SRI International

Final Report Project 7408
Covering the Period November 1983 to October 1985

December 1988

ENHANCED HUMAN PERFORMANCE INVESTIGATIONS (U)

By: EDWIN C. MAY

Prepared for:

JEAN V. SMITH
CONTRACTING OFFICER

SRI Project 7408

SG1A

WARNING NOTICE
RESTRICTED DISSEMINATION TO THOSE WITH VERIFIED ACCESS TO THE [redacted] PROJECT

Approved by:

MURRAY J. BARON, Director
Geoscience and Engineering Center

Copy 2 of 3 Copies

This document consists 3 of pages
SRI/GF-0284a

CLASSIFIED BY: HQ, USAMRDC (SGRD-ZA)
DECLASSIFY ON: OADR

NOT RELEASABLE TO FOREIGN NATIONALS

SECRET

333 Ravenswood Ave. • Menlo Park, CA 94025
(415) 326-6200 • TWX: 910-373-2046 • Telex: 334 486

I. (U) Objective

(U) The objective of this program was to provide an overview of the current state of psychoenergetics research and, based upon this assessment, to recommend avenues of approach for future investigations.

II. (U) Background

(U) Psychoenergetic research can be divided into two major areas of interest:

(1) Informational Processes

(2) Causal Processes.

Each of these areas can be subdivided further into training, screening, and fundamentals such as various type of functional correlates (e.g., psychological, physiological, and physical).

(S/NF) During FY 1985, SRI International completed a retrospective analysis of a substantial body of open and classified literature in order to assess existence issues, research questions and potential applications of the previously reported activity in these areas. Subsequently, part of this analysis produced two reports that outlined an improved remote viewing analysis technique and provided a meta-analysis of the random number generator literature. (These two reports are included as Appendix A and B, respectively.) What follows are the recommendations, for a three-phase multi-year research effort.

III. (U) Recommendation

A. (U) Phase I—Knowledge Building

(U) Phase I is considered to be a knowledge building effort. During this phase, SRI recommends that some form of technical oversight be included in order to provide guidelines on research protocols, to assess the credibility of the research, and to provide insight into new directions for future research. This phase should be as wide in scope as resources allow. More focused research should be delayed until a knowledge base is established. Table 1 shows the specific areas that are recommended for consideration as research items for Phase I.

Table 1
(U) PHASE I RECOMMENDED RESEARCH AREAS

Topic	Description
Informational Processes	
Analysis	A quantitative remote viewing (RV) analysis technique.
Training	Novice and advanced RV training methodologies.
Screening	Techniques to identify good remote viewers.
Physical Correlates	A search for RV correlates to the physical environment.
Personality Correlates	A search for personality traits in good remote viewers.
Physiological Correlates	A search for physiological correlates to RV.
Medical Correlates	Monitor medical conditions of all viewers.
Feedback	Determine the role of feedback in RV experiments.
Spatial Search	Determine if items can be located in space.
Temporal Search	Determine if events can located in time.
Causal Processes	
Micro-remote Action	Remote action (RA) on random number generators.
Intuitive Data Sorting	Test the Intuitive Data Sorting Model.
Macro-remote Action	Test a variety of physical systems as RA targets.
Correlates	As above, determine correlates to RA.
General	
Information Services	Develop a user-accessible library system.

UNCLASSIFIED

(U) While some of the items shown in Table 1 can be considered beyond existence issues and thus should be considered during Phase II, the predominant effort is toward knowledge building.

B. (U) Phase II—Development

(U) During Phase II, research areas from the Phase I effort that yielded incontrovertible evidence for their existence, will be expanded. With the assistance of a technical oversight committee, hypotheses will formulated and tested.

(S/NF) Those areas under Phase I that showed the most promise, will be expanded toward a potential application area. For example, if a physiological measure could be found that correlated strongly with excellent remote viewing, then that measure could be used to improve intelligence applications.

C. (U) Phase III—Applications

(S/NF) While continuing Phases I and II on specific items of interest, Phase III will be devoted toward applications of interest for the DoD. This activity should include at least two parts:

(1) Applications research--Formulate and test hypotheses that are specific with regard to potential applications.

(2) Application testing--Under actual operational conditions, conduct psychoenergetic activity to assess field utility.

IV. Financial Report

(S/NF) During FY 1985 a total of $1,240 K was allocated to contract DAMD17-83-C-3106 for the psychoenergetic investigation and review. All moneys were expended in accomplishing the stated objective.

APPENDIX A

A FIGURE OF MERIT ANALYSIS FOR FREE-REPONSE

(This Appendix Is Unclassified)

A FIGURE OF MERIT ANALYSIS FOR FREE-RESPONSE MATERIAL

by

E. C. May

B. S. Humphrey

C. Mathews

SRI International, Menlo Park, CA

ABSTRACT: A simplified automated procedure is suggested for the analysis of free-response material. As in earlier similar procedures, the target and response materials are coded as yes/no answers to a set of questions (descriptors). By definition, this coding defines the complete target and response information. The accuracy of the response is defined as the percent of the target material that is correctly described (i.e., number of correct response bits divided by the number of target bits = 1). The reliability of the response is defined as the percent of the response that is correct (i.e. the number of correct response bits divided by the total number of response bits = 1). The figure of merit is the product of the accuracy and reliability. The advantages and weaknesses of the figure of merit are discussed with examples.

INTRODUCTION

With the increased use of computers in parapsychology laboratories, it has become possible to consider more complex methods of analysis to provide deeper insight into the mechanisms of the phenomena. The Engineering Anomalies Research Laboratory, Princeton University, provided a major advancement in the analysis of free-response material (Jahn, Dunne and Jahn, 1980).

THE PRINCETON EVALUATION PROCEDURE (PEP) - A BRIEF REVIEW

In general, the Princeton Evaluation Procedure (PEP) is based on comparing *a priori,* quantitatively-defined target information with similarly quantitatively-defined response information. So defined, the PEP applies various methods of mathematical comparisons to arrive at a meaningful assessment score for remote viewing responses.

Target Information

The definition of a particular target site (usually outdoor sites in and around Princeton, New Jersey) is contained in the yes/no answers to a set of questions called descriptors. These descriptors are designed in such a way as to characterize the typical Princeton target. Each descriptor bit is weighted by its *a priori* probability of occurrence in a large target pool. By definition, the only target information that is to be considered for analysis, is that which is contained completely in the yes/no answers to the descriptor questions (with their associated set of descriptor weights) for the site in question. For example, one descriptor from the Princeton list, "Are any animals, birds, fish, major insects, or figures of these *significant* in the scene?" defines the animal content of the site. The question would be answered "yes" for a zoo and a pet store target, but "no" in all probability for a typical campus building target. Similarly, a set of yes/no responses (30 for the PEP) constitutes the target information.

Response Definition

The descriptor list for the target sites is used as a definition of the response as well. For a given remote viewing session, the remote viewer (or an analyst who is blind to the target site) attempts to answer the 30 questions on the basis of that single response only. In the example above, it would be necessary for a viewer (or analyst) to decide whether or not a particular verbal passage or a quick sketch could be interpreted as depicting animals. For some responses this might be an easy task, e.g. "I get a picture of a cow." Most responses, however, are somewhat ambiguous and require a judgment, e.g. "I see a farm." Nonetheless, the yes/no answers to the 30 questions constitute the only response information that are used in the analysis.

Analysis

For a given response/target combination, the information is contained exclusively in the yes/no answers to the descriptors. Two binary numbers (30 bits long each for PEP) are constructed, one for the target and one for the response descriptor questions, respectively. A "yes" answer is considered a binary "1," while a "no" answer is considered a binary "0." The resulting two, 30-bit binary numbers can then be compared by a variety of mathematical techniques involving use of the weighting factors, to form a score for that specific remote viewing session. For a series of sessions, a quantitative assessment is made by comparing a given response (matched to its corresponding target site) against the scores that are computed by matching the response to all other targets used in the series. This procedure has the added advantage of a built-in, within-group control. In other words, this assessment determines the uniqueness of the target/response match as compared with all other possible matches for the series.

Advantages of the PEP

There are a number of obvious and proven advantages (Dunne, Jahn, and Nelson 1983) of the Princeton Evaluation Procedure:

- *Automation* – Rapid and accurate analysis of a large number of free-response sessions can be accomplished with ease.
- *Archives* – With the aid of computer database management, large numbers of free-response sessions can be organized and maintained in a usable manner.
- *Control* – The cross-target scoring procedure provides a powerful built-in within-group control.
- *Use* – PEP is widely distributed and provides a commonalty of analysis procedure across laboratories.

Disadvantages of the PEP

There are actually very few disadvantages to PEP. A common problem that has been observed before (Dunne, Jahn, 1982) arises in the "granularity" of the descriptor list. With any finite list of binary-type descriptors, it is always possible that a response will appear to be correct with "analogue" analysis procedures but will be evaluated as incorrect with the "digital" approach. Another disadvantage of PEP (also noted above, *op cit*) is that any given descriptor list is likely to be applicable *only* to a given target pool type (i.e., Princeton area natural sites, *National Geographic* magazine photographs, etc.). Lastly, one of PEP's strong points--namely, the cross-match, built-in, within-group control--is also potentially one of its weaknesses.

Since nearly all of the various PEP scoring algorithms involve bit-by-bit weighting, which is based upon relative probability of occurrences, a given response/target score depends not only upon the correctness of the response, but also upon the nature of the remaining targets in the pool. Thus, a score for a given session depends upon the quality of response **and** the target pool. The following hypothetical example illustrates this dependency: a given target has 10 of 30 bits present; furthermore, a few bits (e.g. 3) are particularly rare when compared to the remaining bits (i.e. they possess comparatively large weighting factors). Let us assume that two different viewers provide responses to this target and that each asserts 8 descriptors in the response, 6 of which are correct. If the first viewer's response contains only one of the rare bits, while the other viewer's response contains all three, the second viewer's score will be considerably larger as a consequence of the weighting factors.

Such a scoring discrepancy forces us to define what the **purpose** of the remote viewing session is. If the goal is to demonstrate the existence of psi phenomena, then the PEP is a perfectly adequate system of analysis, and it exhibits all of the advantages described above. If the goal, however, is to demonstrate correlation effects (e.g., correlation of free-response material with personality, physiology, environment, etc.), then the scoring difficulties described above confound the correlation measurement.

To summarize, a target pool dependent scoring procedure provides an important measure of a viewer's ability to *discriminate* from among a number of possible targets. (The

second viewer in the example above, for instance, would receive a higher score because his/her response is more unique to the target pool.) The target pool dependent scoring algorithm is less applicable, however, as an independent absolute measure of target contact—a necessary condition for correlation studies.

If we remove the within-group control to eliminate a source of variance for a correlation measurement that is potentially unrelated to psi ability, we are obligated to provide some other form of control to demonstrate a deviation from mean chance expectation.

FIGURE OF MERIT ANALYSIS

The Figure of Merit analysis (FMA) was developed to address the problems associated with correlation studies and to provide a novel form of control.

Target Information

As in the PEP, the Figure of Merit analysis quantifies the target material into binary numbers corresponding to yes/no answers to a set of descriptors. Our descriptor list was developed on the basis of the target material *(National Geographic* magazine photographs), and on the basis of responses that might be expected *a priori* for our novice remote viewers. Table 1 shows the 20 descriptors that were used for the photon production experiment (Hubbard, May, and Puthoff, 1985). The questions are strongly oriented toward outdoor gestalts, typical of *National Geographic* magazine material. The horizontal lines separating the descriptors into groups of three are provided as an aid for translating binary numbers (derived from the yes/no answers to the questions) into an octal shorthand notation.

A self-consistency check is performed on each coded target, and a set of logically consistent rules must be developed for a given descriptor list. One such example for the list (shown in Table 1) involves bits 13 and 14. While it is possible to have a land/water interface that is not a river, canal, or channel, the reverse (i.e. to have a river, canal, or channel without having a land/water interface) is not possible *by definition*. Thus, if a target analyst asserted bit 14 without asserting bit 13, we could consider this an error in coding and assert bit 13. It is beyond the scope of this paper to provide all the logical consistency rules, but most of them are obvious from Table 1. Naturally, these rules *must* be defined in advance of any experimentation.

Table 1

DESCRIPTOR-BIT DEFINITION

Bit No.	Descriptor
1	Is any significant part of the scene hectic, chaotic, congested, or cluttered?
2	Does a single major object or structure dominate the scene?
3	Is the central focus or predominant ambience of the scene primarily natural rather than artificial or manmade?
4	Do the effects of the weather appear to be a significant part of the scene? (e.g., as in the presence of snow or ice, evidence of erosion, etc.)
5	Is the scene predominantly colorful, characterized by a profusion of color, by a strikingly contrasting combination of colors, or by outstanding, brightly-colored objects (e.g., flowers, stained-glass windows, etc.--not normally blue sky, green grass, or usual building color)?
6	Is a mountain, hill, or cliff, or a range of mountains, hills, or cliffs a significant feature of the scene?
7	Is a volcano a significant part of the scene?
8	Are buildings or other manmade structures a significant part of the scene?
9	Is a city a significant part of the scene?
10	Is a town, village, or isolated settlement or outpost a significant feature of the scene?
11	Are ruins a significant part of the scene?
12	Is a large expanse of water--specifically an ocean, sea, gulf, lake, or bay--a significant aspect of the scene?
13	Is a land/water interface a significant part of the scene?
14	Is a river, canal, or channel a significant part of the scene?
15	Is a waterfall a significant part of the scene?
16	Is a port or harbor a significant part of the scene?
17	Is an island a significant part of the scene?
18	Is a swamp, jungle, marsh, or verdant or heavy foliage a significant part of the scene?
19	Is a flat aspect to the landscape a significant part of the scene?
20	Is a desert a significant part of the scene, or is the scene predominately dry to the point of being arid?

Response Definition

The descriptor list shown in Table 1 is applied in exactly the same way in order to define each remote viewing response. In the SRI program, remote viewers do *not* fill in the descriptor list; rather, this task is performed by an analyst who is blind to the target. However, a set of *a priori* defined guidelines must be established in order to aid the analyst in consistently interpreting the responses.

Analysis

The target-pool independent scoring algorithm makes an assessment of the accuracy and reliability of a single response when matched only against the target material used in the session. As described above, the target and response materials are defined as the yes/no answers to a descriptor list (Table 1). Once the session material is coded into binary, we define session reliability and accuracy as follows:

$$\text{Accuracy} = \frac{\text{number of correct response bits}}{\text{number of target bits} = 1}$$

$$\text{Reliability} = \frac{\text{number of correct response bits}}{\text{number of response bits} = 1}$$

In other words, the accuracy is the fraction of the target material that is correctly perceived, and the reliability is the fraction of the response that is correct.

Neither of these measures, by themselves, is sufficient for a meaningful assessment. For example, in the hypothetical situation in which the viewer simply reads the *Encyclopedia Britannica* as his/her response, it is certain that the accuracy would be 1.0 simply because all possible target descriptors would have been mentioned. This would not be compelling evidence of psi. Similarly, in a response consisting of one correct word, the reliability would be 1.0, with little evidence of psi as well. We define the figure of merit (FM) as:

$$\text{Figure of Merit} = \text{Accuracy} \times \text{Reliability}$$

The figure of merit, which ranges between zero and one, provides an accurate assessment of the response. In the example above where the *Encyclopedia Britannica* is the response, the FM will be low. Although the accuracy is one, the fraction of the response that is correct (i.e. the reliability) will be very small. Likewise, in the example of a single correct word as a response, the reliability is one, but the accuracy is low.

A figure of merit can be calculated for each session. For a series of sessions, the FM may be used to assess a viewer's progress on either a session-by-session or descriptor-by-descriptor basis or both.

ABSOLUTE FIGURE OF MERIT - A METHOD OF CONTROL

We have obtained an estimate of the meaning of FM on an absolute basis. Given the hypothetical situation in which ten viewers contribute 50 sessions each to a remote viewing series, a figure of merit can be calculated by the above technique for each session. If we add the number of responses for all viewers for each of the descriptor bits, we can obtain an estimate as to "response/analysis" bias that may have occurred during the series. For example, if bit number 1 were asserted 40 times in 500 sessions, we can assume on the average for this series (accounting for all known and unknown conditions) that the probability that bit 1 will be asserted in a given response is 40/500 or 0.08. By repeating this calculation for each of the descriptor bits, we can determine the probability of occurrence for all bits under *exactly* the same conditions that were used in the series. Since this procedure displays all response/analysis biases that may have developed during the series, we are able to use this information to construct computer-generated "random" responses, with a total absence of psi functioning, that are subject to *exactly the same* biases that were observed in the series. Therefore, we are able to simulate the ideal control condition, which addresses an important question that is frequently asked by our critics: namely, how would an average viewer respond to a no-target session (i.e. the "monkey on a typewriter" scenario)? A simple bit-by-bit random generation of a response is completely inadequate because it does not account for the response biases observed during the series. The method for producing "random" sessions that *do* account for the biases is described below.

A random number generator is used to create pseudo-responses that are assumed to be devoid of psi functioning. Each bit in a given pseudo-response is generated from the empirical "bias" described above. Once the complete response is generated, the same logical consistency rules (described above) are applied to finalize the pseudo-response. By this technique, a large set of pseudo-responses containing no psi information can be generated. To use these pseudo-responses, we must select, on a random basis, targets from the same set that were used during the series from which the biases were observed. A complete pseudo-session consists of a single pseudo-response and a single randomly selected target. The standard figure of merit analysis is applied to all of the pseudo-sessions in order to calculate figures of merit that have, by definition, no psi content. The resulting FMs are fit with a gaussian distribution to provide an estimate of the mean and standard deviation FM for random data.

Figure 1 shows the results of one such fit for a total of 300 pseudo-sessions, using the remote viewings from a photon-production experiment (Hubbard, May, and Puthoff, 1985) as the bias data. From the chi-square, we note that a gaussian is a correct function to use for the fit. Since the gaussian is truncated at zero figure of merit, we must modify the usual z-score techniques to provide p-values for the individual session figure of merits. By definition, the

probability of observing a figure of merit, f_0, or greater is the area under the FM-gaussian for $f \geq f_0$ divided by the total area under the FM-gaussian. An exact p-value is calculated as follows:

Define the minimum value of a Z-like statistic as

$$Z_{min} = - \frac{\mu}{\sigma} ,$$

where μ and σ are the mean and standard deviation of the best-fit gaussian respectively ($\mu = 0.132$ and $\sigma = 0.163$ in the example). Define a second Z-like statistic as

$$Z_0 = \frac{f_0 - \mu}{\sigma} ,$$

where f_0 is the observed figure of merit. Let P_{min} and P_0 be the p-values calculated in the usual way assuming Z_{min} and Z_0 were valid z-scores. Then, the correct p-value is given by

$$\text{p-value} \leq \frac{P_0}{P_{min}} .$$

Utts and May (1985) have provided an exact method for combining p-values to enable an overall series evaluation. For mean p-values calculated for a series greater than .1, and the number of sessions greater than 6, a close approximation for the combined Z-score is given by (Edgington, 1972)

$$Z_{combined} = (0.50 - \overline{p}) \times \sqrt{12\,N} ,$$

where $\overline{p}$ is the average p-value for N sessions.

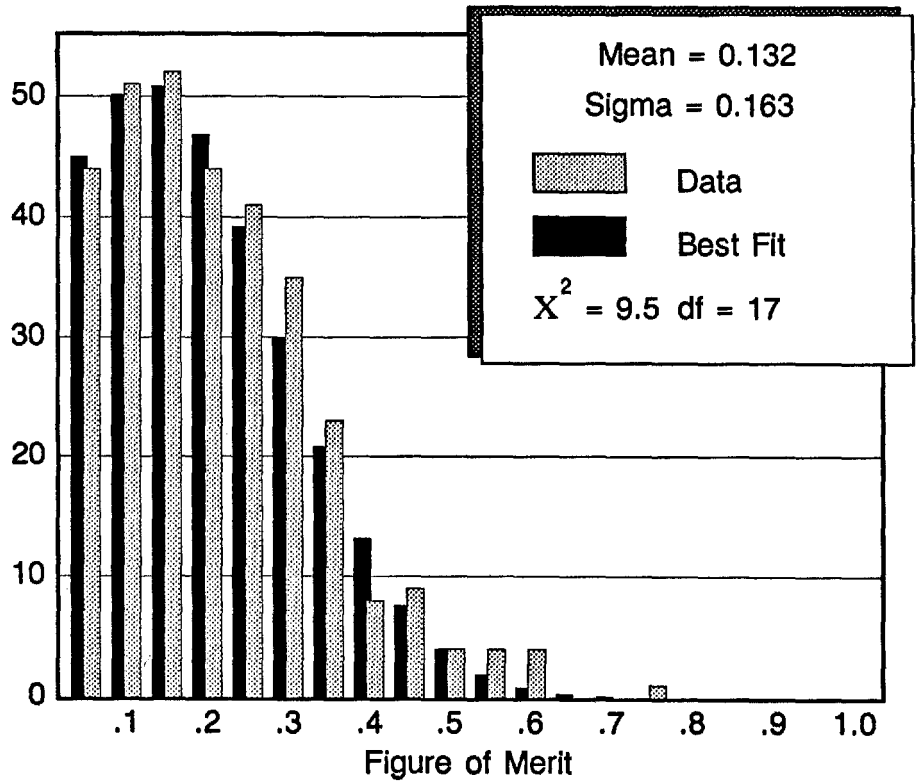


Figure 1 BEST-FIT GAUSSIAN TO CONTROL FMs

CONCLUSIONS AND SUGGESTIONS FOR EXTENSIONS

We are proposing a target-pool independent method (figure of merit analysis) for scoring free-response material. The FMA provides a number of advantages over previous methods.

- Figures of merit can be used in correlation studies.
- FMA provides a novel technique for free-response controls.
- Target pool independent exact p-values can be computed for each free-response session.
- Since the FM is computed by simple counting, the computer coding burden is sharply reduced.

Because of the lack of descriptor bit independence (and thus a need for logically consistent rules) the effective number of descriptor bits is reduced. We are presently investigating a way to utilize a hierarchical descriptor list: that is, each level of the hierarchy consists of a variable number of *independent* descriptors. Finally, the ideal descriptor list would include *arbitrary* weighting factors for the level of hierarchy as well as for the individual descriptors within the level.

REFERENCES

Dunne, B.J., Jahn, R.G., and Nelson, R.D., "Precognitive Remote Perception," Engineering Anomalies Research Laboratory, School of Engineering/Applied Science, Princeton University, Princeton NJ, Technical Note PEAR 83003 (August 1983)

Hubbard, G.S., May, E.C., and Puthoff, H.E., "Possible Production of Photons During a Remote Viewing Task: Preliminary Results," Proceedings of the 28th Convention of the Parapsychological Association, (Radin, D.I, ed.) Tufts University, Medford, MA, (August 1985)

Jahn, R.G., Dunne, B.J., and Jahn, E.G., "Analytical Judging Procedure for Remote Perception Experiments," *The Journal of Parapsychology,* Vol. 44, No. 3, pp. 207-231 (September 1980).

Utts, J.M., and May, E.C., "An Exact Method for Combining P-values," Proceedings of the 28th Convention of the Parapsychological Association, (Radin, D.I, ed.) Tufts University, Medford, MA, (August 1985)

APPENDIX B

PSI EXPERIMENTS WITH RANDOM NUMBER GENERATORS; META-ANALYSIS PART 1

(This Appendix Is Unclassified)

Psi Experiments with Random Number Generators: Meta-Analysis Part 1

Dean I. Radin
Edwin C. May
Martha J. Thomson

SRI International
Menlo Park, California

ABSTRACT: A meta-analysis of 332 psi experiments involving binary random number generators is described. The combined binomial probability for data reported in 56 references published from 1969-1984 is $p \approx 10^{-43}$. A "filedrawer" analysis reveals that over 4500 additional, nonsignificant, unpublished or unretrieved studies would be required to bring the overall result down to a nonsignificant level. Using a novel approach, we estimate the actual size of the "filedrawer" to be 95 studies. Adding the equivalent of 95 nonsignificant studies to the existing data results in $p \approx 10^{-18}$, while a meta-analysis of 98 reported control studies results in $p \approx .78$. An analysis of variance indicates that experimenters' mean z scores are significantly different from each other. We discuss an approach and propose criteria for performing a quality-weighted analysis on the existing data. We conclude that the *prima facie* evidence supports the notion that observers' intentions can affect the statistical properties of truly random number generators.

INTRODUCTION

This is Part 1 of a two part meta-analysis of psi experiments involving truly random number generators (RNG) published from 1969-1984. This part describes the results of a "first-pass" analysis, in which the published data was taken at face value. Part 2 will report on a quality-weighted analysis in which the results of each experiment (in terms of z score) will be evaluated on each of a dozen criteria to produce an adjusted z score reflecting that experiment's overall quality.

Background: On the scent of a trail

When Albert Einstein was asked about his way of thinking, he reportedly replied, "All I have is the stubbornness of a mule; no, that's not quite all, I also have a nose" (Bower, 1985, p.330). What he meant was that he was not only extraordinarily obstinate in tracking down solutions to problems, he was also able to sniff out when he was on the right track. The centennial anniversary

of the American Society for Psychical Research, celebrated this year (1985), clearly demonstrates that parapsychologists have displayed Einstein's stubbornness over the years. One question we might ask after 100 years, however, is whether the parapsychological nose has been sniffing along a clearly defined trail, and if so, is the trail likely to grow more fragrant or more noxious as we progress?

There is evidence that the nose has not been shirking its duty. This can be seen in the single most predictable feature found in the parapsychological literature, that is, the perennial call for a replicable experiment. The ideal experiment is supposed to produce a significant result regardless of the phase of the moon, the price of pork bellies, and the experimenter's shoe size. This Quest for replicable experiments is by no means unique to parapsychology, however. Social and behavioral scientists in general have been acutely aware of the slow progress in the "softer" sciences as compared to the natural sciences such as physics, chemistry, and biology. In experimental psychology, for example, Epstein (1980) has stated,

> Psychological research is rapidly approaching a crisis as the result of extremely inefficient procedures for establishing replicable generalizations. The traditional solution of attempting to obtain a high degree of control in the laboratory is often ineffective because much human behavior is so sensitive to incidental sources of stimulation that adequate control cannot be achieved.... Not only are experimental findings often difficult to replicate when there are the slightest alterations in conditions, but even attempts at exact replication frequently fail. (p. 790)

Many observers of parapsychology (both within and outside the field) claim that the repeatable parapsychological experiment does not exist. For example, Beloff (1977) has written, "There is still no repeatable [psi] experiment on the basis of which any competent investigator can verify a given phenomenon for himself" (p.759). Critics of the field have pointed to the lack of replicability as perhaps the single most serious problem in parapsychology (e.g. Kurtz, 1981, p.12). In response, proponents often point to significant psi studies involving ESP card-guessing (Honorton, 1975), ganzfeld stimulation (Honorton, 1978), remote perception (Dunne, Jahn, and Nelson, 1983), and RNGs (May, Hubbard and Humphrey, 1980) to indicate that there are some significant replications.

The problem is that from different perspectives the proponents and critics are both right. There are indeed many psi experiments that have been repeated, but whether they are considered *robust, successful replications* is the crux of the debate. One of the primary reasons for this debate, in our opinion, is because the traditional approach of assessing the results of a set of related studies is by descriptive literature review. Within parapsychology there are many excellent examples of such reviews (e.g. Carpenter, 1977; Palmer, 1982; Rush, 1982; Schmeidler, 1984; Stanford, 1977; Stanford, 1984). Unfortunately, what one has typically learned after studying such a review is a hodge-podge of variables, conditions, and p-values. Rarely is one left with a quantitative statement of the degree of significance obtained in the studies as a whole.

Addressing this issue empirically, Cooper and Rosenthal (1980) demonstrated that when knowledgeable individuals are instructed to make judgments about the overall significance of a set of studies based on their reading of a comprehensive, descriptive literature review, it is possible for

them to draw conclusions that are *completely the opposite* of the results obtained when the same studies are summarized by more explicit, quantitative methods.

Given the difficulties in assessing evidence from existing psi studies, is the replication trail likely to be heading -- to reinvoke our metaphor -- towards a flowering meadow or something decidedly less pleasant? In general, we believe that the prospects are aromatic. In the last few years, quantitative techniques of combining and comparing research results in systematic ways have been developed -- called *meta-analysis* (Rosenthal, 1984) -- that show great promise in demonstrating that some areas of social science have been progressing much better than previously thought. In parapsychology, initial meta-analyses applied to ganzfeld research (Honorton, 1985), hypnotic induction (Schechter, 1984), RNG studies (May, Hubbard and Humphrey, 1980; Nelson, Dunne and Jahn, 1984; Tart, 1983), and remote viewing (Dunne, Jahn, and Nelson, 1983) have shown that the overall evidence for these psi phenomena is actually quite strong.

Because meta-analysis involves the aggregation of results of numerous studies, several criticisms of this technique have been raised (Rosenthal, 1984, p.124-132). Perhaps the three categories of criticism most pertinant to review of parapsychological data are the following: First, authors may tend to report only the studies with significant results and leave the nonsignificant studies unpublished (called the *filedrawer* problem). Second, the meta-analysis combines poorer quality studies with better studies. And third, meta-analysis may be comparing "apples and oranges" by combining different experiments studying different variables.

The first two problems may inflate the estimate of an overall effect; the third criticism may make the overall summary difficult or impossible to interpret. In the present meta-analysis, however, we actually are interested in whether these psi experiments have borne *fruit*, not whether they have borne specific flavors of apples or oranges. In other words, we are not concerned with whether hypnotic induction, say, has an effect on RNG outputs, but whether there is evidence for *any psi effect* on RNG outputs. Thus, in this investigation we have concentrated on the filedrawer issue (in this report) and the quality of studies (to be described in Part 2 of this study).

OVERVIEW OF A TYPICAL RNG EXPERIMENT

The typical psi experiment with RNGs involves three main components: An observer (e.g. a human, goldfish, cat or dog), a truly random number generator based on radioactive decay or electronic noise, and an experimental task linking the observer with the device, such as a video game, a set of instructions, a need to keep a heat lamp on or avoid a shock, and so on. The aim of these experiments is to show that the instructions (when humans are involved) or the induced need (when animals or plants are involved) are associated in some way -- but not necessarily causally -- to the statistical output of the RNG.

For example, say an RNG was designed to produce 100 random bits at the press of a button. An individual in this experiment might see a digital display of the number of 1's (called hits) produced immediately after he or she pressed a button. The instructions in the experiment would typically be to get as many hits as possible for each button press. The results of many presses, or trials, would

then be evaluated statistically, where under the null hypothesis an average of 50 hits would be expected by chance. If the average number of hits over thousands of repetitions were say, 52, this deviation from chance would be interpreted as evidence of a psi effect (provided that the probability of observing this deviation was less than 1 in 20).

PROCEDURE

Because we were ultimately interested in testing among several different models of mechanisms possibly operating in these RNG experiments, in Part 1 of this meta-analysis (this paper) we surveyed the parapsychological literature with two goals in mind: First, we wanted to see whether the aggregated result of the RNG experiments showed evidence for an anomalous effect. And second, we needed the details of these experiments for use in evaluating a model of the underlying mechanism. [Our modeling effort is discussed in May, Radin, Hubbard, Humphrey and Utts (1985).]

Source of references

We searched through the five major English language parapsychological journals[1] over the years 1969 to 1984. We also included the (refereed) *Proceedings of Presented Papers* for the Annual Parapsychological Association Conventions (1971 and 1984), and a report published by the Princeton Engineering Anomalies Research Laboratory (Nelson, Dunne and Jahn, 1984). The literature search was started in the year 1969 because that was the year Helmut Schmidt (1969) published the seminal RNG study that has since spawned many replications.

Defining "an experiment"

One of the difficulties faced in reviewing the articles for this meta-analysis was to decide what constituted an experiment. In most papers, authors analyze their data repeatedly in various ways, sometimes as *a priori* analyses, sometimes as *post hoc* afterthoughts. Even in cases of planned analyses, there are many ways of interpreting which of several conditions is the "real" experiment. How we decide what is an experiment is important to the meta-analysis for two main reasons: First, the meta-analytic statistical power depends on the number of experiments we find; and second, the z scores are different depending on how we break down the reported results.

To illustrate the difficulty of deciding what an experiment is, consider this example. Say an author uses three different groups of 10 percipients each (e.g. meditators, truck drivers and athletes) and subjects each group to two different conditions (e.g. mental imagery vs. muscular tension) in a study on psi-conducive states. The results can be broken into one big, combined experiment, six experiments (3 groups x 2 conditions), two experiments (2 conditions), three

1. These are the *Journal of Parapsychology, European Journal of Parapsychology, Research in Parapsychology, Journal of the Society for Psychical Research, Journal of the American Society for Psychical Research.*

experiments (3 groups), 30 experiments (subject by subject analysis), and so on. How do we decide what to use?

We resolved this issue for this first-pass analysis in the following way: For cases where there were multiple hypotheses under test and multiple analyses of the data, we chose as the experimental unit the largest possible accumulation of data compatible with a single "direction of effort" assigned to the subjects. A clearly defined direction of effort meant that the experimental protocol required either more 1's or more 0's from the RNG to successfully complete the assigned task, regardless of whether or not the subjects actually knew their task in detail.

Say, for example, a hypothesis predicted that group A would score higher than group B, and it was stated that "higher" meant more 1 bits. Then we would take this study as two experiments: Group A's and group B's scores. In this particular case, since group A was *predicted* to score higher than group B, if in fact the difference between z(A) and z(B) were significant, then both z scores would be taken as *positive*, regardless of the reported z's. Thus if z(A) = 1.5, z(B) = −1.0, then the z-score difference between them would be significant one-tailed with $z_{diff} = 1.77$. If the number of trials run in each case were 10000, then the number of hits assigned per experiment would be hits(A)= 5075 and hits(B) = 5050, which are both positive deviations; similarly the z scores would be recorded as z(A) = 1.5, z(B) = 1.0. If z(A) = 1.2 and z(B) = −1.0, the z scores would be recorded as originally reported since z_{diff} is not significant. The same would be true if z(A) = −2.0 and z(B) = 2.0. (Fortunately, such problems of interpretation were not often encountered in the survey.)

As another example, if groups A, B, and C all tried to influence an RNG in a particular way, and no predictions were made as to interactions, then their overall result would be combined as one experiment. In this way, we attempted to emphasize in the meta-analysis the underlying question of whether or not observers could influence or otherwise affect the statistical output of an RNG *according to the stated intention of the experimenter.*

Results of literature review

We found 73 pertinent references in the journals and reports.[2] These references included 381 experiments contributed by 38 different principal investigators, representing about 10 different laboratories around the world. We say "about 10 laboratories" because over the years labs have come and gone, researchers have moved among different labs, and in many cases, one or two individuals at an academic or private research institution are considered a "laboratory."

Breakdown of experiments Of the 381 experiments found, 332 (in 56 remaining references) were described as using binary generators based on either radioactive decay or electronic noise. For this meta-analysis, we considered only studies using binary RNGs (or any study in which the hit rate was defined or could be interpreted as 50%) for three reasons: First, since 87% of the experiments (332/381) employed binary generators, we felt that this sample was representative of the entire RNG database; second, for the sake of simplicity; and, third because the test of a model

2. These references are listed under the heading "Meta-Analysis" in the references at the end of this paper.

we developed (May, Radin *et al*, 1985) requires binary statistics. In addition, to avoid the possibility that reported p values or z scores were rounded up, whenever possible we recorded the reported number of *trials* (bits generated in an experiment) and *hits* (number of times the designated bit was obtained) in these experiments.

Of these 332 binary experiments, 188 were reported in journals and conference proceedings, of which 58 were reported significant at $p < .05$, 2-tailed, (against 9.4 expected by chance). The probability of observing 58 significant studies out of 188 is less than 10^{-57}. We refer to this body of data as the "survey." The remaining 144 experiments were obtained from the Princeton Engineering Anomalies Research Laboratory (Dunne, Jahn and Nelson, 1982). Of these experiments, 13 were significant 2-tailed, resulting in $p < .04$ (corrected for continuity). We refer to these experiments as the "Princeton" data.

Experiments with incomplete descriptions Of the 188 survey studies, 30 were simulated by Monte Carlo techniques because the experiment was reported as nonsignificant but neither the z score nor the number of trials and hits were provided. To perform the simulation, we had a pseudorandom generator (cf. May, Humphrey and Hubbard, 1980) choose a z score at random from a normal distribution [N(0,1)], but bounded between 10^{-25} to 1.64 and -10^{-25} to -1.64.[3] In five additional studies, the results were reported as significant and p or z values were provided, but the number of trials or hits were not given. For these five studies, since the z score was known or could be calculated from a p value, the trials or hits (whichever was missing) were calculated.

Table 1 shows a breakdown of the number of experiments reported in each of the seven reference sources we used. It is clear that the reports provided in the *Research in Parapsychology* series are not as detailed as one might have wished, but it is not surprising since the contents of this reference are only abstracts of the full papers presented at the annual Parapsychological Association conventions.

Table 1. Experiment breakdown by source of reference.

Reference	Experiments with full detail	Experiments with partial detail
Journal of the American Society for Psychical Research	5	
European Journal of Parapsychology	6	
Journal of the Society for Psychical Research	9	
Proceedings of the Parapsychological Association*	32	
Journal of Parapsychology	49	1
Research in Parapsychology	52	34
Princeton Engineering Anomalies Research Lab	144	

* for the years 1971 and 1984

3. We did not generate z scores of zero because this data was ultimately used in an evaluation of our model (May, Radin et al, 1985), in which log(z) is taken.

In summary, of the 332 experiments we considered (188 from the survey and 144 from Princeton), 71 were reported significant at $p < .05$, 2-tailed, for an overall binomial probability of $p < 5.4 \times 10^{-43}$.

ADDRESSING CRITICISMS OF THE DATA

Taken as *prima facie* evidence, one might think that this body of published data provides indisputable evidence that an anomaly exists. But there are numerous reasons why the data may be suspect. The main criticisms (Akers, 1984; Hansel, 1980; Hyman, 1985; Kurtz, 1981) include

1. Results are due to chance
2. Basic statistical assumptions are violated
3. Only significant studies are published
4. Experiments are not replicable
5. RNGs are nonrandom
6. Poorer studies are included with better studies

Let us consider each of these six steps as successive filters for the reliability of the data. If each criticism can be satisfactorially refuted or countered, then a persuasive case for an anomalous effect can be made.

1. Results are due to chance

In any one experiment we cannot establish the reality of a phenomenon, regardless of the significance level, unless strong theoretical predictions have preceeded the experiments. For example, the recent experiments suggesting that Bell's inequality is violated (e.g. Aspect, Dalibard, and Roger, 1982; Aspect, Grangier and Roger, 1982) have been widely accepted within the physics community on the basis of only a few empirical studies despite its profound implications on our view of the nature of reality (cf. d'Espagnat, 1979; Mermin, 1985; Rohrlich, 1983). Parapsychology, however, has had the disadvantage of not having a firm theoretical base on which to stand. Thus the nature of the claim (any claimed psi effect) understandably requires extremely persuasive evidence.

One wonders how statistically strong an effect must be to bring about a consensual agreement within the scientific community that a psi effect on RNGs is real. Would $p < 10^{-43}$ be sufficient? If this figure were revised to take into account all of the criticisms noted above, and the end result were say, 10^{-5}, would that be sufficient? Clearly an overall $p = .1$ would not satisfy anyone, so there is a decision curve related to this question. This curve is probably different according to individual prejudices and predilictions, but the resolution of this question is beyond the scope of the present paper. Note that if an anomaly did exist, it would not *necessarily* imply that psi was the mediating factor. Such an anomaly may, for example, reveal some heretofore unknown statistical peculiarities about random numbers.

2. *Basic statistical assumptions are violated*

This criticism incorporates such problems as the improper application of statistics to a particular experimental design, violation of assumptions of independence, performing multiple analyses on the same data, and so on. In this meta-analysis, one of the reasons we only considered binary generators was to simplify the statistical assumptions to the point where we could avoid many such problems. Another reason was to avoid the "apples vs. oranges" comparison problem we mentioned earlier. Because we were interested only in RNG experiments that reported (or where we could calculate) the number of hits and trials, we were in fact comparing apples only with apples (actually bits with bits). While it is true that there were many different psychological and physiological conditions involved in these experiments, as well as human and non-human subjects, the underlying question we asked was the same for each experiment: What was the behavior of the RNG as compared to the pre-specified direction of effort defined in the experimental task?

The statistics in these RNG experiments are described by the well understood binomial distribution, and the central limit theorem allows us to use the normal approximation to further simplify the statistical treatment for the range of trials observed in the data (200 to 2 million trials in a single experiment).

Violation of the assumption of independence can be the downfall of an otherwise tightly controlled experiment. In the present case, however, the random events are based on sources that are quantum-mechanical (QM) in nature -- radioactive decay of alpha, beta, or gamma particles, or electronic noise from various semiconductor devices such as tunnel diodes. QM theory states that random numbers based on QM events are *in principle* indeterminant and therefore independent of each other, *provided that the RNG device is properly designed and constructed.*[4]

In this meta-analysis, under the null hypothesis of no psi effect we can assume independence of random bits. Note that the assumption of independence among bits does not override proper concern about whether the RNGs used in the experiments produced bits with equal probabilities. It is entirely possible, for example, to produce bits that are completely independent, but with $p(1) = .6$ and $p(0) = .4$. This is addressed in point 5 below.

3. *Only significant studies are published -- the Filedrawer problem*

The filedrawer problem, in which only significant studies are reported and the nonsignificant studies languish in filedrawers, will inflate the results of a meta-analysis because there will be too many small p values (or equivalently, to many large z scores). To address this problem, we followed a procedure proposed by Rosenthal (1984, p. 108), in which the average z score for all combined studies is applied to the formula:

4. Note that some of the diodes used in noise-based RNGs are not QM in nature. RNGs that use avalanche diodes, for example, derive their noise from fluctuations in charge carrier multiplication, which can be described by classical electromagnetic theory.

$$X = \frac{K[K\bar{Z}^2 - 2.72]}{2.72} \qquad (1)$$

where K is the number of studies combined, Z is the mean Z obtained for the K studies, and X is the number of *new, filed, or unretrieved studies averaging null results* required to bring the new overall p level to a designated level. The value 2.72 in equation (1) is the square of 1.65, the z value for p = .05 (the p level that Rosenthal uses). To make our filedrawer estimate more conservative, we chose a 2-tailed p = .05, z = 1.96. Thus the formula we used was

$$X = \frac{K[K\bar{Z}^2 - 3.92]}{3.92} \qquad (2)$$

We shall consider the Princeton studies separately from the rest of the survey because we have good reason to believe that all of the Princeton data was, in fact, published, thus their data has no filedrawer problem. [(Publishing all data is a part of the Princeton Laboratory's philosophy (Jahn, 1982)].

In the 188 survey experiments, the mean z score = 0.738. A mean z of this value over 188 experiments produces an overall z = 10.114, for a 2-tailed $p < 4.9 \times 10^{-24}$ (see Table 1). Note that this method of estimating the overall probability is more accurate than determining the binomial probability of 71 successes out of 188 samples at p < .05, as described earlier in this paper. Applying Z = .738 and K = 188 to formula (2) results in X = 4723. This means that 4723 additional studies averaging null results would have to be filed away in researchers' filedrawers to bring the overall z score down to a 2-tailed nonsignificant level.

According to Rosenthal (1984), the number X has different meanings depending on the research context. In some areas of research (say genetic engineering), perhaps 10 or 12 unpublished or unretrieved studies might be considered reasonable. In other areas (say child development), perhaps 200 to 500 filedrawer studies might be a reasonable estimate. Rosenthal (1984, p.110) proposes the following general guideline: "Perhaps we could regard as robust to the file drawer problem any combined results for which the tolerance level (X) reaches 5 K + 10."

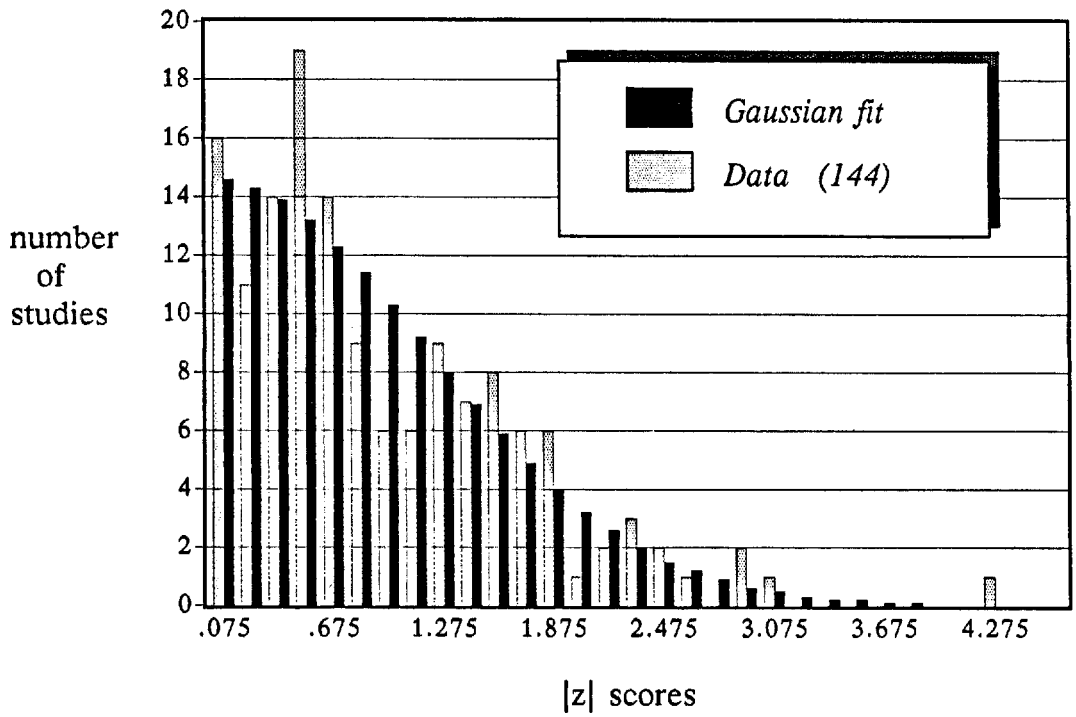
Thus -- not counting the Princeton data -- since X is more than 25 times larger than the observed number of studies, we could state, based on Rosenthal's guideline, that the observed effect *is robust*. Indeed, for this many unpublished or unretrieved studies to exist it would have required each of 10 parapsychology laboratories to have continuously produced nonsignificant studies at the rate of 2.6 *per month* over the 15 years surveyed. This is an unlikely scenario given the limited number of researchers performing these experiments over the years and the time and effort typically required to perform a single study.

If we apply the same procedure to the Princeton data of 144 studies, we find the mean z = .339, overall z = 4.063, and $p < 4.85 \times 10^{-5}$. Plugging these values into formula 2, we find we would need X = 476 additional unpublished or unretrieved studies averaging null results. But as previously

mentioned (*vide supra*), the Princeton lab has claimed that they have no unpublished or filed studies, thus this estimate of the filedrawer size is purely academic.

Another way of looking at the Princeton data is shown in Figure 1. This shows a histogram of the absolute value of the observed z scores in light-colored bars, and a best Gaussian fit in dark bars. As is apparent from the figure, the observed z scores are a good Gaussian fit, but the standard deviation of the fit is not 1.0, as one would expect under the null hypothesis of z scores chosen at random from a normal distribution, but rather the best fit Gaussian standard deviation is 1.17. A variance test between these two variances results in $z = 2.90$, $p < .004$ (2-tailed). Thus the distribution of z scores is significantly altered from that expected by chance. This interesting effect is discussed in more detail by Jahn, Nelson and Dunne (1985) and May, Radin, Hubbard, Humphrey, and Utts (1985).

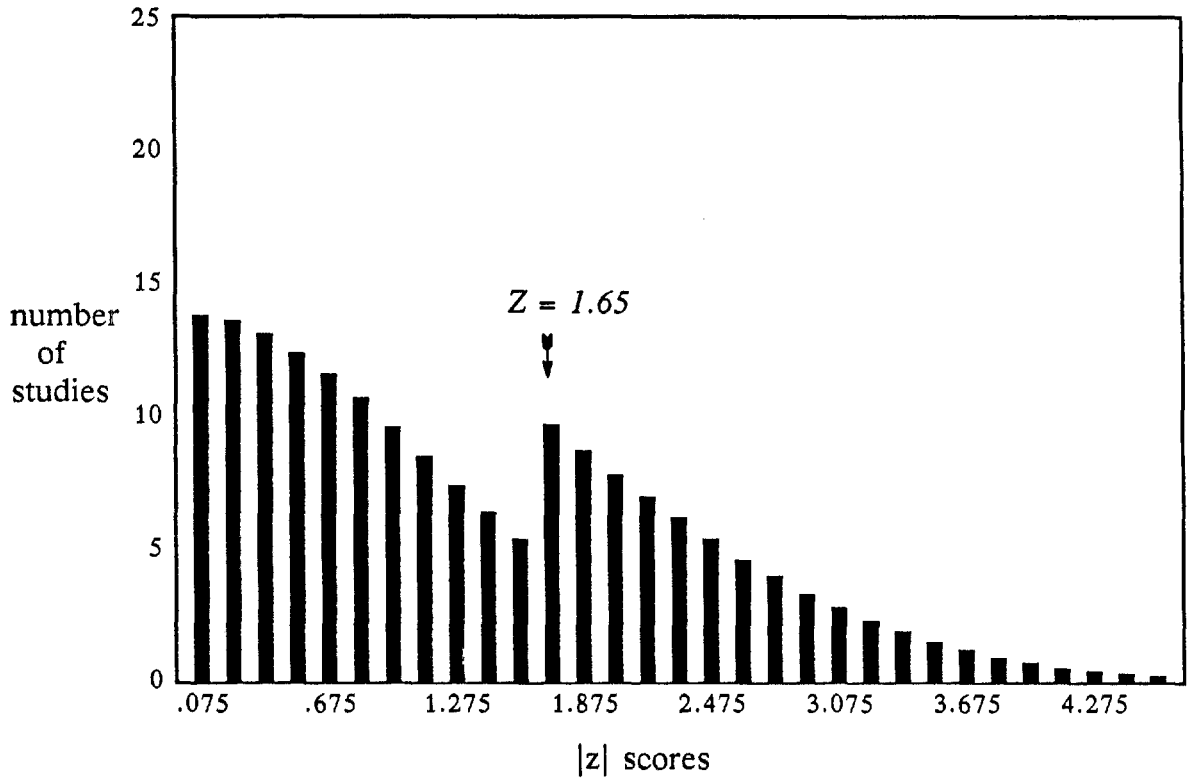
Figure 1. |z| score distribution for Princeton data



Estimating the actual number of filedrawer studies What if we wished to make an estimate of the *actual* size of the filedrawer for the rest of the survey data? We would not be surprised to learn, for example, that there are indeed some unpublished or unretrieved nonsignificant studies we may have missed in our survey. To do this, we postulated what a z-score distribution might look like if there were a filedrawer problem. Figure 2 (next page) shows a histogram of the absolute value of

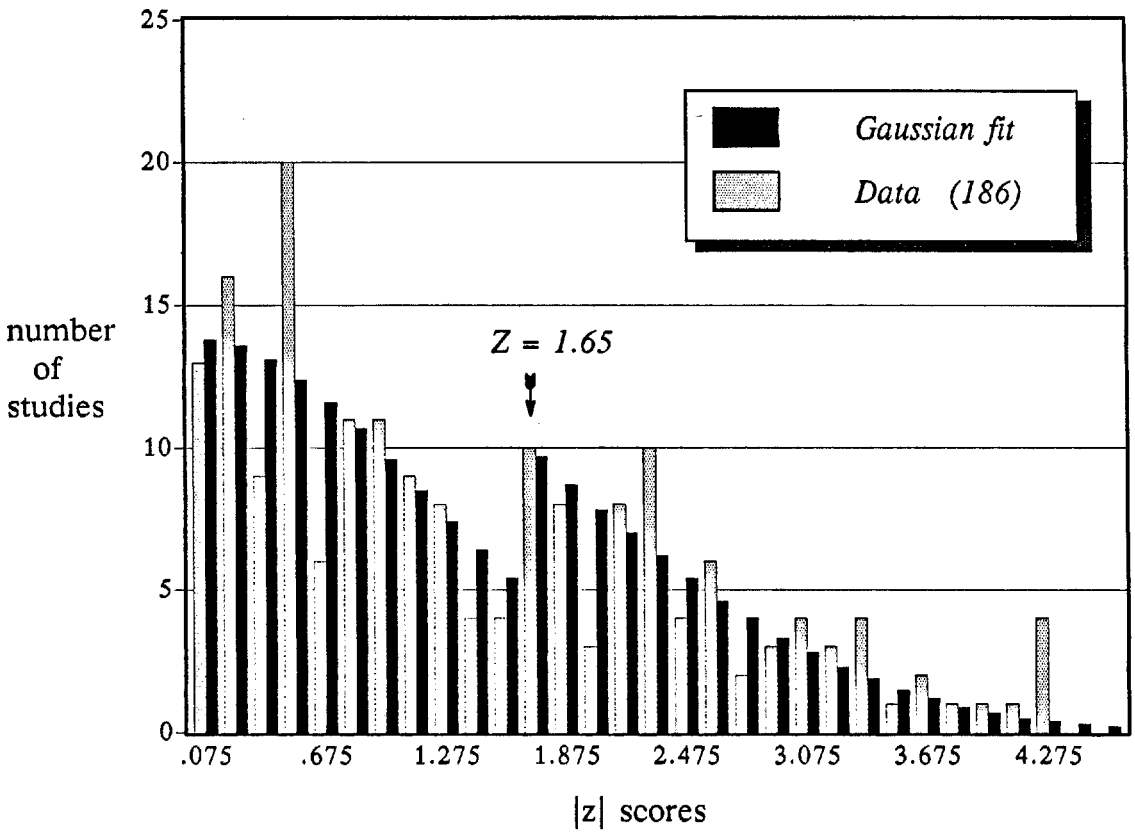
hypothetical z scores with a filedrawer problem. Notice the discontinuity at the magic number z = 1.65 ($p < .05$), which is what one would expect if nonsignificant studies remained unpublished.

Figure 2. |z| score distribution with filedrawer problem



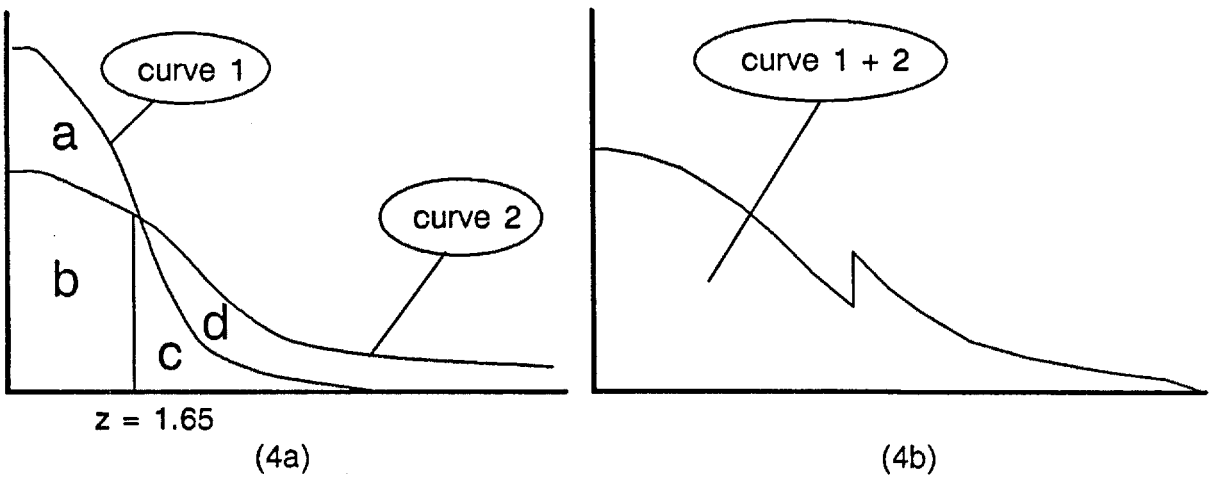
In Figure 3, we plot a histogram of z scores from the 188 survey studies. We also plot a double Gaussian curve, assuming that the observed z-score curve is actually the *sum* of two Gaussians. The resulting two-Gaussian curve is a good fit to the data; in fact, the sum of two Gaussians is a significantly better fit than a single Gaussian curve ($z_{diff} = 1.718$, $p < .04$, 1-tailed, determined by transforming chi-square goodness-of-fit values for one vs. two Gaussian fits into z scores, and comparing those two z scores.)

Figure 3. |z| score distribution for 186 survey experiments



In Figure 4, we show how we estimated the actual number of filedrawer studies. We have assumed that the observed curve (Figure 3 above) is the sum of two Gaussians (Figure 4b), shown as two separate curves (1 and 2) in Figure 4a. We obtained estimates of the amplitude and variance of these curves by allowing a computer-based curve-fitting routine the freedom to vary the amplitude and variance of each curve so the obtained fit to the curve shown in Figure 3 would be the best possible. Under these conditions, the standard deviation (sd) of curve 1 was found to be 0.9256 and sd of curve 2 was 2.024.

Figure 4. Method of estimating filedrawer size (see text)



Now that we had a full description of curves 1 and 2, we assumed that the area labeled "b" in Figure 4a was the number of observed studies with $|z| < 1.65$ (188 – 76 = 112), that area "c + d" was composed of 76 observed studies with $|z| \geq 1.65$, and that the total area "a + b + c + d" was calculated at 283 studies [5]. Doing the subtraction 283 – 112 – 76 = 95, we estimate *95 unreported or unretrieved nonsignficant studies in the actual filedrawer*. We believe that this number is a more realistic estimate than the 4700 studies determined by equation (2). In fact, 95 studies would require each of 10 parapsychology laboratories to have filed only about 0.6 studies *per year* over the 15 year survey period (as opposed to 2.6 *per month*, as 4700 studies would require).

Now if we combine the 188 observed survey studies with 95 new, nonsignificant z scores (generated by Monte Carlo technique with z chosen at random from a normal distribution, and bounded between 10^{-25} and ± 1.64), we find of the 283 resulting studies, mean z = .462, overall z = 7.768, and $p < 8.03 \times 10^{-15}$. Again applying formula (2) to the new values (for the sake of curiosity), we find X = 4078 additional nonsignificant studies needed to bring this overall p value down to p = .05, 2-tailed.

Finally, combining all survey, newly estimated, and Princeton studies (188+144), we find that for the 425 total studies the mean z = .420, overall z = 8.684, and $p < 3.9 \times 10^{-18}$. Applying formula (2), we find we would need 7778 additional nonsignificant studies in the filedrawer. Thus, from several different perspectives, it seems that the filedrawer issue is not as serious a problem as many have thought.

5. This calculation was based on the curve-fitted standard deviations for the two Gaussian curves and the observed number of studies in areas b and c + d.

Incidentally, testing the standard deviation of the z scores observed in these 425 studies (sd = 1.414) against the expected variance of 1.0 for a normal, unperturbed z distribution, results in a chi-square value of 853.7 (424 df), for a $p < 5.9 \times 10^{-34}$. Table 1 (below) and 7 (at end of the paper) summarize these findings.

Table 1. Summary of z score analyses

Source of reference	studies (N)	$\bar{z}$	$\frac{\Sigma z's}{\sqrt{N}}$	p (2-tail)	z score sd	variance test against $\sigma = 1$ χ^2	p (2-tail)
Survey	188	0.738	10.114	4.9×10^{-24}	1.739	568.5	4.9×10^{-47}
Princeton	144	0.339	4.063	4.9×10^{-5}	1.184	201.9	0.001
Estimated (simulated) filedrawer	95	−0.084	−0.820	0.412	0.661	41.5	0.51
Combined	425	0.420	8.684	3.9×10^{-18}	1.414	853.7	5.9×10^{-34}

4. Experiments are not replicable

Occasional significant effects may be impressive, but the existence of the claimed anomaly cannot be established on the basis of results reported by only a few individuals.[6] The same effect must be replicated by many others. Is it true, as Kurtz (1980) claims, that

> The basic problem ... is the *lack of replicability* by other experimenters. Apparently, some experimenters -- a relative few -- are able to get similar results, but most are unable to do so. (Italics in the original, p.12)

In fact, of the 332 experiments we considered, 78.6% failed to reach significant levels. It is hardly surprising, then, that on the basis of examining individual experiments it is easy to reach the conclusion that the effect is elusive and non-replicable. At this failure rate, nearly 4 out of 5 experiments will fail to reject the null hypothesis. (Of course, if just chance were operating, 19 out of 20 experiments would fail to reject the null hypothesis.)

6. Actually, compared to experimental psychology, experimental parapsychology is in much better shape as far as replication rates go. Honorton (1975), for instance, describes a study by Bozarth and Roberts (1972), who, in a survey of 1334 articles from psychology journals, found only eight articles involving replications of previously published work. In this present meta-analysis alone, parapsychology is a factor of 40 ahead of psychology.

Another reason why it may be difficult to produce significant experiments at will is the well-known "experimenter effect" (Rosenthal, 1976). This effect is ubiquitous to all the sciences, but parapsychology seems to be especially vulnerable (see, e.g. White, 1977). The experimenter effect may help explain why some critics of parapsychology claim that they have *never* obtained significant results in their attempts to replicate psi experiments (e.g. Kurtz, 1981, p.16; Neher, 1980, p.147). Of course, the odds of *never* obtaining a significant study can be astronomical, depending on the number of studies conducted. Unfortunately, critics rarely report the number and details of their claimed replications, so a good estimate of the probability of their never seeing a significant result cannot be made.

It should be noted that experimenter effect is only one of many confounding problems involved in the quest for the significant replication. For example, selection of subjects, experimenters, task conditions, experimental protocols, statistical procedures, environmental conditions, feedback techniques and generation of random numbers are all reflected in the ultimate outcome of an experiment. Regardless of how well controlled an experiment may be, a change in any one of these factors will affect the entire experiment in a complex, poorly understood way.

In any case, experimenter bias is unavoidable, and we must rely on well-controlled experiments with features like automated data recording to help eliminate this bias. In spite of tight controls, however, it is known that even parapsychologists who would like to replicate RNG studies cannot guarantee significant results. Thus, critics would perhaps claim that any reported significant studies are due more to unconscious or intentional experimenter bias (i.e. fraud or carelessness) rather than there being a real effect.

To address the issue of what effect different experimenters may have had in the reported RNG experiments, we ran two analyses on the survey data. The first involved calculating the overall z score obtained by each principal investigator; the second was a test of the homogeneity of mean z scores reported by different investigators.

Combined z score results Table 2 shows a combined z and mean z calculated for each of 28 different principal investigators. This list is comprised of only those studies where sufficient detail was published for us to calculate z scores from the number of trials and hits in an experiment (332 total – 35 partially detailed experiments = 297 experiments). The z(overall) scores per investigator were calculated by summing the z scores for all experiments contributed by that investigator and dividing by the square root of the number of experiments. In effect, this weights each experiment equally, regardless of the number of trials (bits) actually used in the experiment. (The number of trials run in these experiments ranged between 144 and 2 million.)

Table 2: Overall z score per investigator

Principal investigator [1]	References	Experiments	z(overall)
Andre	1	4	2.413
Bierman	1	2	3.899
Braud	2	4	3.760
Broughton	1	4	-0.470
Debes	1	8	0.356
Dunne	1	144	4.063
Edge	1	10	0.369
Giesler	1	12	2.694
Heseltine	4	19	-0.386
Hill	1	1	2.950
Honorton	4	14	1.523
Houtkooper	1	4	0.981
Jungerman	1	1	2.332
Kelly	1	2	3.366
Matas	1	2	0.513
May [2]	1	1	-2.384
Millar	2	2	-0.875
Morris	1	5	1.835
Morrison	1	3	1.342
Palmer	1	1	1.750
Pantas	1	4	1.525
Radin	1	4	4.343
Randall	1	6	-0.029
Schechter	1	2	-1.060
Schmeidler	1	1	-1.273
Schmidt	9	30	13.224
Shafer	1	2	-1.440
Winnett	1	5	-0.089
TOTAL	44	297	8.548

[1] This is the name of the first author as listed in the references.

[2] The study by May, Humphrey, and Hubbard (1980) is not included in this survey because their sequential analysis data collection technique is not amenable to z score analysis.

As seen in Table 2, the overall z scores for these investigators ranged between -2.384 to 13.224. The grand total z score, obtained by summing the 28 z scores and dividing by $\sqrt{28}$ is z = 8.548, for an overall $p < 1.27 \times 10^{-17}$ (2-tailed). If we remove Schmidt's 30 studies, since he obtained the largest overall z score and is responsible for the largest number of references in our survey, we find the grand total z = 6.160, $p < 7.31 \times 10^{-10}$ (2-tailed). If we *also* remove the Princeton data, which comprise nearly half of the reported experiments, we get a grand total z = 5.480, $p < 4.25 \times 10^{-8}$ (2-tailed). Thus, after removing the two largest contributors to the database, we are left with a fairly impressive overall result: Odds against chance of about 1 in 23,000,000. In addition, we find that 39% (11/28) of the experimenters obtained overall 2-tailed significance and 68% (19/28) obtained positive z scores.

Test for homogeneity of effect size Do different experimenters tend to observe about the same effects in their experiments? Or are there some individuals who consistently obtain significant results and others do not? In the present context, to test for homogeneity of effect size among

different experimenters, we believe it makes more sense to test the individual *z scores* obtained in each experiment rather than use effect sizes such as d, d', r, or so on, as discussed by Rosenthal (1984) and others.

The reason is the following: Effect size may be defined as

significance test = [effect size] x [size of study]

where "significance test" can be a z, t, r, chi-square, or any other statistical test. In the studies we found in the literature, it is clear that *if* the effect size were *constant* regardless of the size of the study (say, N trials), we should be observing enormous z scores when N is even moderately large. For example, if an investigator ran a study with N = 100 and obtained a z score = 2.0, this would imply that the effect size (defined as $r = 2\Delta p = z/\sqrt{N}$ for a binary RNG) would be $r = 2.0/\sqrt{100} = 2.0/10.0 = .2$. If this effect size were constant, then if we ran the same experiment again but with N = 10000, the z score for this experiment would be $z = (2\Delta p)\ \sqrt{10000} = .2(\ 100) = 20.0$. Z scores of this magnitude are simply not reported in individual experiments, thus our effect size is almost certainly n-dependent. Indeed, this phenomenon has been observed repeatedly in a variety of experiments and has been called a *goal-directed* effect (e.g. Kennedy, 1978; May, Radin *et al*, 1985; Schmidt, 1974).

To take the effect size n-dependence into account, we must multiply the effect size by a function of the size of the study, which brings us back to a significance test, as noted above. For the sake of convenience, we can use the z score calculated for each experiment. To see whether different experimenters reported about the same magnitude z scores, we performed an analysis of variance; the results are shown in Table 3 (on the next page).

It is clear from the results of the ANOVA that different experimenters do indeed obtain different mean z scores, although with 25% (7/28) of the principal investigators reporting mean z scores greater than 2 or less than -2, it is not the case that only one or two experimenters have obtained large mean z scores.

Table 3: Results of one-way analysis of variance

Grand mean

N	MEAN Z	SD	SE
297	0.5979	1.5823	0.0918

Person	*N*	*MEAN Z*	*SD*	*SE*
Andre	4	1.2065	1.9091	0.9546
Bierman	2	2.7570	1.3863	0.9803
Braud	4	1.8797	0.9373	0.4687
Broughten	4	-0.2347	0.3048	0.1524
Debes	8	0.1260	1.8205	0.6437
Dunne	144	0.3386	1.1842	0.0987
Edge	10	0.1166	2.0067	0.6346
Giesler	12	0.7778	0.8011	0.2313
Heseltine	19	-0.0885	1.7124	0.3928
Hill	1	2.9498		
Honorton	14	0.4071	1.1328	0.3028
Houtkooper	4	0.4906	1.4944	0.7472
Jungerman	1	2.3322		
Kelly	2	2.3799	0.3015	0.2132
Matas	2	0.3625	2.9522	2.0875
May	1	-2.3841		
Millar	2	-0.6187	1.6406	1.1601
Morris	5	0.8206	1.0562	0.4723
Morrison	3	0.7746	0.4926	0.2844
Palmer	1	1.7500		
Pantas	4	0.7625	2.4453	1.2226
Radin	4	2.1712	0.8822	0.4411
Randall	6	-0.0120	1.1753	0.4798
Schechter	2	-0.7496	4.2411	2.9989
Schmeidler	1	-1.2728		
Schmidt	30	2.4144	2.0341	0.3714
Shafer	2	-1.0178	1.1158	0.7890
Winnett	5	-0.0396	0.5795	0.2592

SOURCE	SS	df	MS	F	p
person	197.9629	27	7.3320	3.631	2.78×10^{-8}
error	543.1467	269	2.0191		

To see whether the mean z score might be related to the number of experiments each investigator ran, we performed a correlation between N and MEAN Z (in Table 3). Results were as follows:

Correlation	r-squared	t(21)	p
-0.0185	0.0003	-0.0941	0.9257

In summary, taking the data taken at face value (i.e. not weighted by quality analysis), we can make two statements: First, considering all available data, there *do* appear to be significant differences among mean z scores obtained by different experimenters. Second, there is a nonsignificant correlation between the number of experiments run by principal investigators and their mean z scores. So to return to the question at the beginning of this section: Do different experimenters obtain about the same results? The answer is *no* -- experimenters in this survey showed mean z scores ranging from -2.38 to 2.95. As to the question of whether only one or two individuals may be responsible for the overall significance, the answer is also *no*; 25% of the experimenters in our survey reported mean z scores beyond 2 and -2.

5. RNGs were nonrandom

This criticism may be addressed by examining the results of control studies reported in the literature. The results shown in Table 4 were compiled from 14 of the 44 detailed references referred to in Table 3, and were contributed by the following twelve authors: Dunne (Princeton), 57 control studies; Schmidt, 23; Broughten, 8; Braud, 2; and one each for Bierman, Hill, May, Millar, Morris, Schechter, Honorton, and Palmer. The other references did not report control results in detail and could not be used.

Table 4: Combined control studies

Data	Number of control studies	Σz's	$\bar{z}$	overall z	p (2-tail)	sd
Survey	41	-0.012	-0.0003	-0.002	0.999	1.036
Princeton	57	2.829	0.0496	0.375	0.708	.806
Combined	98	2.817	0.0287	0.285	0.776	.905

A variance test of the observed standard deviation (sd = .905) against the expected variance of 1.0 for 98 samples results in a chi-square = 80.2645 (97 df), z = -1.22, and p < .222 (2-tailed). Thus, for the references where control runs were described in sufficient detail to determine the

number of binary hits and trials, there is no evidence of systematic (mean or variance) bias in the RNG equipment.

QUALITY ANALYSIS: A PROPOSAL

In this section, we address how we plan to judge the quality of the published experiments. Quality analysis in effect adds a weighting factor to each experiment's reported z, t, or p value, depending on the assessed quality of that experiment. To avoid making a subjective quality assessment for each experiment, criteria and associated weights can be defined such that if a criterion is met, the weight associated with that criterion is added to that experiment's overall weighting factor. Rosenthal (1984, p.46–48) describes a variety of factors one might want to consider when performing quality analyses, but it is clear that the choice of weighting criteria depends on the research context. For the present analysis, Table 5 shows our initial proposal for criteria and associated weights; these are explained following the table.

Table 5: Weighting criteria for RNG quality analysis

Criteria*	Weighting factors	
	With data	Without data
Controls		
local control runs	30	15
global control runs	20	10
other control/random tests	10	5
target bit oscillation	10	5
Data Integrity		
automatic hit/trial counters	5	
tamper resistant equipment	5	
automatic data recording	10	
Statistical Integrity		
pre-specified analysis	10	
fixed run lengths	10	
direction of effort stated	5	
Subject type		
ordinary subjects	10	
special subjects	4	
experimenter as subject	2	
Reporting clarity		
fully reported hits or trials and z	10	
report of z, p, or t only	4	
report of other statistics	-2	
"significant" only	2	
"nonsignificant" only	4	

* See text for explanation of criteria.

Explanation of RNG weighting criteria

Controls In Table 5, a "local" control means the equipment was checked for randomness as part of the experimental protocol. A typical design is to have an experimental run followed by a

control run equivalent in all respects to the experiment run, but where the subject applies no "effort" to the task or is absent. A "global" control means the equipment (RNG, computer, etc.) was tested under the same conditions as used in the experiment, but separate from the experimental sessions. "Other" control or randomness tests meant that some reference was made to control runs or randomness tests, but the detailed results were either (a) not in the report or (b) the explanation of the controls were referenced or related to a description in another article. The columns labeled "With data" and "Without data" show different weights assigned to control runs depending on whether actual data were reported. "Target bit oscillation" means the assigned "hit" bit alternated with each newly generated bit to counterbalance any possible RNG bias.

Data Integrity The "automatic hit/trial counters" criterion is satisfied if the RNG equipment has an automated method of keeping track of hits and trials. "Automatic data recording" requires use of punched paper tape, magnetic tape, computer disk, or so on, to automatically record the data collected in the experiment. There are instances in the literature (especially in reports from the early 1970's) where the automatic counter criterion is met, but not automatic data recording. "Tamper resistant equipment" requires either that the RNG was (a) in a locked laboratory and inaccessible to subjects at any time, (b) the experiment was under the immediate supervision of an experimenter, (c) the equipment had a "fail-safe" or interlock system that prevented disruption of or tampering with the data collection process, or (d) the device was a computer with software data protection such as a password, protected files, or so on.

Statistical Integrity "Pre-specified analysis" means it is clear from the report that the statistical analysis method was defined before data was collected. "Fixed run lengths" means the total number of trials was specified *in advance* of data collection. "Direction of effort stated" requires that it was clear whether the planned test was one-tailed or two-tailed, and what direction of "effort" subjects were to aim for during the experiment.

Subject Integrity This category checked whether the subjects used in the experiment were ordinary, selected or special in some other way, or the experimenter(s). Stronger weight was applied to unselected subjects because it was felt they would have less invested in the experimental outcome and would be less likely to intentionally or unintentionally interfere with the equipment or procedures.

Reporting Integrity If the report included the actual number of trials and hits, or the number of trials and a z, p, or t score, this was assigned the greatest weight. If it included only z, p, or t scores, this was assigned less weight. Report of any other statistics that we had to transform into the equivalent of z scores were assigned the lowest weight. In addition, reports consisting only of the statement "significant," without supporting data, were assigned a weight of 2 and similarly, the statement "nonsignificant" was assigned a weight of 4.

Method of calculating quality-weighted analysis

The weighting factor per experiment would be calculated as follows: If the criterion was clearly *present* in the published report, the associated weighting factor would be added to that experiment's

weight. If the criterion was not met, the weight assigned for that factor would be zero (0). The sum of the individual weights would be the overall weight per experiment, and the final overall weighted z score is then calculated as

$$\text{Weighted } Z = \frac{\sum w_i z_i}{\sqrt{\sum w_i^2}} \qquad (5)$$

Thus the minimum weight per experiment would be 0 if there were no mention of control tests, no description indicating that data collection was protected in some way, no evidence that statistical tests were pre-planned, insufficient report on who the subjects were, and no report of results. The maximum weight would be 125 (sum of three control weights, three data integrity weights, three statistical integrity weights, use of ordinary subjects, and full report of data).

Weighting the filedrawer estimate

We propose to weight our estimated 95 nonsignificant filedrawer studies with the average weight found in the rest of the studies. This proposal has a potential criticism, however. Our means of estimating the filedrawer size depends on the observed z score distribution. Since the individual z scores depend on the weighting factors (which were in effect all 1's in the analysis reported in this paper), the unweighted filedrawer estimate may be smaller than a similar estimate made with weighted z scores, thus inflating the final results. In response to this criticism, we would point out that the quality analysis is actually orthogonal to the filedrawer estimate because the actual magnitude of a z score does not change with our quality analysis, instead the *importance* of the z score is affected, and the importance of a z score is not considered in our filedrawer estimation method, only in the final estimate of overall significance.

In addition, by adding a group of nonsignificant studies (the filedrawer estimate by definition is composed of nonsignificant studies) into a pool of z scores that have *already* been weighted according to quality, we are in effect creating an *ultra*-conservative test. A case could be made, for instance, on why a filedrawer estimate should not be added into a quality-weighted analysis at all, but to take the conservative approach given the nature of the claim, we will pool the 95 estimated studies along with the quality-weighted z scores.

Defining experiments in the Quality Analysis

Although adequate for a first-pass analysis, the method of selecting experiments described above would be less than perfect for a quality-weighted analysis. The main objection that could be raised is that the decision on what constitutes the subjects' "direction of effort" is dependent on the reviewer's interpretation of the experimental procedure. In many articles, we took educated guesses to decide what were the actual conditions, what were the subjects' intentions, did the authors in fact predict in advance the outcome, and so on.

To address this problem in Part 2 of this meta-analysis, we will actually be performing *two* separate meta-analyses. The first will take into account the *minimum* number of experiments that we decide is a reasonable partitioning, and the second meta-analysis will be for the *maximum* number of experiments. The two end results will be compared, and the more conservative of the two will be used as the overall result.

Deciding on a range of possible experiments allows us to form an "uncertainty" factor for each reference. If a reference's maximum-minimum experiment range is large as compared to the average observed range, we must consider that the quality of that *reference*, at least for our purposes, is poor. We plan on presenting a breakdown of each reference's uncertainly in the Part 2 meta-analysis to judge how clear each reference was in this study.

Example of reference source quality analysis

In Table 5 we present an example of a preliminary quality analysis applied to the source of reference. We assigned arbitrary weights according to our perception of the quality of average papers published in each parapsychological reference source (not counting the Princeton data). Then, after making guesses for these weights, we calculated a combined z score contributed by each journal and compared it to a weighted z score according to equation (5). As seen in Table 5, the original combined z score dropped by 4 orders of magnitude in significance, but the weighted z score is still quite significant. We expect that the wider range of quality weights, as we have proposed above, will make a larger difference in a weighted analysis, but it would appear that most of the reports would have to be extremely poor in quality to nullify the overall p value.

Table 5. Exploratory quality analysis of reference sources

Reference	Studies	Overall z	p(2-tail)	Assigned weight
Journal of Parapsychology	49	7.055	3.30×10^{-13}	10
Proceedings of the PA	32	5.036	4.76×10^{-7}	2
Research in Parapsychology	52	4.052	5.08×10^{-5}	1
Journal of the ASPR	5	4.105	4.04×10^{-5}	8
European Journal of Parapsychology	6	3.052	0.002	8
Journal of the SPR	9	1.692	0.091	5
Combined unweighted result =		10.27	$p < 9.45 \times 10^{-25}$	
Combined weighted result =		9.53	$p < 1.60 \times 10^{-21}$	

Example of chronological analysis

In Table 6 we show an analysis of variance of the 297 detailed experiments grouped according to year of publication.

Table 6. Chronological analysis of variance

SOURCE: grand mean

year	N	MEAN Z	SD	SE
	297	0.5979	1.5823	0.0918

SOURCE: year

year	N	MEAN Z	SD	SE
1970	5	0.8247	3.0969	1.3850
1971	6	0.6292	2.3180	0.9463
1972	9	1.3565	1.4253	0.4751
1973	6	4.1239	1.4665	0.5987
1974	10	1.1539	1.9879	0.6286
1975	9	1.5804	2.1841	0.7280
1976	17	0.7366	1.5784	0.3828
1977	23	0.5695	1.8333	0.3823
1978	9	-0.2520	1.1482	0.3827
1979	7	0.6012	1.2325	0.4658
1980	5	-1.1411	1.3480	0.6029
1981	7	2.3437	0.7836	0.2962
1982	164	0.3098	1.2595	0.0983
1983	1	1.7500		
1984	19	0.8492	1.0779	0.2473

SOURCE	SS	df	MS	F	p
year	151.2565	14	10.8040	5.165	1.14×10^{-6}
error	589.8531	282	2.0917		

This ANOVA shows that mean z scores differ significantly from year to year. We then looked for trends in z scores by performing a correlation between year and mean z. Results showed that $r = -0.205$, $t(13) = -0.756$, $p = 0.463$, i.e. there was no significant correlation between year of publication and mean z score observed for that year.

A planned quality vs. z score correlational study

Once we perform the quality analysis and have a list of raw z scores and associated quality weights, we plan on performing a correlation between these pairs of numbers. If the correlation is significantly negative, it would suggest that the better the quality of a study, the lower the z score. This would be in accordance with what some critics have claimed, namely that "there is a strong tendency for the rate of success to increase with the number of obvious defects" (Hyman, 1983, p.23). If a significant positive correlation is seen, however, this criticism can be refuted.

CONCLUSION

In an initial meta-analysis of psi experiments involving binary RNGs, we have identified 332 experiments published over the years 1969-1984 in 56 references. Based on an analysis of 188 of these experiments reported in parapsychological journals, we estimated the actual number of nonsignificant, unreported or unretrieved experiments to be 95. We found a total of 98 reported control studies in 14 of these references. A summary of the meta-analytic results is shown in Table 7 (on the following page).

In agreement with a hypothesis of a "psi effect" on RNGs, the combined data indicate that, in the aggregate, the experimental conditions resulted in anomalous statistical behavior of the RNG in the direction of effort specified by the task, and the control conditions resulted in expected binomial statistics for both mean z scores and standard deviations.

The combined data shows an interesting effect on the distribution of z scores. We find that in the experimental condition the mean z score has been increased significantly from chance expectation, which in the present context is in accord with the underlying hypothesis that the z score will shift according to the direction of the subject's effort. In the control condition we find the z mean shifted slightly, but not significantly so. We also find that the standard deviation of the combined distribution of experimental z scores has become significantly fatter than chance expectation, and that the combined control standard deviation is as expected. Both of these effects -- a shifting of the mean and fattening of the standard deviation, are accounted for in a model discussed by May, Radin *et al* (1985).

Part 2 of this study will report on a quality-weighted analysis of this same data. By weighting each study according to a semi-objective quality assessment scale, we will address the major criticisms of such experiments in a quantified way, and the overall experimental vs. control result will provide a basis for discussion on whether or not this anomaly is, in fact, real.

We urge readers to comment on and criticize the method described here, and especially on the proposed weighting criteria presented in the Quality Analysis section above. We plan to gain a consensus opinion among informed scientists on what constitutes an agreeable, conservative weighting scheme *before* we perform the quality analysis. In this way, the combined results observed in the weighted data will be less subject to *post hoc* debate over the adequacy of the

analysis method, we will avoid an enormous amount of pointless work, and we can proceed with constructive discussion. We are especially interested in comparing the quality weights proposed by parapsychologists, critics of psi research, and "neutral" scientists, as this may give us a clue as to what is considered to be important in establishing consensus agreement among these different groups.

Table 7. Summary of RNG meta-analysis

Source	studies (N)	$\bar{z}$	$\frac{\Sigma z's}{\sqrt{N}}$	p (2-tail)	standard deviation of z scores	variance test against $\sigma = 1$ χ^2	p (2-tail)
SURVEY							
Experiment	188	0.738	10.114	4.9×10^{-24}	1.739	568.5	4.9×10^{-47}
Control	41	-0.0003	-0.002	0.999	1.036	44.0	0.62
PRINCETON							
Experiment	144	0.339	4.063	4.9×10^{-5}	1.184	201.9	0.001
Control	57	0.050	0.375	0.708	.806	37.0	0.05 *
FILEDRAWER ESTIMATE							
Experiment	95	-0.084	-0.820	0.412	0.661	41.5	0.51
COMBINED							
Experiment	427	0.420	8.684	3.9×10^{-18}	1.414	853.7	5.9×10^{-34}
Control	98	0.029	0.285	0.776	.905	80.3	0.22

* This "too small" variance in the control data is compatible with a model proposed by May, Radin et al (1985) and is also discussed by Jahn, Nelson and Dunne (1985).

REFERENCES

General References

Akers, C. Methodological criticisms of parapsychology. In S. Krippner (Ed.), *Advances in parapsychological research, Volume 4.* Jefferson, NC: McFarland & Company, Inc., 1984.

Aspect, A., Dalibard, J. and Roger, G. Experimental test of Bell's inequalities using time-varying analyzers, *Physical review letters,* 49, 1982, 1804-1807.

Aspect, A., Grangier, P., and Roger, G. Experimental realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: A new violation of Bell's inequalities. *Physical review letters,* 49, 1982, 91-94.

Beloff, J. Parapsychology and philosophy. In B. Wolman (Ed.), *Handbook of parapsychology*, New York: Van Nostrand, 1977.

Bower, B. Getting into Einstein's brain. *Science News*, Vol. 127, No. 21, May 25, 1985, p. 330.

Bozarth, J. D. and Roberts, R. R. Signifying significant significance. *American Psychologist*, Vol. 27, 1972, 774-775.

Carpenter, J. C. Intrasubject and subject-agent effects in ESP experiments. In B. Wolman (Ed.), *Handbook of parapsychology*, New York: Van Nostrand, 1977.

Cooper, H. M. & Rosenthal, R. Statistical versus traditional procedures for summarizing research findings. *Psychological bulletin*, Vol. 87, 1980, 442-449.

d'Espagnat, B. The quantum theory and reality. *Scientific American*, November 1979, 158-181.

Dunne, B. J., Jahn, R. G., Nelson, R. D. An REG experiment with large data-base capability, II: Effects of sample size and various operators. *Technical Note PEAR 82001*, Princeton Engineering Anomalies Research Laboratory, Princeton University, School of Engineering / Applied Science, 1982.

Dunne, B. J., Jahn, R. G., Nelson, R. D. Precognitive remote perception. *Technical Note PEAR 83003*, Princeton Engineering Anomalies Research Laboratory, Princeton University, School of Engineering / Applied Science, 1983.

Epstein, S. The stability of behavior. II. Implications for psychological research. *American Psychologist,* 35, 9, September 1980, 790-806.

Fisher, R. A. *Statistical methods for research workers,* (2nd ed.). London: Oliver & Boyd, 1928.

Hansel, C. E. M. ESP and parapsychology: A critical reevaluation. Buffalo, NY: Prometheus Books, 1980.

Honorton, C. Error some place! *Journal of communication*, Vol. 25:1, Winter 1975.

Honorton, C. Replicability, experimenter influence, and parapsychology: An empirical context for the study of mind. Paper presented at the meeting of the American Association for the Advancement of Science, Washington, D. C., 1978.

Honorton, C. Meta-analysis of psi ganzfeld research: A response to Hyman. *Journal of Parapsychology*, Vol. 49, No. 1, 1985, 51-92.

Hyman, R. Does the ganzfeld experiment answer the critics' objections? In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in Parapsychology 1982*. Metuchen, NJ: Scarecrow Press, 1983, 21-23.

Hyman, R. The ganzfeld psi experiment: A critical appraisal. *Journal of Parapsychology*, 49, 1985, 3-50.

Jahn, R. G. The persistent paradox of psychic phenomena: An engineering perspective. *Proceedings of the IEEE*, Vol. 70, No. 2, February 1982.

Jahn, R. G., Nelson, R. D., and Dunne, B. J. Variance effects in REG series score distributions. *Proceedings of the 28th Annual Parapsychological Association Convention,* Tufts University, Medford, Massachusetts, August 12-16, 1985.

Kennedy, J. E. The role of task complexity in PK: A review. *Journal of Parapsychology*, 42, 1978, 89-122 .

Kurtz, P. Is parapsychology a science? In K. Frazier, *Paranormal borderlands of science.* Buffalo, NY: Prometheus Books, 1981.

May, E. C., Humphrey, B. S., and Hubbard, G. S. Electronic system perturbation techniques. SRI International *Final Report*, September 30, 1980.

May, E. C., Radin, D. I., Hubbard, G. S., Humphrey, B. S. and Utts, J. M. Psi experiments with random number generators: An informational model. *Proceedings of the Presented Papers of the 28th Annual Parapsychological Association Convention,* Tufts University, Medford, Massachusetts, August 12-16, 1985.

Mermin, N. D. Is the moon there when nobody looks? Reality and the quantum theory. *Physics today*, April 1985, 38-47.

Neher, A. *The psychology of transcendence,* Englewood Cliffs, NJ: Prentice-Hall, 1980.

Nelson, R. D., Dunne, B. J. and Jahn, R. G. 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.

Palmer, J. ESP research findings: 1976-1978. In S. Krippner (Ed.), *Advances in parapsychological research, Volume 3.* New York: Plenum Press, 1982.

Rohrlich, F. Facing quantum mechanical reality. *Science,* Vol. 221, No. 4617, September 23, 1983, 1251-1255.

Rosenthal, R. *Experimenter effects in behavioral research,* (rev. ed.). New York: Irvington, 1976.

Rosenthal, R. *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage Publications, 1984.

Rush, J. H. Problems and methods in psychokinesis research. In S. Krippner (Ed.), *Advances in parapsychological research, Volume 3.* New York: Plenum Press, 1982.

Schechter, E. I. Hypnotic induction vs. control conditions: Illustrating an approach to the evaluation of replicability in parapsychological data. *Journal of the American Society for Psychical Research*, 78, 1-28, 1984.

Schmeidler, G. R. Psychokinesis: The basic problem, research methods, and findings. In S. Krippner (Ed.), *Advances in parapsychological research, Volume 4.* Jefferson, NC: McFarland & Company, Inc., 1984.

Schmidt, H. Comparison of PK action on two different random number generators. *Journal of Parapsychology*, 1974, 38, 47–55.

Stanford, R. G. Experimental psychokinesis: A review from diverse perspectives. In B. B. Wolman (Ed.), *Handbook of parapsychology*, NY: Van Nostrand Reinhold Company, 1977.

Stanford, R. G. Recent ganzfeld–ESP research: A survey and critical analysis. In S. Krippner (Ed.), *Advances in parapsychological research, Volume 4.* Jefferson, NC: McFarland & Company, Inc., 1984.

Tart, C. T. Laboratory PK: Frequency of manifestation and resemblance to precognition. In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in Parapsychology 1982.* Metuchen, NJ: Scarecrow Press, 1983, 101–102.

White, R. A. The influence of the experimenter motivation, attitudes and methods of handling subjects in psi test results. In B. B. Wolman (Ed.), *Handbook of parapsychology*, NY: Van Nostrand Reinhold Company, 1977.

Meta–Analysis References

Note: The following references describe psi experiments using RNGs in various ways. In this list, the following codes are used: A ★ means the reference was used in our binary RNG meta–analysis; a ☞ means the reference was also used, but the experiments mentioned in the report were simulated due to lack of sufficient detail; ★☞ means this report contained both detailed and non–detailed studies.

★ **Andre, E.** Confirmation of PK action on electronic equipment. *Journal of Parapsychology*, 1972, 36, 283–293.

Bierman, R. J., and Wout, N. V. T. The performance of healers in PK tests with different RNG feedback algorithms. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1976.* Metuchen, NJ: Scarecrow Press, 1977, 131–133.

★ **Bierman, R. J., and Houtkooper, J. M.** Exploratory PK tests with a programmable high speed random number generator. *European Journal of Parapsychology*, 1975, 1, 3–14.

☞ **Braud, W.** Allobiofeedback: Immediate feedback for a psychokinetic influence upon another person's physiology. In W. G. Roll (Ed.), *Research in Parapsychology 1977.* Metuchen, NJ: Scarecrow Press, 1978, 123–134.

★ **Braud, L., and Braud, W.** Psychokinetic effects upon a random event generator under conditions of limited feedback to volunteers and experimenter. *Journal of the Society for Psychical Research*, 1979, 50, 21–30.

☞ **Braud, W., and Schroeter, W.** Psi tests with Algernon, a computer oracle. In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in Parapsychology 1982.* Metuchen, NJ: Scarecrow Press, 1983, 163–165.

★☞ **Braud, W. G., Smith, G., Andrew, K., and Willis, S.** Psychokinetic influences on random number generators during evocation of "analytic" vs. "nonanalytic" modes of information processing. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1975.* Metuchen, NJ: Scarecrow Press, 1976, 85–88.

☞ **Broughton, R. S., and Millar, B.** A PK experiment with a covert release-of-effort test. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1976*. Metuchen, NJ: Scarecrow Press, 1977, 28-30.

☞ **Broughton, R., Millar, B., Beloff, J., and Wilson, K.** A PK investigation of the experimenter effect and its psi-based component. In W. G. Roll (Ed.), *Research in Parapsychology 1977*. Metuchen, NJ: Scarecrow Press, 1978, 41-48.

★ **Broughton, R. S., Millar, B., and Johnson, M.** An investigation into the use of aversion therapy techniques for the operant control of PK production in humans. *Proceedings of the Parapsychological Association Convention*, 1979, 1-18.

Broughton, R. S. and Perlstrom, J. R. Results of a special subject in a computerized PK game. *Proceedings of the Parapsychological Association Convention*, 1984, 411-419.

Camstra, B. PK conditioning. In W. G. Roll, R. L. Morris, & J. D. Morris (Eds.), *Research in Parapsychology* 1972. Metuchen, NJ: Scarecrow Press, 1973, 25-27.

☞ **Davis, J. W. and Morrison, M. D.** A test of the Schmidt model's prediction concerning multiple feedback in a PK test. In W. G. Roll (Ed.), *Research in Parapsychology* 1977. Metuchen, NJ: Scarecrow Press, 1978, 163-168.

★ **Debes, J. and Morris, R. L.** Comparison of striving and nonstriving instructional sets in a PK study. *Journal of Parapsychology*, 1982, 46, 297-312.

★ **Dunne, B. J., Jahn, R. G., and Nelson, R. D.** An REG experiment with large data-base capability. In W. G. Roll, R. L. Morris, & R. A. White (Ed.), *Research in Parapsychology* 1981. Metuchen, NJ: Scarecrow Press, 1982, 50-51. Main reference is **Nelson, R. D., Dunne, B. J. and Jahn, R. G.** 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.

Dunne, B. J., Jahn, R. G., and Nelson, R. D. An REG experiment with large data-base capability, II: effects of sample size and various operators. In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in Parapsychology* 1982. Metuchen, NJ: Scarecrow Press, 1983, 154-157.

★ **Edge, H. L.** Plant PK on an RNG and the experimenter effect. In W. G. Roll (Ed.), *Research in Parapsychology* 1977. Metuchen, NJ: Scarecrow Press, 1978, 169-174.

★ **Giesler, P. V.** Differential micro-PK effects among Afro-Brazilian Caboclo and Candomble cultist using trance-significant symbols as targets. *Proceedings of the Parapsychological Association Convention*, 1984, 87-105.

★ **Heseltine, G. L.** Electronic random number generator operation associated with EEG activity. *Journal of Parapsychology*, 1977, 41, 103-118.

★ **Heseltine, G. L., and Mayer-Oakes, S. A.** Electronic random generator operation and EEG activity: further studies. *Journal of Parapsychology*, 1978, 42, 123-136.

★ **Heseltine, G. L. and Kirk, J. H.** Examination of a majority-vote technique. *Journal of Parapsychology*, 1980, 44, 167-176.

★ **Heseltine, G. L.** PK success during structured and non structured RNG operation. *Proceedings of the Parapsychological Association Convention*, 1984, 379-388.

★ **Hill, S.** PK effects by a single subject on a binary random number generator based on electronic noise. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology* 1976. Metuchen, NJ: Scarecrow Press, 1977, 26-28.

★ **Honorton, C.** Effects of meditation and feedback on psychokinetic performance: a pilot study with an instructor of transcendental meditation.In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology* 1976. Metuchen, NJ: Scarecrow Press, 1977, 95-97.

★ **Honorton, C., Barker, P., and Sondow, N.** Feedback and participant-selection parameters in a computer RNG study. In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in Parapsychology* 1982. Metuchen, NJ: Scarecrow Press, 1983, 157-159.

★ **Honorton, C. and Barksdale, W.** PK performance with waking suggestions for muscle tension versus relaxation. *Journal of the American Society for Psychical Research*, 1972, 66, 208-214.

☞ **Honorton, C. and Tremmel, L.** Psi correlates of volition: a preliminary test of Eccles' "Neurophysiological Hypothesis" of mind-brain interaction. In W. G. Roll (Ed.), *Research in Parapsychology* 1978. Metuchen, NJ: Scarecrow Press, 1979, 36-38.

★ **Honorton, C. and May, E. C.** Volitional control in a psychokinetic task with auditory and visual feedback. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology* 1975. Metuchen, NJ: Scarecrow Press, 1976, 90-91.

★ **Houtkooper, J. M.** A study of repeated retroactive psychokinesis in relation to direct and random PK effects. *European Journal of Parapsychology*, 1977, 4, 1-20.

★ **Jungerman, R. L., and Jungerman, J. A.** Computer controlled random number generator PK tests. In W. G. Roll (Ed.), *Research in Parapsychology* 1977. Metuchen, NJ: Scarecrow Press, 1978, 157-162.

★ **Kelly, E. F. and Kanthamani, B. K.** A subject's efforts toward voluntary control. *Journal of Parapsychology*, 1972, 36, 185-197.

Levi, A. The influence of imagery and feedback on PK effects. *Journal of Parapsychology*, 1979, 43, 275-289.

★ **Matas, F. and Pantas, L.** A PK experiment comparing meditating versus nonmeditating subjects. *Proceedings of the Parapsychological Association Convention*, 1971, 12-13.

★ **May, E. C. and Honorton, C.** A dynamic PK experiment with Ingo Swann. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology* 1975. Metuchen, NJ: Scarecrow Press, 1976, 88-89.

Millar, B. A covert PK test of a successful psi experimenter. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology* 1976. Metuchen, NJ: Scarecrow Press, 1977, 111-113.

★ **Millar, B. and Broughton, R.** A preliminary PK experiment with a novel computer-linked high speed random number generator. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology* 1975. Metuchen, NJ: Scarecrow Press, 1976, 83-84.

★☞ **Millar, B. and Mackenzie, P.** A test of intentional versus unintentional PK. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology* 1976. Metuchen, NJ: Scarecrow Press, 1977, 32-35.

★ **Morrison, M. D., and Davis, J. W.** PK with immediate, delayed, and multiple feedback: a test of the Schmidt model's predictions. In W. G. Roll (Ed.), *Research in Parapsychology* 1978. Metuchen, NJ: Scarecrow Press, 1979, 117-120.

★ **Morris, R. L., Nanko, M., and Phillips, D.** A comparison of two popularly advocated visual imagery strategies in a psychokinesis task. *Journal of Parapsychology*, 1982, 46, 1-16.

★☞ **Palmer, J. and Kramer, W.** Internal state and temporal factors in RNG PK. In R. A. White & R. S. Broughton (Eds.), *Research in Parapsychology 1983*. Metuchen, NJ: Scarecrow Press, 1984, 28–30.

★ **Pantas, L.** PK scoring under preferred and nonpreferred conditions. *Proceedings of the Parapsychological Association Convention*, 1971, 47–49.

★☞ **Radin, D. I.** Mental influence on machine-generated random events: six experiments. In W. G. Roll, R. L. Morris, & R. W. White (Eds.), *Research in Parapsychology* 1981. Metuchen, NJ: Scarecrow Press, 1982, 141–142.

★ **Randall, J. L.** An extended series of ESP and PK tests with three English schoolboys. *Journal of the Society for Psychical Research*, 1974, 47, 485–494.

☞ **Schechter, E. I., Honorton, C, Barker, P., and Varvoglis, M. P.** Relationships between participant traits and scores on two computer-controlled RNG-PK games. In R. A. White & R. S. Broughton (Eds.), *Research in Parapsychology 1983*. Metuchen, NJ: Scarecrow Press, 1984, 32–33.

☞ **Schechter, E., Barker, P., and Varvoglis, M. P.** A second study with the "psi ball" RNG-PK game. In R. A. White & R. S. Broughton (Eds.), *Research in Parapsychology 1983*. Metuchen, NJ: Scarecrow Press, 1984, 93–94.

★ **Schechter, E. I., Barker, P., and Varvoglis, M.** A preliminary study with a PK game involving distraction from the psi task. In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in Parapsychology* 1982. Metuchen, NJ: Scarecrow Press, 1983, 152–154.

★ **Schmeidler, G. R. and Borchardt, R.** Psi scores with random and pseudo-random targets. In W. G. Roll (Ed.), *Research in Parapsychology 1980*. Metuchen, NJ: Scarecrow Press, 1981, 45–47.

Schmidt, H. Precognition of a quantum process. *Journal of Parapsychology*, 1969, 33, 99–108.

★ **Schmidt, H.** A PK test with electronic equipment. *Journal of Parapsychology*, 1970a, 34, 175–181.

★☞ **Schmidt, H.** PK experiments with animals as subjects. *Journal of Parapsychology*, 1970b, 34, 255–261.

Schmidt, H. An attempt to increase the efficiency of PK testing by an increase in the generation speed. In W. G. Roll, R. L. Morris, & J. D. Morris (Eds.), *Research in Parapsychology* 1972. Metuchen, NJ: Scarecrow Press, 1973a, 65–67.

★ **Schmidt, H.** PK tests with a high-speed random number generator. *Journal of Parapsychology*, 1973b, 37, 105–118.

★ **Schmidt, H.** PK effect on random time intervals. In W. G. Roll, R. L. Morris, & J. D. Morris (Eds.), *Research in Parapsychology* 1973. Metuchen, NJ: Scarecrow Press, 1974a, 46–48.

★ **Schmidt, H.** Comparison of PK action on two different random number generators. *Journal of Parapsychology*, 1974b, 38, 47–55.

Schmidt, H. Observation of subconscious PK effects with and without time displacement. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology* 1974. Metuchen, NJ: Scarecrow Press, 1975, 116–121.

★ **Schmidt, H.** PK experiment with repeated, time displaced feedback. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology* 1975. Metuchen, NJ: Scarecrow Press, 1976a, 107–109.

★ **Schmidt, H.** PK effect on pre-recorded targets. *Journal of the American Society for Psychical Research*, 1976b, 70, 267-291.

★ **Schmidt, H.** A take-home test in PK with pre-recorded targets. In W. G. Roll (Ed.), *Research in Parapsychology* 1977. Metuchen, NJ: Scarecrow Press, 1978, 31-36.

Schmidt, H. Use of stroboscopic light as rewarding feedback in a PK test with prerecorded and momentarily-generated random events. In W. G. Roll (Ed.), *Research in Parapsychology* 1978. Metuchen, NJ: Scarecrow Press, 1979a, 115-117.

☞ **Schmidt, H.** Search for psi fluctuations in a PK test with cockroaches. In W. G. Roll (Ed.), *Research in Parapsychology 1978*. Metuchen, NJ: Scarecrow Press, 1979b, 77-78.

★ **Schmidt, H.** PK tests with pre-recorded and pre-inspected seed numbers. *Journal of Parapsychology*, 1981, 45, 87-98.

★ **Schmidt, H.** Addition effect for PK on pre-recorded targets. *Proceedings of the Parapsychological Association Convention*, 1984, 136-139.

Schmidt, H. and Pantas, L. Psi tests with psychologically equivalent conditions and internally different machines. *Proceedings of the Parapsychological Association Convention*, 1971, 49-51.

Schmidt, H. and Pantas, L. Psi tests with internally different machines. *Journal of Parapsychology*, 1972, 222-232.

★ **Schmidt, H. and Terry, J. C.** Search for a relationship between brainwaves and PK performance. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology* 1976. Metuchen, NJ: Scarecrow Press, 1977, 30-32.

★ **Shafer, M. G.** A PK experiment with random and pseudorandom targets. In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in Parapsychology* 1982. Metuchen, NJ: Scarecrow Press, 1983, 64-66.

Stanford, R. G., Zenhausern, R., Taylor, A., and Dwyer, M. A. Psychokinesis as psi-mediated instrumental response. *Journal of the American Society for Psychical Research*, 1975, 69, 127-133.

Stanford, R. G. "Associative activation of the unconscious" and "visualization" as methods for influencing the PK target: a second study. *Journal of the American Society for Psychical Research*, 1981, 75, 229-240.

Tart, C. T. Are prepared random sequences and real time random generators interchangeable? In W. G. Roll, J. Beloff, & J. McAllister (Eds.), *Research in Parapsychology* 1980. Metuchen, NJ: Scarecrow Press, 1981, 43-47.

Terry, J. and Schmidt, H. Conscious and subconscious PK tests with pre-recorded targets. In W. G. Roll (Ed.), *Research in Parapsychology* 1977. Metuchen, NJ: Scarecrow Press, 1978, 36-41.

☞ **Talbert, R. and Debes, J.** Time-displacement psychokinetic effects on a random-number generator using varying amounts of feedback. In W. G. Roll, R. L. Morris, & R. A. White (Eds.), *Research in Parapsychology* 1981. Metuchen, NJ: Scarecrow Press, 1982, 58-61.

☞ **Varvoglis, M. P. and McCarthy, D.** Psychokinesis, intentionality, and the attentional object. In W. G. Roll, R. L. Morris, & R. A. White (Eds.), *Research in Parapsychology* 1981. Metuchen, NJ: Scarecrow Press, 1982, 51-55.

★ **Winnett, R. and Honorton, C.** Effects of meditation and feedback on psychokinetic performance: results with practitioners of Ajapa yoga. In J. D. Morris, W. G. Roll, & R. L.

Morris (Eds.), *Research in Parapsychology* 1976. Metuchen, NJ: Scarecrow Press, 1977, 97–98.

Approved For Release 2000/08/08 : CIA-RDP96-00789R002200390001-5