Dr.Ulrich Timm **Freiburg**

Kommentar zu Bösch, H., Steinkamp, F. & Boller, E: "Examining psychokinesis .....-a meta-analysis." *Psychological Bulletin, Vol.132,* 2006

Nach einem Kurzvortrag am 4.11.06 auf dem XXII. WGFP-Workshop in Offenburg

Thema: Modifizierte Berechnung des globalen Signifikanztests und seine Korrektur für den Publikationsbias

Die statistischen Berechnungen und Interpretationen in dem Artikel von Bösch et al.(2006) bedürfen in mehreren Punkten einer Korrektur, in denen sie von unrealistischen Annahmen über die Struktur parapsychologischer Experimente ausgehen. Im folgenden beschränkt sich die Diskussion auf den besonders wichtigen globalen Signifikanztest und seine Korrektur für einen etwaigen Publikationsbias.

1. Optimale Gewichtung der zu aggregierenden Resultate in Metaanalysen

In Metaanalysen wird meist ein gewöhnlicher Z-Test benutzt, um zu prüfen, ob überhaupt ein Gesamteffekt nachweisbar ist. Dieser Test berechnet zunächst die Abweichung der (aus den k einzelnen ES_i abgeleiteten) mittleren Effektgröße ES* von ihrem H_0-Erwartungswert ES_0. Diese Differenz wird dann durch den zugehörigen Standardfehler $\sqrt{V_{ES*}}$ dividiert (der aus den entsprechend gemittelten einzelnen Fehlervarianzen V_{ESi} hervorgeht).

(1) $$\mathbf{Z_\Sigma} = (ES^* - ES_0) / \sqrt{V_{ES*}}$$

Das Resultat des Z-Tests hängt entscheidend von der Gewichtung w_i ab, mit der die einzelnen ES_i-Werte in den gewogenen Mittelwert ES* eingehen:

(2) $$\mathbf{ES^*} = \Sigma(w_i\, ES_i) / \Sigma w_i$$

Als statistisch optimales Gewicht w_i einer individuellen ES_i gilt der Kehrwert der Standardfehlervarianz V_{ESi}, da dadurch die globale Fehlervarianz V_{ES*} in Formel (1) minimiert und Z_Σ maximiert wird:

(3) $$\mathbf{w_i} = 1 / V_{ESi}$$

Die Variation der Gewichte w_i wird bei Psi-Experimenten primär von den Stichprobenumfängen n_i bestimmt, zu denen sich die w_i meist proportional verhalten. Hinzu kommt der Einfluss etwaiger variierender Trefferwahrscheinlichkeiten p_o, der wiederum vom benutzten ES-Index abhängt. Weitere Unterschiede der intraexperimentellen Varianz spielen dagegen (anders als sonst) keine Rolle, sofern das standardmäßige Einstichprobendesign benutzt wird, das mit einer konstanten (aus der Binomialverteilung abgeleiteten) theoretischen Varianz arbeitet. Solange p_o

sich nicht ändert oder wenn für ES_i der besonders verbreitete Index $r_i=Z_i/\sqrt{n_i}$ gewählt wird, kann deshalb (2) folgendermaßen vereinfacht werden:

(4) **ES*** = $\Sigma(n_i\ ES_i) / \Sigma\ n_i$

2. Beeinflussung der Effektgröße von Psi-Experimenten durch den Decline-Effekt

Diese Gewichtung ist jedoch nur dann statistisch optimal, wenn die ES von der Trialzahl **n** (die im Folgenden der Bit-Zahl gleichgesetzt wird) unabhängig ist. Genau das ist aber bei Psi-Experimenten nach gängiger Ansicht nicht der Fall, u.a. deshalb, weil dort diverse Decline-Effekte berichtet werden, die statistisch nicht besser oder schlechter gesichert sind als der Psi-Effekt selbst. Von diesen wird im Folgenden nur der intraexperimentelle Decline behandelt. Zu seinem Verständnis muss zunächst einmal **n** als Produkt der Vpn-Zahl **N** und der Trialzahl pro Vp **m** dargestellt werden. - Ein intraexperimenteller Decline lässt sich dann so definieren, dass entweder die momentane ES der Vpn während des Versuchsablaufs absinkt (=personenbezogener Decline) oder (seltener), dass die ES zeitlich aufeinanderfolgender (und mit verschiedenen Vpn durchgeführter) experimenteller Abschnitte immer niedriger wird (=experimentbezogener Decline). Es handelt sich also primär um einen temporalen Effekt, der nur durch eine differenzierte intraexperimentelle Auswertung feststellbar ist.

Trotzdem kann der Decline-Effekt indirekt auch aus den Resultaten einer Metaanalyse ablesbar sein, nämlich dann, wenn dort die ES der einzelnen Experimente mit der Trialzahl m (oder sogar der Vpn-Zahl N) negativ korreliert. Das liegt daran, dass bei intraexperimentell absinkender ES die mittlere ES eines Experiments ebenfalls absinken muss. Allerdings ist diese metaanalytische Diagnose nicht eindeutig, da die negative Korrelation zwischen ES und m (oder N) auch auf stationären Versuchsbedingungen beruhen könnte, die mit zunehmender Trial- (oder Vpn-)Zahl ungünstiger werden. Bei PK-Experimenten mit RNGs kann dies z.B. eine Verkürzung der Trialdauer **t** sein, die zwar (bei unveränderter Versuchszeit) zu einem oft astronomischen Anstieg von m führt, zugleich aber leistungshemmend wirken könnte. (Aus den Daten der hier diskutierten Metaanalyse ergibt sich in der Tat eine – über die Experimente als Einheiten berechnete – Korrelation r[m,ES]= -.13, die auf r[m,t]= -.66 und r[t,ES]= .20 zurückgeht; sie kann durch Auspartialisierung des Einflusses von **t** in frappierender Weise auf null gebracht werden, was **Ertel** in seinem Beitrag ausführlich behandelt.) Da ein solcher Scheindecline datenmäßig oft nicht von einem echten Decline unterscheidbar ist, wird er im folgenden als struktureller Decline mit dem temporalen Decline zusammengefasst, so dass die operationale Definition des Decline als Korrelation der ES mit m oder N volle Gültigkeit erhält. - Nach dieser Ausweitung des Begriffs ist die Forderung noch begründeter, ein empirisch so auffälliges Faktum wie den Decline bei der metaanalytischen Datengewichtung zu berücksichtigen.

3. Modifizierte Gewichtung zur Kompensation des Decline-Effekts

Da beim so definierten Decline-Effekt die ES der Experimente als monotone Funktion von m (und eventuell N) abnimmt, kann dieses Defizit nur durch eine Gewichtung ausgeglichen werden, die (über ihre normale Funktion hinaus) zusätzlich mit m (und eventuell N) abnimmt. Die normale Gewichtung nimmt aber nach (4) proportional zu n=m•N zu, so dass diese Zunahme verringert werden muss. Als einfachste Verringerung bietet sich dann eine reduzierte Zunahme mit der Wurzel aus m (und eventuell N) an. Je nachdem, ob man nur den n-bezogenen oder auch einen N-bezogenen Decline berücksichtigen will, erhält man so zwei verschiedene Alternativen zur w_i-Definition in (3). Von diesen ist die (weniger veränderte) erste vorzuziehen, während sich die zweite durch verblüffende Einfachheit auszeichnet:

(5) $\mathbf{w'_i} = w_i / \sqrt{m_i} = w_i \sqrt{N_i} / \sqrt{n_i}$

(6) $\mathbf{w''_i} = w_i / \sqrt{n_i}$

Die gleichen Voraussetzungen, die oben von (3) zu (4) geführt haben, führen auch hier zu zwei vereinfachten Formeln für ES*:

(7) $\mathbf{ES*'} = \Sigma(\sqrt{m_i}\, N_i\, ESi) / \Sigma(\sqrt{m_i}\, N_i) = \Sigma(\sqrt{n_i}\, \sqrt{N_i}\, ESi) / \Sigma(\sqrt{n_i}\, \sqrt{N_i})$

(8) $\mathbf{ES*''} = \Sigma(\sqrt{n_i}\, ES_i) / \Sigma(\sqrt{n_i})$

Wenn ES_i über den Index r_i berechnet wurde, kann man übrigens ES_i durch $Z_i/\sqrt{n_i}$ ersetzen und so ES* nach (4), (7), (8) als Linearkombination der Z_i-Werte darstellen. Bei (8) ist das besonders interessant, weil hier eine ungewogene Summe der Z_i entsteht, die einer gewogenen Summe der ES_i entspricht. Dadurch wird auch für den Laien evident, dass in $\Sigma(Z_i)$ eine implizite Gewichtung enthalten ist. Da die gleiche Summe in „Stouffer's Z" vorkommt, muss auch in diesem eine Gewichtung enthalten sein, die etwa in der Mitte zwischen einer Gewichtung der ES_i mit n_i und ihrer Nichtgewichtung liegt. Zugleich wird klar, dass dieses Verfahren der natürliche Signifikanztest für eine nach (8) gewogene ES* ist:

(9) $\mathbf{Z_{\Sigma}''} = \Sigma(Z_i) / \sqrt{k}$

Wäre der Erwartungswert aller Z_i bei existierendem Psi-Effekt unabhängig von n_i (was von einigen Parapsychologen tatsächlich diskutiert wurde!), so wäre die (sonst wenig fundierte) Gewichtung in „Stouffer's Z" sogar statistisch optimal.

4) Gegenargumente und alternative Gewichtskorrekturen

Man könnte gegen diese Vorschläge einwenden, dass der Decline-Effekt nur die Wirkung einer – in Abhängigkeit von n (bzw. m oder N) variierenden – Versuchsbedingung sei, die eine normale Moderatorvariable darstelle, weshalb an der Gewichtung nichts zu ändern sei. Das wäre aber ein unpassendes Argument, da es sich hier um eine der wenigen Variablen handelt, die sich nicht nur während des Experiments, sondern auch danach - bei der Gewichtung - auswirken. Dadurch potenziert sich ihre Wirkung in ungerechtfertigter Weise, was durch die vorgeschlagene Gewichtskorrektur lediglich kompensiert wird. (Der interexperimenteller Decline, der auch häufig diskutiert wird, wirkt sich nicht auf die Gewichtung aus, weshalb er oben aus der Betrachtung ausgeklammert wurde.)

Allerdings gibt es bei den RNG-PK-Experimenten, falls der Decline auf die Verkürzung der Trialdauer zurückgeht, noch eine Alternative zu den obigen Gewichtungsvorschlägen, die von **Ertel** in seinem Beitrag vertreten wird. Man könnte nämlich die Trialdauer t (die hier ja negativ mit der Trialzahl m korreliert!) zur alternativen Stichprobeneinheit der Experimente machen und so von m auf ein Maß der Versuchsdauer $m'=t\bullet m$ übergehen, das noch mit N zu multiplizieren wäre, um das endgültige Maß $n'=t\bullet n$ zu erhalten. Die ES würde dann bei kleinem t zunehmen, womit ihre Abnahme bei kurzer Trialdauer kompensiert und ihre negative Korrelation mit m genauso beseitigt wäre wie durch die (oben erwähnte) partielle Korrelation. Gleichzeitig müssten aber mehrere andere Statistiken in unübersichtlicher Weise angepasst werden und V_{ESi} sowie die Gewichtung nach (3) würden sich so ändern, dass die gewogene Summierung der ES_i nach (2) oder (4) zu einem Resultat käme, das auch auf einem übersichtlicheren und direkteren Weg

erlangt werden könnte. Dieser besteht darin, Ertels Multiplikation von n mit t einfach als Gewichtskorrektur zu deuten, durch die ein zusätzlicher Faktor $\mathbf{t_i}$ in (3), (2) und (4) eingeführt wird. (Das entspräche einer Auspartialisierung des Einflusses von t aus der Variablen m.). Nur dann würde t auch den Charakter einer (geradezu exemplarischen) Moderatorvariablen behalten.

Diese Gewichtung ist dort, wo die Voraussetzungen dafür gegeben sind, durchaus diskutabel. (Natürlich muss die Größe t bekannt sein, die aber nur bei 162 der 380 ausgewerteten Experimente - als Variable „bits pro Sek." - vorliegt!) Die Gewichtung des Referenten hat jedoch den Vorteil, keine zusätzlichen empirischen Daten zu benötigen und auf verschiedene Arten des Decline anwendbar zu sein, ohne dass dieser vorher identifiziert werden muss. Sie eignet sich deshalb für alle Arten von Psi-Experimenten.

5. Weitere Argumente für eine gleichmäßigere Gewichtung

Es ist zuzugeben, dass die bisherigen Ableitungen und Vorschläge alle auf der Annahme eines Decline-Effekts beruhen. Aber es gibt noch drei weitere Argumente für eine alternative Gewichtung bei parapsychologischen Metaanalysen, die nichts mit diesem Effekt zu tun haben:
a) Die Unterschiede in den Trial- (bzw. Bit-)Zahlen sind bei Psi-Experimenten im allgemeinen und RNG-Experimenten im besonderen so extrem, dass bei einer n-proportionalen Gewichtung fast immer ganz wenige Experimente (oder nur ein einziges!) das Gesamtresultat bestimmen. Aus der Metaanalyse von Bösch et al. geht das deutlich hervor: Hier verhält sich n(max) zu n(min) wie 10^8 zu 1, während alle oben definierten Gewichte (einschließlich demjenigen von **Ertel)** nur auf 10^5 kommen! (Eliminiert man die drei größten Experimente, wie es die Autoren – in methodisch fragwürdiger Weise – getan haben, so ist ihr Verhältnis immer noch 10^7 zu 1.) Daher weisen auf ihrer Tabelle 6, wo die Experimente (nach 7 verschiedenen Moderatorvariablen) in Teilgruppen aufgeteilt sind, diejenigen Gruppen, in denen sich jeweils die drei größten Experimente befinden, immer die gleichen Werte ES* und Z_Σ auf wie die Gesamtheit aller Experimente!
b) In einem so kontroversen und so stark von Inkonsistenzen geprägten Gebiet wie der Parapsychologie sind Anhänger und Gegner in gleicher Weise daran interessiert, ihr Urteil auf einer Vielzahl unabhängiger Resultate aufzubauen, und wollen gerade nicht von einzelnen – möglicherweise verfälschten – Mammutstichproben überrannt werden. Hier das richtige Maß für eine angemessene Gewichtung zu finden, ist zwar kein eindeutig lösbares Problem, aber der obige Ansatz ist sicher demjenigen der Autoren vorzuziehen.
c) Schließlich ist darauf hinzuweisen, dass jede Gewichtung statistisch statthaft ist, sofern sie unabhängig von den aktuellen Daten a priori festgelegt wird. Letzteres ist bei den Vorschlägen des Referenten nachweislich der Fall. Im übrigen verliert eine nicht optimale Gewichtung nur etwas an Power, so dass der globale Z-Test konservativer ausfällt. In diesem Sinne wäre eine gleichmäßigere Gewichtung sogar vertretbar, wenn keinerlei Decline-Effekte existierten.

6. Anwendung der Gewichtungskorrektur auf die Metaanalyse von Bösch et al.

Damit dürfte der Standpunkt des Referenten ausreichend begründet sein, und es können (nach 4jähriger Wartezeit auf die Daten, deren alternative Auswertung bereits 2002 vorgeschlagen wurde) endlich die Resultate der modifizierten Signifikanzprüfung mitgeteilt werden. Berechnet wurde z_Σ gemäß (1) und (2) mit Benutzung der Gewichtung in (7) und (8). Dabei resultieren $\mathbf{Z'_\Sigma=8,5}$ ($\mathbf{P= 10^{-17}}$) und $\mathbf{Z''_\Sigma=13,1}$ ($\mathbf{P= 10^{-38}}$). Diese Werte sind mit den z-Werten der Autoren nach der klassischen FEM-Methode zu vergleichen, die von ihnen primär (nämlich schon

2002) benutzt wurde und zu der auch die vorne angeführten Formeln passen. Auf ihrer Tabelle 4 findet man hierzu den z-Wert **-3,67** (für alle 380 Experimente) und einen alternativen Wert **+3,59** (nach Weglassung der drei Experimente mit maximalem n, die übereinstimmend ein negatives Resultat erbrachten!). Die absurde Differenz zwischen den beiden Berechnungen weist eindeutig auf eine Gewichtung hin, die große Stichproben zu sehr begünstigt. Dass die Z-Werte bei der Gewichtung des Referenten deutlich höher ausfallen, liegt natürlich an der negativen Korrelation zwischen ES_i und n_i und am verringerten Einfluss von n_i auf die neu definierten Gewichte.

7. Vergleich mit der Gewichtung beim REM (Random Effect Model)

Nun haben die Autoren zusätzlich zum FEM in ihrem Artikel von 2006 noch eine stark abweichende Methode benutzt, nämlich das REM. Auch dieses führt zu einer weniger extremen Gewichtung, erreicht dieses Ziel aber auf einem ganz anderen Weg als der Referent. Das REM fügt nämlich der Standardfehlervarianz V_{ESi} bei jedem Experiment ein konstantes Korrekturglied hinzu, das die empirisch (aus der Metaanalyse) feststellbare interexperimentelle Varianz V'_{ES} enthält. (Da V'_{ES} oft so groß ist, dass beim statistischen Vergleich mit der mittleren V_{ESi} eine extreme Heterogenität ermittelt wird, die leider nicht hinlänglich durch Moderatorvariablen zu erklären ist, macht das REM quasi aus der Not eine Tugend und korrigiert die Fehlervarianz „passend" nach oben.) Wegen Gleichheit des Korrekturglieds bei allen Experimenten variiert dann w_i (das wie beim FEM nach (3) berechnet wird) weniger als beim FEM, wobei die Gewichtung um so gleichmäßiger ausfällt, je stärker V'_{ES} über der mittleren V_{ESi} liegt.

Im konkreten Fall muss demnach (wegen der exzeptionellen Heterogenität) eine Gewichtung resultieren, die viel ausgeglichener ist als beim FEM und (mit 10^6 zu 1) fast an die oben vorgeschlagene heranreicht. Dagegen muss Z_Σ beim REM stets kleiner ausfallen als bei der Methode des Referenten. Denn die generelle Erhöhung der V_{ESi}-Werte führt zwangsläufig zu einer erhöhten Schätzung für V_{ES*} in (1) und zu einem relativ niedrigen Z_Σ=2,47. Immerhin geht aus dem positiven Vorzeichen (das bei der FEM noch negativ war!) eindeutig hervor, dass die Gewichtung bei großen Stichproben erheblich verringert wurde.

Es scheint sich also beim REM um eine Art Danae-Geschenk zu handeln, bei dem die wünschenswerte (und hier nicht einmal an das Postulat eines Decline-Effekts gebundene!) Veränderung der Gewichtung durch einen sehr konservativen Signifikanztest erkauft werden muss. Doch in dieser Vermutung verbirgt sich bei Psi-Experimenten ein Irrtum, der auch den Autoren unterlaufen sein dürfte: Wie oben ausgeführt wurde, wird beim parapsychologischen Einstichprobendesign V_{ESi} aus einem theoretischen Parameter abgeleitet. (Die Autoren haben sich allerdings schon beim FEM nicht genau daran gehalten, was dort aber kaum praktische Auswirkungen hatte.) Die „Verbesserung" eines bekannten Parameters durch irgendeine empirische Schätzung macht nun offensichtlich keinen Sinn! Die Berechnung von Z_Σ kann hier also auch bei einer REM-Gewichtung nach dem üblichen Schema erfolgen, so dass lediglich die veränderten Gewichte w_i in (2) (und in alle anderen relevanten Formeln) einzusetzen sind. Führt man diese Berechnung nachträglich durch, so wird Z_Σ um den Faktor 1,503 größer und nimmt den Wert **3,71** an, der (mit **P=.0001**) den Resultaten des Referenten deutlich näher kommt.

Wenn man sich für diese Modifikation der REM-Auswertung entscheidet, kann man auf die obigen Argumente für eine alternative Gewichtung fast schon verzichten. So wird etwa das Decline-Argument durch das generellere Argument der Heterogenität ersetzt, auf die das REM automatisch mit einer gleichmäßigeren Gewichtung reagiert.

8. Eine neue Formel zur Schätzung der Zahl unpublizierter Psi-Experimente

Nachdem mittels verschiedener Methoden die Werte von Z_{Σ} deutlich nach oben korrigiert worden sind, wird auch die Publikationsbias-Hypothese der Autoren immer unwahrscheinlicher. Mit dieser wollten sie a) die unerwartete negative Korrelation zwischen ES und n und b) die ermittelte globale Signifikanz auf die Elimination insignifikanter Resultate zurückführen. Der Verdacht a) wurde (wie erwähnt) bereits dadurch irrelevant, dass die betreffende Korrelation auf den Einfluss der Variablen „Trialdauer" zurückführbar ist (vgl. den Beitrag von **Ertel**). Übrig bleibt der Verdacht b), der theoretisch natürlich immer erhoben werden kann. Zu seiner – wenigstens annähernd überzeugenden – statistischen Prüfung gehört jedoch eine plausible empirische Schätzung der tatsächlichen Eliminationsquote. Eine solche können die Autoren leider nicht vorlegen, während der Referent bereits auf dem WGFP-Workshop 2002 eine interessante Methode vorgetragen hat, die diesem Anspruch eher gerecht wird. Diese instrumentalisiert gewissermaßen den sog. Psi-Missing-Effekt, der als negative bzw. instruktionswidrige Trefferabweichung zu definieren ist und von den meisten Parapsychologen als (wenn auch seltenes) Faktum akzeptiert wird. Sie kann folgendermaßen skizziert werden:

a) Es wird vorausgesetzt, dass echtes und zugleich signifikantes Psi-Missing bei ganzen Experimenten sehr selten auftritt, weil dort etwaige negative Resultate einzelner Vpn meist durch die positiven Resultate anderer Vpn überkompensiert oder mindestens ausbalanciert werden.
b) Demgemäß wird (vereinfachend) angenommen, dass unterhalb des oberen 5%-Signifikanzbereichs keine echten Effekte mehr vorkommen. Bei Dreiteilung der Verteilung möglicher Resultate in einen oberen signifikanten Bereich (**O** mit $P_O \leq .05$), einen unteren signifikanten Bereich (**U** mit $P_U \geq .95$) und einen insignifikanten mittleren Bereich (**M**), sollten daher in den Bereichen M und U die Fallzahlen in demjenigen Verhältnis zueinander stehen, das unter H_0 bei einer Normalverteilung zu erwarten ist.
c) Es wird ferner (vereinfachend) angenommen, dass neuere Autoren bei ihrem Experiment Psi-Missing nicht prinzipiell ausschließen und daher nur Resultate des mittleren Bereichs M (gelegentlich) eliminieren und nicht publizieren. Die Zahl der eliminierten Fälle kann dann geschätzt werden als diejenige Zahl (X), die M hinzugefügt werden muss, um das erwartete Zahlenverhältnis zum unteren Bereich U wieder herzustellen.
d) Hierzu ist nur eine einfache lineare Gleichung zu lösen mit dem Resultat:

(10) $X = U (P_U - P_O) / (1 - P_U) - M$ | $\mathbf{X = 17{,}01 * U - M}$ [wenn $P_O = 1 - P_U = .05$]

(Für X kann auch ein Konfidenzintervall bestimmt werden, das zwar mit k abnimmt, aber bei üblichem k immer noch recht hoch ist.)
e) Durch Ausschluss des oberen Bereichs aus der Berechnung von X wird erreicht, dass dort vorhandene echte Effekte die Schätzung nicht beeinflussen und daher nicht erhöhen können. Bei anderen verbreiteten Schätzmethoden ist dagegen genau das der Fall, so dass sie nur bei fehlenden echten Effekten richtig sein können. - Sollte im Bereich U echtes Psi-Missing vorkommen (z.B. bei Eperimenten mit nur einer Vp), wird X überhöht und die Korrektur konservativ. Wurden auch aus U Fälle eliminiert, so ist die Korrektur zu gering, weshalb bei diesem Verdacht die Signifikanzgrenze erhöht werden sollte.
f) Die ergänzten Fälle besitzen beim Z-Test meist einen mittleren Wert von 0 und können leicht in den globalen Signifikanztest einbezogen werden, am einfachsten dann, wenn die z-Werte direkt summiert werden. Der korrigierte Summen z-Test für k Fälle lautet (bei $P_O = 1 - P_U$):

(11) $\mathbf{Z_{\Sigma}^* = Z_{\Sigma} / (1 + x / k)^{1/2}}$

9. Anwendung der Eliminationskorrektur auf die Metaanalyse von Bösch et al.

Nach dieser Methode wurden nun für die 380 Experimente folgende Resultate erzielt:

Zahl der Fälle im unteren (U) und mittleren (M) Bereich:	U=27	M=252
Erwartungswerte nach der Normalverteilung:	E(U)=15,5	E(M)=263,5
Signifikanz der Abweichung davon (= Sign. d. Eliminationseffekts):		P(eins.)=.002
Geschätzte Zahl der aus Bereich M eliminierten Fälle (X):		X=207
Eliminationstendenz (ELT) (bezogen auf die Fälle M):		ELT=X/(X+M)=.45
Eliminationsquote (ELQ) (bezogen auf alle k Fälle):		ELQ=X/(X+U+M+O)=X/(X+K)=.35
Globaler Signifikanztest A [$\Sigma(\sqrt{n_i}\ ES_i)$]	Z_Σ [A] = 13,1	P = 10^{-38}
Test A nach Eliminierung von 6 extrem hohen z-Werten	Z_Σ [A-6] = 10,8	P = 10^{-26}
Globaler Signifikanztest B [$\Sigma(\sqrt{m_i}\ N_i\ ESi)$]	Z_Σ [B] = 8,5	P = 10^{-17}
Test A nach Hinzufügung der X eliminierten Fälle	Z_Σ* [A] = 10,5	P = 10^{-25}
Desgl. mit Eliminierung von 6 extrem hohen z-Werten	Z_Σ* [A-6] = 8,7	P = 10^{-17}
Test B nach Hinzufügung der X eliminierten Fälle	Z_Σ* [B] = 6,8	P = 10^{-11}

---

10. Unabhängigkeit der Eliminationsquote von der Stichprobengröße (n)

Die korrigierten $Z_{\Sigma*}$ sind alle so gut, dass trotz einer sehr signifikanten Zahl eliminierter Fälle (X=207) die globale Signifikanz der ES schwerlich zurückgewiesen werden kann.

Aber auch die Hypothese der Autoren, dass bei Experimenten mit kleinen Stichproben (n) eine höhere Eliminationstendenz bestehen würde (mit der sie die negative Korrelation zwischen ES und n erklären wollten), lässt sich nicht bestätigen. Zur Überprüfung wurden die Häufigkeitsverhältnisse M:U für die 4 von den Autoren definierten Größenklassen mittels Chi²-Test verglichen, was zu einem reinen Zufallsresultat (P=.60) führt. Auch der Vergleich anderer Teilgruppen ergab überwiegend keine Signifikanz. (Trotz des signifikanten Eliminationsunterschieds bei den Variablen „Publikationsjahr" und „Selektierte Probanden" bleibt hier die Signifikanz von Teilgruppen auch nach einer Eliminationskorrektur erhalten.)

Unterschiede in der Eliminationstendenz zwischen Teilstichproben der Experimente

4-klassige Variable „Stichprobenumfang"(n)	P=.60
3-klassige Variable „Zahl der Probanden"	P=.45
4-klassige Variable „Publikationsjahr"	P=.02

2-klassige Variable „Selektierte/nicht selekt.Probanden“	P=.035 (eins.) / r=.14
2-klassige Variable „Pilotstudie/normales Experiment“	P=.11 (eins.)
2-klassige Variable „Exper.v. H.Schmidt/nicht von ihm“	P=.23 (eins.)

Fazit:

Die Berechnungen des Referenten, in denen eine neue statistische Korrektur des Publikationsbias enthalten ist, führen zu einer so hohen Gesamtsignifikanz (Z=6,8 / $P=10^{-11}$), dass der Versuch der Autoren, die Effekte in allen RNG-PK-Experimente auf Datenelimination zurückzuführen, unhaltbar ist.