

**In the eye of the beholder: Reply to Wilson and Shadish (2006) and Radin,
Nelson, Dobyms, and Houtkooper (2006)**

Holger Bösch

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

Fiona Steinkamp

Department of Psychology, University of Edinburgh, Edinburgh UK

Emil Boller

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

This article may not exactly replicate the final version
published in the [Psychological Bulletin](#). It is not the copy of record.

© 2006 American Psychological Association

Final Article: Bösch, H.; Steinkamp, F.; & Boller, E. (2006). In the Eye
of the Beholder: Reply to Wilson and Shadish (2006) and Radin, Nelson,
Dobyms, and Houtkooper (2006). *Psychological Bulletin*, 132, 533-537

Abstract

Our meta-analysis, which demonstrated (i) a small, but highly significant overall effect, (ii) a small study effect, and (iii) extreme heterogeneity, has provoked widely differing responses. After considering our respondents' concerns about the possible effects of psychological moderator variables, the potential for missing data, and the difficulties inherent in any meta-analytic data, we reaffirm our view that publication bias is the most parsimonious model to account for all three 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 our results are best and most parsimoniously interpreted.

Keywords: meta-analysis, parapsychology, psychokinesis, random number generator, small-study effect, publication bias, Monte Carlo simulation,

In the eye of the beholder: Reply to Wilson and Shadish (2006) and Radin, Nelson, Dobyms, and Houtkooper (2006)

The effect of human intention on random number generators (RNGs) is either genuine or it is not. The widely differing responses from Wilson & Shadish (2006) (WS) and Radin, Nelson, Dobyms, & Houtkooper (2006) (RNDH) suggest that currently any conclusion about the evidence lies in the eye of the beholder. This situation is unlikely to change any time soon. It would be desirable in the future for parapsychology experimenters to submit to trial registries prespecified protocols detailing (i) their proposed primary and secondary analyses and (ii) 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 is that it was not parapsychological, the data may provide insight into how publication bias can result in the demonstration of (i) a very small (but misleading) overall effect, (ii) a remarkable variability of effect size, and (iii) a small-study effect. The statement by RNDH (2006) that the “existing studies provide evidence for psychokinesis” (p.2) obscures the fact that this very small overall effect might be an artifact.

If the answer is that the effect is 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 WS (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, and nor is it our position, to resolve this issue, interesting and important as it is. Further, in several places, WS (2006) 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 Dobyms, Dunne, Jahn, & 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 importantly, the largest studies still confirmed our finding that larger studies produce smaller effect sizes.

Wilson & Shadish (2006) described our conclusion that sample size is the most important moderator as “suspect” (p.7). Nevertheless, three findings support our conclusion: (i) smaller studies revealed larger effect sizes, (ii) 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 (iii) with one exception, the subsample including the largest studies had the smallest effect size (see Table 6 in our meta-analysis, described in footnote 13).

Below, we will 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 a 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 participant (eg, requesting the participant to tense themselves up physically or asking them to enter a meditative state) was experimentally manipulated in only a third of the studies and only 30% of studies employed psychological measures. Moreover, the physical and mental states that were manipulated, the methods used to induce them, and the psychological measures employed 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 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 participant constantly keeps the same intention throughout and conforms to the experimental protocol. Since one cannot control the participant's intention, it is possible that the participant could change their intention in the middle of a session without the experimenter knowing. For example, if a participant sees feedback depicting that the PK effect is apparently going in the direction opposite to their intention, they may switch their 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 RNDH (2006, p 3f) illustrates the problem particularly clearly. Even if 1 trial is influenced in every 20, then in 1000 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, since 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, RNDH 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

WS (2006) correctly note that the observed effects are very small and that the overall results from the fixed and random effect 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 RNDH's 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 WS (2006) suggest, but due to the data themselves.

WS (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 WS 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

WS (2006, p. 8) find that they are “left without any good reason to reject an effect opposite to intention”. However, they thereby ignore the results of the random effect model and, even more importantly, they ignore the small study effect. 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. 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, Steinkamp, & Boller, 2006, p. 30) 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 first of the three largest studies entered the cumulative analysis. More importantly 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. 30). 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 the current 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 & Nelson (2003) suggested in their meta-analysis that the z -score is constant across RNG studies. To demonstrate this, they calculated the average z -score of RNG studies published up until their initial meta-analysis in 1987 ($\bar{z} = .73$, $SE = .09$) and compared it with the average z -score of RNG studies published after 1987 ($\bar{z} = .61$, $SE = .14$). Because the difference was not statistically significant ($t(179) = .71$, $p = .48$), Radin & 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 & Nelson in their first meta-analysis (Radin & Nelson, 1989) 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} is not significant when the three largest studies are removed ($r(377) = -.02$, $p = .66$) we acknowledge that "an argument for the [constant z -score] model might be made" (Bösch et al., 2006, p. 49).

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 appears 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 is significant, $F(3,376) = 2.64, p = .049$. Therefore, the constant z -score hypothesis appears not to hold.

Table 1

Average z-score of Sample Size Quartiles

		mean	<i>SE</i>	Min	max
	<i>n</i>	<i>z</i> -score	<i>z</i> -score	<i>z</i> -score	<i>z</i> -score
Overall	380 (377)	.67 (.70)	.095 (.095)	-5.0	10.68
Sample size					
(Q1) Smallest	95	1.05	.194	-3.42	10.28
(Q2) Small	95	.75	.196	-2.68	10.68
(Q3) Large	96	.56	.191	-4.29	7.74
(Q4) Largest	94 (91)	.32 (.41)	.174 (.171)	-5.00	5.66

Note. The numbers in brackets indicate the results when the three largest studies were removed from the sample.

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 meta-regression model which 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 upon 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) remarks, “choices of sample size, dependent measures, statistical tests, and the like” (Greenwood, 1975, p. 7) affect the results of any given study and consequently may also affect both the urgency or likelihood with which a report is written as well as the slant given when writing up the report. In his “model of research-publication system”, Greenwood also addresses the problem of intermediate analyses part-way 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 a non-significant subgroup analysis. All 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 (ICMJE), a group of editors 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 outcome and the target sample size (De Angelis 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 database¹¹. 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 the relevant journals and other publications. Nevertheless, we find it very unlikely that our literature search potentially missed 59 studies as suggested by RNDH, although RNDH's ad hoc survey was addressing "non-reported experiments" (p. 9). If no report of the study has been written, no search strategy will ever return them 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 "non-reported experiments" were written up but not published, it is not clear to us how RNDH 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 1544 "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

¹ It should be noted that this problem ultimately results in the necessity of registering not only primary research but also meta-analyses, for meta-analysts too could analyze different samples until a few significant ones are found, or 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.

outcome of a study, which, as we illustrated above, may have been influenced by a number of independent factors, influencing the likelihood of its publication at all the various stages, i.e. (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 is hidden in the file drawer. Publication bias is a subtle effect operating on different levels, some of which, such as editorial decisions to publish a paper, 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” (RNDH, p. 8) 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. 40). However, such improvements to the fit would not increase the plausibility of the approach, since 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 non-publication) 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 since 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 will be up to the individual reader to decide whether or not 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 responses by SH, RNDH, and our own reply demonstrate so very well.

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.
- 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
- Boller, E.; & Bösch, H. (2000). Reliability and correlations of PK performance in a multivariate experiment. In *The Parapsychological Association 43st Annual Convention: Proceedings of Presented Papers* (pp. 380-382). Durham, NC: Parapsychological Association.
- De Angelis, 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.

- 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.
- 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* 8 (pp. 88-125). Jefferson, NC: Mc Farland Company.
- 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., Nelson, R., Dobyns, Y., & Houtkooper, J. (2006). Reexamining psychokinesis: Comment on the Bösch, Steinkamp and Boller (2006) Meta-Analysis. *Psychological Bulletin*, 132, 529-532.
- 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, United Kingdom: Churchill Livingstone.

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

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.

Author note

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

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