Australian Journal of Parapsychology
2006, Volume 6, Number 1, pp. 35-53

Technical Paper No. 11

Meta-Analysis in Parapsychology: I. The Ganzfeld Domain[1]

LANCE STORM

ABSTRACT: The present article is a review of the ganzfeld meta-analytic literature. It is found that significant results were obtained in all but one ganzfeld meta-analysis—that of J. Milton and R. Wiseman (1999). However, with combinatorial re-construction of the available databases and the uncovering of 11 studies overlooked by Milton and Wiseman, L. Storm and S. Ertel (2001) reconfirmed that the ganzfeld was still the paradigm that delivered one of the highest effect sizes of all the experimental domains in parapsychology. More recent studies support this finding. Parapsychologist and pioneer of ganzfeld research, Charles Honorton (Honorton, 1985) said that the ganzfeld demonstrates a "significant psi effect" (p. 81), and the evidence in the present article supports that claim.

INTRODUCTION TO META-ANALYSIS

Meta-analysis, a term coined by G. V. Glass in the 1970s, is "the statistical analysis of the summary findings of many empirical studies" (Glass, McGaw, & Smith, 1981, p. 21). By combining the data from a number of different studies, meta-analysis produces an increased data-set that may be analysed to yield a more powerful result. Glass et al. (1981) describe the characteristics of meta-analysis in the following way:

[1] The author thanks the Bial Foundation for support in the preparation of this article.

1. "Meta-analysis is quantitative"—it organizes and extracts "information from large masses of data that are nearly incomprehensible by other means." (p. 22).

2. "Meta-analysis does not prejudge research findings in terms of research quality." This policy runs counter to the actions of critics of meta-analysis who would like to exclude many studies on the methodological grounds that they are poorly designed or treatments are badly implemented, even though "evidence is never given to support the assumption that these deficiencies . . . influence their findings." (p. 22)

3. "Meta-analysis seeks general conclusions," and "aims to derive a useful generalization that does not do violence to a more useful contingent or interactive conclusion." (pp. 22-23)

(Naturally, these assumptions have been criticised, and these assumptions will be addressed in the next section.)

The importance and usefulness of the 'meta-analytic' approach was demonstrated decades ago by prominent statisticians, psychologists, and medical researchers (Cochran, 1954; Edgington, 1972a, 1972b; Fisher, 1932; Mantel & Haenszel, 1959; Mosteller & Bush, 1954; Tippett, 1931). The flexibility and diversity of applications of meta-analysis have also been well reviewed and described.[1]

Bullock and Svyantek (1985) point out that meta-analysis is "objectively verifiable, using measured concepts, quantitative data, and statistical analysis" (p. 112). As opposed to the traditional literature review, where the reviewer is prone to his or her own "subjective interpretation of results across studies" (p. 112), the reviewer's selection bias in a meta-analytic review is more explicit, since the inclusion and exclusion of particular domains (i.e., 'experimental types' in the context of the present study) is so readily apparent. The advantage for the reader is that the philosophical persuasion and/or general intention of the author can be determined with ease so that the overall merits of the work can be more easily assessed.

---

[1] For examples, see Glass et al.'s (1981, pp. 25-26) meta-analyses of studies relating to the efficacy of psychotherapy in treatment of asthma and alcoholism. See also the meta-analyses by Cook et al. (1992) of juvenile delinquency studies, and studies on psychoeducational care with adult surgical patients, among others.

As mentioned, meta-analysis produces results that are not attainable by other means (at least, not without difficulty), but even these results may not end the debate over a specific area of controversy. Meta-analysis can never be conclusive. It provides a critical examination of the current status of research in a given area. However, by finding flaws in current meta-analytic techniques, and by identifying the more successful experimental domains through up-to-date meta-analytic procedures, newer, more focussed directions for parapsychology can be established, which, in the case of the parapsychologist, may help "settle the question" of whether or not psi exists (Broughton, 1991, p. 284).

PARAPSYCHOLOGY AND META-ANALYSIS

As will be seen in the next section and Part II of this series (in the next issue of *Australian Journal of Parapsychology*), meta-analysis has been a godsend for parapsychologists. Specifically, the direct benefits of meta-analysis for parapsychology become clear once it is understood that meta-analysis is the means by which the "signal" of the psi effect can be 'distilled' from the "noise" of chance with greater sensitivity than in any single study (Broughton, 1991, p. 281-282).[2]

Closely related to meta-analysis is the concept of replication. Rosenthal (1986) lists three main reasons for the so-called failure of many single studies to elicit significant psi effects, after previous successes with the same experimental design:

1. Pseudo-failure to replicate due to a poor consideration of the appropriateness of the statistical test(s) used. (See the illuminating example of a comparison between effect size and significance levels given by Rosenthal, 1986, pp. 317-318.) Essentially, the investigator may be too dependent on the significance (or not) of the *p* value.

[2] In fact, the presence ('signal') of a paranormal effect can never be specifically demonstrated as and when it occurs, even if 100% of trials are hits. For example, an experiment by Rhine (1937/1950, pp. 74-76) showed 25 hits in 25 trials, but in such an experiment there is no means of distinguishing the successful calls that suggest psi effects from those calls attributable to chance. The statistical inference of a psi effect is always made in the knowledge that the effect is produced within a stochastic framework.

2. Pseudo-successful replications, where *p* values are less than .05, but the effect sizes are significantly different, and therefore have not been replicated.

3. Successful replication of Type II error—not finding an effect when one was present due to small sample size, or very weak effects.

Consideration of the replication issue has led to an expansion of the meta-analytic methodology and a refinement of its technique (see Storm, 2005). Replication cannot be guaranteed, but it does come in forms other than rejection of the Null hypothesis (for example, replication of effect sizes, *z* scores, etc.).

Meta-analysis is fast becoming the *only* acceptable evidence that psi might well be an "anomalous effect in need of an explanation" (Utts, 1991, p. 363). However, there is now a growing suspicion of meta-analysis, not simply because it provides suggestive evidence of psi, but because in controversial non-parapsychological fields, such as psychotherapy, meta-analysis is also providing favourable evidence (Smith & Glass, 1977). Thus have some critics (Bandura, 1978; Eysenck, 1978; Oakes, 1986; Shapiro & Shapiro, 1977) focused on the construct validity of meta-analysis. Arguments are made that data from methodologically flawed experiments are thrown into the 'soup' with the data from better-designed experiments, therefore corrupting the result. Different procedures in ostensibly the same kind of experiment may also yield a tainted result.

Of course, the general implication derived by critics (that meta-analysis is a flawed procedure) also applies to meta-analysis in parapsychology. But these criticisms have not been ignored, and improvements have been made to control for the confounding factors. Rosenthal (1984) advocates differential weighting (i.e., effect size values are adjusted) as an effective way of dealing with "variation in the quality of research" (p. 127). Hence the 'blocking' procedure is used to code experiments according to their 'quality' and type (i.e., methodology, hypothesis, etc.).

Sample size and the population from which the sample is drawn are also critical considerations. Credit is given to studies if sample size is specified in advance, as well as the nature of the analyses—pre-planned or post hoc. Acceptable randomisation methods are also credited, and even the date of the experiment and the identity of the investigator are now important criteria in meta-analytic studies (Broughton, 1991, p. 283). Although these procedures may be seen as subjective, some degree of qualitative assessment can be made about studies, and these assessments are converted to numerical values to arrive at a more objective, albeit pseudo-precise,

numerical result that is still seen as a gain over previous methods which did not consider study quality.

One important criticism made against meta-analysis in parapsychological research (applicable to meta-analysis in general) is that significant results are inevitable, since the majority of studies used in the analysis have significant results already. On the other hand, studies with nonsignificant results are never or are rarely published and, therefore, cannot be included in the meta-analysis (Hyman, 1985). This problem is often referred to as 'selective reporting'. There are three factors that belie this criticism.

First, parapsychology journals go to great efforts to publish studies with nonsignificant results—they tend not to end up in the 'file-drawer'. In 1975 the Parapsychological Association Council adopted a policy of opposing the exclusive publication of studies with positive outcomes. Thereby, "negative findings have been routinely reported at the association's meetings and in its affiliated publications" since that date (Bem & Honorton, 1994, p. 6; see also Honorton, 1985, p. 66.)

Second, estimates can be made which account for unpublished studies. Usually, the number of nonsignificant studies that would be needed to reduce a significant meta-analytic result to a chance outcome is shown to be far in excess of that which would be possible for the few researchers in the field of parapsychology (Broughton, 1991, p. 286; Utts, 1991, pp. 370, 372, 375-376).

Third, the so-called funnel-plot technique allows meta-analysts the means by which all the studies used in the meta-analysis can be distributed and presented on a two-axis array (effect size on the x-axis, and number of studies, *N*, on the y-axis), usually appearing as a scatter of data-points that look like an inverted funnel-shape evenly distributed around a mean effect size value. The funnel shape results from the general rule that effect sizes tend to approach zero as *N* increases. If the funnel-plot is asymmetrical, the researcher can determine how many studies (and their effect size values) are *theoretically* missing in order to produce a symmetrical plot.

Rosenthal (1984) has also addressed many criticisms levelled at meta-analysis, including exaggeration of significance levels, which can be a problem since increasing the number of studies in a meta-analysis increases the probability of rejecting the Null hypothesis. Rosenthal argues that when the Null is false it ought to be rejected, but notes that if such a characteristic of meta-analytic procedure increases its accuracy and decreases the likelihood of Type II errors, it must be an advantage. Alternatively, should it be possible that meta-analysis increases the likelihood of Type I errors (i.e., when the Null is really true), an increase in the number of studies does

not increase the probability of rejecting the Null, or the size of the estimated effect.

Oakes (1986) questions the validity of a procedure that doesn't really test for a 'directional hypothesis' (e.g., "do men perform better than women?" p. 162). For parapsychology, at least, domain-specific meta-analyses clearly refer to directional hypotheses. Generally, what are sought in the meta-analyses of these specific domains are significant effect sizes, significant differences in psi performance (above or below mean chance expectation), or significant differences in scoring between comparison groups.

The dubious value of "meta-meta-analysis," criticised by Oakes (1986, p. 162), but endorsed by Glass (1981, cited in Oakes, 1986, p. 162) as a reasonable undertaking, is also acknowledged (Glass, McGaw, & Smith, 1981, believe it is a valid exercise to mix studies on "apples and oranges," p. 218, if your hypotheses are about fruit!). Such a 'method' may be superfluous and even of no validity, furnishing (it would seem) a rather vague, nebulous, and therefore, ambiguous finding, whether significant or not.

As it happens, parapsychology is bereft of such adventurous undertakings. The idea, for example, that ganzfeld studies could be meta-analysed with dice-throwing studies is untenable in the extreme. For example, a significant result in one single meta-analysis of both domains combined would not draw out the possible effect size difference between these two domains. Should there be a significant effect size difference, it would only draw out the distinction that either two types of psi were in operation, or the paradigms are more conducive to psi in one domain, and less conducive in the other. Thus there are limits to how far we can apply Glass et al.'s fruit analogy.

Having described the current status of meta-analysis, and argued that meta-analysis has a relative degree of validity, and therefore, that there is scientific value in its processes, the following section is a presentation of the findings of the meta-analyses of ganzfeld studies dating back to the mid-1970s.

REVIEW OF THE GANZFELD META-ANALYSES

The meta-analyses now reviewed used as their sources bibliographic databases for parapsychology, and all the parapsychological journals, including publications of technical reports, conference proceedings, and manuscripts, etc., and, in some cases, physics journals (e.g., Radin and Nelson, 1989).

The Ganzfeld Procedure

The ganzfeld is a form of free-response test—'free response' being a term that "describes any test of ESP in which the range of possible targets is relatively unlimited and is unknown to the percipient" (Thalbourne, 2003, p. 44). The target is not restricted to a few choices, but can be almost anything, thus hopefully reducing the risk of boredom so common in forced-choice experiments because free responses "more nearly resemble the conditions of spontaneous psi occurrences" (Burdick & Kelly, 1977, p. 109).

The ganzfeld is a "special type of environment (or the technique for producing it) consisting of homogeneous, un-patterned sensory stimulation" to the eyes and ears of the participant who is usually in "a state of bodily comfort" (Thalbourne, 2003, p. 45). A number of investigators pioneered the technique in the 1970s (Braud, Wood, & Braud, 1975; Honorton & Harper, 1974; Parker, 1975).

Procedurally, the eyes of the participant are covered with halved ping-pong balls illuminated by a uniform source of light (usually of a single wavelength, such as red light). A uniform auditory signal of "white" noise (full-range audio signal; see Thalbourne, 2003, p. 45, and Utts, 1991, p. 369) or "pink" noise (high-frequency filtered sound; see Stanford, 1979, p. 253) is channelled through headphones to the ears. The participant reclines on a chair or lies on a bed. This technique has remained essentially the same since the 1970s.

The Ganzfeld Experiments

The first major meta-analytic study in parapsychology started in 1981 when Hyman (1985) began evaluating 42 ganzfeld psi studies conducted during the period 1974 to 1982. Hyman initially chose the ganzfeld studies because they supposedly held a "high level of research sophistication and rigor" (Hyman, 1985, p. 4)—a claim that Hyman was to criticise heavily.

A drawn-out debate ensued between Hyman and Honorton, since they arrived at conflicting conclusions from the same data set. Hyman first argued that the "alleged" 55% success rate of 42 studies determined from a vote-count made by Honorton (Hyman, 1985, p. 5) was inflated due to the fact that many of the studies were not independent (they were more like subsets of ongoing experiments).

Hyman also cited evidence that suggested there was bias in how the studies were reported. For example, some studies were not planned as such, but were "given this status retrospectively just because they yielded significant results" (Hyman, 1985, p. 16). Hyman reduced the success rate

to 31% (he actually argued for less than 30%, given that there must be unsuccessful but unknown ganzfeld studies yet to be considered—the 'file-drawer' problem mentioned above).

Hyman (1985) further criticised many of the studies for their multiple analyses (e.g., use of a number of measures of ESP), which gave increased opportunity for a good result, especially since investigators were not adjusting their criterion significance levels according to the number of statistical tests they performed. He also claimed that independence had been violated in some meta-analytic studies because "agents were friends of the percipient . . . [or were even] members of laboratory staff" (1985, p. 26).

Of interest is Child's (1986, pp. 337-343) comment on a procedural flaw where pooling results based on groups or conditions can actually conceal an effect rather than erroneously identify one. Optimal randomization could not be assumed for such studies. Child indicated that hit-rates can vary systematically from individual to individual, or group to group, so that "genuinely high performances of some may well be buried by the chance performance of many others" (1986, p. 339).

Honorton (1985) accepted the criticism of multiple analysis, and he applied a Bonferroni correction across all studies. He found that only 45% of the 42 studies were significant—not 55% which he originally claimed (but 45% was still higher than Hyman's lower estimate of 31%). Honorton then used the proportion of direct-hits as a common index, since it was the most common measure in the studies (also the most conservative). A total of 28 studies using direct hits alone were thus employed in the meta-analysis, 7 (25%) of which were independently significant at $p \leq .01$, and 12 (43%) of which were significant at $p \leq .05$ (see Table 1).

Honorton (1985) noted that of the 28 studies, 23 (82%) had positive *z* scores. (The probability of this outcome is shown in Table 1, along with the number of studies, also expressed as percentages, and their significance levels.) Honorton reported a composite Stouffer Z score of 6.60 across the 28 studies.[3] Table 1 also includes effect size measures as π values because of their ease of interpretation—π "depends simply on k, the number of alternative choices available, and P, the raw proportion of hits"[4] (Rosenthal & Rubin, 1989, p. 333). Using the mean effect size (*ES*) formula, $\Sigma[z/\sqrt{n}]/k$, the *ES* for the 28 studies was .26.

---

[3] "Stouffer's *Z* is found by dividing the sum of the *z* scores for the individual studies by the square root of the number of studies" (Rosenthal, 1978, p. 6).

[4] $\pi = P(k - 1)/[1 + P(k - 2)]$. Bem and Honorton (1994, p. 8) point out the advantage this measure has in providing a "straightforward intuitive interpretation" of the effect size, because π is the "proportion correct, transformed to a two-choice standard situation" so that $P_{MCE} = P_{test} = .50$ (Rosenthal & Rubin, 1989, p. 333).

Table 1
Meta-Analysis[5] of the 28 Direct-Hit Ganzfeld Studies (and their Subgroups)

Number of studies	Independent *p* value	Proportion of Hits	Effect Size (π)	Stouffer *Z*	Probability (*p*)
7 (25%)	.01	0.47	0.73	8.63	9.80×10^{-9}
12 (43%)	.05	0.46	0.72	10.46	3.50×10^{-9}
23 (82%)	n/a[a]	0.40	0.67	8.42	4.60×10^{-4}
28 (100%)	n/a[b]	0.38[c]	0.65	6.60	2.10×10^{-11}

Notes: [a] Studies with positive z scores (Exact binomial test with $p = q = .5$)
[b] Includes five studies with negative *z* scores
[c] Rosenthal's (1986, p. 333) more conservative estimate of the proportion of hits is "about 1/3." Bem and Honorton (1994, p. 8) give a value of .35, and thus calculate π as .62

Honorton (1985, p. 59) also calculated a more conservative estimate of significance by including 10 additional blind-judging studies that did not report direct-hit information. Assuming a mean z score for these 10 studies of zero, Stouffer $Z = 5.67$ ($p = 7.30 \times 10^{-9}$). Such a probability still indicates how extremely unlikely it would be that these successful ganzfeld studies were all the result of chance.

Using the blocking technique, six of the ten independent investigators who produced these studies achieved significant results, so that neither a specific investigator, nor a specific laboratory was single-handedly responsible for the significant results. The suggestion of a file-drawer problem was also rendered less plausible by the fact that 15 nonsignificant and unknown studies would have to exist for every one of the 28 direct hit studies to reduce the result to a chance outcome.

But other problems had to be addressed. Hyman (1985, pp. 30-35) found that flaws correlated positively with significant results. He identified 12 major flaws, such as inadequate randomisation of targets, and failure to use a duplicate set of targets for judges. Hyman (1985, pp. 35-36) used cluster and factor analysis on these 12 flaws, combining them into 3 new variables: General Security, Statistics, and Controls—upon which were conducted several analyses. The most detailed (factor) analysis was one

[5] Most of these data come from Honorton (1985), or are calculated from the data provided in that study.

consisting of supposedly 17 variables, from which emerged four factors (actually, there were only 16 variables, according to Saunders, 1985, p. 97). Utts (1991) paraphrased the findings of this analysis:

> From these [four factors], Hyman concluded that security had increased over the years, that the significance level tended to be inflated the most for the most complex studies and that both effect size and level of significance were correlated with the existence of flaws. (p. 371)

Hyman's adjusted figure for the number of successful studies which would not possess these flaws was 27% of all the studies considered—"well within the statistical neighborhood of the 25% chance rate" (1985, p. 37). Honorton acknowledged the problems with the studies, but Saunders (1985), on behalf of Honorton, repudiated Hyman's "meaningless" analysis and its "logical problems" (Saunders, 1985, p. 87). Saunders found a violation in statistical procedure in Hyman's factor analysis: "the size of the available database marginally suffices to support [only] one factor" (p. 87).

Hyman had performed a multiple analysis that included the three flaws just mentioned, but out of nine potential flaws (giving 84 sets of three) Hyman (conveniently) selected the set that correlated highly with effect size. Thus the impression was given that, as Hyman implied (1985, p. 37), effect size was a function of procedural flaws (the more flaws in an experiment, the higher the effect size). Saunders noted that Hyman's multiple correlations that resulted from selective testing should be regarded as nonsignificant, rendering Hyman's adjusted figure of 27% meaningless.

Rather than continue the debate, Hyman and Honorton produced a "Joint Communiqué" (Hyman & Honorton, 1986) addressing fundamental issues in parapsychological experimentation. The Communiqué recommended that "more stringent standards" be implemented in experiments, which should also be conducted by a "broader range of investigators" (p. 351). Utts (1991 listed these standards as including:

> controls against any kind of sensory leakage, thorough testing and documentation of randomization methods used, better reporting of judging and feedback protocols, control for multiple analyses [and statistics] and advance specification of number of trials and type of experiment. (p. 371)

Hyman and Honorton (1986) also believed that meta-analysis had a growing role in the evaluation of "research quality and the assessment of moderating variables" (p. 361).

A number of researchers commented on the "Joint Communiqué" and most were in general agreement with its recommendations, though all had unique points to make about the state of affairs of parapsychology. Hövelmann (1986, p. 366) felt that the participant should be left alone at the judging stage (no presence of the experimenter), since even the non-verbal behavior of a non-blind experimenter may have an influence on the judging outcome. Usually, the use of 'blind' experimenters avoids this problem.

Palmer (1986, p. 379) argued that the presence of Hyman's identified flaws in a ganzfeld experiment does not mean replication of positive results will continue in the future. He added that the absence of flaws would not necessarily guarantee positive results either. Nor should it be assumed that failure to replicate when the flaws are removed means that past successes were due to the presence of the flaws.

Stanford (1986, p. 384) expressed his unease about both Hyman's and Honorton's readiness to make a cause célèbre out of the ganzfeld by yielding it up to the National Science Foundation for extensive replicability studies. The worst-case scenario of dismal failure could damage the field of parapsychology, not to mention the careers and professional lives of parapsychologists. It is too early for parapsychologists to be so confident when ganzfeld-ESP success looks more like "art" than "science" (p. 386).

Utts (1986), as a statistician, felt that power considerations must be undertaken more often in experimental design since the "replicability" problem in so many experiments can be due to poor consideration of the sample sizes needed in certain experiments. Real effects can be lost if N is too small, whereas a larger N (as in meta-analytic studies) only increases the chance of getting a significant result.

Rosenthal's (1986) commentary on the meta-analysis also focused on replication, and so included a consideration of effect size. He used Cohen's *h*, which is the transformed proportion of direct hits.[6] Rosenthal calculated that 23 (82%) of the direct hit studies had effect sizes greater than zero. The mean effect size *h* was 0.28, corresponding to the significant direct-hit-rate of 0.38 reported in Table 1 (where $P_{MCE} = .25$). Rosenthal recommended that effect size be the preferred measure of replication success over and above that of significance testing which has "nothing to do with success of replication" (1986, p. 334).

[6] Cohen's $h = 2(\arcsin \sqrt{p'} - \arcsin \sqrt{p})$, where p' is the proportion of observed direct hits, and p is the proportion of expected number of direct hits.

Milton (1997) conducted one recent meta-analysis relevant to the ganzfeld domain. She meta-analysed 46 free-response[7] studies (including 42 ganzfeld studies) to determine which measure—direct hits or sums of ranks—was the more sensitive of the two. The cumulative result of the 46 studies was significant, suggesting a psi effect. In considering only effect sizes and *p* values, sums of ranks "outperformed" direct hits (Milton, 1997, p. 227). However, there was no statistically significant degree of difference between the two measures. Milton called for caution until further research might be more conclusive about which of the two techniques should be considered 'superior', since the nonsignificant difference may have been a chance result.

The Autoganzfeld Experiments

The automated ganzfeld (i.e., "autoganzfeld") procedure was adopted as a more rigorous approach to psi testing, while still maintaining the ganzfeld paradigm. It came into being as a proactive response to the "Joint Communiqué" recommendations. Thus some strict guidelines were implemented in the autoganzfeld, the major one being the introduction of a computer-controlled, randomly-selected, presented, and scored target, which was therefore unknown to all those involved in the experiment except the sender. Feedback is eventually given to the receiver in the form of the correct choice. As in the ordinary ganzfeld, targets can be "dynamic" (short scenes from movies, cartoons, and documentaries, etc.), or "static" (photographs, art prints, advertisements, etc.).

A series of 11 autoganzfeld experiments was conducted by eight experimenters during the period 1983-1989 (see Honorton et al., 1990). As reported in Bem and Honorton (1994) there were 106 direct hits in 329 trials for 10 of these studies yielding a 32.2% hit-rate (P_{MCE} = .25), Stouffer $Z = 2.61$, $p = 4.50 \times 10^{-3}$ (mean $ES = 0.117$; $\pi = 0.59$). The π value for this series of experiments is comparable to the π value of .62 given by Bem and Honorton (1994, p. 8) for the 28 direct-hit non-automated ganzfeld studies. The eleventh study, which used dynamic targets exclusively, had the highest hit-rate (54%) and was in fact significantly higher than any of the other ten studies. The study was rejected by Bem and Honorton (1994) due to "response biases" (pp. 11-12).

---

[7] The free-response method describes any test of ESP using a relatively unlimited range of possible targets, thus permitting the participant to "respond freely with whatever impressions come to mind" (Thalbourne, 2003, p. 44). The participant may, for example, respond by drawing a pictorial representation of the target.

The hit-rate for all the 'dynamic' target studies (164 sessions) was 37% ($\pi = 0.74$), a considerable difference compared to the hit-rate for all the 'static' target studies (165 sessions) of only 27% ($\pi = 0.52$). Bem and Honorton (1994, p. 12) note these differences as suggestive evidence that, generally, dynamic targets may be "more effective than static targets." There was also a distinct difference in success rate between experienced participants and novices, suggesting that experienced participants (those previously tested) yield better results than novices (experienced participants' hit-rate: 37%, $\pi = 0.64$; novices' hit-rate: 32.5%, $\pi = 0.59$).

When Rosenthal (1986, p. 333) adjusted for the flaws in the earlier ganzfeld studies, he arrived at a conservative estimate of "about 1/3," thus reducing the original 38% hit-rate to a hit-rate roughly equivalent to the 10 autoganzfeld studies of 32.2%. The ordinary ganzfeld and the autoganzfeld appeared to be equally effective, since they produced effect sizes in roughly the same vicinity ($\pi = 0.62$, $\pi = 0.59$, respectively). After two nonsignificant performance comparisons (on effect sizes and z scores) between Honorton's (1985) database of 28 studies and the new Honorton et al. (1990) database of 11 autoganzfeld studies, Honorton et al. (1990, p. 99) combined the two databases into a 39-study database, Stouffer $Z = 7.53$, $p = 9.00 \times 10^{-14}$ (Cohen's $h = 0.28$).

One criticism levelled at the autoganzfeld meta-analysis was that the eleven experiments were conducted by only eight experimenters, all of whom were at the same laboratory. Consequently, Milton and Wiseman (1999) conducted a meta-analysis of new ganzfeld studies dating from 1987 to 1997. Studies prior to 1987 were not used because it was assumed that investigators needed time to familiarise themselves with Hyman and Honorton's (1986) guidelines so that earlier studies would be too flawed for serious consideration in a meta-analysis.

Milton and Wiseman (1999) deemed suitable for analysis thirty studies by "10 different principal authors from 7 laboratories" (p. 388). They calculated a Stouffer Z of 0.70, $p = .24$, one-tailed ($ES = 0.013$), and concluded that a significant psi effect for the ganzfeld had not been replicated by a "broader range of researchers" (p. 391).

Storm and Ertel (2001) singled out Milton and Wiseman's (1999) main finding of a nonsignificant *ES* of 0.013 and disputed its derivation. Storm and Ertel argued that Milton and Wiseman did not adopt a 'responsible' attitude in their meta-analysis. A thorough meta-analysis requires a comprehensive literature search and an accumulative approach to the available databases (this approach was not adequately demonstrated in Milton and Wiseman's paper). Arbitrary exclusion rules and unjustifiable, *a posteriori* periods of analysis should be considered unacceptable in any meta-analysis.

Storm and Ertel (2001) found 11 pre-Communiqué studies not previously meta-analysed, and after step-by-step performance comparisons, combined them with the three ganzfeld databases currently extant: Honorton's (1985) database of 28 studies, Bem and Honorton's (1994) databases of 10 studies, and Milton and Wiseman's (1999) database of 30 studies. The resulting 79-study database had a significant mean *ES* of 0.14 (Z = 5.66, p = 7.78 x 10^{-9}). Milton and Wiseman's negative conclusion about the failure of the ganzfeld to replicate is rather misleading and premature as it is pertinent to a limited pool of only 30 studies.

In reply to Storm and Ertel (2001), Milton and Wiseman (2001) argued that the 11 pre-Communiqué studies used in their meta-analysis should not have been used at all because they were poor in quality due to their ostensible "methodological problems" (p. 434). Thus, Milton and Wiseman clearly ignored Storm and Ertel's (2001) performance comparisons of (a) pre-Communiqué authors with post-Communiqué authors, and (b) pre-Communiqué studies with post-Communiqué studies, both of which yielded no statistical evidence that the guidelines in the Communiqué had any "influence on effect size outcomes" (p. 430). Logically, there was no indication that the mean effect size of the pre-Communiqué database was 'inflated' (i.e., an artifact of flaws) because it compared favourably with the allegedly 'flawless' post-Communiqué studies. It follows that there was no evidence that the mean effect size of the post-Communiqué database was 'deflated' due to the removal of these flaws. Apropos to these findings, Palmer (1986) had warned earlier, that false conclusions can be drawn on account of, and by appeal to, the Communiqué's guidelines—it should not be assumed that "past successes were due to the presence of the flaws" (p. 379).

Milton and Wiseman's (2001, p. 436) only other major criticism concerned the lack of conservative calculations of some z scores for studies in the 11-study database. In fact, only 3 of the 11 studies needed adjustment, thus reducing the quality-weighted mean z score from 0.32 (*ES* = 0.14; Stouffer Z = 1.06, p = .144) to 0.26 (*ES* = 0.13; Stouffer Z = 0.87, p = .192). The 11-study database is still not significantly different from Honorton's (1985) 28-study database, $t(37)$ = 0.61, p = .543, two-tailed. The Old Ganzfeld Database of 39 studies (i.e., 28 + 11) can still be formed. It has a mean z of 0.97 (*ES* = .225; Stouffer Z = 6.05, p = 7.24 x 10^{-10}; cf. Storm & Ertel's, 2001, p. 429, original data for the Old Ganzfeld Database: mean z of 0.99, *ES* = .227; Stouffer Z = 6.15, p = 3.93 x 10^{-10}).

The 'Old' (Pre-Communiqué) database (N = 39 studies) and the 'New' (Post-Communiqué) database (N = 40 studies; i.e., 10 + 30) are significantly different, $t(77)$ = 3.04, p = .003, ω^2 = 0.09, but the omega squared value (9%) is now exactly that of the critical value stipulated in

Storm and Ertel's (2001) paper. The difference might be considered important, but Cohen's (1988) test, as originally applied by Storm and Ertel, was again not significant. When the two databases are combined the 79-study database has a mean *z* score only slightly reduced from 0.64 to 0.63 ($ES = 0.14$; Stouffer $Z = 5.59$, $p = 1.14 \times 10^{-8}$).

As of 2004, there is a total of 88 ganzfeld/autoganzfeld studies with an accumulated hit-rate of 1008 hits out of 3145 trials (32%) and a corresponding *p* value in the order of 3.45×10^{-20}. There would need to be over 2000 unpublished studies lying around in file-drawers, all with null results, to reduce this significant hit-rate to chance (Radin, 2006). This largest-ever database has not only unified the ganzfeld paradigm, but also indicates that over three decades of ganzfeld/autoganzfeld work has not been in vain. The ganzfeld procedure might yet prove to be the ideal paradigm that Honorton hoped it might be for finding "strong evidence for psychic functioning" (Milton & Wiseman, 1999, p. 391).

CONCLUSION

Results from the above reviews of the meta-analytic literature indicate, as Honorton believed, that the ganzfeld represents an encouraging step toward replicability of psi effects. The paradigm has been fraught with controversy, dispute, and ongoing debate. The major difficulty lay in establishing ground-rules for conducting ganzfeld experiments that all researchers could agree upon. Currently, a joint communiqué exists that features guidelines for ganzfeld design and procedure that researchers adhere to, by and large. But as Storm and Ertel (1999) pointed out, the instigation of the communiqué in 1986 does not mean that pre-1986 studies should ever be considered unreliable and flawed.

It is a well-accepted fact that parapsychologists conduct extremely rigorous experiments with very much tighter controls compared to other disciplines (cf. Sheldrake, 1998)—not because psi is so elusive, but because of the controversial nature of psi that invariably compels non-parapsychologists to attack parapsychology at its core (i.e., the way parapsychologists design their experiments). It has reached the point where conventional explanations like sensory leakage, selective reporting, and outright fraud, are not only passé, but also insulting to the professionally minded parapsychologist.

In a follow-up article, reviews of the meta-analyses continue into the other (non-ganzfeld) domains: (i) biological systems, (ii) forced-choice, (iii) free-response, (iv) dice-throwing, (v) Micro-PK (RNG), and (vi) dream-ESP.

REFERENCES

Bandura, A. (1978). On paradigms and recycled ideologies. *Cognitive Therapy and Research*, *2*, 79-103.

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

Braud, W. G., Wood, R., & 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*, 105-113.

Broughton, R. S. (1991). *Parapsychology: The controversial science.* New York: Ballantine.

Bullock, R. J., & Svyantek, D. J. (1985). Analyzing meta-analysis: Potential problems, an unsuccessful replication, and evaluation criteria. *Journal of Applied Psychology*, *70*, 108-115.

Burdick, D. S., & Kelly, E. F. (1977). Statistical methods in parapsychological research. In B. B. Wolman (Ed.), *Handbook of parapsychology* (pp. 81-130). New York: Van Nostrand Reinhold.

Child, I. L. (1986). Comments on the ganzfeld controversy. *Journal of Parapsychology*, *50*, 337-343.

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

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T. A., & Mosteller, F. (1992). *Meta-analysis for explanation.* New York: Russell Sage Foundation.

Edgington, E. S. (1972a). An additive method for combining probability values from independent experiments. *Journal of Psychology*, *80*, 351-363.

Edgington, E. S. (1972b). A normal curve method for combining probability values from independent experiments. *Journal of Psychology*, *82*, 85-89.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, *33*, 517.

Fisher, R. A. (1932). *Statistical methods for research workers*. London: Oliver and Boyd.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* London: Sage.

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

Honorton, C., & Harper, S. (1974). Psi-mediated imagery and ideation in an experimental procedure for regulating perceptual input. *Journal of the American Society for Psychical Research*, *68*, 156-168.

Honorton, C., Berger, R. E., Varvoglis, M. P., Quant, M., Derr, P., Schechter, E. I., & Ferrari, D. C. (1990). Psi communication in the ganzfeld: Experiments with an automated testing system and a comparison with a meta-analysis of earlier studies. *Journal of Parapsychology*, *54*, 99-139.

Hövelmann, G. H. (1986). Beyond the ganzfeld debate. *Journal of Parapsychology*, *50*, 364-370.

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

Hyman, R., & Honorton, C. (1986). Joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology*, *50*, 351-364.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Milton, J. (1997). A meta-analytic comparison of the sensitivity of direct hits and sums of ranks as outcome measures for free-response studies. *Journal of Parapsychology*, *61*, 227-237.

Milton, J., & Wiseman, R. (1999). Does psi exist? Lack of replication of an anomalous process of information transfer. *Psychological Bulletin*, *125*, 387-391.

Milton, J., & Wiseman, R. (2001). Does psi exist? Reply to Storm and Ertel (2001). *Psychological Bulletin*, *127*, 434-438.

Mosteller, F., & Bush, R. R. (1954). Selected quantitative techniques. In G. Lindsay (Ed.), *Handbook of social psychology.* Cambridge, MA: Addison-Wesley.

Oakes, M. (1986). *Statistical inference: A commentary for the social sciences.* Chichester, UK: John Wiley.

Palmer, J. (1986). Comments on the "Joint Communiqué." *Journal of Parapsychology*, *50*, 377-381.

Parker, A. (1975). Some findings relevant to the change in state hypothesis. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology 1974* (pp. 40-42). Metuchen, NJ: Scarecrow Press.

Radin, D. (2006). *Entangled minds*. New York: Paraview/Pocket.

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

Rhine, J. B. (1937/1950). *Frontiers of the mind.* Harmondsworth, Middlesex: Pelican/Penguin.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin*, *85*, 185-193.

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

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

Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin*, *106*, 332-337.

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

Shapiro, D. A., & Shapiro, D. (1977). The 'double standard' in evaluation of psychotherapies. *Bulletin of the British Psychological Society*, *30*, 209-210.

Sheldrake, R. (1998). Experimenter effects in scientific research. How widely are they neglected? *Journal of Scientific Exploration*, *12*(1), 73-78.

Smith, M., & Glass, G. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*, 752-760.

Stanford, R. G. (1979). The influence of auditory ganzfeld characteristics upon free-response ESP performance. *Journal of the American Society for Psychical Research*, *73*, 253-272.

Stanford, R. G. (1986). Commentary on the Hyman-Honorton joint communiqué. *Journal of Parapsychology*, *50*, 383-388.

Storm, L. (2005). A socioempirical perspective on skepticism about psi. In M. A. Thalbourne & L. Storm (Eds.), *Parapsychology in the 21st century: Essays on the future of psychical research* (pp. 275-304). Jefferson, NC: McFarland.

Storm, L., & Ertel, S. (2001). Does psi exist? Comments on Milton and Wiseman's (1999) meta-analysis of ganzfeld research. *Psychological Bulletin*, *127*, 424-433.

Thalbourne, M. A. (2003). *A glossary of terms used in parapsychology* (2nd ed.). Charlottesville, VA: Puente.

Tippett, L. H. C. (1931). *The method of statistics.* London: Williams and Norgate.

Utts, J. (1986). The ganzfeld debate: A statistician's perspective. *Journal of Parapsychology*, *50*, 394-402.
Utts, J. (1991). Replication and meta-analysis in parapsychology. *Statistical Science*, *6*, 363-378.

Anomalistic Psychology Research Unit
School of Psychology
University of Adelaide
Adelaide SA 5005
AUSTRALIA

Email: lance.storm@ adelaide.edu.au