portunity to contribute to scientific knowledge. It was an accountant [6] who had the fine idea to put a Band-Aid on the sole of this foot before a firewalk. Let us use the investigation of paranormal phenomena as a tool to educate the public in the nature of scientific thought. We will do this by encouraging them to join us in the endeavor, supporting them in their enthusiasms, and offering them our experience and knowledge.

1 Beauchamp, H. K., Fire-walking ceremonies in India. J. Soc. psych. Res. *8* (1900) 312.
2 Brown, G., Burniston, A report of three experimental fire-walks by Ahmed Hussain and Others. Bulletin IV, University of London Council for Psychical Investigation, London, 1938.
3 Coe, M. R. Jr, Fire-walking and related behaviors. Psychol. Rec. *7* (1957) 101–110.
4 Curzon, F. L., The Leidenfrost phenomenon. Am. J. Phys. *46* (1978) 825–828.
5 Darling, C. R., Fire-walking. Nature *136* (1935) 521.
6 Dennett, M. R., Firewalking: Reality or illusion. The Skeptical Inquirer *10* (1985) 36–40.
7 Doherty, J., Hot feat: Firewalkers of the world. Sci. Dig. *90* (1982) 66–71.
8 Feigen, G. M., Bucky Fuller and the Firewalk. Saturday Review *12* (1969) 22–23.
9 Feinberg, L., Fire Walking in Ceylon. Atlant. Mon. *203* (1959) 73–76.
10 Finn, R., Ghostbusters. Engng Sci. *48* (1985) 2–7.
11 Fodor, J. A., Bever, T. G., and Garrett, M. F., The Psychology of Language. McGraw-Hill, New York 1974.
12 Fonseka, C., Fire-walking: A scientific investigation. Ceylon med. J. *17* (June 1971) 104–109.
13 Freeman, J. M., Trial by Fire. Nat. Hist., N.Y. *83* (January 1974) 54–63.
14 General Meeting. J. Soc. psych. Res. *7* (1899) 146–148.
15 Kahnemann, D., Slovic, P., and Tversky, A. (Eds), Judgment under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge 1982.
16 Kane, S M., Holiness ritual fire handling. Ethos *10* (1982) 369–385.
17 Lang, A. The Fire Walk. Proc. Soc. psych. Res. *36* (1900) 2–15.
18 Langley, S. P., [Letters to the Editor] The Fire-Walk Ceremony in Tahiti. Nature *64* (1900) 397–399.
19 Leikind, B. J., and McCarthy, W. J., An investigation of firewalking. The Skeptical Inquirer *10* (1985) 23–34.
20 Lewis, L. E., The fire-walking Hindus of Singapore. Natn. geogr. Mag. *59* (1931) 513–522.
21 Loftus, E. F., and Palmer, J. C., Reconstruction of automobile destruction: An example of the interaction between language and memory. J. verb. Learn. verb. Behav. *13* (1973) 585–589.
22 Malhotra, K. C., and Klohmne, S. B., Fire walk ceremony at the village Hanuman Takli, Maharashtra. East. Anthrop *33* (1980) 83–88.
23 Marden, L., The islands called Fiji. Natn. geogr. Mag. *114* (1958) 526–561.
24 McCarthy, W. J., and Leikind, B. J., Walking on Fire: Feat of Mind? Psychology Today *20* (1986) 10–12.
25 Melzack, R., and Wall, P., Pain mechanisms: A new theory. Science *50* (1965) 971–979.
26 Miller, G. A., and McNeill, D., Psycholinguistics, in: Handbook of Social Psychology. Eds G. Lindzey and E. Aronson. Addison-Wesley, Reading, Massachusetts 1969.
27 Myers, J., Social Psychology. McGraw-Hill, New York 1983.
28 Obeyesekere, G., The Fire-walkers of Kataragama. J. Studies *37* (May 1978) 457–476.
29 Pannett, C. A., Demonstration of Firewalking. Nature *136* (1935) 468.
30 Pennebaker, J. W., The Psychology of Physical Symptoms. Springer Verlag, New York 1982.
31 Price, H., Walking through fire. The Listener *18* (1935) 470–473.
32 Price, H., Fire-walking experiments. Br. med. J. *2* (1935) 586.
33 Price, H., A report of two Experimental Fire-Walks by Kuda Bux and others. Bulletin II, University of London Council for Psychical Investigation, London 1936.
34 Rao, S., How do people walk on fire? Science Today, Nov. (1981) 69.
35 Ross, L., and Sicoly, F., Egocentric biases in availability and attribution. J. person. soc. Psychol. *47* (1979) 322–336.
36 Schank, R., and Abelson, R. P., Scripts, Plans and Understanding: An Inquiry into Human Knowledge Structures. Erlbaum, Hillsdale, New Jersey 1977.
37 Taylor, S. E., Health Psychology, Random House, New York 1986.
38 Thomas, E. C., Fire-walking. Nature *137* (1936) 213–215.
39 Thomas, M. C., Copra-ship Voyage to Fiji's Outlying Islands. Natl geogr. Mag. 98 (1950) 121–140.
40 Walker, J., Drops of water dance on a hot skillet and the experimenter walks on hot coals. Scient. Am. *237* (August 1977) 126–131.
41 Whorf, B. L., Language, Thought and Reality. Ed. J. B. Carroll. MIT Press, Cambridge, Massachusetts 1956.

0014-4754/88/040310-06$1.50 + 0.20/0

Psi experiments: Do the best parapsychological experiments justify the claims for psi?

R. Hyman

Department of Psychology, University of Oregon, Eugene (Oregon 97403, USA)

Summary. Since the founding of the Society of Psychical Research in 1982, psychical researchers have, in each generation, generated research reports which they believed justified the existence of paranormal phenomena. Throughout this period the scientific establishment has either rejected or ignored such claims. The parapsychologists, with some justification, complained that their claims were being rejected without the benefit of a fair hearing. This paper asks the question of how well the best contemporary evidence for psi – the term used to designate ESP and psychokinetic phenomena – stands up to fair and unbiased appraisal. The results of the scrutiny of the three most widely heralded programs of research – the remote viewing experiments, the psi ganzfeld research, and the work with random number generators – indicates that parapsychological research falls short of the professed standards of the field. In particular, the available reports indicate that randomization is often inadequate, multiple statistical testing without adjustment for significance levels is prevalent, possibilities for sensory leakage are not uniformly prevented, errors in use of statistical tests are much too common, and documentation is typically inadequate. Although the responsible critic cannot argue that these observed departures from optimal experimental procedures have been the sole cause of the reported findings, it is reasonable to demand that the parapsychologists produce consistently significant findings from experiments that are methodologically adequate before their claims are taken seriously.
Key words. Parapsychology; remote viewing; ganzfeld; RNG; PK.

Introduction

The Society for Psychical Research was founded in London in 1882 to scientifically investigate 'that large group of debatable phenomena designated by such terms as mesmeric, psychical, and Spiritualistic' [16]. From the founding of this society through the present, the major psychical researchers have carried out investigations which they claim have met the standards of scientific evidence and whose results support the conclusion that paranormal phenomena have been demonstrated. However, during this same period the majority of the scientific establishment has either ignored or rejected the claims of the psychical researchers. Along with their rejection of the claims, the skeptical scientists have insisted that the evidence was inadequate.

Today, parapsychologists (the contemporary term for psychical researchers) conduct experiments which they publish in several refereed parapsychological journals. These parapsychologists, for the most part, have been trained in one or more of the recognized natural or social scientific disciplines. The experiments feature the same types of methodological controls, sophisticated instrumentation, and statistical analyses that one finds in the more orthodox scientific disciplines.

Despite this apparent sophistication in methodology and continual publication of experimental findings purporting to confirm the existence of psi (ESP and psychokinetic phenomena), the majority of the scientific establishment still does not accept the claims. Indeed, in many ways the relationship between psychical research and the scientific establishment has remained the same from 1882 to the present. Each new generation of psychical investigators puts before the scientific community experimental results which it claims proves the existence of paranormal phenomena. And each new generation of orthodox scientists either ignores the evidence or dismisses it out of hand.

One aspect of this relatively static relationship that particularly frustrates the parapsychologists is that the members of the scientific community often judge the parapsychological claims without firsthand knowledge of the experimental evidence. Very few of the scientific critics have examined even one of the many experimental reports on psychic phenomena. Even fewer, if any, have examined the bulk of the parapsychological literature that appears regularly in *The Journal of Parapsychology, The Journal of the Society for Psychical Research, The Journal of the American Society for Psychical Research,* and *The European Journal of Parapsychology.*

Consequently, parapsychologists have justification for their complaint that the scientific community is dismissing their claims without a fair hearing. The parapsychologists have consistently maintained that their research has to conform to standards that are more stringent than one finds in psychology and other, related fields of inquiry. And they have asserted that if scientists would examine their experimental reports with an open mind the scientific community would have to admit that the evidence justifies the parapsychological claims.

This paper deals with the question, 'Does an impartial and objective evaluation of the best contemporary research in parapsychology justify the claims for paranormal phenomena'? In the past few years a few critics have taken the time and trouble to carefully evaluate some of the best research programs in current parapsychology. Two critics, Akers [1] and this author [4], applied systematic criteria for evaluating the quality of a well-defined and carefully selected segment of the parapsychological literature.

The critical evaluation of the parapsychological literature strongly suggests that the quality falls far short of what the parapsychological community has believed and claimed. As a result of independent surveys, both Akers, [1] a former parapsychologist, and I [4], an outside critic, came to the same conclusion: The best parapsychological research is sufficiently flawed that it cannot support any conclusions about the existence of paranormal phenomena.

What constitutes evidence for psi?

Parapsychologists use the term 'psi' to refer 'to a person's extrasensorimotor communication with environment. Psi includes ESP and PK'. (The definitions in this paragraph can be found in the back of any issue of *The Journal of Parapsychology.*) By 'ESP' or extrasensory perception parapsychologists refer to the 'Experience of, or response to, a target object, state, event, or influence without sensory contact.' By 'PK' or psychokinesis, parapsychologist refer to 'The extramotor aspect of psi; a direct (i.e., mental but nonmuscular) influence exerted by the subject on an external physical process, condition, or object'. Although these definitions are neat and tidy, no acceptable positive theory of psi exists. Parapsychologists recognize that until such a theory has been developed, both the definition of, and evidence for, psi have to be negative.

The parapsychologist John Palmer has characterized the situation in the following words: "How do parapsychologists define psi?... One [definition] which I think most of us would accept is the following: Psi is a statistically significant departure of results from those expected by chance under circumstances that mimic exchanges of information between living organisms and their environment, *provided that,* A) proper statistical models and methods are used to evaluate the significance, and B) reasonable precautions have been taken to eliminate sensory cues and other experimental artifacts" [12].

To this characterization I would add (in addition to his A and B) that C) reasonable steps have been taken to insure that the assumptions of the statistical model being used have been met (assumptions such as independence of sampling units, appropriate randomization of targets, and the like).

This characterization, then, implies the minimal criteria needed to demonstrate psi. Requirement 'C' indicates that special care must be taken to insure that any statistical conclusions are valid. Requirement 'A' is needed to assure us that the observed number of successful 'hits' is significantly better than chance. And requirement 'B' certifies that the non-chance results could not be due to sensory cues or other artefacts. If an experiment can be faulted on any of these three requirements, then both parapsychologists and their critics should agree that it cannot be used as evidence for psi.

At this general level, then I think that parapsychologists and their critics can agree. To demonstrate psi, an experiment must be so designed and conducted that the statistical tests can be legitimately interpreted and that the possibilities for normal sensory communication between target and subject have been eliminated. When disagreement arises, it rarely, if ever, concerns these three general requirements. Rather, disagreements tend to focus an whether one or more of these requirements was adequately achieved in a given experiment.

The reason why the experimental approach has become dominant in parapsychology is just because it provides the only known way to guarantee that these minimal criteria have in fact been met. Only the well-controlled experiment can provide us with all the safeguards needed to eliminate possible alternatives to psi. Data obtained from non-experimental situations or from incompletely controlled experiments always allow for non-paranormal explanations. Some of these non-paranormal alternatives may be more or less plausible. And, indeed parapsychologists have often put forth flawed data as evidence for psi on the grounds that the conceivable normal alternatives were highly implausible. Such arguments replace reliance on strict experimental con-

trol with arguments based on plausibility. Unfortunately, plausibility is a highly subjective criterion. And to skeptics, even highly implausible possibilities may appear relatively more likely than the even greater implausibility (to them) of the paranormal.

The criteria I have been discussing up to now are 'local' in the sense that they apply to the evaluation of a single experiment. As is true in other areas of scientific inquiry, most of the explicit concern with methodology deals with adequacy of the individual experiment. Such emphasis on local criteria seems inconsistent with the fact that scientific inquiry is a cumulative process. A single experiment, no matter how well conducted, never suffices to establish a conclusion, especially one that is novel or surprising in the light of contemporary theories. Rather, laws and theories are based on the cumulative trends of many experiments carried out in several independent laboratories and which tend to converge upon a coherent set of findings.

I will refer to standards which apply to groups of experiments as 'global' criteria. With one major exception, most of the debates on the adequacy of parapsychological experiments, have focussed on local criteria such as the adequacy of the statistics and the possibilities for sensory leakage. The major exception has been the issue of replicability. Another exception has been the suspicion that many unsuccessful experiments go unreported, thereby inflating the apparent proportion of successful outcomes. This latter concern is known as the 'file-drawer problem'.

But global criteria include much more than the replicability and file-drawer problems. There is the matter of lawfulness, It is one thing to argue that a certain proportion of the experiments in a given area yields significant departures from chance. It is another, and scientifically more important, question whether the results of the different experiments yield consistent and lawful patterns. A related matter is cumulativeness. Do today's experimental findings build upon and extend the findings of the previous generations of the experimenters in the field? Parapsychology seems to be alone among the various areas of inquiry which claim scientific status in that it lacks such cumulativeness. As I have argued elsewhere, each new generation of parapsychologists discards the findings of the previous generation and essentially starts from scratch with new paradigms [5]. Other important, but even vaguer, global criteria could be mentioned such as theoretical productivity and paradigmicity.

Although global criteria, in the long run, play a more important role in the development and acceptance of scientific claims, they are much more difficult to specify than the local criteria. Indeed, until the recent efforts to devise objective procedures for summarizing the trends in a body of research, the assessment of global criteria was an informal and subjective matter. Despite the development of tools for combining probabilities over sets of experiments and for performing meta-analyses, the assessment of the impact of a set of experiments, as opposed to the evaluation of the quality of a single experiment, is still a highly subjective and idiosyncratic activity.

An unfortunate consequence of the fact that we lack standardized, objective methods for assessing the impact of a series of experiments is that different reviewers can and do draw opposite conclusions from the same body of research. Consequently we should not be surprised to discover that on the very few occasions when parapsychologists and critics have tried to evaluate the same set of experiments they have disagreed sharply on the conclusions to be drawn.

The interrelationship between local and global criteria has played an interesting role in the debates between critics and parapsychologists. The emphasis on local criteria has led some critics such as Hansel [2] to challenge the parapsychologists to produce a single, fraud-proof experiment that could demonstrate psi. The idea that the existence of psi could be determined by a single experiment, no matter how faithfully it fulfilled all of the local criteria, is quite unrealistic, as the parapsychologists have been quick to point out. No field of scientific inquiry decides such important matters on the basis of a single 'critical' experiment. The key to scientific justification is the ability to produce consistent findings across a number of independent investigations. And this is, of course, a global criterion.

On the other hand, many parapsychologists have resurrected a version of the ancient faggot theory by asserting that global criteria could somehow compensate for inadequacies in local criteria [9]. When recent critics, both from within as well as without the parapsychological community, pointed out serious flaws at the local level in some recent parapsychological experiments, the response was interesting. Although the existence of the flaws was not questioned, the defenders of the experiments argued that the flaws were irrelevant because global criteria suggested that they were not responsible for the successful outcomes. In other words, the existence of deficiencies in individual experiments could be dismissed as irrelevant because of overall patterns which cut across all the experiments. Whether there is any sense in which flaws in individual experiments can be compensated for by systematic patterns in the total set of experiments is highly questionable. At any rate we can be sure that the scientific community is not going to accept the reality of paranormal phenomena on the basis of experiments which are individually flawed.

What can we conclude from the best psi experiments?

Given these preliminary observations on some of the issues involved in evaluating the current status of parapsychological research, what conclusions about psi, if any, are justified on the basis of the best contemporary parapsychological research? Fortunately, there seems to be a consensus in the parapsychological community as to which research programs currently display the strongest case for psi. In this respect, three programs stand out. The experiments on remote viewing, which began in 1972, have received the most publicity outside of the parapsychological community [15]. Within the parapsychological community the so-called ganzfeld experiments have been considered most evidential. Finally, the experiments using random number generators, especially those involving psychokinesis, have been very influential.

My comments on the remote viewing experiments will be brief because Christopher Scott provides a detailed methodological critique in his contribution to this multi-author review. I will also be brief with my comments on the experiments with random number generators because, as will be noted, I can rely upon the systematic survey of a parapsychologist.

The principal basis for my conclusions will be the only two relatively systematic and objective critical surveys of contemporary parapsychological research that are known to me. One was my own attempt to assess the quality of the experiments in the ganzfeld research program [4]. The other is Akers' systematic evaluation of 54 of the best parapsychological experiments which yielded successful outcomes [1].

Remote viewing

As Christopher Scott's critique indicates, the remote viewing experiments have serious weaknesses. Both the experiments with random number generators and the ganzfeld experiments also are characterized by methodological deficiencies. Unlike these latter two cases, however, the deficiences in the remote viewing experiments provide a highly plausible and normal alternative explanation for the results. The flaws in

the ganzfeld and random number generator experiments violate accepted standards for parapsychological research. But they do not necessarily supply a plausible alternative account of the findings.

Just about all the serious weaknesses with the remote viewing experiments can be attributed to two important aspects of the procedure. In each experiment, a single percipient or subject supplies all the data for the sequence of trials. A single trial consists of the percipient, who is isolated with an experimenter, attempting to describe the target site which is being visited by one or more members of a target team. Neither the percipient nor the experimenter who is with him or her has any normal means of knowing the particular target site at the time the description is provided. Immediately after each trial, the target team returns to the laboratory and then returns to the target site with the percipient. This gives the percipient immediate feedback concerning the target side for that trial.

In addition to the feedback after each trial, the remote viewing experiments are characterized by a lack of independence among the successive trials. This comes about because after the experiment is completed a judge is given a set of all the target descriptions and a list of all the targets. These have previously been randomized so that the judge supposedly has no normal clues as to which description goes with which target site. The judge visits each target site in turn. At each site, he rank orders the entire set of descriptions from best to worst in terms of how well each apparently describes the particular target site. It is these rankings which serve as the primary data for the later statistical analyses upon which the conclusions are based.

The combination of the immediate feedback and the interdependence of the trial descriptions creates a fatal flaw in the procedure. It is this flaw that has allowed the sort of overt cuing in the transcripts discussed by Marks and Kammann [8]. This same flaw also produced overly optimistic statistical analyses in the early experiments. And it is this flaw which, I have repeatedly argued, enables the possibility of more pervasive cuing which cannot be corrected by any amount of post hoc analyses or editing.

The underlying flaw that I am talking about involves the lack of independence among the trials. This concept of independence is highly technical and I find that it is not easy to explain it to the uninitiated. However, I think I can provide an example which might make some of its implications evident. Imagine that we are dealing with a remote viewing experiment that consists only of three trials. Three trials would ordinarily be too few for a meaningful statistical analysis, but it is much less complicated to discuss. The basic principles are the same when we add more trials. Now further imagine that the experiment varies from the standard procedure for remote viewing experiments in that the judge does not rank all three descriptions against each target site. Instead, for each trial, the experimenters generate a separate pool of three possible sites, one of which is randomly selected as the actual target. Instead of ranking all three descriptions against each site, the judge visits all the sites in a given target pool and rank orders the three sites in terms of how well each matches the description for that trial.

In this latter procedure, which is the more typical of parapsychological experiments using free responses, there is a one-third chance that the actual target site will receive the top ranking for each trial. Given three trials, this would mean that there would be one chance in 27 of having the actual target site correctly ranked for each trial. Such an outcome would be suggestive that something other than just chance provided the successful matchings.

Now compare the procedure in the preceding two paragraphs with the one that is actually used in remote viewing experiments. Here the judge visits each target site and ranks all three descriptions against each target site. Because the concept of independence can be subtle, the investigator may tend to treat the statistical analysis the same as I did for the preceding method. If the corresponding description turned out to be ranked one for each of the targets, he might calculate the chance level as again one in 27 on the grounds that there was just one chance in three for the description to be ranked first for each of the three sites.

But the probability of correctly ranking all three descriptions is much greater than 1 in 27. This is because of the lack of independence in the judging procedure. Almost certainly the judge will not rank the same description as first for more than one target. Indeed, if the judge believes, as he should, that each description goes with a single target, he is apt to make his judgements accordingly. Because the number of possible rankings is much more restricted in the latter case, the probability of getting all three correct is probably closer to 1 in 6 rather than 1 in 27. In fact, the early remote viewing experiments provided overly optimistic estimates of the significance of the results because they failed to realize the lack of independence among the trials. Kennedy [7] noted that 7 out of the eight earliest remote viewing experiments were reported as yielding significant outcomes when the experimenters used a statistical test which assumed independence among trials. When Kennedy applied a more conservative test that took lack of independence into account the number of significant outcomes was reduced from 7 to 3.

The lack of independence among the trials can be circumvented by using a more conservative statistical test as Kennedy has demonstrated. In theory it might be possible to compensate for the overt clues in the transcripts of the type that Marks and Kammann [8] uncovered by careful editing. In practice it is not clear that this can be accomplished to the satisfaction of all critics. However, as I have repeatedly argued, the possibility for non-paranormal matching of descriptions to target sites is an ineradicable feature of the way the remote viewing experiments are designed.

The combination of immediate feedback and dependence among the trials makes it plausible that the percipient's descriptions will contain sufficient information, for entirely normal reasons, to enable the judge to match descriptions to target sites without error. To see how this could come about, imagine that the first target site was the Hoover Tower on the Stanford University campus. Immediately after having provided her description, the percipient is taken to the target site. At her second trial, it is reasonable to assume that the percipient will avoid describing anything that would specifically match the Hoover Tower. Assume that the second target site is the Palo Alto train station. On the third trial it is highly unlikely that the percipient will include in her description anything that directly matches either the Hoover Tower or the Palo Alto train station. Each successive description can be expected to lack descriptors that uniquely match any of the preceding target sites. In this way, it is easy to believe that the transcripts could be matched successfully to the target sites without having to assume paranormal powers. This flaw is fatal because I do not see anyway of compensating for its effects as long as the experiments continue to be run in the present way. [Against this argument, however, is the fact that well-controlled experiments which allow feedback and interdependence of judge's rankings, but which prevent sensory leakage to judges of target order or transcript positioning have not in practice yielded statistically significant results. – Review Coordinator.]

Experiments with random number generators

Schmidt, a quantum physicist by training, has probably been the individual most responsible for making the experiments with random number generators (RNGs) a major paradigm

in contemporary parapsychology[14]. Schmidt began his experiments in 1969 and has continued doing them ever since. At first, the RNG's were used in ESP type experiments. But quickly the use of RNGs in psychokinesis (PK) experiments became more common. The typical RNG randomly generates a sequence of binary outputs such as 0's and 1's. If the RNG is unbiased, the successive outputs will be independent of one another and the 0's will tend to occur 50% of the time. In a PK experiment, the subject or 'operator' sits before the RNG and tries to mentally influence the output so that the frequency of '1's will either become greater than or less than the expected 50%. During the past 15 years, in fact, many parapsychologists have reported results which they believe prove that operators can mentally bias the output of RNG devices.

The two most influential programs of research on this type of PK, together, account for approximately 60% of all the known experiments with RNGs. In addition to Schmidt's research program, Jahn began a program in the late 1970s, when he was the Dean of the School of Engineering/Applied Science at Princeton University. The research program initiated by these two physical scientists are not only among the most highly respected ones in parapsychology, but are also the most consistently successful in achieving significant outcomes.

When trying to assess the strength of the evidence for PK from such experiments, we should realize that the purported effects are extremely small. Over the years, Schmidt's subjects have 'produced' an average of 50.5% hits in comparison with the expected chance level of 50%. Jahn and his colleagues have come up with an even lower rate of success than Schmidt's. In their 'formal' series of 78 million trials, the percentage of hits in the intended direction was only 50.02%. Such extremely weak effects can yield strong statistical significance when considered over several million trials. But even an extremely weak PK effect, if real, violates some currently accepted principles about the physical universe. An extremely weak effect, one which takes millions of trials to document, may also suggest unknown, but very small biases, which only emerge when we average over an enormous number of trials. The statistical tests that are typically used in scientific experiments, for example, require certain assumptions about how the world operates. A typical experiment in biology or psychology, for example, may use fewer than 100 trials, or sometimes a few hundred trials, but rarely anything much higher. Under these conditions, empirical and theoretical studies by statisticians have found that the statistical tests in use are reasonably 'robust'. By this term, they mean that even though the assumptions underlying the statistical tests do not strictly hold, the results of the tests are still reasonably accurate. In other words even though the underlying distributions are rarely strictly normal and the population variances are not always equal, such departures from the ideal rarely make a serious difference in the interpretation of the statistical tests. But the situation could be quite different when we jump from experiments with relatively large effects and a small number of trials to experiments with extremely small effects and millions of trials. So far as we know, even the slightest departure from the assumptions might suffice to produce artificially significant outcomes.

Such considerations indicate that we have to be even more vigilant to insure that any evidence for PK effects on RNGs was obtained under conditions that stringently adhere to the standards of good scientific practice. Unfortunately, as was the case for remote viewing, the experiments using RNGs have typically been conducted under rather casual and poorly documented circumstances.

By ordinary standards, the accumulated results from RNG experiments seem impressive. A recent survey counted 332 separate 'experiments' published during the years 1969 through 1984[13]. Approximately 57% of these 332 experiments had been reported in refereed journals or conference proceedings. If we assume that these latter experiments have some claim to scientific status, then around 31% of the scientifically reported RNG experiments produced significant results as compared to the 5% that would be expected by chance. (For the entire set of 332 experiments, 21% produced significant results at the traditional 0.05 level.)

Such a success rate suggests that the successful outcomes probably cannot be entirely attributed to the reporting of only successful experiments. Although it does look as if successful experiments in this area are somewhat more likely to be reported than unsuccessful ones, the probability that all the apparent significance could have come about in this manner is quite low.

On the other hand, the use of these results to argue for the reality of PK depends upon how successfully other alternative explanations have been eliminated. The most obvious alternatives have to do with possible biases in outputs of the RNGs or of other factors such as temperture, humidity, and transients upon the outputs. The quality of a RNG experiment has to be judged by how well the investigator has insured that such alternatives have been successfully eliminated.

As May and his colleagues make clear[10] the RNG experiments up to 1979 all are inadequate on one or more key details such as controls for transients, tests of randomness, supervision of the operator, and documentation of procedures. As far as can be determined, this situation has not improved since that report. Today, the most impressive body of research on RNG's both in quantity and apparent quality of the experimentation, is that of Jahn and his colleagues[11]. Jahn's research presents the usual problem that an outsider has in trying to reach a fair conclusion concerning the implications of the reported results. The data base includes an accumulation of trials from a small number of operators over a period of more than six years. The procedures and instrumentation used in the Princeton Anomalies Laboratory are unique. They differ in many ways from those used in other parapsychological laboratories. Indeed, these procedures and technologies have changed over time during the experiments. Presumably, the earlier trials were conducted under conditions which were relatively more informal and less fail-safe than those now currently employed. But it is not clear if the data for the various conditions can be separated and analyzed separately. This problem is especially critical because of the extremely small size of the effect being claimed.

Another problem is that the various operators do not contribute equally to the data base. One subject, for example, is apparently responsible for 23% of the total data base. This subject's hit rate was 50.05%. If we eliminated that subject's data from the data base, the remaining data yield a hit rate of 50.01% which is not only extremely close to 50%, but is no longer significantly different from chance.

Given the implications of the claims being made, it becomes imperative to provide data from these experiments in which the operators have been adequately monitored and the specific experimental arrangements have been consistent throughout. If such experimentally adequate trials continue to yield positive and significant results, then we would want to see replication in independent laboratories.

As is the case with the remote viewing experiments, the RNG experiments, upon close inspection, suffer from a variety of defects at both the local and global level. In the case of remote viewing, the flaws are such that they suggest a plausible, normal alternative to account for the results. We do not have such an obvious plausible alternative to account for the results of the RNG experiments. Most of the experiments lack adequate tests of the randomness of the RNG being

used. But, in most such experiments, it is not clear how any systematic biases could account for the results since such systematic biases presumably would affect control and experimental trials in the same way. Similar problems face the critic who wants to argue that the specific flaws in a given RNG experiment could, in fact, have accounted for the results.

On the other hand, the critic can rightfully insist that is is not his or her responsibility to provide an alternative account. It is sufficient to point out that the given experiment violates one or more criteria that the parapsychologist themselves, have asserted are necessary for a sound parapsychological experiment. Yes, we do not have an obvious account of how inadequate randomization could have produced the reported results. On the other hand, we have no way of assessing the meaning of the reported results unless we can be sure that the reported statistical levels of significance are correct. And we can only have confidence in the statistical conclusions when we know that the underlying assumptions for their interpretation have been met. Yet this is just what we do not know. In other words, the only reasonable conclusion to draw from the existing body of data from the RNG experiments is that we do not know what to make of it. We cannot say that the results were definitely the result of some artifact. On the other hand we cannot say that they were not. The only way we will be able to draw meaningful conclusions from such experiments is when they have been conducted according to the standards that both the parapsychologists and their critics assert ought to be met by any acceptable parapsychological experiment.

The ganzfeld psi experiments

The ganzfeld psi experiments, which began at roughly the same time as the remote viewing experiments, seemingly have produced a stronger case for psi. Because of this, I undertook a careful and systematic review of this data base [4].

The ganzfeld psi experiments are named after the term 'ganzfeld' which was used by Gestalt psychologists to designate the entire or whole visual field. For theoretical purposes, the Gestalt psychologists wanted to create a situation in which the subject or observer could view a homogeneous visual field, one with no imperfections or boundaries. They called such a field the 'ganzfeld'. Later psychologists discovered that when individuals are put into a ganzfeld situation they tend to quickly experience an altered state of consciousness.

In the early 1970s, some parapsychologists decided that the use of the ganzfeld would provide a relatively safe and easy way to create an altered state in their experimental subjects. They believed that such a state was more conducive towards picking up the elusive psi signals. In a typical psi ganzfeld experiment, the subject (or percipient) has halved ping-pong balls taped over his eyes. He then reclines in a comfortable chair while white noise occurs in the earphones attached to his head. A bright light shines in front of his face. When seen through the translucent ping pong balls, the light is experienced as a homogeneous fog-like field. When so prepared, almost all subjects report experiencing a pleasant altered state within 15 minutes.

While an experimenter is preparing the subject for the ganzfeld state, a second experimenter randomly selects a target pool from a large set. The target pool typically consists of four possible targets, usually reproductions of paintings or pictures of travel scenes. One of these four candidates is randomly chosen to be the target for that trial. The target is given to an agent or sender who tries to psychically communicate its substance to the subject who is in the ganzfeld state. After a designated period, the subject is removed from the ganzfeld state and presented with the four candidates from the target pool. The subject then ranks the four candidates in terms of how well each matched the experience of her ganzfeld period. If the actual target is ranked first, the trial is designated as a 'hit'. An actual experiment consists of several trials. In the example, the probability is that one out of every four trials will produce a hit. If the number of hits significantly exceeds the expected 25%, then the result is considered to be evidence for the existence of psi.

What sorts of flaws did I find in this data base [4]? All but three of the 42 experiments used multiple analyses which artificially inflated the chances of obtaining 'significant' outcomes. Only 11 (26%) of the studies contained evidence of having adequately randomized the target selections. As many as 15 (36%) clearly used inferior randomization such as hand shuffling or no randomization at all. The remaining 16 experiments did not supply sufficient information on how they had chosen the targets. As many as 23 of the experiments (55%) used only one target pool which meant that the subject was handed for judging the very same target that the percipient had used. This allowed for the possibility of sensory cuing. Although the argument for psi is mainly a statistical one, the reports of 12 experiments (29%) revealed statistical errors. A variety of other departures from optimum practice were also found.

Honorton's rebuttal questioned many of my flaw assignments, provided a re-analysis of the data base which he claimed overcame many of the statistical weaknesses of the original experiments, and argued that the existing flaws were not sufficient to have accounted for the findings [3]. He does not deny that the experiments departed from optimal design, but he argues that such departures were insufficient to have accounted for the results.

Honorton and I subsequently published a joint paper to emphasize those points on which we agree [6]. Some of the key points of agreement were expressed by us in the following words:

"As to the psi ganzfeld data base, we agree, as our earlier exchanges indicate..., that the experiments as a group departed from ideal standards on aspects such as multiple testing, randomization of targets, controlling for sensory leakage, application of statistical tests, and documentation. Although we probably still differ about the extent and seriousness of these departures, we agree that future psi ganzfeld experiments should be conducted in accordance with these ideals...

"Although we probably still differ on the magnitude of the biases contributed by multiple testing, retrospective experiments, and the file-drawer problem, we agree that the overall significance observed in these studies cannot reasonably be explained by these selective factors. Something beyond selective reporting or inflated significance levels seems to be producing the nonchance outcomes. Moreover, we agree that the significant outcomes have been produced by a number of different investigators.

"Whereas we continue to differ over the degree to which the current ganzfeld data base contributes evidence for psi, we agree that the final verdict awaits the outcome of future psi ganzfeld experiments – ones conducted by a broader range of investigators and according to more stringent standards."

Akers's critique

The systematic evaluation of the contemporary parapsychological literature by Akers [1], a former parapsychologist, is interesting because it used a strategy different from mine. My approach was to evaluate the entire data base of a single research paradigm, including both successful and unsuccessful outcomes. Akers confined himself to the best contemporary ESP experiments which had produced significant results with unselected subjects. I assigned flaws to experiments

without regard for whether each flaw, by itself, could have caused the observed outcome. Akers was more conservative. He charged a flaw to a study only if he thought it could plausibly have been sufficient to produce the observed result. Akers chose a sample of 54 parapsychological experiments from areas of research which had been previously reviewed by one of two major parapsychologists, Honorton or Palmer. The intent was to choose experiments which could be viewed as the currently best evidence for the existence of psi. Akers then evaluated the experiments in terms of a number of possible flaws. An experiment was eliminated as evidential only if it was so seriously defective on a given flaw that the defect could plausibly have caused the observed outcome.
As a result of this exercise, Akers concluded:
"Results from the 54-experiment survey have demonstrated that there are many alternative explanations for ESP phenomena; the choice is not simply between psi and experimenter fraud... The number of experiments flawed on various grounds were as follows: randomization failures (13), sensory leakage (22), subject cheating (12), recording errors (10), classification or scoring errors (9), statistical errors (12), reporting failures (10)... All told, 85% of the experiments were considered flawed (46/54). This leaves eight experiments where no flaws were assigned... Although none of these experiments has a glaring weakness, this does not mean that they are especially strong in either their methods or their results... In conclusion, there were eight experiments conducted with reasonable care, but none of these could be considered as methodologically strong. When all 54 experiments are considered, it can be stated that the research methods are too weak to establish the existence of a paranormal phenomenon" [1].

Conclusions on the scientific evidence

The parapsychologists whom we have cited as well as their critics agree that the best contemporary experiments in parapsychology fall short of acceptable methodological standards. The critics conclude that such data, based on methodologically flawed procedures, cannot justify any conclusions about psi. The parapsychologists argue that, while each experiment is individually flawed, when taken together, the composite body of data suffices to justify that psi exists.
We should distinguish between three degrees of criticism of a given parapsychological finding. The first is what we might refer to as the 'Smoking Gun'. This is the type of criticism that asserts or strongly implies that the observed findings were due to factor X. Such a claim puts the burden of proof on the critic. To back up such a claim, the critic must provide evidence that the results were in fact caused by X. Many of the bitterly contested feuds between critics and proponents have often been the result of the proponent correctly or incorrectly assuming that this type of criticism was being made.
The second type of criticism can be termed the 'Plausible Alternative'. Here the critic does not assert that the result was due to factor X. Instead, he asserts that the obtained result *could have been* due to factor X. Such a stance also places a burden on the critic, but one not so stringent as the 'Smoking Gun' assertion. The critic now has to make a plausible case for the possibility that factor X was sufficient to have caused the result. Optional stopping, for example, can bias the results. But bias is a small one and it would be a mistake to assert that an outcome was due to optional stopping if the probability of the outcome is extremely low. Akers's critique [1], which was previously discussed, is an example based on the plausible alternative.
The third type of criticism can be called the 'Dirty Test Tube'. In this case the critic points out that the experiment departs from accepted methodological standards of the field in certain ways. The critic, in this case, does not claim that he knows that the results have been produced by some artifact. Rather, he points out that the results have been obtained under conditions which fail to meet generally accepted standards. The strength of this latter type of critique is that test tubes should not be dirty when doing careful and important scientific research. To the extent that the test tubes were dirty, it suggests that the experiment was not carried out according to acceptable standards. Consequently, the results remain suspect even though the critic cannot demonstrate that the dirt in the test tubes was sufficient to have produced the outcome [4].
It is in this latter sense, the Dirty Test Tube sense, that the best parapsychological experiments fall short. To be fair, we do not have a 'smoking gun', nor have we demonstrated a plausible alternative. But we suspect that even the parapsychological community must be concerned to discover that their best experiments still fall far short of the methodological adequacy which they themselves would profess.
Under the circumstances, it seems reasonable to withhold judgment at this time. But can we not make at least a provisional conclusion based on the tendencies in the data that have been accumulated up to now? We would argue that it would be counterproductive to even try to draw a tentative conclusion because the data lack robustness. By this term, we mean that even relatively small changes in the data base are capable of reversing any conclusion we might wish to make. For example, Honorton [3] and I [4] differed on whether to assign a flaw on randomization to Sargent's experiments. With Honorton's assignment, the studies with adequate randomization do not differ in significance of outcome from those with inadequate randomization. On Hyman's assignment, the experiments with inadequate randomization significantly have more successful outcomes than do those with inadequate randomization. A simple disagreement on one experimenter's research can thus make a huge difference as to whether we conclude that this flaw contributed or did not contribute to the observed outcomes. Several other similar examples could be cited to illustrate the extreme sensitivity of this data base to slight changes in flaw assignments.

1 Akers, C., Methodological criticisms of parapsychology, in: Advances in Parapsychological Research, vol. 4. Ed. S. Krippner. McFarland, Jefferson, N.C. 1984.
2 Hansel, C. E. M., ESP and parapsychology: A critical re-evaluation. Prometheus Books, Buffalo, N.Y. 1980.
3 Honorton, C., Meta-analysis of psi ganzfeld research: a response to Hyman. J. Parapsychol. *49* (1985) 351–364.
4 Hyman, R., The ganzfeld psi experiment: a critical appraisal. J. Parapsychol. *49* (1985) 3–49.
5 Hyman, R., A critical historical overview of parapsychology, in: A Skeptic's Handbook of Parapsychology, pp. 3–96. Ed. P. Kurtz. Prometheus Books, Buffalo, N.Y. 1985.
6 Hyman, R., and Honorton, C., A joint communique: the psi ganzfeld controversy. J. Parapsychol. *50* (1986) 351–364.
7 Kennedy, J. E., Methodological problems in free-response ESP experiments. J. Am. Soc. psych. Res. *73* (1979) 1–15.
8 Marks, D. F., and Kamman, R., Information transmission in remote viewing. Nature *274* (1978) 680–681.
9 Marks, D. F., Investigating the paranormal. Nature *320* (1986) 119–124.
10 May, E. C., Humphrey, B. S., and Hubbard, G. S., Electronic System Perturbation, Techniques. SRI International (Final Report) Menlo Park, CA 1980.
11 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, Princeton University 1984.
12 Palmer, J., In defense of parapsychology: a reply to James E. Alcock. Zetetic Scholar (1983) 39–70.

13 Radin, D. I., May, E. C., and Thomson, M. J., Psi experiments with random number generators: meta-analysis part 1. SRI International, Menlo Park, CA 1985.
14 Schmidt, H., A PK test with electronic equipment. J. Parapsych. *34* (1970) 175–181.
15 Targ, R., and Puthoff, H. E., Information transmission under conditions of sensory shielding. Nature *252* (1974) 602–607.
16 The Society for Psychical Research: objects of the society. Proc. Soc. psych. Res. *1* (1882–83) 3–6.

0014-4754/88/040315-08$1.50 + 0.20/0

Remote viewing

C. Scott

60 Highgate Hill, London N19 5NQ (England)

Summary. Remote viewing is the supposed faculty which enables a percipient, sited in a closed room, to describe the perceptions of a remote agent visiting an unknown target site. To provide convincing demonstration of such a faculty poses a range of experimental and practical problems, especially if feedback to the percipient is allowed after each trial. The precautions needed are elaborate and troublesome; many potential loopholes have to be plugged and there will be strong temptations to relax standards, requiring exceptional discipline and dedication by the experimenters. Most reports of remote viewing experiments are rather superficial and do not permit assessment of the experimental procedures with confidence; in many cases there is clear evidence of particular loopholes left unclosed. Any serious appraisal of the evidence would have to go beyond the reports. Meanwhile the published evidence is far from compelling, and certainly insufficient to justify overthrow of well-established scientific principles.
Key words. Remote viewing; ESP; feedback; data selection; bias; fraud; statistics; methodology.

Introduction

Parapsychology has its fashions and, like all fashions, they tend to move in cycles. The hypnotists of the late 19th century devoted much attention to a phenomenon they called 'travelling clairvoyance': the hypnotic subject was asked to send his mind to a distant place and describe what he saw. It was widely believed that the state of hypnosis conferred extrasensory powers and this was said to be demonstrated when the subject reported veridical information about the location visited which he could not have known by normal means[4].

In the mid-1970s this idea, without the hypnosis, was resuscitated by two California research workers, Targ and Puthoff, under the name of 'remote viewing' (sometimes called 'remote perception'). Targ and Puthoff[18, 19] put their subject, or 'percipient', in their laboratory with a tape recorder while an 'agent' was sent to one or more randomly selected locations. At a pre-arranged time the agent viewed the chosen location and the percipient recorded his impressions. The recordings were transcribed, the experiment was repeated for many different locations, and judges were presented with the set of transcripts together with the list of locations, and asked to match one against the other. Striking successes were claimed. In a short time remote viewing jumped to the forefront of the armoury of methods used by parapsychologists to demonstrate ESP.

New fashions do not arise simply by chance. One of the attractions of remote viewing was its fund-raising potentiality. If a percipient sitting in a laboratory in California can send his mind anywhere in the world, then obviously he can send it also to Moscow and visit the most secret files in the Kremlin. Some researchers were not slow to point this out to the Department of Defense where the thought duly stimulated the issuance of some lucrative research contracts.

Another practical advantage of the method is the ease with which one can demonstrate face validity. A single transcript typically includes scores of statements, while on the other side a target location contains a great number of objects, shapes, sights, sounds and associations. The opportunity to find a striking correspondence somewhere among such a mass of material is large – and more important, much larger than it seems to be to the casual inquirer. Marks and Kammann[9] describe the phenomenon they call 'subjective validation', which leads the observer to seek correspondences between the transcript and the target and when he finds them (as, almost inevitably, he will) to overlook the size of the parent set of comparisons from which the successful correspondences have been selected. Unlike their Victorian forbears, modern parapsychologists are not so naive as to believe that the mere citation of a selection of striking resemblances between statements in the transcript and features of the target provides an adequate demonstration of the reality of remote viewing. They admit that, for this purpose, only a blind matching method is scientifically acceptable. Nevertheless, once this admission has been made, they feel entitled to go ahead with just such a presentation of selected correspondences by way of examples. In doing so, no author has yet considered it necessary to attempt to give an estimate of the size of the parent set of comparisons. The effect on the reader (or the viewer – see particularly the BBC's *Horizon* programme of 1983, 'The Case of ESP') is overwhelming: taken out of context, the hits appear to outstrip any possibility of chance coincidence. Thus the face validity of the remote viewing technique is a valuable feature for the experimenter wishing to gain the support of a non-specialist audience – a category which generally includes TV producers and some Defense Department project evaluators. It also helps to sustain the morale of those working on the project, in sharp contrast to the older card-guessing techniques which are notoriously boring for those involved.

Setting aside the anecdotal presentation of selected hits, how exactly can the reality of remote viewing be demonstrated?

Selection and definition of targets

First, it is clear that the agent must not be free to choose the target location according to his whim. It has been known to psychologists for a long time that seemingly random choices are in reality influenced by a variety of response preferences which may be shared by many or all respondents. If targets were chosen at the whim of the agent, any coincidence observed between the target and the subject's 'perception' could be attributed to common response preferences be-