Statistical Science
1991, Vol. 6, No. 4, 363–403

Replication and Meta-Analysis in Parapsychology

Jessica Utts

Abstract. Parapsychology, the laboratory study of psychic phenomena, has had its history interwoven with that of statistics. Many of the controversies in parapsychology have focused on statistical issues, and statistical models have played an integral role in the experimental work. Recently, parapsychologists have been using meta-analysis as a tool for synthesizing large bodies of work. This paper presents an overview of the use of statistics in parapsychology and offers a summary of the meta-analyses that have been conducted. It begins with some anecdotal information about the involvement of statistics and statisticians with the early history of parapsychology. Next, it is argued that most nonstatisticians do not appreciate the connection between power and "successful" replication of experimental effects. Returning to parapsychology, a particular experimental regime is examined by summarizing an extended debate over the interpretation of the results. A new set of experiments designed to resolve the debate is then reviewed. Finally, meta-analyses from several areas of parapsychology are summarized. It is concluded that the overall evidence indicates that there is an anomalous effect in need of an explanation.

Key words and phrases: Effect size, psychic research, statistical controversies, randomness, vote-counting.

1. INTRODUCTION

In a June 1990 Gallup Poll, 49% of the 1236 respondents claimed to believe in extrasensory perception (ESP), and one in four claimed to have had a personal experience involving telepathy (Gallup and Newport, 1991). Other surveys have shown even higher percentages; the University of Chicago's National Opinion Research Center recently surveyed 1473 adults, of which 67% claimed that they had experienced ESP (Greeley, 1987).

Public opinion is a poor arbiter of science, however, and experience is a poor substitute for the scientific method. For more than a century, small numbers of scientists have been conducting laboratory experiments to study phenomena such as telepathy, clairvoyance and precognition, collectively known as "psi" abilities. This paper will examine some of that work, as well as some of the statistical controversies it has generated.

Jessica Utts is Associate Professor, Division of Statistics, University of California at Davis, 469 Kerr Hall, Davis, California 95616.

Parapsychology, as this field is called, has been a source of controversy throughout its history. Strong beliefs tend to be resistant to change even in the face of data, and many people, scientists included, seem to have made up their minds on the question without examining any empirical data at all. A critic of parapsychology recently acknowledged that "The level of the debate during the past 130 years has been an embarrassment for anyone who would like to believe that scholars and scientists adhere to standards of rationality and fair play" (Hyman, 1985a, page 89). While much of the controversy has focused on poor experimental design and potential fraud, there have been attacks and defenses of the statistical methods as well, sometimes calling into question the very foundations of probability and statistical inference.

Most of the criticisms have been leveled by psychologists. For example, a 1988 report of the U.S. National Academy of Sciences concluded that "The committee finds no scientific justification from research conducted over a period of 130 years for the existence of parapsychological phenomena" (Druckman and Swets, 1988, page 22). The chapter on parapsychology was written by a subcommittee

chaired by a psychologist who had published a similar conclusion prior to his appointment to the committee (Hyman, 1985a, page 7). There were no parapsychologists involved with the writing of the report. Resulting accusations of bias (Palmer, Honorton and Utts, 1989) led U.S. Senator Claiborne Pell to request that the Congressional Office of Technology Assessment (OTA) conduct an investigation with a more balanced group. A one-day workshop was held on September 30, 1988, bringing together parapsychologists, critics and experts in some related fields (including the author of this paper). The report concluded that parapsychology needs "a fairer hearing across a broader spectrum of the scientific community, so that emotionality does not impede objective assessment of experimental results" (Office of Technology Assessment, 1989).

It is in the spirit of the OTA report that this article is written. After Section 2, which offers an anecdotal account of the role of statisticians and statistics in parapsychology, the discussion turns to the more general question of replication of experimental results. Section 3 illustrates how replication has been (mis)interpreted by scientists in many fields. Returning to parapsychology in Section 4, a particular experimental regime called the "ganzfeld" is described, and an extended debate about the interpretation of the experimental results is discussed. Section 5 examines a meta-analysis of recent ganzfeld experiments designed to resolve the debate. Finally, Section 6 contains a brief account of meta-analyses that have been conducted in other areas of parapsychology, and conclusions are given in Section 7.

2. STATISTICS AND PARAPSYCHOLOGY

Parapsychology had its beginnings in the investigation of purported mediums and other anecdotal claims in the late 19th century. The Society for Psychical Research was founded in Britain in 1882, and its American counterpart was founded in Boston in 1884. While these organizations and their members were primarily involved with investigating anecdotal material, a few of the early researchers were already conducting "forced-choice" experiments such as card-guessing. (Forced-choice experiments are like multiple choice tests; on each trial the subject must guess from a small, known set of possibilities.) Notable among these was Nobel Laureate Charles Richet, who is generally credited with being the first to recognize that probability theory could be applied to card-guessing experiments (Rhine, 1977, page 26; Richet, 1884).

F. Y. Edgeworth, partly in response to what he considered to be incorrect analyses of these experiments, offered one of the earliest treatises on the statistical evaluation of forced-choice experiments in two articles published in the *Proceedings of the Society for Psychical Research* (Edgeworth, 1885, 1886). Unfortunately, as noted by Mauskopf and McVaugh (1979) in their historical account of the period, Edgeworth's papers were "perhaps too difficult for their immediate audience" (page 105).

Edgeworth began his analysis by using Bayes' theorem to derive the formula for the posterior probability that chance was operating, given the data. He then continued with an argument "savouring more of Bernoulli than Bayes" in which "it is consonant, I submit, to experience, to put 1/2 both for α and β," that is, for both the prior probability that chance alone was operating, and the prior probability that "there should have been some additional agency." He then reasoned (using a Taylor series expansion of the posterior probability formula) that if there were a large probability of observing the data given that some additional agency was at work, and a small objective probability of the data under chance, then the latter (binomial) probability "may be taken as a rough measure of the sought *a posteriori* probability in favour of mere chance" (page 195). Edgeworth concluded his article by applying his method to some data published previously in the same journal. He found the probability against chance to be 0.99996, which he said "may fairly be regarded as physical certainty" (page 199). He concluded:

> Such is the evidence which the calculus of probabilities affords as to the existence of an agency other than mere chance. The calculus is silent as to the nature of that agency—whether it is more likely to be vulgar illusion or extraordinary law. That is a question to be decided, not by formulae and figures, but by general philosophy and common sense [page 199].

Both the statistical arguments and the experimental controls in these early experiments were somewhat loose. For example, Edgeworth treated as binomial an experiment in which one person chose a string of eight letters and another attempted to guess the string. Since it has long been understood that people are poor random number (or letter) generators, there is no statistical basis for analyzing such an experiment. Nonetheless, Edgeworth and his contemporaries set the stage for the use of controlled experiments with statistical evaluation in laboratory parapsychology. An interesting historical account of Edgeworth's involvement and the role telepathy experiments played in the early history of randomization and experimental design is provided by Hacking (1988).

One of the first American researchers to use statistical methods in parapsychology was John Edgar Coover, who was the Thomas Welton Stanford Psychical Research Fellow in the Psychology Department at Stanford University from 1912 to 1937 (Dommeyer, 1975). In 1917, Coover published a large volume summarizing his work (Coover, 1917). Coover believed that his results were consistent with chance, but others have argued that Coover's definition of significance was too strict (Dommeyer, 1975). For example, in one evaluation of his telepathy experiments, Coover found a two-tailed p-value of 0.0062. He concluded, "Since this value, then, lies within the field of chance deviation, although the probability of its occurrence by chance is fairly low, it cannot be accepted as a decisive indication of some cause beyond chance which operated in favor of success in guessing" (Coover, 1917, page 82). On the next page, he made it explicit that he would require a p-value of 0.0000221 to declare that something other than chance was operating.

It was during the summer of 1930, with the card-guessing experiments of J. B. Rhine at Duke University, that parapsychology began to take hold as a laboratory science. Rhine's laboratory still exists under the name of the Foundation for Research on the Nature of Man, housed at the edge of the Duke University campus.

It wasn't long after Rhine published his first book, *Extrasensory Perception* in 1934, that the attacks on his methodology began. Since his claims were wholly based on statistical analyses of his experiments, the statistical methods were closely scrutinized by critics anxious to find a conventional explanation for Rhine's positive results.

The most persistent critic was a psychologist from McGill University named Chester Kellogg (Mauskopf and McVaugh, 1979). Kellogg's main argument was that Rhine was using the binomial distribution (and normal approximation) on a series of trials that were not independent. The experiments in question consisted of having a subject guess the order of a deck of 25 cards, with five each of five symbols, so technically Kellogg was correct.

By 1937, several mathematicians and statisticians had come to Rhine's aid. Mauskopf and McVaugh (1979) speculated that since statistics was itself a young discipline, "a number of statisticians were equally outraged by Kellogg, whose arguments they saw as discrediting *their* profession" (page 258). The major technical work, which acknowledged that Kellogg's criticisms were accurate but did little to change the significance of the results, was conducted by Charles Stuart and Joseph A. Greenwood and published in the first volume of the *Journal of Parapsychology* (Stuart and Greenwood, 1937). Stuart, who had been an undergraduate in mathematics at Duke, was one of Rhine's early subjects and continued to work with him as a researcher until Stuart's death in 1947. Greenwood was a Duke mathematician, who apparently converted to a statistician at the urging of Rhine.

Another prominent figure who was distressed with Kellogg's attack was E. V. Huntington, a mathematician at Harvard. After corresponding with Rhine, Huntington decided that, rather than further confuse the public with a technical reply to Kellogg's arguments, a simple statement should be made to the effect that the mathematical issues in Rhine's work had been resolved. Huntington must have successfully convinced his former student, Burton Camp of Wesleyan, that this was a wise approach. Camp was the 1937 President of IMS. When the annual meetings were held in December of 1937 (jointly with AMS and AAAS), Camp released a statement to the press that read:

> Dr. Rhine's investigations have two aspects: experimental and statistical. On the experimental side mathematicians, of course, have nothing to say. On the statistical side, however, recent mathematical work has established the fact that, assuming that the experiments have been properly performed, the statistical analysis is essentially valid. If the Rhine investigation is to be fairly attacked, it must be on other than mathematical grounds [Camp, 1937].

One statistician who did emerge as a critic was William Feller. In a talk at the Duke Mathematical Seminar on April 24, 1940, Feller raised three criticisms to Rhine's work (Feller, 1940). They had been raised before by others (and continue to be raised even today). The first was that inadequate shuffling of the cards resulted in additional information from one series to the next. The second was what is now known as the "file-drawer effect," namely, that if one combines the results of published studies only, there is sure to be a bias in favor of successful studies. The third was that the results were enhanced by the use of optional stopping, that is, by not specifying the number of trials in advance. All three of these criticisms were addressed in a rejoinder by Greenwood and Stuart (1940), but Feller was never convinced. Even in its third edition published in 1968, his book *An Introduction to Probability Theory and Its Applications* still contains his conclusion about Greenwood and Stuart: "Both their arithmetic and their experiments have a distinct tinge of the supernatural" (Feller, 1968, page 407). In his discussion of Feller's position, Diaconis (1978) remarked, "I believe

Feller was confused... he seemed to have decided the opposition was wrong and that was that."

Several statisticians have contributed to the literature in parapsychology to greater or lesser degrees. T. N. E. Greville developed applicable statistical methods for many of the experiments in parapsychology and was Statistical Editor of the *Journal of Parapsychology* (with J. A. Greenwood) from its start in 1937 through Volume 31 in 1967; Fisher (1924, 1929) addressed some specific problems in card-guessing experiments; Wilks (1965a, b) described various statistical methods for parapsychology; Lindley (1957) presented a Bayesian analysis of some parapsychology data; and Diaconis (1978) pointed out some problems with certain experiments and presented a method for analyzing experiments when feedback is given.

Occasionally, attacks on parapsychology have taken the form of attacks on statistical inference in general, at least as it is applied to real data. Spencer-Brown (1957) attempted to show that true randomness is impossible, at least in finite sequences, and that this could be the explanation for the results in parapsychology. That argument reemerged in a recent debate on the role of randomness in parapsychology, initiated by psychologist J. Barnard Gilmore (Gilmore, 1989, 1990; Utts, 1989; Palmer, 1989, 1990). Gilmore stated that "The agnostic statistician, advising on research in psi, should take account of the possible inappropriateness of classical inferential statistics" (1989, page 338). In his second paper, Gilmore reviewed several non-psi studies showing purportedly random systems that do not behave as they should under randomness (e.g., Iversen, Longcor, Mosteller, Gilbert and Youtz, 1971; Spencer-Brown, 1957). Gilmore concluded that "Anomalous data... should not be found nearly so often if classical statistics offers a valid model of reality" (1990, page 54), thus rejecting the use of classical statistical inference for real-world applications in general.

3. REPLICATION

Implicit and explicit in the literature on parapsychology is the assumption that, in order to truly establish itself, the field needs to find a repeatable experiment. For example, Diaconis (1978) started the summary of his article in *Science* with the words "In search of repeatable ESP experiments, modern investigators..." (page 131). On October 28–29, 1983, the 32nd International Conference of the Parapsychology Foundation was held in San Antonio, Texas, to address "The Repeatability Problem in Parapsychology." The Conference Proceedings (Shapin and Coly, 1985) reflect the diverse views among parapsychologists on the nature of the problem. Honorton (1985a) and Rao (1985), for example, both argued that strict replication is uncommon in *most* branches of science and that parapsychology should not be singled out as unique in this regard. Other authors expressed disappointment in the lack of a single repeatable experiment in parapsychology, with titles such as "Unrepeatability: Parapsychology's Only Finding" (Blackmore, 1985), and "Research Strategies for Dealing with Unstable Phenomena" (Beloff, 1985).

It has never been clear, however, just exactly what would constitute acceptable evidence of a repeatable experiment. In the early days of investigation, the major critics "insisted that it would be sufficient for Rhine and Soal to convince them of ESP if a parapsychologist could perform successfully a single 'fraud-proof' experiment" (Hyman, 1985a, page 71). However, as soon as well-designed experiments showing statistical significance emerged, the critics realized that a single experiment could be statistically significant just by chance. British psychologist C. E. M. Hansel quantified the new expectation, that the experiment should be repeated a few times, as follows:

> If a result is significant at the .01 level and this result is not due to chance but to information reaching the subject, it may be expected that by making two further sets of trials the antichance odds of one hundred to one will be increased to around a million to one, thus enabling the effects of ESP—or whatever is responsible for the original result—to manifest itself to such an extent that there will be little doubt that the result is not due to chance [Hansel, 1980, page 298].

In other words, three consecutive experiments at $p \leq 0.01$ would convince Hansel that something other than chance was at work.

This argument implies that if a particular experiment produces a statistically significant result, but subsequent replications fail to attain significance, then the original result was probably due to chance, or at least remains unconvincing. The problem with this line of reasoning is that there is no consideration given to sample size or power. Only an experiment with extremely high power should be expected to be "successful" three times in succession.

It is perhaps a failure of the way statistics is taught that many scientists do not understand the importance of power in defining successful replication. To illustrate this point, psychologists Tversky and Kahnemann (1982) distributed a questionnaire

to their colleagues at a professional meeting, with the question:

> An investigator has reported a result that you consider implausible. He ran 15 subjects, and reported a significant value, $t = 2.46$. Another investigator has attempted to duplicate his procedure, and he obtained a nonsignificant value of t with the same number of subjects. The direction was the same in both sets of data. You are reviewing the literature. What is the highest value of t in the second set of data that you would describe as a failure to replicate? [1982, page 28].

In reporting their results, Tversky and Kahnemann stated:

> The majority of our respondents regarded $t = 1.70$ as a failure to replicate. If the data of two such studies ($t = 2.46$ and $t = 1.70$) are pooled, the value of t for the combined data is about 3.00 (assuming equal variances). Thus, we are faced with a paradoxical state of affairs, in which the same data that would increase our confidence in the finding when viewed as part of the original study, shake our confidence when viewed as an independent study [1982, page 28].

At a recent presentation to the History and Philosophy of Science Seminar at the University of California at Davis, I asked the following question. Two scientists, Professors A and B, each have a theory they would like to demonstrate. Each plans to run a fixed number of Bernoulli trials and then test H_0: $p = 0.25$ versus H_a: $p > 0.25$. Professor A has access to large numbers of students each semester to use as subjects. In his first experiment, he runs 100 subjects, and there are 33 successes ($p = 0.04$, one-tailed). Knowing the importance of replication, Professor A runs an additional 100 subjects as a second experiment. He finds 36 successes ($p = 0.009$, one-tailed).

Professor B only teaches small classes. Each quarter, she runs an experiment on her students to test her theory. She carries out ten studies this way, with the results in Table 1.

I asked the audience by a show of hands to indicate whether or not they felt the scientists had successfully demonstrated their theories. Professor A's theory received overwhelming support, with approximately 20 votes, while Professor B's theory received only one vote.

If you aggregate the results of the experiments for each professor, you will notice that each conducted 200 trials, and Professor B actually demonstrated a *higher* level of success than Professor A, with 71 as opposed to 69 successful trials. The one-tailed p-values for the combined trials are 0.0017 for Professor A and 0.0006 for Professor B.

To address the question of replication more explicitly, I also posed the following scenario. In December of 1987, it was decided to prematurely terminate a study on the effects of aspirin in reducing heart attacks because the data were so convincing (see, e.g., Greenhouse and Greenhouse, 1988; Rosenthal, 1990a). The physician-subjects had been randomly assigned to take aspirin or a placebo. There were 104 heart attacks among the 11,037 subjects in the aspirin group, and 189 heart attacks among the 11,034 subjects in the placebo group (chi-square = 25.01, $p < 0.00001$).

After showing the results of that study, I presented the audience with two hypothetical experiments conducted to try to replicate the original result, with outcomes in Table 2.

I asked the audience to indicate which one they thought was a more successful replication. The audience chose the second one, as would most journal editors, because of the "significant p-value." In fact, the *first* replication has almost exactly the same proportion of heart attacks in the two groups as the original study and is thus a very close replication of that result. The second replication has

TABLE 1
Attempted replciations for professor B

n	Number of successes	One-tailed p-value
10	4	0.22
15	6	0.15
17	6	0.23
25	8	0.17
30	10	0.20
40	13	0.18
18	7	0.14
10	5	0.08
15	5	0.31
20	7	0.21

TABLE 2
Hypothetical replications of the aspirin / heart attack study

	Replication #1 Heart attack		Replication #2 Heart attack	
	Yes	No	Yes	No
Aspirin	11	1156	20	2314
Placebo	19	1090	48	2170
Chi-square	2.596, $p = 0.11$		13.206, $p = 0.0003$	

very *different* proportions, and in fact the relative risk from the second study is not even contained in a 95% confidence interval for relative risk from the original study. The *magnitude* of the effect has been much more closely matched by the "nonsignificant" replication.

Fortunately, psychologists are beginning to notice that replication is not as straightforward as they were originally led to believe. A special issue of the *Journal of Social Behavior and Personality* was entirely devoted to the question of replication (Neuliep, 1990). In one of the articles, Rosenthal cautioned his colleagues: "Given the levels of statistical power at which we normally operate, we have no right to expect the proportion of significant results that we typically do expect, even if in nature there is a very real and very important effect" (Rosenthal, 1990b, page 16).

Jacob Cohen, in his insightful article titled "Things I Have Learned (So Far)," identified another misconception common among social scientists: "Despite widespread misconceptions to the contrary, the rejection of a given null hypothesis gives us no basis for estimating the probability that a replication of the research will again result in rejecting that null hypothesis" (Cohen, 1990, page 1307).

Cohen and Rosenthal both advocate the use of effect sizes as opposed to significance levels when defining the strength of an experimental effect. In general, effect sizes measure the amount by which the data deviate from the null hypothesis in terms of standardized units. For instance, the effect size for a two-sample t-test is usually defined to be the difference in the two means, divided by the standard deviation for the control group. This measure can be compared across studies without the dependence on sample size inherent in significance levels. (Of course there will still be variability in the sample effect sizes, decreasing as a function of sample size.) Comparison of effect sizes across studies is one of the major components of meta-analysis.

Similar arguments have recently been made in the medical literature. For example, Gardner and Altman (1986) stated that the use of p-values "to define two alternative outcomes—significant and not significant—is not helpful and encourages lazy thinking" (page 746). They advocated the use of confidence intervals instead.

As discussed in the next section, the arguments used to conclude that parapsychology has failed to demonstrate a replicable effect hinge on these misconceptions of replication and failure to examine power. A more appropriate analysis would compare the effect sizes for similar experiments across experimenters and across time to see if there have been consistent effects of the same magnitude. Rosenthal also advocates this view of replication:

> The traditional view of replication focuses on significance level as the relevant summary statistic of a study and evaluates the success of a replication in a dichotomous fashion. The newer, more useful view of replication focuses on effect size as the more important summary statistic of a study and evaluates the success of a replication not in a dichotomous but in a continuous fashion [Rosenthal, 1990b, page 28].

The dichotomous view of replication has been used throughout the history of parapsychology, by both parapsychologists and critics (Utts, 1988). For example, the National Academy of Sciences report critically evaluated "significant" experiments, but entirely ignored "nonsignificant" experiments.

In the next three sections, we will examine some of the results in parapsychology using the broader, more appropriate definition of replication. In doing so, we will show that the results are far more interesting than the critics would have us believe.

4. THE GANZFELD DEBATE IN PARAPSYCHOLOGY

An extensive debate took place in the mid-1980s between a parapsychologist and critic, questioning whether or not a particular body of parapsychological data had demonstrated psi abilities. The experiments in question were all conducted using the ganzfeld setting (described below). Several authors were invited to write commentaries on the debate. As a result, this data base has been more thoroughly analyzed by both critics and proponents than any other and provides a good source for studying replication in parapsychology.

The debate concluded with a detailed series of recommendations for further experiments, and left open the question of whether or not psi abilities had been demonstrated. A new series of experiments that followed the recommendations were conducted over the next few years. The results of the new experiments will be presented in Section 5.

4.1 Free-Response Experiments

Recent experiments in parapsychology tend to use more complex target material than the cards and dice used in the early investigations, partially to alleviate boredom on the part of the subjects and partially because they are thought to "more nearly resemble the conditions of spontaneous psi occurrences" (Burdick and Kelly, 1977, page 109). These experiments fall under the general heading of "free-response" experiments, because the subject is asked to give a verbal or written description of the

target, rather than being forced to make a choice from a small discrete set of possibilities. Various types of target material have been used, including pictures, short segments of movies on video tapes, actual locations and small objects.

Despite the more complex target material, the statistical methods used to analyze these experiments are similar to those for forced-choice experiments. A typical experiment proceeds as follows. Before conducting any trials, a large pool of potential targets is assembled, usually in packets of four. Similarity of targets within a packet is kept to a minimum, for reasons made clear below. At the start of an experimental session, after the subject is sequestered in an isolated room, a target is selected at random from the pool. A sender is placed in another room with the target. The subject is asked to provide a verbal or written description of what he or she thinks is in the target, knowing only that it is a photograph, an object, etc.

After the subject's description has been recorded and secured against the potential for later alteration, a judge (who may or may not be the subject) is given a copy of the subject's description and the four possible targets that were in the packet with the correct target. A properly conducted experiment either uses video tapes or has two identical sets of target material and uses the duplicate set for this part of the process, to ensure that clues such as fingerprints don't give away the answer. Based on the subject's description, and of course on a blind basis, the judge is asked to either rank the four choices from most to least likely to have been the target, or to select the one from the four that seems to best match the subject's description. If ranks are used, the statistical analysis proceeds by summing the ranks over a series of trials and comparing the sum to what would be expected by chance. If the selection method is used, a "direct hit" occurs if the correct target is chosen, and the number of direct hits over a series of trials is compared to the number expected in a binomial experiment with $p = 0.25$.

Note that the subjects' *responses* cannot be considered to be "random" in any sense, so probability assessments are based on the random selection of the target and decoys. In a correctly designed experiment, the probability of a direct hit by chance is 0.25 on each trial, regardless of the response, and the trials are independent. These and other issues related to analyzing free-response experiments are discussed by Utts (1991).

4.2 The Psi Ganzfeld Experiments

The ganzfeld procedure is a particular kind of free-response experiment utilizing a perceptual isolation technique originally developed by Gestalt psychologists for other purposes. Evidence from spontaneous case studies and experimental work had led parapsychologists to a model proposing that psychic functioning may be masked by sensory input and by inattention to internal states (Honorton, 1977). The ganzfeld procedure was specifically designed to test whether or not reduction of external "noise" would enhance psi performance.

In these experiments, the subject is placed in a comfortable reclining chair in an acoustically shielded room. To create a mild form of sensory deprivation, the subject wears headphones through which white noise is played, and stares into a constant field of red light. This is achieved by taping halved translucent ping-pong balls over the eyes and then illuminating the room with red light. In the psi ganzfeld experiments, the subject speaks into a microphone and attempts to describe the target material being observed by the sender in a distant room.

At the 1982 Annual Meeting of the Parapsychological Association, a debate took place over the degree to which the results of the psi ganzfeld experiments constituted evidence of psi abilities. Psychologist and critic Ray Hyman and parapsychologist Charles Honorton each analyzed the results of all known psi ganzfeld experiments to date, and they reached strikingly different conclusions (Honorton, 1985b; Hyman, 1985b). The debate continued with the publication of their arguments in separate articles in the March 1985 issue of the *Journal of Parapsychology*. Finally, in the December 1986 issue of the *Journal of Parapsychology*, Hyman and Honorton (1986) wrote a joint article in which they highlighted their agreements and disagreements and outlined detailed criteria for future experiments. That same issue contained commentaries on the debate by 10 other authors.

The data base analyzed by Hyman and Honorton (1986) consisted of results taken from 34 reports written by a total of 47 authors. Honorton counted 42 separate experiments described in the reports, of which 28 reported enough information to determine the number of direct hits achieved. Twenty three of the studies (55%) were classified by Honorton as having achieved statistical significance at 0.05.

4.3 The Vote-Counting Debate

Vote-counting is the term commonly used for the technique of drawing inferences about an experimental effect by counting the number of significant versus nonsignificant studies of the effect. Hedges and Olkin (1985) give a detailed analysis of the inadequacy of this method, showing that it is more and more likely to make the wrong decision as the

number of studies increases. While Hyman acknowledged that "vote-counting raises many problems" (Hyman, 1985b, page 8), he nonetheless spent half of his critique of the ganzfeld studies showing why Honorton's count of 55% was wrong.

Hyman's first complaint was that several of the studies contained multiple conditions, each of which should be considered as a separate study. Using this definition he counted 80 studies (thus further reducing the sample sizes of the individual studies), of which 25 (31%) were "successful." Honorton's response to this was to invite readers to examine the studies and decide for themselves if the varying conditions constituted separate experiments.

Hyman next postulated that there was selection bias, so that significant studies were more likely to be reported. He raised some important issues about how pilot studies may be terminated and not reported if they don't show significant results, or may at least be subject to optional stopping, allowing the experimenter to determine the number of trials. He also presented a chi-square analysis that "suggests a tendency to report studies with a small sample only if they have significant results" (Hyman, 1985b, page 14), but I have questioned his analysis elsewhere (Utts, 1986, page 397).

Honorton refuted Hyman's argument with four rejoinders (Honorton, 1985b, page 66). In addition to reinterpreting Hyman's chi-square analysis, Honorton pointed out that the Parapsychological Association has an official policy encouraging the publication of nonsignificant results in its journals and proceedings, that a large number of reported ganzfeld studies did not achieve statistical significance and that there would have to be 15 studies in the "file-drawer" for every one reported to cancel out the observed significant results.

The remainder of Hyman's vote-counting analysis consisted of showing that the effective error rate for each study was actually much higher than the nominal 5%. For example, each study could have been analyzed using the direct hit measure, the sum of ranks measure or one of two other measures used for free-response analyses. Hyman carried out a simulation study that showed the true error rate would be 0.22 if "significance" was defined by requiring at least one of these four measures to achieve the 0.05 level. He suggested several other ways in which multiple testing could occur and concluded that the effective error rate in each experiment was not the nominal 0.05, but rather was probably close to the 31% he had determined to be the actual success rate in his vote-count.

Honorton acknowledged that there was a multiple testing problem, but he had a two-fold response. First, he applied a Bonferroni correction and found that the number of significant studies (using his definition of a study) only dropped from 55% to 45%. Next, he proposed that a uniform index of success be applied to all studies. He used the number of direct hits, since it was by far the most commonly reported measure and was the measure used in the first published psi ganzfeld study. He then conducted a detailed analysis of the 28 studies reporting direct hits and found that 43% were significant at 0.05 on that measure alone. Further, he showed that significant effects were reported by six of the 10 independent investigators and thus were not due to just one or two investigators or laboratories. He also noted that success rates were very similar for reports published in refereed journals and those published in unrefereed monographs and abstracts.

While Hyman's arguments identified issues such as selective reporting and optional stopping that should be considered in any meta-analysis, the dependence of significance levels on sample size makes the vote-counting technique almost useless for assessing the magnitude of the effect. Consider, for example, the 24 studies where the direct hit measure was reported and the chance probability of a direct hit was 0.25, the most common type of study in the data base. (There were four direct hit studies with other chance probabilities and 14 that did not report direct hits.) Of the 24 studies, 13 (54%) were "nonsignificant" at $\alpha = 0.05$, one-tailed. But if the 367 trials in these "failed replications" are combined, there are 106 direct hits, $z = 1.66$, and $p = 0.0485$, one tailed. This is reminiscent of the dilemma of Professor B in Section 3.

Power is typically very low for these studies. The median sample size for the studies reporting direct hits was 28. If there is a real effect and it increases the success probability from the chance 0.25 to an actual 0.33 (a value whose rationale will be made clear below), the power for a study with 28 trials is only 0.181 (Utts, 1986). It should be no surprise that there is a "repeatability" problem in parapsychology.

4.4 Flaw Analysis and Future Recommendations

The second half of Hyman's paper consisted of a "Meta-Analysis of Flaws and Successful Outcomes" (1985b, page 30), designed to explore whether or not various measures of success were related to specific flaws in the experiments. While many critics have argued that the results in parapsychology can be explained by experimental flaws, Hyman's analysis was the first to attempt to quantify the relationship between flaws and significant results.

Hyman identified 12 potential flaws in the ganzfeld experiments, such as inadequate random-

ization, multiple tests used without adjusting the significance level (thus inflating the significance level from the nominal 5%) and failure to use a duplicate set of targets for the judging process (thus allowing possible clues such as fingerprints). Using cluster and factor analyses, the 12 binary flaw variables were combined into three new variables, which Hyman named General Security, Statistics and Controls.

Several analyses were then conducted. The one reported with the most detail is a factor analysis utilizing 17 variables for each of 36 studies. Four factors emerged from the analysis. From these, Hyman concluded that security had increased over the years, that the significance level tended to be inflated the most for the most complex studies and that both effect size and level of significance were correlated with the existence of flaws.

Following his factor analysis, Hyman picked the three flaws that seemed to be most highly correlated with success, which were inadequate attention to both randomization and documentation and the potential for ordinary communication between the sender and receiver. A regression equation was then computed using each of the three flaws as dummy variables, and the effect size for the experiment as the dependent variable. From this equation, Hyman concluded that a study without these three flaws would be predicted to have a hit rate of 27%. He concluded that this is "well within the statistical neighborhood of the 25% chance rate" (1985b, page 37), and thus "the ganzfeld psi data base, despite initial impressions, is inadequate either to support the contention of a repeatable study or to demonstrate the reality of psi" (page 38).

Honorton discounted both Hyman's flaw classification and his analysis. He did not deny that flaws existed, but he objected that Hyman's analysis was faulty and impossible to interpret. Honorton asked psychometrician David Saunders to write an Appendix to his article, evaluating Hyman's analysis. Saunders first criticized Hyman's use of a factor analysis with 17 variables (many of which were dichotomous) and only 36 cases and concluded that "the entire analysis is meaningless" (Saunders, 1985, page 87). He then noted that Hyman's choice of the three flaws to include in his regression analysis constituted a clear case of multiple analysis, since there were 84 possible sets of three that could have been selected (out of nine potential flaws), and Hyman chose the set most highly correlated with effect size. Again, Saunders concluded that "any interpretation drawn from [the regression analysis] must be regarded as meaningless" (1985, page 88).

Hyman's results were also contradicted by Harris and Rosenthal (1988b) in an analysis requested by Hyman in his capacity as Chair of the National Academy of Sciences' Subcommittee on Parapsychology. Using Hyman's flaw classifications and a multivariate analysis, Harris and Rosenthal concluded that "Our analysis of the effects of flaws on study outcome lends no support to the hypothesis that ganzfeld research results are a significant function of the set of flaw variables" (1988b, page 3).

Hyman and Honorton were in the process of preparing papers for a second round of debate when they were invited to lunch together at the 1986 Meeting of the Parapsychological Association. They discovered that they were in general agreement on several major issues, and they decided to coauthor a "Joint Communiqué" (Hyman and Honorton, 1986). It is clear from their paper that they both thought it was more important to set the stage for future experimentation than to continue the technical arguments over the current data base. In the abstract to their paper, they wrote:

> We agree that there is an overall significant effect in this data base that cannot reasonably be explained by selective reporting or multiple analysis. We continue to differ over the degree to which the effect constitutes evidence for psi, but we agree that the final verdict awaits the outcome of future experiments conducted by a broader range of investigators and according to more stringent standards [page 351].

The paper then outlined what these standards should be. They included controls against any kind of sensory leakage, thorough testing and documentation of randomization methods used, better reporting of judging and feedback protocols, control for multiple analyses and advance specification of number of trials and type of experiment. Indeed, any area of research could benefit from such a careful list of procedural recommendations.

4.5 Rosenthal's Meta-Analysis

The same issue of the *Journal of Parapsychology* in which the Joint Communiqué appeared also carried commentaries on the debate by 10 separate authors. In his commentary, psychologist Robert Rosenthal, one of the pioneers of meta-analysis in psychology, summarized the aspects of Hyman's and Honorton's work that would typically be included in a meta-analysis (Rosenthal, 1986). It is worth reviewing Rosenthal's results so that they can be used as a basis of comparison for the more recent psi ganzfeld studies reported in Section 5.

Rosenthal, like Hyman and Honorton, focused only on the 28 studies for which direct hits were known. He chose to use an effect size measure

called Cohen's h, which is the difference between the arcsin transformed proportions of direct hits that were observed and expected:

$$h = 2(\arcsin \sqrt{\hat{p}} - \arcsin \sqrt{p}).$$

One advantage of this measure over the difference in raw proportions is that it can be used to compare experiments with different chance hit rates.

If the observed and expected numbers of hits were identical, the effect size would be zero. Of the 28 studies, 23 (82%) had effect sizes greater than zero, with a median effect size of 0.32 and a mean of 0.28. These correspond to direct hit rates of 0.40 and 0.38 respectively, when 0.25 is expected by chance. A 95% confidence interval for the true effect size is from 0.11 to 0.45, corresponding to direct hit rates of from 0.30 to 0.46 when chance is 0.25.

A common technique in meta-analysis is to calculate a "combined z," found by summing the individual z scores and dividing by the square root of the number of studies. The result should have a standard normal distribution if each z score has a standard normal distribution. For the ganzfeld studies, Rosenthal reported a combined z of 6.60 with a p-value of 3.37×10^{-11}. He also reiterated Honorton's file-drawer assessment by calculating that there would have to be 423 studies unreported to negate the significant effect in the 28 direct hit studies.

Finally, Rosenthal acknowledged that, because of the flaws in the data base and the potential for at least a small file-drawer effect, the true average effect size was probably closer to 0.18 than 0.28. He concluded, "Thus, when the accuracy rate expected under the null is 1/4, we might estimate the obtained accuracy rate to be about 1/3" (1986, page 333). This is the value used for the earlier power calculation.

It is worth mentioning that Rosenthal was commissioned by the National Academy of Sciences to prepare a background paper to accompany its 1988 report on parapsychology. That paper (Harris and Rosenthal, 1988a) contained much of the same analysis as his commentary summarized above. Ironically, the discussion of the ganzfeld work in the National Academy Report focused on Hyman's 1985 analysis, but never mentioned the work it had commissioned Rosenthal to perform, which contradicted the final conclusion in the report.

5. A META-ANALYSIS OF RECENT GANZFELD EXPERIMENTS

After the initial exchange with Hyman at the 1982 Parapsychological Association Meeting, Honorton and his colleagues developed an automated ganzfeld experiment that was designed to eliminate the methodological flaws identified by Hyman. The execution and reporting of the experiments followed the detailed guidelines agreed upon by Hyman and Honorton.

Using this "autoganzfeld" experiment, 11 experimental series were conducted by eight experimenters between February 1983 and September 1989, when the equipment had to be dismantled due to lack of funding. In this section, the results of these experiments are summarized and compared to the earlier ganzfeld studies. Much of the information is derived from Honorton et al. (1990).

5.1 The Automated Ganzfeld Procedure

Like earlier ganzfeld studies, the "autoganzfeld" experiments require four participants. The first is the Receiver (R), who attempts to identify the target material being observed by the Sender (S). The Experimenter (E) prepares R for the task, elicits the response from R and supervises R's judging of the response against the four potential targets. (Judging is double blind; E does not know which is the correct target.) The fourth participant is the lab assistant (LA) whose only task is to instruct the computer to randomly select the target. No one involved in the experiment knows the identity of the target.

Both R and S are sequestered in sound-isolated, electrically shielded rooms. R is prepared as in earlier ganzfeld studies, with white noise and a field of red light. In a nonadjacent room, S watches the target material on a television and can hear R's target description ("mentation") as it is being given. The mentation is also tape recorded.

The judging process takes place immediately after the 30-minute sending period. On a TV monitor in the isolated room, R views the four choices from the target pack that contains the actual target. R is asked to rate each one according to how closely it matches the ganzfeld mentation. The ratings are converted to ranks and, if the correct target is ranked first, a direct hit is scored. The entire process is automatically recorded by the computer. The computer then displays the correct choice to R as feedback.

There were 160 preselected targets, used with replacement, in 10 of the 11 series. They were arranged in packets of four, and the decoys for a given target were always the remaining three in the same set. Thus, even if a particular target in a set were consistently favored by Rs, the probability of a direct hit under the null hypothesis would remain at 1/4. Popular targets should be no more

likely to be selected by the computer's random number generator than any of the others in the set. The selection of the target by the computer is the only source of randomness in these experiments. This is an important point, and one that is often misunderstood. (See Utts, 1991, for elucidation.)

Eighty of the targets were "dynamic," consisting of scenes from movies, documentaries and cartoons; 80 were "static," consisting of photographs, art prints and advertisements. The four targets within each set were all of the same type. Earlier studies indicated that dynamic targets were more likely to produce successful results, and one of the goals of the new experiments was to test that theory.

The randomization procedure used to select the target and the order of presentation for judging was thoroughly tested before and during the experiments. A detailed description is given by Honorton et al. (1990, pages 118–120).

Three of the 11 series were pilot series, five were formal series with novice receivers, and three were formal series with experienced receivers. The last series with experienced receivers was the only one that did not use the 160 targets. Instead, it used only one set of four dynamic targets in which one target had previously received several first place ranks and one had never received a first place rank. The receivers, none of whom had had prior exposure to that target pack, were not aware that only one target pack was being used. They each contributed one session only to the series. This will be called the "special series" in what follows.

Except for two of the pilot series, numbers of trials were planned in advance for each series. Unfortunately, three of the formal series were not yet completed when the funding ran out, including the special series, and one pilot study with advance planning was terminated early when the experimenter relocated. There were no unreported trials during the 6-year period under review, so there was no "file drawer."

Overall, there were 183 Rs who contributed only one trial and 58 who contributed more than one, for a total of 241 participants and 355 trials. Only 23 Rs had previously participated in ganzfeld experiments, and 194 Rs (81%) had never participated in any parapsychological research.

5.2 Results

While acknowledging that no probabilistic conclusions can be drawn from qualitative data, Honorton et al. (1990) included several examples of session excerpts that Rs identified as providing the basis for their target rating. To give a flavor for the dream-like quality of the mentation and the amount of information that can be lost by only assigning a rank, the first example is reproduced here. The target was a painting by Salvador Dali called "Christ Crucified." The correct target received a first place rank. The part of the mentation R used to make this assessment read:

> . . . I think of guides, like spirit guides, leading me and I come into a court with a king. It's quiet . . . . It's like heaven. The king is something like Jesus. Woman. Now I'm just sort of summersaulting through heaven . . . . Brooding . . . . Aztecs, the Sun God . . . . High priest . . . .Fear . . . . Graves. Woman. Prayer . . . . Funeral . . . . Dark. Death . . . . Souls . . . . Ten Commandments. Moses . . . . [Honorton et al., 1990].

Over all 11 series, there were 122 direct hits in the 355 trials, for a hit rate of 34.4% (exact binomial p-value = 0.00005) when 25% were expected by chance. Cohen's h is 0.20, and a 95% confidence interval for the overall hit rate is from 0.30 to 0.39. This calculation assumes, of course, that the probability of a direct hit is constant and independent across trials, an assumption that may be questionable except under the null hypothesis of no psi abilities.

Honorton et al. (1990) also calculated effect sizes for each of the 11 series and each of the eight experimenters. All but one of the series (the first novice series) had positive effect sizes, as did all of the experimenters.

The special series with experienced Rs had an exceptionally high effect size with $h = 0.81$, corresponding to 16 direct hits out of 25 trials (64%), but the remaining series and the experimenters had relatively homogeneous effect sizes given the amount of variability expected by chance. If the special series is removed, the overall hit rate is 32.1%, $h = 0.16$. Thus, the positive effects are not due to just one series or one experimenter.

Of the 218 trials contributed by novices, 71 were direct hits (32.5%, $h = 0.17$), compared with 51 hits in the 137 trials by those with prior ganzfeld experience (37%, $h = 0.26$). The hit rates and effect sizes were 31% ($h = 0.14$) for the combined pilot series, 32.5% ($h = 0.17$) for the combined formal novice series, and 41.5% ($h = 0.35$) for the combined experienced series. The last figure drops to 31.6% if the outlier series is removed. Finally, without the outlier series the hit rate for the combined series where all of the planned trials were completed was 31.2% ($h = 0.14$), while it was 35% ($h = 0.22$) for the combined series that were terminated early. Thus, optional stopping cannot account for the positive effect.

There were two interesting comparisons that had been suggested by earlier work and were preplanned in these experiments. The first was to compare results for trials with dynamic targets with those for static targets. In the 190 dynamic target sessions there were 77 direct hits (40%, $h = 0.32$) and for the static targets there were 45 hits in 165 trials (27%, $h = 0.05$), thus indicating that dynamic targets produced far more successful results.

The second comparison of interest was whether or not the sender was a friend of the receiver. This was a choice the receiver could make. If he or she did not bring a friend, a lab member acted as sender. There were 211 trials with friends as senders (some of whom were also lab staff), resulting in 76 direct hits (36%, $h = 0.24$). Four trials used no sender. The remaining 140 trials used nonfriend lab staff as senders and resulted in 46 direct hits (33%, $h = 0.18$). Thus, trials with friends as senders were slightly more successful than those without.

Consonant with the definition of replication based on consistent effect sizes, it is informative to compare the autoganzfeld experiments with the direct hit studies in the previous data base. The overall success rates are extremely similar. The overall direct hit rate was 34.4% for the autoganzfeld studies and was 38% for the comparable direct hit studies in the earlier meta-analysis. Rosenthal's (1986) adjustment for flaws had placed a more conservative estimate at 33%, very close to the observed 34.4% in the new studies.

One limitation of this work is that the autoganzfeld studies, while conducted by eight experimenters, all used the same equipment in the same laboratory. Unfortunately, the level of funding available in parapsychology and the cost in time and equipment to conduct proper experiments make it difficult to amass large amounts of data across laboratories. Another autoganzfeld laboratory is currently being constructed at the University of Edinburgh in Scotland, so interlaboratory comparisons may be possible in the near future.

Based on the effect size observed to date, large samples are needed to achieve reasonable power. If there is a constant effect across all trials, resulting in 33% direct hits when 25% are expected by chance, to achieve a one-tailed significance level of 0.05 with 95% probability would require 345 sessions.

We end this section by returning to the aspirin and heart attack example in Section 3 and expanding a comparison noted by Atkinson, Atkinson, Smith and Bem (1990, page 237). Computing the equivalent of Cohen's h for comparing observed heart attack rates in the aspirin and placebo groups results in $h = 0.068$. Thus, the effect size observed in the ganzfeld data base is triple the much publicized effect of aspirin on heart attacks.

6. OTHER META-ANALYSES IN PARAPSYCHOLOGY

Four additional meta-analyses have been conducted in various areas of parapsychology since the original ganzfeld meta-analyses were reported. Three of the four analyses focused on evidence of psi abilities, while the fourth examined the relationship between extroversion and psychic functioning. In this section, each of the four analyses will be briefly summarized.

There are only a handful of English-language journals and proceedings in parapsychology, so retrieval of the relevant studies in each of the four cases was simple to accomplish by searching those sources in detail and by searching other bibliographic data bases for keywords.

Each analysis included an overall summary, an analysis of the quality of the studies versus the size of the effect and a "file-drawer" analysis to determine the possible number of unreported studies. Three of the four also contained comparisons across various conditions.

6.1 Forced-Choice Precognition Experiments

Honorton and Ferrari (1989) analyzed forced-choice experiments conducted from 1935 to 1987, in which the target material was randomly selected *after* the subject had attempted to predict what it would be. The time delay in selecting the target ranged from under a second to one year. Target material included items as diverse as ESP cards and automated random number generators. Two investigators, S. G. Soal and Walter J. Levy, were not included because some of their work has been suspected to be fraudulent.

Overall Results. There were 309 studies reported by 62 senior authors, including more than 50,000 subjects and nearly two million individual trials. Honorton and Ferrari used $z/\sqrt{n}$ as the measure of effect size (ES) for each study, where n was the number of Bernoulli trials in the study. They reported a mean ES of 0.020, and a mean z-score of 0.65 over all studies. They also reported a combined z of 11.41, $p = 6.3 \times 10^{-25}$. Some 30% (92) of the studies were statistically significant at $\alpha = 0.05$. The mean ES per investigator was 0.033, and the significant results were not due to just a few investigators.

Quality. Eight dichotomous quality measures were assigned to each study, resulting in possible

scores from zero for the lowest quality, to eight for the highest. They included features such as adequate randomization, preplanned analysis and automated recording of the results. The correlation between study quality and effect size was 0.081, indicating a slight tendency for higher quality studies to be more successful, contrary to claims by critics that the opposite would be true. There was a clear relationship between quality and year of publication, presumably because over the years experimenters in parapsychology have responded to suggestions from critics for improving their methodology.

File Drawer. Following Rosenthal (1984), the authors calculated the "fail-safe N" indicating the number of unreported studies that would have to be sitting in file drawers in order to negate the significant effect. They found $N = 14{,}268$, or a ratio of 46 unreported studies for each one reported. They also followed a suggestion by Dawes, Landman and Williams (1984) and computed the mean z for all studies with $z > 1.65$. If such studies were a random sample from the upper 5% tail of a $N(0, 1)$ distribution, the mean z would be 2.06. In this case it was 3.61. They concluded that selective reporting could not explain these results.

Comparisons. Four variables were identified that appeared to have a systematic relationship to study outcome. The first was that the 25 studies using subjects selected on the basis of good past performance were more successful than the 223 using unselected subjects, with mean effect sizes of 0.051 and 0.008, respectively. Second, the 97 studies testing subjects individually were more successful than the 105 studies that used group testing; mean effect sizes were 0.021 and 0.004, respectively. Timing of feedback was the third moderating variable, but information was only available for 104 studies. The 15 studies that never told the subjects what the targets were had a mean effect size of −0.001. Feedback after each trial produced the best results, the mean ES for the 47 studies was 0.035. Feedback after each set of trials resulted in mean ES of 0.023 (21 studies), while delayed feedback (also 21 studies) yielded a mean ES of only 0.009. There is a clear ordering; as the gap between time of feedback and time of the actual guesses decreased, effect sizes increased.

The fourth variable was the time interval between the subject's guess and the actual target selection, available for 144 studies. The best results were for the 31 studies that generated targets less than a second after the guess (mean $ES = 0.045$), while the worst were for the seven studies that delayed target selection by at least a month (mean $ES = 0.001$). The mean effect sizes showed a clear trend, decreasing in order as the time interval increased from minutes to hours to days to weeks to months.

6.2 Attempts to Influence Random Physical Systems

Radin and Nelson (1989) examined studies designed to test the hypothesis that "The statistical output of an electronic RNG [random number generator] is correlated with observer intention in accordance with prespecified instructions" (page 1502). These experiments typically involve RNGs based on radioactive decay, electronic noise or pseudorandom number sequences seeded with true random sources. Usually the subject is instructed to try to influence the results of a string of binary trials by mental intention alone. A typical protocol would ask a subject to press a button (thus starting the collection of a fixed-length sequence of bits), and then try to influence the random source to produce more zeroes or more ones. A run might consist of three successive button presses, one each in which the desired result was more zeroes or more ones, and one as a control with no conscious intention. A z score would then be computed for each button press.

The 832 studies in the analysis were conducted from 1959 to 1987 and included 235 "control" studies, in which the output of the RNGs were recorded but there was no conscious intention involved. These were usually conducted before and during the experimental series, as tests of the RNGs.

Results. The effect size measure used was again $z/\sqrt{n}$, where z was positive if more bits of the specified type were achieved. The mean effect size for control studies was not significantly different from zero (-1.0×10^{-5}). The mean effect size for the experimental studies was also very small, 3.2×10^{-4}, but it was significantly higher than the mean ES for the control studies ($z = 4.1$).

Quality. Sixteen quality measures were defined and assigned to each study, under the four general categories of procedures, statistics, data and the RNG device. A score of 16 reflected the highest quality. The authors regressed mean effect size on mean quality for each investigator and found a slope of 2.5×10^{-5} with standard error of 3.2×10^{-5}, indicating little relationship between quality and outcome. They also calculated a weighted mean effect size, using quality scores as weights, and found that it was very similar to the unweighted mean ES. They concluded that "differences in methodological quality are not significant predictors of effect size" (page 1507).

File Drawer. Radin and Nelson used several methods for estimating the number of unreported

studies (pages 1508–1510). Their estimates ranged from 200 to 1000 based on models assuming that all significant studies were reported. They calculated the fail-safe N to be 54,000.

6.3 Attempts to Influence Dice

Radin and Ferrari (1991) examined 148 studies, published from 1935 to 1987, designed to test whether or not consciousness can influence the results of tossing dice. They also found 31 "control" studies in which no conscious intention was involved.

Results. The effect size measure used was $z/\sqrt{n}$, where z was based on the number of throws in which the die landed with the desired face (or faces) up, in n throws. The weighted mean ES for the experimental studies was 0.0122 with a standard error of 0.00062; for the control studies the mean and standard error were 0.00093 and 0.00255, respectively. Weights for each study were determined by quality, giving more weight to high-quality studies. Combined z scores for the experimental and control studies were reported by Radin and Ferrari to be 18.2 and 0.18, respectively.

Quality. Eleven dichotomous quality measures were assigned, ranging from automated recording to whether or not control studies were interspersed with the experimental studies. The final quality score for each study combined these with information on method of tossing the dice, and with source of subject (defined below). A regression of quality score versus effect size resulted in a slope of -0.002, with a standard error of 0.0011. However, when effect sizes were weighted by sample size, there was a significant relationship between quality and effect size, leading Radin and Ferrari to conclude that higher-quality studies produced lower weighted effect sizes.

File Drawer. Radin and Ferrari calculated Rosenthal's fail-safe, N for this analysis to be 17,974. Using the assumption that all significant studies were reported, they estimated the number of unreported studies to be 1152. As a final assessment, they compared studies published before and after 1975, when the *Journal of Parapsychology* adopted an official policy of publishing nonsignificant results. They concluded, based on that analysis, that more nonsignificant studies were published after 1975, and thus "We must consider the overall (1935–1987) data base as suspect with respect to the filedrawer problem."

Comparisons. Radin and Ferrari noted that there was bias in both the experimental and control studies across die face. Six was the face most likely to come up, consistent with the observation that it has the least mass. Therefore, they examined results for the subset of 69 studies in which targets were evenly balanced among the six faces. They still found a significant effect, with mean and standard error for effect size of 8.6×10^{-3} and 1.1×10^{-3}, respectively. The combined z was 7.617 for these studies.

They also compared effect sizes across types of subjects used in the studies, categorizing them as unselected, experimenter and other subjects, experimenter as sole subject, and specially selected subjects. Like Honorton and Ferrari (1989), they found the highest mean ES for studies with selected subjects; it was approximately 0.02, more than twice that for unselected subjects.

6.4 Extroversion and ESP Performance

Honorton, Ferrari and Bem (1991) conducted a meta-analysis to examine the relationship between scores on tests of extroversion and scores on psi-related tasks. They found 60 studies by 17 investigators, conducted from 1945 to 1983.

Results. The effect size measure used for this analysis was the correlation between each subject's extroversion score and ESP score. A variety of measures had been used for both scores across studies, so various correlation coefficients were used. Nonetheless, a stem and leaf diagram of the correlations showed an approximate bell shape with mean and standard deviation of 0.19 and 0.26, respectively, and with an additional outlier at $r = 0.91$. Honorton et al. reported that when weighted by degrees of freedom, the weighted mean r was 0.14, with a 95% confidence interval covering 0.10 to 0.19.

Forced-Choice versus Free-Response Results. Because forced-choice and free-response tests differ qualitatively, Honorton et al. chose to examine their relationship to extroversion separately. They found that for free-response studies there was a significant correlation between extroversion and ESP scores, with mean $r = 0.20$ and $z = 4.46$. Further, this effect was homogeneous across both investigators and extroversion scales.

For forced-choice studies, there was a significant correlation between ESP and extroversion, but only for those studies that reported the ESP results to the subjects *before* measuring extroversion. Honorton et al. speculated that the relationship was an artifact, in which extroversion scores were temporarily inflated as a result of positive feedback on ESP performance.

Confirmation with New Data Following the extroversion/ESP meta-analysis, Honorton et al. attempted to confirm the relationship using the autoganzfeld data base. Extroversion scores based on the Myers–Briggs Type Indicator were available for 221 of the 241 subjects who had participated in autoganzfeld studies.

The correlation between extroversion scores and ganzfeld rating scores was $r = 0.18$, with a 95% confidence interval from 0.05 to 0.30. This is consistent with the mean correlation of $r = 0.20$ for free-response experiments, determined from the meta-analysis. These correlations indicate that extroverted subjects can produce higher scores in free-response ESP tests.

7. CONCLUSIONS

Parapsychologists often make a distinction between "proof-oriented research" and "process-oriented research." The former is typically conducted to test the hypothesis that psi abilities exist, while the latter is designed to answer questions about how psychic functioning works. Proof-oriented research has dominated the literature in parapsychology. Unfortunately, many of the studies used small samples and would thus be nonsignificant even if a moderate-sized effect exists.

The recent focus on meta-analysis in parapsychology has revealed that there are small but consistently nonzero effects across studies, experimenters and laboratories. The sizes of the effects in forced-choice studies appear to be comparable to those reported in some medical studies that had been heralded as breakthroughs. (See Section 5; also Honorton and Ferrari, 1989, page 301.) Free-response studies show effect sizes of far greater magnitude.

A promising direction for future process-oriented research is to examine the causes of individual differences in psychic functioning. The ESP/extroversion meta-analysis is a step in that direction.

In keeping with the idea of individual differences, Bayes and empirical Bayes methods would appear to make more sense than the classical inference methods commonly used, since they would allow individual abilities and beliefs to be modeled. Jeffreys (1990) reported a Bayesian analysis of some of the RNG experiments and showed that conclusions were closely tied to prior beliefs even though hundreds of thousands of trials were available.

It may be that the nonzero effects observed in the meta-analyses can be explained by something other than ESP, such as shortcomings in our understanding of randomness and independence. Nonetheless, there is an anomaly that needs an explanation. As I have argued elsewhere (Utts, 1987), research in parapsychology should receive more support from the scientific community. If ESP does not exist, there is little to be lost by erring in the direction of further research, which may in fact uncover other anomalies. If ESP does exist, there is much to be lost by not doing process-oriented research, and much to be gained by discovering how to enhance and apply these abilities to important world problems.

ACKNOWLEDGMENTS

I would like to thank Deborah Delanoy, Charles Honorton, Wesley Johnson, Scott Plous and an anonymous reviewer for their helpful comments on an earlier draft of this paper, and Robert Rosenthal and Charles Honorton for discussions that helped clarify details.

REFERENCES

ATKINSON, R. L., ATKINSON, R. C., SMITH, E. E. and BEM, D. J. (1990). *Introduction to Psychology*, 10th ed. Harcourt Brace Jovanovich, San Diego.

BELOFF, J. (1985). Research strategies for dealing with unstable phenomena. In *The Repeatability Problem in Parapsychology* (B. Shapin and L. Coly, eds.) 1-21. Parapsychology Foundation, New York.

BLACKMORE, S. J. (1985). Unrepeatability: Parapsychology's only finding. In *The Repeatability Problem in Parapsychology* (B. Shapin and L. Coly, eds.) 183-206. Parapsychology Foundation, New York.

BURDICK, D. S. and KELLY, E. F. (1977). Statistical methods in parapsychological research. In *Handbook of Parapsychology* (B. B. Wolman, ed.) 81-130. Van Nostrand Reinhold, New York.

CAMP, B. H. (1937). (Statement in Notes Section.) *Journal of Parapsychology* **1** 305.

COHEN, J. (1990). Things I have learned (so far). *American Psychologist* **45** 1304-1312.

COOVER, J. E. (1917). *Experiments in Psychical Research at Leland Stanford Junior University*. Stanford Univ.

DAWES, R. M., LANDMAN, J. and WILLIAMS, J. (1984). Reply to Kurosawa. *American Psychologist* **39** 74-75.

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

DOMMEYER, F. C. (1975). Psychical research at Stanford University. *Journal of Parapsychology* **39** 173-205.

DRUCKMAN, D. and SWETS, J. A., eds. (1988) *Enhancing Human Performance: Issues, Theories, and Techniques*. National Academy Press, Washington, D.C.

EDGEWORTH, F. Y. (1885). The calculus of probabilities applied to psychical research. In *Proceedings of the Society for Psychical Research* **3** 190-199.

EDGEWORTH, F. Y. (1886). The calculus of probabilities applied to psychical research. II. In *Proceedings of the Society for Psychical Research* **4** 189-208.

FELLER, W. K. (1940). Statistical aspects of ESP. *Journal of Parapsychology* **4** 271-297.

FELLER, W. K. (1968). *An Introduction to Probability Theory and Its Applications* **1**, 3rd ed. Wiley, New York.

FISHER, R. A. (1924). A method of scoring coincidences in tests with playing cards. In *Proceedings of the Society for Psychical Research* **34** 181-185.

FISHER, R. A. (1929). The statistical method in psychical research. In *Proceedings of the Society for Psychical Research* **39** 189-192.

GALLUP, G. H., JR., and NEWPORT, F. (1991). Belief in paranormal phenomena among adult Americans. *Skeptical Inquirer* **15** 137-146.

GARDNER, M. J. and ALTMAN, D. G. (1986). Confidence intervals rather than *p*-values: Estimation rather than hypothesis testing. *British Medical Journal* **292** 746-750.

GILMORE, J. B. (1989). Randomness and the search for psi. *Journal of Parapsychology* **53** 309-340.

GILMORE, J. B. (1990). Anomalous significance in pararandom and psi-free domains. *Journal of Parapsychology* **54** 53-58.

GREELEY, A. (1987). Mysticism goes mainstream. *American Health* **7** 47-49.

GREENHOUSE, J. B. and GREENHOUSE, S. W. (1988). An aspirin a day...? *Chance* **1** 24-31.

GREENWOOD, J. A. and STUART, C. E. (1940). A review of Dr. Feller's critique. *Journal of Parapsychology* **4** 299-319.

HACKING, I. (1988). Telepathy: Origins of randomization in experimental design. *Isis* **79** 427-451.

HANSEL, C. E. M. (1980). *ESP and Parapsychology: A Critical Re-evaluation*. Prometheus Books, Buffalo, N.Y.

HARRIS, M. J. and ROSENTHAL, R. (1988a). *Interpersonal Expectancy Effects and Human Performance Research*. National Academy Press, Washington, D.C.

HARRIS, M. J. and ROSENTHAL, R. (1988b). *Postscript to Interpersonal Expectancy Effects and Human Performance Research*. National Academy Press, Washington, D.C.

HEDGES, L. V. and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. Academic, Orlando, Fla.

HONORTON, C. (1977). Psi and internal attention states. In *Handbook of Parapsychology* (B. B. Wolman, ed.) 435-472. Van Nostrand Reinhold, New York.

HONORTON, C. (1985a). How to evaluate and improve the replicability of parapsychological effects. In *The Repeatability Problem in Parapsychology* (B. Shapin and L. Coly, eds.) 238-255. Parapsychology Foundation, New York.

HONORTON, C. (1985b). Meta-analysis of psi ganzfeld research: A response to Hyman. *Journal of Parapsychology* **49** 51-91.

HONORTON, C., BERGER, R. E., VARVOGLIS, M. P., QUANT, M., DERR, P., SCHECHTER, E. I. and FERRARI, D. C. (1990). Psi communication in the ganzfeld: Experiments with an automated testing system and a comparison with a meta-analysis of earlier studies. *Journal of Parapsychology* **54** 99-139.

HONORTON, C. and FERRARI, D. C. (1989). "Future telling": A meta-analysis of forced-choice precognition experiments, 1935-1987. *Journal of Parapsychology* **53** 281-308.

HONORTON, C., FERRARI, D. C. and BEM, D. J. (1991). Extraversion and ESP performance: A meta-analysis and a new confirmation. *Research in Parapsychology 1990*. The Scarecrow Press, Metuchen, N.J. To appear.

HYMAN, R. (1985a). A critical overview of parapsychology. In *A Skeptic's Handbook of Parapsychology* (P. Kurtz, ed.) 1-96. Prometheus Books, Buffalo, N.Y.

HYMAN, R. (1985b). The ganzfeld psi experiment: A critical appraisal. *Journal of Parapsychology* **49** 3-49.

HYMAN, R. and HONORTON, C. (1986). Joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology* **50** 351-364.

IVERSEN, G. R., LONGCOR, W. H., MOSTELLER, F., GILBERT, J. P. and YOUTZ, C. (1971). Bias and runs in dice throwing and recording: A few million throws. *Psychometrika* **36** 1-19.

JEFFREYS, W. H. (1990). Bayesian analysis of random event generator data. *Journal of Scientific Exploration* **4** 153-169.

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika* **44** 187-192.

MAUSKOPF, S. H. and MCVAUGH, M. (1979). *The Elusive Science: Origins of Experimental Psychical Research*. Johns Hopkins Univ. Press.

MCVAUGH, M. R. and MAUSKOPF, S. H. (1976). J. B. Rhine's *Extrasensory Perception* and its background in psychical research. *Isis* **67** 161-189.

NEULIEP, J. W., ed. (1990). Handbook of replication research in the behavioral and social sciences. *Journal of Social Behavior and Personality* **5** (4) 1-510.

OFFICE OF TECHNOLOGY ASSESSMENT (1989). Report of a workshop on experimental parapsychology. *Journal of the American Society for Psychical Research* **83** 317-339.

PALMER, J. (1989). A reply to Gilmore. *Journal of Parapsychology* **53** 341-344.

PALMER, J. (1990). Reply to Gilmore: Round two. *Journal of Parapsychology* **54** 59-61.

PALMER, J. A., HONORTON, C. and UTTS, J. (1989). Reply to the National Research Council study on parapsychology. *Journal of the American Society for Psychical Research* **83** 31-49.

RADIN, D. I. and FERRARI, D. C. (1991). Effects of consciousness on the fall of dice: A meta-analysis. *Journal of Scientific Exploration* **5** 61-83.

RADIN, D. I. and NELSON, R. D. (1989). Evidence for consciousness-related anomalies in random physical systems. *Foundations of Physics* **19** 1499-1514.

RAO, K. R. (1985). Replication in conventional and controversial sciences. In *The Repeatability Problem in Parapsychology* (B. Shapin and L. Coly, eds.) 22-41. Parapsychology Foundation, New York.

RHINE, J. B. (1934). *Extrasensory Perception*. Boston Society for Psychical Research, Boston. (Reprinted by Branden Press, 1964.)

RHINE, J. B. (1977). History of experimental studies. In *Handbook of Parapsychology* (B. B. Wolman, ed.) 25-47. Van Nostrand Reinhold, New York.

RICHET, C. (1884). La suggestion mentale et le calcul des probabilités. *Revue Philosophique* **18** 608-674.

ROSENTHAL, R. (1984). *Meta-Analytic Procedures for Social Research*. Sage, Beverly Hills.

ROSENTHAL, R. (1986). Meta-analytic procedures and the nature of replication: The ganzfeld debate. *Journal of Parapsychology* **50** 315-336.

ROSENTHAL, R. (1990a). How are we doing in soft psychology? *American Psychologist* **45** 775-777.

ROSENTHAL, R. (1990b). Replication in behavioral research. *Journal of Social Behavior and Personality* **5** 1-30.

SAUNDERS, D. R. (1985). On Hyman's factor analysis. *Journal of Parapsychology* **49** 86-88.

SHAPIN, B. and COLY, L., eds. (1985). *The Repeatability Problem in Parapsychology*. Parapsychology Foundation, New York.

SPENCER-BROWN, G. (1957). *Probability and Scientific Inference*. Longmans Green, London and New York.

STUART, C. E. and GREENWOOD, J. A. (1937). A review of criticisms of the mathematical evaluation of ESP data. *Journal of Parapsychology* **1** 295-304.

TVERSKY, A. and KAHNEMAN, D. (1982). Belief in the law of small numbers. In *Judgment Under Uncertainty: Heuristics and Biases* (D. Kahneman, P. Slovic and A. Tversky, eds.) 23-31. Cambridge Univ. Press.

UTTS, J. (1986). The ganzfeld debate: A statistician's perspective. *Journal of Parapsychology* **50** 395-402.

UTTS, J. (1987). Psi, statistics, and society. *Behavioral and Brain Sciences* **10** 615-616.

UTTS, J. (1988). Successful replication versus statistical significance. *Journal of Parapsychology* **52** 305-320.

UTTS, J. (1989). Randomness and randomization tests: A reply to Gilmore. *Journal of Parapsychology* **53** 345-351.

UTTS, J. (1991). Analyzing free-response data: A progress report. In *Psi Research Methodology: A Re-examination* (L. Coly, ed.). Parapsychology Foundation, New York. To appear.

WILKS, S. S. (1965a). Statistical aspects of expeirments in telepath. *N.Y. Statistician* **16** (6) 1-3.

WILKS, S. S. (1965b). Statistical aspects of experiments in telepathy. *N.Y. Statistician* **16** (7) 4-6.

Comment

M. J. Bayarri and James Berger

1. INTRODUCTION

There are many fascinating issues discussed in this paper. Several concern parapsychology itself and the interpretation of statistical methodology therein. We are not experts in parapsychology, and so have only one comment concerning such matters: In Section 3 we briefly discuss the need to switch from P-values to Bayes factors in discussing evidence concerning parapsychology.

A more general issue raised in the paper is that of replication. It is quite illuminating to consider the issue of replication from a Bayesian perspective, and this is done in Section 2 of our discussion.

2. REPLICATION

Many insightful observations concerning replication are given in the article, and these spurred us to determine if they could be quantified within Bayesian reasoning. Quantification requires clear delineation of the possible purposes of replication, and at least two are obvious. The first is simple reduction of random error, achieved by obtaining more observations from the replication. The second purpose is to search for possible bias in the original experiment. We use "bias" in a loose sense here, to refer to any of the huge number of ways in which the effects being measured by the experiment can differ from the actual effects of interest. Thus a clinical trial without a placebo can suffer a placebo "bias"; a survey can suffer a "bias" due to the sampling frame being unrepresentative of the actual population; and possible sources of bias in parapsychological experiments have been extensively discussed.

Replication to Reduce Random Error

If the sole goal of replication of an experiment is to reduce random error, matters are very straightforward. Reviewing the Bayesian way of studying this issue is, however, useful and will be done through the following simple example.

EXAMPLE 1. Consider the example from Tversky and Kahnemann (1982), in which an experiment results in a standardized test statistic of $z_1 = 2.46$. (We will assume normality to keep computations trivial.) The question is: What is the highest value of z_2 in a second set of data that would be considered a failure to replicate? Two possible precise versions of this question are: Question 1: What is the probability of observing z_2 for which the null hypothesis would be rejected in the replicated experiment? Question 2: What value of z_2 would leave one's overall opinion about the null hypothesis unchanged?

Consider the simple case where $Z_1 \sim N(z_1 \mid \theta, 1)$ and (independently) $Z_2 \sim N(z_2 \mid \theta, 1)$, where θ is the mean and 1 is the standard deviation of the normal distribution. Note that we are considering the case in which no experimental bias is suspected and so the means for each experiment are assumed to be the same.

Suppose that it is desired to test $H_0: \theta \le 0$ versus $H_1: \theta > 0$, and suppose that initial prior opinion about θ can be described by the noninformative prior $\pi(\theta) = 1$. We consider the one-sided testing problem with a constant prior in this section, because it is known that then the posterior probability of H_0, to be denoted by $P(H_0 \mid \text{data})$, equals the P-value, allowing us to avoid complications arising from differences between Bayesian and classical answers.

After observing $z_1 = 2.46$, the posterior distribution of θ is

$$\pi(\theta \mid z_1) = N(\theta \mid 2.46, 1).$$

Question 1 then has the answer (using predictive Bayesian reasoning)

$$\begin{aligned} &P(\text{rejecting at level } \alpha \mid z_1) \\ &\quad = \int_{c_\alpha}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-1/2(z_2-\theta)^2} \pi(\theta \mid z_1)\, d\theta\, dz_2 \\ &\quad = 1 - \Phi\left(\frac{c_\alpha - 2.46}{\sqrt{2}}\right), \end{aligned}$$

where Φ is the standard normal cdf and c_α is the (one-sided) critical value corresponding to the level, α, of the test. For instance, if $\alpha = 0.05$, then this probability equals 0.7178, demonstrating that there is a quite substantial probability that the second experiment will fail to reject. If α is chosen to be the observed significance level from the first experiment, so that $c_\alpha = z_1$, then the probability that the

M. J. Bayarri is Titular Professor, Department of Statistics and Operations Research, University of Valencia, Avenida Dr. Moliner 50, 46100 Burjassot, Valencia, Spain. James Berger is the Richard M. Brumfield Distinguished Professor of Statistics, Purdue University, West Lafayette, Indiana 47907.

second experiment will reject is just 1/2. This is nothing but a statement of the well-known martingale property of Bayesianism, that what you "expect" to see in the future is just what you know today. In a sense, therefore, question 1 is exposed as being uninteresting.

Question 2 more properly focuses on the fact that the stated goal of replication here is simply to reduce uncertainty in stated conclusions. The answer to the question follows immediately from noting that the posterior from the combined data (z_1, z_2) is

$$\pi(\theta \mid z_1, z_2) = N(\theta \mid (z_1 + z_2)/2, 1/\sqrt{2}),$$

so that

$$P(H_0 \mid \text{data}) = \Phi(-(z_1 + z_2)/\sqrt{2}).$$

Setting this equal to $P(H_0 \mid z_1)$ and solving for z_2 yields $z_2 = (\sqrt{2} - 1)z_1 = 1.02$. Any value of z_2 greater than this will increase the total evidence against H_0, while any value smaller than 1.02 will decrease the evidence.

Replication to Detect Bias

The aspirin example dramatically raises the issue of bias detection as a motive for replication. Professor Utts observes that replication 1 gives results that are fully compatible with those of the original study, which could be interpreted as suggesting that there is no bias in the original study, while replication 2 would raise serious concerns of bias. We became very interested in the implicit suggestion that replication 2 would thus lead to less overall evidence against the null hypothesis than would replication 1, even though in isolation replication 2 was much more "significant" than was replication 1. In attempting to see if this is so, we considered the Bayesian approach to study of bias within the framework of the aspirin example.

EXAMPLE 2. For simplicity in the aspiring example, we reduce consideration to

θ = true difference in heart attack rates between aspirin and placebo populations multiplied by 1000;

Y = difference in observed heart attack rates between aspirin and placebo groups in original study multiplied by 1000;

X_i = difference in observed heart attack rates between aspirin and placebo groups in Replication i multiplied by 1000.

We assume that the replication studies are extremely well designed and implemented, so that one is very confident that the X_i have mean θ. Using normal approximations for convenience, the data can be summarized as

$$X_1 \sim N(x_1 \mid \theta, 4.82), \quad X_2 \sim N(x_2 \mid \theta, 3.63)$$

with actual observations $x_1 = 7.704$ and $x_2 = 13.07$.

Consider now the bias issue. We assume that the original experiment is somewhat suspect in this regard, and we will model bias by defining the mean of Y to be

$$\eta = \theta + \beta,$$

where β is the unknown bias. Then the data in the original experiment can be summarized by

$$Y \sim N(y \mid \eta, 1.54),$$

with the actual observation being $y = 7.707$.

Bayesian analysis requires specification of a prior distribution, $\pi(\beta)$, for the suspected amount of bias. Of particular interest then are the posterior distribution of β, assuming replication i has been performed, given by

$$\pi(\beta \mid y, x_i) \propto \pi(\beta)\exp\left\{-\frac{1}{2(1.54^2 + \sigma_i^2)}\left[\beta - (y - x_i)\right]^2\right\},$$

where σ_i^2 is the variance (4.82 or 3.63) from replication i; and the posterior probability of H_0, given by

$$P(H_0 \mid y, x_i) = \int_{-\infty}^{\infty} \Phi\left(-\frac{\sigma_i}{1.54\sqrt{\sigma_i^2 + 1.54^2}}(y - \beta) - \frac{1.54}{\sigma_i\sqrt{\sigma_i^2 + 1.54^2}}x_i\right)\pi(\beta \mid y, x_i)\, d\beta.$$

Recall that our goal here was to see if Bayesian analysis can reproduce the intuition that the original experiment could be trusted if replication 1 had been done, while it could not be trusted (in spite of its much larger sample size) had replication 2 been performed. Establishing this requires finding a prior distribution $\pi(\beta)$ for which $\pi(\beta \mid y, x_1)$ has little effect on $P(H_0 \mid y, x_1)$, but $\pi(\beta \mid y, x_2)$ has a large effect on $P(H_0 \mid y, x_2)$. To achieve the first objective, $\pi(\beta)$ must be tightly concentrated near zero. To achieve the second, $\pi(\beta)$ must be such that large $|y - x_2|$, which suggests presence of a large bias, can result in a substantial shift of posterior mass for β away from zero.

A sensible candidate for the prior density $\pi(\beta)$ is the Cauchy $(0, V)$ density

$$\pi_V(\beta) = \frac{1}{\pi V\left[1 + (\theta/V)^2\right]}.$$

Flat-tailed densities, such as this, are well known to have the property that when discordant data is observed (e.g., when $(|y - x_2|$ is large), substantial mass shifts away from the prior center towards the likelihood center. It is easy to see that a normal prior for β can not have the desired behavior.

Our first surprise in consideration of these priors was how small V needed to be chosen in order for $P(H_0 \mid y, x_1)$ to be unaffected by the bias. For instance, even with $V = 1.54/100$ (recall that 1.54 was the standard deviation of Y from the original experiment), computation yields $P(H_0 \mid y, x_1) = 4.3 \times 10^{-5}$, compared with the P-value (and posterior probability from the original experiment assuming no bias) of 2.8×10^{-7}. There is a clear lesson here; even very small suspicions of bias can drastically alter a small P-value. Note that replication 1 is very consistent with the presence of no bias, and so the posterior distribution for the bias remains tightly concentrated near zero; for instance, the mean of the posterior for β is then 7.2×10^{-6}, and the standard deviation is 0.25.

When we turned attention to replication 2, we found that it did not seriously change the prior perceptions of bias. Examination quickly revealed the reason; even the maximum likelihood estimate of the bias is no more than 1.4 standard deviations from zero, which is not enough to change strong prior beliefs. We, therefore, considered a third experiment, defined in Table 1. Transforming to approximate normality, as before, yields

$$X_3 \sim N(x_3 \mid \theta, 3.48),$$

with $x_3 = 22.72$ being the actual observation. The maximum likelihood estimate of bias is now 3.95 standard deviations from zero, so there is potential for a substantial change in opinion about the bias.

Sure enough, computation when $V = 1.54/100$ yields that $E[\beta \mid y, x_3] = -4.9$ with (posterior) standard deviation equal to 6.62, which is a dramatic shift from prior opinion (that β is Cauchy (0, 1.54/100)). The effect of this is to essentially ignore the original experiment in overall assessments of evidence. For instance, $P(H_0 \mid y, x_3) = 3.81 \times 10^{-11}$, which is very close to $P(H_0 \mid x_3) = 3.29 \times 10^{-11}$. Note that, if β were set equal to zero, the overall posterior probability of H_0 (and P-value) would be 2.62×10^{-13}.

TABLE 1
Frequency of heart attacks in replication 3

	Yes	No
Aspirin	5	2309
Placebo	54	2116

Thus Bayesian reasoning can reproduce the intuition that replication which indicates bias can cast considerable doubt on the original experiment, while replication which provides no evidence of bias leaves evidence from the original experiment intact. Such behavior seems only obtainable, however, with flat-tailed priors for bias (such as the Cauchy) that are very concentrated (in comparison with the experimental standard deviation) near zero.

3. *P*-VALUES OR BAYES FACTORS?

Parapsychology experiments usually consider testing of H_0: No parapsychological effect exists. Such null hypotheses are often realistically represented as point nulls (see Berger and Delampady, 1987, for the reason that care must be taken in such representation), in which case it is known that there is a large difference between P-values and posterior probabilities (see Berger and Delampady, 1987, for review). The article by Jefferys (1990) dramatically illustrates this, showing that a very small P-value can actually correspond to evidence for H_0 when considered from a Bayesian perspective. (This is very related to the famous "Jeffreys" paradox.) The argument in favor of the Bayesian approach here is very strong, since it can be shown that the conflict holds for virtually any sensible prior distribution; a Bayesian answer can be wrong if the prior information turns out to be inaccurate, but a Bayesian answer that holds for all sensible priors is unassailable.

Since P-values simply cannot be viewed as meaningful in these situations, we found it of interest to reconsider the example in Section 5 from a Bayes factor perspective. We considered only analysis of the overall totals, that is, $x = 122$ successes out of $n = 355$ trials. Assuming a simple Bernoulli trial model with success probability θ, the goal is to test $H_0{:}\theta = 1/4$ versus $H_1{:}\theta \neq 1/4$.

To determine the Bayes factor here, one must specify $g(\theta)$, the conditional prior density on H_1. Consider choosing g to be uniform and symmetric, that is,

$$G_r(\theta) = \begin{cases} \dfrac{1}{2r}, & \text{for } \dfrac{1}{4} - r \leq \theta \leq \dfrac{1}{4} + r, \\ 0, & \text{otherwise.} \end{cases}$$

Crudely, r could be considered to be the maximum change in success probability that one would expect given that ESP exists. Also, these distributions are the "extreme points" over the class of symmetric unimodal conditional densities, so answers that hold over this class are also representative of answers over a much larger class. Note that here $r \leq 0.25$ (because $0 \leq \theta \leq 1$); for the given data the $\theta > 0.5$ are essentially irrelevant, but if it were deemed important to take them into account one could use the more sophisticated binomial analysis in Berger and Delampady (1987).

For g_r, the Bayes factor of H_1 to H_0, which is to be interpreted as the relative odds for the hypotheses provided by the data, is given by

$$B(r) = \frac{(1/(2r))\int_{.25-r}^{.25+r} \theta^{122}(1-\theta)^{355-122}\,d\theta}{(1/4)^{122}(1-1/4)^{355-122}}$$

$$\cong \frac{1}{2r}(63.13) \cdot \left[\Phi\left(\frac{r-.0937}{.0252}\right) + \Phi\left(\frac{-(r+.0937)}{.0252}\right)\right].$$

This is graphed in Figure 1.

The P-value for this problem was 0.00005, indicating overwhelming evidence against H_0 from a classical perspective. In contrast to the situation studied by Jefferys (1990), the Bayes factor here does not completely reverse the conclusion, showing that there are very reasonable values of r for which the evidence against H_0 is moderately strong, for example 100/1 or 200/1. Of course, this evidence is probably not of sufficient strength to overcome strong prior opinions against H_0 (one obtains final posterior odds by multiplying prior odds by the Bayes factor). To properly assess strength of evidence, we feel that such Bayes factor computations should become standard in parapsychology.

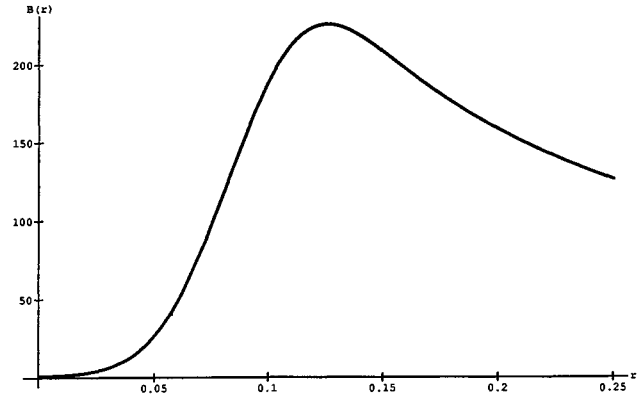

FIG. 1. *The Bayes factor of H_1 to H_0 as a function of r, the maximum change in success probability that is expected given that ESP exists, for the ganzfeld experiment.*

As mentioned by Professor Utts, Bayesian methods have additional potential in situations such as this, by allowing unrealistic models of iid trials to be replaced by hierarchical models reflecting differing abilities among subjects.

ACKNOWLEDGMENTS

M. J. Bayarri's research was supported in part by the Spanish Ministry of Education and Science under DGICYT Grant BE91-038, while visiting Purdue University. James Berger's research was supported by NSF Grant DMS-89-23071.

Comment

Ree Dawson

This paper offers readers interested in statistical science multiple views of the controversial history of parapsychology and how statistics has contributed to its development. It first provides an account of how both design and inferential aspects of statistics have been pivotal issues in evaluating the outcomes of experiments that study psi abilities. It then emphasizes how the idea of science as replication has been key in this field in which results have not been conclusive or consistent and thus meta-analysis has been at the heart of the literature in parapsychology. The author not only reviews past debate on how to interpret repeated psi studies, but also provides very detailed information on the Honorton–Hyman argument, a nice illustration of the challenges of resolving such de-

Ree Dawson is Senior Statistician, New England Biomedical Research Foundation, and Statistical Consultant, RFE/RL Research Institute. Her mailing address is 177 Morrison Avenue, Somerville, Massachusetts 02144.

bate. This debate is also a good example of how statistical criticism can be part of the scientific process and lead to better experiments and, in general, better science.

The remainder of the paper addresses technical issues of meta-analysis, drawing upon recent research in parapsychology for an in-depth application. Through a series of examples, the author presents a convincing argument that power issues cannot be overlooked in successive replications and that comparison of effect sizes provides a richer alternative to the dichotomous measure inherent in the use of p-values. This is particularly relevant when the potential effect size is small and resources are limited, as seems to be the case for psi studies.

The concluding section briefly mentions Bayesian techniques. As noted by the author, Bayes (or empirical Bayes) methodology seems to make sense for research in parapsychology. This discussion examines possible Bayesian approaches to meta-analysis in this field.

BAYES MODELS FOR PARAPSYCHOLOGY

The notion of repeatability maps well into the Bayesian set-up in which experiments, viewed as a random sample from some superpopulation of experiments, are assumed to be exchangeable. When subjects can also be viewed as an approximately random sample from some population, it is appropriate to pool them across experiments. Otherwise, analyses that partially pool information according to experimental heterogeneity need to be considered. Empirical and hierarchical Bayes methods offer a flexible modeling framework for such analyses, relying on empirical or subjective sources to determine the degree of pooling. These richer methods can be particularly useful to meta-analysis of experiments in parapsychology conducted under potentially diverse conditions.

For the recent ganzfeld series, assuming them to be independent binomially distributed as discussed in Section 5, the data can be summed (pooled) across series to estimate a common hit rate. Honorton et al. (1990) assessed the homogeneity of effects across the 11 series using a chi-square test that compares individual effect sizes to the weighted mean effect. The chi-square statistic $\chi^2_{10} = 16.25$, not statistically significant ($p = 0.093$), largely reflects the contribution of the last "special" series (contributes 9.2 units to the χ^2_{10} value), and to a lesser extent the novice series with a negative effect (contributes 2.5 units). The outlier series can be dropped from the analysis to provide a more conservative estimate of the presence of psi effects for this data (this result is reported in Section 5). For the remaining 10 series, the chi-square value $\chi^2_9 = 7.01$ strongly favors homogeneity, although more than one-third of its value is due to the novice series (number 4 in Table 1). This pattern points to the potential usefulness of a richer model to accommodate series that may be distinct from the others. For the earlier ganzfeld data analyzed by Honorton (1985b), the appeal of a Bayes or other model that recognizes the heterogeneity across studies is clear cut: $\chi^2_{23} = 56.6$, $p = 0.0001$, where only those studies with common chance hit rate have been included (see Table 2).

Historic reliance on voting-count approaches to determine the presence of psi effects makes it natural to consider Bayes models that focus on the ensemble of experimental effects from parapsychological studies, rather than individual estimates. Recent work in parapsychology that compares effect sizes across studies, rather than estimating separate study effects, reinforces the need to examine this type of model. Louis (1984) develops Bayes and empirical Bayes methods for problems that consider the ensemble of parameter values to be the primary goal, for example, multiple comparisons. For the simple compound normal model, $Y_i \sim N(\theta_i, 1)$, $\theta_i \sim N(\mu, \tau^2)$, the standard Bayes estimates (posterior means)

$$\theta_i^* = \mu + D(Y_i - \mu) \quad \text{and} \quad D = \frac{\tau^2}{1 + \tau^2}$$

where the θ_i represent experimental effects of interest, are modified approximately to

$$\theta_i^l \approx \mu + \sqrt{D}\,(Y_i - \mu)$$

when an ensemble loss function is assumed. The new estimates adjust the shrinkage factor D so that their sample mean and variance match the posterior expectation and variance of the θ's. Similar results are obtained when the model is gener-

TABLE 1
Recent ganzfeld series

Series type	N Trials	Hit rate	Y_i	σ_i
Pilot	22	0.36	−0.58	0.44
Pilot	9	0.33	−0.71	0.71
Pilot	36	0.28	−0.94	0.37
Novice	50	0.24	−1.15	0.33
Novice	50	0.36	−0.58	0.30
Novice	50	0.30	−0.85	0.31
Novice	50	0.36	−0.58	0.30
Novice	6	0.67	0.71	0.87
Experienced	7	0.43	−0.28	0.76
Experienced	50	0.30	−0.85	0.31
Experienced	25	0.64	0.58	0.42
Overall	355	0.34		

TABLE 2
Earlier ganzfeld studies

N Trials	Hit rate	Y_i	σ_i
32	0.44	−0.24	0.36
7	0.86	1.82	1.09
30	0.43	−0.28	0.37
30	0.23	−1.21	0.43
20	0.10	−2.20	0.75
10	0.90	2.20	1.05
10	0.40	−0.41	0.65
28	0.29	−0.90	0.42
10	0.40	−0.41	0.65
20	0.35	−0.62	0.47
26	0.31	−0.80	0.42
20	0.45	−0.20	0.45
20	0.45	−0.20	0.45
30	0.53	0.12	0.37
36	0.33	−0.71	0.35
32	0.28	−0.94	0.39
40	0.28	−0.94	0.35
26	0.46	−0.16	0.39
20	0.60	0.41	0.46
100	0.41	−0.36	0.20
40	0.33	−0.71	0.34
27	0.41	−0.36	0.39
60	0.45	−0.20	0.26
48	0.21	−1.33	0.35
722	.38		

alized to the case of unequal variances, $Y_i \sim N(\theta_i, \sigma_i^2)$.

For the above model, the fraction of θ_i^l above (or below) a cut point C is a consistent estimate of the fraction of $\theta_i > C$ (or $\theta_i < C$). Thus, the use of ensemble, rather than component-wise, loss can help detect when individual effects are above a specified threshold by chance. For the meta-analysis of ganzfeld experiments, the observed binomial proportions transformed on the logit (or arcsin$\sqrt{}$) scale can be modeled in this framework. Letting d_i and m_i denote the number of direct hits and misses respectively for the ith experiment, and p_i as the corresponding population proportion of direct hits, the Y_i are the observed logits

$$Y_i = \log(d_i/m_i)$$

and σ_i^2, estimated by maximum likelihood as $1/d_i + 1/m_i$, is the variance of Y_i conditional on $\theta_i = \text{logit}(p_i)$. The threshold logit $(0.25) \approx 1.10$ can be used to identify the number of experiments for which the proportion of direct hits exceeds that expected by chance.

Table 1 shows Y_i and σ_i for the 11 ganzfeld series. All but one of the series are well above the threshold; Y_4 marginally falls below −1.10. Any shrinkage toward a common hit rate will lead to an estimate, θ_4^* or θ_4^l, above the threshold. The use of ensemble loss (with its consistency property) provides more convincing support that all $\theta_i > -1.10$, although posterior estimates of uncertainty are needed to fully calibrate this. For the earlier ganzfeld data in Table 2, ensemble loss can similarly be used to determine the number of studies with $\theta_i < -1.10$ and specifically whether the negative effects of studies 4 and 24 ($Y_4 = -1.21$ and $Y_{24} = -1.33$) occurred as a result of chance fluctuation.

Features of the ganzfeld data in Section 5, such as the outlier series, suggest that further elaboration of the basic Bayesian set-up may be necessary for some meta-analyses in parapsychology. Hierarchical models provide a natural framework to specify these elaborations and explore how results change with the prior specification. This type of sensitivity analysis can expose whether conclusions are closely tied to prior beliefs, as observed by Jeffreys for RNG data (see Section 7). Quantifying the influence of model components deemed to be more subjective or less certain is important to broad acceptance of results as evidence of psi performance (or lack thereof).

Consider the initial model commonly used for Bayesian analysis of discrete data:

$$Y_i \mid p_i, n_i \sim B(p_i, n_i),$$
$$\theta_i \sim N(\mu, \tau^2), \quad \theta_i = \text{logit}(p_i),$$

with noninformative priors assumed for μ and τ^2 (e.g., log τ locally uniform). The distinctiveness of the last "special" series and, in general, the different types of series (pilot versus formal, novice versus experienced) raises the question of whether the experimental effects follow a normal distribution. Weighted normal plots (Ryan and Dempster, 1984) can be used to graphically diagnose the adequacy of second-stage normality (see Dempster, Selwyn and Weeks, 1983, for examples with binary response and normal superpopulation).

Alternatively, if nonnormality is suspected, the model can be revised to include some sort of heavy-tailed prior to accommodate possibly outlying series or studies. West (1985) incorporates additional scale parameters, one for each component of the model (experiment), that flexibly adapt to a typical θ_i and discount their influence on posterior estimates, thus avoiding under- or over-shrinkage due to such θ_i. For example, the second stage can specify the prior as a scale mixture of normals:

$$\theta_i \sim N(\mu, \tau^2\gamma_i^{-1}),$$
$$k\gamma_i \sim \chi_k^2,$$
$$v\tau^{-2} \sim \chi_v^2.$$

This approach for the prior is similar to others for

maximum likelihood estimation that modify the sampling error distribution to yield estimates that are "robust" against outlying observations.

Like its maximum likelihood counterparts, in addition to the robust effect estimates θ_i^*, the Bayes model provides (posterior) scale estimates γ_i^*. These can be interpreted as the weight given to the data for each θ_i in the analysis and are useful to diagnosing which model components (series or studies) are unusual and how they influence the shrinkage. When more complex groupings among the θ_i are suspected, for example, bimodal distribution of studies from different sites or experimenters, other mixture specifications can be used to further relax the shrinkage toward a common value.

For the 11 ganzfeld series, the last "outlier" series, quite distinct from the others (hit rate = 0.64), is moderately precise ($N = 25$). Omitting it from the analysis causes the overall hit rate to drop from 0.344 to 0.321. The scale mixture model is a compromise between these two values (on the logit scale), discounting the influence of series 11 on the estimated posterior common hit rate used for shrinkage. The scale factor γ_{11}^*, an indication of how separate θ_{11} is from the other parameters, also causes θ_{11}^* to be shrunk less toward the common hit rate than other, more homogeneous θ_i, giving more weight to individual information for that series (see West, 1985). The heterogeneity of the earlier ganzfeld data is more pronounced, and studies are taken from a variety of sources over time. For these data, the γ_i^* can be used to explore atypical studies (e.g., study 6, with hit rate = 0.90, contributes more than 25% to the χ_{23}^2 value for homogeneity) and groupings among effects, as well as protect the analysis from misspecification of second-stage normality.

Variation among ganzfeld series or studies and the degree to which pooling or shrinking is appropriate can be investigated further by considering a range of priors for τ^2. If the marginal likelihood of τ^2 dominates the prior specification, then results should not vary as the prior for τ^2 is varied. Otherwise, it is important to identify the degree to which subjective information about interexperimental variability influences the conclusions. This sensitivity analysis is a Bayesian enrichment of the simpler test of homogeneity directed toward determining whether or not complete pooling is appropriate.

To assess how well heterogeneity among historical control groups is determined by the data. Dempster, Selwyn and Weeks (1983) propose three priors for τ^2 in the logistic-normal model. The prior distributions range from strongly favoring individual estimates, $p(\tau^2)d\tau \propto \tau^{-1}$, to the uniform reference prior $p(\tau^2)d\tau \propto \tau^{-2}$, flat on the log τ scale, to strongly favoring complete pooling, $p(\tau^2)d\tau \propto \tau^{-3}$ (the latter forcing complete pooling for the compound normal model; see Morris, 1983). For their two examples, the results (estimates of linear treatment effects) are largely insensitive to variation in the prior distribution, but the number of studies in each example was large (70 and 19 studies available for pooling). For the 11 ganzfeld series, τ^2 may be less well determined by the data. The posterior estimate of τ^2 and its sensitivity to $p(\tau^2)d\tau$ will also depend on whether individual scale parameters are incorporated into the model. Discounting the influence of the last series will both shift the marginal likelihood toward smaller values of τ^2 and concentrate it more in that region.

The issue of objective assessment of experiment results is one that extends well beyond the field of parapsychology, and this paper provides insight into issues surrounding the analysis and interpretation of small effects from related studies. Bayes methods can contribute to such meta-analyses in two ways. They permit experimental and subjective evidence to be formally combined to determine the presence or absence of effects that are not clear cut or controversial (e.g., psi abilities). They can also help uncover sources and degree of uncertainty in the scientific conclusions.

Comment

Persi Diaconis

In my experience, parapsychologists use statistics extremely carefully. The plethora of widely significant p-values in the many thousands of published parapsychological studies must give us pause for thought. Either something spooky is going on, or it is possible for a field to exist on error and artifact for over 100 years. The present paper offers a useful review by an expert and a glimpse at some tantalizing new studies.

My reaction is that the studies are crucially flawed. Since my reasons are somewhat unusual, I will try to spell them out.

I have found it impossible to usefully judge what actually went on in a parapsychology trial from their published record. Time after time, skeptics have gone to watch trials and found subtle and not-so-subtle errors. Since the field has so far failed to produce a replicable phenomena, it seems to me that any trial that asks us to take its findings seriously should include full participation by qualified skeptics. Without a magician and/or knowledgeable psychologist skilled at running experiments with human subjects, I don't think a serious effort is being made.

I recognize that this is an unorthodox set of requirements. In fact, one cannot judge what "really goes on" in studies in most areas, and it is impossible to demand wide replicability in others. Finally, defining "qualified skeptic" is difficult. In defense, most areas have many easily replicable experiments and many have their findings explained and connected by unifying theories. It simply seems clear that when making claims at such extraordinary variance with our daily experience, claims that have been made and washed away so often in the past, such extraordinary measures are mandatory before one has the right to ask outsiders to spend their time in review. The papers cited in Section 5 do not actively involve qualified skeptics, and I do not feel they have earned the right to our serious attention.

The points I have made above are not new. Many appear in the present article. This does not diminish their utility nor applicability to the most recent studies.

Parapsychology is worth serious study. First, there may be something there, and I marvel at the patience and drive of people like Jessica Utts and Ray Hyman. Second, if it is wrong, it offers a truly alarming massive case study of how statistics can mislead and be misused. Third, it offers marvelous combinatorial and inferential problems. Chung, Diaconis, Graham and Mallows (1981), Diaconis and Graham (1981) and Samaniego and Utts (1983) offer examples not cited in the text. Finally, our budding statistics students are fascinated by its claims; the present paper gives a responsible overview providing background for a spectacular classroom presentation.

Persi Diaconis is Professor of Mathematics at Harvard University, Science Center, 1 Oxford Street, Cambridge, Massachusetts 02138.

Comment: Parapsychology — On the Margins of Science?

Joel B. Greenhouse

Professor Utts reviews and synthesizes a large body of experimental literature as well as the scientific controversy involved in the attempt to establish the existence of paranormal phenomena. The organization and clarity of her presentation are noteworthy. Although I do not believe that this paper will necessarily change anyone's views regarding the existence of paranormal phenomena, it does raise very interesting questions about the process by which new ideas are either accepted or rejected by the scientific community. As students of science, we believe that scientific discovery

Joel B. Greenhouse is Associate Professor of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890.

advances methodically and objectively through the accumulation of knowledge (or the rejection of false knowledge) derived from the implementation of the scientific method. But, as we will see, there is more to the acceptance of new scientific discoveries than the systematic accumulation and evaluation of facts. The recognition that there is a social process involved with the acceptance or rejection of scientific knowledge has been the subject of study of sociologists for some time. The scientific community's rejection of the existence of paranormal phenomena is an excellent case study of this process (Allison, 1979; Collins and Pinch, 1979).

Implicit in Professor Utts' presentation and paramount to the acceptance of parapsychology as a legitimate science are the description and documentation of the professionalization of the field of parapsychology. It is true that many researchers in the field have university appointments; there are organized professional societies for the advancement of parapsychology; there are journals with rigorous standards for published research; the field has received funding from federal agencies; and parapsychology has received recognition from other professional societies, such as the IMS and the American Association for the Advancement of Science (Collins and Pinch, 1979). Nevertheless, most readers of *Statistical Science* would agree that parapsychology is not accepted as part of orthodox science and is considered by most of the scientific community to be on the margins of science, at best (Allison, 1979; Collins and Pinch, 1979). Why is this the case? Professor Utts believes that it is because people have not examined the data. She states that "Strong beliefs tend to be resistant to change even in the face of data, and many people, scientists included, seem to have made up their minds on the question without examining any empirical data at all."

The history of science is replete with examples of resistance by the established scientific community to new discoveries. A challenging problem for science is to understand the process by which a new theory or discovery becomes accepted by the community of scientists and, likewise, to characterize the nature of the resistance to new ideas. Barber (1961) suggests that there are many different sources of resistance to scientific discovery. In 1900, for example, Karl Pearson met resistance to his use of statistics in applications to biological problems, illustrating a source of resistance due to the use of a particular methodology. The Royal Society informed Pearson that future papers submitted to the Society for publication must keep the mathematics separate from the biological applications.

Another obvious source of resistance to new scientific ideas, and the one referred to by Professor Utts above, is the prevailing substantive beliefs and theories held by scientists at any given time. Barber offers the opposition to Copernicus and his heliocentric theory and to Mendel's theory of genetic inheritance as examples of how, because of preconceived ideas, theories and values, scientists are not as open-minded to new advances as one might think they should be. It was R. A. Fisher who said that each generation seems to have found in Mendel's paper only what it expected to find and ignored what did not conform to its own expectations (Fisher, 1936).

Pearson's response to the antimathematical prejudice expressed by the Royal Society was to establish with Galton's support a new journal, *Biometrika*, to encourage the use of mathematics in biology. Galton (1901) wrote an article for the first issue of the journal, explaining the need for this new voice of "mutual encouragement and support" for mathematics in biology and saying that "a new science cannot depend on a welcome from the followers of the older ones, and [therefore]... it is advisable to establish a special Journal for Biometry." Lavoisier understood the role of preconceived beliefs as a source of resistance when he wrote in 1785,

> I do not expect my ideas to be adopted all at once. The human mind gets creased into a way of seeing things. Those who have envisaged nature according to a certain point of view during much of their career, rise only with difficulty to new ideas. (Barber, 1961.)

I suspect that this paper by Professor Utts synthesizing the accumulation of research results supporting the existence of paranormal phenomena will continue to be received with skepticism by the orthodox scientific community "even after examining the data." In part, this resistance is due to the popular perception of the association between parapsychology and the occult (Allison, 1979) and due to the continued suspicion and documentation of fraud in parapsychology (Diaconis, 1978). An additional and important source of resistance to the evidence presented by Professor Utts, however, is the lack of a model to explain the phenomena. Psychic phenomena are unexplainable by any current scientific theory and, furthermore, directly contradict the laws of physics. Acceptance of psi implies the rejection of a large body of accumulated evidence explaining the physical and biological world as we know it. Thus, even though the effect size for a relationship between aspirin and the prevention of heart attacks is three times smaller than the effect size observed in the ganzfeld data

base, it is the existence of a biological mechanism to explain the effectiveness of aspirin that accounts, in part, for acceptance of this relationship.

In evaluating the evidence in favor of the existence of paranormal phenomena, it is necessary to consider alternative explanations or hypotheses for the results and, as noted by Cornfield (1959), "If important alternative hypotheses are compatible with available evidence, then the question is unsettled, even if the evidence is experimental" (see also Platt, 1964). Many of the experimental results reported by Professor Utts need to be considered in the context of explanations other than the existence of paranormal phenomena. Consider the following examples:

(1) In the various psi experiments that Professor Utts discusses, the null hypothesis is a simple chance model. However, as noted by Diaconis (1978) in a critique of parapsychological research, "In complex, badly controlled experiments simple chance models cannot be seriously considered as tenable explanations: hence, rejection of such models is not of particular interest." Diaconis shows that the underlying probabilistic model in many of these experiments (even those that are well controlled) is much more complicated than chance.

(2) The role that experimenter expectancy plays in the reporting and interpreting of results cannot be underestimated. Rosenthal (1966), based on a meta-analysis of the effects of experimenters' expectancies on the results of their research, found that experimenters tended to get the results they expected to get. Clearly this is an important potential confounder in parapsychological research. Professor Utts comments on a debate between Honorton and Hyman, parapsychologist and critic, respectively, regarding evidence for psi abilities, and, although not necessarily a result of experimenter expectancy, describes how "... each analyzed the results of all known psi ganzfeld experiments to date, and reached strikingly different conclusions."

(3) What is an acceptable response in these experiments? What constitutes a direct hit? What if the response is close, who decides whether or not that constitutes a hit (see (2) above)? In an example of a response of a Receiver in an automated ganzfeld procedure, Professor Utts describes the "dream-like quality of the mentation." Someone must evaluate these stream-of-consciousness responses to determine what is a hit. An important methodological question is: How sensitive are the results to different definitions of a hit?

(4) In describing the results of different meta-analyses, Professor Utts is careful to raise questions about the role of publication bias. Publication bias or "the file-drawer problem" arises when only statistically significant findings get published, while statistically nonsignificant studies sit unreported in investigators' file drawers. Typically, Rosenthal's method (1979) is used to calculate the "fail-safe N," that is, the number of unreported studies that would have to be sitting in file-drawers in order to negate the significant effect. Iyengar and Greenhouse (1988) describe a modification of Rosenthal's method, however, that gives a fail-safe N that is often an order of magnitude smaller than Rosenthal's method, suggesting that the sensitivity of the results of meta-analyses of psi experiments to unpublished negative studies is greater than is currently believed.

Even if parapsychology is thought to be on the margins of science by the scientific community, parapsychologists should not be held to a different standard of evidence to support their findings than orthodox scientists, but like other scientists they must be concerned with spurious effects and the effects of extraneous variables. The experimental results summarized by Professor Utts appear to be sensitive to the effect of alternative hypotheses like the ones described above. Sensitivity analyses, which question, for example, how large of an effect due to experimenter expectancy there would have to be to account for the effect sizes being reported in the psi experiments, are not addressed here. Again, the ability to account for and eliminate the role of alternative hypotheses in explaining the observed relationship between aspirin and the prevention of heart attacks is another reason for the acceptance of these results.

A major new technology discussed by Professor Utts in synthesizing the experimental parapsychology literature is meta-analysis. Until recently, the quantitative review and synthesis of a research literature, that is, meta-analysis, was considered by many to be a questionable research tool (Wachter, 1988). Resistance by statisticians to meta-analysis is interesting because, historically, many prominent statisticians found the combining of information from independent studies to be an important and useful methodology (see, e.g., Fisher, 1932; Cochran, 1954; Mosteller and Bush, 1954; Mantel and Haenszel, 1959). Perhaps the more recent skepticism about meta-analysis is because of its use as a tool to advance discoveries that themselves were the objects of resistance, such as the efficacy of psychotherapy (Smith and Glass, 1977) and now the existence of paranormal phenomena. It is an interesting problem for the history of science to explore why and when in the development of a

of a discipline it turns to meta-analysis to answer research questions or to resolve controversy (e.g., Greenhouse et al., 1990).

One argument for combining information from different studies is that a more powerful result can be obtained than from a single study. This objective is implicit in the use of meta-analysis in parapsychology and is the force behind Professor Utts' paper. The issue is that by combining many small studies consisting of small effects there is a gain in power to find an overall statistically significant effect. It is true that the meta-analyses reported by Professor Utts find extremely small p-values, but the estimate of the overall effect size is still small. As noted earlier, because of the small magnitude of the overall effect size, the possibility that other extraneous variables might account for the relationship remains.

Professor Utts, however, also illustrates the use of meta-analysis to investigate how studies differ and to characterize the influence of difficult covariates or moderating variables on the combined estimate of effect size. For example, she compares the mean effect size of studies where subjects were selected on the basis of good past performance to studies where the subjects were unselected, and she compares the mean effect size of studies with feedback to studies without feedback. To me, this latter use of meta-analysis highlights the more valuable and important contribution of the methodology. Specifically, the value of quantitative methods for research synthesis is in assessing the potential effects of study characteristics and to quantify the sources of heterogeneity in a research domain, that is, to study systematically the effects of extraneous variables. Tom Chalmers and his group at Harvard have used meta-analysis in just this way not only to advance the understanding of the effectiveness of medical therapies but also to study the characteristics of good research in medicine, in particular, the randomized controlled clinical trial. (See Mosteller and Chalmers, 1991, for a review of this work.)

Professor Utts should be congratulated for her courage in contributing her time and statistical expertise to a field struggling on the margins of science, and for her skill in synthesizing a large body of experimental literature. I have found her paper to be quite stimulating, raising many interesting issues about how science progresses or does not progress.

ACKNOWLEDGMENT

This work was supported in part by MHCRC grant MH30915 and MH15758 from the National Institute of Mental Health, and CA54852 from the National Cancer Institute. I would like to acknowledge stimulating discussions with Professors Larry Hedges, Michael Meyer, Ingram Olkin, Teddy Seidenfeld and Larry Wasserman, and thank them for their patience and encouragement while preparing this discussion.

Comment

Ray Hyman

Utts concludes that "there is an anomaly that needs explanation." She bases this conclusion on the ganzfeld experiments and four meta-analyses of parapsychological studies. She argues that both Honorton and Rosenthal have successfully refuted my critique of the ganzfeld experiments. The meta-analyses apparently show effects that cannot be explained away by unreported experiments nor over-analysis of the data. Furthermore, effect size does not correlate with the rated quality of the experiment.

Neither time nor space is available to respond in detail to her argument. Instead, I will point to some of my concerns. I will do so by focusing on those parts of Utts' discussion that involve me. Understandably, I disagree with her assertions that both Honorton and Rosenthal successfully refuted my criticisms of the ganzfeld experiments.

Her treatment of both the ganzfeld debate and the National Research Council's report suggests that Utts has relied on second-hand reports of the data. Some of her statements are simply inaccurate. Others suggest that she has not carefully read what my critics and I have written. This remoteness from the actual experiments and details of the arguments may partially account for her optimistic assessment of the results. Her paper takes

Ray Hyman is Professor of Psychology, University of Oregon, Eugene, Oregon 97403.

the reported data at face value and focuses on the statistical interpretation of these data.

Both the statistical interpretation of the results of an individual experiment and of the results of a meta-analysis are based on a model of an ideal world. In this ideal world, effect sizes have a tractable and known distribution and the points in the sample space are independent samples from a coherent population. The appropriateness of any statistical application in a given context is an empirical matter. That is why such issues as the adequacy of randomization, the non-independence of experiments in a meta-analysis and the over-analysis of data are central to the debate. The optimistic conclusions from the meta-analyses assume that the effect sizes are unbiased estimates from independent experiments and have nicely behaved distributional properties.

Before my detailed assessment of all the available ganzfeld experiments through 1981, I accepted the assertions by parapsychologists that their experiments were of high quality in terms of statistical and experimental methodology. I was surprised to find that the ganzfeld experiments, widely heralded as the best exemplar of a successful research program in parapsychology, were characterized by obvious possibilities for sensory leakage, inadequate randomization, over-analysis and other departures from parapsychology's own professed standards. One response was to argue that I had exaggerated the number of flaws. But even internal critics agreed that the rate of defects in the ganzfeld data base was too high.

The other response, implicit in Utts' discussion of the ganzfeld experiments and the meta-analyses, was to admit the existence of the flaws but to deny their importance. The parapsychologists doing the meta-analysis would rate each experiment for quality on one or more attributes. Then, if the null hypothesis of no correlation between effect size and quality were upheld, the investigators concluded that the results could not be attributed to defects in methodology.

This retrospective sanctification using statistical controls to compensate for inadequate experimental controls has many problems. The quality ratings are not blind. As the differences between myself and Honorton reveal, such ratings are highly subjective. Although I tried my best to restrict my ratings to what I thought were objective and easily codeable indicators, my quality ratings provide a different picture than do those of Honorton. Honorton, I am sure, believes he was just as objective in assigning his ratings as I believe I was.

Another problem is the number of different properties that are rated. Honorton's ratings of quality omitted many attributes that I included in my ratings. Even in those cases where we used the same indicators to make our assessments, we differed because of our scaling. For example, on adequacy of randomization I used a simple dichotomy. Either the experimenter clearly indicated using an appropriate randomization procedure or he did not. Honorton converted this to a trichotomous scale. He distinguished between a clearly inadequate procedure such as hand-shuffling and failure to report how the randomization was done. He then assigned the lowest rating to failure to describe the randomization. In his scheme, clearly inadequate randomization was of higher quality than failure to describe the procedure. Although we agreed on which experiments had adequate randomization, inadequate randomization or inadequate documentation, the different ways these were ordered produced important differences between us in how randomization related to effect size. These are just some of the reasons why the finding of no correlation between effect size and rated quality does not justify concluding that the observed flaws had no effect.

I will now consider some of Utts' assertions and hope that I can go into more detail in another forum. Utts discusses the conclusions of the National Research Council's Committee on Techniques for the Enhancement of Human Performance. I was chairperson of that committee's subcommittee on paranormal phenomena. She wrongly states that we restricted our evaluation only to significant studies. I do not know how she got such an impression since we based our analysis on meta-analyses whenever these were available. The two major inputs for the committee's evaluation were a lengthy evaluation of contemporary parapsychology experiments by John Palmer and an independent assessment of these experiments by James Alcock. Our sponsors, the Army Research Institute had commissioned the report from the parapsychologist John Palmer. They specifically asked our committee to provide a second opinion from a non-parapsychological perspective. They were most interested in the experiments on remote viewing and random number generators. We decided to add the ganzfeld experiments. Alcock was instructed, in making his evaluation, to restrict himself to the same experiments in these categories that Palmer had chosen. In this way, the experiments we evaluated, which included both significant and nonsignificant ones, were, in effect, selected for us by a prominent parapsychologist.

Utts mistakenly asserts that my subcommittee on parapsychology commissioned Harris and Rosenthal to evaluate parapsychology experiments for

us. Harris and Rosenthal were commissioned by our evaluation subcommittee to write a paper on evaluation issues, especially those related to experimenter effects. On their own initiative, Harris and Rosenthal surveyed a number of data bases to illustrate the application of methodological procedures such as meta-analysis. As one illustration, they included a meta-analysis of the subsample of ganzfeld experiments used by Honorton in his rebuttal to my critique.

Because Harris and Rosenthal did not themselves do a first-hand evaluation of the ganzfeld experiments, and because they used Honorton's ratings for their illustration, I did not refer to their analysis when I wrote my draft for the chapter on the paranormal. Rosenthal told me, in a letter, that he had arbitrarily used Honorton's ratings rather than mine because they were the most recent available. I assumed that Harris and Rosenthal were using Honorton's sample and ratings to illustrate meta-analytic procedures. I did not believe they were making a substantive contribution to the debate.

Only after the committee's complete report was in the hands of the editors did someone become concerned that Harris and Rosenthal had come to a conclusion on the ganzfeld experiments different from the committee. Apparently one or more committee members contacted Rosenthal and asked him to explain why he and Harris were dissenting.

Because some committee members believed that we should deal with this apparent discrepancy, I contacted Rosenthal and pointed out if he had used my ratings with *the very same analysis* he had applied to Honorton's ratings, he would have reached a conclusion opposite to what Harris and he had asserted. I did this, not to suggest my ratings were necessarily more trustworthy than Honorton's, but to point out how fragile any conclusions were based on this small and limited sample. Indeed, the data were so lacking in robustness that the difference between my rating and Honorton's rating of one investigator (Sargent) on one attribute (randomization) sufficed to reverse the conclusions Harris and Rosenthal made about the correlation between quality and effect size.

Harris and Rosenthal responded by adding a footnote to their paper. In this footnote, they reported an analysis using my ratings rather than Honorton's. This analysis, they concluded, still supported the null hypothesis of no correlation between quality and effect size. They used 6 of my 12 dichotomous ratings of flaws as predictors and the z score and effect size as criterion variables in both multiple regression and canonical correlation analyses. They reported an "adjusted" canonical correlation between criterion variables and flaws of "only" 0.46. A true correlation of this magnitude would be impressive given the nature and split of the dichotomous variables. But, because it was not statistically significant, Harris and Rosenthal concluded that there was no relationship between quality and effect size. A canonical correlation on this sample of 28 nonindependent cases, of course, has virtually no chance of being significant, even if it were of much greater magnitude.

What this amounts to is that the alleged contradictory conclusions of Harris and Rosenthal are based on a meta-analysis that supports Honorton's position when Honorton's ratings are used and supports my position when my ratings are used. Nothing substantive comes from this, and it is redundant with what Honorton and I have already published. Harris and Rosenthal's footnote adds nothing because it supports the null hypothesis with a statistical test that has no power against a reasonably sized alternative. It is ironic that Utts, after emphasizing the importance of considering statistical power, places so much reliance on the outcome of a powerless test.

(I should add that the recurrent charge that the NRC committee completely ignored Harris and Rosenthal's conclusions is not strictly correct. I wrote a response to the Harris and Rosenthal paper that was included in the same supplementary volume that contains their commissioned paper.)

Utts' discussion of the ganzfeld debate, as I have indicated, also shows unfamiliarity with details. She cites my factor analysis and Saunders' critique as if these somehow jeopardized the conclusions I drew. Again, the matter is too complex to discuss adequately in this forum. The "factor analysis" she is talking about is discussed in a few pages of my critique. I introduced it as a convenient way to summarize my conclusions, *none of which depended on this analysis*. I agree with what Saunders has to say about the limitations of factor analysis in this context. Unfortunately, Saunders bases his criticism on wrong assumptions about what I did and why I did it. His dismissal of the results as "meaningless" is based on mistaken algebra. I included as dummy variables five experimenters in the factor analysis. Because an experimenter can only appear on one variable, this necessarily forces the average intercorrelation among the experimenter variables to be negative. Saunders falsely asserts that this negative correlation must be -1. If he were correct, this would make the results meaningless. But he could be correct only if there were just two investigators and that each one accounted for 50% of the experiments. In my case, as I made sure to check ahead of time, the use of five

experimenters, each of whom contributed only a few studies to the data base, produced a mildly negative intercorrelation of −0.147. To make sure even that small correlation did not distort the results, I did the factor analysis with and without the dummy variables. The same factors were obtained in both cases.

However, I do not wish to defend this factor analysis. None of my conclusions depend on it. I would agree with any editor who insisted that I omit it from the paper on the grounds of redundancy. I am discussing it here as another example that suggests that Utts is not familiar with some relevant details in literature she discusses.

CONCLUSIONS

Utts may be correct. There may indeed be an anomaly in the parapsychological findings. Anomalies may also exist in non-parapsychological domains. The question is when is an anomaly worth taking seriously. The anomaly that Utts has in mind, if it exists, can be described only as a departure from a generalized statistical model. From the evidence she presents, we might conclude that we are dealing with a variety of different anomalies instead of one coherent phenomenon. Clearly, the reported effect sizes for the experiments with random number generators are orders of magnitude lower than those for the ganzfeld experiments. Even within the same experimental domain, the effect sizes do not come from the same population. The effects sizes obtained by Jahn are much smaller than those obtained by Schmidt with similar experiments on random number generators. In the ganzfeld experiments, experimenters differ significantly in the effect sizes each obtains.

This problem of what effect sizes are and what they are measuring points to a problem for parapsychologists. In other fields of science such as astronomy, an "anomaly" is a very precisely specified departure from a well-established substantive theory. When Leverrier discovered Neptune by studying the perturbations in the orbit of Uranus, he was able to characterize the anomaly as a very precise departure of a specific kind from the orbit expected on the basis of Newtonian mechanics. He knew exactly what he had to account for.

The "anomaly" or "anomalies" that Utts talks about are different. We do not know what it is that we are asked to account for other than something that sometimes produces nonchance departures from a statistical model, whose appropriateness is itself open to question.

The case rests on a handful of meta-analyses that suggest effect sizes different from zero and uncorrelated with some non-blindly determined indices of quality. For a variety of reasons, these retrospective attempts to find evidence for paranormal phenomena are problematical. At best, they should provide the basis for parapsychologists designing prospective studies in which they can specify, in advance, the complete sample space and the critical region. When they get to the point where they can specify this along with some boundary conditions and make some reasonable predictions, then they will have demonstrated something worthy of our attention.

In this context, I agree with Utts that Honorton's recent report of his automated ganzfeld experiments is a step in the right direction. He used the ganzfeld meta-analyses and the criticisms of the existing data base to design better experiments and make some predictions. Although he and Utts believe that the findings of meaningful effect sizes in the dynamic targets and a lack of a nonzero effect size in the static targets are somehow consistent with previous ganzfeld results, I disagree. I believe the static targets are closer in spirit to the original data base. But this is a minor criticism.

Honorton's experiments have produced intriguing results. If, as Utts suggests, independent laboratories can produce similar results with the same relationships and with the same attention to rigorous methodology, then parapsychology may indeed have finally captured its elusive quarry. Of course, on several previous occasions in its century-plus history, parapsychology has felt it was on the threshold of a breakthrough. The breakthrough never materialized. We will have to patiently wait to see if the current situation is any different.

Comment

Robert L. Morris

Experimental sciences by their nature have found it relatively easy to deal with simple closed systems. When they come to study more complex, open systems, however, they have more difficulty in generating testable models, must rely more on multivariate approaches, have more diversity from experiment to experiment (and thus more difficulty in constructing replication attempts), have more noise in the data, and more difficulty in constructing a linkage between concept and measurement. Data gatherers and other researchers are more likely to be part of the system themselves. Examples include ecology, economics, social psychology and parapsychology. Parapsychology can be regarded as the study of apparent new means of communication, or transfer of influence, between organism and environment. Any observer attempting to decide whether or not such psychic communication has taken place is one of several elements in a complex open system composed of an indefinite number of interactive features. The system can be modeled, as has been done elsewhere (e.g., Morris, 1986) such as to organise our understanding of how observers can be misled by themselves, or by deliberate frauds. Parapsychologists designing experimental studies must take extreme care to ensure that the elements in the experimental system do not interact in unanticipated ways to produce artifact or encourage fraudulent procedures. When researchers follow up the findings of others, they must ensure that the new experimental system sufficiently resembles the earlier one, regarding its important components and their potential interactions. Specifying sufficient resemblance is more difficult in complex and open systems, and in areas of research using novel methodologies.

As a result, parapsychology and other such areas may well profit from the application of modern meta-analysis, and meta-analytic methods may in turn profit from being given a good stiff workout by controversial data bases, as suggested by Jessica Utts in her article. Parapsychology would appear to gain from meta-analytic techniques, in at least three important areas.

First, in assessing the question of replication rate, the new focus on effect size and confidence intervals rather than arbitrarily chosen significance levels seems to indicate much greater consistency in the findings than has previously been claimed.

Second, when one codes the individual studies for flaws and relates flaw abundance with effect size, there appears to be little correlation for all but one data base. This contradicts the frequent assertion that parapsychological results disappear when methodology is tightened. Additional evidence on this point is the series of studies by Honorton and associates using an automated ganzfeld procedure, apparently better conducted than any of the previous research, which nevertheless obtained an effect size very similar to that of the earlier more diverse data base.

Third, meta-analysis allows researchers to look at moderator variables, to build a clearer picture of the conditions that appear to produce the strongest effects. Research in any real scientific discipline must be cumulative, with later researchers building on the work of those who preceded them. If our earlier successes and failures have meaning, they should help us obtain increasingly consistent, clearer results. If psychic ability exists and is sufficiently stable that it can be manifest in controlled experimental studies, then moderator variables should be present in groups of studies that would indicate conditions most favourable and least favourable to the production of large effect sizes. From the analyses presented by Utts, for instance, it seems evident that group studies tend to produce poor results and, however convenient it may be to conduct them, future researchers should apparently focus much more on individual testing. When doing ganzfeld studies, it appears best to work with dynamic rather than static target material and with experienced participants rather than novices. If such results are valid, then future researchers who wish to get strong results now have a better idea of what procedures to select to increase the likelihood of so doing, what elements in the experimental system seem most relevant. The proportion of studies obtaining positive results should therefore increase.

However, the situation may be more complex than the somewhat ideal version painted above. As noted earlier, meta-analysis may learn from parapsychology as well as vice versa. Parapsychological data may well give meta-analytic techniques a good workout and will certainly pose some challenges. None of the cited meta-analyses, as described above, apparently employed more than one judge or

Robert L. Morris occupies the Koestler Chair of Parapsychology in the Department of Psychology at the University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, United Kingdom.

evaluator. Certainly none of them cited any correlation values between evaluators, and the correlations between judges of research quality in other social sciences tend to be "at best around .50," according to Hunter and Schmidt (1990, page 497). Although Honorton and Hyman reported a relatively high correlation of 0.77 between themselves, they were each doing their own study and their flaw analyses did reach somewhat different conclusions, as noted by Utts. Other than Hyman, the evaluators cited by Utts tend to be positively oriented toward parapsychology; roughly speaking, all evaluators doing flaw analyses found what they might hope to find, with the exception of the PK dice data base. Were evaluators blind as to study outcome when coding flaws? No comment is made on this aspect. The above studies need to be replicated, with multiple (and blind) evaluators and reported indices of evaluator agreement. Ideally, evaluator attitude should be assessed and taken into account as well. A study with all hostile evaluators may report very high evaluator correlations, yet be a less valid study than one that employs a range of evaluators and reports lower correlations among evaluators.

But what constitutes a replication of a meta-analysis? As with experimental replications, it may be important to distinguish between exact and conceptual replications. In the former, a replicator would attempt to match all salient features of the initial analysis, from the selection of reports to the coding of features to the statistical tests employed, such as to verify that the stated original protocol had been followed faithfully and that a similar outcome results. For conceptual replication, replicators would take the stated outcome of the meta-analysis and attempt their own independent analysis, with their own initial report selection criteria, coding criteria and strategy for statistical testing, to see if similar conclusions resulted. Conceptual replication allows more room for bias and resultant debate when findings differ, but when results are similar they can be assumed to have more legitimacy. Given the strong and surprising (for many) conclusions reached in the meta-analysis reported by Utts, it is quite likely that others with strong views on parapsychology will attempt to replicate, hoping for clear confirmation or disconfirmation. The diversity of methods they are likely to employ and the resultant debates should provide a good opportunity for airing the many conceptual problems still present in meta-analysis. If results differ on moderator variables, there can come to be empirical resolution of the differences as further results unfold. With regard to flaw analysis, such analyses have already focused attention in ganzfeld research on the abundance of existing faults and how to avoid them. If results are as strong under well-controlled conditions as under sloppy ones, then additional research such as that done by Honorton and associates under tight conditions should continue to produce positive results.

In addition to the replication issue, there are some other problems that need to be addressed. So far, the assessment of moderator variables has been univariate, whereas a multivariate approach would seem more likely to produce a clearer picture. Moderator variables may covary, with each other or with flaws. For instance, in the dice data higher effect sizes were found for flawed studies and for studies with selected subjects. Did studies using special subjects use weaker procedures?

Given the importance attached to effect size and incorporating estimates of effect size in designing studies for power, we must be careful not to assume that effect size is independent of number of trials or subjects unless we have empirical reason to do so. Effect size may decrease with larger N if experimenters are stressed or bored towards the end of a long study or if there are too many trials to be conducted within a short period of time and subjects are given less time to absorb their instructions or to complete their tasks. On one occasion there is presentation of an estimated "true average effect size," (0.18 rather than 0.28) without also presenting an estimate of effect size dispersal. Future investigators should have some sense of how the likelihood that they will obtain a hit rate of 1/3 (where 1/4 is expected) will vary in accordance with conditions.

There are a few additional quibbles with particular points. In Utts' example experiment with Professor A versus Professor B, sex of professor is a possible confounding variable. When Honorton omitted studies that did not report direct hits as a measure, he may have biased his sample. Were there studies omitted that could have reported direct hits but declined to do so, conceivably because they looked at that measure, saw no results and dropped it? This objection is only with regard to the initial meta-analysis and is not relevant for the later series of studies which all used direct hits. In Honorton's meta-analysis of forced-choice precognition experiments, the comparison variables of feedback delay and time interval to target selection appear to be confounded. Studies delaying target selection cannot provide trial by trial feedback, for instance. Also, I am unsure about using an approximation to Cohen's h for assessing the effect size for the aspirin study. There would appear to be a very striking effect, with the aspirin condition heart attack rate only 55% that of the rate for the placebo condition. How was the expected proportion of

misses estimated; perhaps Cohen's h greatly underestimates effect size when very low probability events (less than 1 in 50 for heart attack in the placebo condition and less than 1 in a 100 for aspirin) are involved. I'm not a statistician and thus don't know if there is a relevant literature on this point.

The above objections should not detract from the overall value of the Utts survey. The findings she reports will need to be replicated; but even as is, they provide a challenge to some of the cherished arguments of counteradvocates, yet also challenge serious researchers to use these findings effectively as guidelines for future studies.

Comment

Frederick Mosteller

Dr. Utts's discussion stimulates me to offer some comments that bear on her topic but do not, in the main, fall into an agree-disagree mode. My references refer to her bibliography.

Let me recommend J. Edgar Coover's work to statisticians who would like to read about a pretty sequence of experiments developed and executed well before Fisher's book on experimental design appeared. Most of the standard kinds of ESP experiments (though not the ganzfeld) are carried out and reported in this 1917 book. Coover even began looking into the amount of information contained in cues such as whispers. He also worked at exposing mediums. I found the book most impressive. As Utts says in her article, the question of significance level was a puzzling one, and one we still cannot solve even though some fields seem to have standardized on 0.05.

When Feller's comments on Stuart and Greenwood's sampling experiments came out in the first edition of his book, I was surprised. Feller devotes a problem to the results of generating 25 symbols from the set a, b, c, d and e (page 45, first edition) using random numbers with 0 and 1 corresponding to a, 2 and 3 to b, etc. He asks the student to find out how often the 25 produce 5 of each symbol. He asks the student to check the results using random number tables. The answer seems to be about 1 chance in 500. In a footnote Feller then says "They [random numbers] are occasionally extraordinarily obliging: c.f. J. A. Greenwood and E. E. Stuart, Review of Dr. Feller's Critique, *Journal of Parapsychology*, vol. 4 (1940), pp. 298–319, in particular p. 306." The 25 symbols of 5 kinds, 5 of each, correspond to the cards in a parapsychology deck.

The point of page 306 is that Greenwood and Stuart on that page claim to have generated two random orders of such a deck using Tippett's table of random numbers. Apparently Feller thought that it would have taken them a long time to do it. If one assumes that Feller's way of generating a random shuffle is required, then it would indeed be unreasonable to suppose that the experiments could be carried out quickly. I wondered then whether Feller thought this was the only way to produce a random order to such a deck of cards. If you happen to know how to shuffle a deck efficiently using random numbers, it is hard to believe that others do not know. I decided to test it out and so I proposed to a class of 90 people in mathematical statistics that we find a way of using random numbers to shuffle a deck of cards. Although they were familiar with random numbers, they could not come up with a way of doing it, nor did anyone after class come in with a workable idea though several students made proposals. I concluded that inventing such a shuffling technique was a hard problem and that maybe Feller just did not know how at the time of writing the footnote. My face-to-face attempts to verify this failed because his response was evasive. I also recall Feller speaking at a scientific meeting where someone had complained about mistakes in published papers. He said essentially that we won't have any literature if mistakes are disallowed and further claimed that he always had mistakes in his own papers, hard as he tried to avoid them. It was fun to hear him speak.

Although I find Utts's discussion of replication engaging as a problem in human perception, I do always feel that people should not be expected to carry out difficult mathematical exercises in their head, off the cuff, without computers, textbooks or advisors. The kind of problem treated requires careful formulation and then careful analysis. Even

Frederick Mosteller is Roger I. Lee Professor of Mathematical Statistics, Emeritus, at Harvard University and Director of the Technology Assessment Group in the Harvard School of Public Health. His mailing address is Department of Statistics, Harvard University, Science Center, 1 Oxford Street, Cambridge, Massachusetts 02138.

after a careful analysis is completed, there can be vigorous reasonable arguments about the appropriateness of the formulation and its analysis. These investigations leave me reinforced with the belief that people cannot do hard mathematical problems in their heads, rather than with an attitude toward or against ESP investigations.

When I first became aware of the work of Rhine and others, the concept seemed to me to be very important and I asked a psychologist friend why more psychologists didn't study this field. He responded that there were too many ways to do these experiments in a poorly controlled manner. At the time, I had just discovered that when viewed with light coming from a certain angle, I could read the backs of the cards of my parapsychology deck as clearly as the faces. While preparing these remarks in 1991, I found a note on page 305 of volume 1 of *The Journal of Parapsychology* (1937) indicating that imperfections in the cards precluded their use in unscreened situations, but that improvements were on the way. Thus I sympathize with Utts's conclusion that much is to be gained by studying how to carry out such work well. If there is no ESP, then we want to be able to carry out null experiments and get no effect, otherwise we cannot put much belief in work on small effects in non-ESP situations. If there is ESP, that is exciting. However, thus far it does not look as if it will replace the telephone.

Rejoinder

Jessica Utts

I would like to thank this distinguished group of discussants for their thought-provoking contributions. They have raised many interesting and diverse issues. Certain points, such as Professor Mosteller's enlightening account of Feller's position, require no further comment. Other points indicate the need for clarification and elaboration of my original material. Issues raised by Professors Diaconis and Hyman and subsequent conversations with Robert Rosenthal and Charles Honorton have led me to consider the topic of "Satisfying the Skeptics." Since the conclusion in my paper was not that psychic phenomena have been proved, but rather that there is an anomalous effect that needs to be explained, comments by several of the discussants led me to address the question "Should Psi Research be Ignored by the Scientific Community?" Finally, each of the discussants addressed replication and modeling issues. The last part of my rejoinder comments on some of these ideas and discusses them in the context of parapsychology.

CLARIFICATION AND ELABORATION

Since my paper was a survey of hundreds of experiments and many published reports, I could obviously not provide all of the details to accompany this overview. However, there were details lacking in my paper that have led to legitimate questions and misunderstandings from several of the discussants. In this section, I address specific points raised by Professors Diaconis, Greenhouse, Hyman and Morris, by either clarifying my original statements or by adding more information from the original reports.

Points Raised by Diaconis

Diaconis raised the point that qualified skeptics and magicians should be active participants in parapsychology experiments. I will discuss this general concept in the next section, but elaborate here on the steps that were taken in this regard for the autoganzfeld experiments described in Section 5 of my paper. As reported by Honorton et al. (1990):

> Two experts on the simulation of psi ability have examined the autoganzfeld system and protocol. Ford Kross has been a professional mentalist [a magician who simulates psychic abilities] for over 20 years... Mr. Kross has provided us with the following statement: "In my professional capacity as a mentalist, I have reviewed Psychophysical Research Laboratories' automated ganzfeld system and found it to provide excellent security against deception by subjects." We have received similar comments from Daryl Bem, Professor of Psychology at Cornell University. Professor Bem is well known for his research in social and personality psychology. He is also a member of the Psychic Entertainers Association and has performed for many years as a mentalist. He vis-

ited PRL for several days and was a subject in Series 101" [pages 134–135].

Honorton has also informed me (personal communication, July 25, 1991) that several self-proclaimed skeptics have visited his laboratory and received demonstrations of the autoganzfeld procedure and that no one expressed any concern with the security arrangements.

This may not completely satisfy Professor Diaconis' objections, but it does indicate a serious effort on the part of the researchers to involve such people. Further, the original publication of the research in Section 5 followed the reporting criteria established by Hyman and Honorton (1986), thus providing much more detail for the reader than the earlier published records to which Professor Diaconis alludes.

Points Raised by Greenhouse

Greenhouse enumerated four items that offer alternative explanations for the observed anomalous effects. Three of these (items 2–4) will be addressed in this section by elaborating on the details provided in my paper. His item 1 will be addressed in a later section.

Item 2 on his list questioned the role of experimenter expectancy effects as a potential confounder in parapsychological research. While the expectations of the experimenter may influence the *reporting* of results, the ganzfeld experiments (as well as other psi experiments) are conducted in such a way that experimenter expectancy cannot account for the results themselves. Rosenthal, who Greenhouse cites as the expert in this area, addressed this in his background paper for the National Research Council (Harris and Rosenthal, 1988a) and concluded that the ganzfeld studies were adequately controlled in this regard. He also visited the autoganzfeld laboratory and was given a demonstration of that procedure.

Greenhouse's item 3, the question of what constitutes a direct hit, was addressed in my paper but perhaps needs elaboration. Although free-response experiments do generate substantial amounts of subjective data, the statistical analysis requires that the results for each trial be condensed into a single measure of whether or not a direct hit was achieved. This is done by presenting four choices to a judge (who of course does not know the correct answer) and asking the judge to decide which of the four best matches the subject's response. If the judge picks the target, a direct hit has occurred.

It is true that different judges may differ on their opinions of whether or not there has been a direct hit on any given trial, but in all cases the statistical question is the same. Under the null hypothesis, since the target is randomly selected from the four possibilities presented, the probability of a direct hit is 0.25 regardless of who does the judging. Thus, the observed anomalous effects cannot be explained by assuming there was an overoptimistic judge.

If Professor Greenhouse is suggesting that the source of judging may be a moderating variable that determines the magnitude of the demonstrated anomalous effect, I agree. The parapsychologists have considered this issue in the context of whether or not subjects should serve as judges for their own sessions, with differing opinions in different laboratories. This is an example of an area that has been suggested for further research.

Finally, Greenhouse raised the question of the accuracy of the file-drawer estimates used in the reported meta-analyses. I agree that it is instructive to examine the file-drawer estimate using more than one model. As an example, consider the 39 studies from the direct hit and autoganzfeld data bases. Rosenthal's fail-safe N estimates that there would have to be 371 studies in the file-drawer to account for the results. In contrast, the method proposed by Iyengar and Greenhouse gives a file-drawer estimate of 258 studies. Even this estimate is unrealistically large for a discipline with as few researchers as parapsychology. Given that the average number of trials per experiment is 30, this would represent almost 8000 unreported trials, and at least that many hours of work.

There are pros and cons to any method of estimating the number of unreported studies, and the actual practices of the discipline in question should be taken into account. Recognizing publication bias as an issue, the Parapsychological Association has had an official policy since 1975 against the selective reporting of positive results. Of the original ganzfeld studies reported in Section 4 of my paper, less than half were significant, and it is a matter of record that there are many nonsignificant studies and "failed replications" published in all areas of psi research. Further, the autoganzfeld database reported in Section 5 has no file-drawer. Given the publication practices and the size of the field, the proposed file-drawer cannot account for the observed effects.

Points Raised by Hyman

One of my goals in writing this paper was to present a fair account of recent work and debate in parapsychology. Thus, I was disturbed that Hyman, who has devoted much of his career to the study of parapsychology, and who had first-hand knowledge of the original published reports, be-

lieved that some of my statements were inaccurate and indicated that I had not carefully read the reports. I will address some of his specific objections and show that, except where noted, the accuracy of my original statements can be verified by further elaboration and clarification, with due apology for whatever necessary details were lacking in my original report.

Most of our points of disagreement concern the National Academy of Sciences (National Research Council) report *Enhancing Human Performance* (Druckman and Swets, 1988). This report evaluated several controversial areas, including parapsychology. Professor Hyman chaired the Parapsychology Subcommittee. Several background papers were commissioned to accompany this report, available from the "Publication on Demand Program" of the National Academy Press. One of the papers was written by Harris and Rosenthal, and entitled "Human Performance Research: An Overview."

Professor Hyman alleged that "Utts mistakenly asserts that my subcommittee on parapsychology commissioned Harris and Rosenthal to evaluate parapsychology experiments for us . . . ." I cannot find a statement in my paper that asserts that Harris and Rosenthal were commissioned by the subcommittee, nor can I find a statement that asserts that they were asked to evaluate parapsychology experiments. Nonetheless, I believe our substantive disagreement results from the fact that the work by Harris and Rosenthal was written in two parts, both of which I referenced in my paper. They were written several months apart, but published together, and each had its own history.

The first part (Harris and Rosenthal, 1988a) is the one to which I referred with the words "Rosenthal was commissioned by the National Academy of Sciences to prepare a background paper to accompany its 1988 report on parapsychology" (p. 372). According to Rosenthal (personal communication, July 23, 1991) he was asked to prepare a background paper to address evaluation issues and experimenter effects to accompany the report in five specific areas of research, including parapsychology.

The second part was a "Postscript" to the commissioned paper (Harris and Rosenthal, 1988b), and this is the one to which I referred on page 371 as "requested by Hyman in his capacity as Chair of the National Academy of Sciences' Subcommittee on Parapsychology." (It is probably this wording that led Professor Hyman to his erroneous allegation.) The postscript began with the words "We have been asked to respond to a letter from Ray Hyman, chair of the subcommittee on parapsychology, in which he raises questions about the presence and consequence of methodological flaws in the ganzfeld studies . . . ."

In reference to this postscript, I stand corrected on a technical point, because Hyman himself did not request the response to his own letter. As noted by Palmer, Honorton and Utts (1989), the postscript was added because:

> At one stage of the process, John Swets, Chair of the Committee, actually phoned Rosenthal and asked him to withdraw the parapsychology section of his [commissioned] paper. When Rosenthal declined, Swets and Druckman then requested that Rosenthal respond to criticisms that Hyman had included in a July 30, 1987 letter to Rosenthal [page 38].

A related issue on which I would like to elaborate concerns the correlation between flaws and success in the original ganzfeld data base. Hyman has misunderstood both my position and that of Harris and Rosenthal. He believes that I implicitly denied the importance of the flaws, so I will make my position explicit. I do not think there is any evidence that the experimental results were due to the identified flaws. The flaw analysis was clearly useful for delineating acceptable criteria for future experiments. Several experiments were conducted using those criteria. The results were similar to the original experiments. I believe that this indicates an anomaly in need of an explanation.

In discussing the paper and postscript by Harris and Rosenthal, Hyman stated that "The alleged contradictory conclusions [to the National Research Council report] of Harris and Rosenthal are based on a meta-analysis that supports Honorton's position when Honorton's [flaw] ratings are used and supports my position when my ratings are used." He believes that Harris and Rosenthal (and I) failed to see this point because the low power of the test associated with their analysis was not taken into account.

The analysis in question was based on a canonical correlation between flaw ratings and measures of successful outcome for the ganzfeld studies. The canonical correlation was 0.46, a value Hyman finds to be impressive. What he has failed to take into account however, is that a canonical correlation gives only the *magnitude* of the relationship, and not the *direction*. A careful reading of Harris and Rosenthal (1988b) reveals that their analysis actually *contradicted* the idea that the flaws could account for the successful ganzfeld results, since "Interestingly, three of the six flaw variables correlated positively with the flaw canonical variable

and with the outcome canonical variable but three correlated *negatively*" (page 2, italics added). Rosenthal (personal communication, July 23, 1991) verified that this was indeed the point he was trying to make. Readers who are interested in drawing their own conclusions from first-hand analyses can find Hyman's original flaw codings in an Appendix to his paper (Hyman, 1985, pages 44–49).

Finally, in my paper, I stated that the parapsychology chapter of the National Research Council report critically evaluated statistically significant experiments, but not those that were nonsignificant. Professor Hyman "does not know how [I] got such an impression," so I will clarify by outlining some of the material reviewed in that report. There were surveys of three major areas of psi research: remote viewing (a particular type of free-response experiment), experiments with random number generators, and the ganzfeld experiments. As an example of where I got the impression that they evaluated only significant studies, consider the section on remote viewing. It began by referencing a published list of 28 studies. Fifteen of these were immediately discounted, since "only 13 . . . were published under refereed auspices" (Druckman and Swets, 1988, page 179). Four more were then dismissed, since "Of the 13 scientifically reported experiments, 9 are classified as successful" (page 179). The report continued by discussing these nine experiments, never again mentioning any of the remaining 19 studies. The other sections of the report placed similar emphasis on significant studies. I did not think this was a valid statistical method for surveying a large body of research.

Minor Point Raised by Morris

The final clarification I would like to offer concerns the minor point raised by Professor Morris, that "When Honorton omitted studies that did not report direct hits as a measure, he may have biased his sample." This possibility was explicitly addressed by Honorton (1985, page 59). He examined what would happen if z-scores of zero were inserted for the 10 studies for which the number of direct hits was not measured, but could have been. He found that even with this conservative scenario, the combined z-score only dropped from 6.60 to 5.67.

SATISFYING THE SKEPTICS

Parapsychology is probably the only scientific discipline for which there is an organization of skeptics trying to discredit its work. The Committee for the Scientific Investigation of Claims of the Paranormal (CSICOP) was established in 1976 by philosopher Paul Kurtz and sociologist Marcello Truzzi when "Kurtz became convinced that the time was ripe for a more active crusade against parapsychology and other pseudo-scientists" (Pinch and Collins, 1984, page 527). Truzzi resigned from the organization the next year (as did Professor Diaconis) "because of what he saw as the growing danger of the committee's excessive negative zeal at the expense of responsible scholarship" (Collins and Pinch, 1982, page 84). In an advertising brochure for their publication *The Skeptical Inquirer*, CSICOP made clear its belief that paranormal phenomena are worthy of scientific attention only to the extent that scientists can fight the growing interest in them. Part of the text of the brochure read: "Why the sudden explosion of interest, even among some otherwise sensible people, in all sorts of paranormal 'happenings'? . . . Ten years ago, scientists started to fight back. They set up an organization—The Committee for the Scientific Investigation of Claims of the Paranormal."

During the six years that I have been working with parapsychologists, they have repeatedly expressed their frustration with the unwillingness of the skeptics to specify what would constitute acceptable evidence, or even to delineate criteria for an acceptable experiment. The Hyman and Honorton Joint Communiqué was seen as the first major step in that direction, especially since Hyman was the Chair of the Parapsychology Subcommittee of CSICOP.

Hyman and Honorton (1986) devoted eight pages to "Recommendations for Future Psi Experiments," carefully outlining details for how the experiments should be conducted and reported. Honorton and his colleagues then conducted several hundred trials using these specific criteria and found essentially the same effect sizes as in earlier work for both the overall effect and effects with moderator variables taken into account. I would expect Professor Hyman to be very interested in the results of these experiments he helped to create. While he did acknowledge that they "have produced intriguing results," it is both surprising and disappointing that he spent only a scant two paragraphs at the end of his discussion on these results.

Instead, Hyman seems to be proposing yet another set of requirements to be satisfied before parapsychology should be taken seriously. It is difficult to sort out what those requirements should be from his account: "[They should] specify, in advance, the complete sample space and the critical region. When they get to the point where they can specify this along with some boundary conditions and make some reasonable predictions, then they

will have demonstrated something worthy of our attention."

Diaconis believes that psi experiments do not deserve serious attention unless they actively involve skeptics. Presumably, he is concerned with subject or experimenter fraud, or with improperly controlled experiments. There are numerous documented cases of fraud and trickery in purported psychic phenomena. Some of these were observed by Diaconis and reported in his article in *Science*. Such cases have mainly been revealed when investigators attempted to verify the claims of individual psychic practitioners in quasi-experimental or uncontrolled conditions. These instances have received considerable attention, probably because the claims are so sensational, the fraud is so easy to detect by a skilled observer and they are an easy target for skeptics looking for a way to discredit psychic phenomena. As noted by Hansen (1990), "Parapsychology has long been tainted by the fraudulent behavior of a few of those claiming psychic abilities" (page 25).

Control against deception by subjects in the laboratory has been discussed extensively in the parapsychological literature (see, e.g., Morris, 1986, and Hansen, 1990). Properly designed experiments should preclude the possibility of such fraud. Hyman and Honorton (1986, page 355) explicitly discussed precautions to be taken in the ganzfeld experiments, all of which were followed in the autoganzfeld experiments. Further the controlled laboratory experiments discussed in my paper usually used a large number of subjects, a situation that minimizes the possibility that the results were due to fraud on the part of a few subjects. As for the possibility of experimenter fraud, it is of course an issue in all areas of science. There have been a few such instances in parapsychology, but since parapsychologists tend to be aware of this possibility, they were generally detected and exposed by insiders in the field.

It is not clear whether or not Diaconis is suggesting that a magician or "qualified skeptic" needs to be present at all times during a laboratory experiment. I believe that it would be more productive for such consultation to occur during the design phase, and during the implementation of some pilot sessions. This is essentially what was done for the autoganzfeld experiments, in which Professor Hyman, a skeptic as well as an accomplished magician, participated in the specification of design criteria, and mentalists Bem and Kross observed experimental sessions. Bem is also a well-respected experimental psychologist.

While I believe that the skeptics, particularly some of the more knowledgeable members of CSICOP, have served a useful role in helping to improve experiments, their counter-advocacy stance is counterproductive. If they are truly interested in resolving the question of whether or not psi abilities exist, I would expect them to encourage evaluation and experimentation by unbiased, skilled experimenters. Instead, they seem to be trying to discourage such interest by providing a moving target of requirements that must be satisfied first.

SHOULD PSI RESEARCH BE IGNORED BY THE SCIENTIFIC COMMUNITY?

In the conclusion of my paper, I argued that the scientific community should pay more attention to the experimental results in parapsychology. I was not suggesting that the accumulated evidence constitutes proof of psi abilities, but rather that it indicates that there is indeed an anomalous effect that needs an explanation. Greenhouse noted that my paper will not necessarily change anyone's view about the existence of paranormal phenomena, an observation with which I agree. However, I hope it will change some views about the importance of further investigation.

Mosteller and Diaconis both acknowledged that there are reasons for statisticians to be interested in studying the anomalous effects, regardless of whether or not psi is real. As noted by Mosteller, "If there is no ESP, then we want to be able to carry out null experiments and get no effect, otherwise we cannot put much belief in work on small effects in non-ESP situations." Diaconis concluded that "Parapsychology is worthy of serious study" partly because "If it is wrong, it offers a truly alarming massive case study of how statistics can mislead and be misused."

Greenhouse noted several sociological reasons for the resistance of the scientific community to accepting parapsychological phenomena. One of these is that they directly contradict the laws of physics. However, this assertion is not uniformly accepted by physicists (see, e.g., Oteri, 1975), and some of the leading parapsychological researchers hold Ph.D.s in physics.

Another reason cited by Greenhouse, and supported by Hyman, is that psychic phenomena are currently unexplainable by a unified scientific theory. But that is precisely the reason for more intensive investigation. The history of science and medicine is replete with examples where empirical departures from expectation led to important findings or theoretical models. For example, the causal connection between cigarette smoking and lung cancer was established only after years of statisti-

cal studies, resulting from the observation by one physician that his lung cancer patients who smoked did not recover at the same rate as those who did not. There are many medications in common use for which there is still no medical explanation for their observed therapeutic effectiveness, but that does not prohibit their use.

There are also examples where a coherent theory of a phenomenon was impossible because the requisite background information was missing. For instance, the current theory of endorphins as an explanation for the success of acupuncture would have been impossible before the discovery of endorphins in the 1970s.

Mosteller's observation that ESP will not replace the telephone leads to the question of whether or not psi abilities are of any use even if they do exist, since the effects are relatively small. Again, a look at history is instructive. For example, in 1938 *Fortune Magazine* reported that "At present, few scientists foresee any serious or practical use for atomic energy."

Greenhouse implied that I think parapsychology is not accepted by more of the scientific community only because they have not examined the data, but this misses the main point I was trying to make. The point is that individual scientists are willing to express an opinion without any reference to data. The interesting sociological question is why they are so resistant to examining the data. One of the major reasons is undoubtedly the perception identified by Greenhouse that there is some connection between parapsychology and the occult, or worse, religious beliefs. Since religion is clearly not in the realm of science, the very thought that parapsychology might be a science leads to what psychologists call "cognitive dissonance." As noted by Griffin (1988), "People feel unpleasantly aroused when two cognitions are dissonant—when they contradict one another" (page 33). Griffin continued by observing that there are also external reasons for scientists to discount the evidence, since "It is generally easier to be a skeptic in the face of novel evidence; skeptics may be overly conservative, but they are rarely held up to ridicule" (page 34).

In summary, while it may be safer and more consonant with their beliefs for individual scientists to ignore the observed anomalous effects, the scientific community should be concerned with finding an explanation. The explanations proposed by Greenhouse and others are simply not tenable.

REPLICATION AND MODELING

Parapsychology is one of the few areas where a point null hypothesis makes some sense. We can specify what should happen if there is no such thing as ESP by using simple binomial models, either to find p-values or Bayes factors. As noted by Mosteller, if there is no ESP, or other nonstatistical explanation for an effect, we should be able to carry out null experiments and get no effect. Otherwise, we should be worried about using these simple models for other applications.

Greenhouse, in his first alternative explanation for the results, questioned the use of these simple models, but his criticisms do not seem relevant to the experiments discussed in Section 5 of my paper. The experiments to which he referred were either poorly controlled, in which case no statistical analysis could be valid, or were specifically designed to incorporate trial by trial feedback in such a way that the analysis needed to account for the added information. Models and analyses for such experiments can be found in the references given at the end of Diaconis' discussion.

For the remainder of this discussion, I will confine myself to models appropriate for experiments such as the autoganzfeld described in Section 5. It is this scenario for which Bayarri and Berger computed Bayes factors, and for which Dawson discussed possible Bayesian models.

If ESP does exist, it is undoubtedly a gross oversimplification to use a simple non-null binomial model for these experiments. In addition to potential differences in ability among subjects, there were also observed differences due to dynamic versus static targets, whether or not the sender was a friend, and how the receiver scored on measures of extraversion. All of these differences were anticipated in advance and could be incorporated into models as covariates.

It is nonetheless instructive to examine the Bayes factor computed by Bayarri and Berger for the simple non-null binomial model. First, the observed anomalous effects would be less interesting if the Bayes factor was small for reasonable values of r, as it was for the random number generator experiments analyzed by Jefferys (1990), most of which purported to measure psychokinesis instead of ESP. Second, the Bayes factor provides a rough measure of the strength of the evidence against the null hypothesis and is a much more sensible summary than the p-value. The Bayes factors provided by Bayarri and Berger are probably more conservative, in the sense of favoring the null hypothesis, than those that would result from priors elicited from parapsychologists, but are probably reasonable for those who know nothing about past observed effects. I expect tht most parapsychologists would not opt for a prior symmetric around chance, but would still choose one with some mass below

chance. The final reason it is instructive to examine these Bayes factors is that they provide a quantitative challenge to skeptics to be explicit about their prior probabilities for the null and alternative hypotheses.

Dawson discussed the use of more complex Bayesian models for the analysis of the autoganzfeld data. She proposed a hierarchical model where the number of successes for each experiment followed a binomial distribution with hit rate p_i, and logit(p_i) came from a normal distribution with noninformative priors for the mean and variance. She then expanded this model to include heavier tails by allowing an additional scale parameter for each experiment. Her rationale for this expanded model was that there were clear outlier series in the data.

The hierarchical model proposed by Dawson is a reasonable place to start given only that there were several experiments trying to measure the same effect, conducted by different investigators. In the autoganzfeld database, the model could be expanded to incorporate the additional information available. Each experiment contained some sessions with static targets and some with dynamic targets, some sessions in which the sender and receiver were friends and others in which they were not and some information about the extraversion score of the receiver. All of this information could be included by defining the individual session as the unit of analysis, and including a vector of covariates for each session. It would then make sense to construct a logistic regression model with a component for each experiment, following the model proposed by Dawson, and a term $X\beta$ to include the covariates. A prior distribution for β could include information from earlier ganzfeld studies. The advantage of using a Bayesian approach over a simple logistic regression is that information could be continually updated. Some of the recent work in Bayesian design could then be incorporated so that future trials make use of the best conditions.

Several of the discussants addressed the concept of replication. I agree with Mosteller's implication that it was unwise for the audience in my seminar to respond to my replication questions so quickly, and that was precisely my point. Most nonstatisticians do not seem to understand the complexity of the replication question. Parenthetically, when I posed the same scenario to an audience of statisticians, very few were willing to offer a quick opinion.

Bayarri and Berger provided an insightful discussion of the purpose of replication, offering quantitative answers to questions that were implicit in my discussion. Their analyses suggest some alternatives to power analysis that might be considered when designing a new study to try to replicate a questionable result.

Morris addressed the question of what constitutes a replication of a meta-analysis. He distinguished between exact and conceptual replications. Using his distinction, the autoganzfeld meta-analysis could be viewed as a conceptual replication of the earlier ganzfeld meta-analysis. He noted that when such a conceptual replication offers results similar to those of the original meta-analysis, it lends legitimacy to the original results, as was the case with the autoganzfeld meta-analysis.

Greenhouse and Morris both noted the value of meta-analysis as a method of comparing different conditions, and I endorse that view. Conditions found to produce different effects in one meta-analysis could be explicitly studied in a conceptual replication. One of the intriguing results of the autoganzfeld experiments was that they supported the distinction between effect sizes for dynamic versus static targets found in the earlier ganzfeld work, and they supported the relationship between ESP and extraversion found in the meta-analysis by Honorton, Ferrari and Bem (1990).

Most modern parapsychologists, as indicated by Morris, recognize that demonstrating the validity of their preliminary findings will depend on identifying and utilizing "moderator variables" in future studies. The use of such variables will require more complicated statistical models than the simple binomial models used in the past. Further, models are needed for combining results from several different experiments, that don't oversimplify at the expense of lost information.

In conclusion, the anomalous effect that persists throughout the work reviewed in my paper will be better understood only after further experimentation that takes into account the complexity of the system. More realistic, and thus more complex, models will be needed to analyze the results of those experiments. This presents a challenge that I hope will be welcomed by the statistics community.

ADDITIONAL REFERENCES

ALLISON, P. (1979). Experimental parapsychology as a rejected science. *The Sociological Review Monograph* **27** 271-291.

BARBER, B. (1961). Resistance by scientists to scientific discovery. *Science* **134** 596-602.

BERGER, J. O. and DELAMPADY, M. (1987). Testing precise hypotheses (with discussion). *Statist. Sci.* **2** 317-352.

CHUNG, F. R. K., DIACONIS, P., GRAHAM, R. L. and MALLOWS, C. L. (1981). On the permanents of compliments of the direct sum of identity matrices. *Adv. Appl. Math.* **2** 121-137.

COCHRAN, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10** 101–129.

COLLINS, H. and PINCH, T. (1979). The construction of the paranormal: Nothing unscientific is happening. *The Sociological Review Monograph* **27** 237–270.

COLLINS, H. M. and PINCH, T. J. (1982). *Frames of Meaning: The Social Construction of Extraordinary Science.* Routledge & Kegan Paul, London.

CORNFIELD, J. (1959). Principles of research. *American Journal of Mental Deficiency* **64** 240–252.

DEMPSTER, A. P., SELWYN, M. R. and WEEKS, B. J. (1983). Combining historical and randomized controls for assessing trends in proportions. *J. Amer. Statist. Assoc.* **78** 221–227.

DIACONIS, P. and GRAHAM, R. L. (1981). The analysis of sequential experiments with feedback to subjects. *Ann. Statist.* **9** 236–244.

FISHER, R. A. (1932). *Statistical Methods for Research Workers,* 4th ed. Oliver and Boyd, London.

FISHER, R. A. (1935). Has Mendel's work been rediscovered? *Ann. of Sci.* **1** 116–137.

GALTON, F. (1901–2). Biometry. *Biometrika* **1** 7–10.

GREENHOUSE, J., FROMM, D., IYENGAR, S., DEW, M. A., HOLLAND, A. and KASS, R. (1990). Case study: The effects of rehabilitation therapy for aphasia. In *The Future of Meta-Analysis* (K. W. Wachter and M. L. Straf, eds.) 31–32. Russell Sage Foundation, New York.

GRIFFIN, D. (1988). Intuitive judgment and the evaluation of evidence. In *Enhancing Human Performance: Issues, Theories and Techniques Background Papers—Part I.* National Academy Press, Washington, D.C.

HANSEN, G. (1990). Deception by subjects in psi research. *Journal of the American Society for Psychical Research* **84** 25–80.

HUNTER, J. and SCHMIDT, F. (1990). *Methods of Meta-Analysis.* Sage, London.

IYENGAR, S. and GREENHOUSE, J. (1988). Selection models and the file drawer problem (with discussion). *Statist. Sci.* **3** 109–135.

LOUIS, T. A. (1984). Estimating an ensemble of parameters using Bayes and empirical Bayes methods. *J. Amer. Statist. Assoc.* **79** 393–398.

MANTEL, N. and HAENSZEL, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22** 719–748.

MORRIS, C. (1983). Parametric empirical Bayes inference: Theory and applications (rejoinder) *J. Amer. Statist. Assoc.* **78** 47–65.

MORRIS, R. L. (1986). What psi is not: The necessity for experiments. In *Foundations of Parapsychology* (H. L. Edge, R. L. Morris, J. H. Rush and J. Palmer, eds.) 70–110. Routledge & Kegan Paul, London.

MOSTELLER, F. and BUSH R. R. (1954). Selected quantitative techniques. In *Handbook of Social Psychology* (G. Lindzey, ed.) **1** 289–334. Addison-Wesley, Cambridge, Mass.

MOSTELLER, F. and CHALMERS, T. (1991). Progress and problems in meta-analysis. *Statist. Sci.* To appear.

OTERI, L., ed. (1975). *Quantum Physics and Parapsychology.* Parapsychology Foundation, New York.

PINCH, T. J. and COLLINS, H. M. (1984). Private science and public knowledge: The Committee for the Scientific Investigation of Claims of the Paranormal and its use of the literature. *Social Studies of Science* **14** 521–546.

PLATT, J. R. (1964). Strong inference. *Science* **146** 347–353.

ROSENTHAL, R. (1966). *Experimenter Effects in Behavioral Research.* Appleton-Century-Crofts, New York.

ROSENTHAL, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin* **86** 638–641.

RYAN, L. M. and DEMPSTER, A. P. (1984). Weighted normal plots. Technical Report 394Z, Dana-Farber Cancer Inst., Boston, Mass.

SAMANIEGO, F. J. and UTTS, J. (1983). Evaluating performance in continuous experiments with feedback to subjects. *Psychometrika* **48** 195–209.

SMITH, M. and GLASS, G. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist* **32** 752–760.

WACHTER, K. (1988). Disturbed by meta-analysis? *Science* **241** 1407–1408.

WEST, M. (1985). Generalized linear models: Scale parameters, outlier accommodation and prior distributions. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.) 531–558. North-Holland Amsterdam.