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 2.98×10^{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 1.56×10^{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 4.54×10^{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 1.4×10^9 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 π, 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 1.3×10^5 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 4.54×10^{11} bits with an aggregate z of -4.03 to nonsignificance (two-tailed) must contain at least 1.47×10^{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 π (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 π 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 ■