Journal of Scientific Exploration, Vol. 20, No. 3, pp. 402–419, 2006 0892-3310/06

COMMENTARZES

A Critique of the Parapsychological Random Number Generator Meta-Analyses of Radin and Nelson

Abstract — No statistical significance is quoted in the often-cited 1989 meta-analysis by Radin and Nelson of anomalous direct mental interaction with random number generators. Authors citing it have quoted z-scores ranging from 4.1 to 15.76. The combined statistical significance turns out to be acutely sensitive to the method used to combine the data, and there is at least one method which gives a non-significant result. The sensitivity is due to small studies with large hit rates, which in turn is at least partly due to publication bias, for which there is both direct and indirect evidence. Publication bias can account for only part of the long tails of the z distribution, but the remainder could well be the results of methodological problems, since, with the possible exception of the Princeton Engineering Anomalies Research data, overall methodological quality is quite poor. Methodological quality turns out to be confounded with publication bias, so the influence of poor quality on the results cannot be ruled out. Given the problems with existing data in this field, convincing evidence for a real effect can only be provided by new experiments free of reporting biases and methodological issues.

Keywords: meta-analysis — publication bias — criticism of parapsychology

Introduction

The 1989 meta-analysis by Radin and Nelson (R&N) of anomalous mental interaction with random event/number generators (R&N, 1989) has been cited a good deal, in this journal and elsewhere. Oddly, it does not quote a statistical significance of its primary result, though a z-score of about 6.8 can be read from one of the graphs. Other authors have quoted z-scores for this same dataset ranging from 15.76 (Palmer et al., 1989: 39) to 4.1 (Utts, 1991: 375). For that matter, the original authors have quoted values ranging from 15 (Nelson, 1999: 15) to "more than 12" (R&N, 2003: 41) to "on the order of 6" (R&N, 1988: 26). Statistical significance in these meta-analyses can clearly be arrived at by a variety of methods which yield a wide variety of results. The results vary so widely because of some serious problems with the data, which I will also discuss.

Background

The experiments in question involve electronic random number generators (RNG's), which generate random numbers using either mathematical algorithms

("pseudo"-random numbers) or random physical processes, such as radioactive decay or noise in reverse-biased diodes. The experiments were intended to test for direct mental interaction of the subject(s) with the RNG. Representative tasks include guessing which random number would be generated next, or biasing the average of the numbers generated. For much of the literature, each random event generates one binary digit (bit), but devices have also been used where each event generates an integer from, say, 1 to 4. If the result is the one the subject desired, the outcome is considered a success or "hit", and the overall result of a given experiment is a proportion of hits (a "hit rate"), which may be converted into a z-score. The meta-analyses of R&N combine results from a collection of experiments to determine whether the hit rate is inconsistent with chance.

The 1989 Meta-Analysis and Later Accounts of It

R&N's 1989 meta-analysis is a continuation of an earlier meta-analysis (Radin et al., 1985) covering the years 1969–1984. The same data as the final peer-reviewed 1989 paper were previously discussed in a conference proceeding (R&N, 1988), and some details can be found there which are not provided in the final 1989 paper. In later years, Radin's book *The Conscious Universe* (Radin, 1997) and R&N's follow-up meta-analysis (R&N, 2003) supply more details for and alternative analyses of (ostensibly) the data from the 1989 paper.

R&N (1989) analyzed a group of 597 experimental and 235 control studies dated from 1959 to 1987. The unit of analysis was a "study", and a given paper could be treated as many studies if it provided information separately for various subgroups of the analysis. As noted above, this paper does not quote any composite z-score. I will discuss below some methods which have been used to combine this dataset, and because Dean Radin has graciously provided me with the 1989 data, I will show results of some methods for which results have not been previously published.

Breaking up the Princeton Engineering Anomalies Research Data

It is not noted in the 1989 paper, but 284 of the 597 experimental studies (48%) and 129 of the 235 control studies (55%) come from one experimental program, that of Princeton Engineering Anomalies Research (PEAR). While the work of other groups comes almost entirely from published accounts, breaking up the PEAR data in this way could not have been done using the published reports and would have required direct access to the PEAR data. The most dramatic effect of subdividing the PEAR data is a smoother z histogram—Figure 1, where I treat the PEAR data as 1 study, is much rougher than Figure 2 in the 1989 paper, where the PEAR data are treated as 284 studies, making some of the artifacts less obvious. Breaking up the PEAR data also has some effect on the statistical significance obtained in some methods of combining the data. It

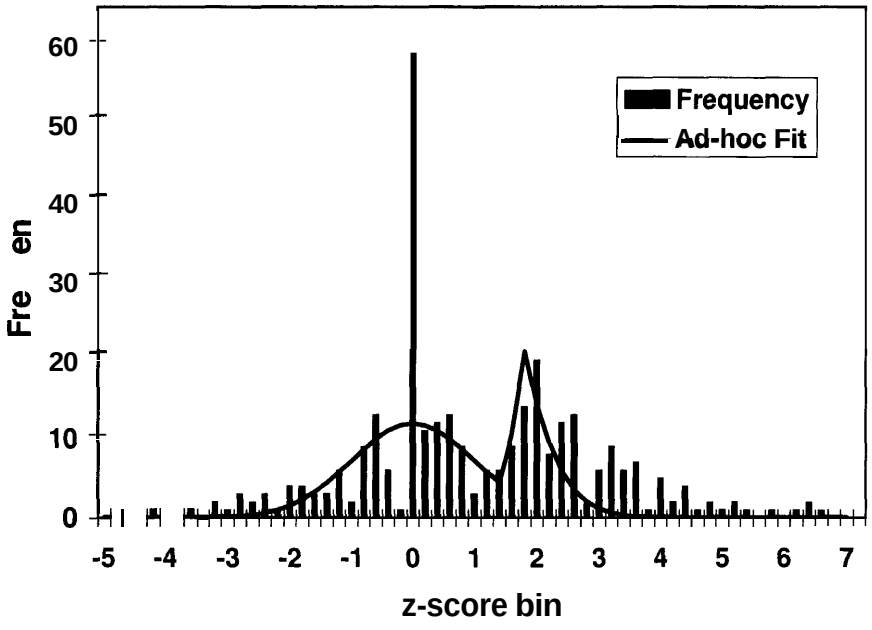

Fig. 1. Histogram of the z-score distribution for the 1989 dataset, with ad-hoc fit of null effect plus publication bias. The PEAR studies are here treated as one study, and studies reported simply as "non-significant" were assigned z = 0.

also potentially affects analysis of the effect of quality, since the PEAR studies were rated higher than any others.

Sensitivity to Method

Table 1 summarizes the results of several methods for combining the data, which are discussed further in the Appendix. The main reason that results vary dramatically from paper to paper is that the method used to combine the data changes from paper to paper. R&N (1989) show a significance equivalent to the pooling-hits method, while Radin (1997) quotes an (incorrect) unweighted average hit rate (yet quotes odds against chance apparently based on the Stouffer z instead), whereas Nelson (1999) simply quotes the Stouffer z.

Unlike examples in meta-analysis textbooks, where different methods tend to give very similar results, here both the final z-score and the composite hit rate vary quite dramatically, from astronomically significant to not significant. The largest z's come from methods which do not account for study size by giving extra weight to large studies. Accounting for between-studies variation reduces the z-scores further, even, in one case, to non-significance. Some methods do not give a composite hit rate, which is a serious shortcoming, because it gives no guidance for designing future experiments. Further, arguably, since the original

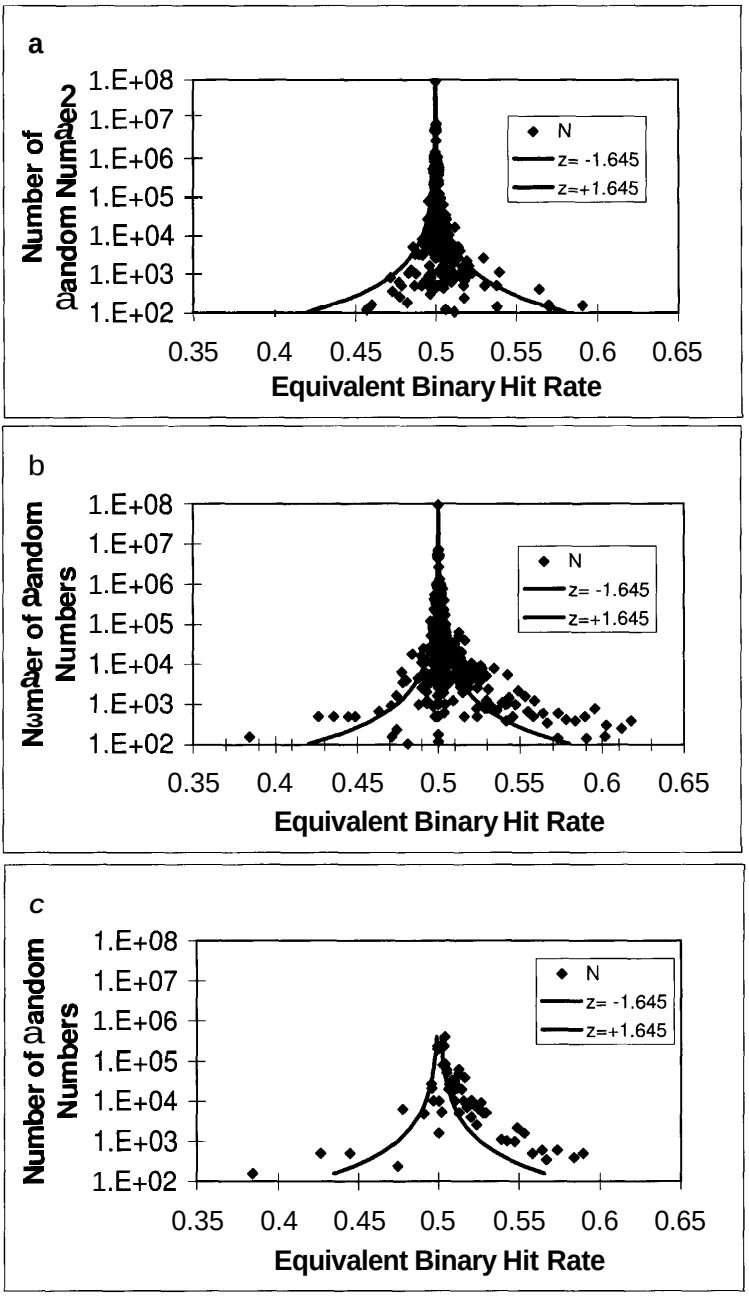

Fig. 2. (a) Funnel plot with simulated z-scores. (b) Funnel plot for all 1989 data with known study size. (c) Funnel plot for studies with principal author Helmut Schmidt.

TABLE 1
Results of Several Methods for Combining the 1989 Data

Method	Accounts for study size	Accounts for between-studies variation	Hit rate	SE of hit rate	z	Probability of z (one tail)
Stouffer z	No	No	—	—	15.8	3×10^{-56}
Testing mean z	No	Yes	—	—	9.8	6×10^{-23}
Pooling hits	Yes	No	0.50016	0.000023	6.8	3×10^{-12}
Pooling hits, correction for non-unit variance	Yes	Yes	0.50016	0.000037	4.2	1×10^{-5}
Unweighted average hit rate	No	Yes	0.50449	0.00076	5.8	3×10^{-9}
Random-effects model	Yes	Yes	0.50033	0.00026	1.3	0.096

hypothesis behind the experiments is that subjects can enhance hit rates, methods not based on hit rate are not tests of the original hypothesis.

The fact that the composite z drops dramatically when weighting by square-root of study size indicates that there is a substantial component of small studies with large hit rates which are not replicated in large studies. This means that the data are heterogeneous, i.e., the variation in hit rates is much greater than expected from sampling error, i.e., the studies are not all measuring the same thing. Heterogeneity is a serious problem, because the composite result may involve the mean of a multimodal distribution, and a standard deviation based only on sampling error is incorrect. The data are in fact so heterogeneous that a χ^2 test of the pooling-hits method has a probability of 1×10^{-130}. Even when arandom-effects model is used in an attempt to account for variation between studies, the χ^2 probability is only 8×10^{-74}. The between-studies variation is clearly enormous and not understood, though as I will discuss later, at least some of it is due to publication bias.

In considering the various composite z-scores discussed above, it is worth noting that Matthews (1999) has argued that hypotheses approached with moderate skepticism should not be accepted unless $p < 1.3 \times 10^{-5}$, which corresponds to a one-tailed z-score of 4.2; for high skepticism, he suggests $p < 1.0 \times 10^{-6}$, which corresponds to a one-tailed z-score of 4.75. Of course, the usual $p = 0.05$ criterion corresponds to a one-tailed z-score of only 1.645.

Changes to the Dataset

While both *The Conscious Universe* (Radin, 1997) and the 2003 paper (R&N, 2003) refer to the 597-study dataset of the 1989 paper, it turns out that they really refer to two rather different datasets, each of which is different from that of the original 1989 paper. The 1989 dataset had only 313 non-PEAR studies, while *The Conscious Universe* and the 2003-paper datasets have 339, so to get 597 studies, some PEAR studies would have had to be dropped. Comparing the non-PEAR portions of these datasets, I found that 30 studies had been added (involving plant and animal subjects) and four studies had been dropped. The

four dropped studies had nonsense hit rates (>1), but oddly, one such study was left in the database.

Comparing the data used for the 2003 paper to those for The Conscious Universe, I also found 15 z's had changed from negative to positive, and 15 from positive to negative. These sign changes are particularly puzzling. One might assume that the sign changes would be corrections, but I compared five of them to the original papers, and in my opinion, all five were correct in the original version.

The changes in dataset account for the discrepancy in Stouffer z of "more than 12" quoted in the 2003 paper for the 1989 data, versus 15.76 for the original dataset. The sign changes have a material effect on the quality analysis, as discussed below.

Nonsense Hit Rates

The hit rate of 0.509 quoted for the 1989 data in both The Conscious Universe and the 2003 paper is much larger than any hit rate shown in Table 1. Partly this is because the datasets are somewhat different from the original 1989 dataset, but it turns out that mostly it is because the calculation that gives 0.509 includes a number of nonsense hit rates. Radin calculates the hit rate as $0.5 \cdot (1 + z_i/\sqrt{n_i})$, where z_i and n_i are the z-score and number of random numbers of the ith study. Unfortunately, when n_i is unknown, R&N code this as $n_i = 1$, so if $z_i > 1$, the hit rate can be greater than 1, which of course is nonsense. Using the original 1989 data and excluding nonsense hit rates, I find that the unweighted average hit rate is only 0.50449.

Evidence of Publication Bias and Other Artifacts

One well-known source of bias and heterogeneity in meta-analysis is publication bias, where studies with undesirable (e.g., statistically insignificant) outcomes are less likely to be published. In their 1989 paper R&N argue that publication bias is unlikely to cause all of the statistical significance, based on fitting two ad-hoc models to the data plus an estimate of the "filedrawer" of unpublished studies, using Rosenthal's (1991) assumption that unpublished studies have $z = 0$ on average. Scargle (2000) notes that small errors in fitting the high tail can result in large errors when extrapolating to the body of the distribution, and he questions R&N's use in one case of an exponential model for the tail. Scargle also notes that Rosenthal's assumption of $z = 0$ for unpublished studies does not make sense, since if all unpublished studies have $z < 1.645$, then the filedrawer is biased and the average z of unpublished studies is -0.1085, not 0. Using Rosenthal's method, R&N estimate a "failsafe" number of 54,000 unpublished studies, but had they used $z = -0.1085$ instead of $z = 0$ for unpublished studies, the estimate would have been only 2681. There are clearly pitfalls in estimating the number of unpublished studies using just the average z and number of published studies.

R&N quote Rosenthal's criterion that an effect is robust if the estimate of unpublished studies is more than five times the number published. This criterion is largely ad-hoc, however, and is expected to vary by field. In some fields, the marginal effort and expense of conducting a study is so high that very few studies ever go unreported. The RNG studies are probably not in that category, since the time it takes to conduct another study can be fairly modest. In one hour, even a very slow RNG that generates one numberlsecond can generate a study of 3600 numbers, and a typical moderate-speed device generating 1000 numbers/second can produce a study of 3.6 million.

There are standard meta-analytic techniques for checking for publication bias. One is the funnel plot, a scatter plot of effect size versus study size (I will use hit rate, which is closely related to R&N's effect size). Because sampling error on the hit rate is proportional to $1/\sqrt{n_i}$, the plot should have the shape of an inverted funnel, as shown in Figure 2a (generated by pairing actual study sizes from the 1989 data with random standard normal z-scores). The 1989 data are plotted in Figure 2b, with lines showing the threshold for one-tailed significance. A log scale has been used so that the whole range of study sizes can be seen, and studies with unknown sample sizes have been omitted.

The funnel of Figure 2b is highly asymmetric, which is often a warning sign of publication bias. A well-known test for publication bias (Begg & Mazumdar, 1994) gives a highly significant z-score of 7.2 for these data. The asymmetry comes primarily from a long curved band of small studies with large effect sizes. This band seems to follow the lines showing the threshold for one-tailed significance, shown by the solid line. A similar band can be seen even more clearly in Figure 2c, which shows the subset of data for the second most prolific principal author, Helmut Schmidt (PEAR is the most prolific group, but their study sizes cluster tightly around a few fixed values, making a scatter plot difficult to interpret).

Asymmetry in a funnel plot can be caused by things other than publication bias, such as a real difference in subjects or methodology between small studies and large ones. One would not expect such effects to cause the hit rate to vary along lines of constant statistical significance, as they do here. Also, there is **direct** evidence for publication bias in at least one original paper (Schmidt, 1976), where it is explicitly stated that complete records were not kept for some studies which did not produce positive results (this of course would not have been an issue if R&N did not use the remaining studies, but they did).

Kennedy (2001) argues that experimenters could use their own psychic powers to influence the outcome of a study to be statistically significant. This would explain the discontinuity at $z = 1.645$, but publication bias also explains it, is non-paranormal and well-known to occur in other fields, and in my view is therefore the preferred explanation.

There are other methodological problems which mimic publication bias but require far fewer unpublished studies. For example, in many of the original papers, the number of trials is not fixed ahead of time, so experiments might be

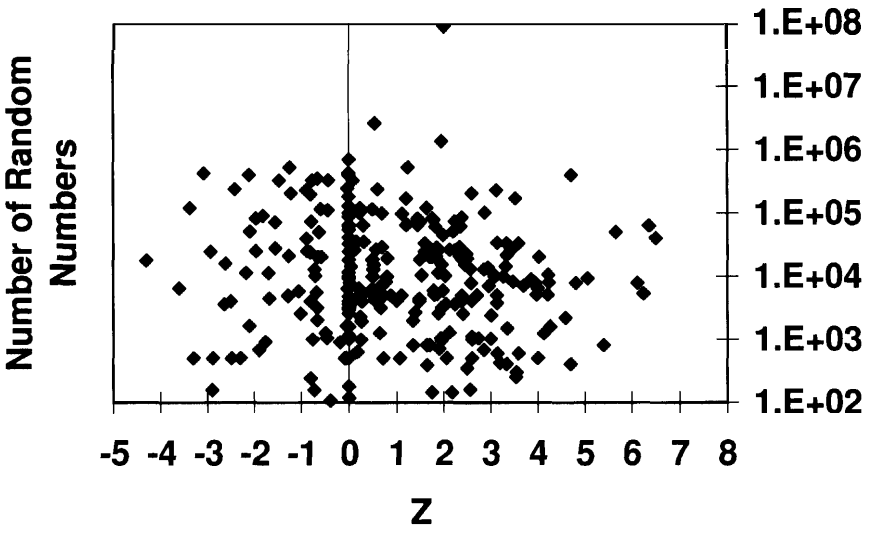

Fig. 3. **Study size (n) versus z-score, PEAR data treated as one study.**

run until a fixed statistical significance is obtained. In some studies, the direction of intention is not clearly defined at the start (Alcock, 1989), leaving experimenters latitude to set the signs of z-scores to the desired direction, thus cutting the required number of unpublished studies by a factor of two.

The histogram fitted by R&N is much smoother because they broke the PEAR data up into 284 separate studies and distributed the studies with unknown but non-significant *z*'s randomly between $z = -1.645$ and $z = 1.645$.[1] With the PEAR data treated as one study and the "non-significant" unknown-z studies shown at $z = 0$ (as originally coded by R&N), as shown in Figure 1, the histogram is much more ragged.

One odd feature of the z distribution is the very small number of studies between about 0 and -0.4. It is perhaps more clearly visible in Figure 3, a plot of *n* versus z. The gap is almost certainly an artifact, but the cause is puzzling. Perhaps it represents a tendency of experimenters to report mildly negative results simply as "non-significant", rather than quoting an exact z. If this is the case, the artifact biases the average *z* and hit rate higher, because when an exact z is not quoted, R&N assign the study $z = 0$.

Unfortunately, there is no completely objective way to correct for the various artifacts, but it is nevertheless important to understand the z distribution as best we can. Figure 1 shows a somewhat ad-hoc fit of two normal distributions with mean 0 and standard deviation 1 (i.e., a null effect), one which extends over the whole graph and one of which is cut off at $z < 1.645$ to simulate a null effect with publication bias.[2] There is qualitative agreement over most of the graph, but the publication-bias curve drops too fast to account for the entire high tail, and there is a small excess low tail as well. To give some idea of the size of the

various components, the fit gives a total of 210 studies, of which 60 are from the truncated curve (implying a filedrawer of 1200 unpublished studies, if no other bias contributes). In addition, there is an excess over the fit of 47 studies at $z = 0$, since I have preserved R&N's coding for studies reported simply as non-significant. There is an excess over the fit of nine studies with $z < -2$ and of 69 studies with $z > 2$. Since certain bins have had to be excluded (to avoid the artifact near $z = 0$ and bins whose contents are too small to treat as Gaussian), these numbers and the fit should not be taken too literally. They are intended to illustrate that the z distribution bears a resemblance to a null effect plus publication bias, plus some unexplained tails (which may well owe their existence to poor methodology).

The 1989 paper describes a method of trimming the data for heterogeneity, and presents several results based on this trim. Duplicating their method, I find that for this dataset it is roughly equivalent to a trim at z of $\pm$ 2.2, and is therefore not sufficient to exclude the publication-bias region.

Quality

Perfect methodological quality is a logical necessity in parapsychology, since a paranormal effect can only be demonstrated if *all* possible non-paranormal causes are ruled out. As we will see, most studies are far from ideal. Rather than confine the meta-analysis to perfect studies, R&N have rated studies according to 16 quality factors, and they argue that an analysis of the quality data shows no relation between effect size and quality.

Even apart from some curious features of the data which I will discuss presently, there are shortcomings to R&N's approach. First, any analysis where independent variables (like quality factors) are *controlled* is necessarily more convincing than any analysis where they vary by happenstance. Also, the quality judges were not blind to the outcomes of studies (Radin, 2004), so the ratings may be biased. Further, in R&N's analysis, studies which explicitly note that they do *not* have a given factor are treated the same as if the factor is not mentioned. The analysis might be more sensitive if for each factor it compared only the known-good studies to the known-bad ones, rather than lumping the unknown and the bad together. It is also worth noting that in the past, methodological flaws have proven difficult to assess from published reports, and finding them has depended on reports of independent skilled observers who were present during the experiment (Diaconis, 1978), scrutiny which occurred in probably only a small fraction of studies. One might also question whether the 16-factor scheme can adequately capture the wide variety of methodological flaws seen in the literature (as described in, e.g., Alcock, 1989).

Dean Radin has supplied me with the quality data, except those for the PEAR studies, which he is unable to locate. The average quality score in these data (formed for each study by adding the scores for the 16 factors) is quite low—only 4.4 out of a possible 16. Many quality features are surprisingly rare.

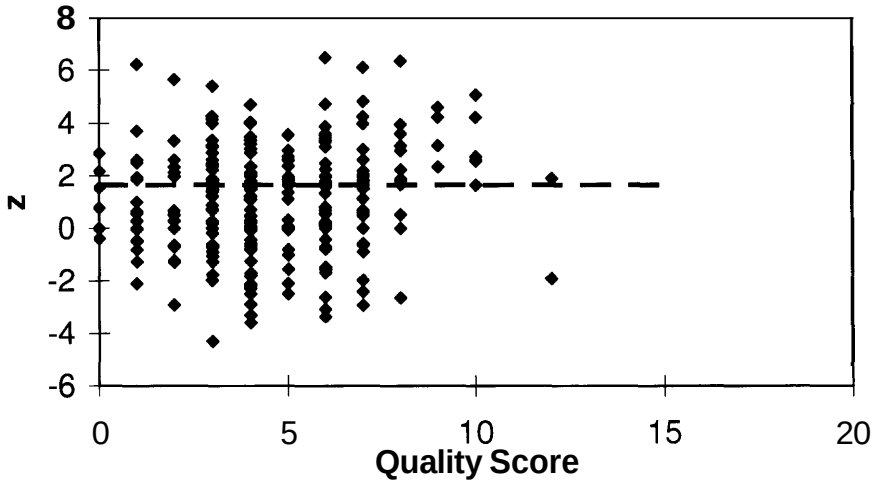

Fig. 4. z-score versus quality score. The dashed line indicates the threshold for one-tailed statistical significance.

Automatic recording, often touted as one of the chief advantages of RNG's versus earlier experiments in parapsychology, is known present in only 57% of the non-PEAR studies. Control tests were noted in only 56%, and randomness checks in only 13%, a serious problem, since both are needed to be sure that outcomes are not just the results of quirks of the RNG. Alternating targets on successive trials can somewhat mitigate the effects of a biased RNG, but it is known to have been done in only 11.6% of non-PEAR studies. Data are known to have been archived in only 2%, and double-checked in only 13%, and data selection is known to have been prevented in only 7%. R&N include a factor for fixed run lengths (known present in only 65% of studies), but this does not indicate that the size of the experiment was fixed in advance.

In the 1989 paper, the effect sizes and quality scores are averaged for each experimenter before fitting (thus losing information). I am unable to directly verify their fits, since I do not have the PEAR data, but surprisingly, fitting the non-PEAR data gives very different slopes to those in the paper. For example, weighting by n, I get a slope of $-1.7 \times 10^{-3} \pm 1.2 \times 10^{-3}$, which is not significant, but is still two orders of magnitude larger than their $-2.5 \times 10^{-5} \pm 3.2 \times 10^{-5}$. In my fit, an extrapolation to the perfect study (quality of 16) actually gives a negative effect size.

The analysis is done in a different way in The Conscious Universe—a simple correlation coefficient is calculated between hit rate and quality. Radin concludes that hit rates are not related to quality, but the correlation coefficient he quotes is 0.165 (degrees of freedom [d.o.f.] = 339), which is significantly positive ($p = 0.002$, two-tailed).[3] The positive value is very interesting, because it means that high-quality studies have a higher average hit rate. In Figure 4, we

see how the positive value comes about. The highest-quality studies have a cutoff near $z = 1.645$, strongly suggesting that publication bias is confounded with quality. It is as if authors resort to publication bias when other sources of positive results are eliminated by methodological safeguards. In any case, confounding of quality with publication bias makes any analysis of the quality data unreliable.

The average quality for studies in the high tail (here taken as $z > 1.645$) is a very poor 5.1. This is a bit higher than the dataset taken as a whole, but this is at least partly due to the confounding with publication bias. Pallikari (2003) attempts to explain the high tail as a complicated psychokinetic effect, but given the low quality, I feel it is premature to accept a paranormal explanation.

It is interesting to note that the sign changes in the 2003-paper dataset have a substantial effect on the correlation between hit rate and quality. Had The Conscious Universe used the same (incorrect) signs as the 2003 paper, the correlation would have dropped to 0.01.

Control Studies

Even the control studies do not behave as expected, and show evidence of what are probably reporting biases. The standard deviation of the control-study z-scores is only 0.82, significantly less than the expected 1 ($p = 5 \times 10^{-5}$). There is a small effect from R&N assigning $z = 0$ to all control studies which are noted simply as "non-significant", but replacing these with random z's brings the standard deviation up to only 0.87. The PEAR control data are dramatically deficient in the tails, as has been noted by Jahn and Dunne (1987). Even the non-PEAR data have too small a standard deviation (0.69, or 0.82 with the 0's replaced by random standard normal z's).

Repeatability

There are a number of ways one can argue that the effect purportedly seen in this meta-analysis is not repeatable. The χ^2 test mentioned earlier checks whether studies repeatedly measured the same hit rate, differing only due to sampling error. It fails dramatically ($\chi^2 = 1730$ for 596 d.o.f., $p = 10^{-130}$). If there were any repeatability of hit rates across studies of different sizes, then in Figure 3 one would expect to see z-scores growing as the square-root of study size, but no such pattern is evident. If the hit rate of 50.9% quoted by R&N were repeatable, the pattern would be very obvious.

Interestingly, only 33 of the 66 authors[4] have ever published a statistically significant study. It is also worth noting that when pooling hits and correcting for non-unit variance, the overall result can be made non-significant by removing just three authors: PEAR, Helmut Schmidt, plus one of several others.

Radin (1997) has argued that the continued success of the PEAR program validates the results of the 1989 meta-analysis. By this reasoning, the subsequent

failure by PEAR and others to replicate the original PEAR findings (Jahn et al., 2000) argues that the meta-analysis is not valid.

Concluding Remarks on the 1989 Paper

We have seen that the statistical significance of the final result depends strongly on the method used to combine the data. This comes about because of artifacts, including publication bias, for which there is both direct and indirect evidence. There is no objective way to correct for the artifacts. There are more studies in the tails than can be accounted for by publication bias alone. Unfortunately, one can only guess at the underlying distribution that causes the tails, because only a portion, possibly a very small portion, is visible, due to the publication-bias cutoff and possible overlap with the null distribution. The low methodological quality of the studies in the tails suggests that they may be the result of poor quality. R&N's conclusion that quality has no effect suffers from, among other things, the fact that quality is confounded with publication bias. Even if one ignores between-study variation and publication bias, as long as study size is accounted for, the effect must be extremely small.

In short, almost 30 years of work in this field has arguably failed to conclusively demonstrate anything anomalous.

The 2003 Meta-Analysis

In 2003 R&N published a meta-analysis of an updated dataset consisting of the 1989 paper's dataset plus 176 new studies found by Steinkamp and collaborators. Because Steinkamp et al. have not yet published a full analysis, it would be inappropriate for me to attempt to re-analyze these data, so I will confine myself to remarks on the published analyses.

Despite the 1989 paper's claim of including "all known studies", a new search by Steinkamp et al. found 84 studies in the same time period as the 1989 paper (R&N, 2003: 41), but not included in it (a 25% increase, if the PEAR studies are counted as one datapoint, as is done in the 2003 paper). The missed studies are primarily from conference proceedings and non-English-language journals, according to the 2003 paper. Also, as discussed above (but not noted in the 2003 paper), some changes have been made to the 1989 portion of the dataset.

The 2003 paper gives the Stouffer *z* value of 16.1 as its primary result and quotes no other z-score. The Stouffer *z* is quoted as if it is the significance of the average hit rate, but in the 1989 data, the significance of the average hit rate is far lower than the Stouffer *z*. The change of analysis method is also odd because in 1997, Radin's The *Conscious Universe* had shown the significance that was actually calculated from an average hit rate, and Nelson had co-authored a paper that in part argues against using Stouffer's method to combine PEAR data (Jahn et al., 1997). Unlike the 1989 paper, the 2003 paper does not show control studies at all.

R&N find that *z* is independent of sample size, which they interpret to mean

that the results cannot be explained by a "simple, linear, force-like mechanism", or a "simple causal process". This seems a very bold interpretation to me. Lack of correlation could result from a null effect, or a null effect plus publication bias. Even if the effect were real, the correlation would be very small, and possibly non-significant, if the hit rate were also very small. The hit rate quoted by R&N (50.7%) is so large, on the other hand, that the correlation would be highly significant if that hit rate applied to large studies as well as small.

The 2003 paper provides graphs of the cumulative Stouffer *z* through time. There are two periods of about 10 years each during which the cumulative *z* is practically flat (unchanged) (the periods are roughly 1975–1985 and 1990–2000). There are thus two decades which have a much lower average z-score than the other two, which is not what one would expect from a stationary, repeatable effect.

Steinkamp et al. (2002a,b) have published an abstract of a preliminary analysis of their own for the same time period, but a somewhat more restricted dataset (including only human subjects, non-pseudo RNG's, and forward-in-time studies with unambiguous intention). In terms of studies, their dataset is not dramatically smaller (357 experimental studies, versus 515 for R&N), but it does appear to be much smaller in terms of total random numbers. They calculate Rosenthal and Rubin's (1991a,b) π by pooling hits, and quote $n = 0.50003$ for experimental, and the *same* value for control studies. For control studies, π drops to 0.49999 once one large control run is removed. The statistical significance of π is $z = 2.7$ (Steinkamp et al., 2002b), far lower than the Stouffer *z* of 13.09.

Conclusion

Dean Radin and Roger Nelson have conducted an important survey of the RNG literature, which must have involved a prodigious amount of labor. Unfortunately I must disagree with their conclusion that "Meta-analysis of 515 RNG experiments conducted by 91 researchers over a span of 41 years indicates the presence of a small magnitude, but statistically highly significant and repeatable mind-matter interaction effect" (R&N, 2003).

In my view their conclusion is questionable on every point. We have seen in data from the 1989 paper that the statistical significance is questionable, varying widely depending on the method used to combine the data. This wide variation is due at least in part to artifacts, including publication bias. Even if the result is significant, the significance is partly the result of artifacts, and there is no objective way to adequately correct for these. The hit rate given in the 2003 paper and characterized as "small" is a simple average, with no accounting for study size, and is far higher than hit rates obtained from combining the data in other ways. The z-scores do not increase with study size as one would expect if there were a consistent effect of the size quoted by R&N. The repeatability is also questionable, based on the 1989 data, given the huge heterogeneity, the fact

that half the authors have never published a significant effect, etc. Overall quality is very low, and the quality analysis is unreliable because of confounding with publication bias. I would argue that 41 years of work in this field has produced no convincing evidence for a real effect.

The 1989 data strongly suggest that there is a substantial component of null results in the data, but the question of whether humans can influence electronic RNG's cannot be resolved with complete certainty until publication and other reporting biases are eliminated and methodological quality is improved. To eliminate publication bias, parapsychologists could set up pre-registration systems analogous to those which have been proposed for medical clinical trials, so that the detailed design of an experiment is fixed before it is conducted. To avoid methodological pitfalls, it would be useful to compile a complete list of methodological best practices which future experiments should follow rigorously. It is also important for purposes of future meta-analysis for experimenters to always report full numerical details of results for both experimental and control studies, rather than simply reporting that some results were "non-significant".

M. H. SCHUB
P.O. Box *6301*
Minneapolis, MN *55406*

Notes

[1] Distributing the studies with unknown but non-significant *z*'s randomly over $-1.645 < z < 1.645$, as done in the 1989 paper's histogram, cuts off two tails but uses the threshold value for one-tailed significance. Using a symmetric cutoff also assumes an unbiased sample, just as in the unbiased-filedrawer calculation, whereas these studies would be biased if only one tail were cut off, as may well be the case.

[2] The edge null-effect-plus-publication-bias curve is not perfectly vertical because I have kept the same binning as in the 1989 paper (for ease of comparison), so the bin edge is not exactly at $z = 1.645$. The fit result for this overlapping bin has been adjusted for this.

[3] The degrees of freedom (339) indicate that the PEAR studies were treated as one study, not broken up extensively as in the Stouffer z calculation. Breaking up the data might well change the conclusion in this case, because the PEAR data have a small effect size and high quality. I was not supplied with the PEAR quality data, but The Conscious Universe says that all PEAR studies had scores greater than 12, i.e., as it turns out, greater than any non-PEAR study.

[4] R&N quote a total 68 different investigators, but they appear to be multiple-counting; in their data I find only 66. They also quote 152 references, but in their data I find only 140 non-PEAR references. Their data do not list ref-

erences for the PEAR data, but they would have to account for a surprisingly-high 12 papers for their count to be correct.

Acknowledgments

I am indebted to Dean Radin and Roger Nelson for providing me with data from the 1989 meta-analysis and for helpful discussions and corrections, and Fiona Steinkamp and Emil Boller for helpful discussions and corrections.

References

Alcock, J. (1989). A comprehensive review of major empirical studies in parapsychology involving random event generators or remote viewing. National Academy Press. Available at http://www.nap.edu/books/POD276/html/602.html. Accessed 31 January 2004.

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

Dawes, R. M. (1987). Random generators, ganzfelds, analysis and theory. *Behavioral and Brain Sciences, 10,* 581–582.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Clinical Trials,* 7, 177–188.

Diaconis, P. (1978). Statistical problems in ESP research. *Science, 201,* 131–136.

Jahn, R. D., & Dunne, B. J. (1987). *Margins of Reality—The Role of Consciousness in the Physical World.* Harcourt, Brace, Jovanovich.

Jahn, R. D., Dunne, B. J., Nelson, R. D., Dobyns, Y. H., & Bradish, G. J. (1997). Correlation of random binary sequences with pre-stated operator intention: A review of a 12-year program. *Journal of Scientific Exploration, 11,* 345–367.

Jahn, R. D., Dunne, B., Bradish, G., Dobyns, Y., Lettieri, A., Nelson, R., Mischo, J., Boller, E., Bosch, H., Vaitl, D., Houtkooper, J., & Walter, B. (2000). Mind/machine interaction consortium: PortREG replication experiments. *Journal of Scientific Exploration, 14,* 499–555.

Kennedy, J. E. (2001). Why is psi so elusive? A review and proposed model. *Journal of Parapsychology, 65,* 219–246.

Matthews, R. A. J. (1999). Significance levels for assessment of anomalous phenomena. *Journal of Scientific Exploration, 13,* 1–7.

Nelson, R. D. (1999). The physical basis of intentional healing systems (PEAR Technical Report 99001). Princeton Engineering Anomalies Research, Princeton University, School of Engineering/Applied Science. In Jonas, W. B., & Levin, J. S. (Eds.), *Textbook of Complimentary and Alternative Medicine, Part I.* Williams and Wilkins.

Pallikari, F. (2003). Must the "magic" of psychokinesis hinder precise scientific measurement? *Journal of Consciousness Studies, 10*(1–7), 199–219.

Palmer, J. A., Honorton, C., & Utts, J. (1989). Reply to the National Research Council study on parapsychology. *Journal of the American Society for Psychical Research, 83,* 31–49.

Radin, D. I. (1997). *The Conscious Universe.* HarperEdge.

Radin, D. I. (2004). Personal communication.

Radin, D. I., May, E. C., & Thomson, M. J. (1985). Psi experiments with random number generators: Meta-analysis part 1. In Wiener, D. H., & Radin, D. I. (Eds.), *Research in Parapsychology 1985.* Metuchen, NJ: Scarecrow Press.

Radin, D. I., & Nelson, R. D. (1988). Statistically robust anomalous effects: Replication in random event generator experiments.In Henckle, L., & Berger, R. E. (Eds.), *Research in Parapsychology 1988.* Metuchen, NJ: Scarecrow Press.

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

Radin, D. I., & Nelson, R. D. (2003). A meta-analysis of mind-matter interaction experiments from 1959 to 2000. In Jonas, W. B., & Crawford, C. C. (Eds.), *Healing, Intention and Energy Medicine: Research Methods and Clinical Applications* (pp. 39–48). Edinburgh, UK: Churchill Livingstone. A nearly identical paper can be found on the Internet at: www.boundaryinstitute.org/articles/rngma.pdf. Accessed 6 December 2003.

Rosenthal, R. R. (1991). *Meta-Analytic Procedures for Social Research,* Newbury Park, CA: SAGE Publications.
Rosenthal, R. R., & Rubin, D. B. (1991a). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin, 106,* 332–337.
Rosenthal, R. R., & Rubin, D. B. (1991b). Further issues in effect size estimation for one-sample multiple-choice-type data. *Psychological Bulletin, 109,* 351–352.
Scargle, J. (2000). Publication bias: The "file-drawer" problem in scientific inference. *Journal of Scientific Exploration, 14,* 91–106.
Schmidt, H. (1976). PK effect on pre-recorded targets. *Journal of the American Society for Psychical Research, 70,* 267–291.
Shaffer, J. P. (1991). Comment on "Effect size estimation for one-sample multiple-choice-type data: design, analysis, and meta-analysis" by Rosenthal and Rubin. *Psychological Bulletin, 109,* 348–350.
Steinkamp, F., Boller, E., & Bosch, H. (2002a). Experiments examining the possibility of human intention interacting with random number generators: A preliminary meta-analysis [Abstract]. *Journal of Parapsychology, 66,* 238–239.
Steinkamp, F., Boller, E., & Bosch, H. (2002b). Experiments examining the possibility of human intention interacting with random number generators: A preliminary meta-analysis. *Proceedings of the 45th Annual Convention of the Parapsychological Association, Aug. 2002.* Unpublished.
Utts, J. (1991). Replication and meta-analysis in parapsychology. *Statistical Science, 6,* 363403.

Appendix—Comments on Various Ways to Combine the Data

In what follows, all numerical results are based on the data used in the 1989 paper.

1. Stouffer's Method: Earlier and later meta-analyses (Radin et al., 1985; R&N, 2003) present z-scores calculated by Stouffer's method (Rosenthal, 1991). A composite z-score is calculated by summing the z-scores for the individual studies, then dividing by the square-root of the number of studies K: $z_s = \sum z_i / \sqrt{K} = \bar{z}\sqrt{K}$. It is a simple test of the hypothesis $\bar{z} = 0$, under the assumption that the standard deviation s_z is exactly 1 (it is trivially derived from the familiar $t = \bar{z}\sqrt{K}/s_z$ by setting s_z to 1).

The number of experimental studies K = 597, and and $\bar{z} = 0.645$, so $s_z = 0.645 \cdot \sqrt{597} = 15.76$. This is the value quoted by Palmer et al. (1989) and is close to the one quoted by Nelson (1999).

The Stouffer z gives equal weight to all studies, even though the sample size (the number of random numbers per study) varies over six orders of magnitude. Equal weights also means that papers which are broken up into more studies get more weight. One extreme example is the PEAR data, which account for 258 of the 597 experimental studies in the 1989 paper (43%), and therefore carry 43% of the weight in the Stouffer z (this fact is not noted in the 1989 paper, but was later reported in Radin [1997]). Collapsing the PEAR data to a single study would have increased the Stouffer z from 15.76 to 16.95.

Unfortunately, s_z is 1.601, quite a bit larger than assumed, and this should be corrected for, as is done in the next method.

It is worth noting that the z distribution is significantly skewed (R&N, 1988), so non-normality is a concern.

2. Method of Testing Mean z: Substituting the real-world standard deviation into the expression above gives $t = \bar{z}\sqrt{K}/s_z = 15.76/1.601 = 9.8$. This is what

Rosenthal (1991) calls the "method of testing mean z" (the Stouffer z is of course also a test of mean z, but I will use Rosenthal's terminology). It can be thought of as correcting the Stouffer z for between-studies variation by assuming that the z's are all drawn from the same distribution with the real-data standard deviation. This is a t statistic, but can be treated as a z in this case due to the large number of degrees of freedom.

3. *Effect Size/Weighting* z's by $\sqrt{n_i}$/*Pooling* Hits: The 1989 paper combines results by defining an effect size $z_i/\sqrt{n_i}$ and weighting by n_i, where n_i is the number of random numbers in the ith study. The procedure is exactly equivalent to weighting the z-scores by $\sqrt{n_i}$. The results are not quoted, but are shown in a graph in the paper (their figure 3); the net effect size is 3.18×10^{-4} with a standard error of 0.47×10^{-4}, from which I calculate a z-score of 6.8. This is a long way from the raw Stouffer z, which will come as a surprise to readers who are used to textbook examples where several different methods of combining results all give the same significance. All that has changed from the Stouffer z is the relative weight of studies, so the big drop in the result means that the Stouffer z was being pulled up substantially by small studies with large z-scores. This surprising sort of heterogeneity was noted by Dawes (1987) in the data from the earlier meta-analysis (Radin et al., 1985).

Radin calculates hit rates as $0.5 \cdot (1 + z_i/\sqrt{n_i})$, and in this scheme it can be shown that the significance for the pooled hit rate (average of hit rate weighted by $\sqrt{n_i}$) is identical to that for weighting z-scores by $\sqrt{n_i}$. The composite hit rate, which is easier to interpret than R&N's effect size, turns out to be 0.50016 with a standard error of 0.0000233. Calculating a composite hit rate requires converting to a common basis (in this case, that for a binary RNG), and Radin's is not the only way to do this (Rosenthal & Rubin, 1991a,b; Shaffer, 1991).

These methods account for study size but still assume no between-studies variation, apart from sampling error. The hypothesis being tested is that the weighted average z is 0 (or equivalently, that the weighted average hit rate is 50%).

4. Weighting z's by $\sqrt{n_i}$ and Correcting for Non-unit Variance: One way to correct the above weighted average for the effect of non-unit variance is to assume that all the z_i are drawn from the same distribution with the observed standard error of $\sigma_z = 1.601$ This gives $6.8/1.601 = 4.2$, which is close to the value of 4.1 quoted by Utts (1991).

5. Averaging Hit Rates: The Conscious Universe (Radin, 1997) quotes a hit rate "per study", indicating that hit rates are calculated for each study and then averaged. From the graph in the book (his figure 8.5), I read an average hit rate of 0.509, with a standard error (SE) of 0.0021, for a z-score of about 4.3. The 2003 paper also refers to a hit rate of 0.509 for the 1989 dataset. As discussed above, these numbers are actually for a somewhat different dataset from that used for the 1989 paper, and are also influenced by the inclusion of one non-sensical hit rate which is greater than 1. Using only hit rates <1, I find a considerably lower hit rate of 0.50449 with smaller SE of 0.00077, for a z-score of 5.8. This method does not account for study size, but does account for

between-study variation in its SE. This unweighted average hit rate, even excluding the nonsense hit rates, is far higher than the pooled hit rate, again indicating substantial influence from small studies with large hit rates. The hypothesis tested is that the unweighted average hit rate is 50%.

6. *Random-Effects* Model: **A** standard meta-analytic method of coping with heterogeneity is to use a random-effects model, which models both within-study sampling variance and variance between study effect sizes. Unlike some methods discussed above, where the *z*'s are assumed to come from the same distribution, here the assumption is that the effect sizes come from two rounds of sampling. First, a study effect size is chosen from a distribution with the between-studies variance, and then the observed hit rate is chosen from a distribution with the within-study variance. Using the non-iterated estimator of DerSimonian and Laird (1986), I find a final hit rate of 0.50033, with an SE of 0.00026, for a z-score of 1.3, which is non-significant. The hypothesis tested is that the weighted average hit rate is 50%, where the weights account for both sampling error and between-studies variation.