*Journal of Scientific Exploration,* Vol. 13, No. 4, pp.615–638, 1999 0892-3310/99


# The Significance of Statistics in Mind-Matter Research

JESSICA UTTS

*Division of Statistics, One Shields Ave.*
*University of California, Davis, CA 95616*

**Abstract** — Statistical methods are designed to detect and measure relationships and effects in situations where results cannot be identically replicated because of natural variability in the measurements of interest. They are generally used as an intermediate step between anecdotal evidence and the determination of causal explanations. Many anomalous phenomena, such as remote viewing or the possible effect of prayer on healing, are amenable to rigorous study. Statistical methods play a major role in making appropriate conclusions from those studies. This paper examines the role statistics can play in summarizing and drawing conclusions from individual and collective studies. Two examples of using meta-analysis to assess evidence are presented and compared. One is a conventional example relating the use of antiplatelets to reduced vascular disease, and the other is an example from mind-matter research, illustrating results of ganzfeld and remote viewing experiments.

*Keywords:* statistical evidence — *p*-values — meta-analysis — repeatability

## 1. Statistics and Anomalous Phenomena

As with any domain, the ease with which anomalous phenomena can be studied using traditional scientific methods depends on the type of purported evidence for the phenomena. The evidence tends to fall into two categories. In one category, including areas such as alien abductions and reincarnation, evidence is completely anecdotal and it is not possible to design situations that invite these phenomena to occur on demand. The second category, of concern in this paper, includes topics that can be invited to occur on demand. This category includes purported abilities such as telepathy, clairvoyance or precognition, the possibility of distant healing through prayer (*e.g.* Sicher *et al.*, 1998), and so on. The common theme is that the phenomena can be requested in randomized controlled experiments, and the results can be measured and compared to what would be expected by chance alone. It is this type of situation for which statistical methods are generally applicable.

## 2. Statistics and the Scientific Process

Throughout this paper the terms "statistics" and "statistical methods" are used in the broad context of an academic subject area including the design,

data collection, and analysis of studies involving randomization or natural variability. A standard definition is:

> Statistics is a collection of procedures and principles for gaining and processing information in order to make decisions when faced with uncertainty (Utts, 1999, p. 3).

The scientific process is generally considered to occur in two phases, one of discovery and one of justification (*e.g.* Hanson, 1958). Statistical methods most often play an important role in the discovery phase. These methods are an intermediate step between the anecdotal evidence or theoretical speculations that lead to discovery research, and the justification phase of the research process in which elaborated theories and comprehensive understanding are established.

Whether in medicine, parapsychology or some other field, most discovery research is initiated because anecdotal evidence, theory based on previous research, or analogies from other domains suggest a possible relationship or effect. For instance, there have been reports of precognitive visions and dreams throughout recorded history, so researchers are attempting to reproduce the precognitive effect in the laboratory. In medicine, theory would suggest that aspirin and similar drugs might help reduce the chances of a heart attack because they tend to thin the blood. So researchers have designed randomized controlled experiments to compare the use of aspirin-type drugs to placebos for reducing the occurrence of vascular disease (*e.g.* Antiplatelet Trialists Collaboration, 1988). Prior research on cortical pathways in the brain led psychologists to predict that listening to classical music might enhance spatial-temporal reasoning. So they designed a randomized experiment to test that hypothesis, and indeed found better spatial abilities in participants after listening to Mozart than after silence or listening to a relaxation tape (Rauscher, Shaw and Ky, 1993). The "cause" of the effect is not clear. Scientists are continuing the discovery phase by investigating the impact of different types of musical experience on spatial reasoning (such as listening to music or teaching children to play an instrument, *e.g.* Rauscher *et al.*, 1997) in order to formulate more specific theories.

In each case, the justification phase of research would follow only after reasonable theories had been formulated based on the statistical results of the discovery phase. For example, the discovery phase for the reduction in heart attacks after taking aspirin has included a variety of studies using different drug formulations and doses, various vascular diseases and levels of health, and so on. The justification phase will come after enough evidence has been accumulated to speculate about physiological causes, and will be based mainly on biochemical knowledge rather than statistical methods. The discovery phase of research in precognition might lead to modified theories, which could then be solidified in the justification phase. This distinction illustrates an important point about statistical methods, which is that they cannot be used to prove any-

thing definitively. There is always an element of uncertainty in results based on statistical methods. These results can suggest causal pathways, but cannot verify them conclusively.

### 3. Why Use Statistics?

There seems to be a misconception among some scientists about the role of statistical methods in science, and specifically about the situations for which statistical methods are most useful. That misconception has sometimes been used in an attempt to negate the evidence for anomalous phenomena. For example, Hyman, in his review of the U.S. government's remote viewing program, wrote:

> Only parapsychology claims to be a science on the basis of phenomena whose presence can be detected only by rejecting a null hypothesis (Hyman, 1996, p. 38).

It is the role of statistics to identify and quantify important effects and relationships before any explanation has been found, and one of the most common means for doing so is to use empirical data to reject a "null hypothesis" that there is no relationship or no effect. There are countless scientific advances that would not have been possible without the use of such statistical methods. Typically, these advances are made in the discovery phase when anecdotal evidence or scientific theory suggests that a relationship or effect might exist, and studies are designed to test the extent to which that can be verified statistically. Only after such studies have indicated that there is almost certainly a relationship do scientists begin to search for a cause or justification. For instance, the link between smoking and lung cancer was first explored when an astute physician noticed that his lung cancer patients tended to be smokers. Numerous studies were then done to explore the link between smoking behavior and subsequent lung cancer, and a statistical relationship was established long before a causal mechanism was determined (*e.g.* Doll and Hill, 1950; Moore and McCabe, 1999, p. 211).

Statistical methods are only useful in situations for which exact replication is not possible. Unlike some experimental domains in the physical sciences, studies relying on statistical methods involve natural variability in the system, and thus the results cannot be precisely predicted or repeated from one experiment to the next. Even if there is a physiological explanation for the results, natural variability in humans or other systems create natural variability in outcomes. For instance, a particular drug may lower blood pressure for known reasons, but it will not lower everyone's blood pressure by the same amount, or even have the exact same effect every day on any given individual.

Statistical methods are designed to measure and incorporate natural variability among individuals to determine what relationships or trends hold for

the aggregate or on average. Here are some of the kinds of situations for which statistical methods are or are not useful:

- They are clearly not needed to determine a relationship that holds every time, such as the blinking response to a wisp of air in the eye, or the fact that a book will drop if you release it in mid-air.
- They are not needed once a causal mechanism is understood even if a relationship does not hold every time, such as trying to determine whether or not pollen causes hay fever or sex causes pregnancy.
- They are useful to indicate the *existence* of a relationship or effect that does not occur every time or in every individual and that does not already have a causal explanation. For instance, the use of aspirin to reduce the risk of heart attacks was established statistically over ten years ago, but it is only recently that causal explanations have been explored.
- They are useful to establish the average *magnitude* of effects and relationships that do not occur every time. One simple example is the batting average for a baseball player. Finding the probability of hitting a ball or a home run is akin to finding the probability of a "hit" in a remote viewing experiment. In each case "hits" happen a certain proportion of the time, but no one can predict in advance when they will occur.

In summary, whereas sometimes scientific research starts with a causal theory and proceeds to verify it with data, statistical methods are most useful in situations where the process happens in reverse. There may be speculation about possible relationships stemming from observations or theories, but the focus is on learning from data. Quite commonly, a relationship is established with near certainty based on large amounts of data (such as the relationship between smoking and lung cancer) before a causal mechanism is determined or even explored. The remainder of this paper discusses details of this process.

## 4. What Constitutes Statistical Evidence?

There are a number of statistical methods that are used to infer the existence of relationships and estimate their strength. The two most commonly used methods for single studies are hypothesis testing and confidence intervals. These two inferential methods for single studies have been standard practice for many decades. In recent years there has been a trend towards using statistical methods to examine the accumulated evidence across many studies on the same topic. In the past, reviews of a collection of studies were subjective and qualitative but the recent trend is towards quantitative methods, which collectively are called "meta-analysis." There is some debate about whether or not meta-analysis provides better evidence than one large well-implemented study on a topic (Bailar, 1997) but there is no doubt that meta-analysis can provide a more complete picture than individual small studies, as will be illustrated by example in Section 5.2.

Replication is at the heart of any science relying on experimental evidence

because any single study potentially could have unrecognized flaws that produce spurious results. (Consider, for example, attempted replications of cold fusion.) However, the meaning of replication is different for studies on living systems, requiring inferential statistical methods, than it is for studies that are supposed to have fixed and predictable outcomes. Variability among individuals can mask real differences or relationships that hold for the aggregate, and will result in somewhat different outcomes even when a study is replicated under similar conditions. If the natural variability is small and the relationship or difference is strong then similar results should emerge from each study. But when the variability is large, the relationship is weak or the effect is rare, the variability may mask the relationship in all but very large studies. For instance, because lung cancer rates are low for both smokers and non-smokers, we would not expect to see smokers develop more lung cancer than non-smokers in every small group of both, even though lung cancer rates for smokers are at least nine times what they are for non-smokers (*e.g.* Taubes, 1993). It is only when examining the trend across studies (or conducting one very large study) that the higher lung cancer rates in smokers would become obvious.

Before considering some simple methods used to examine evidence through combining studies, a brief overview of standard inferential statistics is provided. It is important to understand these methods in order to understand the extensions of them used in meta-analysis.

### *4.1. Hypothesis Testing*

For many decades hypothesis testing was the core of statistical methodology. If the results of the hypothesis test used in a study were "statistically significant" the study was determined to be a success. Unfortunately, if the results were not statistically significant the study was often deemed a failure, and the effect under consideration was thought not to exist. Before explaining why that reasoning is flawed, a brief review of hypothesis testing is required. The procedure follows four basic steps:

1. Establish two competing hypotheses about one or more factors:

   *Null Hypothesis:* There is no relationship, no difference, no effect, nothing of interest, only chance results.

   *Alternative Hypothesis:* There is a relationship, difference or effect of interest.

   It is obviously the goal of most research studies to conclude that the alternative hypothesis is true, since the null hypothesis represents the status quo that would be accepted without any new knowledge or data.

2. Collect data from a sample of individuals, representative of the larger population about which the hypotheses have been proposed.

3. Use the data to calculate a "test statistic" and resulting "*p*-value." The *p*-value is one of the most misunderstood concepts in statistics. It is calculated by *assuming* that the null hypothesis is true, then finding the probability of observing data as contrary to it as that which has just been observed, or more so. It does *not* represent the probability that the null hypothesis is actually true given the observed data, something it is commonly misinterpreted to mean. It only represents how unlikely the observed data or something more contrary to the null hypothesis would be if the null hypothesis were actually true.

4. Use the *p*-value to make a conclusion. Though perhaps not wise in all cases, it is standard practice to conclude that the null hypothesis is false ("reject the null hypothesis") and the alternative is true ("accept the alternative hypothesis") if the *p*-value is 0.05 or less. The notation for this cutoff value in general is $\alpha$ so the standard practice is to reject the null hypothesis if the *p*-value is less than $\alpha$. Otherwise, the proper conclusion is simply that the null hypothesis cannot be rejected based on the evidence presented.

It is rarely prudent to actually "accept" the null hypothesis. The reasoning behind this imbalance between the two hypotheses is based on the types of erroneous conclusions that could be made. The selected value $\alpha$ by definition, provides control over the probability of erroneously rejecting the null hypothesis when it is true (called a type 1 error) and thus declaring an effect exists when it does not. As mentioned, this probability is usually set at 0.05. There is no similar control over the probability of making the mistake of accepting the null hypothesis when it is false, and thus declaring that there is no effect when in fact there is. This mistake is called a type 2 error. It is denoted by $\beta$, which is an unknowable value because it depends on the unknown magnitude of the real effect. (If we knew that magnitude we would not be testing it.)

The probability that the null hypothesis is rejected and thus the alternative hypothesis is accepted is called the *power* of a test. When the null hypothesis is true, the power of the test is the probability of a type 1 error ($\alpha$), and is designated by the researcher (usually at 0.05). The more interesting situation is when the alternative hypothesis is true, in which case the power is the probability of correctly detecting that fact and concluding that there is an effect or relationship. Numerically in this case the power is $1-\beta$ because $\beta$ represents the probability of *not* detecting a real effect when it exists. There is a trade-off between making the two types of errors that is reflected by the choice of $\alpha$. Larger values of $\alpha$ make it easier to reject the null hypothesis, thus increasing the power to detect a real relationship. But the trade-off is that larger $\alpha$ values mean it is easier to conclude a relationship exists in situations where it does not.

The power of a test is also closely tied to the sample size. A very large sample provides a test powerful enough to reject the null hypothesis even when the

effect is very small. In contrast, a small sample has very little power to reject the null hypothesis even if the effect is moderately large. Thus, the way to enhance the power of a test without having to increase the probability of a type 1 error is to increase the amount of data collected.

For example, a very large study conducted by the Steering Committee of the Physicians' Health Study Research Group (1988) compared heart attack rates for 22,071 male physicians who were randomly assigned to take either aspirin or a placebo tablet every other day for five years. The null hypothesis was that taking aspirin has no more effect on heart attack rates than taking a placebo. The alternative hypothesis was that taking aspirin does have an impact on heart attack rates.

There were 17.13 heart attacks per 1000 men in the group taking the placebo, but only 9.42 heart attacks per 1000 men in the group taking aspirin. The *p*-value for this study was extremely small, indicating that if there really is no impact from taking aspirin, a difference as large as the one observed in this study (or larger) would be extremely unlikely to occur in a study with 22,071 participants. In fact the study was terminated earlier than planned because the results were so striking. (Note that stopping a study early is statistically not justified because it could be stopped at a time favorable to the alternative hypothesis. This "optional stopping" was one basis for criticism of early parapsychological experiments. However, ethical issues sometimes outweigh statistical issues in medical trials.)

The estimated "odds ratio" from the study, giving the odds of having a heart attack when taking a placebo compared with when taking aspirin, is 17.13/9.42 = 1.8. (Technically, this is the "relative risk" and the odds ratio is slightly more complicated. However, they are numerically very similar for events with low probability like that of having a heart attack.) At 1.8, the odds of having a heart attack were almost double when taking a placebo compared with when taking aspirin. This is obviously a striking effect with important medical applications.

However, suppose that only one-tenth as many physicians, or about 2200, had participated in the study but that the heart attack rates per 1000 had still been about 17 for the placebo group and 9 for the aspirin group. Although the odds ratio from the data would still indicate that almost twice as many men had heart attacks when taking a placebo compared with when taking aspirin, the evidence would not have been convincing at all using traditional reasoning. The *p*-value would have been about 0.09, which is generally not considered small enough to reject the null hypothesis. Notice that it would not be reasonable to conclude that there is *no* impact from taking aspirin. The justifiable conclusion would be that there is not enough evidence in this study to conclude whether or not aspirin has an impact on heart attack rates, even though the numbers are suggestive. The *p*-value demonstrates that chance alone would result in differences this large or larger about 9% of the time based on samples of 2200.

This example illustrates one of the problems associated with methods commonly used to do qualitative reviews of groups of studies. Often, those reviews simply count how many studies achieved statistically significant results (rejected the null hypothesis), and discount any study that did not do so. Or, worse, often such reviews count these "unsuccessful" studies as evidence that no relationship or effect exists. This technique, called "vote counting" will lead to erroneous conclusions particularly when a series of small studies have been conducted, each of which had very low power to detect a real effect. At the very least, such reviews should consider the role of power when interpreting the results.

### *4.2. Confidence Intervals*

One of the disadvantages of hypothesis testing is that the results do not provide a measure of the magnitude of the relationship or effect. For instance, in the example of the impact of aspirin on heart attack rates, the results of the hypothesis test would only provide information that the $p$-value is almost zero. The heart attack rates for the two groups and the odds ratio would not be provided as part of the standard information accompanying the test.

The other major method of statistical inference is to construct a "confidence interval." The interval is computed from sample data that are supposed to be representative of a larger population. There is a numerical measure in the population about which information is desired, such as the odds ratio for heart attack rates if men were to take a placebo *vs.* if they were to take aspirin. The measure has been computed for the sample, but the sample value will obviously not equal the desired population value exactly. The confidence interval provides a range of numbers that almost certainly does cover the unknown population value. As would be expected, the larger the number of participants in the sample, the shorter the resulting interval will be. A "confidence level," typically 95% or 99% accompanies the interval, indicating how likely it is that it actually covers the unknown population value.

For example, in the study measuring heart attack rates, there was an observed odds ratio of about 1.8, meaning that the rate of heart attacks for those taking a placebo was about 1.8 times what it was for those taking aspirin. A 95% confidence interval for the odds ratio ranges from 1.63 to 2.59, indicating that if all men similar to these were to take a placebo, the heart attack rate could be as little as 1.63 times what it would be if they had taken aspirin instead, or it could be as high as 2.59. The likelihood that the interval covers the true, unknowable odds ratio is about 95%. The interval could also be presented in the other direction, giving a 95% confidence interval for the reduced odds of getting a heart attack after taking aspirin, ranging from 0.39 to 0.61.

Notice that the information provided by the confidence interval is more interesting than the results of the hypothesis test, because the interval provides a numerical assessment of *how much* the risk of heart attack is reduced by taking aspirin. Individuals can then decide if the reduced risk is worth the effort, ex-

pense and possible side effects of taking aspirin. Further, because the range of values is completely above one, the confidence interval provides evidence that there really is a beneficial impact of taking aspirin. In other words, the confidence interval can be used to reject the null hypothesis that there is no difference in heart attack rates after taking aspirin *vs.* a placebo. In general, in situations involving a single population parameter, this simple method can be used for determining whether or not the result is "statistically significant" based on the confidence interval. If the value given in the null hypothesis is covered by the computed 95% confidence interval, then the null hypothesis would not be rejected using the standard criterion of 0.05. If the null value is not in the interval, the null hypothesis is rejected.

As with hypothesis testing, if the sample had been much smaller, the confidence interval results would have been less precise. For instance, if there had been one-tenth as many men in this study, but similar heart attack rates, the 95% confidence interval for the difference would have ranged from a *reverse* difference in which the odds of a heart attack after placebo were only 0.9 times what they were with aspirin, to a beneficial odds of 3.9 times the rate after placebo than after aspirin. The impact of taking aspirin would have been inconclusive, just as it was in this hypothetical situation using hypothesis testing. Even though most of the interval is in the direction indicating that aspirin reduces the heart attack rate, the interval includes values indicating no effect or a negative effect, and those possibilities cannot be ruled out.

In summary, the results of a single study can sometimes be used to conclude that a relationship or effect is "statistically significant" and to estimate its magnitude. However, if the results of one study are not "statistically significant," it does not mean that there is no relationship or effect in the population, especially if the sample size for the study was small. Further, no single study can provide conclusive evidence for a relationship or effect. It is only through replication of the direction and magnitude of an effect across studies that we can determine whether a relationship or effect has been statistically demonstrated.

## 5. Combining Evidence

It has already been noted that studies in which statistical methods are most useful are those involving measurements with natural variability and lack of perfect replication. It is also the case that most studies that use statistical methods involve relationships or effects that are not obvious to the naked eye, and thus that are not strong enough to be confirmed without large amounts of data. Strong relationships that are clearly visible have generally been tested and confirmed by now. For instance, long ago humans determined what substances were poisonous, but only recently have we begun to identify and confirm that eating certain foods can alter our chances of getting various diseases. Since most studies currently conducted involve these smaller effects, the power of these studies to detect real relationships is low unless they amass large

amounts of data. Therefore, in many cases it is only by combining data across studies that enough evidence can be accumulated to make conclusions.

### *5.1. Meta-Analysis*

Meta-analysis is a collection of techniques for quantitatively combining similar studies. Often the subject of controversy (*e.g.* Bailar, 1997; Mann, 1990; Utts, 1999, p. 430), this collection of techniques is powerful when used correctly but can be misleading when used incorrectly.

There are a number of benefits to meta-analysis (*e.g.* Rosenthal, 1991; Utts, 1999, p. 427), but two are most relevant to this paper. First, by combining a large number of small studies all designed to measure the same thing, there begins to be enough data to reach a conclusion that the individual studies were not large enough to justify. Second, by combining studies that use slightly different procedures, treatments, participants, and so on, conclusions can be made about how results are the same or different when these factors are changed.

Almost all meta-analyses have a common fundamental plan. The first step is to identify an effect or relationship of interest and summarize it in the form of a population parameter. A number of studies designed to measure this effect are then identified, for which the outcome of each study can be quantified as the sample version of the parameter of interest. These sample statistics are then evaluated by comparing them to what would be expected by chance if there were no effect or relationship, and by determining whether or not they remain consistent when other factors change.

For instance, in studies of remote viewing, the population effect of interest is whether or not remote viewing works better than would be expected by chance. This effect is quantified as the probability that a judge would be able to pick the correct target from a set of four (or five) possibilities, based on the response of the remote viewer. Studies in which judges were actually given this task are collected, and the actual proportion of sessions for which the judge picked the correct target is recorded for each study, as the sample statistic of interest. These can then be compared to the chance probability of 1/4 (if there were four choices), and can also be grouped and compared based on whether targets were pictures *vs.* actual locations or events, and so on. Thus, from a number of small inconclusive studies, conclusions can be reached by weighing the contribution of each study appropriately. Sometimes studies are weighted according to how large they were, and sometimes they are weighted based on how well they were conducted.

There are many decisions to be made when conducting a meta-analysis, such as whether to include only peer-reviewed studies, whether to combine data across conditions, and so on (*e.g.* Utts, 1999, Chapter 24). Thus two meta-analyses of the same subject area could easily reach different conclusions. The more similar are the studies being combined, and the more careful one is to identify all appropriate studies, the more reliable are the results of a

meta-analysis. Like most statistical procedures, meta-analyses can be done well or poorly, and the results must be interpreted accordingly.

One of the simplest and most informative tools that can be used in meta-analysis is to construct a graph depicting confidence intervals for the specified parameter for each of the studies. If the studies are similar, the results can be combined into a single confidence interval as well. We illustrate this procedure for a medical example and then for an example from parapsychology.

### *5.2. A Classic Example: Antiplatelets and Vascular Disease*

A good example of the graphical display of confidence intervals is provided by Mann (1990) in a news story about meta-analysis in *Science* magazine. The original meta-analysis appeared in the *British Medical Journal* (Antiplatelet Trialists' Collaboration, 1988). The analysis combines the results of 25 clinical studies to determine whether or not there is a relationship between vascular disease (such as heart attacks and strokes) and antiplatelet agents such as aspirin for people who had already had an incident. The relationship was measured using the odds ratio for having a heart attack or stroke given that one is taking an antiplatelet agent compared with taking a placebo. If the antiplatelet drugs have no effect, the odds ratio should be about 1.0. An odds ratio below 1.0 indicates that the drugs have a beneficial effect, while an odds ratio above 1.0 indicates a detrimental effect.

The results of the analysis are displayed in Figure 1. Each study is presented on its own line, with the horizontal line depicting a 95% confidence interval for the odds ratio based on that study. The vertical line at 1.0 represents the chance odds ratio of 1.0. The vertical line to the left of it represents the best single estimate for the odds ratio, which is 0.75, based on the combined data from all studies.

The studies are divided into three types depending on whether they were treating stroke patients (cerebrovascular), heart attack patients (myocardial infarction), or patients with angina pain. After each set of studies a 95% confidence interval is presented for the combined results of that type. Finally, at the bottom of the graph, a 95% confidence interval is shown for the combined results of all 25 studies. If the results of the three types of studies had been very different, it would not have been advisable to combine them.

Notice that very few of the individual studies were "statistically significant" as illustrated by the fact that their confidence intervals cover the chance odds ratio of 1.0. A naive "vote-count" of the number of studies for which the null hypothesis (that antiplatelets have no effect) could be rejected, would appear as if there was little or no use for these drugs in preventing vascular disease reoccurrence. In contrast, the graphical analysis, originally presented by the Antiplatelet Trialists' Collaboration (1988), makes it strikingly obvious that these drugs do work to reduce the odds of a second attack. For the combined data the results indicate that the odds of a second occurrence are reduced by about 25%

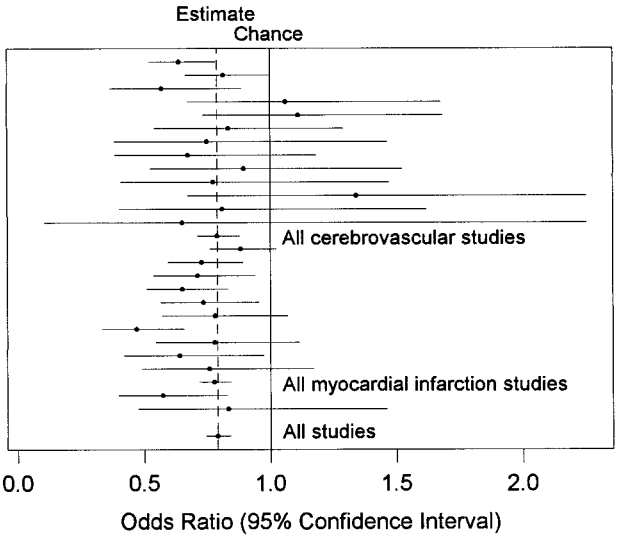


Fig. 1. Odds ratios and 95% confidence intervals for 25 studies comparing aspirin and placebo, from Antiplatelet Trialists Collaboration (1988). Lines (in order from top) represent 13 cerebrovascular studies, then combined data, 10 myocardial infarction studies, then combined data, 2 additional studies, then combined data from all studies. Solid vertical line represents chance (odds ratio of 1.0), dashed vertical line represents odds ratio of 0.75, the best estimate from the combined data.

as a result of taking antiplatelets. In fact the conclusion from the abstract of the original article was clear:

> Thus antiplatelet treatment can reduce the incidence of serious vascular events by about a quarter among a wide range of patients at particular risk of occlusive vascular disease (Antiplatelet Trialists' Collaboration, 1988, p. 320).

### *5.3. Ganzfeld Studies*

Meta-analysis has been used in a number of areas in parapsychology, but was first applied to ganzfeld studies, and an overview and history of this work serves as a good example of the use of meta-analysis in parapsychology. Ganzfeld studies were introduced to parapsychology in the early 1970s and were the subject of a debate in the early 1980s between parapsychologist Charles Honorton and skeptic Ray Hyman (Honorton, 1985; Hyman, 1985; Hyman and Honorton, 1986). The debate focused on meta-analysis of the ganzfeld studies that had been done up until that time. Much of the disagreement between Honorton and Hyman followed from their impressions of

whether the studies contained too many flaws to allow conclusions to be drawn.

In a paper written shortly after the Hyman-Honorton exchange, Akers (1985) criticized the use of meta-analysis as a retrospective method in parapsychology. He noted that as long as there is disagreement about the quality of individual studies, there will be disagreement about meta-analyses of them. He suggested that agreement for protocols be reached *before* conducting future experiments:

> It is time for parapsychologists and their critics to shift attention away from past research and to focus on designs for future research. If experimental standards can be agreed upon before this research begins, much of the methodological controversy can be avoided (p. 624).

As a result of the debate Hyman and Honorton did just that. They agreed to a set of experimental conditions that would alleviate flaws identified in the original collection of studies. Honorton constructed a laboratory that met those conditions, called the Psychophysical Research Laboratories (PRL) and a new series of studies were conducted.

Detailed descriptions of ganzfeld studies are given elsewhere (*e.g.* Bem and Honorton, 1994; Honorton *et al.*, 1990), and only a short description leading to the relevant statistics is given here. A typical study proceeds by asking a receiver to describe a target, usually being simultaneously viewed by a sender isolated from the receiver. The targets in the studies considered here were either photographs or short video segments. For each session the target is randomly selected from a larger set. Three "decoys," that could equally well have been randomly selected to be the target, are bundled with the target for judging purposes. The receiver completes the description and then is shown the four choices (actual target and three decoys), and is asked to identify the one most likely to have been the target based on the receiver's statements during the session. Of course the receivers and everyone in contact with them are blind as to the actual target at this stage.

By chance the receiver has a one in four chance of getting the correct answer, called a "direct hit." This probability is based on the random selection of the target, so it does not depend on what the receiver said. The population parameter of interest is thus the probability that the target selected during judging will be the correct target. The null hypothesis being tested is that this probability is indeed one in four. Presumably, if some sort of psychic functioning is operating, that probability is greater than the chance value of 1/4. The sample statistic used to test this hypothesis is the percentage of direct hits over a series of sessions in a study. The number of sessions in these studies tends to be relatively small, so that it is common, as with the studies of antiplatelets and vascular disease, to find wide confidence intervals and inconclusive results.

The analysis in this paper will be restricted to a subset of the original set of studies included in the 1980s debate, and then to the studies conducted at PRL using the improved consensus protocols. Table 1 lists the original ganzfeld studies considered in the Hyman-Honorton debate, with a few caveats. First, studies with fewer than 20 sessions were combined (Studies 7, 16, 17 and 19), because the statistical method used to construct the confidence intervals is not appropriate with fewer than 20 sessions. Second, studies by one experimenter were removed because they have been criticized as possibly being fraudulent (Blackmore, 1987). Inclusion of those studies would increase rather than decrease the strength of the results, so their removal makes it more difficult rather than easier to reject the null hypothesis. Finally, only studies that provided the number of direct hits based on four choices were included. (Some studies used a different method of collecting data.)

Table 1 includes the study number as assigned by Honorton (1985, Table A1, p. 84), the number of sessions, number of direct hits, proportion of direct hits, and a 95% confidence interval for the estimated probability of a direct hit for the larger population the sessions in the study represent. Figure 2 is a graphical display of the confidence intervals, similar to the display in Figure 1. The vertical line at 0.25 represents chance, and the vertical line at 0.38 represents the combined estimate of the probability of a direct hit based on the results from all of the studies.

Notice that five of the individual studies and the combined data from the small studies have confidence intervals entirely above chance, since the lower ends are still above 0.25. In other words, five of the studies were statistically significant and the remaining seven were not. However, the data from the combined studies clearly indicate that the probability of a direct hit is different from the chance value of 0.25. In fact, a 95% confidence interval extends from 0.34 to 0.43.

TABLE 1
Original Ganzfeld Studies

| Study # | # of Sessions | # of Direct Hits | Proportion Hits | 95% Confidence Interval |
|---|---|---|---|---|
| 1 | 32 | 14 | 0.44 | 0.27 to 0.61 |
| 8 | 30 | 13 | 0.43 | 0.26 to 0.61 |
| 11 | 30 | 7 | 0.23 | 0.08 to 0.38 |
| 12 | 20 | 2 | 0.10 | 0.00 to 0.23 |
| 18 | 28 | 8 | 0.29 | 0.12 to 0.45 |
| 21 | 20 | 7 | 0.35 | 0.14 to 0.56 |
| 31 | 20 | 12 | 0.60 | 0.39 to 0.81 |
| 33 | 100 | 41 | 0.41 | 0.31 to 0.51 |
| 34 | 40 | 13 | 0.33 | 0.18 to 0.47 |
| 38 | 27 | 11 | 0.41 | 0.22 to 0.59 |
| 39 | 60 | 27 | 0.45 | 0.32 to 0.58 |
| 41 | 48 | 10 | 0.21 | 0.09 to 0.32 |
| 7,16,17,19 | 37 | 23 | 0.62 | 0.47 to 0.78 |
| ALL | 492 | 188 | 0.38 | 0.34 to 0.43 |

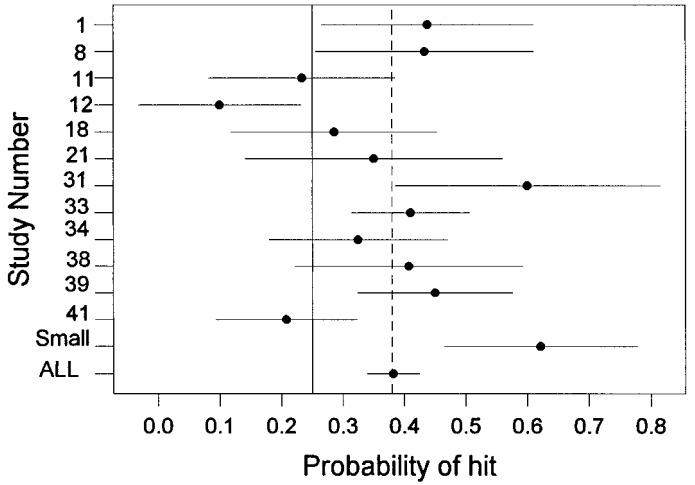


Fig. 2. Proportion of hits and 95% confidence intervals for ganzfeld studies listed in Table 1, and for all studies combined. Solid vertical line represents chance (hit rate of 0.25), dashed vertical line represents hit rate of 0.38, the best estimate from the combined data.

As mentioned, in the debate between Hyman and Honorton a number of potential flaws were identified that could have inflated the success rates in these studies for reasons other than the intended explanation that psychic functioning is possible. Further studies conducted by Honorton and others at PRL were supposed to have eliminated those flaws. Therefore, the results of those studies probably represent a more realistic assessment of the probability of achieving a hit in this ganzfeld procedure.

Table 2 lists the studies conducted at PRL with the same information listed in Table 1 for the earlier studies. The study identification numbers in the first column are the ones originally assigned at PRL (Bem and Honorton, 1994). Notice that all but two of the confidence intervals cover the chance value of 0.25, and thus except for those two studies, the results of the individual studies were not statistically significant. However, when the sessions are all combined the resulting confidence interval is well above chance, ranging from 0.29 to 0.39.

There have been other ganzfeld studies since the ones conducted at PRL, and it is not the intention here to do a thorough meta-analysis of all ganzfeld studies. Milton and Wiseman (1997) have recently attempted to conduct a meta-analysis including more ganzfeld experiments, but their analysis suffers from problems that have not been sufficiently resolved. For instance, because many parapsychologists concluded that the earlier data had already been sufficient for "proof-oriented" research, more recent studies have deviated from the original conditions to look for possible correlates that might impact results.

TABLE 2
PRL Ganzfeld Studies

| Study # | # of Sessions | # of Direct Hits | Proportion Hits | 95% Confidence Interval |
|---|---|---|---|---|
| 1 | 22 | 8 | 0.36 | 0.16 to 0.56 |
| 2 | 9 | 3 | 0.33 | 0.03 to 0.63 |
| 3 | 35 | 10 | 0.29 | 0.14 to 0.44 |
| 101 | 50 | 12 | 0.24 | 0.13 to 0.35 |
| 102 | 50 | 18 | 0.36 | 0.23 to 0.49 |
| 103 | 50 | 15 | 0.30 | 0.17 to 0.43 |
| 104 | 50 | 18 | 0.36 | 0.23 to 0.49 |
| 105 | 6 | 4 | 0.67 | 0.29 to 1.00 |
| 201 | 7 | 3 | 0.43 | 0.06 to 0.80 |
| 301 | 50 | 15 | 0.30 | 0.17 to 0.43 |
| 302 | 25 | 16 | 0.64 | 0.45 to 0.83 |
| ALL | 355 | 122 | 0.34 | 0.29 to 0.39 |

Some of these studies were intentionally designed to include favorable and unfavorable conditions. Others used new target types (such as musical compositions) or new protocols (such as randomly determining whether or not to have a sender) to see if these changes would influence the results. Therefore, not surprisingly, Milton and Wiseman found much lower overall combined hit rates than were found in the earlier studies.

There is a statistical paradigm that allows the preconceived biases of different assessors of information to be taken into account, called Bayesian statistical methods. Matthews (1999) has recently argued that this is a more appropriate way to assess results in parapsychology. Utts, Johnson and Suess (1999) examine the ganzfeld studies from this perspective, and find that the results are still strongly inconsistent with chance.

### *5.4. Is There a "Repeatability Problem" in Parapsychology?*

Throughout the literature in parapsychology there is concern expressed about the "repeatability" of psi effects. For example, an entire conference was held in 1983 entitled "The Repeatability Problem in Parapsychology" (Shapin and Coly, 1985). But much of the concern is unwarranted when the problem is considered from a correct statistical perspective. "Repeatability" has often been misinterpreted to mean that a statistically significant effect should be able to be produced in every experiment, and that psi has not occurred in an experiment unless the results are statistically significant. This interpretation ignores the issue of power. For example, suppose the true probability of a hit in each ganzfeld session is about 0.33, when 0.25 is expected by chance, and the usual criterion of 0.05 is used to define statistical significance. Then in an experiment with ten trials the power will be only 0.073, with 50 trials the power will be 0.27, and even with 100 trials the power will only be 0.54 (Utts, 1988, 1989). In other words, ganzfeld experiments of the size typically conducted

should not be "successful" very often, even if the true effect is that the target can be correctly guessed about a third of the time.

A more appropriate definition of repeatability of an effect is that the estimated magnitude of the effect (odds ratio, hit rate, and so on) falls within the same range from one repetition of an experiment to the next. This definition leads to a different kind of problem that has been discussed in parapsychology, called the "decline effect" (*e.g.* Bierman, 1994). It is based on the observation that the magnitude of psi effects sometimes declines over time, possibly even reaching chance levels. One possible statistical explanation for this effect is "regression toward the mean," a common occurrence in statistical studies. If numerous research studies are done, then by chance some of them will have results that are much stronger than the true effect in the population. Those studies will be published and receive attention. When additional replications are done, chance dictates that the results will be unlikely to again reach those unusual extremes. Thus, the results that are initially higher than their population equivalent will both receive initial attention and fail to replicate in subsequent attempts. Other possible explanations include the fact that replications may be conducted by different researchers and they may introduce subtle or overt changes that dampen the effect. The decline is more commonly observed across researchers than within a single lab.

It may be that there is some other, more interesting explanation for the observed decline effect. In fact, it is not just in parapsychology that the decline effect has appeared. For instance, LeLorier *et al.* (1997) compared 40 different medical treatments for which both a meta-analysis and a large, controlled randomized experiment were available. They found that the agreement between the results of the meta-analysis and the subsequent clinical trial was only moderate, although the results were usually at least in the same direction. For certain medical treatments, the effect in the large clinical trial represented a decline from what would be expected based on the meta-analysis. Further investigation of this phenomenon may provide interesting clues for mind-matter research.

## 6. Making Conclusions Based on Data

The use of statistical methods cannot provide conclusive proof of the physical reasons causing a relationship or an effect. For example, the antiplatelet and vascular disease studies were based on randomized controlled trials, which means that all factors other than the blind, random assignment of the antiplatelet or placebo should have been similar across groups. Because of this, we can reasonably infer that the reduction in the occurrence of vascular disease was actually *caused* by taking the antiplatelet. But that still tells us nothing about the physical process in the human body causing that relationship. The answer to that question is not going to be found in statistics.

Similarly, meta-analyses of ganzfeld and other parapsychological experiments can lead to convincing evidence that the null hypothesis is not true, but

cannot provide answers about the mechanism causing the non-chance results. That leads to an unresolvable debate about whether or not psi exists because the conclusion that psychic functioning is at work rests on ruling out all other possible explanations for the non-chance results. Skeptics are often convinced that an alternative explanation is more probable, even if they cannot identify what it is.

There are, however, some steps that can be taken to strengthen the evidence for one explanation over another. One technique is to look at studies that are similar in concept but have procedural differences, and see if the results are similar. Another is to look for consistent magnitude of effects when other factors are changed. For example, in establishing the causal relationship between smoking and lung cancer, researchers noted that a few decades after men started smoking in large numbers during World War I, they started having higher lung cancer rates. But women started smoking in large numbers during World War II, and indeed it was a few decades after that time that women started having higher lung cancer rates.

In 1995, at the request of the U.S. Congress, skeptic Ray Hyman and I were asked to examine the data produced by U.S. government experiments in remote viewing. One question of interest was whether or not the data supported the conclusion that psychic functioning is possible. Because the data were produced in only two laboratories working under the direction of the same researchers, it was prudent to include additional data from other laboratories. In doing so, I concluded that:

> Using the standards applied to any other area of science [that uses statistics], it is concluded that psychic functioning has been well established (Utts, 1996).

This conclusion was based on comparing and finding consistent results between the government studies in remote viewing and a number of ganzfeld studies. The ganzfeld procedure and remote viewing are very different in methodology, yet very similar in that they both purport to measure whether or not information can be attained through some mechanism other than the normal five senses. Thus, if a methodological flaw in one procedure were responsible for the results, it would be unlikely that similar results would be produced by the other procedure. For instance, in the ganzfeld procedure there is often one-way voice communication leading from the receiver to the sender, and thus one might argue that somehow information is transmitted in the opposite direction to the receiver. But in remote viewing experiments the receiver is completely isolated in a location distant from anyone who knows the correct target, so if similar results are found under the two procedures, the mundane explanation loses credibility.

The same procedure used to compare and combine the different types of antiplatelet studies can be used to compare and combine the remote viewing and ganzfeld studies. It was through this procedure that I reached my conclusion.

Table 3 presents the results I considered in the report. Again, these do not constitute a full meta-analysis of available studies, but rather, the studies that were readily available in the summer of 1995 when I was preparing my report. In addition to the PRL ganzfeld studies, results were provided from laboratories at the University of Amsterdam (Bierman, 1995), the University of Edinburgh (Morris *et al*., 1995) and the Institute for Parapsychology in North Carolina (Broughton and Alexander, 1995).

The remote viewing studies were separated into those conducted at SRI International, where the government program resided until 1990, and those conducted at Science Applications International Corporation (SAIC) where it resided from 1990 to 1994. The results of the remote viewing experiments were based on a sum-of-ranks statistic rather than on the number of direct hits from a four-choice target set. This method is more powerful than the direct hit method under some alternative hypotheses and less powerful under others, and there is no way to do a direct conversion from one statistic to the other. (See Hansen and Utts, 1987, for a discussion and comparison of the use of the two methods.) Therefore, the direct hit probability equivalents for the remote viewing experiments shown in Table 3 were computed by calculating the direct hit rate that would have resulted in the same $p$-value as that reported for the sum-of-ranks results.

The important feature of the data presented in Table 3 is the consistency of the probability of a direct hit or equivalent. Across experimental regimes and laboratories, the probability of a hit remains consistently in the range of about one in three, when one in four would be expected by chance. It is this consistency that lends credence to the possibility that the results are actually the result of psychic functioning rather than some hidden experimental flaw.

## 7. Acceptance of Statistical Results

The intent in this paper was not to provide a conclusive overview of the evidence for psychic functioning, a task for which a much more comprehensive treatment is needed and has been provided elsewhere (see *e.g.* Utts, 1991, or Radin, 1997.) Rather, the intent was to show the extent to which statistical methods and meta-analysis can be used to establish relationships and effects as

TABLE 3
Ganzfeld and Remote Viewing Results

| Laboratory | Type | # of Sessions | Direct Hit Rate or Equivalent | 95% Confidence Interval |
|---|---|---|---|---|
| SRI | Remote viewing | 966 | 0.34 | 0.31 to 0.37 |
| SAIC | Remote viewing | 455 | 0.35 | 0.31 to 0.39 |
| PRL | Ganzfeld | 355 | 0.34 | 0.29 to 0.39 |
| Amsterdam | Ganzfeld | 124 | 0.37 | 0.29 to 0.45 |
| Edinburgh | Ganzfeld | 97 | 0.33 | 0.24 to 0.42 |
| Rhine | Ganzfeld | 100 | 0.33 | 0.24 to 0.42 |
| ALL | | 2097 | 0.34 | 0.32 to 0.36 |

a precursor to finding causal explanations. Utilizing an example from the medical literature and an example from parapsychology illustrates the parallels across subject areas. In general, the steps used to establish relationships and effects can be summarized as follows:

- A strong non-chance effect or relationship exists in the accumulated data, based on hypothesis testing, confidence intervals, or some other method.
- The magnitude of the effect is similar across laboratories, or to the extent that it differs, reasonable explanations are found.
- Consistent patterns are found when other factors are changed or manipulated, such as the observed reduction in lung cancer rates when number of cigarettes smoked is reduced.
- Alternative explanations for the results are ruled out.

Research results in psychic phenomena have been criticized using standards that are much more stringent than those applied to other realms, and in fact, using standards that are sometimes seen as positive aspects in other realms. Here are some criticisms, accompanied by quotes from the original antiplatelet report (Antiplatelet Trialists' Collaboration, 1988, referred to here as ATC). The first three criticisms are taken from Hyman (1996) in his response to Utts (1996), but they are representative of criticisms commonly cited by other skeptics. The final three criticisms are generic ones, often implied rather than stated directly:

1. "Parapsychology is the only field of scientific inquiry that does not have even one exemplar that can be assigned to students with the expectation that they will observe the original results... The phenomena that can be observed with the standard exemplars do not require sensitive statistical rejections of the null hypothesis based on many trials to announce their presence (Hyman, 1996, p. 49)."

   As mentioned earlier, most interesting discoveries today involve small effects that require large amounts of data to verify. So this problem is not unique to parapsychology, it is true of any area in which the effect will only be detected with a large number of individuals. No reasonable student project could conclusively establish the link between antiplatelets and vascular disease. According to ATC, pp. 320–321:

   "Though such risk reductions might be of some practical relevance, however, they are surprisingly easy to miss, even in some of the largest currently available clinical trials. If, for example, such an effect exists, then even if 2000 patients were randomized there would be an even chance of getting a false negative result... that is, of failing to achieve convincing levels of statistical significance."

2. "No other science, so far as I know, would draw conclusions about the existence of phenomena solely on the basis of statistical findings (Hyman, 1996, p. 48)."

   ATC: "Thus antiplatelet treatment can reduce the incidence of serious vascular events by about a quarter among a wide range of patients at particular risk of occlusive vascular disease (p. 320)."

   This conclusion, quoted from the abstract of the *British Medical Journal* report, is based solely on the statistical results of the meta-analysis. In fact it is quite common in the medical literature to find conclusive statements based only on statistical findings.

3. "Where parapsychologists see consistency, I see inconsistency. The ganzfeld studies are premised on the idea that viewers must be in altered state for successful results. The remote viewing studies use viewers in a normal state (Hyman, 1996, p. 57)."

   ATC: "The trials were very heterogeneous, including a range of ages, a range of different diseases, a range of treatments, and so on (p. 322)."

   In fact it is the consistent odds ratio across types of vascular disease and types of antiplatelets that lends credence to the causal nature of the relationship between them. Similarly, it is the consistency of results in ganzfeld and remote viewing studies that lends credence to the idea that information is being gained by means other than the five senses in both domains.

4. The parapsychologists conducting this research have a desired outcome, so they could be subtly influencing the results.

   In fact almost all research is done with a vested interest in the outcome, but good experimental design and rigid methodological controls minimize the impact of that interest. Who conducted and funded the antiplatelet meta-analysis? "The final meeting of collaborators was supported not only by the [British] Medical Research Council and Imperial Cancer Research Fund but also by the Aspirin Foundation, Rhône-Poulenc Santé, Reckitt and Colman, Bayer, Eli Lilly, Beechams, and the United Kingdom Chest, Heart and Stroke Association (ATC, p. 331)." Certainly many of these funders had a vested interest in showing that aspirin and related drugs have a beneficial effect.

5. Data should be automatically (computer) recorded because if there is any potential for the investigators to change it the results cannot be trusted.

   In fact, parapsychological experiments use much more stringent methods and controls than almost any other area, and the researchers are keenly aware that they must continue to do so to avoid allegations of fraud. For the antiplatelet meta-analysis, ATC notes that "The main

results were obtained from the principal investigators in most cases. In some trials the data obtained differed slightly from the data originally published (p. 323)." An admission such as this one would completely negate any study in parapsychology in the eyes of most skeptics.

6. If there is any potential explanation other than psychic functioning, no matter how remote, it should be accepted.

   Very few areas of study could survive such an attitude. For instance, how do we know that the participants in the antiplatelet studies did not have their pills analyzed to determine if they were placebos? If they did, then those taking antiplatelets would have the benefit of knowing they were taking an active ingredient, while those taking the placebo would know they were not. Perhaps that is the reason for the observed difference. Of course that potential explanation is absurd, but it is no less so than many of the attempted explanations for results in parapsychology.

In summary, how are the remote viewing and ganzfeld results different from the antiplatelet and vascular disease conclusions?

- The psi experiments produced *stronger* results than the antiplatelet experiments, in terms of the magnitude of the effect. There is a 36% increase in the probability of a hit over chance, from 25% to 34%. There is a 25% reduction in the probability of a vascular problem after taking antiplatelets.
- The antiplatelet studies had *more opportunity for fraud* and experimenter effects than did the psi experiments.
- The antiplatelet studies were at least as likely to be *funded and conducted by those with a vested interest in the outcome* as were the psi experiments.
- In both cases, the experiments were heterogeneous in terms of experimental methods and characteristics of the participants.

All of this leads to one interesting question: Why are millions of heart attack and stroke patients consuming antiplatelets on a regular basis, while the results of the psi experiments are only marginally known and acknowledged by the scientific community? The answer may have many aspects, but surely it does not lie in the statistical methods.

**Acknowledgments**

I would like to thank Dr. Harald Atmanspacher for organizing the conference from which this paper evolved and for discussions and comments that significantly improved this work. I would also like to thank two anonymous referees for their insightful and helpful remarks.

## References


Akers, C. (1985). Can meta-analysis resolve the ESP-controversy? In P. Kurtz (Ed.), *A Skeptic's Handbook of Parapsychology*, New York: Prometheus Books, 611–630.

Antiplatelet Trialists' Collaboration (1988). Secondary prevention of vascular disease by prolonged antiplatelet treatment. *British Medical Journal (Clinical Research Ed.)*, 296, 6618, 320–331.

Bailar, J. C. (1997). The promise and problems of meta-analysis. *The New England Journal of Medicine*, 337, 8, 559–560.

Bem, D. J. and Honorton, C. (1994). Does psi exist — Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin*, 115, 4–18.

Bierman, D. J. (1994). The decline of the ganzfeld: Patterns in elusiveness? *Proceedings of the 37th Annual Parapsychological Association Convention*, 321–322.

Bierman, D. J. (1995). The Amsterdam Ganzfeld Series III & IV: Target clip emotionality, effect sizes and openness. *Proceedings of the 38th Annual Parapsychological Association Convention*, 27–37.

Blackmore, S. (1987). A report of a visit to Carl Sargent's laboratory. *Journal of the Society for Psychical Research*, 54, 186–198.

Broughton, R. and Alexander, C. (1995). Autoganzfeld II: The first 100 sessions. *Proceedings of the 38th Annual Parapsychological Association Convention*, 53–61.

Doll, R. and Hill, A. B. (1950). Smoking and carcinoma of the lung. *British Medical Journal*, Sept. 30, 739–748.

Hansen, G. P. and Utts, J. (1987). Use of both sum of ranks and direct hits in free-response psi experiments. *Journal of Parapsychology*, 51, 321–335.

Hanson, N. R. (1958). *Patterns of Discovery: An Enquiry into the Conceptual Foundations of Science.* Cambridge: Cambridge University Press.

Honorton, C. (1985). 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, Derr, P., Schechter, E. I. and Ferrari, D. C. (1990). Psi communication in the ganzfeld. *Journal of Parapsychology*, 54, 99–139.

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

Hyman, R. (1996). Evaluation of a program on anomalous mental phenomena. *Journal of Scientific Exploration*, 10, 1, 31–58.

Hyman, R. and Honorton, C. (1986). A joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology*, 50, 351–364.

LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J. and Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine*, 337, 8, 536–542.

Mann, C. (1990). Meta-analysis in the breech. *Science*, 249, 476–480.

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

Milton, J. and Wiseman, R. (1997). Ganzfeld at the crossroads: A meta-analysis of the new generation of studies. *Proceedings of the Parapsychological Association 40th Annual Convention*, 267–282.

Moore, D. S. and McCabe, G. P. (1999). *Introduction to the Practice of Statistics*, 3rd edition. New York: W. H. Freeman and Company.

Morris, R. L., Dalton, K., Delanoy, D. and Watt, C. (1995). Comparison of the sender/no sender condition in the ganzfeld. *Proceedings of the 38th Annual Parapsychological Association Convention*, 244–259.

Radin, D. (1997). *The Conscious Universe: The Scientific Truth of Psychic Phenomena*. New York: Harper Collins.

Rauscher, F. H., Shaw, G. L. and Ky, K. N. (1993). Music and spatial task performance. *Nature*, Oct. 14, 365, 611.

Rauscher, F. H., Shaw, G. L., Levine, L. J., Wright, E. L., Dennis, W. R. and Newcomb, R. L. (1997). Music training causes long-term enhancement of preschool children's spatial-temporal reasoning. *Neurological Research*, 19, 2–8.

Rosenthal, R. (1991). *Meta-Analytic Procedures for Social Research*. Revised edition, Newbury Park, CA: Sage Publications.

Shapen, B. and Coly, L. (Eds.). (1985). *The Repeatability Problem in Parapsychology: Proceedings of an International Conference Held in San Antonio, Texas, 1983*. New York: Parapsychology Foundation.

Sicher, F., Targ, E., Moore II, D. and Smith, H. S. (1998). A randomized double-blind study of the effect of distant healing in a population with advanced AIDS. Report of a small scale study. *Western Journal of Medicine,* 169, 6, 356–363.

The Steering Committee of the Physicians' Health Study Research Group (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine*, Jan. 28, 318, 4, 262–264.

Taubes, G. (1993). Claim of higher risk for women smokers attacked. *Science*, Nov. 26, 262, 1375.

Utts, J. M. (1988). Successful replication versus statistical significance. *Journal of Parapsychology*, 52, 4, 305–320.

Utts, J. M. (1989). Letter to the editor. *Journal of Parapsychology*, 53, 2, 176.

Utts, J. M. (1991). Replication and meta-analysis in parapsychology (with discussion). *Statistical Science*, 6, 4, 363–403.

Utts, J. M. (1996). An Assessment of the evidence for psychic functioning. *Journal of Scientific Exploration* , 10, 1, 3–30.

Utts, J. M. (1999). *Seeing Through Statistics*. 2nd edition, Pacific Grove, CA: Brooks-Cole.

Utts, J. M., Johnson, W. O. and Suess, E. (1999). Bayesian inference for ganzfeld studies. Preprint.