Erwiderung auf die Kommentare von Suitbert Ertel und Ulrich Timm zur RNG-Metaanalyse von Bösch, Steinkamp & Boller (2006):

„Examining psychokinesis: The interaction of human intention with random number generators – A meta-analysis“

anlässlich des WGFP-Workshops 2006 in Offenburg

Emil Boller,

Freiburg, 2007

Die Kritiken von Timm und Ertel an der Metaanalyse von Bösch, Steinkamp und Boller (BSB) werden als unbegründet zurückgewiesen. Besonders Ertels Kritik ist genau genommen eine generelle Kritik an der üblichen Auswertung von RNG-Experimenten und natürlich auch der ersten RNG-Metaanalyse von Radin und Nelson von 1989. Entgegen der Intention von Ertel ist sie Wasser auf die Mühlen der Parapsychologiekritiker, weist sie doch richtig auf einige konzeptionelle und methodische Schwächen der RNG-Forschung hin. Allerdings steht das im Widerspruch zu seinem Versuch die RNG-Forschung als ein erfolgreiches Forschungsprojekt zu stilisieren. Ertel scheint die Tragweite seiner Kritik nicht bewusst zu sein. Wenn auch die BSB-Metaanalyse den Anlass für diese Kritik geliefert haben mag so ist sie nicht unbedingt der richtige Adressat. Ansonsten ist es natürlich zu begrüßen, wenn die RNG-Daten nicht nur ein Dornröschendasein fristen, sondern dazu genutzt werden, verschiedene Annahmen und Modelle zu Psychokinese aber auch Alternativhypothesen zu überprüfen. Indem BSB ihre Daten allen interessierten Forschern zur Verfügung stellen, wird dieser Prozess befördert.

Hintergrund

Die BSB-Metaanalyse (Bösch, Steinkamp & Boller, 2006) knüpft an die Metaanalyse von Radin und Nelson (1989, im folgenden Text auch RN-Metaanalyse genannt) an, was die grundsätzliche Methodik betrifft: Bestimmung von Gesamteffektstärken auf der Basis von nach Studiengröße (operationalisiert als Anzahl der Zufallsereignisse) gewichteten Effektstärken der Einzelstudien. Die BSB-Datenbasis unterscheidet sich in zwei wesentlichen Punkten von der RN-Metaanalyse:
Die Einschlusskriterien waren enger gefasst (z.B. keine Retro-PK, keine Experimente mit Pseudozufallsgeneratoren) und es wurden neue Studien aufgenommen. Dank der umfassenden Literatursuche konnten Studien aufgefunden werden, die, vor 1987 publiziert, nicht in der RN-Metaanalyse berücksichtigt wurden.

Eine weitere Neuerung bestand darin, Moderatoren (Variablen, die potentiell dazu geeignet sind Unterschiede in den Effektstärken verschiedener Studien zu erklären) zu kodieren und zu analysieren. Letzteres wurde Radin und Nelson übrigens schon bei der Präsentation ihrer Ergebnisse anlässlich einer Parapsychological Association(PA)-Konferenz nahe gelegt. Auf der gleichen PA war von Honorton und Ferrari eine Präkognitionsmetaanalyse (die 1989 im Journal of Parapsychology publiziert wurde) vorgestellt worden, die Moderatoren untersuchte. Allerdings griffen Radin und Nelson diese Anregung ganz offensichtlich nicht auf, wie ihre Publikationen von 1989 und 2003 belegen.

Außerdem wurden einige neue metaanalytische Methoden, wie z.B. die Metaregression, erstmals im Rahmen einer parapsychologischen Metaanalyse eingesetzt.

Holger Hartmann (geb. Bösch) und ich waren, als wir damit begannen, die Metaanalyse zu planen, Teil des Mind-Machine-Interaction-Konsortiums, dass in einer groß angelegten und auf drei Forschungseinrichtungen verteilte Studie (Jahn et al. 2000) versuchte die Ergebnisse der Zufallsgeneratorexperimente der Princeton Engineering Anomalies Research-Group (PEAR) zu replizieren. Für die Replikation wurde eine Effektstärke von etwa einer Abweichung auf 10000 Zufallsereignisse angenommen, ein Wert der zwei Größenordnungen kleiner ist als in Dean Radins Buch „The conscious universe“ publiziert. Diese Effektstärke weicht wiederum von der ab, die indirekt der RN-Metaanalyse zu entnehmen ist. Diese Diskrepanzen und die Frage nach den optimalen experimentellen Bedingungen für Nachfolgeexperimente waren der Ausgangspunkt für die Planung und Umsetzung der Metaanalyse.

Befremdlich ist, dass Radin diese gravierenden Diskrepanzen in seinen eigenen Publikationen nicht diskutiert, aber auch, dass innerhalb der parapsychologischen Community davon scheinbar keine Notiz genommen wurde. In parapsychologischen Publikationen ist oft zu lesen, dass die gefundenen Effekte zwar klein, aber hoch signifikant seien. Bei der Planung der notwendigen oder optimalen Stichprobenumfänge neuer Studien wird diesem Umstand jedoch nicht berücksichtigt. Dies hat zur Folge, dass die meisten parapsychologischen Experimente, legt man schon vorhandene Meta-analysen und Effektstärkeschätzungen zugrunde, eine zu geringe Power aufweisen, d.h. die Stich-probenumfänge (Trails) der meisten Experimente ist zu gering um auf deren Basis verlässliche Aussagen treffen zu können. Das fehlende Problembewusstsein bezüglich der Power von Studien und die Neigung die Ergebnisse von Metaanalysen gegenüber der Öffentlichkeit als Beweis für den postulierten Phänomene zu vermarkten statt die gewonnenen Ergebnisse als Grundlage für neue Studien zu betrachten könnte diese Indifferenz möglicherweise erklären.

Kommentar Timm

Globaler Signifikanztest

Timm legt den Schwerpunkt seiner Kritik auf den globalen Signifikanztest. Für eine Metaanalyse, die darauf zielt, Effektstärken zu schätzen und zu modellieren, steht dieser Gesichtspunkt nicht im Mittelpunkt. Nur wenn sich die Effektstärken der Einzelstudien (bezogen auf die Stichproben-theorie) homogen verteilen ist die Analyse mit einer Effektstärkeschätzung und einem globalen Signifikanztest abgeschlossen. Sind jedoch die Effektstärken der Einzelstudien heterogen verteilt, wie das bei den RNG-Studien extrem der Fall ist, stellt sich die Aufgabe, diese Heterogenität aufzuklären, in dem z.B. klar definierte Teilstichproben (Experimente) bestimmt werden, die in sich homogen sind, aber unterschiedliche Effektstärken aufweisen oder Moderatorvariablen identifiziert werden, die einen Beitrag zur Aufklärung der Variabilität leisten. In beiden Fällen ergeben sich komplexere Modelle, für die globale Effekt- und Signifikanzschätzungen nicht mehr adäquat erscheinen. Gelingt es nicht die Ergebnisse der Primärstudien befriedigend zu modellieren, wie das in der BSB-Metaanalyse der Fall ist, stellt sich die Frage, welche Bedeutung unter diesen Bedingungen ein globaler Signifikanztest bzw. die Schätzung einer globalen Effektstärke zukommt.

Gewichtung der Studien

Timm schreibt ganz zu Anfang „Die statistischen Berechnungen und Interpretationen ... bedürfen in mehreren Punkten eine Korrektur, in denen sie von unrealistischen Annahmen über die Struktur parapsychologischer Experimente ausgehen.“ Er bezieht sich dabei auf die Gewichtung der Experimente, die er auf eine andere Weise durchgeführt sehen will. Zur Motivierung führt er den intraexperimentellen Decline an. Zwar werden dieser und andere Declineeffekte in der parapsycho-logischen Gemeinschaft immer wieder diskutiert, aber offensichtlich spielen Declineeffecte in RNG-Experimenten kaum eine Rolle. Bei allen von uns zitierten Arbeiten (gut 370 Referenzen)

kommt das Stichwort nicht ein einziges Mal im Titel vor. PEAR hat in seinen Daten zwar einen Verringerung der Trefferrate von der ersten zur dritten Sitzung gefunden, die sich dann aber bis zur 5. Sitzung wieder auf das Ausgangsniveau erholt und in folgenden Sitzungen erhalten bleibt. Dies ist ein Befund, der die Annahme eines generellen intraexperimentellen (respektive intrapersonellen) Declines nicht unterstützt. Aber auch bei PEAR spielt der sogenannte „series position effect" bei der Auswertung des Gesamteffekts keine Rolle, er wurde als ein unabhängiger Befund behandelt. Zudem konnte der series position effect in einer groß angelegten Replikationsstudie (Jahn et al. 2000) nicht repliziert werden. Wie Timm richtig bemerkt, müsste, wenn der Effekt von den Experimentatoren ernst genommen wird, eine zumindest dem Konzept entsprechende differenzierte Auswertung stattfinden. Solche Auswertungen finden bei RNG-Experimenten nur selten statt und Declineeffekte wurden bei RNG-Experimenten so gut wie nicht systematisch untersucht. Als weiteres Argument führt er an, dass möglicherweise eine oder einige wenige extrem große Studien die Ergebnisse einer Metaanalyse dominieren können und um dies zu vermeiden, sei sein vorgeschlagenes Gewichtungsschema zu bevorzugen sei. Bei seinem Ansatz erhalten im Gegensatz dazu kleinere Studien ein überproportionales Gewicht, wobei kleinere Studien als anfälliger für methodische Fehler aller Art gelten. Bei ernsthaften Bedenken bezüglich der Qualität von Studien besteht in Metaanalysen die Möglichkeit, vorher Qualitätsmerkmale festzulegen, die, wenn sie nicht eingehalten werden, zum Ausschluss von Studien führen. Dies war jedoch bei der BSB-Metaanalyse nicht vorgesehen. Eine Kontrolle der Studienqualität erfolgte über Qualitätsratings. Das Problem mit den drei größten Studien ist übrigens erst durch den Reviewprozess zustande gekommen. Nachdem wir den Artikel zum ersten Mal eingereicht hatten, forderte ein Reviewer[1], dass wir auch neuere Studien (nach August 2000) berücksichtigen sollten. Dieser Forderung hat sich der Editor angeschlossen, was für uns einen enormen zusätzlichen Arbeitsaufwand mit sich brachte. Es mussten weitere Studien kodiert, die Daten neu ausgewertet und der Artikel im Prinzip neu geschrieben werden unter anderem auch wegen der drei größten Studien.

Timm relativiert seine Aussage bezüglich der Korrektur der Auswertungen und Interpretation dann jedoch wieder auf Seite 4 indem er darauf hinweist: „dass jede Gewichtung statistisch statthaft ist, sofern sie unabhängig von den aktuellen Daten a priori festgelegt wird." Dies sollte dann fairerweise auch BSB zugestanden werden. Indem die BSB-Metaanalyse als Ergänzung und Erweiterung der RN-Metaanalyse konzipiert wurde, war auch das Gewichtungsschema a priori festgelegt.
Die Gewichtung in der RN-Metaanalyse und auch der BSB-Metaanalyse hat mit der Stichprobentheorie eine theoretische Basis. Jede Gewichtung von Studien ohne Bezug zur Stichprobentheorie bedarf einer überzeugenden Begründung. Die ist für das von Timm vorgeschlagene Schema nicht zwingend. Um die Qualität von Studien zu berücksichtigen, gibt es anerkannte Methoden (unter anderem, indem man Studien zusätzlich nach ihrer Qualität gewichtet). Was die möglichen Decline-effekte betrifft, sollte erst einmal geklärt werden, wie diese in Primärstudien angemessen zu operationalisieren sind und dann müssen erst einmal entsprechende Primärstudien durchgeführt werden. In einem ersten Schritt könnten auch passende Primärstudien entsprechend reanalysiert werden. Es spricht auch nichts dagegen, zu untersuchen, inwieweit die bisherigen Studien Hinweise für oder gegen Declineeffekte liefern und gegebenenfalls eine spezifische Metaanalyse durchzuführen. Die von Timm vorgeschlagene Gewichtung ist dafür aber zu unspezifisch.

Timms Gewichtung führt zu einem signifikanteren Gesamtergebnis. Allerdings bleibt unklar, was signifikant hier bedeutet. Bei der von BSB gewählten Gewichtung ist es die Signifikanz der geschätzten Effektstärke. Da jedoch die Voraussetzungen für eine globalen Effektstärkeschätzung

1 Ich vermute, dass es der parapsychologische Reviewer war, da bezüglich vermeintlich fehlender Studien später nochmals Einwände kamen, wobei dann der Reviewer allerdings Probleme hatte, seine Behauptung zu belegen. In dem Zusammenhang soll auch nicht verschwiegen werden, dass BSB eine Studie entgangen ist, was insofern peinlich ist, weil diese sich in einer Publikation befand, von der BSB drei Studien kodiert hatten – die besagten größten Studien mit negativen Ergebnissen.

(besonders die extreme Heterogenität) und damit auch den globalen Signifikanztest, fraglich sind, sollte man auch den globalen Parameterschätzungen von BSB nicht zu viel Gewicht beimessen.

Interessant sind Timms Bemerkungen zur impliziten Gewichtung von Studien im Random Effects Modell (REM) und deren Relation zu anderen Gewichtungsschemata.

Filedrawerschätzung

Was nun die Zahl nicht publizierter Studien betrifft, ist sein Ansatz ebenfalls interessant, aber insgesamt ist dieses Thema bei einer Metaanalyse, die die Effekt der Einzelstudien nach den Grundsätzen der Stichprobentheorie gewichtet, nachrangig. Wie die drei größten Studien zeigen, können auch schon einige wenige Studien ein Ergebnis ändern. Anstatt die Anzahl nicht publizierter Studien bzw. der Anzahl Studien, die nicht signifikant sind, abzuschätzen, wäre es bei diesem Ansatz sinnvoller die Anzahl der nicht signifikanten Trials, als den Trialdrawer in Anlehnung an den Filedrawer, zu bestimmen.
Den small studies bias, der den Umstand beschreibt, dass kleine Studien höhere Effekte produzieren als größere Studien, betrachten wir als favorisierte Erklärung des Ergebnisses der BSB-Meta-analyse. Eine Effekt, der auch in der medizinischen Forschung beobachtet werden kann. Ein Effekt übrigens, der weit mehr Ursachen hat als nur den Filedrawer. In Tabelle 11, S. 514 der BSB-Metaanalyse wird ein Überblick über Faktoren gegeben, die einen small studies bias hervorrufen können. Die üblichen Filedrawerabschätzungen können als eine nicht unproblematische Projektion aller möglichen Fehler in Design, Durchführung, Auswertung, Interpretation und Publikation oder Nichtpublikation von Studien und nicht zuletzt auch Fehler der Metaanalysten[2] selbst auf eine Fehlerquelle betrachtet werden.

Kommentar Ertel

Bezüglich der folgender Bemerkung Ertels: „Doch waren die Autoren selbstkritisch genug, um dem zu entsprechen, was Greenhouse & IYengar (1994) Meta-Analytikern ans Herz legt (p. 384).“

> *"At every step in a research synthesis decisions are made that can affect the conclusions and inferences drawn from the analysis... In other words, it is important to check how sensitive the conclusions are to the method of analysis or to changes in the data"*

kann ich nur darauf hinweisen, dass BSB die Daten differenziert ausgewertet und die Ergebnisse ebenso differenziert dargestellt haben, so dass diese Forderung im hohen Maße erfüllt ist. Die Daten wurden nach zwei gängigen metaanalytischen Modellen (fixed effects modell, random effects modell), insgesamt, für Teilstichproben, einer Metaregression unter Berücksichtigung mehrere Moderatorvariablen (auch 2 Varianten) jeweils mit und ohne kritische Datensätze ausgewertet und dargestellt, um nur die wichtigsten Punkte zu nennen.
Was Ertels Vermutung bezüglich eines „anderen oder besseren Modells“ betrifft, so müssen erst noch zukünftige Analysen zeigen, ob das von Ertel vorgeschlagene Modell tatsächlich diese Erwartungen erfüllt. Die BSB-Metaanalyse wurde unter Annahme des Einflussmodells durchgeführt. Daneben gibt es andere Erklärungsansätze für RNG-Experimente, wie z.B. die DAT-Hypothese, auf die BSB in ihrer Metaanalyse kurz eingegangen sind, die aber nicht Gegenstand der BSB-Metaanalyse waren. Ob und inwieweit andere Erklärungsansätze möglicherweise zu zufriedenstellenderen Ergebnissen führen, müssen noch zu erbringende spezifische Metaanalysen zeigen.

2 Nach der Korrektur von **zwei** fehlerhaft kodierten Datensätzen, auf die Schub (2006) hinwies, reduziert sich die in Radins The Conscious Universe (1997) noch mit 0,509 angegebene Effektstärke um mehr als 20% auf nunmehr 0,507 (Radin, Nelson, Dobyns & Houtkooper, 2006).

Operationalisierung der Stichprobengröße

Der zentrale Einwand Ertels gegenüber der BSB-Metaanalyse („Grundfehler", „Fehlentscheidung") ist die von uns verwendete Operationalisierung der Stichprobenumfänge als Anzahl der Bits. Es scheint Ertel nicht klar zu sein, dass das die Maßeinheit ist, die den Auswertungen der Primärstudien zugrunde liegt und nicht eine willkürlich von uns gewählte Größe. Seine Kritik ist somit auch eine Kritik fast aller Primärstudien und kann genauso gegenüber der RN-Metaanalyse von 1989 geäußert werden. Allerdings ist mir bisher kein derartiger Einwand bekannt. Dass die Perspektive von Ertel nicht unbedingt die Sichtweise der Ersteller der Primärstudien wieder spiegelt kommt dadurch zum Ausdruck, dass von 380 Studien nur 140 die für seine Perspektive notwendigen Informationen zur Verfügung stellen.

Seine Darstellung des Versuchsumfangs beschreibt nicht die übliche Auswertung von RNG-Experimenten, da in seiner Darstellung die Anzahl der Bits (hier als kleinste experimentelle Einheit verstanden) fehlt. Seine Darstellung liefe auf die Auswertung von RNG-Experimenten mit einem t-Test hinaus (was auch im Einstichprobendesign möglich ist, wenn wie im Fall von RNG-Experimenten die Verteilungsparameter unter Annahme der Nullhypothese bekannt sind). In der Parapsychologie wird dieser Ansatz normalerweise nur beim Vergleich verschiedener Bedingungen innerhalb eines Experiments (Mehrstichprobendesign oder Messwiederholung) eingesetzt. Dieser Ansatz hat den Vorteil, die Unterschiede zwischen Versuchspersonen, d.h. die empirische Varianz in den Daten zu berücksichtigen. Im Einstichprobendesign, wie den RNG-Studien, ist die Verwendung des t-Tests es aber unüblich und möglicherweise sogar nachteilig, worauf Timm in seinem Kommentar hinweist, wenn er den REM-Ansatz kritisiert (Verwendung der empirischen Variabilität statt der theoretischen). Viele Studien haben zudem eine so geringe Versuchspersonenzahl (96 mit nur einem TN, weitere 107 mit 2 – 10 TN), dass der t-Test und selbst nichtparametrische Äquivalente nicht einsetzbar sind.

Bei statistischen Auswertungen, die stärker einzelne Messergebnisse und die Variabilität zwischen Messergebnissen berücksichtigen tritt ein Aspekt in den Vordergrund, der in der parapsychologischen Forschung allgemein zu wenig Beachtung findet, die Frage nach der Reliabilität der Messungen[3] (als Indikator für einen nicht direkt messbaren, also latenten Faktor). Was ja tatsächlich gemessen oder besser gezählt wird (das allerdings sehr reliabel), sind Häufigkeiten von bestimmten Zufallsereignissen, die erst im experimentellen Kontext als Indikator z.B. für PK interpretiert werden. Unter Annahme der Nullhypothese sollte die Reliabilität übrigens gleich Null sein. Wie viele Messwiederholungen (Trials) hierbei erforderlich sind, hängt von der Psileistung der einzelnen Personen ab. Je geringer die Leistung, desto mehr Trials sind erforderlich, um zu halbwegs zuverlässigen Schätzungen zu kommen. Die Standardauswertung von RNG-Experimenten (parapsychologischen Experimenten) ist unter den Bedingungen einer so gut wie nicht vorhandenen Reliabilität noch die optimalste Auswertungsstrategie. (Zu einer vertiefenden Diskussion dieser Problematik siehe Boller & Bösch 2000). Wenn also Timm und auch Ertel eine personenbezogene Auswertung bevorzugen, sollten sie sich im Klaren darüber sein, dass sie den Teufel mit dem Belzebub austreiben.

In diesem Zusammenhang möchte ich noch auf eine weitere Schwäche von RNG-Experimenten hinweisen. Betrachtet man PK-Experiment als eine Korrelation zwischen einer psychologischen Variablen, der Intention der Versuchsperson und dem Output eines Zufallsgenerators, dann muss man zur Kenntnis nehmen, dass die psychologische Variable, die Intention, überhaupt nicht gemessen, also kontrolliert wird. Sie wird durch den Versuchsleiter oder die Versuchsperson gesetzt und dann nicht weiter kontrolliert. Es wird also ungeprüft angenommen, dass die Versuchspersonen die Intention während eines Versuchsdurchgangs konstant halten. Wenn man, wie bei manchen

3 Was tatsächlich fehlt ist eine Reflektion über das von anderen Disziplinen unkritisch übernommene Messmodell, dass parapsychologischen Experimenten zugrunde liegt.

Psitheorien auch noch annimmt, dass unbewusste Motive eine Rolle spielen, wird die Geschichte noch problematischer.

Ertel fragt: „Warum überhaupt wollen Bösch et al. den Versuchsumfang operationalisieren, was hat er mit dem PK-Effekt zu tun?“ und gibt dann auch prompt die Antwort auf seine rhetorische Frage: „Die Autoren benötigen den Versuchsumfang, um zwischen *big studies* und *small studies* differenzieren zu können. An dieser Differenzierung sind sie interessiert, weil sie einen Beweis brauchen für ihre Annahme, dass ihr Datensatz durch einen Publikationsbias beeinträchtigt wurde.“ Nun, BSB wollten eine Metaanalyse durchführen und dazu braucht es nun mal eine Größe für die einzelnen Versuchsumfänge, wenn Effektstärken bestimmt werden sollen und nicht nur Z-Werte (wofür man übrigens auch wieder die Stichprobenumfänge braucht). Und da ist es nahe liegend die Operationalisierung zu wählen, wie sie weit überwiegend in den Primärstudien verwendet wird. Dass es dann große und kleine Studien gibt, ist nur selbstverständlich. Die extreme Bandbreite, wie sie bei RNG-Experimenten aufgefunden wird, dürfte aber einmalig sein. BSB zu unterstellen, sie hätten das getan, weil sie einen Beweis bräuchten, dass der Datensatz durch einen Publikationsbias beeinträchtigt sei, kann ich mir nur so erklären, dass Ertel wohl mit der Primärliteratur wenig vertraut ist. Allerdings muss ich mich auch fragen, warum Houtkooper und Timm Ertel auf diese Fehlinterpretation nicht aufmerksam gemacht haben und ihn darauf hingewiesen haben, dass er sich hier versteigt, da diese ja nach den Angaben Ertels seinen Text durchgesehen haben.
Später schreibt er noch einmal „dass N_B, die von Bösch et al. gewählte Größe, kein Indikator des Studienumfangs darstellt“. Diese Aussage ist schlicht und einfach falsch. Natürlich ist die von BSB gewählte Operationalisierung ein möglicher Indikator des Studienumfangs, und zwar eine, die in der RNG-Forschung allgemein akzeptiert ist und auch bei der RN-Metaanalyse so angewendet wurde. (Gibt es eine Kritik von Ertel an der RN-Metaanalyse?) Selbst bei Metaanalysen, die auf Z- oder p-Werten basieren, muss Z oder p erst bestimmt werden, und üblicherweise geschieht das in der Parapsychologie über die Anzahl der Einzelereignisse.

Bittempo und Bitmenge sind im hohen Maße korreliert. Eine nachvollziehbare Erklärung, weshalb durch die Verwendung der Bitfrequenz die Asymmetrie im Funnelplot (Abbildung 2), die bei Verwendung der Bitfrequenz statt der Bitmenge erhalten bleibt, nicht als ein Indikator für einen möglichen Publikationsbias interpretiert werden kann, bleibt Ertel schuldig. In der ASW-Forschung gibt es Hinweise, dass auch ohne Erhöhung der Bitfrequenz bei großen Studien die Effektstärke abnimmt. Auch in der medizinischen Forschung, wo es das Problem des Bittempos nicht gibt, ist der small studies bias bekannt. Die Rolle des Bittempos bei diesem Effekt ist somit noch offen.

Versuchsdauer als Operationalisierung des Versuchsumfangs

Ertel schlägt als alternative Operationalisierung des Versuchsumfangs die Versuchsdauer vor. Damit erreicht er zwar, dass Versuchsumfang und PK-Effektgröße nicht einmal andeutungsweise zusammenhängen. Der Zusammenhang wird damit aber nicht aus der Welt geschafft, sondern durch die Wahl seiner Operationalisierung nur erfolgreich auspartialisiert.
Seinem Anliegen gerechter würde aber wahrscheinlich ein Ansatz, bei dem auch der Effekt bezüglich der aufgewendeten Zeit normalisiert wird. Einen Versuch Effekte bezüglich der Zeit zu normalisieren hat Roger Nelson schon 1994 in einem technical report von PEAR dargestellt und auch auf Daten angewendet. Inzwischen (2006) ist diese Arbeit auch im JSE erschienen. Ziel war es, Ergebnisse von verschiedenen experimentellen Paradigmen (z.B. PK, Remote viewing) vergleichbar zu machen und dazu eine Normalisierung der Effekte zu finden, die die Variabilität der Ergebnisse reduziert. Eine Normalisierung nach Zeit (Z/Wurzel(Stunden)) war von 5 verschiedenen untersuchten Maßen am erfolgreichsten und wurde von Nelson als natürliche Effektstärke für parapsychologische Experimente erklärt. Falls solch ein Ansatz zu einer Reduktion der Variabilität in den RNG-Studienergebnissen führt und die so operationalisierte Effektstärke mit der Studiendauer korreliert, wäre tatsächlich etwas gewonnen. Dann wäre ein Modell gefunden, mit

dem die Daten halbwegs in Einklang zu bringen sind. Die klassisch gewonnene Effektstärke gegen die Studiendauer grafisch abzutragen, wie von Ertel praktiziert, ist da zu kurz gegriffen. Es gibt keine Annahme darüber, wie die Daten im Falle des Zutreffens der Nullhypothese oder der Alternativhypothese verteilt sein sollten. Wie die Abbildung 4 von Ertel zeigt, handelt es sich nicht um einen typischen Funnelplot, wo kleine Studien normalerweise am stärksten streuen und die großen Studien nur noch sehr wenig. Zudem streuen auch hier die Einzelergebnisse offensichtlich weiterhin stark (siehe Abbildung 4 bei Ertel).
Die Studienzeit ist als Operationalisierung der Studiengröße kaum geeignet die qualitativen Unterschiede in der Versuchsdurchführung zwischen kleinen und großen Studien abzubilden, wie das zugegebenermaßen auch bei der Anzahl der Einzelereignisse (Bits) der Fall ist. Die Anzahl der Einzelereignisse als Operationalisierung der Studiengröße hat jedoch den Vorteil, dass sie auf der Stichprobentheorie basiert und konform zum statistischen Standardauswertungsmodell ist, das üblicherweise der Auswertung von RNG-Experimenten zugrunde gelegt wird.
Ein wichtiges Ergebnis der BSB-Metaanalyse ist der Zusammenhang zwischen Studiengröße und PK-Effektstärke. Ein empirischer Fakt, der nicht einfach durch den Akt einer gegenteiligen Behauptung aus der Welt geschaffen werden kann. Diesen Zusammenhang unter Berücksichtigung aller relevanten Faktoren[4] zu untersuchen, wird zukünftigen Arbeiten vorbehalten bleiben. Dazu ist eine eigens konzipierte Metaanalyse mit spezifischeren Modellannahmen erforderlich. Außerdem wird es notwendig sein die Primärstudien erneut durchzuarbeiten.

Ertels Versuche den small studies effect wegzudiskutieren werden nicht der Tatsache gerecht, dass wir die RNG-Studien unter Annahme des Einflussmodells einer Metaanalyse unterzogen. Unter dieser Perspektive kann die Feststellung eines small studies effect ohne Abstriche aufrecht erhalten werden.

Selbstkritisch kann ich zum Thema Filedrawer anmerken, dass einer sehr umfänglich dargestellten Teilstudie (als Teil der Replikation des MMI-Konsortiums, Jahn et al. 2000) eine weitere Studie (Boller & Bösch, 2000) gegenüber steht, die bisher unzureichend publiziert ist, und weitere Daten, die wir als Interim bezeichnet haben, gar nicht publiziert sind. Außerdem existiert noch eine Kleinstudie mit 3 Experimenten (wovon eines signifikant war), die ebenfalls nicht publiziert ist.

Publikationsstandards in parapsychologischen Journals

Bezüglich der *Ethical and Professional Standards of the Parapsychological Association* (P. A.) von 1986 ist nur anzumerken, dass die Mehrzahl der RNG-Studien vor diesem Zeitpunkt publiziert worden sind. Allerdings wiesen BSG in ihrer Arbeit darauf hin, dass nach Broughton (1987) schon 1975 das Council of the Parapsychological Association die Unterdrückung nichtsignifikanter Studien in parapsychologischen Journals verwarf. Und tatsächlich hat die Anzahl signifikanter Studien dramatisch von 47% vor 1975 auf 17% und weniger in den Folgejahren abgenommen[5].
Sofern dafür keine andere Erklärung gefunden wird, unterstützt dieser Befund die Annahme, dass besonders in den Anfangsjahren nichtsignifikante Experimente öfters nicht publiziert wurden.
Anbei noch ein längeres Zitat von Langmuir zu Rhines Publikationsgebahren, dass im Rahmen eines Colloquiums at The Knolls Research Laboratory (Pathological Science), Dezember 18, 1953

4 Stichworte hierzu sind die technische Geschwindigkeit, in der Zufallsereignisse produziert werden, Geschwindigkeit, mit der die VPN Zufallsereignisse und/oder Feedback erhält, Trial by Trial Feedback oder summarisches Feedback (z.B. für Trials von 200 Bit), Geschwindigkeit des Feedbacks. Um auf dieser Basis Aussagen treffen zu können, müssen die einzelnen Merkmalskombinationen eine ausreichende Anzahl von Studien vorliegen. In der ASW-Forschung gibt es Hinweise darauf, dass bei großen Studien die Effektstärke auch ohne gravierenden methodischen Unterschieden zwischen großen und kleinen Studien abnimmt.

5 Ein Umstand, der auch Krippner et al. (1993) nicht entgangen ist: „It was also noted that more nonsignificant studies were published after the PA-affiliated journals, in 1975, adopted an official policy of publishing nonsignificant results. This suggests that some earlier nonsignificant studies were never published; hence the necessity of statistical adjustments for possible unpublished studies. “ Zwar bezieht sich diese Aussage auf Würfelexperimente, wie die Ergebnisse von BSB zeigen, kann sie auch auf RNG-Experimente übertragen werden.

zustande kam. Es betrifft zwar Rhines ASW-Forschung, ist aber dennoch aufschlussreich:

> I'll go first, before I get into what Rhine said, and say this: David Langmuir, a nephew of mine, who was in the Atomic Energy Commission, when he was with the Radio Corporation of America a few years ago, he and a group of other young men thought they would like to check up Rhine's work so they got some cards and they spent many evenings together finding how these cards turned up and they got well above 5. They began to get quite excited about it and they kept on, and they kept on, and they were right on the point of writing Rhine about the thing. And they kept on a little longer and things began to fall off, and fall off a little more, and they fell off a little more. And after many, many, many days, they fell down to an average of five--grand average--so they didn't write to Rhine. Now if Rhine had received that information, that this reputable body of men had gone ahead and gotten a value of 8 or 9 or 10 after so many trials, why he would have put it in his book How much of that sort of thing, when you are fed information of that sort by people who are interested--how are you going to weigh the things that are published in the book?
>
> Now an illustration of how it works is this. He told me that, "People don't like me," he said "I took a lot of cards and sealed them up in envelopes and I put a code number on the outside, and I didn't trust anybody to know that code. Nobody!"
>
> (A section of the speech is missing at this point. It evidently described some tests that gave scores below 5.) "... the idea of having this thing sealed up in the cards as though I didn't trust them, and therefore to spite me they made it purposely low."
>
> "Well," I said, "that's interesting--interesting a lot, because you said that you'd published a summary of all of the data that you had. And it comes out to be 7. It is now within your power to take a larger percentage including those cards that are sealed up in those envelopes which could bring the whole thing back down to five. Would you do that?
>
> "Of course not," he said. "That would be dishonest. "
>
> "Why would it be dishonest?"
>
> "The low scores are just as significant as the high ones, aren't they? They proved that there's something there just as much, and therefore it wouldn't be fair. "
>
> I said, "Are you going to count them, are you going to reverse the sign and count them, or count them as credits?"
>
> "No, No," he said.
>
> I said, "What have you done with them? Are they in your book?"
>
> "No."
>
> "Why, I thought you said that all your values were in your book. Why haven't you put those in?"
>
> "Well," he said, "I haven't had time to work them up."
>
> "Well, you know all the results, you told me the results. "
>
> "Well," he said, "I don't give the results out until I've had time to digest them."
>
> I said, "How many of these things have you?" He showed me filing cabinets--a whole row of them. Maybe hundreds of thousands of cards. He has a filing cabinet that contained nothing but these things that were done in sealed up envelopes. And they were the ones that gave the average of five.
>
> Well, we'll let it stand at that. A year or so later, he published a new volume of his book. In that, there's a chapter on the sealed up cards in the (p.10) envelopes and they all come up to around seven. And nothing is said about the fact that for a long time they came down below five. You see, he knows if they come below five, he knows that isn't fair to the public to misrepresent this thing by including those things that prove just as much a positive result as though they came above. It's just a trick of the mind that these people do to try to spite you and of course it wouldn't be fair to publish.

Rhines Bücher zu ASW haben trotz aller Kritik auch dazu geführt, dass andere Wissenschaftler versuchten die Experimente zu replizieren. Was dabei herauskam und welche Folgen das hatte, kann man sich denken. Englische Parapsychologen wunderten sich warum sie jenseits des Atlantiks nicht in der Lage waren die Ergebnisse von Rhine zu replizieren. Wenn tatsächlich negative Resultate zurückgehalten wurden, dann mussten Nachfolger zwangsläufig scheitern.

Die Nichtpublikation von empirischen Befunden ist offensichtlich weiterhin ein Problem, wie folgendes Zitat (Savva, 2006) zeigt: „I was also involved in the informal testing of a number of other parapsychological claims (e.g. telephone telepathy) none of which showed a significant paranormal effect and none of which were written up and communicated to a wider audience."

Selektion psi-begabter Testpersonen

Die Selektion psi-begabter Testpersonen ist ein bisher ungelöstes Problem, das von vielen Faktoren abhängt, unter anderem auch der Operationalisierung des Effekts und der schon angesprochenen Reliabilität als Indikator für einen Psi-Effekt verwendeten Messungen oder auch der postulierten Biddirektionalität von Psi. Klare Kriterien für die Selektion von Versuchspersonen für PK-Experimente gibt es nicht. Die Selektion von Vpn aufgrund ihrer Leistung setzt einen robusten Effekt voraus, der reliabel gemessen werden kann. Genau diese Voraussetzungen sind aber bei RNG-Experimenten nicht erfüllt. Wie Ertels Grafik auf S. 13 zeigt, streuen die Ergebnisse der Studien mit selektierten TN von Z = -5 bis Z = 10. Auch mit selektierten TN ist also der Ausgang eines RNG-Experiments nur bedingt vorhersagbar. Negative Z-Werte sind ein Indikator dafür, dass auch Experimente mit einer oder mehreren selektierten Vpn zu Ergebnissen führen, die nicht mit dem Hypothesen bzw. den Erwartungen der Experimentatoren übereinstimmen. Auch mit selektierten Vpn wird die globale Effektstärke- und Signifikanzschätzung durch die extreme Variabilität der Ergebnisse relativiert. Wie Ertel bei dieser Datenlage zur Einschätzung kommt, „dass die Stabilität der Effekte bei Psi-Begabten für die Existenz und Echtheit der Effekte spricht" und von BSB verlangt, dies zuzugeben, kann ich nur schwer nachvollziehen. Die Effektstärken der Primärstudien verhalten sich anders, als es die mittlere Effekstärke in Ertels Abbildung 6 suggeriert. Modell und Daten stimmen schlecht überein. Da es sich bei den Studien mit selektierten TN überwiegend um kleine Studien handelt, die zudem überwiegend aus der Anfangszeit der RNG-Forschung stammen, ist es nahe liegend einen small studies bias anzunehmen. Nur indem Ertel diese extreme Variabilität ausblendet kann er zur von uns abweichenden Einschätzung kommen.

Weiter ist nicht auszuschließen, dass die Selektion von Vpn wirksam ist, aber nicht weil sie über Psi verfügen sondern andere Fähigkeiten wirksam werden, die bisher zu wenig berücksichtigt wurden. Palmers Reanalyse (1996, 1997) von zwei Schmidtexperimenten, eines davon ist auch in BSB-Metaanalyse eingegangen, kann als Hinweis für solche Effekte interpretiert werden. Ein drastisches Beispiel liefert Radin (2002). Bei einem Internetexperiment-ASW-Experiment waren 2 Versuchspersonen mit sehr vielen Trials und sehr signifikanten Ergebnissen aufgefallen. Eine detaillierte Analyse ergab dann, dass der Pseudozufallsgenerator ungünstig initialisiert wurde und es zu lokalen Abhängigkeiten in der generierten Zufallsfolgen kam. Nachdem dies behoben worden war, verschwand ein zuvor hoch signifikanter Effekt bei den nachfolgenden Durchgängen sowohl bei einer der beiden besonders aktiven Vpn wie auch den übrigen TN.

Zusammenfassung

Die pauschale Kritik, die von Timm und Ertel gegenüber der BSB-Metaanalyse geäußert wird, ist auch auf die RN-Metaanalyse anwendbar. Und nicht nur auf diese sondern sie trifft auch auf die überwiegende Zahl der Primärstudien zu, für deren Ausgestaltung Metaanalytiker nicht verantwortlich gemacht werden können. Wenn fast alle REG-Studien etwas gemeinsam haben, dann die Auswertung auf der Basis des Gesamtzahl der beeinflussten Zufallsereignisse. Nur wenige Studien verfolgen andere Ziele, wie z. B. die Erhöhung der run score variance, die dann auch für die BSB-Untersuchung nicht geeignet waren.

Die Ergebnisse und besonders die Interpretation der Ergebnisse der BSB-Metaanalyse mag den Erwartungen mancher Parapsychologen widersprechen, aber die Ergebnisse sind nun mal trotz Interpretationsspielraumes so, wie sie sind und, was die Gesamteffektstärke- und

Gesamtsignifikanzschätzung betrifft nicht unähnlich den Ergebnissen der RN-Metaanalyse.

Zum Nachwort von Ertels Beitrag

Was nun Ertels strategischen Überlegungen betrifft und den Schaden, den wir der Parapsychologie zugefügt haben sollen, bin ich völlig anderer Ansicht. Kurzfristig mag so eine Strategie funktionieren, auf lange Sicht führt sie meiner Ansicht zu größerem Schaden, indem sie das Vertrauen in die Parapsychologen als Wissenschaftler untergräbt. Die Diskrepanzen in den REG-Metaanalysen von Radin & Nelson 1989 und dem was Radin und Nelson später publizierten sind zu offensichtlich. Es war nur eine Frage der Zeit, bis das jemandem auffallen musste. Tatsächlich wurden Steinkamp und ich schon vor länger Zeit von jemandem kontaktiert, der auf Fehler in der RN- und den Folgemetaanalysen aufmerksam geworden war, sich von Radin die Daten hatte geben lassen und dabei war dazu einen Artikel zu schreiben. Ich hatte mich schon gefragt, was aus diesem Projekt geworden ist, nachdem ich lange nichts mehr davon gehört hatte. Aus Ertels Beitrag konnte ich nun entnehmen, dass der Artikel inzwischen im JSE publiziert wurde (Schub, 2006), begleitet von weiteren Artikeln zur Thematik. Interessanterweise läuft Schubs Beitrag unter Commentaries, d.h. sein Name erscheint (außer den Kopfzeilen) erst am Ende des Artikels, während die Replik darauf von Radin et al. (2006) als regulärer Artikel abgedruckt ist. Schubs Zusammenfassung seiner Ergebnisse kann nur als ernüchternd bezeichnet werden:

> "No statistical significance is quoted in the often-cited 1989 meta-analysis by Radin and Nelson of anomalous direct mental interaction with random number generators. Authors citing it have quoted z-scores ranging from 4.1 to 15.76. The combined statistical significance turns out to be acutely sensitive to the method used to combine the data, and there is at least one method which gives a non-significant result. The sensitivity is due to small studies with large hit rates, which in turn is at least partly due to publication bias, for which there is both direct and indirect evidence. Publication bias can account for only part of the long tails of the z distribution, but the remainder could well be the results of methodological problems, since, with the possible exception of the Princeton Engineering Anomalies Research data, overall methodological quality is quite poor. Methodological quality turns out to be confounded with publication bias, so the influence of poor quality on the results cannot be ruled out. Given the problems with existing data in this field, convincing evidence for a real effect can only be provided by new experiments free of reporting biases and methodological issues."

Weniger beachtet von der interessierten Öffentlichkeit erschien schon zuvor eine weitere Metaanalyse (Ehm, 2005), die auf den Daten von Radin und Nelson (1989) basiert. Diese Metaanalyse ist für die Parapsychologie wahrscheinlich weniger problematisch, weil sie in Mind and Matter erschien. Das Journal existiert erst seit 2003 und dürfte bisher von der parapsychologischen Community kaum wahrgenommen worden sein. Zudem ist der Artikel methodisch sehr anspruchsvoll. Leider ist der Artikel nicht, wie viele andere Artikel von Mind und Matter als PDF-Dokument im Internet zugänglich. Mit einem Ansatz, der die Modellierung der Daten (mittels Simulationen) in den Vordergrund stellt, kommt Ehm aber zu einem mit uns vergleichbaren Ergebnis: „Adjustment for possible selection is found to render the, without such an adjustment significant, experimental effect non-significant." Nach einer persönlichen Mitteilung musste Ehm darum kämpfen, dass der Artikel in Mind und Matter publiziert wurde. Das Editorial von Atmanspacher (2005) zeigt die Schwierigkeiten, die der *Editor-in-Chief* mit dem Artikel hatte:

> „It is important to consider random effects of experiments whenever systems show variability in addition to variance, i.e. results are scattered not only due to measurement error but also due to genuine differences in parameters. Such situations are generic in biology and psychology. With respect to selection effects, it should be pointed out that the well-known topic of publication bias, i.e. publication of significant results only, is only one among many options for selection effects. Ehm addresses selection in a general sense which ultimately includes the selection of model parameters or modeling strategies as well. As an important general result he finds that even an insignificant selection effect can influence the overall significance of an experimental effect seriously.
> From the statistical modeling perspective of Ehm's work, the evidence for anomalous mind-matter correlations found by Radin and Nelson appears less significant. However, other results may be possible

> if other modeling strategies consistent with the data are applied. Moreover, statistical analyses always remain inconclusive in a very basic manner as long as a theoretical understanding of statistically evident effects (if there are any) is missing. The key question that mind-matter experiments of the mentioned kind have been intended to resolve remains open: Are there relations between mind and matter which go beyond the established corpus of scientific knowledge?

Ausblick

Wenn man die BSB-Metaanalyse auf den Punkt bringen will, dann war es der Versuch die REG-Experimente unter Annahme des Einflussmodells einer Metaanalyse zu unterziehen. Die Modellierung der Daten unter diesem Gesichtspunkt hat, wenn alle Studien berücksichtigt werden, zu keinem überzeugenden Ergebnis im Sinne der Bestätigung dieser Annahme geführt. Hauptproblem ist die extrem heterogene Verteilung der Effektstärken, die auch durch die Analyse von Moderatorvariablen nicht aufgeklärt werden konnte. Diese heterogene Verteilung findet sich auch bei Teilstichproben, wie z.B. für Studien mit selektierten Teilnehmern. PK ist damit sicher nicht wahrscheinlicher geworden. Es wäre aber ein Missverständnis anzunehmen die BSB-Metaanalyse sei das letzte Wort in dieser Geschichte. Neben dem Einflussmodell gibt es weitere Modelle, wie z.B. die DAT-Hypothese, oder auch eine zeitbasierte Effektstärkeschätzung, die noch in spezifischen Metaanalysen untersucht werden können. Bestimmte Paradigmen, wie Fieldreg-Studien oder RetroPK, sind noch nicht eigenständig einer Metaanalyse unterzogen worden.
Metaanalysen sind mit einer größeren Anzahl von Entscheidungen verbunden. Unter anderem müssen Einschluß- und Ausschlußkriterien für Studien Auswahl der Studien und die zu untersuchenden Parameter bestimmt werden, aber auch die zugrunde gelegten inhaltlichen und statistischen Modelle. Timm und Ertel liefern Beispiele für alternative Entscheidungen und damit auch alternative Auswertungen. Wie auch immer das Ergebnis solcher Analysen ausfällt, meine Erwartungen sind da gering, müssen positive Befunde letztlich durch neue, hypothesenprüfende Experimente bestätigt werden. Das ist auch das eigentliche Ziel von Metaanalysen, Hinweise für zielgerichtetere zukünftige Primärstudien zu geben. In dieser Hinsicht war die BSB-Metaanalyse nicht besonders erfolgreich, was aber weniger an der Methodik den an den Daten selbst liegt. Vielleicht sind Nachfolgearbeiten da erfolgreicher, es sei ihnen gewünscht.
Wenn die BSB-Metaanalyse, wie auch die Untersuchungen von Schub oder Ehm, bewirken, dass noch einmal gründlich über RNG-Experimente nachgedacht wird, dann waren diese Untersuchungen erfolgreich. Tatsächlich hatten bisher in der Parapsychologie Metaanalysen mit für die Parapsychologie unvorteilhaften Ergebnissen einen größeren Auswirkungen als Metaanalysen, die zu einem positiven Ergebnis gelangten. Bei diesem Nachdenken sollten aber auch die experimentellen Designs und die ihnen zugrunde liegenden Annahmen nicht ausgeblendet werden. Mit der fehlenden Reliabilität von Psimaßen und der fehlenden Messung der Intention habe ich bisher nur zwei Punkte angesprochen. Ein weiterer wichtiger Punkt ist die Funktion und Rolle des Feedbacks, das psychologisch gesehen eigentlich als Distraktor fungiert. Nicht zuletzt gehören all die Zusatzannahmen, denen vielleicht nicht zu unrecht unterstellt wird, sie dienten dazu die Psihypothese gegen Falsifizierung zu immunisieren, wie z.B. die diversen Declineeffekte, die immer wieder postulierte Abhängigkeit des Effekts von psychologischen Variablen der TN, Versuchsleitereffekte und die Psimissingsannahme bzw. Bidirektionalität von Psi, um nur einige zu nennen, gründlich evaluiert.
Aus metaanalytischer Perspektive ist es wünschenswert, wenn für RNG-Experimente (und nicht nur für diese) ein Register für neue Studien eingerichtet wird, ähnlich wie in der Medizin, wo alle wichtigen Parameter vor Studienbeginn zentral erfasst und öffentlich zugänglich gemacht werden. Zusätzlich sollte es Pflicht werden, dass Laborbücher geführt werden, in denen alle Arbeitsschritte einer Studie gründlich dokumentiert werden. Die Daten sollten so, wie sie anfallen, d.h. auf der untersten Ebene und so vollständig wie möglich erfasst, gespeichert und später öffentlich zugänglich gemacht werden. Nur so ist es potentiell möglich zur Überprüfung von Hypothesen auf eine breitere Datenbasis zurückgreifen zu können, bevor sehr aufwendig neue Studien initialisiert

werden. Ich möchte dies anhand der PEAR-Daten kurz erläutern. Der Zufallsgenerator liefert einzelne Bits. Diese werden von einer Software verarbeitet. Ein Teil der Daten, die nicht gebraucht wird, geht völlig verloren. Für die restlichen wird für z.B. 200 einzelne Bits nur die Summe der Einsen gespeichert. Die zugrunde liegende Bitabfolge geht verloren. In Publikationen werden die Daten noch stärker aggregiert präsentiert. Für viele spezielle Fragestellungen sind die publizierten Daten nicht oder nur bedingt brauchbar. Werden die Daten dagegen vollständig dokumentiert, wäre es manchmal möglich, empirische Fragen zu überprüfen, ohne gleich neu Daten erheben zu müssen. Die Aussichten für solche Standards stehen allerdings schlecht. Abgesehen davon, dass ein entsprechendes Problembewusstsein noch nicht entwickelt ist, fehlt es auch an einer Einrichtung, die über die notwendig Autorität und Mittel verfügt und dazu noch Willens ist um solche Standards zu etablieren.

Literatur:

Atmanspacher, H. (2005). Editorial. Mind and Matter, 3(1), 5-7. (http://www.mindmatter.de/mmpdf/editorial3_1.pdf)

Bösch, H., Steinkamp, F. & Boller, E. (2006a). Examining psychokinesis: The interaction of human intention with random number generators - A meta-analysis. Psychological Bulletin. 132, 497-523. (Eine leicht von PB-Version abweichende online zugängliche Version: http://www.ebo.de/publikationen/pk_ma.pdf)

Boller, E., & Bösch, H. (2000). Reliability and correlations of PK performance in a multivariate experiment. In The Parapsychological Association 43rd Annual Convention: Proceedings of presented papers (pp. 380–382). Durham, NC: Parapsychological Association.

Broughton, R. S. (1987). Publication policy and the Journal of Parapsychology. Journal of Parapsychology, 51, 21–32.

Ehm W. (2005). Meta-analysis in mind-matter experiments: a statistical modeling perspective. Mind and Matter 3(1), 85–132.

Ertel, S. (2007). Kritischer Kommentar zu einer Meta-Analyse von Bösch, Steinkamp & Boller: „*Examining psychokinesis: The interaction of human intention with random number generators - A meta-analysis" Psychological Bulletin 2006, 132, 497-523.* Göttingen (http://www.anomalistik.de/sdm_pdfs/ertel-boesch-kritik.pdf)

Honorton, C. und D.C. Ferrari (1989) Future Telling: A Meta-Analysis of Forced-Choice Precognition Experiments, 1935-1987. Journal of Parapsychology 53, 281-308.

Jahn, R. G., Mischo, J., Vaitl, D., Dunne, B. J., Bradish, G. J., Dobyns, Y. H., Boller, E., Bösch, H. Houtkooper, J., & Walter, B. (2000). Mind/machine interaction consortium: PortREG replication experiments. Journal of Scientific Exploration, 14, 499–555. (http://www.princeton.edu/~pear/pdfs/jse_papers/portREG.pdf)

Langmuir, I. (1968) Langmuir's talk on Pathological Science (Colloquium at The Knolls Research Laboratory, December 18, 1953). Transcribed and edited by R. N. Hall (http://www.cs.princeton.edu/~ken/Langmuir/langmuir.htm)

Nelson, R. D. (1994). Effect Size per Hour: A Natural Unit for Interpreting Anomalies Experiments. Technical Note PEAR 94003. Princeton Engineering Anomalies Research, Princeton

University, Princeton, NJ 08544

Nelson, R. D. (2006). Time-normalized yield: A natural unit for effect size in anomalies experiements. Journal of Scientific Exploration, 20, 177-199.

Palmer, J. (1997). Hit-contingent response biases in Helmut Schmidt's automated precognition experiments. Journal of Parapsychology, 61, 135-141. (http://www.findarticles.com/p/articles/mi_m2320/is_n2_v61/ai_20576516)

Palmer, J. (1996). Evaluation of a conventional interpretation of Helmut Schmidt's automated precognition experiments. Journal of Parapsychology, 60, 149-170. (http://www.findarticles.com/p/articles/mi_m2320/is_n2_v60/ai_18960811)

Radin, D. I., (2003). Preliminary analysis of a suite of informal web-based psi experiments. Boundary Institute and Institute of Noetic Sciences. (http://www.boundary.org/articles/GotPsi-public.pdf)

Radin, D. I. (1997). *The conscious universe.* San Francisco: Harper Edge.

Radin, D. I., & Nelson, R. D. (1989). Evidence for consciousness-related anomalies in random physical systems. *Foundations of Physics, 19,* 1499–1514.

Radin, D., Nelson, R., Dobyns, Y. & Houtkooper, J. (2006). Assessing the evidence for mind-matter interaction effects. Journal of Scientific Exploration, 20, 361-374.

Savva, Louie. (2006, Nov. 11) Wy I quit parapsychology. http://www.everythingispointless.com/2006/11/why-i-quit-studying-parapsychology.html

Schub, M.O. (2006). A critique of the parapsychological random number generator meta-analysis of Radin and Nelson. Journal of Scientific Exploration, 20, 402-419

Stanley Krippner, William Braud, Irvin L. Child, John Palmer, K. Ramakrishna Rao, Marilyn Schlitz, Rhea A. White, Jessica Utts. (1993). Demonstration research and meta-analysis in parapsycholog. Journal of Parapsycholog. (http://www.findarticles.com/p/articles/mi_m2320/is_n3_v57/ai_15383545/pg_1)

Timm, U. (2007) Kommentar zu Bösch, H., Steinkamp, F. & Boller, E: "Examining psychokinesis .....-a meta-analysis." Psychological Bulletin, Vol.132, 2006. Nach einem Kurzvortrag vom 4. 11.06 auf dem XXII WGFP-Workshop in Offenburg, Freiburg (http://www.anomalistik.de/sdm_pdfs/timm-boesch-kritik.pdf)