Psychological Bulletin
2006, Vol. 132, No. 4, 497–523


0033-2909/06/$12.00 DOI: 10.1037/0033-2909.132.4.497

# Examining Psychokinesis: The Interaction of Human Intention With Random Number Generators—A Meta-Analysis

Holger Bösch
University Hospital Freiburg

Fiona Steinkamp
University of Edinburgh

Emil Boller
Institute for Border Areas of Psychology and Mental Hygiene

Séance-room and other large-scale psychokinetic phenomena have fascinated humankind 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 correlated with human intention and found a significant but very small overall effect size. 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, and the extreme effect size heterogeneity found could in principle be a result of publication bias.

*Keywords:* meta-analysis, psychokinesis, random number generator, small-study effect, publication bias

During the 1970s, Uri Geller inspired much public interest in phenomena apparently demonstrating the ability of the 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, 1991), 17% of American adults believed in "the ability of the mind to move or bend objects using just mental energy" (p. 138), and 7% 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 in which 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 summarizes how the empirical investigation of this phenomenon developed over the decades and presents a new meta-analysis of a large set of experiments examining the interaction between human intention and random number generators.[1]

## Psi Research

*Psi* phenomena (Thouless, 1942; Thouless & Wiesner, 1946) can be split into two main categories: psychokinesis (PK) and extrasensory perception (ESP). *PK* 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 colleagues 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 9 years (L. E. Rhine & Rhine, 1943), even though both the ESP and PK experiments had been carried out at the same time:

Holger Bösch, Department of Evaluation Research in Complementary Medicine, University Hospital Freiburg, Freiburg, Germany; Fiona Steinkamp, Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom; 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 and 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 article. A table showing the effect sizes and study characteristics for each study in the meta-analysis is available from Holger Bösch.

Correspondence concerning this article should be addressed to Holger Bösch, Department of Evaluation Research in Complementary Medicine, University Hospital Freiburg, Hugstetter Strasse 55, D 79106 Freiburg, Germany. E-mail: holger.boesch@uniklinik-freiburg.de

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

J. B. Rhine and his colleagues did not want to undermine the scientific credibility that they had gained through their pioneering monograph on ESP (Pratt, Rhine, Smith, Stuart, & Greenwood, 1940).

When L. E. Rhine and 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.

### *Verifying Psi*

To verify the existence of psi phenomena, researchers have already conducted 13 meta-analyses (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), 2 of which provide no evidence for psi (Milton & Wiseman, 1999a, 1999b). Only 3 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.

### *Psychology and Psi*

Psychological approaches to psi experiences have also almost exclusively focused on ESP. For example, some researchers have hypothesized 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). Both 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), focus on ESP, although the paranormal belief scale most frequently used in this line of research also has some items on PK (Thalbourne, 1995).

### *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, & Myers, 1885), prompting some outstanding researchers of the time to devote at least part of their careers 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 & 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 (L. E. Rhine & Rhine, 1943). 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 throws of the dice, and recording errors were minimized by either having experimenters 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 the sides with the 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 side (e.g., 1).

In 1962, Girden (1962a) published a comprehensive critique of dice experiments in *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 and Ferrari undertook a meta-analysis of the dice experiments.

## Dice Meta-Analysis

The dice meta-analysis (Radin & Ferrari, 1991) comprised 148 experimental studies and 31 control studies published between 1935 and 1987. In the experimental studies, 2,569 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) under a control condition in which the dice were tossed specifically as control runs (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] The overall effect size, weighted by the inverse of the variance, is small but highly significant ($\bar{\pi}_t$ = .50610, $z$ = 19.68). Radin and Ferrari (1991) calculated that approximately 18,000 null effect studies would be required to reduce the result to a nonsignificant level (Rosenthal, 1979).[3] When the studies were weighted for quality, the effect size decreased considerably ($\Delta z = 5.27$, $p = 1.34 \times 10^{-7}$; see Table 1 for comparison) but was still highly significantly above chance.

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

| Study category and group | $N$ | $\bar{\pi}_t$ | $SE$ | $z$ |
|---|---|---|---|---|
| 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 based on $r = z/\sqrt{N}$ were transformed using $\bar{\pi}_t = 0.5\bar{r} + 0.5$ to achieve comparability.
** $p < .01$, one-tailed. *** $p < .001$, one-tailed.

Radin and Ferrari (1991) 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 side). 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 and Ferrari trimmed the sample until the effect sizes in the balanced database became homogenous, the effect size was reduced to only .50158, and it fell yet further to .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 and Ferrari (1991) neither examined the source(s) of heterogeneity in their meta-analysis nor addressed whether the strong correlation

[2] To compare the meta-analytic findings from the dice and previous random number generator (RNG) meta-analyses with those from our RNG meta-analysis, we converted all effect size measures to the proportion index $\pi$, which we use throughout the article (see the Method section). Although we use a fixed-effects model (FEM) as well as a random-effects model (REM) for our own analyses, the first dice and the first RNG meta-analyses exclusively used a weighted ($1/v$) FEM. Because it is not possible to calculate an REM given only the published data, all analyses 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 $\bar{\pi}_t = 0.5\bar{r} + 0.5$. 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 $\pm 1$. However, the difference between $\bar{\pi}_o$ and $\bar{\pi}_t$ is very small ($< .05\%$) as long as the $z$ values are not extreme ($z < 10$, $p < 1 \times 10^{-10}$). 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 (1979) 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 0. This assumption has been questioned by several authors (e.g., Iyengar & Greenhouse, 1988; Scargle, 2000). If one were to assume instead that the unpublished studies are a random sample of only the nonsignificant studies and that the mean $z$ score of the unpublished studies is $-0.1085$ (Scargle, 2000), then 1,450 studies, rather than 18,000 studies, would be needed to reduce the overall effect to a nonsignificant level.

[4] For this particular subsample, Radin and Ferrari (1991) 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's $z$ ($z_n$) provided in the article (Radin & Ferrari, 1991, Table 2, p. 76) and used $X = (n/2.706)[n(z_n)^2 - 2.706]$ as proposed by Rosenthal, where $z_n = z_n/\sqrt{n}$.

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, 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.

## From Dice to Random Number Generator

With the arrival of computation, dice experiments were slowly replaced by a new approach. Beloff and 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 the use of 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 (e.g., Bohr, 1935; Stapp, 1993), observational theory models psi effects as analogous to the collapse of the state vector, which is believed to be related to the consciousness of the observer (e.g., von Lucadou & Kornwachs, 1977; Schmidt, 1975; Walker, 1974, 1975). During this time, parapsychological modeling 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., Schmidt, 1969) and later by the Princeton Engineering Anomalies Research (PEAR) laboratory at Princeton University (e.g., Jahn, Dunne, & Nelson, 1980).

### *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 to mentally influence the RNG to produce, say, more 1s than 0s for a predefined number of bits. Participants are generally given real-time feedback on 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 counterclockwise 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 toward 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's (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).

### *New Limits*

From a methodological point of view, RNG experiments have many advantages over the older dice experiments. However, with respect to 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. In addition, 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 through the use of 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.

## Previous RNG Meta-Analyses

The first RNG meta-analysis was published by Radin and 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 $\bar{\pi}_o = .50018$ ($SE = .00003$, $z = 6.53$, $p < 1.00 \times 10^{-10}$).[5] The size of the effect did not diminish 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%" (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 presented 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 and using 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 is a very important topic in the current meta-analysis.

Recently, Radin and Nelson (2003) published an update of their earlier (Radin & Nelson, 1989) RNG meta-analysis, adding a further 176 studies to their database. In this update, the PEAR laboratory data were collapsed into one 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 FEM is applied to the data, as was used in the first RNG meta-analysis, the effect size of the updated database becomes $\bar{\pi}_o = .50005$, which is significantly smaller than the effect size of the original RNG meta-analysis ($\Delta z = 4.27$, $p = 1.99 \times 10^{-5}$; 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 independent evaluation of their findings is lacking. The need for a more thoroughgoing meta-analysis of PK RNG experiments is clear.

## Human Intention Interacting With RNGs: A New Meta-Analysis

The meta-analysis presented here was part of a 5-year consortium project on RNG experiments. The consortium comprised research groups from the PEAR laboratory; the Justus Liebig University of Giessen, 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, Mischo, et al., 2000), which form one of the strongest and most influential data sets 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.

---

[5] The meta-analysis provided the overall effect size only in a figure (Radin & Nelson, 1989, Figure 3, p. 1506). Because Dean Radin kindly provided us with the original data, we were able to calculate the 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" (Radin & Nelson, 1989, 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 with respect to the title of the meta-analysis: "Evidence for Consciousness-Related Anomalies in Random Physical Systems."

[7] The difference in effect size between $\bar{\pi}_o$ (i.e., the effect size based on original data) and $\bar{\pi}_t$ (i.e., 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 ($\Delta z = 4.12$, $p = 3.72 \times 10^{-5}$), 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 Psychokinesis Meta-Analyses: Total Samples*

| Study type and meta-analysis | *N* | $\bar{\pi}_o$ | *SE* | *z* | *M* $\pi$ |
|---|---|---|---|---|---|
| Dice | | | | | |
| Radin & Ferrari, 1991, meta-analysis | 148 | .50822 | .00041 | 20.23*** | .51105 |
| Random number generator | | | | | |
| Radin & Nelson, 1989: First meta-analysis | 597 | .50018 | .00003 | 6.53*** | .50414 |
| Radin, 1997: First meta-analysis without PEAR lab data | 339 | .50061 | .00009 | 6.41*** | .50701 |
| Radin & Nelson, 2003: Second meta-analysis | 515 | .50005 | .00001 | 3.81*** | .50568 |

*Note.* The effect size measure $\bar{\pi}_o$ was computed from original data available to the authors. *M* $\pi$ = unweighted, averaged effect size of studies. PEAR = Princeton Engineering Anomalies Research.
*** $p < .001$, one-tailed.

## Method

### 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 through the use of multiple search strategies. The first step involved an extensive manual search at the library and archives of the Institute for Border Areas of Psychology and Mental Hygiene, 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 is 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 University of Utrecht* (1971–1984), *Bulletin PSILOG* (1981–1983), *Journal of the Southern California Society for Psychical Research* (1979–1985), and *Arbeitsberichte Parapsychologie der Technischen Universität Berlin* (1971–1980). Although for some journals the search may seem incomplete, we 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 Association's annual convention appear 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 postconference 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 (Version 1999), a continuously updated, specialized electronic resource of parapsychologically relevant writings (White, 1991), was searched using the keywords *random number generator, RNG, random event generator,* and *REG. Dissertation Abstracts on Disk* (8 CDs; January 1961–June 2004) was searched using four different search strategies. First, the keywords *random number generator, RNG, random event generator, REG, randomness, radioactive, parapsychology, parapsychological, perturbation, psychokinesis, PK, extrasensory perception, ESP, telepathy, precognition,* and *calibration* were used. Second, the keywords *random* and *experiment* were combined with *event, number, noise, anomalous, anomaly, influence, generator, apparatus,* or *binary.* Third, the keyword *machine* was combined with *man* or *mind.* Fourth, the keyword *zener* was combined with *diode.* The search included plural variants of all keywords accordingly. However, not all keywords were indexed for all CDs. The PsycINFO database (June 2004) was searched using three different search strategies. First, the keywords *random number generator, RNG, random event generator, REG, perturbation,* and *psychokinesis* were used. Second, the keyword *machine* was combined with *man* or *mind,* and third, the keyword *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 Radin and Nelson, was searched for reports using true RNGs. To obtain as many relevant unpublished manuscripts as possible, we made visits to three other prolific parapsychology research institutes: the Rhine Research Center, Durham, NC; the PEAR laboratory; and the Koestler Parapsychology Unit at University of Edinburgh. Furthermore, a request for unpublished experiments was placed on an electronic mailing list for professional parapsychologists (Parapsychology Research Forum).

As a final step, the reference sections of all retrieved reports, that is, journal articles, conference proceedings, theses and 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.

### 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 is 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 to influencing the RNG. We also excluded those that (b) used animals or plants as partici-

pants (e.g., Schmidt, 1970b); (c) assessed the possibility of a nonintentional 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 backward in time or, similarly, in which participants observed the same bits a number of times (e.g., Morris, 1982; Schmidt, 1985) and; (e) evaluated whether there was an effect of human intention on a pseudo-RNG (e.g., Radin, 1982).

In addition, experiments were excluded if their outcome could not be transformed into the effect size $\pi$ that was prespecified for this meta-analysis. This excluded studies for 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 are decisions concerning how to perform the literature search and decisions made during the coding procedure. The decisions depend not only 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.

## Defining Studies

Some experiments were described in both published and unpublished reports or 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.*" We defined a study as 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 in which the same data could be split 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 in the experimental database, were split on the basis of 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, Mischo, et al., 2000). These data were excluded from both experimental and control databases because of the inherent ambiguity as to whether the participant attempted an influence during these data-gathering periods. Jahn also argued that these data should be excluded (as cited by 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 with 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 (L. E. Rhine, 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 was used in the meta-analysis.

## Coding Studies

The variables coded covered six main areas:

1. *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, whether the psychological test was taken before experiment (i.e., yes, no, other), comments regarding psychological testing procedure, systematic state manipulation (i.e., no, yes, other), whether state manipulation was verified (i.e., yes, no, other), description of the state manipulation procedure, comments regarding state manipulation, when control data was accumulated (i.e., during experiment, before or after experiment, during and before or after experiment, other), feedback during accumulation of control data (i.e., yes, no, other), and comments regarding control data.

2. *Participant information,* which included participant type (i.e., adults, students, adults or students, 13–18-year-olds, 6–12-year-olds, preschool children, infants or babies, animals, plants, other), species of animal or plant, participant selection (i.e., volunteer paid, volunteer unpaid, semivolunteer, nonvolunteer, experimenter, mixed, other), selection criteria (i.e., none, psychic claimant, prior success in psi experiment, psychological test, prior psychic experiences, practicing meditation or yoga, other), number of participants, and comments regarding participant information.

3. *Experimenter information,* which included whether experimenter was also participant (i.e., yes, no, partially, other), affiliation of first author, whether experimenter was in room with participant (i.e., yes, no, experimenter was participant, sometimes, other), and who initiated individual trial or run (i.e., experimenter, participant, mixed, automatic, other).

4. *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), whether participant was informed about RNG (i.e., no, some details, detailed information, other), direction of intention (i.e., one direction, balanced, other), who chose intention (i.e., experimenter, participant, prespecified, randomized, other), RNG type (i.e., radioactive, noise, mixed with pseudo-RNG, other), type if mixed with pseudo-RNG (i.e., radioactive, noise, other), type of feedback (i.e., visual, auditory, other), timing of participant feedback (i.e., bit by bit, trial by trial, end of run, end of session, end of experiment, false feedback, mixed, other), timing of experimenter feedback (i.e., experimenter first, participant first, experimenter and participant receive feedback at the same time, mixed, other), and comments regarding experimental setting.

5. *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.

6. *Safeguard variables,* which were 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; and *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 3-point scale (*yes* [2], *earlier/other* [1],[8] *no* [0]), with the intermediate value being used either when it was unclear whether the study actually took the safeguard into account or when 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 specializing 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 with 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, that is, the moderators of quality, we prespecified 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, we decided to double-code only reports for which the main coder (Fiona Steinkamp) was unclear about how to code at least one variable. The second independent coder (Emil Boller) 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%–87.8% agreement about the basic information variables, a 76.5%–92.9% agreement about the statistical information, and a 73.4%–88.8% agreement regarding the safeguard variables. With respect to all other variables, the agreement ranged from 69.4% to 92.9%. All differences between the coders were resolved by consulting Holger Bösch, 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 conservative estimates.

## *Analyses*

The effect sizes of individual studies were combined into composite mean weighted effect size measures with an intuitively comprehensible effect size measure suggested by Rosenthal and Rubin (1989) for one-sample data. For $\pi$, a proportion index, 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 probabilities according to the hit rate of an equally likely two-alternative case such as coin flipping (with a fair coin). Thus, if heads in a coin flipping experiment ($k = 2$) wins at a hit rate of 50%, the effect size $\pi = .50$ indicates that heads and tails came down equally often; if the hit rate for heads is 75%, the effect size would be $\pi = .75$. An RNG (or dice) experiment with a 1/6 hit rate ($k = 6$) thus also converts to $\pi = .50$, the mean chance expectation (MCE) of $\pi$. The range of $\pi$, like the range of all probability measures, is from 0 to 1. With $k = 2$, that is, in the two alternatives case, Equation 1 reduces to $\pi = P$.

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

$$SE_{(\pi)} = \frac{\pi(1-\pi)}{\sqrt{N \times P(1-P)}}. \tag{2}$$

It is crucial to understand that in contrast to meta-analyses in psychology and medicine, $N$ (i.e., the number of independent data points) refers to the number of bits accumulated in an RNG study and not to 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 1 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 (first quartile, 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 the MCE.

To combine effect sizes from different studies, we calculated an FEM as well as an REM. The mean effect size ($\bar{\pi}$) 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):

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

where

$$w_i = \frac{1}{SE_{\pi_i}^2}. \tag{4}$$

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

[9] Variables that are rarely reported are generally problematic because it is unclear whether they are just rarely implemented in experiments or 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.

[10] Actually, none of the meta-analyses in parapsychology has so far made use of the number of participants as the number of 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, $r(344) = .18$, $p = .001$.

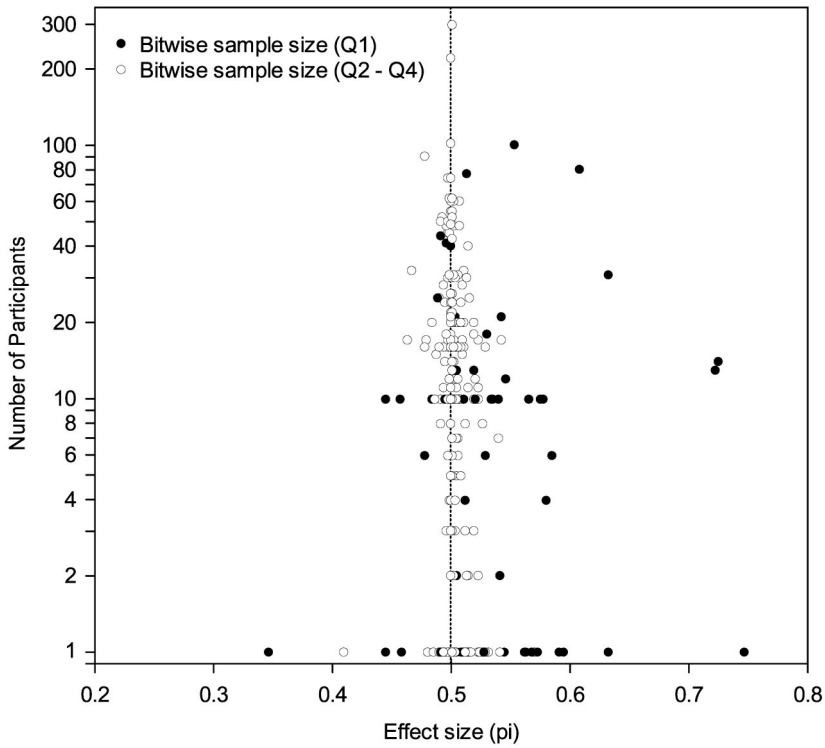


*Figure 1.* Funnel plot intentional studies with respect to the number of participants. The funnel shape of the graph is more evident when the number of participants is plotted using a linear scale. However, using a logarithmic scale stretches the graph in the lower part (fewer number of participants) 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), appears to depart visibly from the center line (range of $\pi = .495$–$.504$). Q = quartile.

To determine whether a sample of $\pi$s shared a common effect size (i.e., was consistent across studies), we calculated a homogeneity statistic $Q$, which has an approximately chi-square 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):

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

and

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

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

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

and

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

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

Generally, the result of the homogeneity statistic is considered crucial with respect to the appropriateness of the statistical model applied. However, a nonsignificant $Q$ value does not guarantee the adequacy of an FEM, and nor does a significant $Q$ value guarantee the adequacy of an REM (e.g., Lipsey & Wilson, 2001). There might be a considerable between-studies variance, suggesting an 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 (as our simulation demonstrates).[11] That is, although theoretically studies should distribute homogeneously, they do not have to, and consequently, the more conservative REM is more appropriate. We therefore

[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.

provide both estimates and several other sensitivity measures to put the data into perspective.

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

$$\Delta z = \frac{(\bar{\pi}_1 - \bar{\pi}_2)}{\sqrt{SE_1^2 + SE_2^2}}. \quad (10)$$

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

To explore the putative impact of moderator and safeguard variables on the effect size and to determine sources of heterogeneity, we carried out two metaregression analyses. Metaregression 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 and Olkin (1985). Two regression models were calculated. In Regression Model 1, sample size, year of publication, and number of participants entered as continuous variables. All other variables were dummy coded. In 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, we used a funnel plot. Three 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 the Discussion section). First, the sample was split into quartiles of sample size. Second, and on the basis of Begg and Mazumdar's (1994) approach, a rank correlation between effect size and sample size was performed. Third, Duval and 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 and Tweedie (2000), we used the $L_o$ 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's (1992) step-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 estimated the weights of the step function on the basis of the available meta-analytical data. However, unlike 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 $\leq .05$ and $> .01$, 50% of studies with a $p$ value between $\leq .10$ and $> .05$, 20% of studies with a $p$ value between $\leq .50$ and $> .10$, and 10% of studies with $p$ value $> .50$ (one-tailed) are "published."[12] Starting with these parameters, we randomly generated uniformly distributed $p$ values, and we calculated the effect sizes for all "published" studies and counted the number of "unpublished" studies. That is, for every 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 (Version 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 (Version 6.0; Steichen, 2004) using user-written Stata commands (from the Stata home page, www.stata.com).

## Results

### Study Characteristics

The basic study characteristics are summarized in Table 3. The heyday of RNG experimentation was in the 1970s, when more than half 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 average study sample size was 787,888,669 bits. However, most studies were much smaller, as indicated by the 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.

### 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 indicated an effect opposite to intention, whereas the effect size of the REM indicated 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 Figure 2), published in a single experimental report (Dobyns, Dunne, & Nelson, 2004). The effect sizes of these three studies, ranging from $\pi = .499989$ to $\pi = .499997$, indicated a result opposite to intention. Without these three studies, both models showed 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 ($n = 20$, $\pi = .75$) 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 at which 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, $Q(380) = 1{,}508.56$, $p = 2.07 \times 10^{-141}$, and

[12] The term *published* is used here very broadly to include publications of conference proceedings and reports that in terms of our literature search were considered unpublished. Of importance, 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.

Table 3
*Basic Study Characteristics: Intentional Studies*

| Characteristic | No. of studies |
|---|---|
| Source of studies | |
| Journal | 277 |
| Conference proceeding | 68 |
| Report | 25 |
| Thesis or dissertation | 8 |
| Book chapter | 2 |
| Number of participants | |
| 1 | 96 |
| >1–10 | 107 |
| >10–20 | 61 |
| >20–30 | 34 |
| >30–40 | 12 |
| >40–50 | 13 |
| >50–60 | 10 |
| >60–70 | 2 |
| >70–80 | 4 |
| >80–90 | 1 |
| >90–100 | 1 |
| >100 | 3 |
| Year of publication | |
| ≤1970 | 14 |
| 1971–1980 | 199 |
| 1981–1990 | 111 |
| 1991–2000 | 40 |
| 2001–2004 | 16 |
| Sample size (bit) | |
| $>10^1$–$10^2$ | 10 |
| $>10^2$–$10^3$ | 62 |
| $>10^3$–$10^4$ | 130 |
| $>10^4$–$10^5$ | 93 |
| $>10^5$–$10^6$ | 41 |
| $>10^6$–$10^7$ | 19 |
| $>10^7$–$10^8$ | 17 |
| $>10^8$–$10^9$ | 5 |
| $>10^9$ | 3 |

remained extremely heterogeneous even when the three largest studies were removed from the sample, $Q(377) = 1{,}489.99$, $p = 2.50 \times 10^{-138}$. This heterogeneity may be the reason for the large difference in effect size between the FEM and REM. Even when the three largest studies were removed, the difference between the two models was highly significant ($\Delta z = 3.34$, $p = .0008$).

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 were 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).

### *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: $\Delta z = -0.22$, $p = .82$; REM: $\Delta z = -1.60$, $p = .11$). Similarly, studies that reported all data did not have different effect sizes from studies that did not report all the data (FEM: $\Delta z = -0.76$, $p = .45$; REM: $\Delta z = -0.41$, $p = .68$). 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: $\Delta z = 0.07$, $p = .94$; REM: $\Delta z = -1.31$, $p = .19$), as well as the difference between the studies that reported all data and those that did not report all the data (FEM: $\Delta z = -0.18$, $p = .86$; REM: $\Delta z = 1.17$, $p = .24$) remained nonsignificant.

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 that had preplanned their split of data was significantly smaller than that of the studies with a post hoc split ($\Delta z = 2.46$, $p = .01$). This finding was mirrored in the REM, in which, again, studies with a preplanned split had a considerably smaller effect size than did studies with a post hoc split ($\Delta z = 5.42$, $p = 6.01 \times 10^{-8}$). 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

Table 4
*Overall Sample Summary Statistics*

| | | Fixed-effects model | | | Random-effects model | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample | *n* | $\bar{\pi}$ | *SE* | *z* | $\bar{\pi}$ | *SE* | *z* | *M* bit | *Mdn* bit | *M* py | *Q* |
| Overall | 380 | .499997 | .000001 | −3.67*** | .500035 | .000014 | 2.47* | 787,888,669 | 8,596 | 1981 | 1,508.56*** |
| Overall − 3 largest | 377 | .500048 | .000013 | 3.59*** | .500286 | .000070 | 4.08*** | 3,707,412 | 8,039 | 1981 | 1,489.99*** |

*Note.* py = publication year.
* $p < .05$. *** $p < .001$.

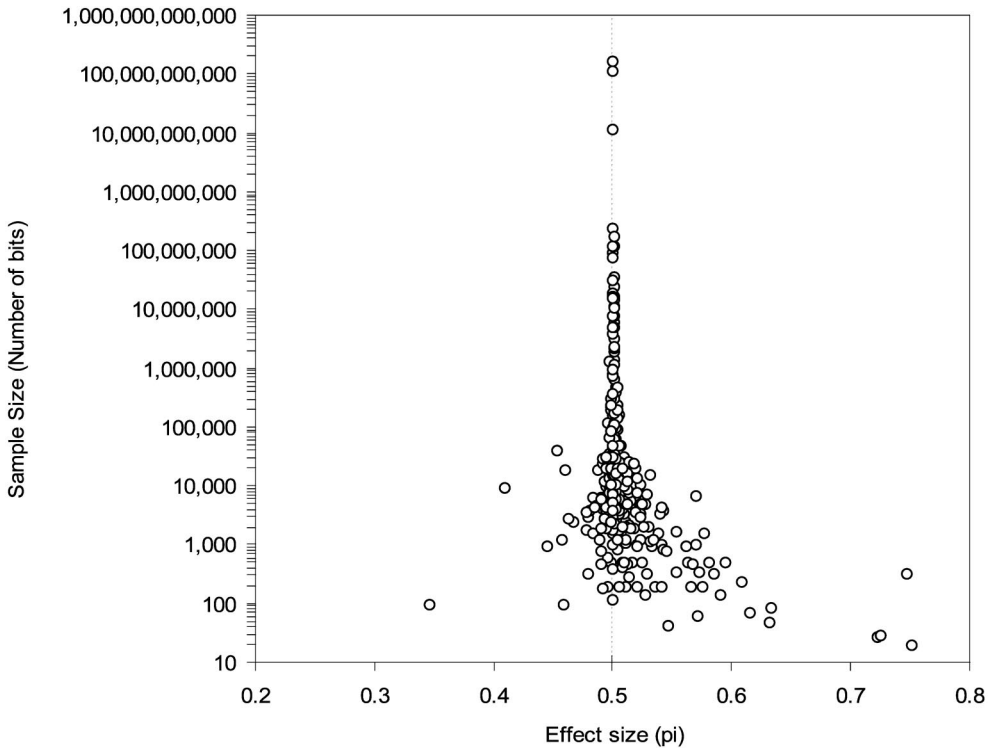


*Figure 2.* Funnel plot intentional studies with respect to number of bits.

effect sizes. As indicated in Table 5, study quality was also positively correlated with year of publication, $r(380) = .29$, $p = 8.27 \times 10^{-9}$, 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.

### *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. Besides sample size and year of publication, we analyzed five additional central moderator variables.

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 one that was 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 (see Table 6).[13] 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 ($\Delta z > 5.00$, $p < 5.74 \times 10^{-7}$). The trend is continuous: The smaller the sample size, the bigger the effect size. Sterne, Gavaghan, and Egger (2000) called this the "small-study effect." The funnel plot (see Figure 2) illustrates the effect. Whereas the bigger studies distribute symmetrically around the overall effect size, the distribution of studies below 10,000 bits is increasingly asymmetrical.

With respect to the mean year of publication, the quartile with the largest studies (Q4) stands out from the other three, smaller-study quartiles. The largest studies were, 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 et al., 2004; Jahn, Mischo, et al., 2000; Nelson, 1994).

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

[13] The smallest effect size is the effect size closest to the theoretical mean value of $\bar{\pi} = .50$. 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 number of participants variable) as when the three largest studies were included.

Table 5
*Safeguard Variables' Summary Statistics*

| Variable and class | *n* | Fixed-effects model $\bar{\pi}$ | *SE* | *z* | Random-effects model $\bar{\pi}$ | *SE* | *z* | *M* bit | *Mdn* bit | *M* py | *Q* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RNG control | | | | | | | | | | | |
| Yes (2) | 269 | .499997[a] | .000001 | −3.67 | .500029 | .000012 | 2.32* | 111,261,910 | 12,288 | 1983 | 911.68*** |
| Earlier (1) | 7 | .499996 | .000051 | −0.08 | .521295 | .993298 | 6.46*** | 13,471,208 | 1,000 | 1982 | 286.75*** |
| No (0) | 104 | .500038 | .000188 | 0.20 | .501101 | .000668 | 1.65* | 85,177 | 4,838 | 1977 | 310.09*** |
| All data reported | | | | | | | | | | | |
| Yes (2) | 311 | .499997[a] | .000001 | −3.68 | .500033 | .000014 | 2.32** | 962,583,297 | 8,192 | 1982 | 1,405.71*** |
| Unclear (1) | 11 | .501074 | .000537 | 2.00* | .500927 | .000882 | 1.05 | 80,726 | 37,000 | 1976 | 16.75 |
| No (0) | 58 | .500063 | .000087 | 0.72 | .500101 | .000163 | 0.62 | 575,876 | 7,750 | 1980 | 81.50 |
| Split of data | | | | | | | | | | | |
| Preplanned (2) | 253 | .499997[b] | .000001 | −3.46 | .500012[a] | .000016 | 0.74 | 113,250,870 | 10,000 | 1982 | 761.78*** |
| Unclear (1) | 50 | .500060 | .000017 | 3.54*** | .500105 | .000067 | 1.58 | 17,356,282 | 19,000 | 1982 | 167.74*** |
| Post hoc (0) | 77 | .499989[a] | .000005 | −2.37 | .504052 | .000745 | 5.54*** | 155,911,422 | 4,600 | 1979 | 562.36*** |
| Safeguard sum score | | | | | | | | | | | |
| 6 (highest) | 159 | .499997[b] | .000001 | −3.47 | .500007[a] | .500007 | 0.47 | 1,801,262,569 | 11,360 | 1984 | 479.52*** |
| 5 | 47 | .500054 | .000016 | 3.36*** | .500132 | .000069 | 1.93* | 20,402,2,900 | 48,000 | 1983 | 206.02*** |
| 4 | 106 | .499989[b] | .000005 | −2.36 | .500472[a] | .000292 | 1.61 | 113,487,404 | 6,400 | 1979 | 405.62*** |
| 3 | 8 | .515664 | .002616 | 5.99*** | .544965 | .511953 | 2.67** | 4,635 | 2,880 | 1978 | 224.87*** |
| 2 | 44 | .499910 | .000297 | −0.30 | .501504 | .001075 | 1.40 | 72,014 | 3,146 | 1977 | 130.55*** |
| 1 | 9 | .500000 | .000250 | 0.00 | .500000 | .000250 | 0.00 | 445,209 | 1,600 | 1976 | 0.00 |
| 0 (lowest) | 7 | .500398 | .000470 | 0.85 | .502072 | .001267 | 1.63 | 161,714 | 25,000 | 1979 | 9.88 |

*Note.* py = publication year; RNG = random number generator.
[a] With the three largest studies removed from the sample, the effect size is significantly larger ($p < .05$, $z > 1.96$) than the mean chance expectation (MCE). [b] With the three largest studies removed from the sample, the effect size is larger than .50 (MCE) but not significantly so.
* $p < .05$. ** $p < .01$. *** $p < .001$.

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: $\Delta z = -1.01$, $p = .31$; REM: $\Delta z = 0.23$, $p = .82$) 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: $\Delta z = 6.77$, $p = 1.26 \times 10^{-11}$; REM: $\Delta z = 3.94$, $p = 8.29 \times 10^{-5}$). The half with the smaller studies ($n = 49$, $M = 810$, $Mdn = 500$) has a much larger effect size (FEM: $\bar{\pi} = .522382$, $SE = .002546$, $z = 8.79$, $p < 1.00 \times 10^{-10}$; REM: $\bar{\pi} = .536425$, $SE = .007216$, $z = 5.05$, $p = 4.48 \times 10^{-7}$) than does the half with the larger studies ($n = 50$, $M = 34{,}011$, $Mdn = 9{,}630$; FEM: $\bar{\pi} = .504926$, $SE = .000398$, $z = 12.38$, $p < 1.00 \times 10^{-10}$; REM: $\bar{\pi} = .507557$, $SE = .001312$, $z = 5.76$, $p = 8.44 \times 10^{-9}$). The difference in 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 1 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 nonselected participants, a claim that 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: $\Delta z = 4.02$, $p = 5.90 \times 10^{-5}$; REM: $\Delta z = 6.85$, $p < 1.00 \times 10^{-10}$) and remains so with the three largest studies removed (FEM: $\Delta z = 3.69$, $p = 2.22 \times 10^{-4}$; REM: $\Delta z = 6.73$, $p < 1.00 \times 10^{-10}$). 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: $M = 187{,}290$, $Mdn = 8{,}000$; unselected: $M = 5{,}369{,}064$, $Mdn = 13{,}968$).

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 to be published if they yield unpromising directions for further study. In this sample, pilot studies are, as one would expect, smaller than formal studies. With respect to their FEM effect size, pilot and formal studies do not differ ($\Delta z = 1.46$, $p = .15$). However, with respect to their REM effect, they differ considerably ($\Delta z = -3.31$, $p = 9.17 \times 10^{-4}$). 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 (FEM: $\bar{\pi} = .500061$, $SE = .000034$, $z = 1.80$,

Table 6
*Moderator Variables' Summary Statistics*

| Variable and class | *n* | Fixed-effects model $\bar{\pi}$ | Fixed-effects model *SE* | Fixed-effects model *z* | Random-effects model $\bar{\pi}$ | Random-effects model *SE* | Random-effects model *z* | *M* bit | *Mdn* bit | *M* py | *Q* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample size (bit) | | | | | | | | | | | |
| Smallest (Q1) | 95 | .519908 | .002070 | 9.61*** | .525523 | .004616 | 5.23*** | 641 | 490 | 1978 | 393.31*** |
| Small (Q2) | 95 | .506320 | .000788 | 8.02*** | .505900 | .001541 | 3.83*** | 4,726 | 4,900 | 1979 | 333.86*** |
| Large (Q3) | 96 | .502087 | .000362 | 5.76*** | .502355 | .000703 | 3.35*** | 21,833 | 20,034 | 1980 | 331.69*** |
| Largest (Q4) | 94 | .499997[a] | .000001 | −3.70*** | .500009[a] | .000008 | 1.19 | 3,185,054,132 | 727,620 | 1989 | 259.46*** |
| Year of publication | | | | | | | | | | | |
| Oldest (Q1) | 99 | .505342 | .000393 | 13.60*** | .511509 | .001505 | 7.65*** | 17,578 | 3,000 | 1972 | 719.66*** |
| Old (Q2) | 96 | .500194 | .000148 | 1.31 | .500811 | .000369 | 2.20* | 119,912 | 6,800 | 1979 | 185.03*** |
| New (Q3) | 103 | .500382 | .000115 | 3.33*** | .500702 | .000307 | 2.28* | 187,156 | 12,288 | 1983 | 230.00*** |
| Newest (Q4) | 82 | .499997[a] | .000001 | −3.73*** | .500003 | .000006 | 0.47 | 3,650,794,697 | 380,000 | 1996 | 175.69*** |
| Number of participants | | | | | | | | | | | |
| One: 1 (Q1) | 96 | .500499 | .000130 | 3.84*** | .503208 | .000610 | 5.26*** | 171,288 | 7,640 | 1981 | 644.17*** |
| Few: 2–10 (Q2) | 107 | .499995[b] | .000001 | −3.53*** | .500025[a] | .000030 | 0.83 | 1,216,285,332 | 5,000 | 1980 | 339.94*** |
| Several: 11–20 (Q3) | 61 | .499997[b] | .000001 | −2.07* | .500190 | .000164 | 1.16 | 2,755,175,923 | 12,288 | 1981 | 169.39*** |
| Many: 21–299 (Q4) | 80 | .500033 | .000015 | 2.14* | .500001 | .000043 | 0.03 | 13,026,064 | 22,446 | 1984 | 140.90*** |
| Unknown | 36 | .500123 | .000044 | 2.80** | .500453 | .000180 | 2.51* | 3,636,208 | 17,875 | 1984 | 183.66*** |
| Participants | | | | | | | | | | | |
| Selected | 59 | .500603 | .000151 | 3.99*** | .506450 | .000939 | 6.87*** | 187,290 | 8,000 | 1977 | 578.98*** |
| Unselected | 261 | .499997[a] | .000001 | −3.69*** | .500020[a] | .000011 | 1.84 | 1,147,069,802 | 15,057 | 1982 | 720.20*** |
| Other | 60 | .500408 | .000422 | 0.97 | .504691 | .001308 | 3.59*** | 23,761 | 1,280 | 1981 | 183.34*** |
| Study status | | | | | | | | | | | |
| Formal | 209 | .499997[a] | .000001 | −3.31*** | .500024 | .000013 | 1.84 | 1,374,014,360 | 12,000 | 1982 | 668.85*** |
| Pilot | 160 | .499990[b] | .000005 | −2.17* | .500493 | .000141 | 3.50*** | 76,366,304 | 7,350 | 1980 | 813.15*** |
| Other | 11 | .500325 | .000157 | 2.07* | .500505 | .000481 | 1.05 | 916,957 | 7,926 | 1979 | 23.09* |
| Feedback | | | | | | | | | | | |
| Visual | 227 | .500030 | .000016 | 1.81 | .500228 | .000092 | 2.48* | 4,149,925 | 6,400 | 1980 | 845.78*** |
| Auditory | 34 | .502377 | .000382 | 6.22*** | .505422 | .001392 | 3.90*** | 51,695 | 18,100 | 1976 | 253.38*** |
| Other | 119 | .499997[a] | .000001 | −3.79*** | .500009 | .000011 | 0.83 | 2,508,015,996 | 20,000 | 1986 | 366.54*** |
| Random sources | | | | | | | | | | | |
| Noise | 228 | .499997[a] | .000001 | −3.68*** | .500026 | .000012 | 2.13* | 1,313,136,638 | 18,375 | 1985 | 913.03*** |
| Radioactive | 93 | .503354 | .000601 | 5.58*** | .509804 | .001778 | 5.51*** | 8,339 | 2,000 | 1974 | 467.69*** |
| Other | 59 | .500945 | .000382 | 2.48* | .501562 | .000633 | 2.47* | 29,920 | 13,600 | 1979 | 93.41** |

*Note.* py = publication year; Q = quartile.
[a] With the three largest studies removed from the sample, the effect size is significantly larger ($p < .05$, $z > 1.96$) than the mean chance expectation (MCE). [b] With the three largest studies removed from the sample, the effect size is larger than .50 (MCE) but not significantly so.
* $p < .05$. ** $p < .01$. *** $p < .001$.

$p = .07$; REM: $\bar{\pi} = .500701$, $SE = .000195$, $z = 3.59$, $p = 3.37 \times 10^{-4}$) studies 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 feedback, and some provide auditory feedback. Besides these two main categories, the coding resulted in a large "other" category, with 119 studies that 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: $\Delta z = 6.14$, $p = 8.49 \times 10^{-10}$; REM: $\Delta z = 3.72$, $p = 1.96 \times 10^{-4}$) but also the studies in the "other" category (FEM: $\Delta z = 6.23$, $p = 4.74 \times 10^{-10}$; REM: $\Delta z = 3.89$, $p = 1.01 \times 10^{-4}$). 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: $\Delta z = 5.59$, $p = 2.28 \times 10^{-8}$; REM: $\Delta z = 5.50$, $p = 3.86 \times 10^{-8}$). And although the effect size of the studies using noise becomes significantly different from the MCE when the three largest studies, all noise based, are removed from the sample (FEM: $\bar{\pi} = .500045$, $SE = .000013$, $z = 3.39$, $p = 7.12 \times 10^{-4}$; 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 that for studies using noise (FEM: $\Delta z = 5.51$, $p = 3.65 \times 10^{-8}$; REM: $\Delta z = 5.41$, $p = 5.41 \times 10^{-8}$). 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 ($M = 6{,}200{,}682$ bit, $Mdn = 17{,}000$ bit) than the sample size of the radioactive-

Table 7
*Summary of the Weighted Metaregression: Regression Model 1 (Sample Size)*

| Variable | Fixed-effects model | | | Random-effects model | | |
|---|---|---|---|---|---|---|
| | *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. RNG = random number generator.
* $p < .05$. ** $p < .01$. *** $p < .001$.

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 2 years above the mean year of publication of the oldest studies in our sample (see Table 6).

### *Metaregression 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: $Q(17) = 121.76$, $p = 7.11 \times 10^{-18}$; REM: $Q(17) = 99.93$, $p = 9.17 \times 10^{-14}$—the unaccounted residual variance is considerable—FEM: $Q(362) = 1{,}386.80$, $p = 1.16 \times 10^{-119}$; REM: $Q(362) = 1{,}361.73$, $p = 1.22 \times 10^{-115}$. 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.[14]

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: $Q(17) = 233.45$, $p = 4.93 \times 10^{-40}$; REM: $Q(17) = 212.19$, $p = 1.00 \times 10^{-35}$—the unaccounted residual variance again remains considerable—FEM: $Q(362) = 1{,}275.12$, $p = 5.84 \times 10^{-102}$; REM: $Q(362) = 1{,}262.44$, $p = 4.48 \times 10^{-100}$—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 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, to adequately assess potential moderators.

### *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 ($r_s = -.33$, $p = 4.38 \times 10^{-11}$) also confirms the asymmetric distribution of effect size. Use of Duval and Tweedie's (2000) trim and fill approach found that 83 studies had to 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: $\bar{\pi} = .500036$, $SE =$

[14] Because of the extremely lopsided distribution of sample size, the log10 transformation that we use throughout the article was also used for the regression analysis. However, exactly the same results occurred when sample size entered the regression model as a continuous variable; that is, the same variables were or were not significant, and even the variance remained identical (FEM: 8.2%; REM: 7.1%).

Table 8
*Summary of the Weighted Metaregression: Regression Model 2 (Sample Size Quartiles)*

| Variable | Fixed-effects model | | | Random-effects model | | |
|---|---|---|---|---|---|---|
| | *B* | $SE_B$ | *z* | *B* | $SE_B$ | *z* |
| Sample size quartiles | −.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. RNG = random number generator.
* $p < .05$. ** $p < .01$. *** $p < .001$.

.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 1 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.

### *Monte Carlo Simulation*

The averaged results of the simulation of 1,000 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 they did 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 1,544 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" 100% and 90% of the time

Table 9
*Step-Weight Function Monte Carlo Simulation of Publication Bias*

| Simulation | *n* | Fixed-effects model | | | | Random-effects model | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\bar{\pi}$ | *SE* | *z* | Δ*z* | $\bar{\pi}$ | *SE* | *z* | Δ*z* | *Q* | Stud |
| Overall | 380 | .500001 | .000001 | 1.29 | −3.51*** | .500024 | .000009 | 2.68** | 0.62 | 631.58*** | 1,544 |
| Sample size | | | | | | | | | | | |
| Smallest (Q1) | 95 | .511582 | .002024 | 5.72*** | 2.88** | .512474 | .002478 | 5.07*** | 2.49* | 125.87* | 389 |
| Small (Q2) | 95 | .504629 | .000746 | 6.20*** | 1.56 | .504705 | .000849 | 5.58*** | 0.68 | 119.01* | 384 |
| Large (Q3) | 96 | .502145 | .000345 | 6.21*** | −0.12 | .502192 | .000393 | 5.61*** | 0.20 | 119.47* | 390 |
| Largest (Q4) | 94 | .500001 | .000001 | 1.27 | −3.51*** | .500009 | .000005 | 1.70 | 0.02 | 153.68*** | 381 |

*Note.* Δ*z* = difference between effect sizes of simulated and experimental data. Stud = number of unpublished studies (simulated). Q = quartile.
* $p < .05$. ** $p < .01$. *** $p < .001$.

and with studies in the second step ($p \leq .05$ and $p > .01$) to be "published" 90%, 80%, and 70% of the time. For each of the 162 ($2 \times 3 \times 3 \times 3 \times 3$) possible combinations of limit values, 1,000 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. The minimum values for all effect sizes and $z$ scores come from a single set of parameters (90% of $p \leq .01$; 70% of $p > .01$ and $p \leq .05$; 40% of $p > .05$ and $p \leq .10$; 10% of $p > .10$ and $p \leq .50$, and 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 negative $z$ 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.

## Discussion

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

### *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 1,000 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 affect 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 is considered 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

Table 10
*Limit Values of the Step-Weight Function Monte Carlo Simulation Dependent on the Initial Weighting (±10%)*

| Variable | Smallest studies (Q1) | Small studies (Q2) | Large studies (Q3) | Largest studies (Q4) | Overall sample |
|---|---|---|---|---|---|
| $\bar{\pi}_f$ | | | | | |
| Min | .504542 | .501709 | .500831 | .500000 | .500000 |
| Max | .523420 | .509265 | .504289 | .500002 | .500002 |
| $z_f$ | | | | | |
| Min | 2.24 | 2.29 | 2.41 | 0.46 | 0.48 |
| 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 |
| $\bar{\pi}_r$ | | | | | |
| Min | .505130 | .501769 | .500862 | .500003 | .500008 |
| Max | .523970 | .509269 | .504291 | .500029 | .500069 |
| $z_r$ | | | | | |
| Min | 1.83 | 1.83 | 1.93 | 0.64 | 0.99 |
| Max | 11.20 | 12.39 | 12.39 | 4.02 | 6.06 |

*Note.* Variables $\bar{\pi}_f$ and $z_f$ are parameter estimates based on a fixed-effects model. Variables $\bar{\pi}_r$ and $z_r$ are parameter estimates based on a random-effects model. Min = minimum; Max = maximum; Stud = number of unpublished studies (simulated).

the alleged concurrent interaction between direct human intention and RNGs. All previous meta-analyses also included nonintentional and nonhuman 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 that, therefore, the findings based on this model must not be considered. However, empirically it is impossible to decide whether the model is adequate. 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 are also not unambiguous results because the findings 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 and Tweedie's (2000) trim and fill approach suggests that the REM $z$ score drops from 4.08 to 2.71 after adding the 73 missing studies. However, the most important finding with respect to 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. Although there is a significant correlation between effect size and study quality, the relationship is too small to account for the overall effect. Moreover, any comprehensive explanation of the data would also have to take into account the extreme heterogeneity and the small-study effect observed in the data.

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 the MCE, and the finding that the control sample was homogeneous, demonstrate that the RNGs were not malfunctioning.

### *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 demonstrated 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. Of 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 are newer studies. Other variables affecting the publication process might also have changed over time. However, we have only modeled a simple selection process, and therefore, these arguments must be considered speculative.

### *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: (a) true heterogeneity, (b) data irregularities, and (c) 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.

*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 because 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.

On the basis of 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 (Schmidt, 1969, 1973, 1974a, 1974b; Schmidt & Pantas, 1972), 50% of the studies that used auditory feedback (Schmidt, 1972, 1973, 1976; Schmidt & Terry, 1977),

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 M. Egger, G. D. Smith, and D. Altman (Eds.), *Systematic Reviews in Health Care: Meta-Analysis in Context,* London: BMJ Books. Copyright 2001 by Blackwell Publishing. Adapted with permission.

and 29% of the studies that used radioactive random sources (Schmidt, 1969, 1978; Schmidt & Pantas, 1972). 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.

*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 exclude 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 are 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.

*Selection biases.* When the inclusion of studies in a meta-analysis is systematically biased in a way that smaller studies with more significant $p$ values, that is, larger effect sizes, are more likely to be included than larger studies with less significant $p$ 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 author's 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 because of 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 (1979) "file drawer" calculation, 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 and Tweedie's (2000) trim and fill approach 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.

### *Monte Carlo Simulation*

The straightforward simulation is in good agreement with all three major findings of this meta-analysis and is particularly persuasive with respect to 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 1,500 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" (as cited in Broughton, 1987, p. 27), a view which at that time was probably not uncommon in other research areas as well. However, in 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.[15]

The number of "unpublished" studies in the simulation reflects not only 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 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 studies were all independently statistically highly significant. There is a highly significant correlation between the study split and effect size, $r(380) = .36$, $p = 5.68 \times 10^{-30}$. Studies from an experimental report that was split into several studies produce larger effect sizes than do studies from an experimental report that 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 wrote 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 argued 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 reasoned that

> A typical experiment, like the one reported in this paper (8 Ss [subjects] 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 [sic] 20 experiments per year. For the past 5 years this could imply that 3000 experiments were done which never reached the outside world (pp. 33f)

With 131,072 random bits, Bierman's experiment is typical with respect to sample size.[16] 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 taken into account here. From this perspective, the proposed 1,500 unpublished studies do not appear to be a wholly unreasonable number. Meta-analyzing RNG studies would certainly be a lot easier if all experimenters registered the hypothesis, the sample size, and the preplanned analyses to an external body before conducting their studies.

### *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 and 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 and 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, and Utts (1985; see also Dobyns, 1996; May, Utts, & Spottiswoode, 1995), 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 push" of the participant, who, according to the model, precognitively scans the future behavior of the RNG and selects the time at which there are "locally deviant subsequences 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, in that one anomalous concept, namely PK, 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 1s and 0s. 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.

## Conclusion

The statistical significance of the overall database provides no directive as to whether the phenomenon is genuine. The difference between the two statistical models used (FEM and REM) and the dependency of the results on three very large studies demonstrate the difficulties regarding these data. If the striking heterogeneity and the small-study effect are taken into account, one must ask

[15] Although the change is particularly interesting for the *Journal of Parapsychology,* these data are not very reliable because almost 60% ($n = 47$) of the studies published in this journal 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 this journal.

[16] The sample size is based on 8 subjects participating in 16 runs with 16 intervals with 64 bits.

Table 12
*Reported Split of Studies per Published Report*

| Study split of report | $n$ | Fixed-effects model | | | Random-effects model | | | $M$ bit | *Mdn* bit | $M$ py | $Q$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\bar{\pi}$ | *SE* | $z$ | $\bar{\pi}$ | *SE* | $z$ | | | | |
| 1 | 30 | .499983 | .000044 | −0.40 | .500205 | .000317 | 0.65 | 4,447,013 | 8,968 | 1985 | 73.23*** |
| 2 | 66 | .499998 | .000045 | −0.04 | .500119 | .000233 | 0.51 | 1,842,341 | 12,517 | 1983 | 190.60*** |
| 3 | 27 | .500017 | .000023 | 0.75 | .500124 | .000124 | 1.00 | 17,490,052 | 24,000 | 1984 | 142.15*** |
| 4 | 80 | .501061 | .000200 | 5.31*** | .503079 | .000712 | 4.32*** | 82,438 | 7,440 | 1979 | 442.74*** |
| 5 | 35 | .500004 | .000083 | 0.05 | .500097 | .000179 | 0.54 | 1,034,624 | 30,000 | 1985 | 51.34* |
| 6[a] | 48 | .499997[b] | .000001 | −3.75*** | .499999[c] | .000005 | −0.01 | 6,219,133,881 | 23,400 | 1984 | 102.60*** |
| 7 | 21 | .500052 | .000048 | 1.09 | .500308 | .000222 | 1.39 | 5,172,284 | 24,247 | 1982 | 131.53*** |
| 8 | 16 | .501382 | .001491 | 0.93 | .502627 | .002727 | 0.96 | 7,024 | 7,552 | 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*** | 42,727 | 10,000 | 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* |

*Note.* py = publication year.
[a] It should be noted that the experimental report with the three by far largest studies (Dobyns, Dunne, & Nelson, 2004) also includes three smaller studies. [b] With the three largest studies removed from the sample, the effect size is significantly larger ($p < .05$, $z > 1.96$) than the mean chance expectation (MCE). [c] With the three largest studies removed from the sample, the effect size is larger than 50 (MCE) but not significantly so.
* $p < .05$. ** $p < .01$. *** $p < .001$.

whether the findings are artifactual or 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 that have to be unpublished is considerable ($N$ = 1,500), Bierman's (1987) 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 remind the reader that these experiments are highly refined operationalizations of a phenomenon that has challenged humankind 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 was 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 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 researchers 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 (1962b) verdict of "not proven" (p. 530), which he mooted more than 40 years ago in the same journal with respect to dice experiments, also holds for human intentionality on RNGs.

## 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*(2), 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*(4), 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.

Bohr, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Physical Review, 48,* 696–702.

*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*(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*(3), 6–16.

*Braud, W. G., & Kirk, J. (1978). Attempt to observe psychokinetic influences upon a random event generator by person–fish teams. *European Journal of Parapsychology, 2*(3), 228–237.

Braude, S. E. (1997). *The limits of influence: Psychokinesis and the philosophy of science* (Rev. 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: British Medical Journal 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, 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: British Medical Journal 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.

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* (Vol. 8, pp. 88–125). Jefferson, NC: McFarland 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.

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.

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.

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* (pp. 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: Proceedings 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, October). Le factor psi séxerce sur un appareil électronique [The psi factor affects an electronic apparatus]. *Psi-Réalité, 1,* 54–62.

Lipsey, M. W., & Wilson, D. B. (2001). *Applied social research methods series: Vol. 49. Practical meta-analysis.* 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.

*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. (1985). Psi experiments with random number generators: An informational model. In *The Parapsychological Association 28th Annual Convention: Proceedings of presented papers* (pp. 235–266). Durham, NC: Parapsychological Association.

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 misjudgment, 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.

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 and 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: Holt.

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.

*Psychophysical Research Laboratories. (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 Convention: 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 *p. 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'Étranger, 18,* 609–674.

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

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.

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 Press.

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 Press.

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* (Tech. Rep. 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. (1974a). Comparison of PK action on two different random number generators. *Journal of Parapsychology, 38,* 47–55.

*Schmidt, H. (1974b). 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). 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. (1975). Toward a mathematical theory of psi. *Journal of the American Society for Psychical Research, 69,* 301–319.

*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. (1993). *Mind, matter, and quantum mechanics.* Berlin: Springer.

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

Steichen, T. J. (2004). Stata (Version 6.0, metatrim module) [Computer software]. Retrieved from http://www.stata.com/stb/stb61/sbe39_2/

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., Egger, M., & Smith, G. D. (2001). Investigating and dealing with publication and other biases. In M. Egger, G. D. Smith, & D. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (pp. 189–208). London: BMJ Books.

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* (Vol. 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: Pro-*

*ceedings 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. L. 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, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21,* 1559–1574.

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

*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.

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.

von Lucadou, W., & 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.

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

Walker, E. H. (1975). Foundations of paraphysical and parapsychological phenomena. In L. Oteri (Ed.), *Quantum physics and parapsychology* (pp. 1–53). New York: Parapsychology Foundation.

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.

Received August 23, 2004
Revision received July 7, 2005
Accepted July 14, 2005 ■

Psychological Bulletin
2006, Vol. 132, No. 4, 524–528


0033-2909/06/$12.00 DOI: 10.1037/0033-2909.132.4.524

# On Blowing Trumpets to the Tulips: To Prove or Not to Prove the Null Hypothesis—Comment on Bösch, Steinkamp, and Boller (2006)

David B. Wilson
George Mason University

William R. Shadish
University of California, Merced

The H. Bösch, F. Steinkamp, and E. Boller (2006) meta-analysis reaches mixed and cautious conclusions about the possibility of psychokinesis. The authors argue that, for both methodological and philosophical reasons, it is nearly impossible to draw any conclusions from this body of research. The authors do not agree that any significant effect at all, no matter how small, is fundamentally important (Bösch et al., 2006, p. 517), and they suggest that psychokinesis researchers focus either on producing larger effects or on specifying the conditions under which they would be willing to accept the null hypothesis.

*Keywords:* psychokinesis, meta-analysis, null hypothesis significance testing

> The physicist George Darwin used to say that once in a while one should do a completely crazy experiment, like blowing the trumpet to the tulips every morning for a month. Probably nothing will happen, but if something did happen, that would be a stupendous discovery. (Hacking, 1983, p. 154)

Bösch, Steinkamp, and Boller (2006) have provided us with a very intriguing and detailed review of the possibility of psychokinesis, whether the mind can directly influence physical matter. The authors synthesized the results of hundreds of studies of whether people can influence whether a 1 or a 0 appears in a sequence of randomly generated 1s and 0s. The results suggest that "a significant but very small overall effect size" was found (p. 497). The authors conclude that this "effect in general, even if incredibly small, is of great fundamental importance" (p. 517) but also "not proven" (p. 517).

The title of our article is intended to point to a fundamental dilemma that lies at the heart of the Bösch et al. review. If we carried out Darwin's experiment and we blew trumpets to the tulips, how would we know whether something happened? Having been to many outdoor concerts, we can attest that the playing of musical instruments does not appear to have a noticeable effect on the surrounding flora (though the fauna do respond). Yet have we looked closely enough? Do we see a slight swaying? Was it the breeze? Was it the vibrations of the music? Perhaps we should do a randomized experiment, blowing trumpets to the tulips in different settings, at repeated intervals, using different players, with very careful measurement of plant motion, survival, or color to see whether something happened. The methodological problems would be enormous, especially given our expectation that any effect on the tulips would likely be extremely small and so very difficult to measure with accuracy. In this imaginary program of research, it is not at all unlikely to expect profoundly equivocal results. A host of methodological problems would provide viable alternative hypotheses to the observed effect. Moreover, from a philosophical point of view, we must wonder what kind or size of effect would be worth finding, and what kinds of findings might cause us to give up our hope that blowing trumpets to the tulips makes something happen. In the end, we may have to realize that we may never know whether something happens when we blow the trumpet to the tulips.

In this commentary, we argue that a convincing answer to the existence of a psychokinesis effect on random number generators (RNGs) may remain as elusive as the answer to the question of whether something happens when we blow trumpets to the tulips, at least in the absence of what George Darwin would call a "stupendous discovery." We maintain that unless a nontrivial psychokinetic effect is found with the RNG methodology, additional research in this paradigm will continue to lead to an equivocal conclusion.

## Methodological Problems

From a methodological perspective, this meta-analysis has many strengths. Bösch et al. did an admirable job searching for and retrieving all available psychokinesis studies, independent of publication status, and used a well-justified eligibility criteria for establishing which studies to include in the synthesis. The statistical methods used mostly (but not always) reflect current practice. Each effect size, in this case the proportion of 1s or 0s, was weighted by the inverse of its variance under both fixed- and random-effects models. The authors attended to the issue of consistency in the findings across studies (homogeneity) and examined potential moderators of the observed inconsistency of effects (heterogeneity). They also explored the plausibility that publication selection bias affected the overall findings and noted that the small effect they found could be the result of publication bias.

Nonetheless, this meta-analysis contains sufficient methodological problems to make us doubt that we can conclude anything at all from this study, except that there is no evidence of any sizable

David B. Wilson, Department of Public and International Affairs, George Mason University; William R. Shadish, School of Social Sciences, Humanities, and Arts, University of California, Merced.

Correspondence concerning this article should be addressed to David B. Wilson, Administration of Justice Program, Department of Public and International Affairs, George Mason University, 10900 University Boulevard, MS 4F4, Manassas, VA 20110. E-mail: dwilsonb@gmu.edu

effect. The fundamental problem is that the observed effect is so small that even the slightest of methodological problems could change the results from significant to nonsignificant, or vice versa. Indeed, this is exactly what happened as the authors varied how they analyzed the data. The 117 reports representing 380 experimental studies produced an equivocal result. The overall fixed-effects mean is slightly less than the null value (0.499997, i.e., slightly fewer 1s than 0s), and the random-effects mean is slightly greater than the null value (0.500035). Both of these mean effects are statistically significant but in opposite directions.

Consider some of the ways in which this miniscule finding is methodologically vulnerable. The authors apply ordinary meta-analytic statistics to these data, statistics that assume that observations within studies are independent. However, the explicit hypothesis of any psychokinesis study is that people can create dependencies within a run of RNG bits. If so, then the standard error of the effect size from each run is likely to be too small. In addition, further dependencies are caused by having bits nested within runs, nested within people, nested within 380 studies, nested in 117 experimental reports, nested in 59 authors, nested in 33 institutions. These dependencies can also artificially reduce the average effect size standard error. The bias caused by both problems would tend toward finding statistical significance that would not occur if the researcher used a proper analysis such as a multilevel or random coefficients model capable of taking multiple levels of dependency into account (e.g., Goldstein, 2003)

Support for this concern comes from the control data, in which "137 control studies yielded a nonsignificant effect size ($\overline{\pi}$ = .499978, $SE$ = .000015, $z = -1.51$, $p = .13$)" (p. 507). Notice that this effect size is similar to the FEM effect size from the RNG intentionality experiments ($\overline{\pi}$ = .499997, $SE$ = .000001, $z = -3.67$, $p = .000243$). The latter is statistically significant partly because the standard error is 15 times smaller in the RNG studies than in the control studies. Might this be because in the control data no human is present to create the dependencies? That would still provide evidence of a psychokinesis effect, albeit in the unintended direction and on both the mean and standard error rather than just the mean, but would also require an analysis that took the dependencies into account to get accurate Type I error rates.

This problem of potential dependencies in the data pervades every single analysis done in the Bösch et al. review. For example, the trim and fill analysis also assumes independence of observations within studies and between studies. Although it is less clear what the effect of dependency would be on trim and fill analysis, ignoring the problem leads to more skepticism about the reliability of the results. What may be needed is a careful review of how analyses have been and should be conducted in both the primary studies and in meta-analyses like Bösch et al.'s to take potential dependencies into account.

To their credit, Bösch et al. expended considerable effort doing sensitivity analyses, primarily by examining categorical and continuous moderators of study effect. They concluded that study sample size is the most important moderator. They have made a plausible case for this conclusion in many respects, but it is also suspect on three counts. First, in many analyses, what they actually varied was not large sample size but rather one publication containing three studies, all of which used large sample sizes, and they compared analyses in which this one publication was either included or excluded. The problem is that this one publication had many other characteristics associated with it as well, such as the same author, laboratory, and presumably similar operations across the three studies. We cannot be sure it is sample size that is the main influence. Second, the initial regression analysis did not support the importance of sample size, which became significant only in a second regression when it was transformed into an ordinal variable with four categories. Why should we think the quartile analysis is better than the original analysis? Third, even in the quartile analysis, many other variables were significant predictors of effect size, even when sample size quartiles were controlled, including year of publication and the use of selected participants, auditory feedback, noise RNG, and an RNG control. Especially given that year of publication, auditory feedback, and RNG control were significant in both the initial and the quartile regressions, their importance may be as great as sample size. Ironically, the authors end their discussion of regression by saying "a very small overall effect size makes it difficult for any regression analysis, or any meta-analysis or any study, to adequately assess potential moderators" (p. 511). Our point exactly, except we add that it also makes it difficult to believe conclusions about the overall effect size.

It is a truism in meta-analysis that larger studies get more weight, because good statistical theory buttresses the assumption that large studies give more accurate parameter estimates. This creates a problem for Bösch et al. because the three largest studies all report an effect opposite to the intended one—attempting to use the mind to create more 1s in the RNG sequence actually had the effect of producing more 0s. They concluded that "an effect opposite to intention cannot be claimed to be a general finding of this meta-analysis" (p. 513). Thus, they went to some effort to argue that "the three studies are considered to be outliers, and the overall effect found in the meta-analysis is considered to be an effect in the direction intended by the participants in the studies" (p. 513). But what exactly does "outlier" mean? Clearly they are outliers in sample size, but they should receive more weight on that account, not less. They are not outliers in the direction of the effect, for the authors themselves report that "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" (p. 513). We are left without any good arguments for believing that the results of these three studies should be given less weight. Therefore, we are left without any good reason to reject an effect opposite to intention.

## Philosophical Problems

Important philosophical problems also plague the entire line of psychokinesis research described in the Bösch et al. review. The willingness of the psi research community to find any effect, no matter how small, "of great fundamental importance" interacts with a basic weakness of null hypothesis significance testing (NHST) to render nearly any result invulnerable to falsification. Popper (1965) argued that the distinction between a scientific and a pseudoscientific theory is its falsifiability or refutability. From his perspective, a scientific theory must be able to make predictions that can be refuted by observations that are genuinely risky: There must be a real possibility of refutation. Later scholars of

science showed convincingly that such falsification can never be as definitive as Popper hoped. Kuhn (1962) pointed out that falsification depends on two assumptions that can never be fully tested. The first is that the causal claim is perfectly specified. But that is never the case. So, many features of both the claim and the test of the claim are debatable—for example, which outcome is of interest, how it is measured, the conditions of treatment, who needs treatment, and all the many other decisions that researchers must make in testing causal relationships. As a result, disconfirmation often leads theorists to respecify part of their causal theory. For example, they might now specify novel conditions that must hold for their theory to be true and that were derived from the apparently disconfirming observations. Second, falsification requires measures that are perfectly valid reflections of the theory being tested. However, most philosophers maintain that all observation is theory laden. It is laden both with intellectual nuances specific to the partially unique scientific understandings of the theory held by the individual or group devising the test and with the experimenters' extrascientific wishes, hopes, aspirations, and broadly shared cultural assumptions and understandings. If measures are not independent of theories, how can they provide independent theory tests, including tests of causal theories? If the possibility of theory-neutral observations is denied, with them disappears the possibility of definitive knowledge both of what seems to confirm a causal claim and of what seems to disconfirm it.

The psychokinesis research synthesized by Bösch et al. is a particularly cogent example of a paradigm that seems virtually invulnerable to falsification. It is difficult to imagine a result that would disprove the psychokinesis hypothesis if effects as small as .500048 and .499997 are theoretically meaningful. Suppose the result had been .50 plus $10^{-100}$? Would that still be "of great fundamental importance" if it were significantly different from zero? Is there any limit on the size of the effect that psychokinesis researchers would find to be "of great fundamental importance"? What effect size would be so small that psychokinesis researchers would agree to give up the hypothesis? The Bösch et al. review also illustrates how nearly any result can be either dismissed or accepted through the use of ancillary respecifications to the central hypothesis. For example, when the three largest studies yielded a result exactly the opposite to the hypothesized psychokinesis effect, Bösch et al. argued that this result was just an artifact and should not be taken to be a valid result from their review (though it is not clear how future RNG researchers could protect themselves from this alleged artifact except by avoiding large sample studies, exactly the opposite of the usual methodological wisdom).

A similar dilemma applies to anyone who wishes to refute the psychokinesis hypothesis and conclude no such effect exists. The reason concerns the fundamental nature of NHST, especially that a failure to reject a null hypothesis does not establish that there is no effect (Berkson, 1938; Boring, 1919; Carver, 1978). Even if this meta-analysis had observed a mean effect size exactly equal to .5, the confidence interval would still be a nonzero range around .5, establishing the possibility of some non-null value for which those who favor the possibility of psychokinesis could argue is grounds for continued search for evidence of the effects of human intention. Indeed, it may always be plausible to argue that ever larger sample sizes may eventually have sufficient power to distinguish between the null and an effect. This is problematic for any social scientific theory that is willing to accept any effect, no matter how small, as confirmation of the theory. Without an unambiguous acceptance of the null, which NHST cannot provide, the hypothesis and by extension the theory remains plausible.

The limits of NHST with respect to the null can be illustrated with a simple computer simulation. We simulated 1,000 meta-analyses, each with 100 studies. All of the studies had a sample size of 10,000 bits (slightly above the median for the studies included in the Bösch et al. meta-analysis). We used a random numbers function built into the Stata (StataCorp, 2001) software package to generate a random series of 0s and 1s around the population parameter set at $\pi = .5$. Across the 1,000 simulated meta-analyses that included 100,000 simulated studies, the estimate of $\pi$ ranged from .49866 to .50141, with a mean effect size (.500009) that was still not exactly equal to .50. The standard error around these estimates was .0005, producing a 95% confidence interval of $\pm$.00098. The mean effect size from this simulation differed from .50 by a larger amount than that observed by Bösch et al., and the 95% confidence interval from the simulation had a wider range than that observed by Bösch et al., despite being simulated from a distribution for which $\pi = .5$. Clearly, because this range includes values that psychokinesis researchers apparently find "of great fundamental importance," 100 (or more) additional studies similar in size to those conducted to date would still not establish the absence of a psychokinesis effect to their satisfaction.

Examination of the control studies further illustrates the problem of knowing what hypothesis to believe. A subset of 137 of the psychokinesis studies included a control run of the RNG (i.e., a run of the experiment with no human intention). The mean effect size for these control runs was .499978, or .000022 from the null value. Although not statistically significant, this effect was more than 7 times larger than the fixed-effects mean for the experimental runs (.000003 from the null value) and roughly half the size of the random-effects mean (.000048 from the null value). In a context such as this in which the effects are very small, any source of bias becomes highly problematic and raises the plausibility that any finding, null or otherwise, might be due to bias and not psychokinesis.

Although, in a strict sense, the null hypothesis can never be accepted, Cook, Gruder, Henningan, and Faly (1979) argued that there are conditions under which the null may be at least provisionally accepted. These are

> (a) when the theoretical conditions necessary for the effect to occur have been explicated, operationalized, and demonstrably met in the research; (b) when all the known plausible countervailing forces have been explicated, operationalized, and demonstrably ruled out; (c) when the statistical analysis is powerful enough to detect at least the theoretically expected maximum effect at a preordained alpha level; and (d) when the manipulations and measures are demonstrably valid. (p. 668)

It is worth considering these criteria in the context of RNG psychokinesis research to clarify the conditions that might have to be present for psychokinesis researchers to accept the conclusion that there is no psychokinesis effect.

Cook et al.'s (1979) first and fourth criteria both address construct validity. In this context, the issue is how well the experimental task of having a human subject try to influence an RNG process represents the construct of psychokinesis. RNG experi-

ments that fail to find evidence for a psychokinesis effect can be discounted as imperfect tests of the theory because they do not represent the construct of psychokinesis sufficiently well. Bösch et al. were aware of this when they concluded that the RNG experiments "may not necessarily bear a direct relation to purported large-scale phenomena" (p. 517) such as those reported in séance rooms. From Popper's (1965) perspective, therefore, RNG studies may not represent risky tests of the psychokinesis hypothesis because they can too easily be dismissed as insufficiently representative of the constructs of interest. Thus, even if RNG evidence fails to clearly establish a psychokinesis effect, this failure may not translate into a refutation of psychokinesis in general. Therefore, the psychokinesis research community needs to decide whether RNG tests of the hypothesis are of sufficient theoretical centrality to warrant continued study. If not, more cogent and risky tests need to be devised.

Cook at al's (1979) second criterion addresses whether all the plausible alternative explanations to a psychokinesis effect have been identified and ruled out. In many respects, this is the strength of the RNG paradigm for testing psychokinesis. Bösch et al. described all the artifacts that were solved by using RNG tests. Here, it seems to us that the key task is distinguishing a confounding artifact from a moderator of theoretical interest. Consider the variables in Bösch et al.'s Table 7. Year of publication is most likely a confounding artifact arising because of technology changes in how studies are conducted. We have no reason to think that the psychokinesis itself has changed over time, either as a function of time or of time-varying covariates such as the potentially mind-numbing effects of TV. Auditory feedback, however, could reflect a phenomenon of theoretical interest about the channels through which psychokinesis might work. Though we are insufficiently versed in psychokinesis research to make such discriminations, we do claim that the question of psychokinesis is unlikely to be settled without attention to them.

Cook et al.'s (1979) third criterion of adequate statistical power presumes that there are statistical effects sufficiently small as to be theoretically equivalent to the null. Though RNG psychokinesis researchers have not specified what such effects might be, researchers in other areas have shown it can be done. For example, Bickman, Lambert, Andrade, and Penaloza (2000) accepted the no difference null in the Fort Bragg continuum of care for children's mental health services evaluation. The statistical power was sufficient to provide a confidence interval centered on zero with a range that did not extend beyond a trivial and theoretically null effect. Another example is equivalence testing in drug trials research. Equivalence testing is used to establish the equivalence of two drugs, such as a generic and brand name version of the same drug (e.g., Jones, Jarvis, Lewis, & Ebbut, 1996), a key ingredient of which is postulating the smallest population effect of theoretical interest. All this suggests two improvements to RNG psychokinesis research: (a) specifying the smallest population effect of theoretical interest and (b) using methods such as equivalency testing rather than an exclusive reliance on NHST.

## An Effect of Fundamental Importance?

Psychokinesis researchers need to better justify the claim that effects of the minute magnitude found in the Bösch et al. review are "of great fundamental importance", for the validity of the claim is not obvious to many of researchers outside psychokinesis. For example, one could argue that this effect is important because of practical application. Imagine using psychokinesis to influence the outcome of tossing silver dollars, winning a dollar if the toss came up heads and losing it otherwise. After 500,000 tosses, which we estimate would take nearly 2 months of nonstop tossing, we might be either $48 ahead or $3 behind. This is not "of great fundamental importance." Perhaps, however, we could find a way to manipulate the binary process of a computer in a way that accomplishes some good end. If this is the claim, Bösch et al. should spell out the supposed details, showing exactly the kind of manipulation that they have in mind, how a human would have any idea what was going on in the computer that they should influence, all connected to how a very small effect on the computer's binary language could result in some practical outcome.

Perhaps there is no practical importance of this result, but the claim is that the result is of fundamental theoretical importance. If so, psychokinesis researchers should spell out the theoretical importance. After all, providing evidence that the mind can influence physical reality would not be new—it is the basis of any number of randomized experiments in medicine that demonstrate an influence of psychological interventions on physical health, ranging from a patient who meditates and achieves better health to a psychologist who applies biofeedback techniques to influence the patient's mind and in turn the patient's body. Or maybe the claim is that the results would be of fundamental theoretical or practical importance if only they were bigger, or more reliably produced, or more generalizable across different operationalizations of psychokinesis. No doubt psychokinesis researchers have already addressed this question, but for those researchers not in the field, the claim as reflected in the Bösch et al. review is not compelling.

## Conclusion

If we had to take a stand on the existence of an RNG psychokinesis effect on the basis of the evidence in Bösch et al., we would probably vote no. The most convincing evidence for the absence of a psychokinesis effect, we believe, comes from the studies with the larger sample sizes (number of bits). The rationale for restricting a meta-analysis to high-powered studies was put forth by Kraemer, Gardner, Brooks, and Yesavage (1998) and is based on the increased likelihood of publication bias for smaller studies. Thus, larger sample size studies are less likely to represent a biased sample. In Bösch et al.'s meta-analysis there were 94 experiments in the top quartile for number of bits. The fixed-effects mean for this set of studies was slightly less than the null value of .5 (.499997) and statistically significant but in the opposite direction of the intended effect. The random-effects mean, perhaps providing a more appropriate test that takes into account the uncertainty reflected in the heterogeneity of the observed effects, was slightly greater than .5 and not statistically significant. Although by no means definitive, given that these larger sample studies could be systematically different from other studies in additional ways than just sample size, nonetheless the largest studies fail to provide support for the psychokinesis hypothesis. Barring good reason to the contrary, we would place our bets on the null hypothesis.

To return to Hacking's quote that started our commentary, we are all for blowing trumpets at tulips once in a while. Indeed, RNG psychokinesis researchers are to be congratulated on the courage

they display in investigating something that may appear so "completely crazy" to much of the scientific and lay community. However, after repeated replications of trumpet blowing with no clearly "stupendous" outcome, the reasonable course of action is to rethink the enterprise or to move on to other research that may be more likely to shed light on the possibility of psychokinesis—or both. Bösch et al.'s meta-analysis cannot reject the possibility of a genuine psychokinesis effect. It seems unlikely, however, that additional studies of the type synthesized by Bosch et al. will ever definitively resolve the issue, so long as any effect, no matter how small or in which direction, is interpreted as support for a psi phenomenon. What is needed are stronger theoretical predictions that can plausibly be refuted by empirical evidence or new research paradigms that can produce bigger effects.

## References

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33,* 526–536.

Bickman, L., Lambert, E. W., Andrade, A. R., & Penaloza, R. V. (2000). The Fort Bragg continuum of care for children and adolescents: Mental health outcomes over 5 years. *Journal of Consulting and Clinical Psychology, 68,* 710–716.

Boring, E. G. (1919). Mathematical vs. scientific importance. *Psychological Bulletin, 16,* 335–338.

Bösch, H., Steinkamp, F., & Boller, E. (2006). Examining psychokinesis: The interaction of human intention with random number generators—A meta-analysis. *Psychological Bulletin, 132,* 497–523.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378–399.

Cook, T. D., Gruder, C. L., Henningan, K. M., & Faly, B. R. (1979). History of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. *Psychological Bulletin, 86,* 662–679.

Goldstein, H. (2003). *Multilevel statistical models.* London: Arnold.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science.* Cambridge, England: Cambridge University Press.

Jones, B., Jarvis, P. Lewis, J. A., & Ebbut, A. F. (1996). Trials to assess equivalence: The importance of rigorous methods. *BMJ (British Medical Journal), 313,* 36–39.

Kraemer, H. C., Gardner, C., Brooks, J. O., III, & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23–31.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Popper, K. R. (1965). *Conjectures and refutations: The growth of scientific knowledge.* New York: Harper & Row.

StataCorp. (2001). Stata Statistical Software: Release 7. College Station, TX: StataCorp.

Received October 18, 2005
Accepted October 19, 2005 ■

Psychological Bulletin
2006, Vol. 132, No. 4, 529–532


0033-2909/06/$12.00 DOI: 10.1037/0033-2909.132.4.529

# Reexamining Psychokinesis: Comment on Bösch, Steinkamp, and Boller (2006)

Dean Radin
Institute of Noetic Sciences

Roger Nelson and York Dobyns
Princeton University

Joop Houtkooper
Justus Liebig University of Giessen

H. Bösch, F. Steinkamp, and E. Boller's (2006) review of the evidence for psychokinesis confirms many of the authors' earlier findings. The authors agree with Bösch et al. that existing studies provide statistical evidence for psychokinesis, that the evidence is generally of high methodological quality, and that effect sizes are distributed heterogeneously. Bösch et al. postulated the heterogeneity is attributable to selective reporting and thus that psychokinesis is "not proven." However, Bösch et al. assumed that effect size is entirely independent of sample size. For these experiments, this assumption is incorrect; it also guarantees heterogeneity. The authors maintain that selective reporting is an implausible explanation for the observed data and hence that these studies provide evidence for a genuine psychokinetic effect.

*Keywords:* meta-analysis, parapsychology, psychokinesis, random number generator

Bösch, Steinkamp, and Boller's (2006) review of the experimental evidence for psychokinesis (PK), specifically, direct mind–matter interactions on the outputs of electronic truly random number generators (RNGs), confirms many of our earlier findings. With Bösch et al. we agree that the existing data indicate the existence of a PK effect, that the studies are generally of high methodological quality, and that effect sizes are distributed heterogeneously. We disagree about the source of the heterogeneity. Bösch et al. proposed that the large variation among effect sizes is due to selective reporting practices, and they present an ad hoc Monte Carlo simulation in support of their contention. We believe there is a more parsimonious explanation for the heterogeneity, namely that Bösch et al.'s initial assumption—effect size is independent of sample size—is incorrect.

On the basis of their assumption, Bösch et al. concluded that the PK hypothesis is "not proven." This verdict refers to situations in which circumstantial evidence is too strong to disregard but also too weak to unambiguously convince. We contend that Bösch et al.'s jury is still out not because the evidence is weak but because their assumption leads to a series of escalating confusions.

Bösch et al. assumed that mental intention acts uniformly on each random bit, regardless of the number of bits generated per sample, the rate at which bits are generated, or the psychological conditions of the task. To illustrate why Bösch et al.'s assumption is fallacious, we provide the following scenarios: Consider that we conduct a study involving 1,000 experienced meditators, each of whom is selected on the basis of his or her performance on a previous, similar PK task. Each participant is asked by a cordial, enthusiastic investigator to engage in a daily intention-focusing practice for 4 weeks in preparation for the experiment, in which he or she will be asked to intentionally influence the generation of a single random bit. Participants are told that the outcome of that random decision will determine the outcome of a meaningful bonus, such as winning a scholarship. Now consider a second study in which a bored student investigator indifferently recruits an arbitrarily selected college sophomore, who is asked to mentally influence 1,000 random bits generated in a millisecond, with no feedback of the results and no consequences regardless of the outcome.

The physical context of these two studies may be identical, using the same RNG and statistics to evaluate the resulting data sets, each of which consists of a total of 1,000 randomly generated bits. But it is clear that the psychological contexts differ radically. If we presume that the only important factor in this type of experiment is the number of bits generated, then the two studies should provide about the same results. But if a significant variable is the amount of time or effort one can apply in focusing mental intention toward each random event, then the former study might result in an effect size orders of magnitude larger than the latter.

Clearly, one's view of what is meant by PK shapes the proper definition of effect size in these studies, and as such, it is important to note that the hypothesis under test is not a proposal about pure physics. Rather, PK proposes an interaction between physics and psychology in which both sides of that relationship are linked in a meaningful way. Thus, Bösch et al.'s major assumption, which may be plausible for an experiment designed to measure the gravitational constant, is inappropriate for a PK experiment.

Dean Radin, Consciousness Research Laboratory, Institute of Noetic Sciences, Petaluma, California; Roger Nelson and York Dobyns, Princeton Engineering Anomalies Research Laboratory, Princeton University; Joop Houtkooper, Center for Psychobiology and Behavioral Medicine, Justus Liebig University of Giessen, Giessen, Germany.

Correspondence concerning this article should be addressed to Dean Radin, Institute of Noetic Sciences, 101 San Antonio Road, Petaluma, CA 94952. E-mail: deanradin@noetic.org

Indeed, if PK operates without regard to psychological factors and effect size is constant regardless of the number of randomly generated bits, then experiments with high levels of statistical significance can easily be produced by simply increasing the number of bits. But the data clearly show that $z$ scores do not increase with increasing sample size. On the other hand, if effect sizes are intimately related to the psychological task, then a better measure of effect size might be associated with the statistical outcome of a single session or, alternatively, the overall $z$ score associated with an entire study. When Bösch et al. assessed Radin and Nelson's (2003) prediction of a constant $z$ score per study, they confirmed that this was indeed the case when the three largest "outlier" studies were excluded, which they argued is the preferred way to analyze these data.

## The "Three Largest Studies"

Because the three largest studies figure prominently in Bösch et al.'s discussion, they merit closer examination. Bösch et al.'s identification of these three studies is enigmatic because the reference as cited, "Dobyns, Dunne, and Nelson (2004)," evidently refers to Dobyns, Dunne, Jahn, and Nelson (2004), but beyond this minor citation error, the reference in question reports two experiments, only one of which Bösch et al. considered. Of the two experiments, one is subdivided into three phases, each generating two data sets per phase, producing a total of seven data sets that can be distinguished as separate studies.

Examination of Bösch et al.'s Table 4 reveals that the three largest studies consisted of a total of $2.98 \times 10^{11}$ bits. This is the number of trials in the "high speed" data sets of the three phases of the first experiment reported in Dobyns et al. (2004). That study was a double-blind experiment in which the results of low- and high-speed bit generation rates were compared with each other. The second experiment, however, which Bösch et al. did not report, was a replication of just the high-speed bit rate design. That experiment consisted of $1.56 \times 10^{11}$ bits and was therefore larger than any individual study considered by Bösch et al..

In Bösch et al.'s Table 3, the three largest studies are reported as each containing over $10^9$ bits. That is true but also a sizable understatement. The studies reported in Dobyns et al. (2004) contain a grand total of $4.54 \times 10^{11}$ bits. The populations reported in Bösch et al.'s Table 4, on the other hand, make it clear that the entirety of the remaining meta-analytic database contains less than $1.4 \times 10^9$ bits. In terms of bits measured for the PK hypothesis, then, the four largest studies contain more than 320 times as much data as all other experiments in the Bösch et al. meta-analysis combined. Bösch et al.'s selection of just three of those four still contains over 210 times as much data as the remaining 377 studies in their meta-analysis.

These four large studies also have an aggregate $z$ equal to $-4.03$. Thus, if one takes seriously Bösch et al.'s hypothesis that PK effects manifest as shifts in the probabilities of individual random bits and that the fundamental variable of interest is $\pi$, then the overwhelming preponderance of data in these large experiments should be taken as definitive. That is, whatever the oddities of interstudy heterogeneity and small-study effects that may appear in the remainder of the meta-analytic database, that remainder comprises less than half a percent of the total available data. In this interpretation, the experiments in question unequivocally demonstrate the existence of a PK effect that is contrary to conscious intention, of high methodological quality, and established to very high levels of statistical confidence.

Moreover, the size of these large studies refutes the plausibility of a file drawer explanation. Bösch et al. argued, for example, on the basis of a suggestion by Bierman (1987), that a typical unpublished RNG experiment may contain $1.3 \times 10^5$ bits and that a file drawer of some 1,500 such studies is plausible if one postulates scores of investigators each generating 20 failed experiments per year. However, the file drawer required to reduce a set of four experiments consisting of $4.54 \times 10^{11}$ bits with an aggregate $z$ of $-4.03$ to nonsignificance (two-tailed) must contain at least $1.47 \times 10^{12}$ bits and would therefore require somewhat over 11 million unpublished, nonsignificant experiments of the scale suggested by Bierman.

The actual import of these large studies is even worse for Bösch et al.'s assumption about the independence of effect size and sample size. Bösch et al. did not mention that the studies in Dobyns et al. (2004) were designed to test the hypothesis that PK could be modeled as a shift in per-bit probabilities and, specifically, that such a shift would not be sensitive to the rate at which bits were collected. The immense size of this database relative to the other RNG studies arises from the use of an RNG designed to generate random bits 10,000 times faster than those previously deployed at the Princeton Engineering Anomalies Research (PEAR) Laboratory (Ibison, 1998). These experiments were identical in protocol and presentation to earlier RNG studies conducted at the PEAR Lab, which had produced positive effects, and therefore the strong negative $z$ score resulting from the large studies demonstrates a clear dependence of the PK effect size on the speed with which random bits are generated. Bösch et al.'s arguments for heterogeneity and small-study effects are based on the premise that there is no such functional dependence. The high-speed RNG experiments reported in Dobyns et al. (2004) refute that premise, invalidating Bösch et al.'s assumption and thereby casting doubt on their conclusions.

## Adequacy of Monte Carlo Model

Bösch et al. asserted that their Monte Carlo model of selective reporting successfully explains ("matches the empirical $z$ score almost perfectly" [p. 514] and "is in good agreement with all three major findings" [p. 515]) the large observed heterogeneity of effect sizes. But such statements, repeated throughout the article, put an overly enthusiastic spin on the actual results of their modeling efforts. As Bösch et al. showed in their Table 9, overall, the file drawer simulation resulted in highly significant underestimates of both the actual (negative) effect size and heterogeneity.

Further, the model's input parameters completely predetermine its outcome, obviating the need for running the simulation. That is, Bösch et al. reported in their Table 9 that the model estimates 1,544 unpublished studies. Because $p$ values are uniformly distributed by construction, a cursory inspection of the acceptance probabilities of their selection model reveals that the model will accept 0.197 of all studies presented to it. Thus, the number of file drawer studies it is expected to produce is $(1 - 0.197)/0.197$ or 4.076 times the number of surviving "published" studies in the postselection population. Their simulated result of $1{,}544/380 = 4.063$ is then,

not surprising, almost exactly in alignment with this expected value.

The simulation selects studies on the basis of $p$ values, which is equivalent to selecting on $z$ scores. For binary data, the relation between $z$, $N$, and $\pi$ (where $N$ is the study size) is simply $z = 2\sqrt{N}(\pi - 0.5)$. Because the expected $z$ for studies generated by the Monte Carlo process is constant for any given set of selection parameters, it follows that an effect size $(\pi - 0.5) \sim 1/\sqrt{N}$ is expected. In other words, a small-study effect with effect sizes proportional to $1/\sqrt{N}$ is built into the very structure of Bösch et al.'s Monte Carlo model.

In fact, a selection model of this sort can produce any desired output distribution by a suitable choice of breakpoints and weight factors. Bösch et al. were pleased by their model's good fit to the observed effect sizes (although it is marginally adequate only for the Random Effects Model on the reduced data set) but unconcerned by the poor fit to the observed heterogeneity. This latter point is important because the observed heterogeneity cannot be fit except by considerably more stringent selection processes than those they consider.

For example, Bösch et al. showed in Table 9 that their Monte Carlo process produces a heterogeneity measure ($Q$) of 631.58; this corresponds to approximately 1.67 units of normalized variance per degree of freedom. In their Table 4, they show the same value for the actual data to be 1,508.56, or 3.98 times the expected interstudy unit variance. The most efficient way to produce such an increase in the variance of the postselection distribution would be to discard those studies contributing least to the variance, that is, those with the smallest values of $|z|$. To achieve the observed level of heterogeneity through selection by this maximally efficient process requires that one drop the study retention fraction from Bösch et al.'s figure of 0.197 to 0.128. This leads to a file drawer some 6.81 times larger than the observed data, or 2,588 studies. To accommodate the observed heterogeneity after factoring in psychological factors and with a bias toward reporting positive results, one would require an even larger file drawer.

## Size of the File Drawer

Bösch et al. proposed, on the basis of Bierman's (1987) thought experiment, that if 30 researchers ran 20 experiments per year for 5 years, each with about 131,000 random bits, then this could plausibly account for the missing studies. Of course, all of those hypothetical studies would have had to escape Bösch et al.'s "very comprehensive search strategy" (p. 515), which included obscure technical reports and conference presentations in many languages. But beyond the hypothetical, in preparation for this commentary we conducted a survey among the members of an online discussion group that includes many of the researchers who have conducted RNG PK studies. The survey revealed that the average number of nonreported experiments per investigator was 1, suggesting that perhaps 59 studies were potentially missed by Bösch et al.'s search strategy. (Some of those missing studies were reportedly statistically significant.) In light of this file drawer estimate based on empirical data and the failure of Bösch et al.'s model to account for both the observed effect sizes and their heterogeneity, their assertion that "publication bias appears to be the easiest and most encompassing explanation for the primary findings of the meta-analysis" (p. 517) is unjustified.

In addition, Bösch et al. demonstrate that Duval and Tweedie's (2000) trim and fill algorithm only marginally changes the results of both the Fixed Effects Model (FEM) and Random Effects Model (REM) models. This independently implies that the original results may be considered robust with respect to selective reporting (Gilbody, Song, Eastwood, & Sutton, 2000).

## Exclusion Criteria

Bösch et al. excluded two thirds of the experimental reports they found. That selection may have introduced important factors that the reader cannot evaluate. In any case, the exclusion of data with a nonbinomial distribution, such as studies based on radioactive decay, is questionable. In the dice studies, for example, a transform was used to convert any $z$ score, and therefore any $p$ value, into the $\pi$ statistic. The same approach could have been used for these excluded cases.

## Experimenters' Regress

It may be useful to draw attention to a temperamental difference relevant to assessing this evidence. Many scientists agree that demonstrating the existence of genuine PK would be of profound importance, and thus, careful consideration of this topic is warranted. But different predilections lead to different assessments of the same evidence. Those scientists who fret over Type I errors insist on proof positive before taking the evidence seriously, whereas those who worry more about Type II errors prefer to take a more affirmative stance to counteract the prejudices invariably faced by anomalies research. Type I preference appears to have led to Bösch et al.'s comment that "this unique experimental approach will gain scientific recognition only when we know *with certainty* what an unbiased funnel plot . . . looks like" (emphasis added; p. 517). This sounds reasonable until it is unpacked, and then it is found to hide an irresolvable paradox.

Collins (1992) called this problem the *experimenters' regress,* a catch-22 that arises when the correct outcome of an experiment is unknown. To settle the question under normal circumstances, in which results are predicted by a well-accepted theory, one can simply compare an experimental outcome to the prediction. If they match, then the experiment was conducted in a proper fashion, and the outcome is regarded as correct. If not, the experiment was flawed. Unfortunately, when it comes to a pretheoretical concept like PK, to judge whether an experiment was performed well, one first needs to know whether PK exists. But to know whether PK exists, one needs to conduct the correct experiment. But to conduct that experiment, one needs a well-accepted theory. And so on, ad infinitum. For Type I scientists, this loop will continue indefinitely and remain unresolved in spite of the application of the most rigorous scientific methods. The stalemate can be broken only by Type II scientists who are willing to entertain the possibility that Nature consists of many curious phenomena, some of which are not yet described by adequate theories.

## Historical Context

Bösch et al.'s opening theme, focusing on dubious tales from séance rooms and claims of spoon bending, considers only a small portion of the relevant historical record. Many scholarly disci-

plines have pondered the role of mind in the physical world, and this topic was of serious interest much earlier than the séance rooms of the late 19th century. For example, in 1627, Francis Bacon, one of the founders of modern empiricism, published a book entitled *Sylva Sylvarum: or A Naturall Historie In Ten Centuries.* In that work, Bacon proposed that mental intention (his term was the "force of imagination") could be studied on objects that "have the lightest and easiest motions," including "the casting of dice." Bacon's recommendation thus presaged by over 300 years the use of dice in investigating PK, illustrating that interest in this topic can be traced to the very origins of the scientific method.

Physicists continue to debate the role of the observer within the framework of modern physical theory. Virtually all of the founders of quantum theory, including Werner Heisenberg, Erwin Schrödinger, and Pascual Jordan, thought deeply about the issue of mind–matter interaction (Jordan, 1960; Wilber, 1984), and this intellectual lineage continues in contemporary physics (Nadeau & Kafatos, 2001; Stapp, 2004). One can even find pronouncements on the topic published in sober magazines like *Scientific American:* "The doctrine that the world is made up of objects whose existence is independent of human consciousness turns out to be in conflict with quantum mechanics and with facts established by experiment" (d'Espagnat, 1979, p. 158). Without belaboring the point, interest in the possibility of mind–matter interaction did not arise because of what may or may not have happened in Victorian parlors, but rather, the problem has ancient origins and it continues to permeate scholarly and scientific interest.

## Conclusion

From the earliest investigations of PK, researchers struggled to understand the physically perplexing but psychologically sensible goal-oriented nature of these phenomena (Schmidt, 1987). After a decade of proof-oriented experiments suggested that something interesting was going on, most researchers later concentrated on process-oriented research in an attempt to understand the interactions between psychological and physical factors. We sympathize with reviewers who assume that mind–matter interaction implies a stationary, uniform effect on each individual random bit, because that is what many earlier researchers also assumed. Unfortunately, that simplistic view is not what nature is revealing in these experiments, so more complex models are required.

We agree with Bösch et al. that the existing experimental database provides high-quality evidence suggestive of a genuine PK effect and that effect sizes are distributed heterogeneously. Bösch et al. proposed that the heterogeneity is due to selective reporting practices, but their ad hoc simulation fails to make a plausible argument in favor of that hypothesis. In addition, a survey among authors of these experiments reveals that the actual file drawer probably amounts to less than 4% of the 1,544 studies estimated by Bösch et al.'s model. We propose that a more satisfactory explanation for the observed heterogeneity is that effect size (per bit) is not independent of sample size. In summary, we believe that the cumulative data are now sufficiently persuasive to advance beyond the timid conclusion of "not proven" and that it is more fruitful to focus on understanding the nature of PK rather than to concentrate solely on the question of existence.

## References

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.

Bösch, H., Steinkamp, F., & Boller, E. (2006). Examining psychokinesis: The interaction of human intention with random number generators—A meta-analysis. *Psychological Bulletin, 132,* 497–523.

Collins, H. M. (1992). *Changing order: Replication and induction in scientific practice* (2nd ed.). Chicago: University of Chicago Press.

d'Espagnat, B. (1979, November). The quantum theory and reality. *Scientific American, 241,* 158–181.

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

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

Gilbody, S. M., Song, F., Eastwood, A . J., & Sutton, A. (2000). The causes, consequences and detection of publication bias in psychiatry. *Acta Psychiatrica Scandinavica, 102,* 241–249.

Ibison, M. (1998). Evidence that anomalous statistical influence depends on the details of the random process. *Journal of Scientific Exploration, 12,* 407–423.

Jordan, P. (1960). Parapsychological implications of research in atomic physics. *International Journal of Parapsychology, 2,* 5.

Nadeau, R., & Kafatos, M. (2001). *The non-local universe: The new physics and matters of the mind.* Oxford, England: Oxford University Press.

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.

Schmidt, H. (1987). The strange properties of psychokinesis. *Journal of Scientific Exploration, 1,* 103–118.

Stapp, H. P. (2004). *Mind, matter and quantum mechanics* (2nd ed.). New York: Springer.

Wilber, K. (1984). *Quantum questions.* Boulder, CO: Shambhala.

Received October 5, 2005
Revision received October 18, 2005
Accepted October 19, 2005 ■

Psychological Bulletin
2006, Vol. 132, No. 4, 533–537


0033-2909/06/$12.00 DOI: 10.1037/0033-2909.132.4.533

# In the Eye of the Beholder: Reply to Wilson and Shadish (2006) and Radin, Nelson, Dobyns, and Houtkooper (2006)

Holger Bösch
University Hospital Freiburg

Fiona Steinkamp
University of Edinburgh

Emil Boller
Institute for Border Areas of Psychology and Mental Hygiene

H. Bösch, F. Steinkamp, and E. Boller's (2006) meta-analysis, which demonstrated (a) a small but highly significant overall effect, (b) a small-study effect, and (c) extreme heterogeneity, has provoked widely differing responses. After considering D. B. Wilson and W. R. Shadish's (2006) and D. Radin, R. Nelson, Y. Dobyns, and J. Houtkooper's (2006) concerns about the possible effects of psychological moderator variables, the potential for missing data, and the difficulties inherent in any meta-analytic data, the authors reaffirm their view that publication bias is the most parsimonious model to account for all 3 findings. However, until compulsory registration of trials occurs, it cannot be proven that the effect is in fact attributable to publication bias, and it remains up to the individual reader to decide how the results are best and most parsimoniously interpreted.

*Keywords:* meta-analysis, psychokinesis, random number generator, small-study effect, publication bias

Either the effect of human intention on random number generators (RNGs) is genuine or it is not. The widely differing responses from Wilson and Shadish (2006) and Radin, Nelson, Dobyns, and Houtkooper (2006) suggest that currently any conclusion about the evidence lies in the eye of the beholder. This situation is unlikely to change anytime soon. It would be desirable in the future for parapsychology experimenters to submit to trial registries prespecified protocols detailing (a) their proposed primary and secondary analyses and (b) the defining characteristics of their forthcoming RNG trials. However, the answer to the question will still remain ambiguous if the data remain poorly replicable. Indeed, we ourselves remain undecided about the precise conclusions to be drawn from the existing data.

If the answer to the question of what the underlying cause was for the significant effect in our meta-analysis (Bösch, Steinkamp, & Boller, 2006) is that it was not parapsychological, the data may provide insight into how publication bias can result in the demonstration of (a) a very small (but misleading) overall effect, (b) a remarkable variability of effect size, and (c) a small-study effect. The statement by Radin et al. (2006) that the "existing studies provide statistical evidence for psychokinesis" (p. 529) obscures the fact that this very small overall effect might be an artifact.

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

Correspondence concerning this article should be addressed to Holger Bösch, Department of Evaluation Research in Complementary Medicine, University Hospital Freiburg, Hugstetterstrasse 55, D79106, Freiburg, Germany. E-mail: holger.boesch@uniklinik-freiburg.de

If the answer is that the effect was parapsychological, it could form the foundation of a new or revised understanding of the abilities of the human mind; it may provoke us to revisit Cartesian dualism or revise our understanding of the nature of matter. It is unlikely that our understanding of the world would remain unchanged.

We agree with Wilson and Shadish (2006) that there is an unresolved problem regarding the point at which an effect is so small that it no longer warrants consideration as a genuine effect, but it was not our aim, nor is it our position, to resolve this issue, interesting and important as it is. Further, in several places, Wilson and Shadish suggest that the limits of our meta-analysis are methodological in nature. However, data coded from primary sources are generally limited, and we view their concerns as a methodological problem of meta-analyses in general, not just with our effort.

Although it is unfortunate that we overlooked the fourth very large study published in Dobyns, Dunne, Jahn, and Nelson (2004), even several such studies would not compromise the findings of our meta-analysis. It is not particularly surprising that three or more very large studies with small effects in the direction opposite from intention change the direction of the overall findings when one uses a fixed-effects model weighted by size of study. More important, the largest studies still confirmed our finding that larger studies produce smaller effect sizes.

Wilson and Shadish (2006) describe our conclusion that sample size is the most important moderator as "suspect" (p. 525). Nevertheless, three findings support our conclusion: (a) Smaller studies revealed larger effect sizes; (b) the cumulative meta-analysis demonstrated that gradually introducing studies according to sample size, starting with the smallest study, brought the effect size closer and closer to the null value; and (c) with one exception, the subsample including the largest studies had the smallest effect size

(see Table 6 in our meta-analysis [Bösch et al., 2006], described in footnote 13).

Below, we focus on what we regard to be the five most important aspects that require clarification and that would most benefit from further consideration.

## Psychological Variables

The meta-analysis was expressly designed to analyze potential moderator variables. To this end, we included a range of items in our coding book, such as whether psychological tests were taken in conjunction with an RNG experiment, including variables concerning the experimental setting and recording technical details about the type of RNG. We did not know in advance whether variables would be reported systematically or whether reports would provide sufficient data for an overall analysis.

As it turned out, the physical or mental state of the participants (e.g., requesting the participants to physically tense up themselves or asking them to enter a meditative state) was experimentally manipulated in only a third of the studies, and only 30% of studies used psychological measures. Moreover, the physical and mental states that were manipulated, the methods used to induce them, and the psychological measures used in this minority of experiments differed greatly across the studies. Consequently, it was not advisable to perform an analysis on these variables.

The one psychological aspect that all of the studies had in common was the presence of human intention. In some experiments, participants decided the direction of intention, and in others, this was done by the experimenter or computer. But, however or whoever decides what the direction of intention should be, the assumption remains that the participants constantly keep the same intention throughout and conform to the experimental protocol. Because one cannot control the participants' intention, it is possible that the participants could change their intention in the middle of a session without the experimenter knowing. For example, if a participant sees feedback depicting that the psychokinesis (PK) effect is apparently going in the direction opposite to his or her intention, the participant may switch his or her intention to encourage the results to go yet further in that direction rather than in the one originally agreed.

Another rarely discussed psychological difficulty with RNG experiments is that of obtaining feedback. The thought experiment suggested by Radin et al. (2006, p. 531) illustrates the problem particularly clearly. Even if 1 trial is influenced in every 20, then in 1,000 trials of "heads or tails," with a very optimistic hit rate of 55% ($\pi = .55$), approximately 550 hits are to be expected. However, of these, 500 are obtained purely by chance. Consequently, many participants will be under the false assumption from their feedback that they have been successful when their apparent success is just a chance result. Feedback in RNG experiments may have the appearance of providing helpful information to participants, but, in fact, it is more of a distractor because chance fluctuations will be more visible than the occasional small effect.

Moreover, in practical terms, in RNG experiments, the signal-to-noise ratio, or the reliability of the effect, is so small that one cannot reasonably expect to find systematic correlations with, for example, psychological state or trait variables (Boller & Bösch, 2000). Previous reviews have also been unable to identify any clear moderator variables (Gissurarson, 1992, 1997; Gissurarson & Morris, 1991; Schmeidler, 1977). In these circumstances, proof of a general effect rather than of psychological correlates may be the best strategy.

Consequently, Radin et al. (2006) are mistaken in their claim that we assumed that mental intention acts uniformly on each random bit, regardless of the number of bits generated per sample, the rate at which bits are generated, or the psychological conditions of the task. We did consider all of these potential moderators but concluded that they were issues that obscured rather than aided the meta-analysis.

## Adequacy of the Statistics Used

Wilson and Shadish (2006) correctly note that the observed effects are very small and that the overall results from the fixed- and random-effects models are individually statistically significant but run in the opposite direction to each other.

The results from the meta-analysis are inconsistent, and this is underlined by the results from the sensitivity analyses. Moreover, the heterogeneity of the database was explicable only in part through moderator variables. Our view is that the overall mean effect and the results of the subsample analyses are difficult to interpret. To this extent, we do not share Radin et al.'s (2006) opinion that our meta-analysis necessarily statistically confirms the existence of a PK effect. However, the inconclusive nature of our findings is not due to methodological problems as Wilson and Shadish (2006) suggest but rather to the data themselves.

Wilson and Shadish (2006) postulate that there are dependencies within and across the studies and that, therefore, ordinary meta-analytical statistics will not suffice. However, if the dependencies surmised by Wilson and Shadish really did exist, they would point to the existence of PK (Boller & Bösch, 2000). Yet PK experiments start out from the null hypothesis that the data are independent of each other. Ordinary meta-analytic statistics may not suffice once there is an extraordinary effect to examine, but currently such an effect has not been established.

## An Effect Opposite to Intention

Wilson and Shadish (2006) find that they are "left without any good reason to reject an effect opposite to intention" (p. 525). However, they thereby ignore the results of the random-effects model, and, even more important, they ignore the small-study effect. In the quartile with the largest studies (Quartile 4), 13 studies produced significant results in the direction intended, and 9 studies produced significant results in the direction opposite to intention. However, this does not imply the existence of an effect opposite to intention. Indeed, overall, 83 studies produced significant results in the direction intended, and only 23 studies produced significant results in the direction opposite to intention. Of course, larger studies merit more weight than smaller studies. Moreover, the small-study effect clearly indicates that the smaller the study the larger the effect in the direction intended. Our cumulative meta-analysis demonstrated (Bösch et al., 2006, p. 506) that the size of the overall effect became progressively smaller as each larger study entered into the analysis. The direction of the effect changed to one opposite to intention at just the point where the 1st of the 3 largest studies entered the cumulative analysis. More important from our point of view, the effect at this

point still continued to approach even more closely the theoretical mean value (Bösch et al., 2006, p. 506). Therefore, if we assume that a genuine effect is present in our meta-analysis, there is no reason to believe that it is one that is opposite to intention.

Of course, the question remains as to why precisely the larger studies should demonstrate an effect opposite to intention. However, questions such as this cannot be answered by our meta-analysis. Moreover, an effect size from a fixed-effects model in which the variance between studies is not taken into account must be interpreted with caution when the effect size distribution is as heterogeneous as it is in our sample.

## The Constant *z*-Score Hypothesis

Radin and Nelson (2003) suggested in their meta-analysis that the *z* score is constant across RNG studies. To demonstrate this constancy, they calculated the average *z* score of RNG studies published up until their initial meta-analysis in 1987 ($\bar{z} = 0.73$, $SE = 0.09$) and compared it with the average *z* score of RNG studies published after 1987 ($\bar{z} = 0.61$, $SE = 0.14$). Because the difference was not statistically significant, $t(179) = 0.71$, $p = .48$, Radin and Nelson (2003) concluded that the meta-analytic evidence for mind–matter interaction effects persists.

The statistical effect that they claim persists is no different from the effect we found in our refined sample of RNG studies. It is an effect operating on bit level, an effect that is assumed to be independent of sample size, as was assumed by Radin and Nelson (1989) in their first meta-analysis of RNG data and in Radin's (1997) popular book, which heavily relied on meta-analysis, including a meta-analysis of PK data, to demonstrate that the effect is genuine.

As we discussed in the limits section of our meta-analysis, it is possible that the use of a standard effect size measure might not be adequate in RNG research. Because the correlation between the studies' *z* score and $\sqrt{N}$ was not significant when the three largest studies were removed, $r(377) = -.02$, $p = .66$, we acknowledged that "an argument for the [constant *z*-score] model might be made" (Bösch et al., 2006, p. 516).

To analyze the constant *z*-score hypothesis, we split our sample of RNG studies into quartiles of sample size and calculated the average *z* score (see Table 1). The trend observed with effect size also appeared for the *z* scores: the larger the sample size, the smaller the average *z* score. An analysis of variance showed that the effect of sample size was significant, $F(3, 376) = 2.64$, $p = .049$. Therefore, the constant *z*-score hypothesis appears not to hold.

This analysis demonstrates that splitting the sample into quartiles can bring out information that would otherwise not come to light. That data reduction can be a successful procedure to pronounce the importance of certain variables has already been shown by our second metaregression model that clearly demonstrated the importance of sample size. Of course, there were other significant moderator variables in addition to sample size, but in terms of level of significance, sample size was by far the most notable.

The finding that the average *z* score (of sample-size quartiles) was related to sample size indicates not only that the constant *z*-score hypothesis does not fit the data but also that our Monte Carlo simulation oversimplified the actual conditions (as we noted in the meta-analysis). As we previously argued, our model is simply a "proof in principle" that publication bias could explain the results; it cannot completely explain the heterogeneity or the distribution of *z* scores. The question remains as to whether any complete explanation can be found for the meta-analytic results.

## Publication Bias

Publication bias is a crucial issue for most sciences and refers to the problem that the probability of a study being published is dependent on the study's *p* value. This bias is affected by several independent factors, as discussed briefly in our meta-analysis (Bösch et al., 2006). Even at a very early stage of the "publication process," at least two steps can be differentiated. First, the data must be analyzed, and second, a report must be written. As Greenwood (1975) remarked, "choices of sample size, dependent measures, statistical tests, and the like" (p. 7) affect the results of any given study and consequently may also affect the urgency or likelihood with which a report is written as well as the slant given when writing the report. In his "model of research-publication system," Greenwood also addressed the problem of intermediate analyses partway through a study that might result in terminating or altering the study. Moreover, subgroup analyses can be conducted post hoc without appropriate indication of their essentially selective nature. A statistically significant subgroup analysis is certainly more likely to end up in a published report than is a nonsignificant subgroup analysis. All of these factors distort the meta-analytic data and misleadingly increase the likelihood of obtaining a significant overall effect as well as adding to the heterogeneity of the database.

Problems such as these could be overcome if there were a trial registry. In medicine, for example, from July 1, 2005, the International Committee of Medical Journal Editors, a group of editors

Table 1
*Mean z Score of Sample-Size Quartiles*

| Sample size | *n* | *M* | *SE* | Minimum | Maximum |
|---|---|---|---|---|---|
| Overall | 380 (377) | 0.67 (0.70) | 0.095 (0.095) | −5.00 | 10.68 |
| Smallest (Q1) | 95 | 1.05 | 0.194 | −3.42 | 10.28 |
| Small (Q2) | 95 | 0.75 | 0.196 | −2.68 | 10.68 |
| Large (Q3) | 96 | 0.56 | 0.191 | −4.29 | 7.74 |
| Largest (Q4) | 94 (91) | 0.32 (0.41) | 0.174 (0.171) | −5.00 | 5.66 |

*Note.* The numbers in parentheses indicate the results when the three largest studies were removed from the sample. Q = quartile.

of high-impact medical journals, has required "as a condition of consideration for publication, registration in a public trials registry" (DeAngelis et al., 2004, p. 1363). This procedure enables researchers to know which studies have been published and which have not. Because these registries require, at a minimum, information about the primary and secondary outcomes and the target sample size (DeAngelis et al., 2005), later redefined analyses cannot be carried out without it being clear that they are at best tertiary analyses. As a side effect, such registries will likely reduce the number of post hoc subgroup analyses and multiple analyses, which are probably the most commonly observed current bad practices in statistical analysis of trial data (Beck-Bornhold & Dubben, 1994).

Meta-analytic results can be distorted not only by these publication biases but also by the selection of publications to insert in the meta-analytic database.[1] Even the most well-intentioned, comprehensive search strategy aimed at including published as well as unpublished manuscripts can be fallible. We do not deny that we inadvertently missed some relevant reports, despite having done our best to contact all researchers in the field and to search through all relevant journals and other publications. Nevertheless, we find it very unlikely that our literature search potentially missed 59 studies as suggested by Radin et al. (2006), although Radin et al.'s ad hoc survey was addressing "nonreported experiments" (p. 531). If no report of the study has been written, no search strategy will ever return it, and there is difficulty in knowing how to go about coding studies that are known about purely by word of mouth. Moreover, even if these "nonreported experiments" were written up but not published, it is not clear to us how Radin et al. can be sure that we had not deemed these reports as failing to meet the inclusion and exclusion criteria for our meta-analysis and deliberately excluded them. We have a list of 225 reports that did not meet our criteria, and it is available to anyone who asks.

The crucial question that arises from our meta-analysis is whether the 1,544 "unpublished" studies in our Monte Carlo simulation could be the result of publication bias. In our opinion, the answer to this question is yes because publication bias relates the outcome of a study, which, as we illustrated above, may have been influenced by a number of independent factors, affecting the likelihood of its publication at all of the various stages, that is, (re)planning, (re)analyzing, (re)writing, or (re)submitting, at which bias can come into play. In our simulation, we did not address these steps in any detail, as it is a whole research area of its own. The way in which experimental reports are split into studies also contributes to publication bias because researchers are more likely to pursue (and finally report) splits of data that are significant and less likely to report nonsignificant analyses. These procedures also artificially inflate heterogeneity. However, although we believe that publication bias is the greatest threat to the data, we do not believe that a large number of reports are hidden in the file drawer. Publication bias is a subtle effect operating on different levels, some of which, such as editorial decisions to publish an article, are not even in the hands of the experimenter.

The Monte Carlo model is too simplistic to depict the real-life publication process. We were surprised to find that our simple model would reproduce the three main effects to such a good degree. The small-study effect demonstrated by the Monte Carlo simulation clearly is not "built into" (Radin et al., 2006, p. 531) the simulation but is a result of publication bias. As we pointed out, "the fit of the simulation can be improved by varying the parameters used and/or by including additional parameters" (Bösch et al., 2006, p. 513). However, such improvements to the fit would not increase the plausibility of the approach because the question is how to prove which parameters best belong to the model. As a result, the hypothesis arises that researchers are less likely to publish nonsignificant, small, more easily conducted studies, preferring to initiate the next study instead, and yet are more likely to publish small studies if they do happen to provide significant results. If this does form part of the publication (or nonpublication) process that occurs, it would also go some way to explaining the heterogeneity of our database. However, this is speculation and not proof. Until other explanations for the data are forthcoming, credit must be given to this simple model because it is potentially able to explain the meta-analytic data with relatively few assumptions.

## Conclusion

In our view, the most important findings from our meta-analysis are the finding of a small but significant overall effect in the experimental data, the existence of a small-study effect, and the extreme heterogeneity of the database. We believe that, ultimately, all of these findings could be explained through publication bias and there is currently no other model available to clarify the data any better. Nevertheless, at this point in time, it is up to individual readers to decide whether they agree with our speculations. The issue will be more easily resolved once trial registers are established and their use required by all major journals. Until that day, the answer will remain in the eye of the beholder, as the comments by Wilson and Shadish (2006) and Radin et al. (2006) and our own reply demonstrate so very well.

[1] It should be noted that this problem ultimately results in the necessity of registering not only primary research but also meta-analyses because meta-analysts too could analyze different samples until a few significant ones are found, they could apply different inclusion criteria until the result is as desired, or a meta-analysis could be discontinued after the initial data have been analyzed if the results look to be unfavorable to the hypothesis.

## References

Beck-Bornhold, H.-P., & Dubben, H.-H. (1994). Potential pitfalls in the use of p-values and in interpretation of significance levels. *Radiotherapy and Oncology, 33,* 171–176.

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.

Bösch, H., Steinkamp, F., & Boller, E. (2006). Examining psychokinesis: The interaction of human intention with random number generators—A meta-analysis. *Psychological Bulletin, 132,* 497–523.

DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., et al. (2004). Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *Journal of the American Medical Association, 292,* 1363–1364.

DeAngelis, C. D., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., et al. (2005). Is this clinical trial fully registered? A statement from the International Committee of Medical Journal Editors. *Lancet, 365,* 1827–1829.

Dobyns, Y. H., Dunne, B. J., Jahn, R. G., & Nelson, R. D. (2004). The

MegaREG experiment: Replication and interpretation. *Journal of Scientific Exploration, 18,* 369–397.

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* (Vol. 8, pp. 88–125). Jefferson, NC: McFarland.

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

Greenwood, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

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

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

Radin, D. 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, Scotland: Churchill Livingstone.

Radin, D., Nelson, R., Dobyns, Y., & Houtkooper, J. (2006). Reexamining psychokinesis: Comment on the Bösch, Steinkamp, and Boller (2006) meta-analysis. *Psychological Bulletin, 132,* 529–532.

Schmeidler, G. R. (1977). Research findings in psychokinesis. In S. Krippner (Ed.), *Advances in parapsychological research: Vol. 1. Psychokinesis* (pp. 79–132). New York: Plenum Press.

Wilson, D. B., & Shadish, W. R. (2006). On blowing trumpets to the tulips: To prove or not to prove the null hypothesis: Comment on Bösch, Steinkamp, and Boller (2006). *Psychological Bulletin, 132,* 524–528.

Received February 8, 2006
Revision received February 22, 2006
Accepted February 22, 2006 ■