John Palmer

ESP in the Ganzfeld

Analysis of a Debate

Abstract: *This paper reviews the debate over the evidence for ESP provided by experiments using the ganzfeld technique, a simple method used to induce a mild altered state of consciousness. The quantitative literature review technique called meta-analysis has played a prominent role in this controversy. The first question addressed by the reviewer is whether the data establish that ESP in the ganzfeld is replicable. Issues discussed include the effect of multiple analyses, the 'file-drawer' problem and statistical errors. The second question asks, if the effect is real, can it be explained by methodological artifacts? Potential flaws discussed include sensory leakage, problems of randomization and participant fraud. The reviewer's first conclusion is that the aggregate database does provide evidence for a genuine psi effect. However, heterogeneity of results across experimenters indicates that the phenomenon is not easily replicable. The second conclusion is that conventional alternative explanations offered for the observed results tend to be conceivable, but even critics sometimes agree that they are implausible.*

One of the more pressing questions in the debate about the reality of psychic phenomena is whether psi effects in the laboratory are repeatable, or replicable. Some critics of psi research have insisted that results be replicable on demand, but this requirement is usually not realistic for research involving psychological processes such as psi. Most commentators are willing to accept statistical replication, which essentially means that replication is successful more frequently than expected by chance. In recent years, a technique called meta-analysis has been widely used in psychology and other sciences to address this question in a relatively objective fashion. Because of the important role that meta-analysis has played in the ESP-ganzfeld debate, a brief tutorial seems appropriate at this point for readers who might not be conversant with the technique.

Correspondence:
John Palmer, Rhine Research Center, 2741 Campus Walk Aveune, Building 500, Durham, NC 27705, USA. *Email: john@rhine.org*

The Basics of Meta-Analysis

Meta-analysts begin with studies gleaned from all reputable published sources they can find that address a common hypothesis or effect. The core procedure is to convert the outcomes to a *p*-value, convert the *p*-value to a *z*-score, and then sum the *z*-scores and divide the sum by the square root of the number of studies. This provides a statistic called a Stouffer *Z*, and the *p*-value associated with the Stouffer *Z* is then consulted to determine if the group of experiments considered as a whole is statistically significant. Meta-analysts also pay great attention to the effect size, a standardized measure of the strength of the relationship. The simplest and most widely applicable way to compute the effect size for a single study is to take the *z* derived from the *p*-value and divide it by the square root of the number of scores. The effect size for the sample as a whole is then simply the mean of the effect sizes for the individual studies. A study is sometimes considered to have been successfully replicated if the effect size of the second study falls within the 95% confidence intervals of the effect size of the first study. Meta-analysts also sometimes consider the dispersion, or heterogeneity, of the individual study outcomes. Significant heterogeneity is sometimes taken as evidence for a lack of replicability. Finally, meta-analysts often try to assess if a sample produces confirmation of a given hypothesis because of methodological artifacts. The most common approach to this issue is to code the individual studies for methodological quality and correlate these quality codes with the effect sizes or *z*-scores of the studies. A significant positive correlation means that methodological artifacts played an important role and inferences that the hypothesis is really true are suspect.

Although meta-analysis is often touted as an 'objective' way to tackle the issues it addresses, and it is indeed an improvement over previous, seat-of-the-pants, methods, there are a number of subjective decisions that meta-analysts routinely must make to implement the procedure. For example, when is a study similar enough, conceptually or methodologically, to what might be called the population norm to be included in the sample? Meta-analysts call this the 'apples and oranges' problem, and there is no recipe that can guarantee consensus when borderline examples are confronted. Another area of possible disagreement is how to derive quality codes. Milton (1996) attempted to develop a set of uniform standards for coding psi studies by polling psi researchers, but there was, as one would expect, a lack of complete agreement about the necessity of eliminating many of the flaws. Some meta-analysts believe that studies should be removed from the sample entirely if they fail to meet some threshold of methodological purity, whereas others maintain that all the studies should be included in the correlations. Finally, and most importantly, what outcome is necessary or sufficient to support the claim of replicability? Is a significant Stouffer *Z* enough? Is a significant Stouffer *Z* unnecessary if the mean effect size from the second sample falls within the 95% confidence intervals of the mean effect size of the first sample? Does significant heterogeneity mean replicability has failed, whatever the

other outcomes? Whether or not one finds the ESP-ganzfeld results to be replicable will depend in part on how one answers these questions.

Meta-analysis requires at least a moderately large group of studies testing the same hypothesis with comparable methodology. Ideally, they should be conceptual if not direct replications. Parapsychologists take the replication issue very seriously, in part because it is demanded by their critics, and so they are very conscientious about replicating both their own studies and related studies of others. This has led to a large number of meta-analyses of various experimental paradigms, including attempts to influence by psychokinesis (PK) mechanical dice throwing, the output of electronic random number generators (RNGs), and biological processes. On the ESP side, there have been meta-analyses of forced-choice precognition studies such as card-guessing, free-response ESP studies not involving the induction of an altered state of consciousness, and the relations between ESP and extraversion and the belief in ESP. For a summary, see Radin (1997). However, the paradigm to which meta-analysis has been applied most extensively involves the application of an altered state induction technique called the ganzfeld to facilitate ESP performance. The debate about this body of research will occupy the remainder of this paper.

The Ganzfeld Procedure

The ganzfeld is a mild form of sensory deprivation or isolation developed originally by psychologists (Bertini, Lewis and Witkin, 1964). The credit for first applying the technique to parapsychology is jointly shared by Charles Honorton, William Braud and Adrian Parker, although it has been most closely identified with Honorton. For Honorton (1978), the technique had the potential to facilitate ESP because it created what he called an 'internal attention state'; in the absence of competing external stimulation, participants are encouraged to focus their attention inward, thereby increasing the likelihood that they will identify subtle mental impressions. In a similar vein it can be argued that the ganzfeld facilitates mental imagery, because the brain is otherwise starved of information. However, Rex Stanford conducted a series of experiments that collectively suggested that ganzfeld success is more probably attributable to its tendency to increase the spontaneity of participants' thought processes than to the creation of an internal attention state (e.g. Stanford *et al.*, 1989).

The ganzfeld is designed to minimize *patterned* sensory stimulation from the environment. In the visual mode, this is accomplished by the low-tech procedure of affixing halves of ping-pong balls over the receiver's eyes and having them look into a red light. This creates a homogenous pinkish-red visual field. In the auditory mode, it is accomplished by having the receiver listen to pink noise through headphones. Pink noise is white noise with the high-frequency components filtered out. This adjustment makes the sound more pleasant, and it resembles the sound of a waterfall. The reason for maintaining unpatterned stimulation, rather than, for example, just having receivers close their eyes, is to maintain a sufficient level of arousal and discourage the receiver from falling asleep. It is also

common in ganzfeld experiments to precede the test period with progressive relaxation suggestions played through headphones. This has the effect of minimizing kinesthetic stimulation, which is also facilitated by having the receiver seated in a comfortable recliner. The relaxation suggestions are customarily followed by hypnotic-like suggestions for success in the ensuing ESP task.

The receiver stays in the ganzfeld environment for from 30 to 45 minutes following cessation of the relaxation suggestions. Receivers are instructed that during this time they should report out loud any mental images or impressions that come to them. These utterances are tape recorded and written down by the experimenter, who is located in another room.

The targets consist of pictorial material, which can be 'static' (e.g. photographs, art prints) or 'dynamic' (e.g. movie clips). It is generally considered desirable that the targets also be emotionally evocative. The target for a given session is randomly selected from a large pool of potential targets. In most ganzfeld experiments a sender in a third room periodically observes the target during the time the receiver is attempting to gain impressions, trying to 'send' the information to the receiver.

Following the reception period, the experimenter, who is blind to the target, reads back to the receiver the notes taken of their utterances, to refresh the receiver's memory. Then the experimenter removes the headphones and ping-pong balls and commences the judging task. The receiver is shown four possible targets, one of which is a duplicate of the real target, and asked to rank them based on their correspondence to his or her impressions. By chance the correct target will be ranked first 25% of the time. This is called a 'direct hit'. If the target is ranked first or second, it is called a 'binary hit', with a chance probability of 50%. Because a ganzfeld session comprises only one trial, success cannot be demonstrated statistically for a single session, although sometimes the correspondence of the mentation to the target is so close in detail that success is obvious. Success is claimed statistically if the mean hit rate or effect size over a group of sessions is significantly greater than chance.

The Ganzfeld Databases

Over one hundred formal ganzfeld experiments have been published over a twenty five year period from 1974 to 1999. I will compartmentalize them into three separate databases, defined in relation to a pivotal set of ten or eleven experiments conducted under the supervision of Charles Honorton at the Psychophysical Research Laboratories in the 1980s. Thus, for this article the databases will be labelled as the 'Pre-PRL', 'PRL' and 'Post-PRL' databases.

The first formal analysis of ganzfeld studies was conducted by critic Ray Hyman. Until Hyman arrived on the scene, criticism of this research focused on attempts to demonstrate that isolated 'crucial' experiments, usually conducted with so-called 'gifted' participants, could conceivably have been caused by trickery by the participant, the experimenter, or both. The main proponent of this approach was C.E.M. Hansel (1989). Hyman (1981) criticized Hansel's approach

as, among other things, unfalsifiable, and advocated a type of criticism more in line with what one finds in mainstream psychology. The ganzfeld debate reflects this new approach, as illustrated by the prominent role played by meta-analysis.

Is the Ganzfeld Replicable?

The meta-analyses address two separate issues, and it is important to keep them separate. The first is the replicability issue and the second is whether, assuming the effect is indeed replicable (in effect, real), can the success be adequately accounted for by methodological artifacts. We will begin with the replicability issue. Is there an effect at all?

The Pre-PRL Database

Hyman (1985) chose the ganzfeld paradigm as representing the best research parapsychologists had to offer and asked Honorton to send him published reports of all the ganzfeld studies published up to that time (1981). The forty two studies Honorton provided constitute the Pre-PRL database. According to Hyman (1985), Honorton had originally claimed that 55% of these studies provided significant positive results. This figure became the jumping-off point for his critique, which sought to demonstrate that the true figure is not significantly greater than the 25% expected by chance.

Multiple Analyses. The most important of Hyman's criticisms concerns multiple analyses. This problem arises when a researcher conducts multiple statistical tests of the same hypothesis and considers the hypothesis confirmed if any one of them is significant. Assuming the tests are all independent of one another, the criterion level of significance, or alpha level, is not the customary 0.05, but 0.05 divided by the number of analyses. Although such analyses are rarely even close to being independent, a modestly significant result still can quickly lose its significance when multiple analyses are taken into account. A good way around the problem is to specify in advance which analysis one considers to be crucial, but that was rarely done in the ganzfeld reports. Multiple analyses could have occurred in a variety of ways. For example, success could be based on direct hits, binary hits, average rank, or some other method. Sometimes the judging was done both by the participants themselves and by outsiders working with transcripts of the participants' mentation reports. Sometimes the mean of the experimental group was compared both to the theoretical mean score and the score obtained by a control group.

Honorton (1985) conceded that multiple analysis was a problem. His solution was to conduct a meta-analysis restricted to the twenty eight studies (from ten laboratories) that reported direct hits, the most common scoring scheme employed in the database. These twenty eight studies yielded a Stouffer Z of 6.60 ($p < 10^{-9}$).[1] Twenty three of the twenty eight studies were in the positive direction, and twelve of the twenty three were statistically significant. It is noteworthy

---

[1] All *p*-values are one-tailed unless noted otherwise.

that the twenty eight direct-hit (DH) studies were also the ones that got the best results. According to Palmer and Broughton (2000), the mean effect size for the twenty eight direct hits studies was 0.263, compared to 0.055 for the remaining eleven studies for which effect sizes could be calculated. The hit rate for the DH studies, based on an adjustment that rendered each study as having a 25% chance of success (four target alternatives) was subsequently calculated to be 35%, with the 95% confidence interval from 28% to 43% (Bem and Honorton, 1994).

The File-Drawer. Another of Hyman's criticisms concerned a point well known to meta-analysts, referred to as the 'file-drawer' problem. The assumption is made that unsuccessful studies tend not to get published, so that the published studies represent a positively biased sample of all those conducted. Hyman speculated that in the present case many of the studies in the file drawer were aborted before completion because the results did not look as though they would be significant. In support of his speculation, Hyman found, contrary to what one would ordinarily expect, that the studies with the smallest sample sizes got the best results. One explanation for such a result is that low-N studies with poor results were back in the file drawer.

The flip side of failing to report low-N studies with poor results is to report low-N studies with good results, studies one would not have reported had the results been poor. Hyman cited two apparent examples of such 'retrospective studies' in the database: one was a composite of sessions conducted for visiting film crews (Honorton, 1976) and the other from a classroom demonstration (Child and Levi, 1979). The latter study, incidentally, yielded significant psi-missing and removing it would have helped support the positive psi hypothesis.

While noting that a survey of ganzfeld researchers by Blackmore (1980) had failed to find much evidence for unpublished ganzfeld studies, as well as the policy of the Parapsychological Association that its affiliated journals not discriminate against publication of negative results, Honorton's most effective rebuttal was to employ a statistical technique developed by meta-analyst Robert Rosenthal (1979) to estimate the number of studies that would need to be in the file-drawer for the overall result of the DH studies to be reduced to non-significance. The number he came up with was 423, or more than 12,000 sessions. Although 12,000 is probably an overestimate because Hyman's aborted studies would have low Ns, the idea that even half that number of sessions were conducted and not reported seems unlikely, particularly since a typical ganzfeld experiment takes about two hours to conduct and the number of parapsychologists is quite small. As for the negative correlation between success and sample size, Honorton noted that it could also be caused by a loss of experimenter enthusiasm as a long study drags on. He cited evidence from two large-N studies that scores declined when the number of sessions per day was increased so the study could be completed within the specified time frame.

Statistical Errors. Although Hyman (1985) included statistical errors among his methodological flaws, these more properly belong under replicability, because they impact whether there is an effect, independently of whether or not that effect is paranormal. Hyman found twelve of the forty two studies to have

statistical errors, the most common of which (four cases) was failing to add probabilities more extreme than the designated one when computing Fisher's exact probability test. Honorton (1985) agreed that six of the suspect twelve studies had statistical errors but made no comment about the other six.

A Meeting of the Minds. Following the publication of the debate described above, Hyman and Honorton (1986) published a 'joint communiqué' summing up their impressions at that time. In this paper, Hyman joins in the following quote:

> Although we probably still differ on the magnitude of the biases contributed by multiple testing, retrospective experiments, and the file-drawer problem, we agree that the overall significance observed in these studies cannot be reasonably explained by these selective factors. Something beyond selective reporting or inflated significance levels seems to be producing the nonchance outcomes. (p. 352)

However, the authors continued to disagree on our second question, whether the results could be attributed to methodological flaws, a topic to be addressed later in this paper.

The PRL Database

The PRL studies were designed by Honorton to remedy the methodological criticisms raised by Hyman (1985; Hyman and Honorton, 1986). This database is by far the most homogenous of the three, as all the experiments were conducted at the same laboratory using the same basic procedure (Bem and Honorton, 1994; Honorton *et al*., 1990). It consists of eleven studies, one of which was removed from some analyses for methodological reasons. The remaining ten comprised 329 trials completed by 240 different receivers. The hit rate was 32%, $z = 2.89$, $p = 0.002$; Stouffer $Z = 2.55$, $p = 0.005$, with a 95% confidence interval ranging from 30% to 35%. Nine of the ten results were in the positive direction, but only one was independently significant.

It is worth mentioning that a significant negative correlation of -0.64, $p < 0.05$, two-tailed, was found between the sample sizes and effect sizes of the ten studies. This mimics the trend found by Hyman (1985) for the Pre-PRL database, but it clearly cannot be attributed to unreported low-N studies, as all trials contributing to the database were reported.

The legitimacy of the statistical significance as such of the PRL database has not been challenged.

The Post-PRL Databases

At the end of their report on the PRL studies (which, incidentally, was published in a prestigious mainstream psychology journal, *Psychological Bulletin*), Bem and Honorton (1994) noted that a final conclusion about the evidentiality of the ganzfeld paradigm would depend upon the success of subsequent investigators in replicating the PRL studies. This call inspired a number of other parapsychologists to attempt their own ganzfeld experiments, conforming to varying degrees with the PRL procedure. Milton and Wiseman (1999) published a meta-analysis

of thirty post-PRL ganzfeld studies that collectively were quite close to chance, with a Stouffer Z of 0.70, $p = 0.24$, and an effect size of 0.013. It thus appeared that the PRL studies had not been successfully replicated.

A Meta-Meta-Analysis. The Milton and Wiseman analysis was criticized by Storm and Ertel (2001), primarily for ignoring the pre-PRL data in their calculations. The authors also came up with a set of eleven ganzfeld studies published between 1982 and 1986 that reported direct hits but had not been considered by any of the previous analysts. (If they had been, they would have been included in the pre-PRL database.) These studies were independently significant with a Stouffer Z of 3.46, but when the studies were weighted for quality of the methodology, the Z dropped dramatically to 1.06.

Storm and Ertel reasoned that the best measure of overall success in the ganzfeld would be a figure based on all the databases combined, provided that they did not differ significantly among themselves. Although the extended pre-PRL ('old') database was found to have produced significantly higher scores than the PRL and post-PRL ('new') databases by an ω^2 method suggested by Hays (1963), they did not differ significantly by another method suggested more recently by Cohen (1988, p.179), and the authors decided this latter criterion was preferable. Combining the twenty eight pre-PRL direct hit studies identified by Honorton, their own eleven pre-PRL studies, the ten PRL studies, and the thirty post-PRL studies, they arrived at a sample of seventy nine studies. The Stouffer Z was 5.66, $p = 7.8 \times 10^{-9}$, with an effect size of 0.138.

Storm and Ertel also pointed out that the possibility of below-chance scores (so-called psi-missing) needs to be taken into account. Although such results are uncommon in ganzfeld experiments, they do occur occasionally. The authors cited statistical evidence that both the pre-PRL and post-PRL databases were significantly heterogeneous, a point I will return to later in this paper.

In response, Milton and Wiseman (2001) argued that it was illegitimate to include the pre-PRL database because of the methodological weaknesses pointed out by Hyman (1985), and if it was to be incorporated, the non-DH studies should have been included as well, as such studies were included in the post-PRL sample. They also complained that Storm and Ertel applied quality codes only to the eleven new pre-PRL studies and the codes themselves were inadequate. They pointed out further that the heterogeneity statistic did not say anything directly about psi-missing. Finally, they complained that z-scores for individual studies were based on the normal approximation to the binomial, even in cases where the sample sizes were too small to justify this procedure.

Updating the Post-PRL Database. Milton (1999) published a new meta-analysis updating the one she published in the *Psychological Bulletin* (Milton and Wiseman, 1999). She included eight new studies, but she excluded a highly significant study by Dalton (1997) because she considered it to be an outlier that unduly weighted the final outcome. This updated post-PRL database of thirty eight studies yielded a marginally significant Stouffer Z of 1.45, $p = 0.074$, which Milton still considered inadequate for claiming that the ganzfeld is replicable.

Yet another updated post-PRL meta-analysis, conducted independently of Milton's, appeared in the literature shortly thereafter (Palmer and Broughton, 2000; Bem, Palmer and Broughton, 2001). This analysis added ten new studies, compared to Milton's eight. One of these was the Dalton (1997) study mentioned above, and another one was a successful study published after Milton's update was completed (Alexander and Broughton, 1999). (There appears to be one other minor difference in the samples but it is not clear what it is, because Milton did not state precisely what studies were included in her analysis.) The Bem, Palmer and Broughton (BPB) meta-analysis yielded a clearly significant Stouffer Z of 2.59, $p = 0.0048$, with an effect size of 0.051 and a hit rate of 30%.

Methodological Standardness. Milton's (1999) updated meta-analysis was followed in the *Journal of Parapsychology* by the publication of an email debate among numerous parapsychologists concerning the status of ESP ganzfeld research (Schmeidler and Edge, 1999). Space limitations prohibit a full review of this debate, but one major point does need to be raised. Several commentators complained that one important reason that the Milton and Wiseman (1999) meta-analysis yielded such poor results is that they included studies that employed non-standard ganzfeld methodology. Recall that this issue also came up in the debate about the pre-PRL database (Honorton, 1985; Hyman, 1985). The present critics cited in particular three studies that yielded below chance results, two in which musical pieces were used as targets (Willin, 1996a, 1996b), and one in which the sender was periodically presented with the target slide sub-liminally, interrupting a PK test (Kanthamani and Palmer, 1993). However, these judgments of standardness were subjective and made with knowledge of the experimental results.

BPB sought to remedy this latter problem experimentally (Palmer and Broughton, 2000; Bem, Palmer and Broughton, 2001). Three of Bem's graduate psychology students were given the method sections only of the forty studies in the updated (by the authors) post-PRL database and asked to rate them for standardness. As a criterion for standardness they were given a section describing 'The Ganzfeld Procedure' from Bem and Honorton's (1994) *Psychological Bulletin* article covering the pre-PRL and PRL databases, and a more detailed description of the method used in the PRL studies (Honorton *et al.*, 1990). As the debate concerned the ganzfeld *procedure*, they were not told to consider how participants were selected. The judges were told to base their ratings on how much they thought the departure from standardness might affect the results, based on the rationale of the ganzfeld. Thus, inconsequential departures from standardness were given little if any weight in the ratings.

The mean standardness rating of the forty studies, averaged over the three judges, on the 1–7 scale (with 7 being maximum standardness) was 5.33. These standardness ratings were positively correlated with ESP effect size, $r_s(38) = 0.31$, $p = 0.024$. The twenty nine 'standard' studies that fell above the midpoint of the scale (4) yielded a Stouffer Z of 3.49, $p = 0.0002$, an effect size of 0.096 and a hit rate of 31%. Nine 'nonstandard' studies falling below the midpoint yielded a negative Stouffer Z of –1.30, *ns*, and a hit rate of only 24%. The two samples

differed significantly ($p = 0.02$).[2] These data clearly show that including the non-standard studies did indeed adversely affect the results of the Milton and Wiseman (1999) meta-analysis.

Only one of the twenty nine standard studies did not report direct hits. Eliminating this study raises the Stouffer Z slightly to 3.78 (Palmer and Broughton, 2000), with the hit rate maintaining itself at 31%. This hit rate falls inside the 95% confidence intervals reported by Bem and Honorton (1994) for both the post-PRL (28%–43%) and the PRL (30%–35%) databases. Note, however, that standardness ratings have never been computed for the pre-PRL database and this database doubtless included a number of studies that would have been rated as non-standard by Bem's judges. As non-standard methodology and failure to report direct hits closely mirrored each other in the post-PRL database, it is likely that the use of non-standard methods in the non-DH studies in the pre-PRL database helps explain why these studies yielded such poor results compared to the DH studies in that database. Thus, it does seem prudent to say that claims of replicability of the ganzfeld should have the caveat 'provided standard methods were employed, as defined by BPB'.

Conclusions

Honorton (1985), Milton and Wiseman (1999), and BPB (Bem, Palmer and Broughton, 2001; Palmer and Broughton, 2000) all used highly conservative methods to obtain the *p*-values for individual studies. The later two used the exact binomial, which actually yields a negative z when the number of hits is exactly at chance, and according to Donald Burdick, Professor of Statistics at Duke University, 'will overestimate ["the probability that random guesses would produce at least as many hits as were obtained . . ."] by a substantial margin' (personal communication, 9 October 2001). For example, even the thirty studies in the Milton and Wiseman meta-analysis produce a significant positive result when a simple *z*-test is applied using individual trials as the unit of analysis.[3] Part of the success of this latter method is that it effectively weights the *z* for each study by the number of trials in the study. All the meta-analyses used unweighted *z*'s. The unweighted method decreased significance for the post-PRL database but increased it for the other databases.

However, the most important factor distinguishing the nonsignificant Milton and Wiseman (1999; Milton, 1999) and the significant BPB (Bem, Palmer and Broughton, 2001; Palmer and Broughton, 2000) meta-analyses of the post-PRL database is the decision to include or not include the highly successful Dalton (1997) study. As I wrote in the email debate (Schmeidler and Edge, 1999), I strongly disapprove of removing outliers from databases, even though it is advocated or at least tolerated by some meta-analysts. My primary reason for this

---

[2] The Willin (1996a, 1996b) studies with musical targets received the lowest average standardness score in the database. However, the Kanthamani and Palmer (1993) study with subliminal sending scored above the midpoint on standardness and thus was included in the standard group.

[3] I thank Dean Radin for bringing this fact to my attention.

view is that removing outliers misrepresents the data, which is the cardinal sin in any data analysis. (In the debate, I made this point in response to someone who was an advocate of the ganzfeld's evidentiality, not to Milton.)

On the other hand, the Dalton data contribute to a statistic that does deserve to be taken seriously, namely heterogeneity. It turns out that all the various pre- and post-PRL databases are significantly heterogeneous (Palmer and Broughton, 2000). For example, the *p*-values for heterogeneity for both the DH studies in the pre-PRL database and the twenty eight standard DH studies in the post-PRL database are 10^{-5}. This implies that unaccounted for factors that vary from study to study influence ganzfeld success. It is clear that we do not have a very good idea what these factors are. Elements of the debate that I have not covered in this paper concern the consistency with which various personality and target-type variables influence ganzfeld success. However, even if some of these variables turn out to be reliable predictors they will not account for enough of the variance to put a major dent in this huge heterogeneity.

In my opinion, the strongest predictor of ESP results generally is the identity of the experimenter (Kennedy and Taddonio, 1976). Whether or not this is demonstrated in the ganzfeld depends on how the effect is analysed. For example, Honorton (1985) noted that six of the ten laboratories in the pre-PRL database reported significant overall positive results, but Rosenthal (1986) found that 'these 10 investigators differed significantly and importantly in the average magnitude of the effects they obtained' (p. 327).

In conclusion, it seems to me that all three databases have provided overall significant evidence of ESP and fit reasonably well within each other's confidence intervals. This is an impressive rate of stability that clearly cannot be attributed to just a handful of investigators, but at the same time, investigator differences do seem to play some role. Furthermore, the successful investigators all come from a fairly narrowly defined population of experimental parapsychologists who may not be very representative of scientific researchers generally. The huge heterogeneity suggests that we still have a lot to learn about the factors that determine ganzfeld success, a theme that has been stressed by the critic Milton (1999). Until we find out what these factors are, it is anyone's guess how widely replicable the ganzfeld will be outside the parapsychological community. It would be immensely valuable for such 'outside' researchers, particularly benevolent neutrals on the subject, to give ganzfeld research a try and help us learn what the crucial variables are.

Is it all Methodological Artifacts?

The collective statistical significance or the replicability of ESP in the ganzfeld mean nothing so far as providing evidence for a paranormal process if the success can be attributed to flaws in the way the studies were conducted. Such methodological artifacts have been a major component of the ganzfeld debate, and it is now time to examine the arguments of the two sides.

The Pre-PRL Database

At the beginning of the paper I mentioned that one aspect of meta-analysis that is vulnerable to subjective bias is the assignment of quality ratings or codes by the meta-analyst. This is particularly true when the analysts are aware of the outcomes of the studies in question and have strong theoretical predispositions. This problem is illustrated dramatically in the debate between Hyman (1985) and Honorton (1985) on the pre-PRL database. Both employed the standard technique described earlier of correlating the quality codes with measures of ESP scoring. As one might expect from the preceding discussion, Hyman found that methodological artifacts did account for the ganzfeld success and Honorton found that they did not.

In addition to the statistical flaws addressed earlier, Hyman (1985) designated five categories of methodological flaws in the pre-PRL database. The first, called *single target*, occurred in the earlier ganzfeld studies when a paper target such as an art print was included in the judging packet after being handled by the sender. It is possible that the receiver could be tipped off to the identity of the target by noting which of the four art prints displayed handling cues, such as fingerprints. I previously conducted a study demonstrating that Dutch college students were able to successfully identify such handling cues when told to look for them (Palmer, 1983). The solution, adopted in later ganzfeld studies, is to use a duplicate judging set. The problem is eliminated in the auto-ganzfeld by presenting the target and judging materials by videotape.

The next two flaws seem for the most part to assume cheating on the part of the sender. A *security* flaw was assigned when an experimenter also served as the sender, rather than having separate experimenters monitor the sender and receiver. A *documentation* flaw was assigned when the report did not make it clear how many senders were friends of the receiver or whether this made a difference in the results. Perhaps, for example, the sender at the time of judging could somehow substitute the picture which best coincided with the receiver's impressions for the actual target. Security and documentation flaws were assigned for a few other isolated problems, many of which were not specified. One that was specified, rolling a clay ball over the target picture, would be better placed in the single-target category, because it involved the mechanism of sensory cues on the target in the judging pack.[4]

The final two flaws involved problems with randomization. This could become a factor if, say, due to inadequate randomization a picture that in general receives high ratings from receivers whether or not it is the target appears more frequently as a target than do other pictures. A *randomization* flaw was assigned if the targets were selected by a suboptimal procedure, such as shuffling a deck of cards, or the method of randomization was not reported. Modern parapsychologists generally use either random number tables (RNTs) or electronic RNGs for target selection. A *feedback* flaw was assigned when the location of the target

---

[4] This happened to be a study I conducted (Palmer and Aued, 1975), and I was careful to roll the clay ball *lightly* over the target picture and checked to be sure it left no visible mark. The clay was hard and never stuck to the picture.

picture within the judging packet was not randomized. Hyman only assigned this flaw if a single judging set was used, but it would appear that the same problem would occur even if a duplicate set was used for judging. It presumably would occur if, say, receivers are biased to choose the first picture they see in the judging packet as the target, and due to poor randomization the target picture actually appears first in the packet more frequently than in other locations.

Hyman submitted his flaws to a factor and a cluster analysis. Each revealed the same three factors or clusters, one of which correlated significantly with the ESP measures. This factor included three flaws as predictors: randomization, feedback and documentation, and each of these three correlated significantly with ESP by themselves. He then computed regression equations to show that one would expect chance results from studies in which the three crucial flaws were eliminated.

Honorton (1985) gave a number of specific examples of what he considered to be erroneous coding by Hyman, especially inconsistencies in applying his criteria. For randomization he claimed misclassification of at least five studies; for example, in one case the author (York, 1977) states: 'individual targets . . . selected using a random number generator from each mini-target pool by an otherwise uninvolved assistant' (pp. 48–9). He asserts that seven of the ten studies to which Hyman assigned the feedback flaw 'describe procedures for ordering targets at judging' (Honorton, 1985, p. 75). It is possible that in these cases of disagreement Hyman recognized that an RNG was used or an ordering procedure was applied, but that the specific ways these procedures were carried out were somehow inadequate. Such, however, was not mentioned in his report.

As stated above, when Honorton assigned his own quality codes he found no significant correlations between Hyman's flaw categories and ESP. Although this difference in results is due partly to differences in coding of particular studies, in the case of randomization it is evident that differences in the coding criteria themselves played a role. Both Honorton and Hyman coded the randomization flaw on a three-point scale. Both gave the highest score to studies using RNTs or RNGs. However, whereas Hyman gave the lowest rating to studies that used suboptimal methods such as card shuffling, and intermediate ratings to studies where the method was not reported, Honorton did the reverse. In my opinion, Hyman's method makes more sense.

Honorton (1985) recruited a specialist on factor analysis named David Saunders to write an appendix discussing the validity of Hyman's factor and regression analyses. His most important point was that Hyman was guilty of not properly correcting for multiple analyses. Ironically, this is the same generic criticism Hyman made of a number of studies included in the pre-PRL database.

The differences in coding between the protagonists resulted in Hyman assigning many more flaws than Honorton. By my count Hyman listed ninety nine separate flaws, ranging from ten for security and feedback to thirty one for randomization. I think it is fair to infer that Hyman considers the sheer volume of alleged flaws to be sufficient to discredit the database, regardless of what the correlations show. Honorton did not list specific flaws in his report, but it is safe to

assume that it was much less than ninety nine. He builds his case on examples which appear to show that either Hyman did not follow his own coding criteria or that proper control was achieved, even though not by the specific method Hyman coded for. The latter is particularly an issue for the security and documentation flaws. For instance, Honorton discusses in some detail an experiment by Braud, Wood and Braud (1975) in which precautions were taken to prevent sender cheating even though the experimenter served as the sender. The scenario of the sender substituting a bogus target for the real one presumes that he has access to the judging packet after learning the receiver's mentation, so that he or she knows which picture to substitute. That was not the case in Braud's experiment, and may not have been the case in other studies Hyman cited for this flaw.

Experimenter Fraud? A major contributor to the pre-PRL database, whose results were consistently positive, was Carl Sargent. A critic, Susan Blackmore (1987), was able to observe some of Sargent's ganzfeld sessions and noted protocol violations in the procedure. Sargent used a stack of envelopes from which the target was to be randomly selected. Although the stack was supposed to contain an equal number of each target, Blackmore uncovered some slight departures from this equality. She cited one specific instance in which the discrepancy could have indicated that Sargent as experimenter knew the identity of the target when he attempted to guide the participant to choose that target as his response during judging. Sargent (1987) and his associates (Harley and Matthews, 1987) gave what I consider to be plausible reasons for the discrepancies, and Blackmore did not reply to their replies. In my opinion, the discrepancies represent random errors rather than systematic bias or fraud. In any case, Honorton (1985) reported that the pre-PRL database remains significant when Sargent's data are removed.

The PRL Database

Randomization. The first round of methodological criticisms of the PRL database occurred in a debate by Hyman (1994) and Bem (1994) immediately following the report of the PRL studies in the *Psychological Bulletin* (Bem and Honorton, 1994). Hyman praised the PRL studies as being a significant improvement methodologically over the pre-PRL studies but still had some complaints. The most important involved randomization. Although Honorton used the RNG-based method that received the highest quality codes in the pre-PRL debate and agreed upon in the 'joint communiqué' (Hyman and Honorton, 1986), Hyman took the argument a step further in the PRL debate. He argued in effect that even an adequate randomization procedure could (and indeed almost certainly would) result in pictures or movie clips in the target pool appearing as targets different numbers of times. If it happened that *by chance* the pictures selected most frequently as targets were the same as those tending to receive relative high ratings from receivers, whether they were targets or not, then a spurious excess of hits could result. Hyman said that the number of times each target appeared should have been equalized in advance, with only their order randomized.

Bem (1994) responded by doing a reanalysis of the data. For each trial he computed an adjusted chance probability influenced by receiver response biases. Thus, if the target for a trial happened to be a popular one with all the receivers in the study, the adjusted chance probability would be greater than the original 25%. The adjusted hit rate was 30.7%, only trivially different from the original 31.5% hit rate.

Hyman's criticism is a potentially serious one because it could apply not just to the PRL studies but to all the other ganzfeld studies and, in fact, to most studies in parapsychology that use a procedure where the frequency of each target is free to vary. Parapsychologists almost never perform the kind of control analyses reported by Bem (1994). Has Hyman unwittingly undermined a huge amount of the psi literature? The answer is no, and the reason is that the bias at issue is not systematic. In other words, it could just as easily lead to a deficiency of hits as to an excess of hits, depending upon whether the most frequent targets contradict or match participant response biases. As Hyman himself notes, this effect cancels itself out over a large number of studies. That is one reason we do replications; successes based on such 'lucky' matches of targets and response biases will not replicate very well.

Hyman applied essentially the same criticism to the order of the targets in the judging pack. Recall that a similar issue defined the feedback flaw in Hyman's (1985) critique of the pre-PRL database. The locations of the targets within each judging pack were fixed in the PRL studies, so these locations depended upon which pictures or movie clips were randomly selected as targets. When Bem (1994) examined the distribution of target locations he did find a slight bias ($p < 0.05$), but it was the opposite of the receiver's biases: the receivers preferred the first and fourth locations, whereas the most frequent location of the actual targets was third.

Finally, Hyman found a complex internal effect in the data that he found suspicious, although he did not indicate how it might have produced a spurious conclusion. He noted that hitting was concentrated on trials in which the target picture or movie clip had been a target in previous sessions *and*, as sometimes happened in these studies, the experimenter (who was blind to the target) assisted the receiver in the judging task. Bem noted that Hyman's conditions were most likely to apply to the later studies in the database, in which other methodological refinements were introduced that were responsible for the success. For example, the highly successful 'Julliard' study (Schlitz and Honorton, 1992) restricted to artistic participants was the eighth of the ten experiments.

Sensory Leakage. Wiseman, Smith and Kornbrot (1996) proposed a method by which sensory leakage could have accounted for the successful PRL results. Although the receivers were located in an acoustically isolated chamber, the senders were not. This arrangement allowed for possible auditory stimuli travelling from the sender to the experimenter, rendering the latter non-blind. If the experimenters were honest and they were aware of such cues, they would have aborted the session. As the authors did not want to postulate experimenter fraud, they assumed that the auditory stimuli were subliminal, which also allowed for

the stimuli to be quite weak in amplitude. Unfortunately, the PRL lab had been torn down by the time Wiseman *et al*. devised their scenario, so the actual degree of sound isolation between the sender and experimenter could not be determined. The authors were thus relegated to making estimates. They concluded that in the worst of cases 'sender-to-experimenter leakage *could* . . . have taken place' (Wiseman, Smith and Kornbrot, 1996, p. 119, italics in original). This would have required the sender to be making loud noises (preferably at times when the receiver said something that matched the target — in the autoganzfeld, the sender could hear the receiver's ongoing mentation report through headphones), but senders had been instructed not to make such sounds.

Wiseman *et al*. bolstered their case by noting that sessions in which the experimenter helped the receiver with the judging yielded higher ESP scores than those without experimenter assistance ($p = 0.026$), although the authors note this also would be predicted by an ESP hypothesis. Although the unassisted trials were independently significant, the authors point out that the experimenter reading back to the receiver the notes of their preceding mentation might allow a basis for a contaminated experimenter to unwittingly bias receivers in their judging.

As Wiseman *et al*. point out, a way to settle this matter is to have the transcripts of the suspect sessions re-judged by an outsider. Bem has conducted just such an analysis and says that the results remained significant, although not strongly so (personal communication to Palmer, 15 June 2001). However, too much weight should not be placed on this finding until the report is subjected to peer review and published.

Honorton himself reported another potential sensory leakage problem that potentially contaminated 80% of the trials (Honorton *et al*., 1990). At this point in the process of data collection, the researchers discovered that when an external amplifier was placed between the VCR and the receiver's headphones and the pink noise turned totally off, the soundtrack accompanying the dynamic targets (which were responsible for all the significance) could faintly be detected. Of course, under the conditions pertaining in the actual study (i.e. pink noise and no amplifier) the sound track could not be detected (supraliminally), although perhaps subliminal detection was possible. The results did not decline over the last 20% of the trials, which contradicts what one would expect if this leakage actually took place.

Finally, a weakness of these subliminal perception critiques is that auditory subliminal perception itself has not been well established scientifically — far less so than visual subliminal perception, on which much more research has been conducted.

The Post-PRL Database

This section is easy to write because there have been as yet no published methodological criticisms of the post-PRL studies. The good news for ganzfeld proponents is that the most successful studies are generally the ones that most closely followed the 'standard' PRL protocol, which has come through the wars relatively unscathed, certainly much better than did the pre-PRL studies.

Conclusions

My impression over many years is that critics of parapsychology are very good at providing *conceivable* normal explanations for psi effects but rather poor at providing *plausible* normal explanations for them. This principal is abundantly illustrated in the ganzfeld debate. Even Hyman (1994) acknowledges that his alternative explanations are unlikely to account for the data. He sees them instead as symptoms, presumably of some unidentified, more serious problems. I find this argument to be a *non sequitur*; it doesn't follow that overlooking relatively trivial problems implies overlooking more serious ones. Of course, psi is implausible to many people as well, so in the final analysis readers must decide for themselves whether a unified psi process is more implausible than a plethora of arcane alternatives. Someday psi effects may become so strong and so repeatable that the opposition will simply be overwhelmed, but that day has yet to arrive.

References

Alexander, C.H. and Broughton, R.S. (1999, 'CL1-Ganzfeld study: A look at brain hemisphere differences and scoring in the autoganzfeld', *Proceedings of Presented Papers*: The Parapsychological Association 42nd Annual Convention, pp. 3–18.

Bem, D.J. (1994), 'Response to Hyman', *Psychological Bulletin*, **115**, pp. 25–7.

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

Bem, D.J., Palmer, J. and Broughton, R.S. (2001), 'Updating the ganzfeld database: a victim of its own success?', *Journal of Parapsychology*, **65**, pp. 207–18.

Bertini, M., Lewis, H. and Witkin, H. (1964), 'Some preliminary observations with an experimental procedure for the study of hypnagogic and related phenomena', *Archivo di Psicologia Neurologia e Psichiatra*, **6**, pp. 493–534.

Blackmore, S. (1980), 'The extent of selective reporting of ESP ganzfeld studies', *European Journal of Parapsychology*, **3**, pp. 213–19.

Blackmore, S. (1987), 'A report of a visit to Carl Sargent's laboratory', *Journal of the Society for Psychical Research*, **54**, pp. 186–98.

Braud, W.G., Wood, R. and Braud, L.W. (1975), 'Free-response GESP performance during an experimental hypnagogic state induced by visual and acoustic ganzfeld techniques: a replication and extension', *Journal of the American Society for Psychical Research*, **69**, pp. 105–13.

Child, I.L. and Levi, A. (1979), 'Psi-missing in free-response settings', *Journal of the American Society for Psychical Research*, **73**, pp. 273–89.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (Hillsdale, NJ: Erlbaum).

Dalton, K. (1997), 'Exploring the links: creativity and psi in the ganzfeld', *Proceedings of Presented Papers: The Parapsychological Association 40th Annual Convention*, pp. 119–34.

Hansel, C.E.M. (1989), *The Search for Psychic Power*: ESP and Parapsychology Revisited (Buffalo: Prometheus).

Harley, T. and Matthews, G. (1987), 'Cheating, psi, and the appliance of science: a reply to Blackmore', *Journal of the Society for Psychical Research*, **54**, pp. 199–207.

Hays, W.L. (1963), *Statistics for Psychologists* (New York: Holt, Rinehart & Winston).

Honorton, C. (1976), 'Length of isolation and degree of arousal as probable factors influencing information retrieval in the ganzfeld [Abstract]', in *Research in Parapsychology 1975*, ed. Joanna D. Morris, William G. Roll and Robert L. Morris (Metuchen, NJ: Scarecrow Press).

Honorton, C. (1978), 'Psi and internal attention states: information retrieval in the ganzfeld', in *Psi and States of Awareness*, ed. Betty Shapin and Lisette Coly (New York: Parapsychology Foundation).

Honorton, C. (1985), 'Meta-analysis of ganzfeld research: a response to Hyman', *Journal of Parapsychology*, **49**, pp. 51–9.

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**, pp. 99–139.

Hyman, R. (1981), 'Further comments on Schmidt's PK experiments', *Skeptical Inquirer*, **5** (3), pp. 34–40.

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

Hyman, R. (1994), 'Anomaly or artifact? Comments on Bem and Honorton', *Psychological Bulletin*, **115**, pp. 19–24.

Hyman, R. and Honorton, C. (1986), 'A joint communiqué: the psi ganzfeld controversy', *Journal of Parapsychology*, **50**, pp. 350–64.

Kanthamani, H. and Palmer, J. (1993), 'A ganzfeld experiment with "subliminal sending" ', *Journal of Parapsychology*, **57**, pp. 241–57.

Kennedy, J.E. and Taddonio, J.L. (1976), 'Experimenter effects in parapsychological research', *Journal of Parapsychology*, **40**, pp. 1–33.

Milton, J. (1996), 'Establishing methodological guidelines for ESP studies: a questionnaire survey of experimenters' and critics' consensus', *Journal of Parapsychology*, **60**, pp. 289–334.

Milton, J. (1999), 'Should ganzfeld research continue to be crucial in the search for a replicable psi effect? Part I. Discussion paper and an introduction to an electronic mail discussion', *Journal of Parapsychology*, **63**, pp. 309–33.

Milton, J. and Wiseman, R. (1999), 'Does psi exist? Lack of replication of an anomalous process of information transfer', *Psychological Bulletin*, **125**, pp. 387–91.

Milton, J. and Wiseman, R. (2001), 'Does psi exist? Reply to Storm and Ertel (2001)', *Psychological Bulletin*, **127**, pp. 434–8.

Palmer, J. (1983), 'Sensory contamination of free-response ESP targets: the greasy fingers hypothesis', *Journal of the American Society for Psychical Research*, **77**, pp. 101–13.

Palmer, J. and Aued, I. (1975), 'An ESP test with psychometric objects and the ganzfeld: negative findings [Abstract]', in *Research in Parapsychology 1974*, ed. Joanna D. Morris, William G. Roll and Robert L. Morris (Metuchen, NJ: Scarecrow Press).

Palmer, J. and Broughton, R.S. (2000), 'An updated meta-analysis of post-PRL ESP ganzfeld experiments', *Proceedings of Presented Papers: The Parapsychological Association 43rd Annual Convention*, pp. 224–40.

Radin, D. (1997), *The Conscious Universe: The Scientific Truth of Psychic Phenomena* (San Francisco: HarperEdge).

Rosenthal, R. (1979), 'The "file drawer" problem and tolerance for null results', *Psychological Bulletin*, **86**, pp. 638–41.

Rosenthal, R. (1986), 'Meta-analytic procedures and the nature of replication: the ganzfeld debate', *Journal of Parapsychology*, **50**, pp. 315–36.

Sargent, C. (1987), 'Sceptical fairytales from Bristol', *Journal of the Society for Psychical Research*, **54**, pp. 208–18.

Schlitz, M.J. and Honorton, C. (1992), 'Ganzfeld psi performance within an artistically gifted population', *Journal of the American Society for Psychical Research*, **86**, pp. 83–98.

Schmeidler, G.R. and Edge, H. (1999), 'Should ganzfeld research continue to be crucial in the search for a replicable psi effect? Part II. Edited ganzfeld debate', *Journal of Parapsychology*, **63**, pp. 335–88.

Stanford, R.G., Frank, S., Kass, G. and Skoll, S. (1989), 'Ganzfeld as an ESP-favorable setting: Part II. Prediction of ESP-task performance through verbal-transcript measures of spontaneity, suboptimal arousal, and internal attention state', *Journal of Parapsychology*, **53**, pp. 95–124.

Storm L. and Ertel S. (2001), 'Does psi exist? Milton and Wiseman's (1999) meta-analysis of ganzfeld research', *Psychological Bulletin*, **127**, pp. 424–33.

Willin, M.J. (1996a), 'A ganzfeld experiment using musical targets', *Journal of the Society for Psychical Research*, **61**, pp. 1–17.

Willin, M.J. (1996b), 'A ganzfeld experiment using musical targets with previous high scorers from the general population', *Journal of the Society for Psychical Research*, **61**, pp. 103–6.

Wiseman, R., Smith, M. and Kornbrot, D. (1996), 'Exploring possible sender-to-experimenter acoustic leakage in the PRL autoganzfeld experiments', *Journal of Parapsychology*, **60**, pp. 97–128.

York, M. (1977), 'The Defense Mechanism Test (DMT) as an indicator of psychic performance as measured by a free-response clairvoyance test using a ganzfeld technique [Abstract]', in *Research in Parapsychology 1976*, ed. William G. Roll and Robert L. Morris (Metuchen, NJ: Scarecrow Press).