*Reprinted from FOUNDATIONS OF PHYSICS

Vol. 19, No. 12, December 1989
Printed in Belgium

Evidence for Consciousness-Related Anomalies in Random Physical Systems

Dean I. Radin[1] and Roger D. Nelson[2]

Received May 6, 1988; revised June 12, 1989

Speculations about the role of consciousness in physical systems are frequently observed in the literature concerned with the interpretation of quantum mechanics. While only three experimental investigations can be found on this topic in physics journals, more than 800 relevant experiments have been reported in the literature of parapsychology. A well-defined body of empirical evidence from this domain was reviewed using meta-analytic techniques to assess methodological quality and overall effect size. Results showed effects conforming to chance expectation in control conditions and unequivocal non-chance effects in experimental conditions. This quantitative literature review agrees with the findings of two earlier reviews, suggesting the existence of some form of consciousness-related anomaly in random physical systems.

1. INTRODUCTION

The nature of the relationship between human consciousness and the physical world has intrigued philosophers for millenia. In this century, speculations about mind–body interactions persist, often contributed by physicists in discussions of the measurement problem in quantum mechanics. Virtually all of the founders of quantum theory—Planck, de Broglie, Heisenberg, Schrödinger, Einstein—considered this subject in depth,[1] and contemporary physicists continue this tradition.[2–7]

[1] Department of Psychology, Princeton University, Princeton, New Jersey 08544. Present address: Contel Technology Center, 15000 Conference Center Drive, P.O. Box 10814, Chantilly, Virginia 22021-3808.

[2] Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, New Jersey 08544.

0015-9018/89/1200-1499$06.00/0

The following expression of the problem can be found in a recent interpretation of quantum theory:

> If conscious choice can decide what particular observation I measure, and therefore into what states my consciousness splits, might not conscious choice also be able to influence the outcome of the measurement? One possible place where mind may influence matter is in quantum effects. Experiments on whether it is possible to affect the decay rates of nuclei by thinking suitable thoughts would presumably be easy to perform, and might be worth doing.[8]

Given the distinguished history of speculations about the role of consciousness in quantum mechanics, one might expect that the physics literature would contain a sizable body of empirical data on this topic. A search, however, reveals only three studies.

The first is in an article by Hall, Kim, McElroy, and Shimony, who reported an experiment "based upon taking seriously the proposal that the reduction of the wave packet is due to a mind–body interaction, in which both of the interacting systems are changed."[9] This experiment examined whether one person could detect if another person had previously observed a quantum mechanical event (gamma emission from sodium-22 atoms). The idea was based on the supposition that if person A's observation actually changes the physical state of a system, then when person B observes the same system later, B's experience may be different according to whether A has or has not looked at the system. Hall *et al.*'s results, based on a total of 554 trials, did not support the hypothesis; the observed number of "hits" obtained in their experiment was precisely the number expected by chance (277), while the variance of their measurements was significantly smaller than expected ($p < 0.05$).[9]

The second study is referred to by Hall *et al.*, who end their article by pointing out that a similar, unpublished experiment using cobalt-57 as the source was successful (40 hits out of 67 trials).[10]

The third study is a more systematic investigation reported by Jahn and Dunne,[11] who summarize results of over 25 million binary trials collected during seven years of experimentation with random-event generators. These experiments, involving long-term data collection with 33 unselected individuals, provide persuasive, replicable evidence of an anomalous correlation between conscious intention and the output of random number generators.

Thus, of three pertinent experiments referenced in mainstream physics journals, one describes results statistically too close to chance expectation and two describe positive effects.[9–11] Given the theoretical implications of such an effect, it is remarkable that no further experiments of this type can be found in the physics literature; but this is not to say that no such experiments have been performed. In fact, dozens of researchers have

reported conceptually identical experiments in the puzzling and uncertain domain of parapsychology. Perhaps because of the insular nature of scientific disciplines, the vast majority of these experiments are unknown to most scientists. A few critics who have considered this literature have dismissed the experiments as being flawed, nonreplicable, or open to fraud,[12–16] but their assertions are countered by at least two detailed reviews which provide strong statistical support for the existence of anomalous consciousness-related effects with random number generators.[17,18] In this paper, we describe the results of a comprehensive, quantitative meta-analysis which focused on the questions of methodological quality and replicability in these experiments.

2. THE EXPERIMENTS

The experiments involved some form of microelectronic random number generator (RNG), a human observer, and a set of instructions for the observer to attempt to "influence" the RNG to generate particular numbers, or changes in a distribution, solely by intention. RNGs are usually based upon a source of truly random events such as electronic noise, radioactive decay, or randomly seeded pseudorandom sequences.[19] Feedback about the distribution of random events is often provided in the form of a digital display, but audio feedback, computer graphics, and a variety of other mechanisms have also been used. Some of the RNGs described in the literature are technically sophisticated, the best devices employing electromagnetic shielding, environmental failsafe mechanisms triggered by deviant voltages, currents, or temperature, automatic computer-based data recording on magnetic media, redundant hard copy output, periodic randomness calibrations, and so on.[18,20]

RNGs are typically designed to produce a sequence of random bits at the press of a button. After generating a sequence of say, 100 random bits (0's or 1's), the number of 1's in the sequence may be provided as feedback. In an experimental protocol using a binary RNG, a run might consist of an observer being asked to cause the RNG to produce, in three successive button presses, a high number (sum of 1's greater than chance expectation of 50), a low number (less than 50), and a control condition with no directional intention. An experiment might consist of a group of individuals each contributing a hundred such runs, or one individual contributing several thousand runs. Results are usually analyzed by comparing high aim and low aim means against a control mean or theoretical chance expectation.

3. META-ANALYTIC PROCEDURES

The quantitative literature review, also called meta-analysis, has become a valuable tool in the behavioral and social sciences.[21] Meta-analysis is analogous to well-established procedures used in the physical sciences to determine parameters and constants. The technique assesses replication of an effect within a body of studies by examining the distribution of effect sizes.[22–24] In the present context, the null hypothesis (no mental influence on the RNG output) specifies an expected mean effect size of zero. A homogeneous distribution of effect sizes with nonzero mean indicates replication of an effect, and the size of the deviation of the mean from its expected value estimates the magnitude of the effect.

Meta-analyses assume that effects being compared are similar across different experiments, that is, that all studies seek to estimate the same population parameters. Thus the scope of a quantitative review must be strictly delimited to ensure appropriate commonality across the different studies that are combined.[21,25] This can present a nontrivial problem in meta-analytic reviews because replication studies typically investigate a number of variables in addition to those studied in the original experiments. In the present case, because different subjects, experimental protocols, and RNGs were employed within the reviewed literature, some heterogeneity attributable to these factors was expected in the obtained distribution of effect sizes. However, the circumscription for the review required that every study in the database have the same primary goal or hypothesis, and hence estimate the same underlying effect.

Experiments selected for review examined the following hypothesis: The statistical output of an electronic RNG is correlated with observer intention in accordance with prespecified instructions, as indicated by the directional shift of distribution parameters (usually the mean) from expected values.

Because this "directional shift" is most often reported as a standard normal deviate (i.e., Z score) in the reviewed experiments, we determined effect size as a Z score normalized by the square root of the sample size (N), $e = Z/\sqrt{N}$, where N was the total number of individual random events (with probability of a hit at $p = 0.5$, $p = 0.25$, etc.). This effect size measure is equivalent to a Pearson product moment correlation.[21]

3.1. Unit of Analysis

To avoid redundant inclusion of data in a meta-analysis, "units of analysis" are often specified. We employed the following method: If an author distinguished among several experiments reported in a single

article with titles such as "pilot test" or "confirmatory test," or provided independent statistical summaries, each of these studies was coded and quality-assessed separately. If an experiment consisted of two or more conditions comparing different intentions or types of RNG devices, the data were split into separate units of analysis to allow the results to be coded unambiguously. In general, within a given reviewed report, the largest possible aggregation of nonoverlapping data collected under a single intentional aim was defined as the unit of analysis (hereafter called an experiment or study).

For each experiment, a Z score was assigned corresponding to whether the observed result matched the direction of intention. Thus, a negative Z obtained under intention to "aim low" was recorded as a positive score. When sufficient data were provided in a report, Z was calculated from those data and compared with the reported results; the new calculation was used if there was a discrepancy. If only probability levels were reported, these were transformed into the corresponding Z score. For experiments reported only as "nonsignificant," a conservative value of $Z = 0$ was assigned; if the outcome was reported only as "statistically significant," $Z = 1.645$ was assigned; and if sample size was not reported or could not be calculated from the information provided, a special code of $N = 1$ was assigned.

3.2. Assessing Quality

Because the hypothesized anomalous effect is not easily accommodated within the prevailing scientific world-view, it is particularly important to assess the trustworthiness of each reviewed experiment. Unfortunately, estimating experimental quality tends to be a subjective task confounded by prior expectations and beliefs.[26,27] Estimates of inter-judge reliability in assessing the quality of research reports, for example, rarely exceed correlations of 0.5.[28] We addressed this problem by assigning to each experiment a single quality weight derived from a set of sixteen binary (present/absent) criteria. The first author coded and double-checked the coding for all studies; the second author independently coded the first 100 studies. Inter-judge reliability for quality criteria was $r = 0.802$ with 98 degrees of freedom.

These criteria were developed from published criticisms about random-number generator experiments[14,15,29–33] and from expert opinion on important methodological considerations when performing studies involving human behavior.[20,34,35] Collectively, these criteria form a measure of credibility by which to judge the reported data. The criteria assess the integrity of the experiment in four categories—procedures,

statistics, the data, and the RNG device—and they cover virtually all methodological criticisms raised to date. They are (1) control tests noted, (2) local controls conducted, (3) global controls conducted, (4) controls established through the experimental protocol, (5) randomness calibrations conducted, (6) failsafe equipment employed, (7) data automatically recorded, (8) redundant data recording employed, (9) data double checked, (10) data permanently archived, (11) targets alternated on successive trials, (12) data selection prevented by protocol or equipment, (13) fixed run lengths specified, (14) formal experiment declared, (15) tamper-resistant RNG employed, and (16) use of unselected subjects.

Each criterion was coded as being present or absent in the report of an experiment, specifically excluding consideration of previously published descriptions of RNG devices or control tests. This strategy was employed to reflect lower confidence in such experiments since, for example, randomness tests conducted once on an RNG do not guarantee acceptable performance in the same RNG in all future experiments. As a result, assessed quality was conservative, that is, lower than the "true" quality for some experiments, especially those reported only as abstracts or conference proceedings. Using unit weights (which have been shown to be robust in such applications[36]) on each of the sixteen descriptors, the quality rating for an individual experiment was simply the sum of the descriptors. Thus, while a quality score near zero indicated a low quality or poorly reported experiment, a score near sixteen reflected a highly credible experiment.

3.3. Assessing Effect Size

Assume that each of K experiments produces effect size estimates e of a parameter E, based on N samples, and that each e has a known standard error s. The weighted mean effect size is calculated as $e_\cdot = \sum \omega_i e_i / \sum \omega_i$, where $\omega_i = 1/s_i^2 = N_i$, and i ranges from 1 to K. The standard error of $e_\cdot$ is $s_e = (\sum \omega_i)^{-1/2}$. A test for homogeneity for the K estimates of e_i is given by $H_K = \sum \omega_i (e_i - e_\cdot)^2$, where H_K has a chi-square distribution with $K-1$ degrees of freedom.[37] The same procedure can be followed to test for homogeneity of effect size across M independent investigators. In this case, $e_{\cdot j}$ and s_{ej} are calculated per investigator, and the test for homogeneity is performed as $H_M = \sum \omega_j (e_{\cdot j} - e_{\cdot M})^2$, where $e_{\cdot j}$ and ω_j are mean weighted effect size and $1/s_e^2$ per investigator, respectively, $e_{\cdot M} = \sum \omega_j e_{\cdot j} / \sum \omega_j$, and j ranges from 1 to M. H_M has $M-1$ degrees of freedom.

For a quality-weighted analysis, we may determine $e_{\cdot Q} = \sum (Q_i \omega_i e_i) / \sum (Q_i \omega_i)$, where Q_i is the quality assessed for experiment i. The standard error associated with e_Q is $se_Q = (\sum (Q_i^2 \omega_i) / (\sum Q_i \omega_i)^2)^{-1/2}$; the test for homogeneity is similar to that described above. Finally, following

the practice of reviewers in the physical sciences,[23,24] we deleted potential "outlier" studies to obtain a homogeneous distribution of effect sizes and to reduce the possibility that the calculated mean effect size may have been spuriously enlarged by extreme values. The procedure used was as follows: If the homogeneity statistic for all studies was significant (at the $p < 0.05$ level), the study that would produce the largest reduction in this statistic was deleted; this was repeated until the homogeneity statistic had become nonsignificant.

4. RESULTS

On-line bibliographic databases for psychology and physics journals were searched, as was a specialized database covering parapsychological articles, technical reports, conference proceedings and manuscripts. Altogether 152 references were found from 1959 to 1987. These reports described 832 studies conducted by 68 different investigators (597 experimental studies and 235 control studies). Fifty-four experimental and 33 control studies reported only as nonsignificant were assigned $Z = 0$. Six experiments and two control studies coded as $(N = 1, Z > 0)$ were eliminated from further meta-analysis because effect size could not be accurately estimated (this required the elimination of one investigator who reported a single study). Figures 1 and 2 show the distributions of Z scores reported for control and experimental studies, respectively.

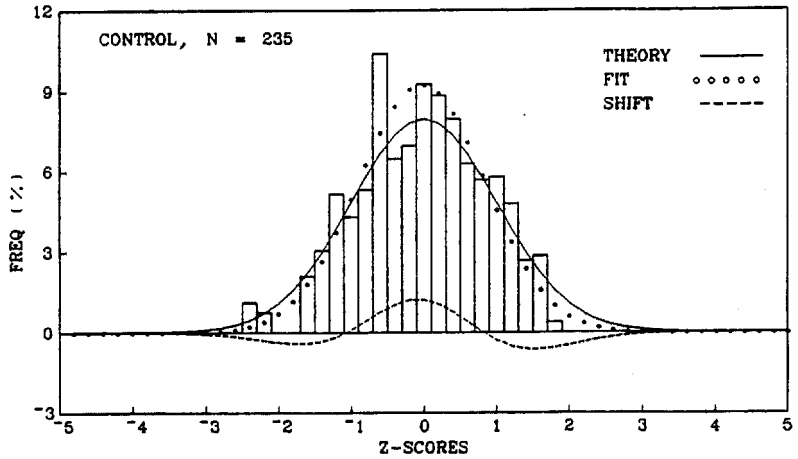

Fig. 1. Distribution of Z scores reported in 235 control studies. Thirty-three of these studies were reported only as "nonsignificant" and were assigned Z scores of zero. To replace the spurious spike at $Z = 0$, those 33 studies were recast as normally distributed Z scores, bounded by ± 1.64, averaging $Z = 0$.

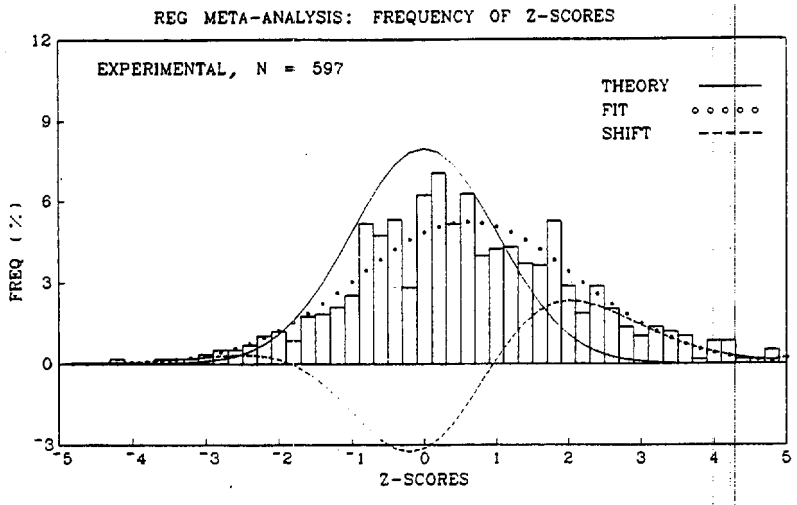

Fig. 2. Distribution of Z scores reported in 597 experimental studies. Fifty-four of these studies were reported as "nonsignificant" and were assigned Z scores of zero. As in Fig. 1, those 54 studies were recast as normally distributed Z scores, bounded by ± 1.64, averaging $Z = 0$.

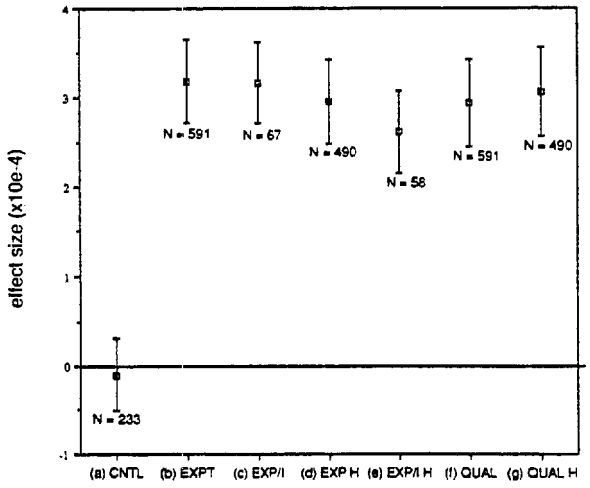

Fig. 3. Mean effect size point estimates ± 1 standard error for (a) control studies and (b) individual experiments; (c) mean effect size per investigator, (d) homogeneous mean effect size for experiments, (e) homogeneous mean effect size per investigator, (f) mean effect size for quality-weighted experiments, and (g) mean effect size for homogeneous quality-weighted experiments.

These results, expressed as overall mean effect sizes, show that control studies conform well to chance expectation (Fig. 3a), and that experimental effects, whether calculated for studies or investigators, deviate significantly from chance expectation (Fig. 3b, 3c). To obtain a homogeneous distribution of effect sizes, it was necessary to delete 17% of individual outlier studies (Fig. 3d) and 13% of mean effect sizes across investigators (Fig. 3e). This may be compared with exemplary physical and social science reviews, where it is sometimes necessary to discard as many as 45% of the studies to achieve a homogeneous effect size distribution.[19] Of individual studies deleted, 77% deviated from the overall mean in the positive direction, and of investigator means deleted, all were positive (i.e., supportive of the experimental hypothesis).

4.1. Effect of Quality

Some critics have postulated that as experimental quality increases in these studies, effect size would decrease, ultimately regressing to the "true" value of zero, i.e., chance results.[12,13,15,32,33,38] We tested this conjecture with two linear regressions of mean effect size vs. mean quality assessed per investigator, one weighted with ω_j as defined above and the other weighted with the number of studies per investigator. The calculated slope for the former is $-2.5 \times 10^{-5} \pm 3.2 \times 10^{-5}$, and for the latter, $-7.6 \times 10^{-4} \pm 3.9 \times 10^{-4}$. These nonsignificant relationships between quality and effect size is typical of meta-analytic findings in other fields,[39,40] suggesting that the present database is not compromised by poor experimental methodology. Another assessment of the effect of quality was obtained by comparing unweighted and quality-weighted effect sizes per experiment (Fig. 3b vs. 3f). These are nearly identical, and the same is true after deleting outliers to obtain a homogeneous quality-weighted distribution (Fig. 3d vs. 3g), confirming that differences in methodological quality are not significant predictors of effect size.

It might be argued that the quality assessment procedure employed here was nonoptimal because some quality criteria are more important than others, so that if appropriate weights were assigned, the quality-weighted effect size might turn out to be quite different. This was tested by Monte Carlo simulation, using sets of 16 weights, one per criterion, randomly selected over the range 0 to 6. A quality-weighted effect size was calculated for the 597 experiments as before, now using the random weights instead of unit weights, and this process was repeated one thousand times, yielding a distribution of possible quality ratings. The average effect size from the simulation was $3.18 \times 10^{-4} \pm 0.15 \times 10^{-4}$, indicating that in this particular database coded by these sixteen criteria,

the probable range of the quality-weighted mean effect size clearly excludes chance expectation of zero.

4.2. The "Filedrawer" Problem

Although accounting for differences in assessed quality does not nullify the effect, it is well known in the behavioral and social sciences that nonsignificant studies are published less often than significant studies (this is called the "filedrawer" problem[21,41–43]). If the number of nonsignificant studies in the filedrawer is large, this reporting bias may seriously inflate the effect size estimated in a meta-analysis. We explored several procedures for estimating the magnitude of this problem and to assess the possibility that the filedrawer problem can sufficiently explain the observed results.

The filedrawer hypothesis implicitly maintains that all or nearly all significant positive results are reported. If positive studies are not balanced by reports of studies having chance and negative outcomes, the empirical Z score distribution should show more than the expected proportion of scores in the positive tail beyond $Z = 1.645$. While no argument can be made that all negative effects are reported, it is interesting to note that the database contains 37 Z scores in the negative tail, where only 30 would be expected by chance. On the other hand, there are 152 scores in the positive tail, about five times as many as expected. The question is whether this excess represents a genuine deviation from the null hypothesis or a defect in reporting or editorial practices.

This question may be addressed by modeling based on the assumption that all significant positive results are reported. A four-parameter fit minimizing the chi-square goodness-of-fit statistic was applied to all observed data with $Z \geqslant 1.645$, using the exponential

$$Y = \frac{1}{\sqrt{2}\,\sigma} e^{-\sqrt{2}\,(|x-\mu|/\sigma)} \tag{1}$$

to simulate the effect of skew or kurtosis in producing the disproportionately long positive tail. This exponential is a probability distribution with the same mean and variance as the normal distribution, but with kurtosis = 3.0.

To begin, the null hypothesis of a (0, 1) normal distribution with no kurtosis was considered. To account for the excess in the positive tail, $N = 585{,}000$ filedrawer studies were required, and the chi-squared statistic remained far too large to indicate a reasonable fit (see Table I). This large N, in comparison with the 597 studies actually reported together with the poor goodness-of-fit statistic, suggests that the assumption of a (0, 1) normal distribution is inappropriate.

Table I. Four-Parameter Fit (E:N, N, Mean, sd) Minimizing Chi-Square (10 df) Goodness-of-Fit Statistic to the Positive Tail of the Observed Z Score Distribution, for Several Exponential:Normal Ratios[a]

Assumption	E:N ratio	N	Mean	sd	Chi-square	p
Normal distribution	0	585,000	0	1	57,867.84	0
(null hypothesis)	1	5,300	0	1	220.97	0
	2	4,800	0	1	167.84	0
	3	4,600	0	1	148.45	0
	10	4,400	0	1	119.69	0
Empirical distribution	0	700	0.145	2.10	23.94	0.008
	1	747	0.345	1.90	16.32	0.091
	2	757	0.445	1.80	14.21	0.164
	3	777	0.445	1.80	11.08	0.226
	10	807	0.445	1.80	11.08	0.351

[a] The null hypothesis is tested by clamping the mean at 0 and the standard deviation at 1, allowing N and E:N to vary. The empirical database is addressed by allowing all four parameters to vary.

Adding simulated kurtosis to a (0, 1) normal distribution by mixing exponential [Eq. (1)] and normal distributions in a 1:1 ratio reduced N by two orders of magnitude, and ratios of 2:1, 3:1, and 10:1 exponential to normal (E:N) yielded further small improvements. However, the chi-squared statistic still indicated a poor fit to the empirical data. Applying the same mixture of exponential and normal distributions, but starting from the observed values of $N = 597$, mean Z score $= 0.645$, and standard deviation $= 1.601$, with the constraint that the mean could only *decrease* from 0.645, resulted in much better fits to the data. Table I shows the results.

This procedure shows that the null hypothesis is unviable, even after allowing a huge filedrawer. The chi-square fit vastly improves with the addition of kurtosis, but only becomes a reasonably good fit when mean and standard deviation are allowed to approximate the empirical values. The filedrawer estimate from this model depends on a number of assumptions (e.g., the true distribution is generally normal, but has a disproportionately large positive tail). It suggests a total number of experimental studies on the order of 800, of which three-fourths have been formally reported.

A somewhat simpler modeling procedure was applied to the data assuming that all studies with significant Z scores in either the positive or negative tail are reported. The model is based on the normal distribution with a standard deviation $= 1$, and estimates the mean and N required to

account for the 152 Z scores in the positive tail and 37 Z scores in the negative tail. This mean-shift model, which ignores the shape of the observed distribution, results in an $N = 1{,}580$ and a mean Z score $= 0.34$.

These modeling efforts suggest that the number of unreported or unretrieved RNG studies falls in the range of 200 to 1,000. A remaining question is, how many filedrawer studies with an average *null* result would be required to reduce the effect to nonsignificance (i.e., $p < 0.05$)? This "failsafe" quantity is 54,000—approximately 90 times the number of studies actually reported. Rosenthal suggests that an effect can be considered robust if the failsafe number is more than five times the observed number of studies.[21]

4.3. Investigators must be faking the data!

5. DISCUSSION

Repeatable experiments are the keystone of experimental science. In practice, repeatability depends upon a host of controllable and uncontrollable ingredients, including factors such as stochastic variation, changes in environmental conditions, difficulties in communicating tacit knowledge employed by successful experimenters,[44] and so on. Difficulties in achieving systematic replication are therefore ubiquitous, from experimental psychology[21,45] to particle physics.[23,24] Of course, this is not to say that systematic replication is impossible in these or other fields, but it may appear to be extraordinarily difficult when experiments are considered individually rather than cumulatively. In the case of the present database, the authors of a recent report issued by the US National Research Council stated that the overall results of the RNG experiments could not be explained by chance,[46] but they questioned the quality and replicability of the research. This meta-analysis shows that effects are not a function of experimental quality, and that the replication rate is as good as that found in exemplary experiments in psychology and physics.

Besides the issue of replicability, five other objections are often raised about the present experiments. These are (a) the effect is inconsistent with prevailing scientific models, (b) the experimental methodology is technically naïve, thus the results are not trustworthy, (c) the experiments are vulnerable to fraud by subjects or by experimenters, (d) skeptics cannot obtain positive results, and (e) there are no adequate theoretical explanations or predictions for the anomalous effect.

These criticisms may be addressed as follows: (a) "Inconsistency with the scientific world-view" is essentially a philosophical argument that carries little weight in the face of repeatable experimental evidence, as suggested by the present and two corroborating meta-analyses.[17,18]

Indeed, if the "inconsistency" argument were sufficient to discount anomalous findings, we would have ignored much of the motivation leading to the development of quantum mechanics. (b) The "naïve methodology" argument was empirically addressed by the assessment of methodological quality in the present analysis. No significant relationship between quality and effect size was found. (c) Fraud postulated as the explanation of the results is untenable as it would have required widespread collusion among 68 independent investigators. In any case, even severe critics of parapsychological experiments have discounted fraud as a viable explanation.[32] (d) Skeptics often assert that only "believers" obtain positive results in such experiments. However, a thorough literature search finds not a single attempted replication of the RNG experiment by a publicly proclaimed skeptic; thus the assertion is not based on verifiable evidence. Furthermore, skeptics who claim to have attempted replications insist (without providing details or references) that they have never achieved positive results in any of their RNG experiments.[15,47] Such a claim is itself quite remarkable, as the likelihood of never obtaining a statistically significant result by chance in series of experiments can be extremely low, depending on the number of experiments conducted. Unfortunately, because we cannot determine how many experiments skeptics have actually conducted, it is impossible to judge the validity of this criticism.

Finally, (e) the "no theoretical basis" argument is correct, but it does not support a negative conclusion about experimental observation. There are at present no adequate theories, with the possible exception of some interpretations of quantum mechanics,[2,3,8,11] that convincingly explain or predict consciousness-related anomalies in random physical systems. We note, however, that the anomalous effects reviewed in this paper apparently can be operationally predicted under well-specified conditions. For example, when individuals are instructed to "aim" for high (or low) numbers in RNG experiments, it is possible to predict with some small degree of confidence that anomalous positive (or negative) shifts of distribution means will be observed.

6. CONCLUSION

In this paper, we have summarized results of all known experiments testing possible interactions between consciousness and the statistical behavior of random-number generators. The overall effect size obtained in experimental conditions cannot be adequately explained by methodological flaws or selective reporting practices. Therefore, after considering all of the

retrievable evidence, published and unpublished, tempered by all legitimate criticisms raised to date, it is difficult to avoid the conclusion that under certain circumstances, consciousness interacts with random physical systems. Whether this effect will ultimately be established as an overlooked methodological artifact, as a novel bioelectrical perturbation of sensitive electronic devices, or as an empirical contribution to the philosophy of mind, remains to be seen.

ACKNOWLEDGMENTS

This study was supported by major grants from the James S. McDonnell Foundation, Inc. and the John E. Fetzer Foundation, Inc. The authors express their gratitude to Dr. York Dobyns of the Princeton University Engineering Anomalies Laboratory for his assistance with the filedrawer models.

REFERENCES

1. R. G. Jahn and B. J. Dunne, *Margins of Reality* (Harcourt Brace Jovanovich, Orlando, Florida, 1987).
2. B. d'Espagnat, "The quantum theory and reality," *Sci. Am.*, pp. 158–181 (November, 1979).
3. O. Costa de Beauregard, "S-matrix, Feynman zigzag and Einstein correlation," *Phys. Lett.* **67A**, 171–173 (1978).
4. N. D. Mermin, "Is the moon there when nobody looks? Reality and the quantum theory," *Phys. Today*, pp. 38–47 (April, 1985).
5. A. Shimony, "Role of the observer in quantum theory," *Am. J. Phys.* **31**, 755 (1963).
6. E. P. Wigner, "The problem of measurement," *Am. J. Phys.* **31**, 6 (1963).
7. U. Ziemelis, "Quantum-mechanical reality, consciousness and creativity," *Can. Res.* **19**, 62–68 (September, 1986).
8. E. J. Squires, "Many views of one world—an interpretation of quantum theory," *Eur. J. Phys.* **8**, 173 (1987).
9. J. Hall, C. Kim, B. McElroy, and A. Shimony, "Wave-packet reduction as a medium of communication," *Found. Phys.* 7, 759–767 (1977); p. 761.
10. R. Smith, unpublished manuscript, MIT, 1968. (Cited in Ref. 9, p. 767.)
11. R. G. Jahn and B. J. Dunne, "On the quantum mechanics of consciousness, with application to anomalous phenomena," *Found. Phys.* **16**, 721–772 (1986).
12. J. E. Alcock, *Parapsychology: Science or Magic?* (Pergamon Press, Elmsford, New York, 1981), pp. 124–125.
13. M. Gardner, *Science: Good, Bad, and Bogus* (Prometheus Books, Buffalo, New York, 1981).
14. R. Hyman, "Parapsychological research: A tutorial review and critical appraisal," *Proc. IEEE* **74**, 823–849 (1986).
15. P. Kurtz, "Is parapsychology a science?" In *Paranormal Borderlands of Science*, K. Frazier, ed. (Prometheus Books, Buffalo, New York, 1981).

16. D. F. Marks, "Investigating the paranormal," *Nature (London)* **320**, 119–124 (1986).
17. C. Honorton, "Replicability, experimenter influence, and parapsychology: An empirical context for the study of mind," paper presented at the annual meeting of the AAAS, Washington, D.C., 1978.
18. E. C. May, B. S. Humphrey, and G. S. Hubbard, "Electronic system perturbation techniques." SRI International Final Report, September 30, 1980.
19. H. Schmidt, "Precognition of a quantum process," *J. Parapsychol.* **33**, 99–108 (1969); "A PK test with electronic equipment," *J. Parapsychol.* **34**, 175–181 (1970); "Mental influence on random events," *New Sci. Sci. J.* **50**, 757–758 (1971); "PK tests with pre-recorded and pre-inspected seed numbers," *J. Parapsychol.* **45**, 87–98 (1981).
20. R. G. Jahn, "The persistent paradox of psychic phenomena: An engineering perspective," *Proc. IEEE* **70**, 136–170 (1982); R. D. Nelson, B. J. Dunne, and R. G. Jahn, "An REG experiment with large data-base capability, III: Operator-related anomalies," Technical Note PEAR 84003, Princeton Engineering Anomalies Research Laboratory, Princeton University, School of Engineering/Applied Science, September 1984; H. Schmidt, R. Morris, and L. Rudolph, "Channeling evidence for a PK effect to independent observers," *J. Parapsychol.* **50**, 1–16 (1986).
21. R. Rosenthal, *Meta-Analytic Procedures for Social Research* (Sage Publications, Beverly Hills, California, 1984); K. Wachter, "Disturbed by meta-analysis?" *Science* **241**, 1407–1408 (1988). We may note that Cohen's *h*, the difference between control and experimental proportions, is a common effect size measure that might have been used in the present study. This was rejected in favor of *e*, as defined, because some of the reviewed studies reported only final *p* values or only overall *Z* scores; *e* was thus deemed more useful in the present meta-analysis.
22. R. L. Bangert-Drowns, "Review of developments in meta-analytic method," *Psychol. Bull.* **99**, 388–399 (1986).
23. A. H. Rosenfeld, "The particle data group: Growth and operations." *Annu. Rev. Nucl. Sci.* **25**, 555–599 (1975).
24. C. G. Wohl *et al.*, *Rev. Mod. Phys.* **56**, Part II, p. S5 (1984).
25. G. V. Glass, "In defense of generalization," *Behav. Brain Sci.* 3, 394–395 (1978).
26. H. M. Cooper, "Scientific guidelines for conducting integrative reviews," *Rev. Educ. Res.* **52**, 291–302 (1982).
27. R. M. Dawes, "You can't systematize human judgment: Dyslexia." In *New Directions for Methodology of Social and Behavioral Science: Fallible Judgment in Behavorial Research*, R. A. Shweder, ed. (Jossey-Bass, San Francisco, 1980), pp. 67–78.
28. S. D. Gottfredson, "Evaluating psychological research reports: Dimensions, reliability, and correlates of quality judgments," *Am. Psychol.* **33**, 920–934 (1978).
29. C. Akers, "Methodological criticisms of parapsychology." In *Advances in Parapsychological Research*, Vol. 4, S. Krippner, ed. (McFarland, Jefferson, North Carolina, 1984); "Can meta-analysis resolve the ESP controversy?" In *A Skeptic's Handbook of Parapsychology*, P. Kurtz, ed. (Prometheus Books, Buffalo, New York, 1985).
30. J. E. Alcock, "Parapsychology: Science of the anomalous or search for the soul," *Behav. Brain Sci.* **10**, 553–565 (1987).
31. P. Diaconis, "Statistical problems in ESP research," *Science* **201**, 131–136 (1978).
32. C. E. M. Hansel, *ESP and Parapsychology: A Critical Reevaluation* (Prometheus Books, Buffalo, New York, 1980).
33. R. Hyman, "The ganzfeld psi experiment: A critical appraisal," *J. Parapsychol.* **49**, 3–50 (1985).
34. T. X. Barber, *Pitfalls in Human Research: Ten Pivotal Points* (Pergamon Press, Elmsford, New York, 1976).

35. J. B. Rhine, "Comments: 'A new case of experimenter unreliability,'" *J. Parapsychol.* **38**, 215–255 (1974).
36. R. M. Dawes, "The robust beauty of improper linear models in decision making," *Am. Psychol.* **34**, 571–582 (1979).
37. L. V. Hedges, "How hard is hard science, how soft is soft science?" *Am. Psychol.* **42**, 443–455 (1987).
38. C. E. M. Hansel, *ESP: A Scientific Evaluation* (Charles Scribner's Sons, New York, 1966), p. 234.
39. R. Rosenthal and D. B. Rubin, "Interpersonal expectancy effects: The first 345 studies," *Behav. Brain Sci.* **3**, 377–415 (1978).
40. G. V. Glass, B. McGaw, and M. L. Smith, *Meta-analysis in Social Research* (Sage Publications, Beverly Hills, California, 1981).
41. Q. McNemar, "At random: Sense and nonsense," *Am. Psychol.* **15**, 295–300 (1960).
42. S. Iyengar and J. B. Greenhouse, "Selection models and the file-drawer problem," Technical Report 394, Department of Statistics, Carnegie-Mellon University (July, 1987).
43. L. V. Hedges, "Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences," *J. Educ. Stat.* **9**, 61–86 (1984).
44. H. H. Collins, *Changing Order: Replication and Induction in Scientific Practice* (Sage Publications, Beverly Hills, California, 1985).
45. S. Epstein, "The stability of behavior, II: Implications for psychological research," *Am. Psychol.* **35**, 790–806 (1980).
46. D. Druckman and J. A. Swets, eds. *Enhancing Human Performance: Issues, Theories, and Techniques* (National Academy Press, Washington, D.C., 1988), p. 207.
47. A. Neher, *The Psychology of Transcendence* (Prentice-Hall, Englewood Cliffs, New Jersey, 1980), p. 147.

Printed by Catherine Press, Ltd., Tempelhof 41, B-8000 Brugge, Belgium