Metaanalyse wurden 41 bis einschließlich 1991 veröffentlichte Studien einbezogen, in denen mindestens zwei Therapieverfahren direkt verglichen wurden. Neben einem umstrittenen sogenannten Vorzeichen-Test wird als zweites Auswertungsverfahren die Berechnung von Effektstärken und deren Mittelung über korrespondierende Studien hinweg angewandt. Setzt man übliche metaanalytische Standards an (vgl. z.B. BEELMANN & BLIESENER, 1994), begehen die Autoren dabei folgenden Fehler:

GRAWE u. a. „berechneten für alle Studien (...), für jede Behandlungsbedingung und (...) für jedes erhobene Veränderungsmaß für jeden Meßzeitpunkt nach Ende der Therapie Effektstärken ...“ (S. 661). Diese Effektstärken wurden über alle Studien und Maße hinweg gemittelt. Gängiges Prozedere bei der Berechnung einer mittleren Effektstärke ist es jedoch, pro Studie nur eine Effektstärke einzubeziehen, diese jedoch zumindest entsprechend der Stichprobengröße zu gewichten (weitere Gewichtungsfaktoren wie z. B. die Studienqualität sind denkbar). Andernfalls erhalten Ergebnisse aus Studien mit vielen Maßen und Meßzeitpunkten ein unverhältnismäßig hohes Gewicht. So geschehen beim Vergleich der Gesprächspsychotherapie mit der Verhaltenstherapie (S. 664), in der von 723 Effektstärkenvergleichen mehr als die Hälfte (387!) aus einer einzigen von 19 Studien stammen. Daß diese Studie von GRAWE und Mitarbeitern selbst stammt, sei hier nur am Rande bemerkt. Durch diese Vorgehensweise wird der Grundidee der Metaanalyse, nämlich der Minimierung des Stichprobenfehlers, zuwidergehandelt.

Für die statistischen Vergleiche zwischen den einzelnen Therapieschulen mitteln GRAWE u. a. dann glücklicherweise jedoch die Effektstärken zunächst pro Studie und dann über alle Studien hinweg. Warum jedoch auch hier keine Gewichtung der Effektstärken nach Stichprobengröße und Qualitätsmerkmalen der Studie vorgenommen wird, bleibt dem Leser verschlossen. Auch die undifferenzierte Mittelung von Effektstärken zu den verschiedenen Veränderungsbereichen, die nach GRAWE u. a. zu einer konservativen Schätzung der Effektstärken führt, ist ein schlechtes Beispiel für das sogenannte „Äpfel-Birnen-Problem“ der Metaanalyse (s. z. B. BEELMANN & BLIESENER, 1994), das jedoch nicht verfahrensimmanent, sondern *benutzerabhängig* ist. Schließlich vermißt man bei GRAWE u. a. ebenfalls Angaben über die aufgeklärte Varianz einzelner Therapierichtungen, für die dann

[1] Die Autoren dieses Diskussionsbeitrages waren sich zunächst nicht darüber im klaren, daß dieser zuletzt genannte Grund von GRAWE u. a. tatsächlich ernst gemeint war!

Überlegungen über spezifische Stichprobenfehler angestellt werden müßten. Diese Angaben gehören zum gängigen metaanalytischen Methodeninventar.

Resümee

GRAWE u. a. erheben den Anspruch, daß Psychotherapie sich am aktuellen wissenschaftlichen Erkenntnisstand der Psychologie zu orientieren habe. Um so mehr kann von GRAWE u. a. erwartet werden, daß sie als PsychotherapieforscherInnen selbst diesem Anspruch genügen.

Wie unsere Ausführungen nahelegen, ist es GRAWE u. a. nicht gelungen, mit ihrer Untersuchung aktuellen forschungsmethodischen Standards zu genügen. Inwieweit diese Mängel die Ergebnisaussagen und weiteren Schlußfolgerungen von GRAWE u. a. verzerren, läßt sich anhand der vorliegenden Informationen nicht beurteilen. Daß solche Beeinträchtigungen nicht ausgeschlossen werden können, ist angesichts des immensen (zeitlichen, personellen und finanziellen) Forschungsaufwandes dieser umfassend und differenziert angelegten Untersuchung umso bedauerlicher.

Literatur

BEELMANN, A. & BLIESENER, T. (1994). Aktuelle Probleme und Strategien der Metaanalyse. *Psychologische Rundschau, 45*, 211-233.

DIEPGEN, R. (1993). Münchhausen-Statistik. Eine Randbemerkung zur Argumentationsfigur von GRAWE 1992. *Psychologische Rundschau, 44*, 176.

GRAWE, K. (1992). Psychotherapieforschung zu Beginn der neunziger Jahre. *Psychologische Rundschau, 3*, 132-162.

HUNTER, J. E., SCHMIDT, F. L. & JACKSON, G. B. (1982). *Meta-analysis. Cumulating research findings across studies*. Beverly Hills: Sage Publication.

SMITH, M. L., GLASS, G. V. & MILLER, T. I. (1980). *The benefits of psychotherapy.* Baltimore: John Hopkins University Press.

Anschrift der Verfasser:

Heinrich GELDSCHLÄGER & Bernd RUNDE
Westfälische Wilhelms - Universität Münster
Psychologisches Institut IV - Organisationspsychologie
Fliednerstraße 21
48149 Münster