Running Head: Trumpets and Tulips

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

and

William R. Shadish

University of California, Merced

Abstract

The Bösch, Steinkamp and Boller (2006) meta-analysis reaches mixed and cautious conclusions about the possibility of psychokinesis. We argue that, for both methodological and philosophical reasons, it is nearly impossible to draw any conclusions from this body of research. We do not agree that any significant effect at all, no matter how small, is "fundamentally important" (Bösch et al., 2006, p. 2); and we 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

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

> "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" (Bösch et al., 2006, p. 2). The authors conclude that this "effect in general, even if incredibly small, is of great fundamental importance" (Bösch et al., 2006x, p. 51), but also "not proven" (p. 52).

The title of our article is intended to point to a fundamental dilemma that lies at the heart of the Bösch et al. (2006) review. If we carried out Darwin's experiment, and we blew trumpets to the tulips, how would we know if 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 if 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 randomly generated numbers (RNG) 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. (2006) 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 employed 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 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 was slightly less than the null value (0.499997, i.e., slightly fewer 1s than 0s) and the random effects mean was slightly greater than the null value (0.500035). Both of these mean effects were 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 to be to find 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)". 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 in order 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. (2006) 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, 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 the present one in order to take potential dependencies into account.

To their credit, Bösch et al. (2006) expend considerable effort doing sensitivity analyses, primarily by examining categorical and continuous moderators of study effect. They conclude that study sample size is the most important moderator. They have made a plausible case for this conclusion in many respects, but one that 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, comparing analyses in which this one publication is 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 only became significant 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 a RNG control. Especially given that year of publication, auditory feedback, and RNG control were significant in both the initial and the quartile regression, 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, adequately to assess potential

moderators" (Bösch et al., 2006, p. 38). Our point exactly, except we add that it also makes it difficult to believe conclusions about the overall effect size as well.

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. (2006) 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 want to conclude that "an effect opposite to intention cannot be claimed to be a general finding of this meta-analysis" (Bösch et al., 2006, p. 41). Thus they go to some effort to argue that "the three studies are considered to be outliers and the overall effect found in the meta-analysis to be an effect in the direction intended by the participants in the studies" (Bösch et al., 2006, p. 41). 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" (Bösch et al., 2006, p. 41). 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. (2006) 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 pseudo-scientific 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 also with the experimenters' extra-scientific 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. (2006) 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. (2006) review also illustrates how nearly any result can be either dismissed or accepted using 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. (2006) 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 null hypothesis significance testing (NHST), especially that a failure to reject a null hypothesis does not establish that there is no effect (Boring, 1919, Berkson, 1938, Carver, 1978). Even if this meta-analysis had observed a mean effect size exactly equal to .5, the confidence interval would still be a non-zero range around .5, establishing the

possibility of some non-null value that 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 π 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 observed by Bösch et al. (2006), and the 95% confidence interval from the simulation had a wider range than observed by Bösch et al. (2006), despite being simulated from a distribution where $\pi = .5$. 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 problems of knowing what hypothesis to believe. A subset 137 of the psychokinesis studies included a control run of the RNG (i.e., running 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 where 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 a RNG process represents the construct of psychokinesis. RNG experiments 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. (2006) were aware of this when they concluded that the RNG experiments "may not necessarily bear a direct relation to purported large-scale phenomena" (p. 51) such as those reported in séance-rooms. From Popper's 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 or not 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. (2006) describe 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 (2006) 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 television. 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 the 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. (2006) review are "of great fundamental importance" (Bösch et al., 2006, p. 41), for the validity of the claim is not obvious to many of us outside psychokinesis. For example, one could argue that this effect is importance 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 two 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 accomplished some good end. If this is the claim, Bösch et al. (2006) 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 of us not in the field, the claim as reflected in the Bösch et al. (2006) review is not compelling.

Conclusion

If we had to take a stand on the existence of a RNG psychokinesis effect based on the evidence in Bösch et al. (2006), 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 find 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" (Hacking, 1983, p. 154) 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 (2006) 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. (2001). The Fort Bragg continuum of care for children and adolescents: Mental health outcomes over 5 years. *Journal of Consulting & 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, XX,* XXX-XXX.

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, 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.

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

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

Author Note

David B. Wilson, Department of Public and International Affairs, George Mason University and 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 Unveristy, 10900 University Blvd., MS 4F4, Manassas, VA, 20110, dwilsonb@gmu.edu.