Examining Psychokinesis: The Interaction of Human Intention with Random Number Generators. A Meta-Analysis

Holger Bösch

University Hospital Freiburg, Department of Evaluation Research in Complementary Medicine, Freiburg, Germany

Fiona Steinkamp

Department of Psychology, University of Edinburgh, Edinburgh UK

Emil Boller

Institute for Border Areas of Psychology and Mental Hygiene, Freiburg, Germany

4 PSYCHOKINESE

4.1 Examining Psychokinesis: The Interaction of Human Intention with Random Number Generators. A Meta-Analysis*

Holger Bösch[1], Fiona Steinkamp[2], Emil Boller[3]

(1) University Hospital Freiburg, Department of Evaluation Research in Complementary Medicine, Freiburg, Germany

(2) Department of Psychology, University of Edinburgh, Edinburgh UK

(3) Institute for Border Areas of Psychology and Mental Hygiene, Freiburg, Germany

4.1.1 Abstract

Séance-room and other large-scale psychokinetic phenomena have fascinated mankind for decades. Experimental research has reduced these phenomena to attempts to influence (a) the fall of dice and, later, (b) the output of random number generators (RNGs). The meta-analysis combined 380 studies that assessed whether RNG output could correlate with human intention. A significant but very small overall effect size was found. The study effect sizes were strongly and inversely related to sample size and were extremely heterogeneous. A Monte Carlo simulation revealed that the small effect size, the relation between sample size and effect size, as well as the extreme effect size heterogeneity found, could in principle be a result of publication bias.

* Akzeptiert: Psychological Bulletin, 14.07.2005. Diese Publikation wurde von zwei Autorenkollektiven kommentiert. Bösch, Steinkamp & Boller, die Autoren der Originalarbeit, wurden daraufhin zu einer Antwort eingeladen:

Radin, D., Nelson, R., Dobyns, Y., & Houtkooper, J. (in press). Reexamining psychokinesis: Comment on the Bösch, Steinkamp and Boller (in press) Meta-Analysis. *Psychological Bulletin*.

Wilson, D. B., & Shadish, W. R. (in press). On blowing trumpets to the tulips: To prove or not to prove the null hypothesis - Comment on Bösch, Steinkamp and Boller (in press). *Psychological Bulletin*.

Bösch, H., Steinkamp, F., & Boller, E. (in press). In the eye of the beholder: Reply to Wilson & Shadish (2006) and Radin, Nelson, Dobyns and Houtkooper (in press). *Psychological Bulletin*.

4.1.2 Introduction

During the 1970s, Uri Geller inspired much public interest in phenomena apparently demonstrating the ability of mind to exert power over matter in his demonstrations of spoon bending using his alleged psychic ability (Targ & Puthoff, 1977; Wilson, 1976) and lays claim to this ability even now (e.g., Geller, 1998). Belief in this phenomenon is widespread. In 1991 (Gallup & Newport), 17 percent of American adults believed in “the ability of the mind to move or bend objects using just mental energy” (p. 138) and seven percent even claimed that they had “seen somebody moving or bending an object using mental energy” (p. 141).

Unknown to most academics, a large amount of experimental data has accrued testing the hypothesis of a direct connection between the human mind and the physical world. It is one of the very few lines of research where replication is the main and central target, a commitment that some methodologists wish to be the commitment of experimental psychologists in general (e.g., Cohen, 1994; Rosenthal & Rosnow, 1991). This article will summarize how the empirical investigation of this phenomenon developed over the decades and will present a new meta-analysis of a large set of experiments examining the interaction between human intention and random number generators.[1]

4.1.3 Psi Research

Psi phenomena (Thouless, 1942; Thouless & Wiesner, 1946) can be split into two main categories: psychokinesis (PK) and extrasensory perception (ESP). *Psychokinesis* refers to the apparent ability of humans to affect objects solely by the power of the mind, and ESP relates to the apparent ability of humans to acquire information without the mediation of the recognized senses or inference. Many researchers believe that PK and ESP phenomena share a common underlying mechanism (e.g., Pratt, 1949; J. B. Rhine, 1946; Schmeidler, 1982; Stanford, 1978; Thalbourne, in press; Thouless & Wiesner, 1946). Nevertheless, the two phenomena have been treated very differently right from the start of their scientific examination. For instance, whereas J. B. Rhine and his col-

[1] In this article, the term experiment refers to a one-sample approach generally used in psi research (see Method).

leagues at the Psychology Department at Duke University immediately published the results of their first ESP card experiments (Pratt, 1937; Price & Pegram, 1937; J. B. Rhine, 1934, 1936, 1937; L. E. Rhine, 1937), they withheld the results of their first PK experiments for nine years (L. E. Rhine & J. B. Rhine, 1943), even though the ESP and PK experiments had both been carried out at the same time: Rhine and his colleagues did not want to undermine the scientific credibility that they had gained through their pioneering monograph on ESP (Pratt, J. B. Rhine, Smith, Stuart & Greenwood, 1940).

When L. E. Rhine & J. B. Rhine (1943) went public with their early dice experiments, the evidence for PK was based not only on above-chance results, but also on a particular scoring pattern. In those early experiments, participants were asked to throw a prespecified combination of die faces (e.g., a 1 and a 6). The researchers discovered that success declined during longer series of experiments, which was thought to be a pattern suggestive of mental fatigue (Reeves & Rhine, 1943; J. B. Rhine & Humphrey, 1944, 1945). This psychologically plausible pattern of decline seemed to eliminate several counterhypotheses for the positive results obtained, such as die bias or trickery, because they would not lead to such a systematic decline. However, as the number of experimental PK studies and their quality increased, the decline pattern became less important as a means of evidential support for the psi hypothesis.

4.1.3.1 Verifying Psi

In order to verify the existence of psi phenomena, 13 meta-analyses have already been conducted (Bem & Honorton, 1994; Honorton, 1985; Honorton & Ferrari, 1989; Milton, 1993, 1997; Milton & Wiseman, 1999a, 1999b; Radin & Ferrari, 1991; Radin & Nelson, 1989, 2003; Stanford & Stein, 1994; Steinkamp, Milton & Morris, 1998; Storm & Ertel, 2001), two of which provide no evidence for psi (Milton & Wiseman, 1999a, 1999b). Only three meta-analyses on psi data address research on PK (Radin & Ferrari, 1991; Radin & Nelson, 1989, 2003), basically because research on ESP produced a greater diversity of experimental approaches. Although there has been some variety in methods to address PK, such as coin tossing and influencing the outcome of a roulette wheel, these methods have been used only occasionally.

The greater variety of experimental approaches to assess ESP may explain why potential moderators of PK, such as the distance between the participant and the target, as well as various psychological variables, have not been investigated as systematically as alleged moderators of ESP. To date, no PK meta-analysis has reported data on potential moderators and the three main reviews of potential PK moderators (Gissurarson, 1992 & 1997; Gissurarson & Morris, 1991; Schmeidler, 1977) have arrived at inconclusive results.

Nevertheless, three of the ESP meta-analyses have tentatively established potential moderators--significant correlations have been found between ESP and (a) extraversion (Honorton, Ferrari & Bem, 1998), (b) belief in ESP (Lawrence, 1998), and (c) defensiveness (Watt, 1994). It seems to us that there is a general disparity between the experimental investigations of the two categories of psi. From the very beginning, researchers have focused on ESP.

4.1.3.2 Psychology and Psi

Psychological approaches to psi experiences have also almost exclusively focused on ESP. For example, some researchers hypothesize that alleged ESP experiences are the result of delusions and misinterpretations (e.g., Alcock, 1981; Blackmore, 1992; Brugger et al., 1993; Persinger, 2001). A line of research addressing the misinterpretation of alleged PK events was initiated by Langer in 1975 and meta-analyzed once her ideas had been operationalized in various ways (Presson & Benassi, 1996). Personality-oriented research established connections between belief in ESP and personality variables (Irwin, 1993; see also, Dudley, 2000; McGarry & Newberry, 1981; Musch & Ehrenberg, 2002). Experience-oriented approaches to paranormal beliefs, which stress the connection between paranormal belief and paranormal experiences (e.g., Alcock, 1981; Blackmore, 1992; Schouten, 1983) and media-oriented approaches, which examine the connection between paranormal belief and depictions of paranormal events in the media (e.g., Sparks, 1998; Sparks, Hansen & Shah, 1994; Sparks, Nelson & Campbell, 1997) both focus on ESP, although the paranormal belief scale most frequently used in this line of research also has some items on PK (Thalbourne, 1995).

4.1.3.3 The Beginning of the Experimental Approach to Psychokinesis

Reports of séance-room sessions during the late 19th century are filled with claims of extraordinary movements of objects (e.g., Crookes, Horsley, Bull, & Meyers, 1885), prompting some outstanding researchers of the time to devote at least part of their career to determining whether the alleged phenomena were real (e.g., Crookes, 1889; James, 1896; Richet, 1923). In these early days, as in psychology, case studies and field investigations predominated. Experiments using randomization and statistical analysis to draw conclusions were just about to become standard in the empirical sciences (Hacking, 1988). Hence, it is not surprising that in this era experimental approaches and statistical analyses were used only occasionally (e.g., Edgeworth, 1885, 1886; Fisher, 1924; Richet, 1884; Sanger, 1895; Taylor, 1890). Even J. B. Rhine, the founder of the experimental study of psi phenomena, abandoned case studies and field investigations as a means of obtaining scientific proof only after he exposed several mediums as frauds (e.g., J. B. Rhine & L. E. Rhine, 1927). However, after a period of several years when he and his colleagues focused almost solely on ESP research, their interest in PK was reawakened when a gambler visited the laboratory at Duke University and casually mentioned that many gamblers believed they could mentally influence the outcome of a throw of dice. This inspired J. B. Rhine to perform a series of informal experiments using dice. Very soon experiments with dice became the standard approach for investigating PK.

Difficulties in devising an appropriate methodology soon became apparent and improvements in the experimental procedures were quickly implemented. For example, standardized methods were developed for throwing the dice, dice-throwing machines were used to prevent participants from manipulating their throw of the dice, and recording errors were minimized by having experimenters either photograph the outcome of each throw or having a second experimenter independently record the results. Commercial, pipped dice were found to have sides of unequal weight, with the sides with the larger number of excavated pips, such as the 6, being lighter and hence more likely to land uppermost than lower numbers, such as the 1. Consequently, experiments required participants to attempt to score seven with two dice, or used a (counter) balanced design in which the target face alternated from one side of the die (e.g., 6) to the opposite site (e.g., 1).

In 1962, Girden (1962a) published a comprehensive critique of dice experiments in the *Psychological Bulletin*. Among other things, he criticized the experimenters for pooling data as it suited them, and for changing the experimental design once it appeared that results were not going in a favorable direction. He concluded that the results from the early experiments were largely due to the bias in the dice and that the later, better controlled experiments were progressively tending toward nonsignificant results. Although Murphy (1962) disagreed with Girden's conclusion, he did concede that no "ideal" experiment had yet been published that met all six quality criteria--namely one with (a) a sufficiently large sample size; (b) a standardized method of throwing the dice; (c) a balanced design; (d) an objective record of the outcome of the throw; (e) the hypothesis stated in advance; and (f) a prespecified end point.

The controversy about the validity of the dice experiments continued (e.g., Girden, 1962b; Girden & Girden, 1985; Rush, 1977). Over time, experimental and statistical methods improved and, in 1991, Radin & Ferrari undertook a meta-analysis of the dice experiments.

4.1.4 Dice Meta-Analysis

The dice meta-analysis comprised 148 experimental studies and 31 control studies published between 1935 and 1987. In the experimental studies 2569 participants tried mentally to influence 2,592,817 die-casts to land with a predefined die face uppermost. In the control studies, a total of 153,288 dice were tossed (a) without a specific target aim or (b) "the condition was defined as such in the published report" (Radin & Ferrari, 1991, p. 65). The experimental studies were coded for various quality measures, including a number of those mentioned by Girden (1962a). Table 1 provides the main meta-analytic results.[2]

[2] To compare the meta-analytic findings from the dice and previous RNG meta-analyses with those from our RNG meta-analysis, we converted all effect size measures to the proportion index π which we use throughout the paper (see Method). Although we use a fixed effects model as well as a random effects model for our own analyses, the first dice and the first RNG meta-analyses exclusively used a weighted ($1/v$) fixed effects model. Because it is not possible to calculate a random effects model given only the published data, all analyses

Table 1 Main Results of Radin & Ferrari's (1991) Dice Meta-Analysis

	N	$\bar{\pi}_t$	***SE***	***z***
Dice-casts "Influenced"				
All studies	148	.50610	.00031	19.68***
All studies, quality weighted	148	.50362	.00036	10.18***
Balanced studies	69	.50431	.00055	7.83***
Balanced studies, homogenous	59	.50158	.00061	2.60**
Balanced studies, homogenous, quality weighted	59	.50147	.00063	2.33**
Dice-casts Control				
All studies	31	.50047	.00128	0.36***

Note. Published effect sizes on the basis of r = z/√N were transformed using $\bar{\pi}_t = .5\bar{r} + .5$ to achieve comparability.

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. All p-values are one-tailed.

The overall effect size, weighted by the inverse of the variance, is small but highly significant ($\bar{\pi}_t$= .50610, *z* = 19.68). Radin & Ferrari calculated that approximately 18,000 null effect studies would have been required to reduce the result to a nonsignificant level (Rosenthal, 1979).[3] When the studies were

on previous dice and RNG data are exclusively based on fixed effects modeling. We *transformed* the published results, which used the effect size *r*=*z*/sqrt(*n*), using $\bar{\pi}_t = .5\bar{r} + .5$. This transformation is accurate as long as the *z*-values of the individual studies are based on two equally likely alternatives ($p = q = .5$).

However, the *z*-scores of most dice experiments are based on six equally likely alternatives ($p = 1/6$ and $q = 5/6$). Consequently $\bar{\pi}_o$ as computed on the basis of the *original* data and $\bar{\pi}_t$ as computed on the basis of the transformation formula diverge slightly because *r* no longer remains in the limits of +/-1. However, the difference between $\bar{\pi}_o$ and $\bar{\pi}_t$ is very small (< .05%) as long as the *z*-values are not extreme ($z < 10$, $p < 1 * 10^{-10}$). The difference is smaller the closer the value is to the null value of .50, which is the case for all effect sizes presented here.

[3] Rosenthal's approach is based on the assumption that the unpublished studies are a random sample of all conducted studies, that is, the approach assumes that the mean *z*-score of the unpublished studies is zero. This assumption has been questioned by several authors (e.g., Iyengar & Greenhouse, 1988; Scargle, 2000). If one were to assume instead that the

weighted for quality, the effect size decreased considerably ($z_{\Delta} = 5.27$, $p = 1.34 * 10^{-7}$; see Table 1 for comparison), but was still highly significantly above chance.

The authors of the dice meta-analysis found that there were indeed problems regarding die bias, with the effect size of the target face 6 being significantly larger than the effect size of any other target face. They concluded that this bias was sufficient to cast doubt on the whole database. They subsequently reduced their database to only those 69 studies that had correctly controlled for die bias (the “balanced database”, in which the target face had been alternated equally from one side of the die to the opposite site). As shown in Table 1, the resultant effect size remained statistically highly significant, although the effect size decreased considerably. However, the effect sizes of the studies in the balanced database were statistically heterogeneous. When Radin & Ferrari trimmed the sample until the effect sizes in the balanced database became homogenous, the effect size was reduced to only $\bar{\pi}_{t} = .50158$ and fell yet further to $\bar{\pi}_{t} = .50147$ when the 59 studies were weighted for quality. Only 60 unpublished null effect studies are required to bring the balanced, homogenous and quality-weighted studies down to a nonsignificant level.[4] Ultimately, the dice meta-analysis did not advance the controversy over the putative PK effect beyond the verdict of “not proven”, as mooted by Girden (1962b, p. 530) almost 30 years earlier.

Moreover, the meta-analysis has several limitations; Radin & Ferrari neither examined the source(s) of heterogeneity in their meta-analysis, nor addressed whether the strong correlation between effect size and target face disappeared when they trimmed the 79 studies not using a balanced design from the overall sample. The authors did not analyze potential moderator variables. For instance,

unpublished studies were a random sample of the nonsignificant studies only, and that the mean z-score of the unpublished studies were $z = -0.1085$ (Scargle, 2000), then 1450 studies, rather than 18,000 studies, would be needed to reduce the overall effect to a nonsignificant level.

[4] For this particular subsample Radin & Ferrari did not report Rosenthal’s (1979) failsafe number (X), that is the number of unpublished null effects needed to reduce the result to just $p = .05$. We calculated X on the basis of Stouffer z (z_n) provided in the article (Table 2, p. 76) and calculated $X = (n/2.706)[n(\bar{z}_n)^2 - 2.706]$ as proposed by Rosenthal (1979), where $\bar{z}_n = z_n/\sqrt{n}$.

the studies varied considerably regarding the type of feedback given to participants, with some participants gaining no feedback at all; the type of participant who was recruited, with some studies recruiting psychic claimants and other studies recruiting participants with no claim to having any “psychic powers”; and the experimental instructions that were given to participants, with some experiments asking participants to predict which die face would land uppermost in a future die cast thrown by someone other than the participant.

4.1.5 From Dice to Random Number Generator

With the arrival of computation, dice experiments were slowly replaced by a new approach. Beloff & Evans (1961) were the first experimenters to use radioactive decay as a truly random source to be influenced. In the initial experiments, participants would try mentally to slow down or speed up the rate of decay of a radioactive source. The mean disintegration rate of the source subjected to mental influence was then compared with that of a control condition in which there had been no attempt at mental influence.

Soon after this, experiments were devised in which the output from the radioactive source was transformed into bits (1s or 0s) that could be stored on a computer. These devices were known as random number generators (RNGs). Later, RNGs were generally based on avalanche noise (Zener diode) and thermal noise as the source of randomness. During the first decade of RNG research the truly random origin was an important factor for using RNGs (e.g., Beloff & Evans, 1961; Schmidt, 1970a), although the technical feasibility and, in comparison with dice experiments, the much better control over the experimental conditions, played the most important role in conducting RNG experiments (Schmidt, 1992). However, during the 1970s some physicists, inspired by the early RNG experiments, started to model psi phenomena in the framework of quantum physics. Building on the ‘measurement problem’ formulated in the Copenhagen Interpretation, the Observational Theory models psi effects in analogy to the collapse of the state vector, which is believed to be related to the consciousness of the observer (e.g., Lucadou & Kornwachs, 1977; Schmidt, 1975; Walker, 1974, 1975). During this time parapsychological modelling was very productive (for a review, see Stokes, 1987). New models accounting for the putative anomalous effects still evolve (e.g., Houtkooper, 2002, Jeffers, 2003; Shoup, 2002; Stapp, 1994).

During the time that the Observational Theories evolved, PK experiments with dice were almost entirely replaced with PK experiments using RNGs. This line of research was, and continues to be, pursued by many experimenters, but predominantly by Schmidt (e.g., 1969), and later by the Princeton Engineering Anomalies Research (PEAR) laboratory at Princeton University (e.g., Jahn, Dunne & Nelson, 1980).

4.1.5.1 RNG Experiments

In a typical PK RNG-experiment, a participant presses a button to start the accumulation of experimental data. The participant's task is mentally to influence the RNG to produce, say, more 1s than 0s for a predefined number of bits. Participants are generally given real-time feedback of their ongoing performance. The feedback can take a variety of forms. For example, it may consist in the lighting of lamps "moving" in a clockwise or counter clockwise direction, or in clicks provided to the right or left ear, depending on whether the RNG produces a 1 or a 0. Today, feedback is generally software implemented and is primarily visual. If the RNG is based on a truly random source, it should generate 1s and 0s an equal number of times. However, because small drifts cannot be totally eliminated, experimental precautions such as the use of XOR filters or balanced designs in which participants alternate their aim towards a 1 or a 0 from run to run are still required.

RNG experiments have many advantages over the earlier dice experiments, making it much easier to perform quality research with much less effort. Computerization alone meant that many of Girden (1962a) and Murphy's (1962) concerns about methodological quality could be overcome. If we return to Murphy's list of six methodological criteria, then (a) unlike with manual throws of dice, RNGs made it possible to conduct experiments with large sample sizes in a short space of time; (b) the RNG was completely impersonal--unlike the dice, it was not open to any classical (normal human) biasing of its output; (c) balanced designs were still necessary due to potential drifts in the RNG; (d) the output of the RNG could be stored automatically by computer, thus eliminating recording errors that may have been present in the dice experiments; (e) like the dice experiments, the hypotheses still had to be formulated in advance; and (f) like the dice experiments, optional stopping, that is arbitrarily terminating the experiment at a point of statistical significance, could still be a potential problem. Thus, RNG research entailed that, in practical terms, researchers no longer had to be concerned about alleged weak points (a), (b) and (d).

4.1.5.2 New Limits

From a methodological point of view, RNG experiments have many advantages over the older dice experiments. However, in respect of ecological validity, RNG experiments have some failings. Originally, the PK effect to be assessed was macroscopic and visual. Experimentalists then reduced séance-room PK, first to PK on dice, and then to PK on a random source in an RNG. But, as some commentators have argued, PK may not be reducible to a microscopic or quantum level (e.g., Braude, 1997). Moreover, psychologically a dice experiment is very different from an RNG experiment. Most people have played with dice, but few have had prior experience with RNGs. Additionally, an RNG is a complicated technical gadget from which the output must be computed before feedback can be presented. Complex operations are performed within the RNG before the random physical process results in a sequence of 1s and 0s. The output and the fundamental physical process are generally only partly correlated, that is, the output is at some remove from the fundamental physical process. Nevertheless, the ease with which PK data can be accumulated using an RNG has led to PK RNG experiments forming a substantial proportion of available data. Three related meta-analyses of these data have already been published.

4.1.6 Previous RNG Meta-Analyses

The first RNG meta-analysis was published by Radin & Nelson (1989) in *Foundations of Physics*. This meta-analysis of 597 experimental studies published between 1959 and 1987 found a small but significant effect of $\bar{\pi}_o = .50018$ ($SE = .00003$, $z = 6.53$, $p < 1 * 10^{-10}$).[5] The size of the effect did not di-

[5] The meta-analysis provided the overall effect size only in a figure (Fig. 3, p. 1506). Because its first author kindly provided us with the original data, we were able to calculate the

minish when the studies were weighted for quality or when they were trimmed by 101 studies to render the database homogenous.

The limitations of this meta-analysis are very similar to the limitations of the dice meta-analysis. The authors did not examine the source(s) of heterogeneity and did not specify definite and conclusive inclusion and exclusion criteria.[6] The authors took a very inclusive approach. Participants in the included studies varied from humans to cockroaches (Schmidt, 1979), feedback ranged from no feedback at all to the administration of electric shocks, and the meta-analysis included not only studies using true RNGs, which are RNGs based on true random sources such as electronic noise or radioactive decay, but also those using pseudo RNGs (e.g., Radin, 1982), which are based on deterministic algorithms. However, the authors did not discuss the extreme variance in the distribution of the studies' *z*-scores and did not assess any potential moderator variables, which were also two limitations of the dice meta-analysis. Nevertheless, this first RNG meta-analysis served to justify further experimentation and analyses with the PK RNG approach.

Almost 10 years later, in his book aimed at a popular audience, Radin (1997) recalculated the effect size of the first RNG meta-analysis claiming that the "overall experimental effect, calculated per study, was about 51 percent" (p. 141). However, this newly calculated effect size is two orders of magnitude larger than the effect size of the first RNG meta-analysis (50.018%). The increase has two sources. First, Radin removed the 258 PEAR laboratory studies included in the first meta-analysis (without discussing why) and second, he pre-

overall effect size and the relevant statistics.

[6] Although the authors state that they selected experiments examining the hypothesis, that "the statistical output of an electronic RNG is correlated with observer intention in accordance with prespecified instructions, as indicated by the directional shift of distribution parameters (usually the mean) from expected values" (p. 1502), this statement cannot be considered definite. The meta-analysis included experiments with animals (e.g. cockroaches), which puts into question the use of the term "observer intention", and included experiments using pseudo RNGs, that is, RNGs based on deterministic mathematical algorithms, which puts into question the term "electronic RNG". That the meta-analysis suffers from vaguely defined inclusion and missing exclusion criteria is particularly evident in respect to the title of the meta-analysis: "Evidence for consciousness-related anomalies in random physical systems".

sented simple mean values instead of weighted means as presented 10 years earlier. The use of simple mean values in meta-analyses is generally discredited (e.g., Shadish & Haddock, 1994), because it does not reflect the more accurate estimates of effect size provided by larger studies. In the case of the data presented in Radin's book, the difference between computing an overall effect size using mean values rather than weighted mean values is dramatic. The removal of the PEAR laboratory studies effectively increased the impact of other small studies that had very large effect sizes. The effect of small studies on the overall outcome will be a very important topic in the current meta-analysis.

Recently, Radin & Nelson (2003) published an update of their earlier (1989) RNG meta-analysis, adding a further 176 studies to their database. In this update, the PEAR laboratory data were collapsed into a new, single data point. The authors reported a simple mean effect size of 50.7%. Presented as such, the data appear to suggest that this updated effect size replicates that found in their first RNG meta-analysis. However, when the weighted fixed effects model (FEM) is applied to the data, as was used in the first RNG meta-analysis, the effect size of the updated database becomes $\overline{\pi}_o$= .50005, which is significantly smaller than the effect size of the original RNG meta-analysis (z_Δ = 4.27, p = 1.99 * 10^{-5}; see Table 2 for comparison).[7] One reason for the difference is the increase in sample size of the more recent experiments, which also have a concomitant decline in effect size.

Like the other meta-analyses, the updated 2003 meta-analysis did not investigate any potential moderator variables and no inclusion and exclusion criteria were specified; it also did not include a heterogeneity test of the database. All three meta-analyses were conducted by related research teams and thus an inde-

[7] The difference in effect size between $\overline{\pi}_o$, that is the effect size based on original data and $\overline{\pi}_t$, that is the effect size based on the transformed effect size (see Footnote 1) can be seen when the results of the overall dice meta-analysis as presented in Table 1 are compared with the results presented in Table 2. Although the difference is statistically highly significant (z_Δ= 4.12, p = 3.72 * 10^{-5}), the order of magnitude is the same. Because Dean Radin, the first author of the dice meta-analysis, kindly provided us with the basic data files of the dice meta-analysis, this comparison was made possible. However, the data file did not enable us to calculate the effect sizes of the specific subgroups as summarized in Table 1.

Table 2 Previous PK Meta-analyses - Total Samples

	N	$\overline{\pi}_o$	***SE***	***z***	π ***mean***
Dice					
1991 Meta-analysis	148	.50822	.00041	20.23***	.51105
RNG					
1989 First meta-analysis	597	.50018	.00003	6.53***	.50414
1997 First MA without PEAR data	339	.50061	.00009	6.41***	.50701
2000 Second meta-analysis	515	.50005	.00001	3.81***	.50568

Note. The effect size measure $\overline{\pi}_o$ was computed from original data available to the authors. π *mean* = the unweighted averaged effect size of studies.
$^{***}p$ < .001 (one-tailed).

pendent replication of their findings is lacking. The need for a more thoroughgoing meta-analysis of PK RNG experiments is clear.

4.1.7 Human Intention Interacting with Random Number Generators: A New Meta-Analysis

The meta-analysis presented here was part of a five-year consortium project on RNG experiments. The consortium comprised research groups from the PEAR laboratory, USA; the University of Giessen, Germany; and the Institut für Grenzgebiete der Psychologie und Psychohygiene [Institute for Border Areas of Psychology and Mental Hygiene] in Freiburg, Germany. After all three groups in the consortium failed to replicate the shift in the mean value of the PEAR laboratory data (Jahn et al., 2000), which form one of the strongest and most influential datasets in psi research, the question about possible moderating variables in RNG experiments rose to the forefront. Consequently, a meta-analysis was conducted to determine whether the existence of an anomalous interaction could be established between *direct human intention* and the *concurrent output of a true RNG*, and if so, whether there were moderators or other explanations that influenced the apparent connection.

4.1.8 Method

4.1.8.1 Literature Search

The meta-analysis began with a search for any experimental report that examined the possibility of an anomalous connection between the output of an RNG and the presence of a living being. This search was designed to be as comprehensive as possible in the first instance, and to be trimmed later in accordance with our prespecified inclusion and exclusion criteria. Both published and unpublished manuscripts were sought.

A total of 372 experimental reports were retrieved using multiple search strategies. The first step involved an extensive manual search at the library and archives of the Institut für Grenzgebiete der Psychologie und Psychohygiene in Freiburg, Germany, which provides the most comprehensive international collection of literature on psi research. Although, generally, computerized search strategies are crucial, in psi research manual searches are necessary because most of the relevant literature is not or only fragmentarily indexed in common databases such as PsycINFO. Our search included the following journals: *Proceedings of the Parapsychological Association Annual Convention* (1968, 1977-2004), *Research in Parapsychology* (1969-1993), *Journal of Parapsychology* (1959-2003), *Journal of the Society for Psychical Research* (1959-2004), *European Journal of Parapsychology* (1975-2003), *Journal of the American Society for Psychical Research* (1959-2002), *Journal of Scientific Exploration* (1987-2004), *Subtle Energies* (1991-2002), *Journal of Indian Psychology* (1978-2002), *Tijdschrift voor Parapsychologie* (1959-2004), *International Journal of Parapsychology* (1959-1968, 2000, 2001), *Cuadernos de Parapsicologia* (1963-2002), *Revue Métapsychique* (1960-1983), *Australian Parapsychological Review* (1983-2000), *Research letter of the Parapsychological Division of the Psychological Laboratory of Utrecht* (1971-1984), *Bulletin PSILOG* (1981-1983), *Journal of the Southern California Society for Psychical Research* (1979-1985), and the *Arbeitsberichte Parapsychologie der technischen Universität Berlin* (1971-1980). Although for some journals the search seems incomplete, we have always searched the most current issue of the respective journal. Current omissions are generally the result of a journal being behind schedule (e.g., *Journal of the American Society for Psychical Research*). All substantial omissions are the result of journals having stopped or suspended publication (e.g., *International Journal of Parapsychology*). The conference proceedings of the *Parapsychological Associations' Annual Convention* appeared to be the most important single source. Any gaps in the library's holdings of the conference proceedings was compensated for by *Research in Parapsychology*, which is a post-conference volume providing extended abstracts of most conference contributions.

The second step to retrieving studies was the search of three computer-based databases using different search terms and search strategies with regard to the content and the indexing methods of the respective database. The *Psiline Database System* (Vers. 1999), a continuously updated specialized electronic resource of parapsychologically-relevant writings (White, 1991) was searched using the key words *random number generator*, *RNG*, *random event generator* and *REG*. *Dissertation Abstracts on Disc* (8 CDs; Jan. 1961 - Jun. 2004) was searched using four different search strategies. First, the key words r*andom number generator*, *RNG*, r*andom event generator*, *REG, randomness*, *radioactive*, *parapsychology*, *parapsychological*, *perturbation*, p*sychokinesis*, *PK*, *extra-sensory perception*, *ESP*, *telepathy*, *precognition* and *calibration* were used. Second, the key words *random* and *experiment* were combined with *event*, *number*, *noise*, *anomalous*, *anomaly*, *influence*, *generator*, *apparatus* or *binary*. Third, the key word *machine* was combined with *man* or *mind*. Fourth, the key word z*ener* was combined with *diode*. The search included plural variants of all key words accordingly. However, not all key words were indexed for all CDs. PsycINFO (Jun. 2004) was searched using three different search strategies. First the key words r*andom number generator*, *RNG*, r*andom event generator*, *REG, perturbation* and *psychokinesis* were used. Second, the key word *machine* was combined with *man* or *mind*, and third, the key word *random* was combined with *calibration* and *radioactive*.

The reference list of the first RNG meta-analysis (Radin & Nelson, 1989), which was kindly provided to us by the authors, was searched for reports using true RNGs. To obtain as many relevant unpublished manuscripts as possible, visits were made to three other prolific parapsychology research institutes: the

Rhine Research Center, Durham NC; the PEAR laboratory at Princeton University; and the Koestler Parapsychology Unit at Edinburgh University. Furthermore, a request for unpublished experiments was placed on an electronic mailing list for professional parapsychologists (Parapsychology Research Forum [PRF]).

As a final step, the reference sections of all retrieved reports, that is, journal articles, conference proceedings, thesis/dissertations and so forth were searched. The search covered a broad range of languages and included items in Dutch, English, French, German, Italian and Spanish and was otherwise limited only because of lack of further available linguistic expertise.

4.1.8.2 Inclusion and Exclusion Criteria

The final database included only experimental reports that examined the correlation between *direct human intention* and the *concurrent* output of *true RNGs*. Thus, after the comprehensive literature search was conducted, we excluded experiments that: (a) involved, implicitly or explicitly, only an *indirect* intention toward the RNG. For example, telepathy experiments, in which a receiver attempts to gain impressions about the sender's viewing of a target that had been randomly selected by a true RNG, were excluded (e.g., Tart, 1976). Here, the receiver's intention is presumably directed to gaining knowledge about what the sender is viewing, rather than on influencing the RNG; (b) used *animals* or *plants* as participants (e.g., Schmidt, 1970b); (c) assessed the possibility of a *non-intentional*, or only *ambiguously intentional*, effect. For instance, experiments evaluating whether hidden RNGs could be influenced when the participant's intention was directed to another task or another RNG (e.g., Varvoglis & McCarthy, 1986) or experiments with babies as participants (e.g., Bierman, 1985); (d) looked for an effect *backwards in time* or, similarly, in which participants observed the same bits a number of times (e.g., Morris, 1982; Schmidt, 1985); (e) evaluated whether there was an effect of human intention on a *pseudo* RNG (e.g., Radin, 1982).

Additionally, experiments were excluded if their outcome could not be transformed into the effect size $\bar{\pi}$ that was prespecified for this meta-analysis. This excluded studies of which the data are not expected to be binomially distributed. As a result, for example, experiments that compared the rate of radioactive decay in the presence of attempted human influence with that of the same element in the absence of human intention (e.g., Beloff & Evans, 1961), were excluded.

Deciding which experiments to include and which to exclude, even if the criteria are clearly defined, can be as delicate as deciding how to perform the literature search and as decisions made during the coding procedure. The decisions not only depend on the skills of the person who decides but also, and sometimes even more importantly, on the report itself, which may be written ambiguously. Generally, any difficult or potentially contentious decisions were discussed by all three authors. From the 372 experimental reports retrieved, 255 were excluded after applying the inclusion and exclusion criteria.

4.1.8.3 Defining Studies

Some experiments were described in both published and unpublished reports, or both in a full journal article and elsewhere in an abstract. In these cases, all reports of the same experiment were used to obtain information for the coding, but the report with the most details was classified as the “main report”. The main reports often contained more than one “study”. A study was the smallest experimental unit described that did not overlap with other data in the report. This enabled the maximum amount of information to be included. In cases where the same data could be split up in two different ways (e.g., men vs. women or morning sessions vs. afternoon sessions), the split was used that appeared to reflect the author’s greatest interest in designing the study. At the same time the split of data is a very important quality measure. The split is a subgroup analysis, which might be planned a priori or conducted post hoc and interpreted with caution. The reference list of this meta-analysis refers to the main reports only.

Many experimenters performed randomness checks of the RNG to ensure that the apparatus was functioning properly. These control runs were coded in a separate “control” database. Data for these control runs, like the experimental database, were split based on the smallest unit described. In some experiments, data were gathered in the presence of a participant with an instruction to the participant “not to influence” the RNG (e.g., Jahn et al., 2000). These data were excluded from both experimental and control databases due to the inherent ambiguity as to whether the participant is attempting an influence during these data-gathering periods. Jahn also argued that these data should be excluded (Jeffers, 2003).

Although we have coded and analyzed unattended randomness checks as “control” studies, those studies are not the focus of our meta-analysis because all RNG studies included in our meta-analysis are based on a one-sample design, that is, the proportion of empirically accumulated 1s and 0s is compared to that of expected 1s and 0s under the null hypothesis that participants can perform no better than chance. The purpose of control studies is to demonstrate that “without intention” the apparatus produces results (binomially distributed) as expected theoretically. When control study data deviate from the expected value, the experimenter revises the experimental setup looking for variables that may have introduced the bias. An experimenter using an established apparatus therefore need not necessarily generate control data. Control studies in psi research are also fundamentally problematic. If one accepts the possibility of psychic functioning, the “unconscious influence [of the experimenter] can affect and therefore contaminate” control data in general (Rhine L.E., 1970, p. 254).

The split of the 117 experimental reports into studies led to the corpus of 380 experimental and 137 corresponding control studies that were used in the meta-analysis.

4.1.8.4 Coding Studies

The variables coded covered six main areas: (a) *Basic information*, which included study ID number, name of coder, name of first author, year of publication, short description of experimental condition, study status (i.e., formal, pilot, mixed, control), psychological test used (i.e., no, yes--for information, yes--to split participants into groups, yes--but no results reported), use of established psychological test (i.e., yes, no, other), name of psychological test, was the psychological test taken before experiment (i.e., yes, no, other), comments regarding psychological testing procedure, systematic state manipulation (i.e., no, yes, other), was state manipulation verified (i.e., yes, no, other), description of the state manipulation procedure, comments regarding state manipulation, control data accumulated (i.e., during experiment, before/after experiment, during and before/after experiment, other), feedback during accumulation of control data (i.e., yes, no, other), and comments regarding control data; (b) *Participant information*, which included participant type (i.e., adults, students, adults/students, 13-18 year olds, 6-12 year olds, pre-school infants/babies, animals, plants, other), species of animal/plant, participant selection (i.e., volunteer paid, volunteer unpaid, semi-volunteer, non-volunteer, experimenter, mixed, other), selection criteria (i.e., none, psychic claimant, prior success in psi experiment, psychological test, prior psychic experiences, practicing meditation/yoga, other), number of participants, and comments regarding participant information; (c) *Experimenter information*, which included experimenter also participant (i.e., yes, no, partially, other), affiliation of first author, experimenter in room with participant (i.e., yes, no, experimenter was participant, sometimes, other), and initiating individual trial/run (i.e., experimenter, participant, mixed, automatic, other); (d) *Experimental setting*, which included participation (i.e., individually, pairs, group, not systematic, other), experimental definition of experiment (i.e., PK, retro-PK, precognition, clairvoyance, covert psi, mixed, other), participants' understanding of experiment (i.e., PK, retro-PK, precognition, clairvoyance, mixed, other), participant informed about RNG (i.e., no, some details, detailed information, other), direction of intention (i.e., one direction, balanced, other), intention chosen by (i.e., experimenter, participant, prespecified, randomized, other), RNG type (i.e., radioactive, noise, mixed with pseudo RNG, other), what type if mixed with pseudo RNG (i.e., radioactive, noise, other), type of feedback (i.e. visual, auditory, other), timing participant feedback (i.e., bit by bit, trial by trial, end of run, end of session, end of experiment, false feedback, mixed, other), timing experimenter feedback (i.e., experimenter first, participant first, experimenter and participant receive feedback at the same time, mixed, other), and comments regarding experimental setting; (e) *Statistical information*, which included number of bits (per trial), number of bits (per second), number of random events technically generated by RNG (per second), number of bits (per run), number of trials (per run), number of runs (per session), number of bits (per session), number of sessions, total number of bits (sample size), duration of one trial (in seconds), duration of one session (in seconds), theoretical probability of a hit, observed probability of a hit, *z*-score, total number of starting points ("button pushes" during experiment), and comments regarding statistical information; and (f) *Safeguard variables*, which are described in some detail. *RNG control* coded whether any malfunction of the RNG had been ruled out by the study, either by using a balanced design or by performing control runs of the RNG; *all data reported* coded whether the final study size matched the planned size of the study or whether optional stopping or selective reporting may have occurred; *split of data* coded whether the split of data reported was explicitly planned or was potentially post-hoc.

The safeguard variables were ranked on a three point scale (yes [2], earlier[8]/other[1], no[0]) with the intermediate value being coded either when it was unclear whether the study actually took the safeguard into account or where it was only partially taken into account. Because summary scores of safeguard variables are problematic if considered exclusively (e.g., Jüni, Witschi, Bloch, & Egger, 1999), we examined the influence of the safeguard variables both separately and in conjunction with each other.

The Microsoft-Access-based coding form contained 59 variables altogether, and was the result of extensive discussions among the authors and researchers specialized in RNG research via an electronic forum. All variables suggested by previous literature reviews were coded (Gissurarson, 1992 & 1997; Gissurarson & Morris, 1991; Schmeidler, 1977). However, no study was coded for all 59 variables. Control studies for example, were coded only in respect to some basic and statistical information provided, and details about psychological tests that were applied were coded only when such a test was actually used in the experiment. Several of the variables permitted the inclusion of additional comments, which were used to record extra information that may be important for the understanding of the study. This comprehensive coding strategy was applied to obtain a detailed overview of the database as a whole and because, prior to coding the studies, it was not clear which variables would provide enough data for a sensible moderator variable analysis. However, because of the importance of the safeguard variables, i.e., the moderators of quality, we prespecified

[8] When authors referred to previous studies in which the RNG was tested, studies were coded as controlled "earlier".

that the impact of the three safeguard variables would be examined independently of their frequency distribution and that all other variables would be analyzed if at least 50% of the studies could be coded.[9] This procedure was prespecified prior to the coding of the studies.

To save resources only reports for which the main coder (FS) was unclear about how to code at least one variable were double-coded. The second independent coder (EB) was blind to the coding of the main coder. A total of 17 reports (134 studies) were double coded. There was an 87.5% agreement regarding the split of reports into studies, a 73.5% to 87.8% agreement about the basic information variables, a 76.5% to 92.9% agreement about the statistical information, and a 73.4% to 88.8% agreement regarding the safeguard variables. In respect of all other variables the agreement ranged from 69.4% to 92.9%. All differences between the coders were resolved by consulting HB, who made the final decision. These double-coded studies represent those that were more difficult to code than the average study. The intercoder reliability results can therefore be considered as conservative estimates.

4.1.8.5 Analyses

The effect sizes of individual studies were combined into composite mean weighted effect size measures using an intuitively comprehensible effect size measure suggested by Rosenthal & Rubin (1989) for one-sample data. For π, a proportion index (*pi*), the number of alternative choices available is *k*, with *P* as the raw proportion of hits.

$$\pi = \frac{P(k-1)}{1+P(k-2)} \tag{1}$$

The proportion index expresses hit rates of studies with different hit probabili-

[9] Variables which are rarely reported are generally problematic because it is unclear whether they are just rarely implemented in experiments or whether they are reported only when they are found to produce a significant correlation. The number of bits per trial, the number of bits per run, the number of trials per run, the number of runs per session, the number of bits per session and the number of sessions were coded purely to calculate and/or countercheck the total number of bits accumulated (sample size). Some of the more technical details, such as the duration of one session or the duration of one trial, were often not reported.

ties according to the hit rate of an equally likely two alternative case like for example coin flipping (with a fair coin). Thus, if head in a coin flipping experiment (k = 2) wins at a hit rate of 50%, the effect size π = .50 indicates that heads and tails came down equally often; if the hit rate for heads were 75%, the effect size would be π = .75. An RNG (or dice) experiment with a 1/6 hit rate (k = 6) thus also converts to π = .50, the mean chance expectation (*MCE*) of π. The range of π, like the range of all probability measures, is from 0 to 1. With k = 2, that is in the two alternatives case, formula (1) reduces to $\pi = P$.

Following Rosenthal & Rubin (1989), the standard error of π ($SE_{(\pi)}$) was calculated based on a large-sample normal approximation on the basis of the common values P and π, and the total number of trials per experiment, N.

$$SE_{(\pi)} = \frac{\pi(1-\pi)}{\sqrt{N * P(1-P)}} \qquad (2)$$

It is crucial to understand that in contrast to meta-analyses in psychology and medicine N, that is the number of independent data points, refers to the number of bits accumulated in a RNG study and *not* the number of participants.[10] The precision of RNG studies depends only on the number of bits accumulated and not on the number of participants. Several studies (n = 36) did not even provide the number of participants and only very few studies with more than one participant included data on a participant level. Figure 1 illustrates that several studies with comparatively many participants fell far outside the expected range of the funnel plot. All these studies were based on small samples in terms of bits accumulated (Q1) and therefore their effect size estimates are not very accurate. On the other hand, none of the large-scale studies in terms of bits accumulated (Q4) appeared visually to depart from *MCE*.

In order to combine effect sizes from different studies a fixed effects model (FEM) as well as a random effects model (REM) was calculated. The mean ef-

[10] Actually none of the meta-analyses in parapsychology has so far made use of the number of participants as independent data points. Although for some experimental approaches the number of participants and the number of trials, that is the number of attempts to guess correctly or to influence a target system, might be linear, for RNG experiments the correlation between the number of bits accumulated and the number of participants is not linear (r(344) = -.02, p = .75) but rather exponential (r(344) = .18, p = .001).

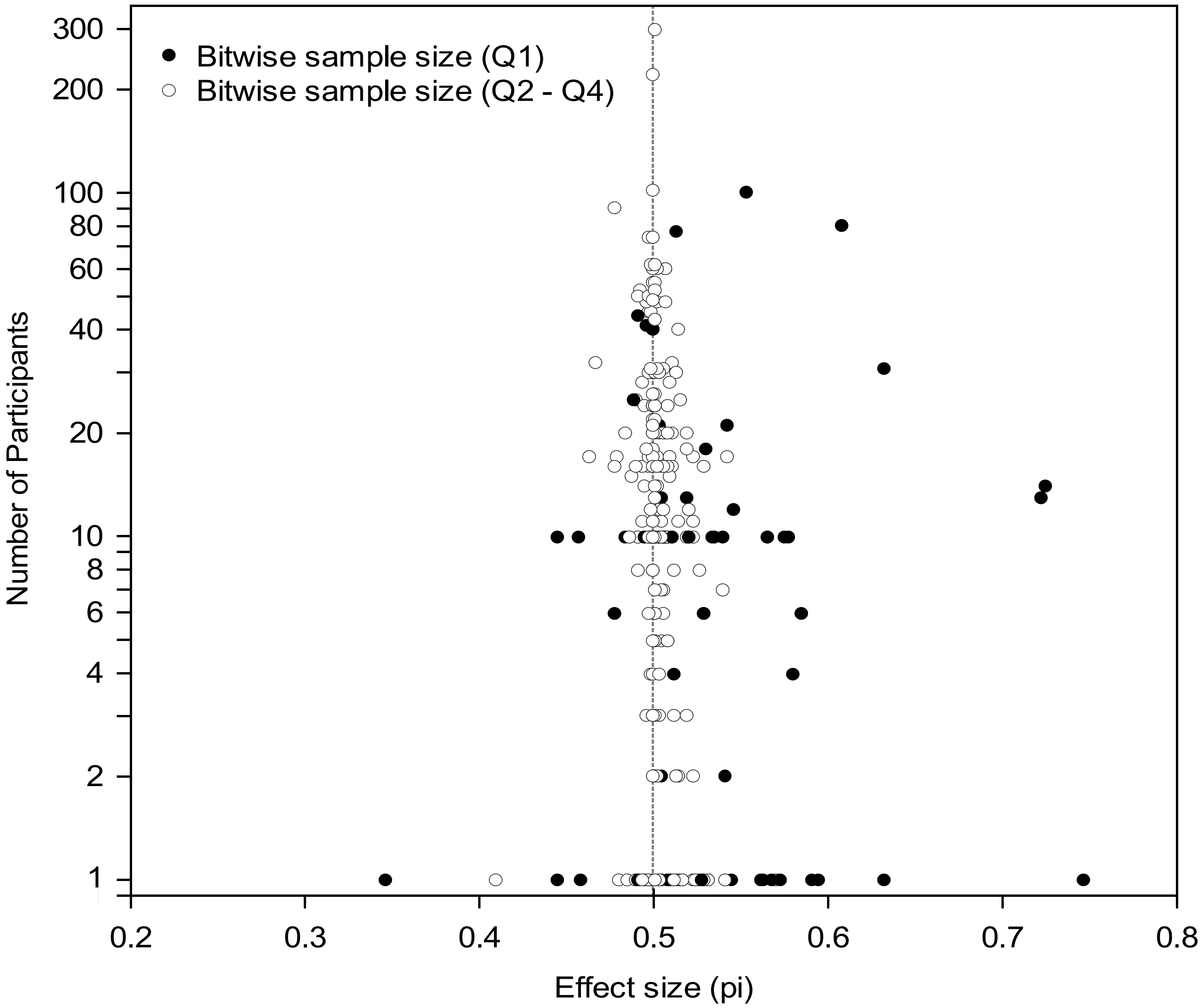

Figure 1. Funnel plot intentional studies in respect of the number of participants, The funnel shape of the graph is more evident when the number of participants is plotted using a linear scale. However, using a logarithmic scale stretches the graph in the lower part (few number of participants) and demonstrates that the large effect sizes come from the studies with the smallest sizes in terms of the number of bits accumulated (Q1, n = 95), which is the appropriate measure of sample size for the studies analyzed here. None of the large scale studies (Q4, n = 94), independently of the number of participants (range = 1-299), appear to depart visibly from the centre line (range π= 0.495-0.504).

fect size ($\overline{\pi}$) of the FEM was computed by weighting each effect size by the inverse of the variance (w_i), where m is the number of effect sizes (e.g., Hedges, 1994).

$$\bar{\pi} = \frac{\sum_{i=1}^{m} w_i \pi_i}{\sum_{i=1}^{m} w_i} \tag{3}$$

$$w_i = \frac{1}{SE^2_{(\pi_i)}} \tag{4}$$

To determine whether a sample of πs shared a common effect size (i.e., was consistent across studies), a homogeneity statistic Q was calculated, which has an approximately χ^2 distribution with m - 1 degrees of freedom (Shadish & Haddock, 1994).

$$Q = \sum_{i=1}^{m} \left(\frac{\pi_i - \bar{\pi}}{SE_{(\pi_i)}} \right)^2 \tag{5}$$

On the basis of the standard error of the combined effect sizes $SE_{(\bar{\pi})}$ a z-score statistic was used to determine the statistical significance of the combined effect sizes (e.g., Hedges, 1994).

$$SE_{(\bar{\pi})} = \frac{1}{\sqrt{\sum_{i=1}^{m} w_i}} \tag{6}$$

$$z = \frac{\bar{\pi} - 0.5}{SE_{(\bar{\pi})}} \tag{7}$$

The REM was estimated taking into account the variance between-studies ($\hat{v}_\theta$) in addition to within-study variance ($SE^2_{(\pi_i)}$) accounted for by the FEM (Shadish & Haddock, 1994).

$$v_i{}^* = SE^2_{\pi_i} + \hat{v}_\theta \tag{8}$$

$$\hat{v}_\theta = \frac{Q - (m-1)}{\sum_{i=1}^{m} w_i - \left(\sum_{i=1}^{m} w_i^2 \Big/ \sum_{i=1}^{m} w_i\right)} \tag{9}$$

To compute the REM, the total variance parameter ($v_i{}^*$) replaced the within study variance parameter ($SE^2_{(\pi_i)}$) in the equations 3-5. The z-score statistic of the REM converts accordingly (equations 6-7).

Generally the result of the homogeneity statistic is considered crucial in respect of the appropriateness of the statistical model applied. However, a nonsignificant Q value does not guarantee the adequacy of a FEM, and nor does a significant Q value guarantee the adequacy of a REM (e.g., Lipsey & Wilson, 2001). There might be a considerable between-studies variance, suggesting a REM. But this variance may not necessarily be the result of a known or unknown experimental moderator variable; for example, it could be due to publication bias[11] (as our simulation will demonstrate). That is, although theoretically studies should distribute homogeneously, they do not have to and consequently the more conservative REM is more appropriate. We therefore provide both estimates and several other sensitivity measures in order to put the data into perspective.

To determine whether the difference between two independent fixed effect size estimates ($\overline{\pi}_1, \overline{\pi}_2$) is significant, a z-score was calculated.

$$z_{\Delta} = \frac{(\overline{\pi}_1 - \overline{\pi}_2)}{\sqrt{SE_1^2 + SE_2^2}} \qquad (10)$$

The difference between two random effect size estimates was computed using the relevant effect size and the total variance parameters (equation 8).

To explore the putative impact of moderator and safeguard variables on the effect size and to determine sources of heterogeneity, two meta-regression analyses were carried out. Meta-regression is a multivariate regression analysis with independent studies as the unit of observation (e.g., Hedges & Vevea, 1998; Thompson & Higgins, 2002; Thompson & Sharp, 1999). We applied a fixed effects as well as a random effects weighted regression analysis with the moderator variables as predictors and effect size as the dependent variable adjusted as described by Hedges & Olkin (1985). Two regression models were calculated. In the Regression Model 1, sample size, year of publication and number of participants entered as continuous variables. All other variables were dummy coded. In the Regression Model 2 sample size was categorized in quartiles. All other variables entered the model according to Regression Model 1.

To illustrate the effect size distribution of studies a funnel plot was used. Three

[11] Mathematically publication bias can be considered a moderator variable, from the perspective of a meta-analyst publication bias is very different from moderators like study quality, experimental setup or participant characteristics.

approaches were taken to examine the hypothesis that the effect size distribution in the funnel plot was symmetrical, that is to test the hypothesis that the effect size was independent of sample size indicating that the sample of studies was not affected by publication or other biases (see Discussion). First, the sample was split into quartiles of sample size. Second, and on the basis of Begg & Mazumdar's (1994) approach, a rank correlation between effect size and sample size was performed. Third, Duval & Tweedie's (2000) trim and fill approach was used to estimate the number of studies causing the asymmetry (trim) and to examine the impact of these studies on the overall effect size (fill). As suggested by Duval & Tweedie (2000), we used the L_0 estimator to obtain the number of studies to be trimmed.

In an attempt to examine publication bias we ran a Monte Carlo simulation based on Hedges (1992) stepped weight function model and simulated a simple selection process. According to this model, the authors', reviewers', and editors' perceived conclusiveness of a p-value is subject to certain "cliff effects" (Hedges, 1992) and this impacts on the likelihood of a study getting published. Hedges (1992) estimates the weights of the step function based on the available meta-analytical data. However, different from Hedges, we used a predefined step-weight function model, because we were primarily interested in seeing whether a simple selection model may in principle account for the small-study effect found.

We assumed that 100% of studies (weight) with a p-value $\leq .01$ (step), 80% of studies with a p-value between $p \leq .05$ and $p > .01$, 50% of studies with a p-value between $p \leq .10$ and $p > .05$, 20% of studies with a p-value between $p \leq .50$ and $p > .10$ and 10% of studies with p-value $> .50$ (one-sided) are "published".[12] Starting with these parameters, we randomly generated uniformly distributed p-values and calculated the effect sizes for all "published" studies and counted the number of "unpublished" studies. That is, for every

[12] The term published is used here very broadly to include publications of conference proceedings and reports which in terms of our literature search were considered unpublished. Importantly, in our discussion of the Monte Carlo simulation, the term "published" also refers to studies obtained by splitting experimental reports into studies. For simplicity, we assumed in the Monte Carlo simulation that the splitting of the 117 reports into 380 experimental studies was subject to the same selection process as the publication process.

study, one random process was used to generate the study's *p*-value and another random process was used to generate its corresponding "limit value" (0-100%). A simulated study with a *p*-value > .50 needed at least to pass the limit value of 90% to be "published". For an "unpublished" study, that is, a study that did not pass the limit value, the whole process started over again with simulating the study's *p*-value. This means that, on the basis of the sample size for each of the 380 studies included in our meta-analysis, we simulated a selective null-effect publication process.

All primary analyses were performed using SPSS (Vers. 11.5) software. The standard meta-analytical procedures not implemented in SPSS were programmed on the basis of available SPSS macros (Lipsey & Wilson, 2001). The trim and fill procedure was performed with STATA (Vers. 6.0) using user-written STATA commands (from the STATA homepage).

4.1.9 Results

4.1.9.1 Study Characteristics

The basic study characteristics are summarized in Table 3. The heyday of RNG experimentation was in the 1970s, when more than half of the studies were published. A quarter of the studies were published in conference proceedings and reports, but most of the studies were published in journals. The number of participants per study varied considerably. Approximately one quarter of studies were conducted with a sole participant and another quarter with up to 10 participants. There were only three studies with more than 100 participants. The sample size of the average study is 787,888,669 bits. However, most studies were much smaller, as indicated by a median sample size of 8,596 bits (see Table 4). Some very large studies considerably increased the average sample size and resulted in an extremely right-skewed distribution of sample size. This variable was therefore log10-transformed. Consequently, a significant linear correlation or regression coefficient of sample size with another variable indicates an underlying exponential relationship. The 117 experimental reports were published by 59 different first authors affiliated with 33 different institutions.

Table 3 Basic Study Characteristics - Intentional Studies

	Studies (*n*)		Studies (*n*)
Source of studies		Year of publication	
Journal	277	≤ 1970	14
Conference proceeding	68	1971 - 1980	199
Report	25	1981 - 1990	111
Thesis/Dissertation	8	1991 - 2000	40
Book Chapter	2	2001 - 2004	16
Number of participants		Sample size (bit)	
1	96	> 10^1 - 10^2	10
> 1 - 10	107	> 10^2 - 10^3	62
> 10 - 20	61	> 10^3 - 10^4	130
> 20 - 30	34	> 10^4 - 10^5	93
> 30 - 40	12	> 10^5 - 10^6	41
> 40 - 50	13	> 10^6 - 10^7	19
> 50 - 60	10	> 10^7 - 10^8	17
> 60 - 70	2	> 10^8 - 10^9	5
> 70 - 80	4	> 10^9	3
> 80 - 90	1		
> 90 - 100	1		
> 100	3		

4.1.9.2 Overall Effect Size

When combined, the overall result of the 380 intentional studies depended on the statistical model applied. The overall effect size of the FEM indicates an effect opposite to intention whereas the effect size of the REM indicates an effect in the intended direction (see Table 4). The considerable difference between the two models was due to the three by far largest studies in the meta-analysis (see

Table 4 Overall Sample Summary Statistics

		Fixed Effects Model (FEM)			**Random Effects Model (REM)**						
Sample	***n***	$\overline{\pi}$	***SE***	***z***	$\overline{\pi}$	***SE***	***z***	***M*** **bit**	***Mdn*** **bit**	***M*** **py**	***Q***
Overall	380	.499997	.000001	-3.67***	.500035	.000014	2.47*	787888669	8596	1981	1508.56***
Overall (-3 largest)	377	.500048	.000013	3.59***	.500286	.000070	4.08***	3707412	8039	1981	1489.99***

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$

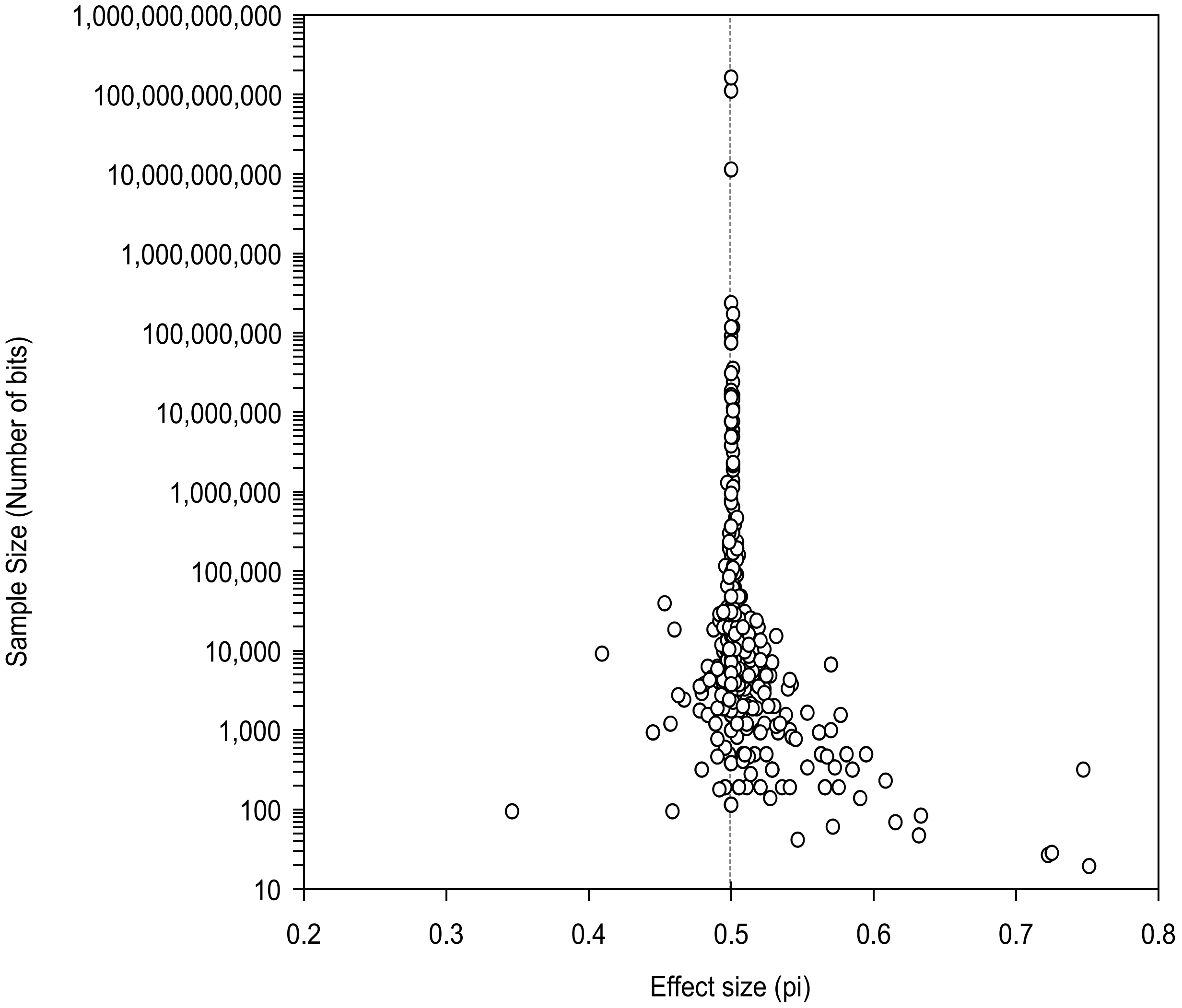

Figure 2. Funnel plot intentional studies.

Figure 2), published in a single experimental report (Dobyns, Dunne & Nelson, 2004). The effect sizes of the three studies ranging from π = .499989 to π = .499997 indicate a result opposite to intention. Without the three studies, both models show a statistically highly significant effect in the intended direction (see Table 4).

When cumulatively calculating the FEM, starting with the smallest study in the sample (n = 20, π = .75) and consecutively adding the next largest study to the sample, the overall effect size of the FEM became progressively closer to the theoretical mean value of $\bar{\pi}$ = .50. The cumulative analysis became opposite to the direction of intention ($\bar{\pi}$ < .50) at the very point where the first of the three largest studies was added to the cumulative sample. However, even as each of the final three studies was added, the overall effect size approached closer and closer to the theoretical mean value.

The studies in the meta-analysis had an extremely heterogeneous effect size distribution ($Q(380) = 1508.56$, $p = 2.07 * 10^{-141}$) and remained extremely heterogeneous even when the three largest studies were removed from the sample ($Q(377) = 1489.99$, $p = 2.50 * 10^{-138}$). This heterogeneity may be the reason for the large difference in effect size between the FEM and REM. Even when the three largest studies are removed, the difference between the two models is highly significant ($z_\Delta = 3.34$, $p = 0.0008$).

Data for one or more control studies were provided in approximately one-third of the reports ($n = 45$). The total of 137 control studies yielded a nonsignificant effect size ($\bar{\pi} = .499978$, $SE = .000015$, $z = -1.51$, $p = .13$). The effect sizes for the FEM and the REM were identical because the control data distributed homogeneously ($Q(136) = 136.34$, $p = .60$). With a median sample size of 50,000 bits and a mean sample size of 8,441,949 bits, the control studies were large in comparison with the intentional studies (see Table 4).

4.1.9.3 Safeguard Variable Analyses

The simple overview of study quality revealed that the quality of studies was high. In the FEM, for each safeguard variable, the effect size of studies with the highest quality rating pointed in the opposite direction to intention (see Table 5). However, when the three largest studies were removed, the effect size for all variables (FEM) showed an effect in the direction of intention and was in good agreement with REM analyses.

Both fixed and random effects analyses suggested that the effect sizes of studies implementing *RNG controls* were similar to those that did not implement the safeguard (FEM: $z_\Delta = -.22$, $p = .82$; REM: $z_\Delta = -1.60$, $p = .11$). Similarly, studies that *reported all data* did not have different effect sizes from studies that did not report all the data (FEM: $z_\Delta = -.76$, $p = .45$; REM: $z_\Delta = -.41$, $p = .68$). When the three largest studies were removed from the FEM analyses, the high quality studies became statistically significant in the intended direction. The difference between the studies implementing RNG controls and those that did not implement the safeguard (FEM: $z_\Delta = .07$, $p = .94$; REM: $z_\Delta = -1.31$, $p = .19$) as well as

Table 5 Safeguard Variables Summary Statistics

		Fixed Effects Model (FEM)			**Random Effects Model (REM)**						
Sample	***n***	$\bar{\pi}$	***SE***	***z***	$\bar{\pi}$	***SE***	***z***	***M* bit**	***Mdn* bit**	***M* py**	***Q***
RNG control											
Yes (2)	269	.499997$_a$	.000001	-3.67	.500029	.000012	2.32*	111261910	12288	1983	911.68***
Earlier (1)	7	.499996	.000051	-0.08	.521295	.993298	6.46***	13471208	1000	1982	286.75***
No (0)	104	.500038	.000188	0.20	.501101	.000668	1.65*	85177	4838	1977	310.09***
All data reported											
Yes (2)	311	.499997$_a$	.000001	-3.68	.500033	.000014	2.32**	962583297	8192	1982	1405.71***
Unclear (1)	11	.501074	.000537	2.00*	.500927	.000882	1.05	80726	37000	1976	16.75
No (0)	58	.500063	.000087	0.72	.500101	.000163	0.62	575876	7750	1980	81.50
Split of data											
Preplanned (2)	253	.499997$_b$	.000001	-3.46	.500012$_a$	.000016	0.74	113250870	10000	1982	761.78***
Unclear (1)	50	.500060	.000017	3.54***	.500105	.000067	1.58	17356282	19000	1982	167.74***
Post hoc(0)	77	.499989$_a$	.000005	-2.37	.504052	.000745	5.54***	155911422	4600	1979	562.36***

(table continues)

Table 5 (continued)

		Fixed Effects Model (FEM)			**Random Effects Model (REM)**			***M***	***Mdn***	***M***	
Sample	***n***	$\bar{\pi}$	***SE***	***z***	$\bar{\pi}$	***SE***	***z***	**bit**	**bit**	**py**	***Q***
Safeguard sum-score											
Sum = 6 (highest)	159	.499997[b]	.000001	-3.47***	.500007[a]	.500007	0.47***	1801262569	11360	1984	479.52***
Sum = 5	47	.500054	.000016	3.36***	.500132	.000069	1.93***	20402900	48000	1983	206.02***
Sum = 4	106	.499989[b]	.000005	-2.36***	.500472[a]	.000292	1.61***	113487404	6400	1979	405.62***
Sum = 3	8	.515664	.002616	5.99***	.544965	.511953	2.67***	4635	2880	1978	224.87***
Sum = 2	44	.499910	.000297	-0.30***	.501504	.001075	1.40***	72014	3146	1977	130.55***
Sum = 1	9	.500000	.000250	0.00***	.500000	.000250	0.00***	445209	1600	1976	.00***
Sum = 0 (lowest)	7	.500398	.000470	0.85***	.502072	.001267	1.63***	161714	25000	1979	9.88***

[a] With the three largest studies removed from the sample, the effect size is significantly larger ($p < .05$, $z > 1.96$) than *MCE*.

[b] With the three largest studies removed from the sample, the effect size is larger than .50 (*MCE*), but not significantly so.

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

the difference between the studies that reported all data and those that did not report all the data (FEM: $z_\Delta = -.18$, $p = .86$; REM: $z_\Delta = 1.17$, $p = .24$) remained non significant.

The *split of data* was reported to be preplanned for almost three quarters of the studies, indicating that "fishing for significance" did not occur in most of the studies in the meta-analysis. In the FEM, the 253 studies with their split of data preplanned yielded a highly significant effect opposite to intention. When the three largest studies were removed, the effect size of the studies which had preplanned their split of data was significantly smaller than that of the studies with a post-hoc split ($z_\Delta = 2.46$, $p = 0.01$). This finding was mirrored in the REM, where, again, studies with a preplanned split had a considerably smaller effect size than studies with a post-hoc split ($z_\Delta = 5.42$, $p = 6.01 * 10^{-8}$). These results indicate that post-hoc splitting of data (artificially) increases effect size.

The sum-score of safety variables indicated (see Table 5) that the majority of studies had adequately implemented the specified safeguards. More than 40% of the studies ($n = 159$) were given the highest rating for each of the three safeguards. The mean rating was 4.6 (*Mdn* = 5). However, there was a small but significant correlation between effect size and safeguard sum-score ($r(380) = .15$, $p = .004$) indicating that lower quality studies produced larger effect sizes. As indicated in Table 5, study quality was also positively correlated with year of publication ($r(380) = .29$, $p = 8.27 * 10^{-9}$) and sample size ($r(380) = .19$, $p = .0002$), that is, high quality studies had larger sample sizes and were conducted more recently. However, although the correlations were all significant, they were small and must be seen against the fact that the average study quality was very high.

4.1.9.4 Moderator Variable Analyses

Other than sample size and year of publication, few other moderator variables provided enough entries for us to be able to carry out sensible analyses. For instance, 112 studies were coded as having used psychological questionnaires. This was less than a quarter of the studies in our sample. Moreover, only 22 studies used established measures. Beside sample size and year of publication, we analyzed five additional central moderator variables.

Table 6 Moderator Variables Summary Statistics.

		Fixed Effects Model (FEM)			**Random Effects Model (REM)**			***M***	***Mdn***	***M***	
Sample	***n***	$\overline{\pi}$	***SE***	***z***	$\overline{\pi}$	***SE***	***z***	**bit**	**bit**	**py**	***Q***
Sample size (bit)											
(Q1) Smallest	95	.519908	.002070	9.61^{***}	.525523	.004616	5.23^{***}	641	490	1978	393.31^{***}
(Q2) Small	95	.506320	.000788	8.02^{***}	.505900	.001541	3.83^{***}	4726	4900	1979	333.86^{***}
(Q3) Large	96	.502087	.000362	5.76^{***}	.502355	.000703	3.35^{***}	21833	20034	1980	331.69^{***}
(Q4) Largest	94	$.499997_a$	.000001	-3.70^{***}	$.500009_a$	.000008	1.19^{***}	3185054132	727620	1989	259.46^{***}
Year of publication											
(Q1) Oldest	99	.505342	.000393	13.60^{***}	.511509	.001505	7.65^{***}	17578	3000	1972	719.66^{***}
(Q2) Old	96	.500194	.000148	1.31^{***}	.500811	.000369	2.20^{***}	119912	6800	1979	185.03^{***}
(Q3) New	103	.500382	.000115	3.33^{***}	.500702	.000307	2.28^{***}	187156	12288	1983	230.00^{***}
(Q4) Newest	82	$.499997_a$	.000001	-3.73^{***}	.500003	.000006	0.47^{***}	3650794697	380000	1996	175.69^{***}

(table continues)

Table 6 (continued)

		Fixed Effects Model (FEM)			Random Effects Model (REM)						
Sample	***n***	$\overline{\pi}$	***SE***	***z***	$\overline{\pi}$	***SE***	***z***	***M*** **bit**	***Mdn*** **bit**	***M*** **py**	***Q***
Number of participants											
(Q1) One (1)	96	.500499	.000130	3.84^{***}	.503208	.000610	5.26^{***}	171288	7640	1981	644.17^{***}
(Q2) Few (2-10)	107	$.499995_{b}$	.000001	-3.53^{***}	$.500025_{a}$	.000030	0.83	1216285332	5000	1980	339.94^{***}
(Q3) Several (11-20)	61	$.499997_{b}$	.000001	-2.07^{*}	.500190	.000164	1.16	2755175923	12288	1981	169.39^{***}
(Q4) Many (21-299)	80	.500033	.000015	2.14^{*}	.500001	.000043	0.03	13026064	22446	1984	140.90^{***}
Unknown	36	.500123	.000044	2.80^{**}	.500453	.000180	2.51^{*}	3636208	17875	1984	183.66^{***}
Participants											
Selected	59	.500603	.000151	3.99^{***}	.506450	.000939	6.87^{***}	187290	8000	1977	578.98^{***}
Unselected	261	$.499997_{a}$	.000001	-3.69^{***}	$.500020_{a}$	.000011	1.84	1147069802	15057	1982	720.20^{***}
Other	60	.500408	.000422	0.97	.504691	.001308	3.59^{***}	23761	1280	1981	183.34^{***}

(table continues)

Table 6 (continued)

		Fixed Effects Model (FEM)			Random Effects Model (REM)						
Sample	***n***	$\bar{\pi}$	***SE***	***z***	$\bar{\pi}$	***SE***	***z***	***M* bit**	***Mdn* bit**	***M* py**	***Q***
Study status											
Formal	209	.499997[a]	.000001	-3.31***	.500024	.000013	1.84	1374014360	12000	1982	668.85***
Pilot	160	.499990[b]	.000005	-2.17*	.500493	.000141	3.50***	76366304	7350	1980	813.15***
Other	11	.500325	.000157	2.07*	.500505	.000481	1.05	916957	7926	1979	23.09*
Feedback											
Visual	227	.500030	.000016	1.81	.500228	.000092	2.48*	4149925	6400	1980	845.78***
Auditory	34	.502377	.000382	6.22***	.505422	.001392	3.90***	51695	18100	1976	253.38***
Other	119	.499997[a]	.000001	-3,79***	.500009	.000011	0.83	2508015996	20000	1986	366.54***
Random sources											
Noise	228	.499997[a]	.000001	-3.68***	.500026	.000012	2.13*	1313136638	18375	1985	913.03***
Radioactive	93	.503354	.000601	5.58***	.509804	.001778	5.51***	8339	2000	1974	467.69***
Other	59	.500945	.000382	2.48*	.501562	.000633	2.47*	29920	13600	1979	93.41**

[a] With the three largest studies removed from the sample, the effect size is significantly larger ($p < .05$, $z > 1.96$) than *MCE*.

[b] With the three largest studies removed from the sample, the effect size is larger than .50 (*MCE*), but not significantly so.

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Table 6 provides the mean effect sizes associated with sample size, year of publication and the five central moderators. Here too, as with the safeguard variables, in the FEM, any subsample containing at least one of the three largest studies had an effect that was reversed to that of one opposite to intention. This illustrates well that *sample size* is the most important moderator of effect size. Because studies were weighted (according to the inverse of the variance), the three by far largest studies, which also had the smallest effect sizes and a direction opposite to that of the rest of the database, had a large influence on any subsample effect size in which they were included. Consequently, it is important not to place too much emphasis on the apparent reversal of direction in any subsample that includes one or more of the three largest studies. Quite generally, for each moderator, the subsample with the largest sample size is, with only one exception (REM, number of participants Q4), always associated with the smallest effect size[13] (see Table 6). Conversely, studies in the quartile with the smallest studies (Q1) have an effect size that is four orders of magnitude larger than the effect size in the quartile with the largest studies (Q4). The difference is highly significant regardless of whether the FEM or the REM is used and regardless of whether the three largest studies are included or removed from the sample ($z_\Delta > 5.00$, $p < 5.74 * 10^{-7}$). The trend is continuous: the smaller the sample size, the bigger the effect size. Sterne, Gavaghan, & Egger (2000) called this the “small-study effect”. The funnel plot (see Figure 2) illustrates the effect. Whereas the bigger studies distribute symmetrically round the overall effect size, the distribution of studies below 10,000 bits is increasingly asymmetrical.

In respect of the mean year of publication, the largest studies (Q4) stand out from the other three, smaller-study quartiles. The largest studies are, on average, published 9-11 years later than the smaller studies. Most of the big studies, with very small effect sizes, have been published only recently (e.g., Dobyns, Dunne & Nelson, 2004; Jahn et al., 2000; Nelson, 1994).

[13] The smallest effect size is the effect size closest to the theoretical mean value of $\overline{\pi} = .50$. When the three largest studies were removed from the analyses, the subsample with the largest sample size generally still had the smallest effect size, with the same exception (Q4 in the variable *number of participants*) as when the three largest studies were included.

The *year of publication* underpins the importance of sample size for the outcome of the studies (see Table 6). The oldest studies (Q1), which have the smallest sample size, have an effect size that is, depending on the statistical model, at least three orders of magnitude larger than the effect size of the newest studies, which have by far the largest mean sample size of all subsamples in Table 6. The two middle quartiles show no clear cut difference in effect size (FEM: z_Δ = -1.01, p = .31; REM: z_Δ = .23, p = .82) and in sample size. Therefore sample size, and not year of publication, seems to be the important variable. To verify this we median split the subsample of oldest studies (Q4) according to sample size. The effect sizes of the two halves differ highly significantly from each other (FEM: z_Δ = 6.77, p = 1.26 * 10^{-11}; REM: z_Δ = 3.94, p = 8.29 * 10^{-5}). The half with the smaller studies (n = 49, M = 810, Mdn = 500) has a much larger effect size (FEM: $\bar{\pi}$ = .522382, SE = .002546, z = 8.79, p < 1 * 10^{-10}; REM: $\bar{\pi}$ = .536425, SE = .007216, z = 5.05, p = 4.48 * 10^{-7}) than the half with the larger studies (n = 50, M = 34011, Mdn = 9630) which has a considerably smaller effect size (FEM: $\bar{\pi}$ = .504926, SE = .000398, z = 12.38, p < 1 * 10^{-10}; REM: $\bar{\pi}$ = .507557, SE = .001312, z = 5.76, p = 8.44 * 10^{-9}). The mean year of publication in both subsamples with 1972.0 for the half with the smaller studies and 1971.4 for the half with the bigger studies is far too small to account for the difference in effect size. The analysis strongly suggests that sample size is the deciding moderator and not year of publication.

Most studies in the meta-analysis were conducted with only one or only a few (i.e., 2-10) participants (see Table 6). Although Table 6 suggests a connection between the *number of participants* and effect size, because the single participant experiments (Q1) have the largest mean effect size, no correlation was observed between number of participants and effect size (r(344) = -.05, p = .38). This correlation is not affected by the three largest studies in the sample because in terms of the number of participants used they are average (range = 3-11).

The analyses seem to support the claim that *selected participants* perform better than non-selected participants, a claim that had found support in an earlier precognition meta-analysis (Honorton & Ferrari, 1989). As can be seen in Table 6, the effect size of studies with selected participants is considerably larger than that of studies that did not select their participants, for example, on the basis of their prior success in a psi experiment or for being a psychic claimant. The difference between selected and unselected participants is highly significant (FEM: $z_{\Delta} = 4.02$, $p = 5.90 * 10^{-5}$; REM: $z_{\Delta} = 6.85$, $p < 1 * 10^{-10}$) and remains so with the three largest studies removed (FEM: $z_{\Delta} = 3.69$, $p = 2.22 * 10^{-4}$; REM: $z_{\Delta} = 6.73$, $p < 1 * 10^{-10}$). However, the two subsamples differ considerably in sample size. Studies using selected participants were considerably smaller, even when the three largest studies, which used unselected participants, were removed (Selected: $M = 187290$, $Mdn = 8000$; Unselected: $M = 5369064$, $Mdn = 13968$).

Study status is an important moderator in meta-analyses that include both formal and pilot studies. Pilot studies are likely to comprise a selective sample insofar as they tend to be published if they yield significant results (and hence have larger-than-usual effect sizes) and not published if they yield unpromising directions for further study. In this sample pilot studies are, as one would expect, smaller than formal studies. In respect of their FEM effect size, pilot and formal studies do not differ ($z_{\Delta} = 1.46$, $p = 0.15$). However, in respect of their REM effect, they differ considerably ($z_{\Delta} = -3.31$, $p = 9.17 * 10^{-4}$). When the three largest studies are removed, the picture remains the same although the effect sizes of the formal (FEM: $\bar{\pi} = .500043$, $SE = .000015$, $z = 2.96$, $p = .003$; REM: $\bar{\pi} = .500125$, $SE = .000068$, $z = 1.83$, $p = .07$) and pilot studies (FEM: $\bar{\pi} = .500061$, $SE = .000034$, $z = 1.80$, $p = .07$; $\bar{\pi} = .500701$, $SE = .000195$, $z = 3.59$, $p = 3.37 * 10^{-4}$) are larger. The results regarding the study status are not clear-cut, they depend on the chosen statistical model.

The type of *feedback* to the participant in RNG studies has been regarded as an important issue in psi research from its very inception. The majority of RNG studies provide participants with visual and some with auditory feedback. Beside these two main categories, the coding resulted in a large “other” category with 119 studies, which used, for example, alternating visual and auditory feedback or no feedback at all. The result is clear-cut: studies providing exclusively auditory feedback outperform not only the studies using visual feedback (FEM: $z_{\Delta} = 6.14$, $p = 8.49 * 10^{-10}$; REM: $z_{\Delta} = 3.72$, $p = 1.96 * 10^{-4}$), but also the studies in the “other” category (FEM: $z_{\Delta} = 6.23$, $p = 4.74 * 10^{-10}$; REM: $z_{\Delta} = 3.89$,

$p = 1.01 * 10^{-4}$). This finding changes only marginally when the three largest studies, which all belong to the "other" category, are removed from the sample. However, the finding is based on a very small and very heterogeneous sample of smaller studies (see Table 6).

The core of all RNG studies is the *random source*. Although the participants' intention is generally directed (by the instructions given to them) to the feedback and not to the technical details of the RNG, it is the sequence of random numbers produced by the random source that is compared with the theoretical expectation (binominal distribution) and that is therefore allegedly influenced. RNGs can be based on truly random radioactive decay, Zener diode, or occasionally thermal noise. As shown in Table 6, the effect size of studies with RNGs based on radioactive decay is, considerably larger than the effect size of studies using noise (FEM: $z_\Delta = 5.59$, $p = 2.28 * 10^{-8}$; REM: $z_\Delta = 5.50$, $p = 3.86 * 10^{-8}$). And although the effect size of the studies using noise becomes significantly different from *MCE* when the three largest studies, all noise based, are removed from the sample (FEM: $\bar{\pi} = .500045$, $SE = .000013$, $z = 3.39$, $p = 7.12 * 10^{-4}$; REM: $\bar{\pi} = .500174$, $SE = .000059$, $z = 2.93$, $p = .003$), the mean effect size of the studies using radioactive decay remains significantly larger than studies using noise (FEM: $z_\Delta = 5.51$, $p = 3.65 * 10^{-8}$; REM: $z_\Delta = 5.41$, $p = 5.41 * 10^{-8}$). However, this variable, too, is strongly confounded by sample size. Studies using radioactive decay are much smaller than studies using noise (see Table 6). The sample size of noise-based studies without the three largest studies remains considerably larger (M = 6200682 bit, Mdn = 17000 bit) than the sample size of the radioactive based studies. Chronologically, studies with RNGs based on radioactive decay predominated in the very early years of RNG experimentation, as indicated by their mean year of publication, which is just two years above the mean year of publication of the oldest studies in our sample (see Table 6).

4.1.9.5 Meta-Regression Analyses

The first regression model (see Table 7) accounts for 8.1% (FEM) of the variability (REM: 6.8%). Although this model is statistically highly significant (FEM: $Q = 121.76$, $df = 17$, $p = 7.11 * 10^{-18}$; REM: $Q = 99.93$, $df = 17$, $p = 9.17$

Table 7 Summary of the Weighted Meta-Regression - Model 1 (Sample Size)

	Fixed Effects Model (FEM)			**Random Effects Model (REM)**		
Variable	***B***	***SE B***	***z***	***B***	***SE B***	***z***
Sample size (log10)	.000005	.000009	0.55	-.000027	.000021	-1.29
Year of publication	-.000016	.000004	-4.24***	-.000016	.000005	-3.10**
Number of participants	-.000016	.000029	-0.54	-.000079	.000061	-1.30
Selected participants	.000950	.000525	1.81	.000989	.000528	1.87
Unselected participants	-.000055	.000427	-0.13	.000107	.000436	0.24
Formal study	.000834	.000352	2.37*	.000822	.000359	2.29*
Pilot study	.000898	.000354	2.53*	.000806	.000365	2.21*
Visual feedback	-.000046	.000035	-1.30	-.000081	.000060	-1.36
Auditory feedback	.001484	.000438	3.39***	.001423	.000444	3.21**
Noise RNG	-.000303	.000456	-0.66	-.000331	.000464	-0.71
Radioactive RNG	.002154	.000718	3.00**	.002089	.000720	2.90**
RNG control - Yes	.000165	.000074	2.24*	.000130	.000111	1.16
RNG control - No	-.000327	.000246	-1.33	-.000466	.000273	-1.71
All data reported - Yes	-.000493	.000547	-0.90	-.000427	.000554	-0.77
All data reported - No	-.000543	.000557	-0.97	-.000513	.000564	-0.91
Split of data - Preplanned	-.000008	.000038	-0.21	-.000024	.000057	-0.43
Split of data - Post hoc	-.000082	.000073	-1.12	.000001	.000123	0.01
Constant	.532077	.007413	4.33***	.532109	.010064	3.19**

Note. Regression coefficients are unstandardized and are the amount of change in effect size associated with one unit change in the predictor.
$^{*}p$ < .05. $^{**}p$ < .01. $^{***}p$ < .001.

$* 10^{-14}$) the unaccounted residual variance is considerable (FEM: Q = 1386.80, df = 362, p = 1.16 * 10^{-119}; REM: Q = 1361.73, df = 362, p = 1.22 * 10^{-115}). This indicates that important moderator variables were missed in the meta-analysis. Alternatively, if one were to assume that there is no effect of intention on the outcome of RNGs, the significant variables could also indicate that early RNG experiments using a radioactive source and auditory feedback were published only when a large effect size was found. The predominant role of sample size is nevertheless called into question. However, this regression model was based on the assumption of an exponential relationship between sample size and effect size.

The importance of sample size in the meta-analysis is demonstrated by the second regression model (see Table 8), in which sample size is categorized into quartiles. Model 2 indicates that the quartiles of sample size are by far the most important predictor of effect size. The model accounts for 15.5% (FEM) of the variability (REM: 14.4). Although this regression model is statistically highly significant (FEM: $Q = 233.45$, $df = 17$, $p = 4.93 * 10^{-40}$; REM: $Q = 212.19$, $df = 17$, $p = 1.00 * 10^{-35}$) the unaccounted residual variance again remains considerable (FEM: $Q = 1275.12$, $df = 362$, $p = 5.84 * 10^{-102}$; REM: $Q = 1262.44$, $df = 362$, $p = 4.48 * 10^{-100}$) indicating that this model cannot be considered definitive either. However, the second regression model explains twice the variance explained by the first model only because there is indeed a strong relationship between sample size and effect size.

It is evident that both regression models account for only a small proportion of the effect size variability. The meaning of the variables found to be significant predictors of effect size is not clear-cut. Regression analyses cannot establish causal connections and therefore it remains unclear whether the significant variables are predictor variables in the usual sense or whether these variables indicate that the studies were published selectively. A very small overall effect size makes it difficult for any regression analysis, or any meta-analysis or any study, adequately to assess potential moderators.

4.1.9.6 Small-Study Effect

From the distribution of effect sizes in the funnel plot (see Figure 2), and from the split of studies in sample size quartiles (see Table 6), it is evident that the smaller studies in the meta-analysis produce larger effect sizes. The highly significant negative correlation between effect size and sample size ($r_s = -.33$, $p = 4.38 * 10^{-11}$) also confirms the asymmetric distribution of effect size. Use of Duval & Tweedie's (2000) trim and fill approach found that 83 studies had to

Table 8 Summary of the Weighted Meta-Regression - Model 2 (Sample Size Quartiles)

	Fixed Effects Model (FEM)			Random Effects Model (REM)		
Variable	***B***	***SE B***	***z***	***B***	***SE B***	***z***
Sample size (log10)	-.003019	.000285	-10.58***	-.003017	.000286	-10.54***
Year of publication	-.000012	.000004	-3.23**	-.000011	.000004	-2.47*
Number of participants	-.000012	.000027	-0.44	-.000060	.000049	-1.22
Selected participants	.001190	.000525	2.27*	.001173	.000526	2.23*
Unselected participants	.000471	.000429	1.10	.000496	.000432	1.15
Formal study	.000483	.000353	1.37	.000482	.000356	1.35
Pilot study	.000535	.000354	1.51	.000526	.000358	1.47
Visual feedback	-.000052	.000028	-1.87	-.000038	.000043	-0.89
Auditory feedback	.001930	.000440	4.38***	.001924	.000443	4.34***
Noise RNG	.001093	.000475	2.30*	.001046	.000478	2.19*
Radioactive RNG	.000843	.000729	1.16	.000809	.000730	1.11
RNG control - Yes	.000138	.000073	1.91*	.000131	.000091	1.44
RNG control - No	-.000228	.000246	-0.93	-.000261	.000257	-1.01
All data reported - Yes	-.000513	.000547	-0.94	-.000523	.000551	-0.95
All data reported - No	-.000610	.000557	-1.10	-.000617	.000561	-1.10
Split of data - Preplanned	-.000026	.000037	-0.71	-.000049	.000049	-1.01
Split of data - Post hoc	-.000092	.000063	-1.45	-.000128	.000091	-1.41
Constant	.533704	.006989	4.82***	.532772	.008691	3.77***

Note. Regression coefficients are unstandardized and are the amount of change in effect size associated with one unit change in the predictor.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

be filled in so that the distribution became symmetrical ($N = 463$). However, the overall results changed only marginally when the studies were added (FEM: $\bar{\pi} = .499997$, $SE = .000001$, $z = -3.70$, $p = .0002$; REM: $\bar{\pi} = .500036$, $SE = .000016$, $z = 2.16$, $p = .03$). Without the three largest studies, the trim and fill approach found that 73 studies had to be filled in for the distribution to become symmetrical. Adding the 73 studies to the sample (*N* = 450) only marginally changed the result of the FEM (FEM: $\bar{\pi}$ = .500045, *SE* = .000014, *z* = 3.33, *p* = .0009), but the result of the REM dropped more than one standard deviation compared with the overall sample not including the three largest studies (REM: $\bar{\pi}$ = .500229, *SE* = .000084, *z* = 2.71, *p* = .007). However, although the straightforward approach cannot account for the small study effect, it does indicate how the overall picture may change by adding relatively few studies to the overall sample.

4.1.9.7 Monte Carlo Simulation

The averaged results of the simulation of 1000 meta-analyses are shown in Table 9. As can be seen, the effect sizes based on the simulation match well to the overall effect sizes found in the meta-analysis (see Table 4). Although the effect sizes in the quartile with the smallest studies came out significantly smaller than in the meta-analysis reported here, the simulated data replicate the small-study effect evident in the data (see Table 6). The heterogeneity found in the meta-analysis is replicated to only some degree by the simulation. Although the heterogeneity of all quartiles reaches statistical significance, the actual data distribute far more heterogeneously. The simulation found that a total of 1544 studies had to be "unpublished" for these results to appear, that is for every study passing the limit values ("published"), four studies did not pass the limit values ("unpublished").

Although the parameters of our step-weight function model were predefined, the results of any simulation depend on the parameters used. We assessed the sensitivity of our simulation by varying the percentage of "published" studies in the five intervals of the step function in the range of ± 10% from their initial values (when applicable). That is, simulations were run with studies in the first step ($p \leq .01$) to be "published" in 100% and 90% of the time, and with studies in the second step ($p \leq .05$ & $p > .01$) to be "published" in 90%, 80% and 70% of the time. For each of the 162 (2 * 3 * 3 * 3 * 3) possible combinations of limit values, 1000 simulations were run. Table 10 shows that although the values of the six variables vary noticeably, the overall picture in the five categories remains, independently of which initial parameters were used in the simulation.

Table 9 Stepped Weight Function Monte Carlo Simulation of Publication Bias

		Fixed Effects Model (FEM)				**Random Effects Model (REM)**					
	n	$\overline{\pi}$	***SE***	***z***	z_Δ	$\overline{\pi}$	***SE***	***z***	z_Δ	***Q***	***Stud***
Overall	380	.500001	.000001	1.29	-3.51***	.500024	.000009	2.68**	0.62	631.58***	1544
Sample size											
(Q1) Smallest	95	.511582	.002024	5.72***	2.88**	.512474	.002478	5.07***	2.49*	125.87*	389
(Q2) Small	95	.504629	.000746	6.20***	1.56	.504705	.000849	5.58***	0.68	119.01*	384
(Q3) Large	96	.502145	.000345	6.21***	-0.12	.502192	.000393	5.61***	0.20	119.47*	390
(Q4) Largest	94	.500001	.000001	1.27	-3.51***	.500009	.000005	1.70	0.02	153.68***	381

Note. z_Δ = difference between effect size of simulated and experimental data. *Stud* = number of "unpublished" studies (simulated).

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Table 10 Limit Values of the Stepped Weight Function Monte Carlo Simulation in Dependence of the Initial Weighting (± 10%)

Variable	Smallest Studies (Q1)	Small Studies (Q2)	Large Studies (Q3)	Largest Studies (Q4)	Overall Sample
$\overline{\pi}_f$					
min	.504542	.501709	.500831	.500000	.500000
max	.523420	.509265	.504289	.500002	.500002
z_f					
min	2.24	2.29	2.41	.46	.48
max	11.58	12.42	12.42	2.53	2.58
Q					
min	72.55	59.02	59.66	121.30	500.10
max	161.01	157.23	158.53	220.88	921.81
Stud					
Min	224	225	228	223	900
Max	835	835	846	824	3340
$\overline{\pi}_r$					
Min	.505130	.501769	.500862	.500003	.500008
Max	.523970	.509269	.504291	.500029	.500069
z_r					
Min	1.83	1.83	1.93	.64	.99
Max	11.20	12.39	12.39	4.02	6.06

Note. $\overline{\pi}_f$, z_f are parameter estimates based on a FEM. $\overline{\pi}_r$, z_r are parameter estimates based on a REM. *Stud* = number of "unpublished" studies (simulated).

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

The minimum values for all effect sizes and *z*-scores come from a single set of parameters (90% of $p \leq .01$; 70% of $p > .01$ & $p \leq .05$; 40% of $p > .05$ & $p \leq .10$; 10% of $p > .10$ & $p \leq .50$, 20% of $p > .50$). However, this set of parameters does not result in extreme values regarding heterogeneity (*Q*) and unpublished studies (*Stud*), although the results are almost identical (± 1%) to those of our original simulation (see Table 9).

The fit of the simulation can be improved by varying the parameters used and/or by including additional parameters. For example, additional, interdependent limit values could be introduced for studies with extremely negatively *z*-scores or extremely large sample sizes, thus increasing the heterogeneity. However, the straightforward model was introduced to examine whether a simple selection process could produce results similar to those found in the meta-analysis. It cannot prove that the results actually are a function of this or a similar process, although considering the complexity of a very long research process the fit of the model is striking.

4.1.10 Discussion

In summary, the meta-analysis revealed three main findings: (i) a very small overall effect, which, when the three largest studies were omitted, was significant and held independently of which statistical model was applied (ii) a tremendous variability of effect size and (iii) a small-study effect.

4.1.10.1 Statistical Significance

When the three largest studies are removed from the sample, the overall effect size of both statistical models is highly statistically significant and points in the direction of intention. However, when all studies are considered, the FEM effect size points significantly in the direction opposite to intention, whereas the REM effect size points in the direction of intention but only just reaches significance. Although an effect opposite to intention would also be a notable finding, the result is clearly driven by the three largest studies, which are 100 to 1000 times larger than the largest study in the rest of the database (see Figure 2) and which have effect sizes that point in the opposite direction to the other studies. Because the FEM does not take into account the between-studies variance, the (consistent) results of the three largest studies clearly affects the overall result based on the model. Of the 380 studies, 83 produced significant results in the direction intended and 23 studies produced significant results in the direction opposite to intention. In the quartile with the largest studies (Q4) 13 studies produced significant results in the direction intended and 9 studies produced significant results in the direction opposite to intention. Thus, an effect opposite to intention cannot be claimed to be a general finding of this meta-analysis. The three studies are considered to be outliers and the overall effect found in the meta-analysis to be an effect in the direction intended by the participants in the studies.

The statistical significance, as well as the overall effect size, of the combined experimental studies has dropped continuously from the first meta-analysis to the one reported here. This is partially the result of the more recent meta-analyses including newer, larger studies. However, another difference between the current and the previous meta-analyses lies in the application of inclusion and exclusion criteria. We focused exclusively on studies examining the alleged concurrent interaction between direct human intention and RNGs. All previous meta-analyses also included non-intentional and non-human studies. Although this difference might explain the reduction in effect size and significance level, it cannot explain the extreme statistical heterogeneity of the database. This topic was overlooked in the previous RNG meta-analyses.

Because of the tremendous variability of effect size it might be argued that the FEM is not adequate and therefore the findings based on this model must not be considered. However, empirically it is impossible to decide whether the model is adequate or not. As the Monte Carlo simulation demonstrated, the effect size variability could simply be the result of selective publication. No (hidden) moderator variable need be involved. If we assume that there is no effect, the FEM is certainly adequate, at least theoretically.

However, the overall *z*-score of 2.47 for the REM and the *z*-score of 4.08 with the three largest studies excluded, is also not an unambiguous result because the finding must be understood against the background of the extreme, yet unexplained, heterogeneity and the small-study effect. The effect size from the individual analyses of the moderator and safeguard variables, and the corresponding significance level were strongly related to sample size, which confounds the effect. Moreover, Duval & Tweedie’s (2000) trim and fill approach suggests that the REM *z*-score drops from $z = 4.08$ to $z = 2.71$ after adding the 73 missing studies. However, the most important finding in respect of the overall significance level is the strong agreement between the empirical and the simulated data. The overall REM *z*-score of the simulation matches the empirical *z*-score almost perfectly.

The safeguard (quality) analyses indicated that the average study quality is very high and although there is a significant correlation between effect size and lower quality, the relationship is too small to account for the overall effect, especially when the extreme heterogeneity and the small-study effect are considered to be part of the overall effect for which a comprehensive explanation must also account.

The control studies in this meta-analysis were simply used to demonstrate that the RNG output fits the theoretical premise (binominal distribution). The finding that the mean effect size of the control studies does not differ significantly from *MCE* and the finding that the control sample was homogeneous demonstrate that the RNGs were not malfunctioning.

4.1.10.2 Variability of Effect Size

There was an extreme variability of effect size in this meta-analysis. The variability does not seem to be the result of any of the moderator variables examined. None of the moderator variable subsamples was independently homogeneous, not even sample size. The Monte Carlo simulation demonstrated that effect size variability could theoretically be the result of a selection process. It also demonstrates how all three major findings might be linked. However, the heterogeneity in the meta-analysis is much greater than the heterogeneity found in the simulation. From the three major findings discussed here, the worst fit between the simulation and the empirical data is for the heterogeneity. This might be due to the highly idealized boundary conditions of the simulation. The real world publication process is certainly more complex. For example, although we have demonstrated that the effect of publication year is confounded by sample size, older studies are generally much smaller and might have been subject to a quite different selection process than newer studies. Other variables affecting the publication process might also have changed over time. However, we have only modelled a simple selection process and therefore these arguments must be considered speculative.

4.1.10.3 Small-Study Effect

For a similar class of studies it is generally assumed that effect size is independent of sample size. However, it is evident that the effect size in this meta-analysis depends strongly on sample size, as illustrated by the asymmetric distribution of effect sizes in the funnel plot (see Figure 2) and the continuous decline of effect size with increasing sample size.

Table 11 provides a list of potential sources for the small-study effect. The sources fall into three main categories (1) true heterogeneity, (2) data irregularities, and (3) selection biases. Chance, another possible explanation for a small-study effect, seems very unlikely because of the magnitude of the effect and the sample size of the meta-analysis.

4.1.10.3.1 True Heterogeneity

The larger effect sizes of the smaller studies may be due to specific differences in experimental design or setting compared with the larger studies. For instance, smaller studies might be more successful because the participant-experimenter relationship is more intense, or the routine of longer experimental series may make it difficult for the experimenter to maintain enthusiasm in the study. However, explanations such as these remain speculative as long as they are not systematically investigated and meta-analyzed.

From the moderator variables investigated in this meta-analysis, the hypotheses that smaller studies on average tested a different type of participant (selected) and used a different form of feedback (auditory) and random source (radioactive) are the most interesting. This finding is mainly the result of experiments conducted by Schmidt. He carried out 42% of the studies that had selected participants, 50% of the studies that used auditory feedback and 29% of the studies that used radioactive random sources However, our analyses showed that not only these three variables, but also all other variables considered here, are linked to sample size. None of the three variables (and no other variable) distributes homogeneously.

Empirically, true heterogeneity cannot be eliminated as a causal factor for the small-study effect, especially regarding complex interactions, which we have disregarded here. However, the heterogeneity of the moderator-variable subsamples and the clear influence of the role of sample size at all levels of analysis with all probability likely excludes true heterogeneity as the main source of the small-study effect.

4.1.10.3.2 Data Irregularities

A small-study effect may be due to data irregularities threatening the validity of the data. For example, smaller studies might be of poorer methodological quality, thereby artificially raising their effect size compared with that of larger studies. However, the average study quality is very high and although effect size is significantly correlated with study quality the correlation is too small to account for the prominent small-study effect found. Just as the significant moderator variables were unable to be the main source of the small-study effect, the same holds for the safeguard variables. Another form of data irregularity--inadequate analysis--that may explain the small-study effect assumes that smaller trials are generally analyzed with less methodological rigor and therefore are more likely to report “false-positive results”. However, the straightforward and simple effect size measure used for the studies in this meta-analysis and the one-sample approach used in those experiments excludes the possibility of inadequate or erroneous control data clouding experimental comparisons. Another potential source of data irregularity to explain the small-study effect might be that smaller studies are more easily manipulated by fraud than larger studies because, for example, fewer people are involved. However, the number of researchers that would have to be implicated over the years renders this hypothesis very unlikely. In general, none of the data irregularity hypotheses considered appears to explain the small-study effect.

4.1.10.3.3 Selection Biases

When the inclusion of studies in a meta-analysis is systematically biased in a way that smaller studies with larger *p*-values, that is, larger effect sizes, are more likely to be included than larger studies with smaller *p*-values, that is, smaller effect sizes, a small-study effect may be the result. Several well-known selection biases such as publication bias, selective reporting bias, foreign language bias, citation bias and time lag bias may be responsible for a small-study effect (e.g., Egger, Dickersin, & Smith, 2001; Mahoney, 1985).

Biased inclusion criteria refer to biases on the side of the meta-analyst. The two most prominent of these biases are foreign language bias and citation bias. Foreign language bias occurs when significant results are published in well-circulated, high-impact journals in English, whereas nonsignificant findings are published in small journals in the authors' native language. Therefore a meta-analysis including studies solely from journals in English may include a disproportionately large number of significant studies. Citation bias refers to selective quoting. Studies with significant *p*-values are quoted more often and are more likely to be retrieved by the meta-analyst. However, the small-study effect in this meta-analysis is probably not due to these biases due to the inclusion of non-English publications and a very comprehensive search strategy.

The most prominent selection bias to consider in any meta-analysis is publication bias. Publication bias refers to the fact that the probability of a study being published depends to some extent on its *p*-value. Several independent factors affect the publication of a study. Rosenthal's term "file drawer problem" (Rosenthal, 1979) focuses on the author as the main source of publication bias, but there are other issues too. Editors' and reviewers' decisions also affect whether a study is published. The time lag from the completion of a study to its publication might also depend on the *p*-value of the study (e.g., Ioannidis, 1998) and additionally contribute to the selection of studies available. Since the development of Rosenthal's "file drawer" calculation (1979), numerous other methods have been developed to examine the impact of publication bias on meta-analyses (e.g. Dear & Begg, 1992; Duval & Tweedie, 2000; Hedges, 1992; Iyengar & Greenhouse, 1988; Sterne & Egger, 2001). Most of these methods either directly or indirectly address funnel plot asymmetry, which is regarded as evidence for publication bias. Because the asymmetry is clearly related to the small-study effect, Duval & Tweedie's (2000) trim and fill ap-

Table 11 Potential Sources of the Small-Study Effect

True heterogeneity	
	Different intensity/quality
	Different participants
	Different feedback
	Different random source
	Other moderator(s)
Data irregularities	
	Poor methodological design
	Inadequate analysis
	Fraud
Selection biases	
	Biased inclusion criteria
	Publication bias
Chance	

Note. From *Investigating and dealing with publication and other biases* (p. 193), by J. A. C. Sterne, M. Egger and G. D. Smith, 2001. In: *Systematic reviews in health care: Meta-analysis in context*, edited by M. Egger, G. D. Smith and D. Altman, London: BMJ Books. Copyright by Blackwell Publishing. Adapted with permission.

proach can also be regarded as an approach to the small-study effect. However, the approach cannot be regarded as conclusive here because although it demonstrates how the overall picture changes by adding a few studies, it does not account for the small-study effect. In contrast to this, the simulation not only accounts for the small-study effect but also, at least to some degree, reveals a possible explanation for it.

4.1.10.4 Monte Carlo Simulation

The straightforward simulation is in good agreement with all three major findings of this meta-analysis and is particularly persuasive in respect of its fit with the level of effect size and of statistical significance. The small-study effect is evident and independent of the initial parameters of the simulation. Even the heterogeneity is evident, although in weaker form. However, the number of "unpublished" studies required for the fit is potentially the crucial point of contention. The initial reaction may be to think that it is unreasonable to postulate that 1500 RNG studies remain "unpublished". After all, there are very few people conducting this type of research and the funding available for conducting such experiments is miniscule.

However, during the early period of RNG experimentation, many studies may have remained unpublished. For example, J. B. Rhine, the first editor of the *Journal of Parapsychology* (inception in 1937), the leading journal for experimental work in parapsychology, believed "that little can be learned from a report of an experiment that failed to find psi" (Broughton, 1987, p. 27), a view which at that time was probably not uncommon in other research areas as well. However, from 1975, the Council of the Parapsychological Association rejected the policy of suppressing nonsignificant studies in parapsychological journals (Broughton, 1987; Honorton, 1985). The proportion of statistically significant studies ($p < .05$) dropped from 47% in Q1 (1969 - 1974) to 17% (Q2), 13% (Q3) and 10% (Q4) in the subsequent quartiles, suggesting that the policy was implemented.[14]

The number of "unpublished" studies in the simulation not only reflects the publication process but also the splitting of the 117 experimental reports into the 380 studies. We assumed that both processes are subject to the same selection process. This is certainly questionable. For example, one might assume that data from a report are split into several studies in order to demonstrate that a particular condition or variable, such as a particular type of feedback, is statistically more successful than another, even though the overall result, comprising both conditions, does not reach statistical significance. However, Table 12 clearly shows that this is not happening. The reports split into more than 10

[14] Although the change is particularly interesting for the *Journal of Parapsychology* (JP) these data are not very reliable because almost 60% (n = 47) of the studies published in the JP were published prior to 1975 (Q1). However, the overall picture, especially the dramatic drop of significant studies from Q1 to Q2, is also evident in the studies published in the JP.

Table 12 Reported Split of Studies per Published Report

		Fixed Effects Model (FEM)			**Random Effects Model (REM)**						
Study Split of Report	***n***	$\bar{\pi}$	***SE***	***z***	$\bar{\pi}$	***SE***	***z***	***M* bit**	***Mdn* bit**	***M* py**	***Q***
1	30	.499983	.000044	-0.40^{***}	.500205	.000317	0.65^{***}	4447013	8968	1985	73.23^{***}
2	66	.499998	.000045	-0.04^{***}	.500119	.000233	0.51^{***}	1842341	12517	1983	190.60^{***}
3	27	.500017	.000023	0.75^{***}	.500124	.000124	1.00^{***}	17490052	24000	1984	142.15^{***}
4	80	.501061	.000200	5.31^{***}	.503079	.000712	4.32^{***}	82438	7440	1979	442.74^{***}
5	35	.500004	.000083	0.05^{***}	.500097	.000179	0.54^{***}	1034624	30000	1985	51.34^{***}
6[1]	48	.499997[a]	.000001	-3.75^{***}	.499999[b]	.000005	-0.01^{***}	6219133881	23400	1984	102.60^{***}
7	21	.500052	.000048	1.09^{***}	.500308	.000222	1.39^{***}	5172284	24247	1982	131.53^{***}
8	16	.501382	.001491	0.93^{***}	.502627	.002727	0.96^{***}	7024	7552	1980	40.51^{***}
10	20	.510463	.003597	2.91^{***}	.514224	.009038	1.57^{***}	960	960	1972	109.58^{***}
11	11	.505180	.000731	7.08^{***}	.509037	.001890	4.78^{***}	42727	10000	1973	36.70^{***}
12	12	.527704	.010175	2.72^{***}	.527704	.101754	2.72^{***}	200	200	1985	7.14^{***}
14	14	.577050	.010031	7.68^{***}	.583156	.015003	5.54^{***}	196	133	1972	23.83^{***}

[a] With the three largest studies removed from the sample, the effect size is significantly larger ($p < .05$, $z > 1.96$) than *MCE*.

[b] With the three largest studies removed from the sample, the effect size is larger than .50 (*MCE*), but not significantly so.

[1] It should be noted that the experimental report with the three by far largest studies (Dobyns, Dunne & Nelson, 2004) also includes three smaller studies.

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

studies were all independently statistically highly significant. There is a highly significant correlation between the study split and effect size ($r(380) = .36$, $p = 5.68 * 10^{-30}$). Studies from an experimental report that was split into several studies produce larger effect sizes than studies from an experimental report which was split into fewer studies. Moreover, the split of studies appeared to be preplanned for the overwhelming majority of studies (see Table 5), making it difficult to understand how so many unpublished, nonsignificant studies can be missing from the database.

A thought experiment by Bierman (1987) makes it less implausible to think that so many studies could remain unpublished. He writes that “RNG-PK data are very prone of [*sic*] ending up in a file-drawer” (p. 33). To prevent this bias he sent the “hypothesis and the planned explorations” (p. 33) of his experiment to an outside person. Bierman argues that “there are about 30 RIPP (Research Institute for Psi phenomena and Physics) RNG boards in the field” (p. 33), that is, there were approximately 30 RNGs developed at his institute (RIPP) that were run on Apple II computers. He reasons that:

> A typical experiment, like the one reported in this paper (8 Ss and 16 runs per subject), takes about 1000 seconds as far as data-acquisition is concerned. Including proper handling of subjects, such a typical experiment can be done within a week, including data-analysis. Thus using this technology one can easily run a 20 experiments per year. For the past 5 years this could imply that 3000 experiments were done which never reached the outside world. (p. 33f)

With 131,072 random bits, Bierman’s experiment is typical in respect of sample size.[15] Also, RNGs other than RIPP-RNGs are available and the 5 year period Bierman is taking into account is only 1/7 of the period taking into account in here. From this perspective, the proposed 1500 unpublished studies do not appear to be a wholly unreasonable number. Meta-analysing RNG studies would certainly be a lot easier if all experimenters registered the hypothesis, the

[15] The sample size is based on eight subjects participating in 16 runs with 16 intervals with 64 bit respectively.

sample size, and the preplanned analyses to an external body before conducting their studies.

4.1.10.5 Limits

Modeling, as well as meta-analyses, are limited by the assumptions underlying them. One of the main assumptions in undertaking a meta-analysis is the independence of effect size from sample size, an assumption that is inherent in effect size measures. However, the effect might be one in which sample size is not independent of effect size. For example, the *z*-scores of studies could be independent of (the square root of) sample size and constant across studies, as proposed by Radin & Nelson (2003) in their last RNG meta-analysis. In the current meta-analysis, the correlation between the studies *z*-score and $\sqrt{N}$ is significant ($r(380) = -.14$, $p = .006$) but negatively, so our total database does not support the constant *z*-score hypothesis proposed by Radin & Nelson. However, with the three largest studies removed the correlation becomes nonsignificant ($r(377) = -.02$, $p = .66$) and an argument for the model might be made. Nevertheless, the data clearly violate the general assumption behind power analysis, that is, that power increases with sample size. This is also evident from the small-study effect.

Another model, proposed by May, Radin, Hubbard, Humphrey & Utts (1986; see also, May, Utts & Spottiswoode, 1995; Dobyns, 1996), also questions the assumption that effect size is independent of sample size. It assumes that effect size in RNG experiments is a function "of 'correct' decisions based upon statistical glimpses of the future" (p. 261), that is the effect size of a study depends on the number of bits determined by each "button pushes" of the participant, who, according to the model, precognitively scans the future behaviour of the RNG and selects the time at which there are "locally deviant sub-sequences from a longer random sequence" (p. 249). However, a correlation of the number of starting points and effect size with sample size revealed no such relationship ($r(153) = .07$, $p = .36$). A more detailed analysis of one of the largest and most varied databases in the field (the PEAR laboratory database) also failed to confirm the model (Dobyns & Nelson, 1998). Moreover, this model must be considered highly speculative, since one anomalous concept, namely psychokinesis, is replaced by another anomalous concept, namely precognition.

However, most experimental RNG research assumes that intention affects the mean value of the random sequence, for example a shift in the distribution of Ones and Zeros. Although other outcome measures have been suggested to address the possibility of interdependence of data points (e.g., Atmanspacher, Bösch, Boller, Nelson & Scheingraber, 1999; Ehm, 2003; Nelson, 1994; Pallikari & Boller, 1999; Radin, 1989), they have been used only occasionally. Consequently, most RNG experiments have used the *z*-score measure, which assumes that any alleged influence affects the mean value of the random sequence. As a result, the straightforward effect size approach in this meta-analysis is clearly justifiable.

4.1.10.6 Conclusion

The statistical significance of the overall database provides no directive as to whether the phenomenon is genuine or not. The difference between the two statistical models used (FEM and REM), and the dependency of the results on three very large studies, demonstrates the difficulties regarding these data. If the striking heterogeneity and the small-study effect are taken into account, one must ask whether the findings are artifactual or whether all these findings are indicative of a genuine effect.

Publication bias appears to be the easiest and most encompassing explanation for the primary findings of the meta-analysis. The fit achieved by the Monte Carlo simulation was fairly good and clearly underpinned the hypothesis that the findings presented here are a result of publication bias. No other explanation accounted for all major findings (i.e., a striking variability of effect size, and the clearly visible small-study effect). Although the number of studies which have to be unpublished is considerable (*N*=1500), Bierman's thought experiment does make this number appear to be more than possible.

The publication process was clearly selective. The quartile of early RNG studies stands out from the other quartiles in terms of statistical significance and large effect size, during a period of time when RNG sample sizes were relatively small. Modeling this process by introducing additional limit values to early or small studies in the simulation might reduce the unpublished studies to a much smaller number. However, we have not implemented additional parameters in the model because the simulation was implemented primarily to indicate proof of principle. Adding additional parameters to the model will not necessarily increase the persuasive power because almost any model with a large enough number of parameters will eventually fit.

Although we question the conclusions of the preceding RNG meta-analyses, we would like to remind the reader that these experiments are highly refined operationalizations of a phenomenon which has challenged mankind for a long period of time. The dramatic anomalous PK effects reported in séance-rooms were reduced to experiments with electronic devices over a 100-year history of PK experiments. The effects dealt with in RNG experiments are certainly a far cry from those dramatic effects and, even if demonstrable, may not necessarily bear a direct relation to purported large-scale phenomena. PK may not be reducible to a microscopic level. Similarly, even if PK on a microscopic level were regarded as proven, this is a far remove from demonstrating the reality or otherwise of séance-room phenomena.

Further experiments will be conducted. They should be registered. This is the most straightforward solution for determining with any accuracy the rate of publication bias (e.g., Chalmers, 2001; Simes, 1986). It allows subsequent meta-analysts to resolve more firmly the question as to whether the overall effect in RNG experiments is an artifact of publication bias or whether the effect is genuine. The effect in general, even if incredibly small, is of great fundamental importance--if genuine. However, this unique experimental approach will gain scientific recognition only when we know with certainty what an unbiased funnel plot (i.e. a funnel plot that includes all studies that have been undertaken) looks like. If the time comes when the funnel indicates a systematic effect, a model to explain the effect will be more than crucial. Until that time, Girden’s verdict of “not proven” (1962b, p. 530), which he mooted more than 40 years ago in the same journal in respect of dice experiments, also holds for human intentionality on RNGs.

4.1.11 Author note

Holger Bösch, University Hospital Freiburg, Department of Evaluation Research in Complementary Medicine, Freiburg, Germany. Fiona Steinkamp, Department of Psychology, University of Edinburgh, Edinburgh UK. Emil Boller, Institute for Border Areas of Psychology and Mental Hygiene, Freiburg, Germany.

This research was sponsored by the Institute for Border Areas of Psychology and Mental Hygiene and the Samueli Institute. During the design of the meta-analysis and for the coding of the vast majority of the studies in the meta-analysis, all authors were employed at the Institute for Border Areas of Psychology and Mental Hygiene.

We thank Gabriela Böhm & Anna Hack for the manual search of the journals and Sonja Falk for her meticulous data cross checking. We also thank Peter Wilhelm, Andreas Sommer and Nikolaus von Stillfried for their comments and assistance on previous versions of this article. We are particularly thankful to Harald Walach for his help and encouragement during the writing and rewriting of the paper.

A table showing the effect sizes and study characteristics for each study in the meta-analysis is available from Holger Bösch.

4.1.12 References

References marked with an asterisk indicate studies included in the meta-analysis.

Alcock, J. E. (1981). *Parapsychology: Science or magic? A psychological perspective*. Oxford, England: Pergamon Press.

*André, E. (1972). Confirmation of PK action on electronic equipment. *Journal of Parapsychology, 36,* 283-293.Atmanspacher, H., Bösch, H., Boller, E., Nelson, R. D., & Scheingraber, H. (1999). Deviations from physical randomness due to human agent intention? *Chaos, Solitons & Fractals, 10,* 935-952.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50,* 1088-1101.

Beloff, J., & Evans, L. (1961). A radioactivity test of psycho-kinesis. *Journal of the Society for Psychical Research, 41,* 41-46.

Bem, D. J., & Honorton, C. (1994). Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin, 115,* 4-18.

*Berger, R. E. (1986). Psi effects without real-time feedback using a PsiLab // Video game experiment. In *The Parapsychological Association 29th Annual Convention*: *Proceedings of Presented Papers* (pp. 111-128). Durham, NC: Parapsychological Association.

*Berger, R. E. (1988). In search of "psychic signatures" in random data. In: D. H. Weiner & R. L. Morris (Eds.), *Research in Parapsychology 1987* (pp. 81-85). Metuchen, NJ: Scarecrow Press.

Bierman, D. J. (1985). A retro and direct PK test for babies with the manipulation of feedback: A first trial of independent replication using software exchange. *European Journal of Parapsychology, 5,* 373-390.

*Bierman, D. J. (1987). Explorations of some theoretical frameworks using a PK-test environment. In *The Parapsychological Association 30th Annual Convention*: *Proceedings of Presented Papers* (pp. 33-40). Durham, NC: Parapsychological Association.

*Bierman, D. J. (1988). Testing the IDS model with a gifted subject. *Theoretical Parapsychology, 6,* 31-36.

*Bierman, D. J., De Diana, I. P. F., & Houtkooper, J. M. (1976). Preliminary report on the Amsterdam experiments with Matthew Manning. *European Journal of Parapsychology, 1,* 6-16.

*Bierman, D. J., & Houtkooper, J. M. (1975). Exploratory PK tests with a programmable high speed random number generator. *European Journal of Parapsychology, 1,* 3-14.

*Bierman, D. J., & Houtkooper, J. M. (1981). The potential observer effect or the mystery of irreproduceability. *European Journal of Parapsychology, 3,* 345-371.

*Bierman, D. J., & Noortje, V. T. (1977). The performance of healers in PK tests with different RNG feedback algorithms. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1976* (pp. 131-133). Metuchen, NJ: Scarecrow Press.

*Bierman, D. J., & Van Gelderen, W. J. M. (1994). Geomagnetic activity and PK on a low and high trial-rate RNG. In *The Parapsychological Association 37th Annual Convention*: *Proceedings of Presented Papers* (pp. 50-56). Durham, NC: Parapsychological Association.

*Bierman, D. J., & Weiner, D. H. (1980). A preliminary study of the effect of data destruction on the influence of future observers. *Journal of Parapsychology, 44,* 233-243.

Blackmore, S. J. (1992). Psychic experiences: Psychic illusions. *Skeptical Inquirer, 16,* 367-376.

*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.

*Braud, L., & Braud, W. G. (1977). Psychokinetic effects upon a random event generator under conditions of limited feedback to volunteers and experimenter. In *The Parapsychological Association 20th Annual Convention: Proceedings of Presented Papers* (pp. 1-18). Durham, NC: Parapsychological Association.

*Braud, W. G. (1978). Recent investigations of microdynamic psychokinesis, with special emphasis on the roles of feedback, effort and awareness. *European Journal of Parapsychology, 2,* 137-162.

*Braud, W. G. (1981). Psychokinesis experiments with infants and young children. In W. G. Roll & J. Beloff (Eds.), *Research in Parapsychology 1980* (pp. 30-31). Metuchen, NJ: Scarecrow Press.

*Braud, W. G. (1983). Prolonged visualization practice and psychokinesis: A pilot study. In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in Parapsychology 1982* (pp. 187-189). Metuchen, NJ: Scarecrow Press.

*Braud, W. G., & Hartgrove, J. (1976). Clairvoyance and psychokinesis in transcendental meditators and matched control subjects: A preliminary study. *European Journal of Parapsychology, 1,* 6-16.

*Braud, W. G., & Kirk, J. (1977). Attempt to observe psychokinetic influences upon a random event generator by person-fish teams. *European Journal of Parapsychology, 2,* 228-237.

Braude, S. E. (1997). *The limits of influence: Psychokinesis and the philosophy of science* (Ref. ed.). Lanham, MD: University Press of America.

*Breederveld, H. (1988). Towards reproducible experiments in psychokinesis IV. Experiments with an electronic random number generator. *Theoretical Parapsychology, 6,* 43-51.

*Breederveld, H. (1989). The Michels experiments: An attempted replication. *Journal of the Society for Psychical* Research*, 55,* 360-363.

*Breederveld, H. (2001). De Optimal Stopping Strategie XL PK-experimenten met een random number generator. [The optimal stopping strategy XL. PK experiments with a random number generator]. *SRU-Bulletin, 13,* 22-23.

*Broughton, R. S. (1979). An experiment with the head of Jut. *European Journal of Parapsychology, 2,* 337-357.

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

*Broughton, R. S., & Alexander, C. H. (1997). Destruction testing DAT. In *The Parapsychological Association 40th Annual Convention*: *Proceedings of Presented Papers* (pp. 100-104). Durham, NC: Parapsychological Association.

*Broughton, R. S., & Higgins, C. A. (1994). An investigation of micro-PK and geomagnetism. In *The Parapsychological Association 37th Annual Convention: Proceedings of Presented Papers* (pp. 87-94). Durham, NC: Parapsychological Association.

*Broughton, R. S., & Millar, B. (1977). A PK experiment with a covert release-of-effort test. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1976* (pp. 28-30). Metuchen, NJ: Scarecrow Press.

*Broughton, R. S., Millar, B., & Johnson, M. (1981). An investigation into the use of aversion therapy techniques for the operant control of PK production in humans. *European Journal of Parapsychology, 3,* 317-344.

Brugger, P., Regard, M., Landis, T., Cook, N., Krebs, D., & Niederberger, J. (1993). 'Meaningful' patterns in visual noise: Effects of lateral stimulation and the observer's belief in ESP. *Psychopathology, 26,* 261-265.

Chalmers, I. (2001). Using systematic reviews and registers of ongoing trials for scientific and ethical trial design, monitoring, and reporting. In M. Egger, G. D. Smith, & D. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (pp. 429-443). London: BMJ Books.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49,* 997-1003.

*Crandall, J. E. (1993). Effects of extrinsic motivation on PK performance and its relations to state anxiety and extraversion. In *The Parapsychological Association 36th Annual Convention: Proceedings of Presented Papers* (pp. 372-377). Durham, NC: Parapsychological Association.

Crookes, W. (1889). Notes of séances with D. D. Home. *Proceedings of the Society for Psychical Research, 6,* 98-127.

Crookes, W., Horsley, V., Bull, W. C., & Myers, A. T. (1885). Report on an alleged physical phenomenon. *Proceedings of the Society for Psychical Research, 3,* 460-463.

*Curry, C. (1978). *A modularized random number generator: Engineering design and psychic experimentation*. Unpublished master's thesis, Department of Electrical Engineering and Computer Science, School of Engineering and Applied Science, Princeton University.

*Dalton, K. S. (1994). Remotely influenced ESP performance in a computer task: A preliminary study. In *The Parapsychological Association 37th Annual Convention: Proceedings of Presented Papers* (pp. 95-103). Durham, NC: Parapsychological Association.

*Davis, J. W., & Morrison, M. D. (1978). A test of the Schmidt model's prediction concerning multiple feedback in a PK test. In W. G. Roll (Ed.), *Research in Parapsychology 1977* (pp. 163-168). Metuchen, NJ: Scarecrow Press.

Dear, K. B. G., & Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science, 7,* 237-245.

*Debes, J., & Morris, R. L. (1982). Comparison of striving and nonstriving instructional sets in a PK study. *Journal of* Parapsychology*, 46,* 297-312.

Dobyns, Y. H. (1996). Selection versus influence revisited: New method and conclusions. *Journal of Scientific Exploration*, *10*, 253-267.

*Dobyns, Y. H., Dunne, B. J., & Nelson, R. D. (2004). The megaREG experiment: Replication and interpretation. *Journal of Scientific Exploration, 18,* 369-397.

Dobyns, Y. H., & Nelson, R. D. (1998). Empirical evidence against decision augmentation theory. *Journal of Scientific Exploration, 12,* 231-257.

Dudley, R. T. (2000). The relationship between negative affect and paranormal belief. *Personality and Individual Differences, 28,* 315-321.

Duval, S., & Tweedie, R. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95,* 89-98.

Edgeworth, F. Y. (1885). The calculus of probabilities applied to psychical research. *Proceedings of the Society for Psychical Research, 3,* 190-199.

Edgeworth, F. Y. (1886). The calculus of probabilities applied to psychical research. II. *Proceedings of the Society for Psychical Research, 4,* 189-208.

Egger, M., Dickersin, K., & Smith, G. D. (2001). Problems and limitations in conducting systematic reviews. In M. Egger, G. D. Smith, & D. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (pp. 43-68). London: BMJ Books.

Ehm, W. (2003). Pattern count statistics for the analysis of time series in mind-matter studies. *Journal of Scientific Exploration, 17,* 497-520.

Fisher, R. A. (1924). A method of scoring coincidences in tests with playing cards. *Proceedings of the Society for Psychical Research, 34,* 181-185.

Gallup, G., & Newport, F. (1991). Belief in paranormal phenomena among adult Americans. *Skeptical Inquirer, 15,* 137-146.

*Gausmann, U. (2004, June). *ArtREG: Ein Psychokineseexperiment mit visuellen affektiven Reizen [ArtREG: An experiment in psychokinesis with visually affective stimuli]* (Abschließender Forschungsbericht). Freiburg, Germany: Institut für Grenzgebiete der Psychologie und Psychohygiene e.V.

Geller, U. (1998). *Uri Geller's little book of mind-power*. London: Robson.

*Gerding, J. L. F., Wezelman, R., & Bierman, D. J. (1997). The Druten disturbances - Exploratory RSPK research. In *The Parapsychological Association 40th Annual Convention: Proceedings of Presented Papers* (pp. 146-161). Durham, NC: Parapsychological Association.

*Giesler, P. V. (1985). Differential micro-PK effects among Afro-Brazilian cultists: Three studies using trance-significant symbols as targets. *Journal of Parapsychology, 49,* 329-366.

Girden, E. (1962a). A review of psychokinesis (PK). *Psychological Bulletin, 59,* 353-388.

Girden, E. (1962b). A postscript to "A Review of Psychokinesis (PK)". *Psychological Bulletin, 59,* 529-531.

Girden, E., & Girden, E. (1985). Psychokinesis: Fifty years afterward. In P. Kurtz (Eds.), *A Skeptic's Handbook of Parapsychology* (pp. 129-146). Buffalo, NY: Prometheus Books.

*Gissurarson, L. R. (1986). RNG-PK microcomputer "games" overviewed: An experiment with the videogame "PSI INVADERS". *European Journal of Parapsychology, 6,* 199-215.

*Gissurarson, L. R. (1990). Some PK attitudes as determinants of PK performance. *European Journal of Parapsychology, 8,* 112-122.

Gissurarson, L. R. (1992). Studies of methods of enhancing and potentially training psychokinesis: A review. *Journal of the American Society for Psychical Research, 86,* 303-346.

Gissurarson, L. R. (1997). Methods of enhancing PK task performance. In S. Krippner (Ed.), *Advances in Parapsychological Research 8* (pp. 88-125). Jefferson, NC: Mc Farland Company.

*Gissurarson, L. R., & Morris, R. L. (1990). Volition and psychokinesis: Attempts to enhance PK performance through the practice of imagery strategies. *Journal of Parapsychology, 54,* 331-370.

*Gissurarson, L. R., & Morris, R. L. (1991). Examination of six questionnaires as predictors of psychokinesis performance. *Journal of Parapsychology, 55,* 119-145.

Hacking, I. (1988). Telepathy: Origins of randomization in experimental design. *Isis*, 79, 427-451.

*Haraldsson, E. (1970). Subject selection in a machine precognition test. *Journal of Parapsychology,* 34, 182-191.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science, 7,* 246-255.

Hedges, L. V. (1994). Fixed effect models. In L. V. Hedges & H. Cooper (Eds.), *The handbook of research synthesis* (pp. 285-299). New York: Russell Sage Foundation.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486-504.

*Heseltine, G. L. (1977). Electronic random number generator operation associated with EEG activity. *Journal of Parapsychology, 41,* 103-118.

*Heseltine, G. L., & Kirk, J. (1980). Examination of a majority-vote technique. *Journal of Parapsychology, 44,* 167-176.

*Heseltine, G. L., & Mayer-Oakes, S. A. (1978). Electronic random generator operation and EEG activity: Further studies. *Journal of Parapsychology, 42,* 123-136.

*Hill, S. (1977). PK effects by a single subject on a binary random number generator based on electronic noise. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1976* (pp. 26-28). Metuchen, NJ: Scarecrow Press.

*Honorton, C. (1971). Automated forced-choice precognition tests with a "sensitive". *Journal of the American Society for Psychical Research, 65,* 476-481.

*Honorton, C. (1971). Group PK performance with waking suggestions for muscle tension/relaxation and active/passive concentration. *Proceedings of the Parapsychological Association, 8,* 14-15.

*Honorton, C. (1977). Effects of meditation and feedback on psychokinetic performance: A pilot study with an instructor of Transcendental Meditation. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1976* (pp. 95-97). Metuchen, NJ: Scarecrow Press.

Honorton, C. (1985). Meta-analysis of psi ganzfeld research: A response to Hyman. *Journal of Parapsychology, 49,* 51-91.

*Honorton, C. (1987). Precognition and real-time ESP performance in a computer task with an exceptional subject. *Journal of Parapsychology, 51,* 291-320.

*Honorton, C., Barker, P., & Sondow, N. (1983). Feedback and participant-selection parameters in a computer RNG study. In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in Parapsychology 1982* (pp. 157-159). Metuchen, NJ: Scarecrow Press.

*Honorton, C., & Barksdale, W. (1972). PK performance with waking suggestions for muscle tension vs. relaxation. *Journal of the American Society for Psychical Research, 66,* 208-214.

Honorton, C., & Ferrari, D. C. (1989). "Future telling": A meta-analysis of forced-choice precognition experiments, 1935-1987. *Journal of Parapsychology, 53,* 281-308.

Honorton, C., Ferrari, D. C., & Bem, D. J. (1998). Extraversion and ESP performance: A meta-analysis and a new confirmation. *Journal of Parapsychology, 62,* 255-276.

*Honorton, C., & May, E. C. (1976). Volitional control in a psychokinetic task with auditory and visual feedback. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1975* (pp. 90-91). Metuchen, NJ: Scarecrow Press.

*Honorton, C., Ramsey, M., & Cabibbo, C. (1975). Experimenter effects in ESP research. *Journal of the American Society for Psychical Research, 69,* 135-139.

*Honorton, C., & Tremmel, L. (1980). Psitrek: A preliminary effort toward development of psi-conducive computer software. In W. G. Roll (Ed.), *Research in Parapsychology 1979* (pp. 159-161). Metuchen, NJ: Scarecrow Press.

*Houtkooper, J. M. (1976). Psychokinesis, clairvoyance and personality factors. In *The Parapsychological Association 19th Annual Convention*: *Proceedings of Presented Papers* (pp. 1-15). Durham, NC: Parapsychological Association.

*Houtkooper, J. M. (1977). A study of repeated retroactive psychokinesis in relation to direct and random PK effects. *European Journal of Parapsychology, 1,* 1-20.

Houtkooper, J. M. (2002). Arguing for an observational theory of paranormal phenomena. *Journal of Scientific Exploration, 16,* 171-185.

*Houtkooper, J. M. (2004). Exploring volitional strategies in the mind-machine interaction replication. In *The Parapsychological Association 47th Annual Convention: Proceedings of Presented Papers* (pp. 51-65). Durham, NC: Parapsychological Association.

Irwin, H. J. (1993). Belief in the paranormal: A review of the empirical literature. *Journal of the American Society for Psychical Research, 87,* 1-39.

Ioannidis, J. P. (1998). Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *The Journal of the American Medical Association, 279,* 281-286.

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science, 3,* 109-117.

*Jacobs, J. C., Michels, J. A. G., Millar, B., & Millar-De Bruyne, M.-L. F. L. (1987). Building a PK trap: The adaptive trial speed method. In *The Parapsychological Association 30th Annual Convention*: *Proceedings of Presented Papers* 348-370. Durham, NC: Parapsychological Association.

*Jahn, R. G., Dunne, B. J., Dobyns, Y. H., Nelson, R. D., & Bradish, G. J. (2000). ArtREG: A random event experiment utilizing picture-preference feedback. *Journal of Scientific Exploration, 14,* 383-409.

Jahn, R. G., Dunne, B. J., & Nelson, R. D. (1980). *Princeton Engineering Anomalies Research. Program Statement* (Tech. Rep.). Princeton, NJ: Princeton University, School of Engineering/Applied Science.

*Jahn, R. G., Mischo, J., Vaitl, D., Dunne, B. J., Bradish, G. J., Dobyns, Y. H., et al. (2000). Mind/Machine interaction consortium: PortREG replication experiments. *Journal of Scientific Exploration, 14,* 499-555.

James, W. (1896). Psychical research. *Psychological Review, 3,* 649-652.

Jeffers, S. (2003). Physics and claims for anomalous effects related to consciousness. *Journal of Consciousness Studies, 10,* 135-152.

Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *The Journal of the American Medical Association, 282,* 1054-1060.

*Kelly, E. F., & Kanthamani, B. K. (1972). A subject's efforts toward voluntary control. *Journal of* Parapsychology*, 36,* 185-197.

*Kugel, W. (1999). Amplifying precognition: Four experiments with roulette. In *The Parapsychological Association 42nd Annual Convention*: *Pro-*

ceedings of Presented Papers (pp. 136-146). Durham, NC: Parapsychological Association.

*Kugel, W., Bauer, B., & Bock, W. (1979). *Versuchsreihe Telbin [Experimental series Telbin]* (Arbeitsbericht 7). Berlin, Germany: Technische Universität Berlin.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32, 311-328.

Lawrence, T. R. (1998). Gathering in the sheep and goats. A meta-analysis of forced choice sheep-goat ESP studies, 1947-1993. In N. L. Zingrone, M. J. Schlitz, C. S. Alvarado, & J. Milton (Eds.), *Research in Parapsychology 1993* (pp. 27-31). Lanham, MD: Scarecrow Press.

*Lay, B. (1982). *Ein multivariates Psychokinese-Experiment [A multivariate experiment in psychokinesis]*. Unpublished master's thesis, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany.

*Levi, A. (1979). The influence of imagery and feedback on PK effects. *Journal of Parapsychology, 43,* 275-289.

*Lignon, Y., & Faton, L. (1977). Le factor psi séxerce sur un appareil électronique [The psi factor affects an electronic apparatus]. *Psi-Réalité,* 54-62.

Lipsey, M. W., & Wison, D. B. (2001). Practical meta-analysis. Applied Social Research Methods Series (Volume 49). London: Sage.

*Lounds, P. (1993). The influence of psychokinesis on the randomly-generated order of emotive and non-emotive slides. *Journal of the Society for Psychical Research, 59,* 187-193.

Lucadou, W. v., & Kornwachs, K. (1977). Can quantum theory explain paranormal phenomena? In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1976* (pp. 187-191). Metuchen, NJ: Scarecrow Press.

*Mabilleau, P. (1982). Electronic dice: A new way for experimentation in "psiology". *Le Bulletin PSILOG, 2,* 13-14.

Mahoney, M. J. (1985). Open exchange and epistemic progress. *American Psychologist, 40,* 29-39.

*Matas, F., & Pantas, L. (1971). A PK experiment comparing meditating vs. nonmeditating subjects. *Proceedings of the Parapsychological Association, 8,* 12-13.

*May, E. C., & Honorton, C. (1976). A dynamic PK experiment with Ingo Swann. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in* Parapsychology *1975* (pp. 88-89). Metuchen, NJ: Scarecrow Press.

May, E. C., Radin, D. I., Hubbard, G. S., Humphrey, B. S., & Utts, J. M. (1986). Psi experiments with random number generators: An informational model. In D. H. Weiner & D. I. Radin (Eds.), *Research in Parapsychology 1985* (pp. 119-123). Metuchen, NJ: Scarecrow Press.

May, E. C., Utts, J., & Spottiswoode, J. P. (1995). Decision augmentation theory: Toward a model of anomalous mental phenomena. *Journal of Parapsychology, 59,* 195-220.

McGarry, J. J., & Newberry, B. H. (1981). Beliefs in paranormal phenomena and locus of control: A field study. *Journal of Personality and Social Psychology, 41,* 725-736.

*Michels, J. A. G. (1987). Consistent high scoring in self-test PK experiments using a stopping strategy. *Journal of the Society for Psychical Research, 54,* 119-129.

*Millar, B. (1983). Random bit generator experimenten. Millar-replicatie. [Random bit generator experiments. Millar's replication]. *SRU-Bulletin, 8,* 119-123.

*Millar, B., & Broughton, R. S. (1976). A preliminary PK experiment with a novel computer-linked high speed random number generator. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1975* (pp. 83-84). Metuchen, NJ: Scarecrow Press.

*Millar, B., & Mackenzie, P. (1977). A test of intentional vs unintentional PK. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1976* (pp. 32-35). Metuchen, NJ: Scarecrow Press.

Milton, J. (1993). A meta-analysis of waking state of consciousness, free response ESP studies. In *The Parapsychological Association 36th Annual Convention*: *Proceedings of Presented Papers* (pp. 87-104). Durham, NC: Parapsychological Association.

Milton, J. (1997). Meta-Analysis of free-response ESP studies without altered states of consciousness. *Journal of Parapsychology, 61,* 279-319.

Milton, J., & Wiseman, R. (1999a). A meta-analysis of mass-media tests of extrasensory perception. *British Journal of Psychology, 90,* 235-240.

Milton, J., & Wiseman, R. (1999b). Does psi exist? Lack of replication of an anomalous process of information transfer. *Psychological Bulletin, 125,* 387-391.

*Morris, R., Nanko, M., & Phillips, D. (1978). Intentional observer influence upon measurements of a quantum mechanical system: A comparison of two imagery strategies. In *The Parapsychological Association 21st Annual Convention*: *Proceedings of Presented Papers* (pp. 266-275). Durham, NC: Parapsychological Association.

Morris, R. L. (1982). Assessing experimental support for true precognition. *Journal of Parapsychology, 46,* 321-336.

*Morris, R. L., & Garcia-Noriega, C. (1982). Variations in feedback characteristics and PK success. In W. G. Roll, R. L. Morris, & R. A. White (Eds.), *Research in Parapsychology 1981*, (pp. 138-140). Metuchen, NJ: Scarecrow Press.

*Morris, R. L., & Harnaday, J. (1981). An attempt to employ mental practice to facilitate PK. In W. G. Roll & J. Beloff (Eds.), *Research in Parapsychology 1980* (pp. 103-104). Metuchen, NJ: Scarecrow Press.

*Morris, R. L., & Reilly, V. (1980). A failure to obtain results with goal-oriented imagery PK and a random event generator with varying hit probability. In W. G. Roll (Ed.), *Research in Parapsychology 1979* (pp. 166-167). Metuchen, NJ: Scarecrow Press.

*Morrison, M. D., & Davis, J. W. (1978). PK with immediate, delayed, and multiple feedback: A test of the Schmidt model's predictions. In *The Parapsychological Association 21st Annual Convention*: *Proceedings of Presented Papers* (pp. 97-117). Durham, NC: Parapsychological Association.

Murphy, G. (1962). Report on paper by Edward Girden on psychokinesis. *Psychological Bulletin, 59,* 638-641.

Musch, J., & Ehrenberg, K. (2002). Probability misjudgement, cognitive ability, and belief in the paranormal. *British Journal of Psychology, 93,* 177

*Nanko, M. (1981). Use of goal-oriented imagery strategy on a psychokinetic task with "selected" subjects. *Journal of the Southern California Society for Psychical Research, 2,* 1-5.

*Nelson, R. D. (1994). *Effect size per hour: A natural unit for interpreting anomalies experiments* (Tech. Note 94003). Princeton, NJ: Princeton University, Princeton Engineering Anomalies Research (PEAR).

Pallikari, F., & Boller, E. (1999). A rescaled range analysis of random events. *Journal of Scientific Exploration, 13,* 35-40.

*Palmer, J. (1995). External psi influence of ESP task performance. In *The Parapsychological Association 38th Annual Convention*: *Proceedings of Presented Papers* (pp. 270-282). Durham, NC: Parapsychological Association.

*Palmer, J. (1998). ESP and REG PK with Sean Harribance: Three new studies. In *The Parapsychological Association 41st Annual Convention*: *Proceedings of Presented Papers* (pp. 124-134). Durham, NC: Parapsychological Association.

*Palmer, J., & Broughton, R. S. (1995). Performance in a computer task with an exceptional subject: A failure to replicate. In *The Parapsychological Association 38th Annual Convention*: *Proceedings of Presented Papers* (pp. 289-294). Durham, NC: Parapsychological Association.

*Palmer, J., & Perlstrom, J. R. (1986). Random event generator PK in relation to task instructions: A case of "motivated" error? In *The Parapsychological Association 29th Annual Convention*: *Proceedings of Presented Papers* (pp. 131-147). Durham, NC: Parapsychological Association.

*Pantas, L. (1971). PK scoring under preferred and nonpreferred conditions. *Proceedings of the Parapsychological Association, 8,* 47-49.

*Pare, R. (1983). Random bit generator experimenten. Pare-replicatie. [Random bit generator experiments. Pare's replication]. *SRU-Bulletin, 8,* 123-128.

Persinger, M. A. (2001). The neuropsychiatry of paranormal experiences. *Journal of Neuropsychiatry & Clinical Neurosciences, 13,* 515-523.

Pratt, J. G. (1937). Clairvoyant blind matching. *Journal of Parapsychology, 1,* 10-17.

Pratt, J. G. (1949). The meaning of performance curves in ESP and PK test data. *Journal of Parapsychology, 13,* 9-22.

Pratt, J. G., Rhine, J. B., Smith, B. M., Stuart, C. E., & Greenwood, J. A. (1940). *Extra-sensory perception after sixty years: A critical appraisal of* *the research in extra-sensory perception*. New York: Henry Holt and Company.

Presson, P. K., & Benassi, V. A. (1996). Illusion of control: A meta-analytic review. *Journal of Social Behavior & Personality, 11,* 493-510.

Price, M. M., & Pegram, M. H. (1937). Extra-sensory perception among the blind. *Journal of Parapsychology, 1,* 143-155.

*PRL. (1985). *PRL 1984 Annual Report*. Princeton, NJ: Psychophysical Research Laboratories.

Radin, D. I. (1982). Experimental attempts to influence pseudorandom number sequences. *Journal of the American Society for Psychical Research, 76,* 359-374.

Radin, D. I. (1989). Searching for "signatures" in anomalous human-machine interaction data: A neural network approach. *Journal of Scientific Exploration, 3,* 185-200.

*Radin, D. I. (1990). Testing the plausibility of psi-mediated computer system failures. *Journal of Parapsychology, 54,* 1-19.

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

Radin, D. I., & Ferrari, D. C. (1991). Effects of consciousness on the fall of dice: A meta-analysis. *Journal of Scientific Exploration, 5,* 61-83.

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

Radin, D. I., & Nelson, R. D. (2003). Research on mind-matter interactions (MMI): Individual intention. In W. B. Jonas & C. C. Crawford (Eds.), *Healing, Intention and Energy Medicine: Research and Clinical Implications* (pp. 39-48). Edinburgh, England: Churchill Livingstone.

*Randall, J. L. (1974). An extended series of ESP and PK tests with three English schoolboys. *Journal of the Society for Psychical Research, 47,* 485-494.

Reeves, M. P., & Rhine, J. B. (1943). The PK effect: II. A study in declines. *Journal of Parapsychology, 7,* 76-93.

*Reinsel, R. (1987). PK performance as a function of prior stage of sleep and time of night. In *The Parapsychological Association 30th Annual Con-*

vention: *Proceedings of Presented Papers* (pp. 332-347). Durham, NC: Parapsychological Association.

Rhine, J. B. (1934). *Extrasensory perception*. Boston: Boston Society for Psychic Research.

Rhine, J. B. (1936). Some selected experiments in extra-sensory perception. *Journal of Abnormal and Social Psychology, 29,* 151-171.

Rhine, J. B. (1937). The effect of distance in ESP tests. *Journal of Parapsychology, 1,* 172-184.

Rhine, J. B. (1946). Editorial: ESP and PK as "psi phenomena". *Journal of Parapsychology, 10,* 74-75.

Rhine, J. B., & Humphrey, B. M. (1944). The PK effect: Special evidence from hit patterns. I. Quarter distribution of the page. *Journal of Parapsychology, 8,* 18-60.

Rhine, J. B., & Humphrey, B. M. (1945). The PK effect with sixty dice per throw. *Journal of Parapsychology, 9,* 203-218.

Rhine, J. B., & Rhine, L. E. (1927). One evening's observation on the Margery mediumship. *Journal of Abnormal and Social Psychology, 21,* 421.

Rhine, L. E. (1937). Some stimulus variations in extra-sensory perception with child subjects. *Journal of Parapsychology, 1,* 102-113.

Rhine, L. E. (1970). *Mind over matter: Psychokinesis*. New York: Macmillan.

Rhine, L. E., & Rhine, J. B. (1943). The psychokinetic effect: I. The first experiment. *Journal of Parapsychology, 7,* 20-43.

Richet, C. (1884). La suggestion mentale et le calcul des probabilites [Mental suggestion and probability calculation]. *Revue Philosophique de la France et de l'Etranger, 18,* 609-674.

Richet, C. (1923). *Thirty years of physical research. A treatise on metapsychics*. New York: MacMillan Company.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638-641.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research. Methods and data analysis*. (2nd ed.). New York: McGraw-Hill Publishing.

Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice type data: Design, analysis, and meta-analysis. *Psychological Bulletin, 106,* 332-337.

Rush, J. H. (1977). Problems and methods in psychokinesis research. In S. Krippner (Ed.), *Advances in Parapsychological Research 1. Psychokinesis* (pp. 15-78). New York: Plenum.

Sanger, C. P. (1895). Analysis of Mrs. Verrall's card experiments. *Proceedings of the Society for Psychical Research, 11,* 193-197.

Scargle, J. D. (2000). Publication bias: The "file-drawer" problem in scientific inference. *Journal of Scientific Exploration, 14*, 91-106.

*Schechter, E. I., Barker, P., & Varvoglis, M. P. (1983). A preliminary study with a PK game involving distraction from the Psi task. In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in Parapsychology 1982* (pp. 152-154). Metuchen, NJ: Scarecrow Press.

*Schechter, E. I., Honorton, C., Barker, P., & Varvoglis, M. P. (1984). Relationships between participant traits and scores on two computer-controlled RNG-PK games. In R. A. White & R. S. Broughton (Eds.), *Research in Parapsychology 1983* (pp. 32-33). Metuchen, NJ: Scarecrow Press.

Schmeidler, G. R. (1977). Research findings in psychokinesis. In S. Krippner (Ed.), *Advances in Parapsychological Research 1. Psychokinesis* (pp. 79-132). New York: Plenum.

Schmeidler, G. R. (1982). PK research: Findings and theories. In S. Krippner (Ed.), *Advances in Parapsychological Research 3* (pp. 115-146). New York: Plenum Press.

*Schmeidler, G. R., & Borchardt, R. (1981). Psi-scores with random and pseudo-random targets. In W. G. Roll & J. Beloff (Eds.), *Research in Parapsychology 1980* (pp. 45-47). Metuchen, NJ: Scarecrow Press.

*Schmidt, H. (1969). *Anomalous prediction of quantum processes by some human subjects* (D1-82-0821). Seattle, WA: Boeing Scientific Research Laboratories, Plasma Physics Laboratory.

*Schmidt, H. (1970a). A PK test with electronic equipment. *Journal of Parapsychology, 34,* 175-181.

Schmidt, H. (1970b). PK experiments with animals as subjects. *Journal of Parapsychology, 34,* 255-261.

*Schmidt, H. (1972). An attempt to increase the efficiency of PK testing by an increase in the generation speed. In *The Parapsychological Association 15th Annual Convention*: *Proceedings of Presented Papers* (pp. 1-6). Durham, NC: Parapsychological Association.

*Schmidt, H. (1973). PK tests with a high-speed random number generator. *Journal of Parapsychology, 37,* 105-118.

*Schmidt, H. (1974). Comparison of PK action on two different random number generators. *Journal of Parapsychology, 38,* 47-55.

*Schmidt, H. (1974). PK effect on random time intervals. In W. G. Roll, R. L. Morris, & J. D. Morris (Eds.), *Research in Parapsychology 1973* (pp. 46-48). Metuchen, NJ: Scarecrow Press.

Schmidt, H. (1975). Toward a mathematical theory of psi. *Journal of the American Society for Psychical Research, 69,* 301-319.

*Schmidt, H. (1975). PK experiment with repeated, time displaced feedback. In *The Parapsychological Association 18th Annual Convention*: *Proceedings of Presented Papers* (pp. 1-6). Durham, NC: Parapsychological Association.

*Schmidt, H. (1976). PK effects on pre-recorded targets. *Journal of the American Society for Psychical Research, 70,* 267-291.

*Schmidt, H. (1978). Use of stroboscopic light as rewarding feedback in a PK test with pre-recorded and momentarily generated random events. In *The Parapsychological Association 21st Annual Convention*: *Proceedings of Presented Papers* (pp. 85-96). Durham, NC: Parapsychological Association.

Schmidt, H. (1979). Search for psi fluctuations in a PK test with cockroaches. In W. G. Roll (Ed.), *Research in Parapsychology 1978* (pp. 77-78). Metuchen, NJ: Scarecrow Press.

Schmidt, H. (1985). Addition effect for PK on pre-recorded targets. *Journal of Parapsychology, 49,* 229-244.

*Schmidt, H. (1990). Correlation between mental processes and external random events. *Journal of Scientific Exploration, 4,* 233-241.

Schmidt, H. (1992). Progress and problems in psychokinesis research. In B. Rubik (Ed.), *The interrelationship between mind and matter: Proceedings of a conference hosted by the Center for Frontier Studies* (pp. 39-55). Philadelphia: Temple University.

*Schmidt, H., & Pantas, L. (1972). Psi tests with internally different machines. *Journal of Parapsychology, 36,* 222-232.

*Schmidt, H., & Terry, J. (1977). Search for a relationship between brainwaves and PK performance. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1976* (pp. 30-32). Metuchen, NJ: Scarecrow Press.

*Schouten, S. A. (1977). Testing some implications of a PK observational theory. *European Journal of Parapsychology, 1,* 21-31.

Schouten, S. A. (1983). Personal experience and belief in ESP. *The Journal of Psychology, 114,* 219-222.

Shadish, W. R., & Haddock, K. C. (1994). Combining estimates of effect size. In L. V. Hedges & H. Cooper (Eds.), *The handbook of research synthesis* (pp. 261-281). New York: Russell Sage Foundation.

Shoup, R. (2002). Anomalies and constraints: Can clairvoyance, precognition, and psychokinesis be accommodated within known physics? *Journal of Scientific Exploration, 16,* 3-18.

Simes, R. J. (1986). Publication bias: The case for an international registry of clinical trials. *Journal of Clinical Oncology, 4,* 1529-1541.

Sparks, G. G. (1998). Paranormal depictions in the media: How do they affect what people believe? *Skeptical Inquirer, 22,* 35-39.

Sparks, G. G., Hansen, T., & Shah, R. (1994). Do televised depictions of paranormal events influence viewers′ beliefs? *Skeptical Inquirer, 18,* 386-395.

Sparks, G. G., Nelson, C. L., & Campbell, R. G. (1997). The relationship between exposure to televised messages about paranormal phenomena and paranormal beliefs. *Journal of Broadcasting & Electronic Media, 41,* 345-359.

Stanford, R. G. (1978). Toward reinterpreting psi events. *Journal of the American Society for Psychical Research, 72,* 197-214.

*Stanford, R. G. (1981). "Associative activation of the unconscious" and "visualization" as methods for influencing the PK target: A second study. *Journal of the American Society for Psychical Research, 75,* 229-240.

*Stanford, R. G., & Kottoor, T. M. (1985). Disruption of attention and PK-task performance. In *The Parapsychological Association 28th Annual Convention*: *Proceedings of Presented Papers* (pp. 117-132). Durham, NC: Parapsychological Association.

Stanford, R. G., & Stein, A. G. (1994). A meta-analysis of ESP studies contrasting hypnosis and a comparison condition. *Journal of Parapsychology, 58,* 235-269.

Stapp, H. P. (1994). Theoretical model of a purported empirical violation of the predictions of quantum theory. *Physical Review A, 50, 1,* 18-22.

Steinkamp, F., Milton, J., & Morris, R. L. (1998). A meta-analysis of forced-choice experiments comparing clairvoyance and precognition. *Journal of Parapsychology, 62,* 193-218.

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology, 54,* 1046-1055.

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology, 53,* 1119-1129.

Stokes, D. M. (1987). Theoretical parapsychology. In K. Stanley (Ed.), *Advances in Parapsychological Research 5* (pp. 77-189). Jefferson, NC: McFarland.

Storm, L., & Ertel, S. (2001). Does psi exist? Comments on Milton and Wiseman's (1999) meta-analysis of ganzfeld research. *Psychological Bulletin, 127,* 424-433.

*Talbert, R., & Debes, J. (1981). Time-displacement psychokinetic effects on a random number generator using varying amounts of feedback. In *The Parapsychological Association 24th Annual Convention*: *Proceedings of Presented Papers* (pp. 58-61) Durham, NC: Parapsychological Association.

Targ, R., & Puthoff, H. E. (1977). *Mind-reach: Scientists look at psychic ability*. New York: Delacorte Press.

Tart, C. T. (1976). Effects of immediate feedback on ESP performance. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1975* (pp. 80-82). Metuchen, NJ: Scarecrow Press.

Taylor, G. Le M. (1890). Experimental comparison between chance and thought-transference in correspondence of diagrams. *Proceedings of the Society for Psychical Research, 6,* 398-405.

*Tedder, W. (1984). Computer-based long distance ESP: An exploratory examination. In R. A. White & R. S. Broughton (Eds.), *Research in Parapsychology 1983* (pp. 100-101). Metuchen, NJ: Scarecrow Press.

Thalbourne, M. A. (1995). Further studies of the measurement and correlates of belief in the paranormal. *Journal of the American Society for Psychical Research, 89,* 233-247.

Thalbourne, M. A. (in press). *The common thread between ESP and PK*. New York: Parapsychological Foundation.

*Thompson Smith, A. T. (2000). *Anomalous human computer interaction: Relationship to training expectations, absorption, flow, and creativity*. Unpublished doctoral dissertation, Saybrook Graduate School and Research Center, San Francisco, CA.

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 2002,* 1559-1574.

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine, 18,* 2693-2708.

Thouless, R. H. (1942). The present position of experimental research into telepathy and related phenomena. *Proceedings of the Society for Psychical Research, 47,* 1-19.

*Thouless, R. H. (1971). Experiments on psi self-training with Dr. Schmidt´s pre-cognitive apparatus. *Journal* of *the Society for Psychical Research, 46,* 15-21.

Thouless, R. H., & Wiesner, B. P. (1946). The psi processes in normal and "paranormal" psychology. *Proceedings of the Society for Psychical Research, 48,* 177-196.

*Tremmel, L., & Honorton, C. (1980). Directional PK effects with a computer-based random generator system: A preliminary study. In W. G. Roll (Ed.), *Research in Parapsychology 1979* (pp. 69-71). Metuchen, NJ: Scarecrow Press.

*Varvoglis, M. P. (1988). A "psychic contest" using a computer-RNG task in a non-laboratory setting. In *The Parapsychological Association 31st Annual Convention*: *Proceedings of Presented Papers* (pp. 36-52). Durham, NC: Parapsychological Association.

Varvoglis, M. P., & McCarthy, D. (1986). Conscious-purposive focus and PK: RNG activity in relation to awareness, task-orientation, and feedback. *Journal of the American Society for Psychical Research, 80,* 1-29.

*Verbraak, A. (1981). Onafhankelijke random bit generator experimenten - Verbraak-replicatie. [Independent random bit generator experiments - Verbraak's replication]. *SRU-Bulletin, 6,* 134-139.

Walker, E. H. (1974). Consciousness and quantum theory. In J. White (Ed.), *Psychic exploration: A challenge for science* (pp. 544-568). New York: Putnam's.

Walker, E. H. (1975). Foundations of paraphysical and parapsychological phenomena. In L. Oteri (Ed.), *Quantum physics and parapsychology* (pp. 1-53).

Watt, C. A. (1994). Meta-analysis of DMT-ESP studies and an experimental investigation of perceptual defense/vigilance and extrasensory perception. In E. W. Cook & D. L. Delanoy (Eds.), *Research in Parapsychology 1991* (pp. 64-68). Metuchen, NJ: Scarecrow Press.

*Weiner, D. H., & Bierman, D. J. (1979). An observer effect in data analysis? In W. G. Roll (Ed.), *Research in Parapsychology 1978* (pp. 57-58). Metuchen, NJ: Scarecrow Press.

White, R. A. (1991). The psiline database system. *Exceptional Human Experience, 9,* 163-167.

Wilson, C. (1976). *The Geller phenomenon*. London: Aldus Books.

*Winnett, R. (1977). Effects of meditation and feedback on psychokinetic performance: Results with practitioners of Ajapa Yoga. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1976* (pp. 97-98). Metuchen, NJ: Scarecrow Press.