Psychological Bulletin
2001, Vol. 127, No. 3, 424–433


0033-2909/01/$5.00 DOI: 10.1037//0033-2909.127.3.424

# Does Psi Exist? Comments on Milton and Wiseman's (1999) Meta-Analysis of Ganzfeld Research

Lance Storm
Adelaide University

Suitbert Ertel
Georg August Universität

J. Milton and R. Wiseman (1999) attempted to replicate D. Bem and C. Honorton's (1994) meta-analysis, which yielded evidence that the ganzfeld is a suitable method for demonstrating anomalous communication. Using a database of 30 ganzfeld and autoganzfeld studies, Milton and Wiseman's meta-analysis yielded an effect size (*ES*) of only 0.013 (Stouffer $Z = 0.70$, $p = .24$, one-tailed). Thus they failed to replicate Bem and Honorton's finding ($ES = 0.162$, Stouffer $Z = 2.52$, $p = 5.90 \times 10^{-3}$, one-tailed). The authors conducted stepwise performance comparisons between all available databases of ganzfeld research, which were argued not to be lacking in quality. Larger aggregates of such studies were formed, including a database comprising 79 ganzfeld–autoganzfeld studies ($ES = 0.138$, Stouffer $Z = 5.66$, $p = 7.78 \times 10^{-9}$). Thus Bem and Honorton's positive conclusion was confirmed. More accurate population parameters for the ganzfeld and autoganzfeld domains were calculated. Significant bidirectional psi effects were also found in all databases. The ganzfeld appears to be a replicable technique for producing psi effects in the laboratory.

Parapsychologists are often challenged by skeptics to provide evidence that psi exists. Bem and Honorton (1994) defined *psi* as "anomalous processes of information or energy transfer such as telepathy or extrasensory perception that are currently unexplained in terms of known physical or biological mechanisms" (p. 4). One experimental technique designed to test for such anomalous effects is the ganzfeld procedure (the German term *Ganzfeld* means total field), where a "sender" in one room is required to "physically communicate" one of four picture targets or movie-film targets to a "receiver" in another room, who is in the ganzfeld condition of homogeneous sensory stimulation (Milton & Wiseman, 1999, p. 387). This condition is alleged to enhance psychic functioning in the receiver.

Honorton (1985) undertook a meta-analysis of all the ganzfeld studies up to and including January 1982 to determine whether or not there was evidence for an overall psi effect induced by the ganzfeld condition. He arrived at a database of 28 direct-hit (first-rank) ganzfeld studies that demonstrated a "significant psi ganzfeld effect" (p. 81). Expert discussants considered this result an encouraging step toward replicability of psi effects. Indeed, Honorton, in his meta-analysis, had gone to a great deal of effort to rectify the earlier methodological faults of less attentive researchers, such as selective reporting, multiple testing, and so on, as pointed out by Hyman (1985), a member of the Committee of Scientific Investigation of Claims of the Paranormal (CSICOP), an organization that aims at defending orthodox science against wanton paranormal claims.

Ultimately, Hyman and Honorton came to an agreement (Hyman & Honorton, 1986). Their very words from the "Joint Communiqué" were "we agree that there is an overall significant effect in this database that cannot reasonably be explained by selective reporting or multiple analysis" (p. 351). They differed only on the "degree to which the effect constitutes evidence for psi" (p. 351), meaning that the unresolved issue between the two was over the actual size of the effect. In other words, there was agreement between the two that an effect existed. Numerous claims that flaws in Honorton's (1985) meta-analysis still exist have been debunked (Atkinson, Atkinson, Smith, & Bem, 1990; Harris & Rosenthal, 1988a, 1988b; Saunders, 1985; Utts, 1991).

Honorton et al. (1990) followed with a meta-analysis of 11 autoganzfeld studies, which adhered to the guidelines laid down by Hyman and Honorton (1986). The autoganzfeld procedure was designed to avoid the accusations about methodological flaws leveled at the earlier ganzfeld studies by using a computer-controlled target randomization, selection, and judging technique (hence the term *autoganzfeld*).

Bem and Honorton (1994) reduced the Honorton et al. (1990) database to 10 studies by removing one study because of its "response bias."[1] Bem and Honorton (p. 10) reported a hit rate of 32.2% (106 hits in 329 trials, $z = 2.89$, $p = .002$, one-tailed) for these 10 "waterproof" studies. The effect size $\pi$ was .59, where $\pi_{MCE} = .50$ (MCE = mean chance expectation). The $\pi$ value of .59 is equivalent to an *ES* value of 0.16. (In our study, we did not use $\pi$ but used the mean effect size *ES*, an estimate of *r*, as given by the formula $\Sigma [z/\sqrt{n}]/k$).

Only 5 years after Bem and Honorton's (1994) report, Milton and Wiseman (1999) published an account of another meta-analysis of ganzfeld studies. Their intention was to replicate Bem and Honorton's earlier meta-analysis of 10 ganzfeld studies. They

Lance Storm, Department of Psychology, Adelaide University, Adelaide, Australia; Suitbert Ertel, Georg-Elias-Müller Institute for Psychology, Georg August Universität, Göttingen, Germany.

Preparation of this article was supported, in part, by a grant from the Bial Foundation. Lance Storm thanks Michael Thalbourne for advice and helpful comments on drafts of this article and for access to the literature for the meta-analysis. Lance Storm also thanks Bob Willson for statistical advice.

Correspondence concerning this article should be addressed to Lance Storm, Department of Psychology, Adelaide University, Adelaide 5005, Australia. Electronic mail may be sent to lance.storm@psychology.adelaide.edu.au.

[1] Milton and Wiseman (1999) mistakenly reported the hit rate for these 10 studies as "35% ($p = .002$, one-tailed)" (p. 387). The only calculation made by Bem and Honorton (1994, p. 11) of a 35% hit rate was for 9 of the 10 studies "if Studies 104 and 105 are retained as separate studies"—they were originally split into two studies, "104/105(a) and 104/105(b)" (p. 10)—and further, only if Study 101 was excluded because it had a negative effect size ($\pi = .47$, $z = -.30$).

selected for their meta-analysis studies that "began in 1987" and "were published by February 1997" (Milton & Wiseman, 1999, p. 388). Studies already under way prior to 1987 were not included because it was assumed that investigators needed time to familiarize themselves with Hyman and Honorton's (1986) methodological guidelines (the assumption being that earlier studies would be too flawed for serious consideration in a meta-analysis). Thirty studies (including only 7 autoganzfeld studies) by "10 different principal authors from 7 laboratories" were deemed suitable for analysis (Milton & Wiseman, 1999, p. 388).

Milton and Wiseman (1999) calculated a Stouffer $Z$ of 0.70 ($p = .24$, one-tailed), with a mean effect size ($ES$) of 0.013 ($SD = 0.23$). They concluded that a significant psi effect for the ganzfeld had not been replicated by a "broader range of researchers" (Bem & Honorton, 1994, p. 13, and cited in Milton & Wiseman, 1999, p. 391).

## Problems With Milton and Wiseman's Meta-Analysis

### *Unwarranted Questioning of the Existence of Psi*

The title of Milton and Wiseman's (1999) article misrepresents a tenet held by the majority of active researchers in the parapsychological field, namely, that psi exists. Bem and Honorton (1994, p. 4) had good reasons to make use of the rhetorical question "Does psi exist?" because its implicit affirmation (psi does exist), suggested by their positive results, was merely the bottom line of a bulk of accumulated evidence (e.g., see Radin, 1997). In view of this state of affairs, one negative result by Milton and Wiseman (even if it were truly negative) could never remove, nor even touch, the evidence on which an acknowledgment of anomalous information transfer was already based. Milton and Wiseman's reiteration of Bem and Honorton's question, with its meaning reversed, merely furnishes cannon fodder for uninformed psi opponents who will take existential doubts, voiced by insiders, as sufficient grounds to reject the parapsychological subject matter altogether (see a first reaction to Milton & Wiseman's, 1999, article in CSICOP's periodical by Lilienfeld, 1999).

### *Spurious "Judgment Calls"*

Unfortunately, Milton and Wiseman (1999) exacerbated the damage further by making two rather spurious judgment calls. First, they did away with all ganzfeld research prior to Hyman and Honorton's (1986) communiqué, as if this publication—a mere documentation of traditional and uncontroversial research rules—could ever justify downgrading the quality of all research published before 1986 (more on this issue below). In a single stroke, Milton and Wiseman disqualified all pre-1986 studies. Negative results of meta-analyses based on deliberately curtailed databases give rise to beta error (i.e., false negative, or Type II, error). The authors excluded, without any inspection, Honorton's (1985) database of 28 studies. They also ignored that short (albeit productive) "middle" period of ganzfeld experimentation—1982–1986[2]—a period that produced studies that even Honorton had overlooked, as we discovered by searching the major parapsychology journals and other publications. We return to these (11) overlooked studies later.

Second, and surprisingly, Milton and Wiseman (1999) even expressed doubts about the quality of Bem and Honorton's (1994) meta-analysis, which comprised publications compliant with the communiqué's standards! They argued that the failure of Bem and Honorton's study to replicate might be due to (a) Bem and Honorton's "spurious" results, (b) "methodological artifacts" that may have arisen from "very weak sensory leakage," and (c) an "explicitly exploratory strategy of post hoc data selection or mislabeling of a nonsignificant effect" (Milton & Wiseman, 1999, p. 391). These statements suggest there might have been no psi effect at all, not only in Bem and Honorton's study, but also from time immemorial. We note, however, that the authors' boldness is easily rebuffed: "They [Milton & Wiseman] do not consider that their own deviation from B&H [Bem & Honorton, 1994]. . . . [which was] not significant, might be spurious" (S. Ertel, personal communication, October 17, 1999).

### *Negligence of Bidirectionality*

Milton and Wiseman (1999) also failed to look at the issue of bidirectional psi. They took an occurrence of hits above expectancy as the only criterion of value, which leads to the conclusion that their database indicates a decline when compared with Bem and Honorton's (1994) database. However, investigators have long recognized the fact that psi effects can manifest in polarized forms (i.e., as psi hitting and psi missing; see Nash, 1976; Rao, 1965). Timm (1983) advised psi researchers to routinely conduct bidirectional rather than unidirectional tests.

Although Milton and Wiseman's (1999) meta-analysis should have accommodated such a test, we do note that their oversight may have been due to the fact that the bidirectionality issue has not played much of a role in past telepathy research. The reason might be that in telepathy research, with its low numbers of trials per participant, reliable negative deviations, even though probably present for a minority of individual participants, cannot easily be identified. However, negative experimenter effects, ubiquitous in parapsychological research (Kennedy & Taddonio, 1976), might fully reverse telepathic communication effects for certain authors (significant psi missing instead of psi hitting). For some reason or other, quite a few such studies might have been included in Milton and Wiseman's sample. If so (see our analysis below), the presence of significant bidirectional psi effects would contradict Milton and Wiseman's claim that their results suggest "the ganzfeld paradigm cannot at present be seen as constituting strong evidence for psychic functioning" (p. 391).

### *Dynamic Versus Static Targets*

Milton and Wiseman's (1999) meta-analysis can be faulted in other ways. For example, the authors claimed that they were not able to replicate a significant difference between "dynamic" target hit rates and "static" target hit rates ($z = -0.95$, $p = .171$), as originally found by Bem and Honorton (1994). However, Milton and Wiseman's (p. 388) result is based on six studies—five by Broughton and Alexander (1996) and one by Morris, Cunningham, McAlpine, and Taylor (1993). Thus Milton and Wiseman considered only two principal authors in this specific analysis, which

[2] Studies published during the period February 1982 to 1985 were not picked up by Honorton (1985). The latest study in his meta-analysis was published in January 1982.

hardly constitutes a "broader range of researchers" (Milton & Wiseman, 1999, p. 391).

### *Ambience Levels and the Robustness of the Ganzfeld Paradigm*

Milton and Wiseman (1999, p. 390) noted that a warm social ambience should be created and maintained during the ganzfeld experiment, just as had been maintained by Honorton et al. (1990). A warm social ambience acts as a moderator variable in favor of a psi effect (Honorton et al., 1990). Only 20 of Milton and Wiseman's 30 studies (67%) actually reported an attempt to create this environment, whereas the remaining 10 studies (33%) must be given the benefit of the doubt. If a psi-conducive condition, well-known to research experts in this field, were really absent in 33% of the studies meta-analyzed by Milton and Wiseman, this would be grounds for serious concern. However, it is highly improbable, even though not inconceivable, that many of the earlier ganzfeld studies were flawed in this way. In any event, the fact that the ganzfeld effect came through as significant and positive in earlier meta-analyses, perhaps even despite poor social ambience, attests to the robustness of the ganzfeld effect.

Regarding robustness of the ganzfeld effect, Milton's opinion seems to waver (S. Ertel, personal communication, September 23, 1999). She noted that the similarity of mean effect sizes between two databases—Honorton's (1985) 28 studies and Bem and Honorton's (1994) 11 studies—"implies that the 'ganzfeld effect' is fairly robust." Ironically, in this statement, Milton defended the ganzfeld, as well as the two databases on which she later cast doubt. The issue of study quality is raised in more detail next.

## Meta-Analyses of Ganzfeld Studies

Our aim was to replace Milton and Wiseman's (1999) meta-analysis with an analysis based on a unified domain of ganzfeld data. Two principles guided us in this process: (a) Any meta-analysis intending to draw a general conclusion (in this case, the existence of psi) must take into account similar meta-analyses conducted earlier, unless there is evidence leading to the conclusion (or strong suspicion) that earlier such research was flawed (note the references above where the claims of alleged flaws in Honorton's, 1985, ganzfeld studies, and his meta-analyses, have not been successfully defended), and (b) two databases used for two consecutive and independent meta-analyses of studies testing the same hypothesis (in this case telepathy) using comparable procedures (in this case ganzfeld conditions) may be combined to form a larger database if the results of the two meta-analyses do not differ significantly.

We know that ganzfeld studies vary appreciably. One might object that visual and auditory, automatic and nonautomatic, precommuniqué and postcommuniqué ganzfeld studies should not be pooled—"apples and oranges" should not be combined, so to speak. However, it may be noted that meta-analysts do not reasonably hesitate to put apples and oranges in one basket if their hypotheses are about fruit (Glass, McGaw, & Smith, 1981, p. 218)! Milton and Wiseman (1999) acted accordingly when they compared a great majority of their own nonautoganzfeld collection (23 of 30 studies) with Bem and Honorton's (1994) sample of 10 autoganzfeld studies (which contains no other ganzfeld studies).

Our undertaking was not unprecedented (note, e.g., Radin & Ferrari's, 1991, meta-analysis in which over 50 years of dice-throwing studies were combined into one database; for other examples, see Utts, 1991). We believe that unification of databases is necessary if a researcher's goal is to draw general conclusions for larger domains because unification allows for the calculation of more accurate population parameters. As long as the databases remain ostensibly heterogeneous, the ganzfeld comes over as a fragmented experimental domain, whereas unification of the ganzfeld serves to strengthen the domain (cf. Honorton et al., 1990, pp. 127–128). The databases we considered are described in the following sections.

### *Precommuniqué Databases*

These databases were generated prior to Hyman and Honorton's (1986) guidelines and do not include autoganzfeld studies: (a) Honorton's (1985) database of 28 studies (referred to as *H1*) and (b) overlooked studies conducted between February 1982 and 1986, that is, ganzfeld studies not used by Honorton (1985; referred to as *S&E;* our selection procedure is described below).

Our first hypothesis was that H1 does not differ significantly from S&E. When combined (to anticipate the Results section, this turned out to be legitimate), H1 plus S&E is referred to as the *Old Ganzfeld Database.*

### *Postcommuniqué Databases*

Postcommuniqué databases were (a) Bem and Honorton's (1994) database of 10 studies (referred to as *H2a;* note that the *a* suffix signifies an exclusively autoganzfeld database) and (b) Milton and Wiseman's (1999) database of 30 studies (referred to as *M&W*), 7 studies using automatic procedures and 23 using nonautomatic procedures.

Our second hypothesis was that H2a does not differ significantly from M&W. When combined (this turned out to be legitimate), H2a plus M&W is referred to as the *New Ganzfeld Database,* which leads to our third hypothesis: The Old Ganzfeld Database does not differ significantly from the New Ganzfeld Database.

### *Rationale for Testing Hypotheses 1, 2, and 3*

The testing of the above hypotheses is not unprecedented and comes from the example set by Honorton et al. (1990, pp. 127–128), who found it a reasonable exercise to pool their 11 autoganzfeld studies with Honorton's (1985, Table A1, p. 84) 28 ganzfeld studies. Their original intention was "to assess the consistency of results" (p. 127) between the two databases. They used *t* tests and found that the *z* scores were similar between the two groups of studies, as were effect sizes (using an effect size measure called Cohen's *h*), and were then able to calculate "a better estimate of their true population values" (p. 127) by combining the two databases. A database of 39 studies resulted in a Cohen's *h* of .28 ($SD$ = .41) and a $Z$ of 7.53 ($p = 9.00 \times 10^{-14}$; Honorton et al., 1990, p. 99).

In using the independent-samples *t* test, where it is found that there are significant differences between samples, a critical level of variance in the dependent variable accounted for by the grouping

variable will be set at 9%[3] (Hays's, 1963, estimate of omega squared was used as the measure of effect size[4]). Thus, should omega squared fall below .09, the differences between the two samples on the tested variable are regarded as being of no importance and the two samples are combined.

### *Considering Bidirectionality*

As stated above, Milton and Wiseman (1999) took an occurrence of hits above expectancy as the only criterion of value. This might lead one to conclude that their database indicates a decline of psi effects when compared with Bem and Honorton's (1994) database and earlier databases (see Figure 1). However, as a mere inspection of Figure 1 already shows, positive *ES* values in Milton and Wiseman's sample are counterbalanced by a considerable number of negative *ES* values that are significant.

Our fourth hypothesis was that there is evidence of bidirectional psi in all major databases that will be combined following the above tests for database combination, including M&W's database in isolation. We use Timm's (1983, p. 222) Formula 5 for combining a data set of $z$ values to produce a chi-square value, which is then tested for significance.

## Criteria for Constructing the S&E Database

### *Selection Criteria*

The S&E database comprises previously overlooked studies conducted between February 1982 and 1986 (i.e., ganzfeld studies not used by Honorton, 1985, were used in the S&E database, provided they fulfilled the criteria below). The following selection criteria were used.

1. Only direct-hit data were used—the "most conservative" of measures used in "psi ganzfeld research" (Honorton, 1985, p. 54). Thus, (a) studies that used sums of ranks, but did not report the number of direct hits, were not used, and (b) studies reporting binary hits, but not the number of direct hits, were not used.
2. Where more than one hit rate was reported (e.g., as an alternative interpretation of the data, or from multiple judging, or purely as a post hoc construction), the lowest hit rate was used at all times to calculate $z$ scores and effect sizes. This criterion ensures a more conservative estimate of the overall effect size.
3. In studies where all participants were in the ganzfeld condition, but subgroups on different treatment regimes were not set up as controls of each other, the total number of trials and the total number of hits were calculated as single scores. This treatment produces more conservative $z$ scores and corresponding *ES* scores.
4. Studies published twice under alternative authorship were used only once, provided they met all other criteria.
5. Autoganzfeld studies were not used (because they qualify as postcommuniqué studies and are dealt with as such under Hypothesis 2).

### *Quality Criteria*

The following criteria were used to code studies. (Note that the overall quality may be higher than the weighted calculations suggest because some studies, such as abstracts, did not report every detail of the experiment. Thus, the overall calculations may also be conservative.) A criterion was either present (1 point) or absent (0 points), and each study was rated accordingly.

1. The study gave a prespecified methodology, including analyses specified in advance of experimental runs.
2. There was no optional stopping during the experiment.
3. Subject types were coded (based on Radin & Ferrari, 1991, pp. 65–66) as follows: (a) zero points = prior performance–special abilities, (b) ½ point = experimenter as sole participant, (c) ¾ of a point = experimenter and participants participated, and (d) 1 point = unselected participants.
4. Randomized targets were used.

The calculations of $z$ scores, effect sizes, weighted values, and "file-drawer" statistics can be found in the Appendix.

## Results

### *The S&E Database*

Our literature search yielded 11 studies (the S&E database) that fulfilled the above selection criteria (see Table 1). These studies, conducted during the period 1982 to 1986 (even though some were published after 1986), represented eight principal investigators (i.e., eight first authors).

The S&E database has an unweighted *ES* of 0.222 ($SD$ = 0.23) and a Stouffer $Z$ of 3.46 ($p = 2.70 \times 10^{-4}$). On the basis of Rosenthal's (1995, p. 189) file-drawer formula, there would have to be approximately 37 unpublished and nonsignificant studies in existence to reduce this significant finding to a chance result. However, the quality-weighted *ES* is 0.137 (±0.022 of a standard error) with a quality-weighted $Z$ of 1.06 ($p$ = .144; see the Appendix for calculations). (These more conservative results will be used in all relevant calculations involving this database.)

### *Hypothesis 1*

As far as *ES* values of individual studies were concerned, the performance comparison between H1 and S&E showed that the two databases were similar, $t(37) = 0.35$, $p$ = .729, two-tailed. Hypothesis 1 was confirmed. The *ES* for the Old Ganzfeld Database (a combined database) of 28 + 11 = 39 studies is .227 ($SD$ = .34) with a $Z$ of 6.15 ($p = 3.93 \times 10^{-10}$).

There would have to be approximately 507 unpublished, nonsignificant studies in existence to reduce this significant outcome to chance. There are 16 out of 39 studies in this database that have positive and significant $z$ scores (41%), which is well above a chance outcome. (See Table 2 for other results.)

### *Hypothesis 2*

The performance comparison between H2a and M&W did not reveal a significant difference, $t(38) = -1.88$, $p$ = .067, two-tailed. Although the probability value is also interpretable as marginally significant, it still allows us to take H2a and M&W as

[3] We set the critical value at 9% because it is equivalent to the coefficient of determination for a Pearson's product–moment correlation value of .30, which is recognized as the lowest correlation value of "importance" (B. Wilson, personal communication, February 21, 2001.)

[4] Hays (1963) recommended that the estimate of effect size omega squared accompany the result of a $t$ test. A significant $t$ value implies the existence of an association, but omega squared gives an estimate of how strong that association may be. When $t \leq 1$, $\omega^2 = 0$; when $t > 1$, estimated $\omega^2 = (t^2 - 1)/(t^2 + N_1 + N_2 - 1)$, where $N_j$ is the size of each sample.

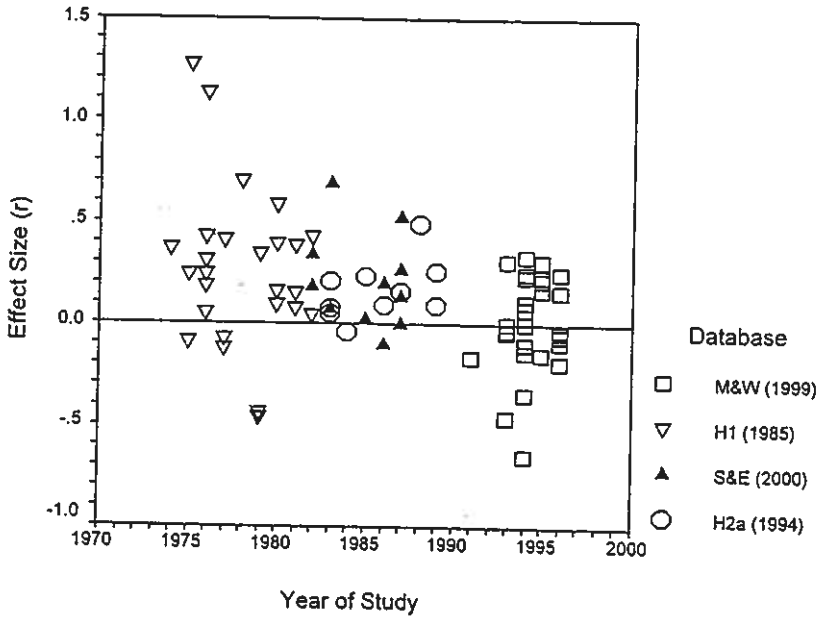


*Figure 1.* Scatter plot showing the distribution of effect sizes for the four databases: Milton and Wiseman (M&W; 1999), 30 studies; Honorton (H1; 1985), 28 studies; Storm and Ertel (S&E), 11 studies; and Bem and Honorton (H2a, 1994), 10 studies. Total: 79 studies.

subsamples of a larger sample. Hypothesis 2 was confirmed. The *ES* for the New Ganzfeld Database (a combined database) of 30 + 10 = 40 studies is .050 (*SD* = .22) with a *Z* of 1.88 ($p$ = .03). There would have to be 12 unpublished, nonsignificant studies in existence to reduce this significant outcome to chance. Eight of the 40 studies in this database have positive and significant *z* scores (20%), which is above a chance outcome. (See Table 3 for other results.)

Note that Hypotheses 1 and 2 would be confirmed again if effect sizes had been subjected to a bidirectional hypothesis by taking absolute values. For H1 and S&E, using absolute *ES* values, $t(37) = 1.12$, $p = .272$, and for H2a and M&W, using absolute *ES* values, $t(38) = 0.17$, $p = .866$.

### *Hypothesis 3*

The Old Ganzfeld Database and the New Ganzfeld Database were compared. The databases were significantly different, $t(77) = 3.15$, $p = .002$, $\omega^2 = .10$. Hypothesis 3, therefore, was not confirmed. Note that this omega squared value indicates that 10% of the variance in the independent variable (i.e., effect size) can be explained as due to the effects of the grouping variable (i.e., the database). This value exceeds our critical value by only 1%. Such a negligible difference might have been ignored, and combination of the two databases could have proceeded. Nevertheless, it was deemed inappropriate to combine the two databases on the basis of our a priori rule.

Table 1
*Number of Trials, z Scores, and Effect Sizes ($r = z/\sqrt{n}$) for the Second Precommuniqué Database (S&E)*

| Study | Trials ($n$) | $z$ | Effect size ($r$) |
|---|---|---|---|
| Bierman (1987) | 16 | 2.07* | 0.52 |
| Bierman et al. (1984) | 32 | 1.02 | 0.18 |
| Braud, Ackles, & Kyles (1984) | 10 | 2.19* | 0.69 |
| Haraldsson & Gissurarson (1985) | 70 | 0.28 | 0.03 |
| Houtkooper, Gissurarson, & Haraldsson (1988–1989) | 40 | 0.00 | 0.00 |
| Milton (1987) | 37 | 1.23 | 0.20 |
| Milton (1988–1989) | 35 | 1.58 | 0.27 |
| Murre et al. (1988) | 41 | 0.81 | 0.13 |
| Sargent (1982) | 20 | 0.79 | 0.18 |
| Sargent & Harley (1982) | 44 | 2.26* | 0.34 |
| Sondow (1987) | 60 | −0.75 | −0.10 |
| Total | 405 | 11.48 | 2.44 |

*Note.* *z* scores and effect sizes are calculated from direct-hit data only.
* $p < .05$.

Table 2
*Performance Comparison: Honorton's (1985) Database (H1) and Our Database (S&E)*

| No. of studies ($k$; database) | Sums of $ES$ ($\Sigma [z/\sqrt{n}]$) | Mean $ES$ ($\Sigma [z/\sqrt{n}]/k$) | $SD$ | Sums of $z$ scores ($\Sigma z$) | Mean $z$ score ($\Sigma z/k$) | Stouffer $Z$ ($\Sigma z/\sqrt{k}$) | $p$ |
|---|---|---|---|---|---|---|---|
| 28 (H1) | 7.35 | .264 | .374 | 34.93 | 1.247 | 6.60 | $2.10 \times 10^{-11}$ |
| 11 (S&E) | 1.51[a] | .137[b] | .229 | 03.50[c] | 0.318[d] | 1.06[c] | .144[c] |
| Total 39 | 8.86 | .227 | .337 | 38.43 | 0.986 | 6.15 | $3.93 \times 10^{-10}$ |

*Note.* $n$ = number of trials per study.
[a] Derived from the quality-weighted mean $ES$. [b] Quality-weighted mean $ES$. [c] Derived from the quality-weighted mean $z$ score. [d] Quality-weighted mean $z$ score.

### *Hypothesis 4*

We applied Timm's (1983, p. 222) Formula 5 ($\chi^2 = \Sigma [Z^2]$, and $df = k$, where $k$ is the number of studies) and obtained the following results for the databases in question: (a) Milton and Wiseman's (1999) database (M&W): $\chi^2(1, N = 30) = 46.66$, $p = .027$; (b) the Old Ganzfeld Database (H1 plus S&E): $\chi^2(1, N = 39) = 130.93$, $p = 7.30 \times 10^{-12}$; (c) the New Ganzfeld Database (M&W plus H2a): $\chi^2(1, N = 40) = 57.61$, $p = .035$; and (d) the Old plus New databases (79 studies): $\chi^2(1, N = 79) = 188.55$, $p = 6.05 \times 10^{-11}$.

Thus, within Milton and Wiseman's (1999) database there is an extreme dispersion of $z$ values not explainable by chance alone. This result and the results for the other three databases actually support the psi hypothesis.

## Post Hoc Analyses

### *A Reconsideration of the Effect Size Difference Between Databases*

Switching back to unidirectional modes of analysis (which was the approach taken by Bem & Honorton, 1994, and Milton & Wiseman, 1999), we have to concede that the mean $ES$ of the Old Ganzfeld Database is larger than the mean $ES$ of the New Ganzfeld Database. The difference is significant, and our above-mentioned principle for guiding the combination of databases, strictly applied, is an obstacle to combining these two databases. But post hoc we found more adaptive rules for database combinations in Cohen's (1988, p. 179) notion of a *critical effect size difference.* According to Cohen, the "differences between proportions [or differences between Cohen's $h$ values, or mean $ES$ values in our case] can be viewed in correlational terms" (p. 179). That is, we can view the $ES$ difference between our two databases as a relationship between two variables, and we decided to test it accordingly. We proposed that, should Cohen's test be successful, the two databases be combined (see the Appendix for the test procedure and the formulas for the calculations) despite some factual decline (as revealed by our result of Hypothesis 3), explanation of which would be another matter.

We therefore proposed a fifth hypothesis (a revised version of Hypothesis 3): that the observed effect size difference between the Old Ganzfeld Database and the New Ganzfeld Database does not differ significantly from the appropriate critical effect size difference. This hypothesis was tested by looking at the mean $ES$ data (as approximate measures of Cohen's $h$).[5]

The obtained $ES$ difference ($h_s = .090$) did not reach the critical $ES$ difference ($h_c = .372$) and was thus still small enough to treat the Old and the New databases as homogeneous and pool them. We defer to the Discussion section an explanation of the $t$ test based on the difference between the Old ($ES$ "high") database and the New ($ES$ "low") database.

The mean $ES$ of the unified Old and New database of 79 studies is 0.138 ($SD = 0.30$) with a Stouffer $Z$ of 5.66 ($p = 7.78 \times 10^{-9}$). There would have to be 857 unpublished, nonsignificant studies hidden away in file drawers to bring this highly significant result down to a chance outcome. Twenty-four of the 79 studies in this database have positive and significant $z$ scores (30%). (See Table 4 for other results.)

Two single-sample $t$ tests (testing mean $ES$ performance of studies within the database, one with studies as units, the other with authors as units) and an independent-samples $t$ test (testing for the difference of mean $ES$ performance of authors between the two databases) helped test our hypothesis further.

A single-sample $t$ test on effect sizes for the 79 individual studies was significant, $t(78) = 4.43$, $p < .001$, two-tailed. When the 29 authors represented in this database were used as units instead of studies, the single-sample $t$-test result was also significant, $t(28) = 2.92$, $p = .007$, two-tailed. Psi thus apparently exists in the unified sample, whichever way we tested it. On the other hand, one might surmise that the grand psi effect of the unified sample is merely due to some "outlier" studies by authors in the Old Ganzfeld Database. (When two single-sample $t$ tests were conducted on the New Ganzfeld Database only—one on study units, and one on author units—there were no significant $t$-test indications of psi in either case.)

The two databases were then compared to test possible effects of the guidelines on principal postcommuniqué authors. An independent-samples $t$ test, which compared Old authors with New authors,[6] was not significant, $t(29) = 1.92$, $p = .065$, $\omega^2 = .08$. The omega squared value indicates that only 8% of the variance was explained. In this case, we reject even the marginally significant difference that may be interpreted from the $t$-test result. This

[5] $ES$ and $h$ are effectively interchangeable ($r = .97$; see Honorton & Ferrari, 1989, p. 283).

[6] Note that two authors (Dick Bierman and Charles Honorton) conducted both pre- and postcommuniqué studies. They are thus represented in both the Old and New databases. The $t$ test on authors, however, is still valid because gauging the effects of the guidelines on Bierman, Berendsen, et al. (1984), Bierman, Bosga, et al. (1993), and Honorton's (1985; Honorton et al., 1990) ganzfeld practices, before and after the guidelines, is pertinent to our argument (hence, $df = 29$).

Table 3
*Performance Comparison: Bem and Honorton's (1994) Database (H2a) and Milton and Wiseman's (1999) Database (M&W)*

| No. of studies ($k$; database) | Sums of $ES$ ($\Sigma [z/\sqrt{n}]$) | Mean $ES$ ($\Sigma [z/\sqrt{n}]/k$) | $SD$ | Sums of $z$ scores ($\Sigma z$) | Mean $z$ score ($\Sigma z/k$) | Stouffer $Z$ ($\Sigma z/\sqrt{k}$) | $p$ |
|---|---|---|---|---|---|---|---|
| 10 (H2a) | 1.62 | .162 | .147 | 34.93 | 1.247 | 6.60 | $2.10 \times 10^{-11}$ |
| 30 (M&W) | 0.39 | .013 | .230 | 3.83 | 0.128 | 0.70 | $2.42 \times 10^{-1}$ |
| Total 40 | 2.01 | .050 | .224 | 11.89 | 0.297 | 1.88 | $3.00 \times 10^{-2}$ |

*Note.* $n$ = number of trials per study.

result suggests that the guidelines had no significant influence on effect size outcomes.

### *Population Parameters*

Having unified the ganzfeld research, by combining a total of 79 Old (ganzfeld) and New (ganzfeld and autoganzfeld) studies, we report the following population parameters for that database. Thirty-nine studies predate the Hyman and Honorton (1986) guidelines, and 40 postdate those guidelines. Thirty-two of 39 Old studies had positive $z$ scores (82%), and 25 of 40 New studies had positive $z$ scores (63%). Overall, 57 studies had positive $z$ scores (72%) of the 79 studies (39 + 40 = 79). The 95% confidence interval (CI) is from 62% to 82%.

The $z$ scores range from −2.30 to 4.02 (mean $z$ = 0.64, $SD$ = 1.37, CI = 0.43 to 1.04), and effect sizes range from −.65 to 1.27 (mean $ES$ = 0.14, $SD$ = 0.30, CI = 0.08 to 0.22). The overall hit rate is 31% (CI = 29% to 35%).

## Discussion

This study scrutinized and amended Milton and Wiseman's (1999) meta-analysis of ganzfeld studies. Milton and Wiseman selected a database of 30 such studies covering the period from 1987 to early 1997. For this particular database, they did not obtain significant hit rates and concluded that Bem and Honorton's (1994) "anomalous process of information transfer" (p. 387) might not exist.

In a number of ways, however, Milton and Wiseman (1999) were lacking in caution. Their conclusion implies that their own selection of 30 studies was superior to previous databases. Thus, they disregarded all pertinent research conducted prior to 1987. Not only was Honorton's (1985) database overlooked but so were other studies preceding the Hyman and Honorton (1986) guidelines. In addition, Milton and Wiseman even contrived possible deficiencies in the procedures applied in the studies included in Bem and Honorton's (1994) database (such as "weak sensory leakage," "post-hoc data selection," "mislabeling of a nonsignificant effect," etc.; Milton & Wiseman, 1999, p. 391).

Ironically, Bem and Honorton's (1994) database is still the only database that Milton and Wiseman (1999) regarded as worthy of a replication trial on the basis of its positive results. Yet they never combined their own database with, say, the methodologically better studies in Honorton's (1985) and Bem and Honorton's databases. Not even a pure autoganzfeld sample of 17 studies was formed, all postdating the guidelines[7]—a point in line even with Milton and Wiseman's misleading restrictions.

Finally, Milton and Wiseman's (1999) ignorance of the fact that the nature of psi has often been shown to be bidirectional caused them to overlook this feature in their database. Psi-missing studies are apparently less frequent in their sample than those with hit-rate deviations, as is always the case whenever psi missing is observed, and we can only speculate why they appeared in Milton and Wiseman's sample so frequently compared with earlier databases. Overly strict controls, automation and other procedural changes, and an increased concern of experimenters about success might elicit subconscious inhibition and corresponding effect reversals. Suffice it to conclude, the Milton and Wiseman database did exhibit bidirectional psi. A note of caution: Psi-missing results (see Figure 1) are placed on the negative side of the zero-axis, thus giving the impression that Milton and Wiseman's data indicate a sudden psi decline. Conclusions on psi decline, however, must always be based on both positive and negative effect directions.

Milton and Wiseman will probably explain our significant mean effect size difference between the Old and the New databases by referring to the Hyman and Honorton (1986) guidelines. Regarding researchers in the Old period, Milton and Wiseman might claim that their evidence of psi in the ganzfeld was actually artifactual, whereas researchers in the New period, having been warned by the guidelines, conducted flawless ganzfeld experiments, so their $ES$ values dropped. We would reply, however, that even if $ES$ values had dropped, because of excluding all possible sources of artifact (sensory leakage, etc.), Milton and Wiseman's (1999) database—alone or in combination with H2a—did manifest psi-typical bidirectional effects.

Moreover, the observed $ES$ decrease in the databases over the course of the two periods cannot safely be explained by invoking, for the Old period, the presence of artifactual sources for deviations from chance and, for the New period, the absence of such sources—after all, because decline effects in psi research existed long before the guidelines, it cannot be assumed that the guidelines are single-handedly responsible for declines in the ganzfeld domain. And we already referred above to the well-known psi-inhibiting effects that result from the introduction of stricter controls, automation, and so on. Such changes are on the increase, possibly because of the guidelines, which might have led to psi-typical effect reversals. In addition, a general fading of psi effects altogether might have occurred recently—temporary declines are almost no less atypical for "anomalous communication" than are, say, declines of ionospheric shortwave reflections for radio communication.

[7] An independent-samples $t$ test was performed on H2a and M&Wa (7 autoganzfeld studies only), but it failed to give a significant result, $t(16) = 1.23$, $p = .235$, two-tailed, equal variances assumed. The 17 autoganzfeld studies were combined: $ES = .117$ ($SD = .17$), $Z = 2.67$ ($p = 3.79 \times 10^{-3}$).

Table 4
*Performance Comparison: The Old Ganzfeld Database (H1 + S&E) and the New Ganzfeld Database (H2a + M&W)*

| No. of studies ($k$; database) | Sums of $ES$ ($\Sigma [z/\sqrt{n}]$) | Mean $ES$ ($\Sigma [z/\sqrt{n}]/k$) | $SD$ | Sums of $z$ scores ($\Sigma z$) | Mean $z$ score ($\Sigma z/k$) | Stouffer $Z$ ($\Sigma z/\sqrt{k}$) | $p$ |
|---|---|---|---|---|---|---|---|
| 39 (H1 + S&E) | 8.86 | .227 | .337 | 38.43 | 0.986 | 6.15 | $3.93 \times 10^{-10}$ |
| 40 (H2a + M&W) | 2.01 | .050 | .224 | 11.89 | 0.297 | 1.88 | $3.00 \times 10^{-2}$ |
| Total 79 | 10.87 | .138 | .301 | 50.32 | 0.637 | 5.66 | $7.58 \times 10^{-9}$ |

*Note.* $n$ = number of trials per study; H1 = Honorton's (1985) database; S&E = our database; H2a = Bem and Honorton's (1994) database; M&W = Milton and Wiseman's (1999) database.

Milton and Wiseman might further object that they should not be blamed for following the example of Bem and Honorton (1994), who also meta-analyzed ganzfeld studies but excluded the H1 database (Bem and Honorton, however, did not give reasons for that exclusion). Thus Milton and Wiseman's (1999) meta-analysis considered the H2a database only, justifying in their minds the questions: "Why should we not also exclude H1 data, and why should we not make conclusions about the ganzfeld based on our (M&W) database only?"

Our response is, first, that Bem and Honorton (1994) merely wanted to demonstrate to skeptical observers that "existence of psi" (p. 4) even if, as their overly skeptical opponents claimed, the H1 database was doubtful. Bem and Honorton expected and found psi effects in postcommuniqué data alone. Milton and Wiseman (1999) acted differently: They did not test the psi hypothesis by using all trustworthy postcommuniqué (i.e., all New) data; rather, they excluded all Bem and Honorton (H2a) data, even though their trustworthiness cannot be challenged (except by arguments from out of the blue; see our critique above).

Second, an extremely skeptical approach to the ganzfeld can be taken by conducting a meta-analysis with quite deliberate concessions, that is, by including in a larger updated database, aside from all "safe" postcommuniqué data, only "safe" studies from the H1 database. The H1 database would be rendered safe by excluding the 11 highest ranking $ES$ studies (of the total of 28) such that the mean $ES$ of the modified H1 database (absolute $ES$) matches the mean $ES$ (absolute $ES$) of all postcommuniqué data. A single-sample $t$ test on this truncated database (17 studies) is still significant if based on the unidirectional notion, $t(16) = 3.98$, $p = .001$, and is also significant if based on the bidirectional notion, $\chi^2(1, N = 17) = 28.64$, $p = .038$. Significant results are also obtained for the truncated 79-study database of 57 studies (excludes S&E), based on the unidirectional notion, $t(56) = 2.90$, $p = .005$, and the bidirectional notion, $\chi^2(1, N = 57) = 86.26$, $p = 7.43 \times 10^{-3}$. (If we included the overlooked studies that form the S&E database, the $t$ and chi-square values would be even higher.)

### *Future Directions for Meta-Analysis in Psi Research*

In our meta-analysis we have identified some salient points that meta-analysts need to consider: (a) Meta-analyses must be based on all available evidence, not on a deliberate narrowing down of data pools; (b) any exclusion of available studies must be justified on empirical grounds; and (c) scattered data must be combined provided they prove to be effectively drawn from the same data population.

In following these points, we compiled three comprehensive databases. We found that they supported the psi hypothesis. The biggest database (79 studies) contains both ganzfeld and autoganzfeld studies (see Table 4). Investigators, when planning future research in the ganzfeld domain, may refer to our tabulated data for determining effect size norms and other expected outcomes.

Our finding that there is homogeneity between seemingly disparate databases may not convince everyone. Others might prefer more restrictive criteria in selecting studies according to their purposes. So long as these tests are unbiased, all combinations of previous research will be more comprehensive and therefore more valid than Milton and Wiseman's (1999) minimal selection. Whatever the selection criteria may be, they must be derived from observation and not be diminished by (or disguised by) judgments of faith that ignore pertinent empirical information. Arbitrary data exclusions (to which we took exception) will hardly find general approval.

### *Conclusion*

Milton and Wiseman's (1999) generalized dismissal of the ganzfeld procedure, which purportedly "does not . . . offer a replicable method for producing ESP in the laboratory" (p. 387), is hardly warranted. Their message appears partial and dangerous:

> It is the message of [i.e., the implicit answer to] their paper's title ["Does Psi Exist? . . ."] that will probably be transmitted around the globe and . . . will have some impact on the world's public opinion, including opinion in scientific circles about the entire parapsychological enterprise. (S. Ertel, personal communication, October 17, 1999)

Twenty-five years of ganzfeld–autoganzfeld work suggest an "anomalous effect in need of an explanation" (Utts, 1991, p. 363). The domain might continue to be an ideal paradigm for pointing the skeptic in the direction of a more positive answer to that often asked question: "Does psi exist?"

A number of ganzfeld experiments have been published since Milton and Wiseman (1999) completed their search (Carpenter, 1999), and accordingly further meta-analyses have been done as well (see Milton, 1999; Palmer & Broughton, 2000; Storm, 2000). These studies hold the key to the ganzfeld's future. It is too early at this stage to anticipate all possible eventualities, but it might turn out that ganzfeld effects keep appearing with lower effect sizes—declines of psi effects are common in laboratories, as are declines of certain drug effects by routine use. However, one statement that we can make with confidence is that results of ganzfeld experiments will keep on being prolific enough to "prompt others to try replicating the psi . . . effect" (Bem & Honorton, 1994, p. 13).

## References

References marked with an asterisk indicate studies included in the meta-analysis.


Atkinson, R. L., Atkinson, R. C., Smith, E. E., & Bem, D. J. (1990). *Introduction to psychology* (10th ed.). New York: Harcourt Brace Jovanovich.

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

* Bierman, D. J. (1987). A test on possible implications of the OT's for ganzfeld research. *European Journal of Parapsychology, 7,* 1–10.

* Bierman, D. J., Berendsen, J., Koenen, C., Kuipers, C., Louman, J., & Maisson, F. (1984). The effects of ganzfeld stimulation and feedback in a clairvoyance task. In R. A. White & R. S. Broughton (Eds.), *Research in parapsychology 1983* (p. 14). Metuchen, NJ: Scarecrow.

Bierman, D. J., Bosga, D. J., Gerding, H., & Wezelman, R. (1993). Anomalous information access in the ganzfeld: Utrecht—Novice Series I and II. In N. L. Zingrone, M. J. Schlitz, C. S. Alvorado, & J. Milton (Eds.), *The Parapsychological Association 36th Annual Convention: Proceedings of presented papers* (pp. 192–203). Durham, NC: Parapsychological Association.

* Braud, L. W., Ackles, L., & Kyles, W. (1984). Free-response GESP performance during ganzfeld stimulation. In R. A. White & R. S. Broughton (Eds.), *Research in parapsychology 1983* (pp. 78–80). Metuchen, NJ: Scarecrow.

Broughton, R. S., & Alexander, C. H. (1996). Autoganzfeld: II. In E. May (Ed.), An attempted replication of the PRL ganzfeld research. In *The Parapsychological Association 39th Annual Convention: Proceedings of presented papers* (pp. 45–56). Durham, NC: Parapsychological Association.

Carpenter, S. (1999). ESP findings send controversial message. *Science News, 156,* 70.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

* Haraldsson, E., & Gissurarson, L. R. (1985). Perceptual defensiveness, ganzfeld and the percipient-order effect. *European Journal of Parapsychology, 6,* 1–17.

Harris, M. J., & Rosenthal, R. (1988a). *Interpersonal expectancy effects and human performance research.* Washington, DC: National Academy Press.

Harris, M. J., & Rosenthal, R. (1988b). *Postscript to interpersonal expectancy effects and human performance research.* Washington, DC: National Academy Press.

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

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

Honorton, C., Berger, R. E., Varvoglis, M. P., Quant, 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.

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

* Houtkooper, J. M., Gissurarson, L. R., & Haraldsson, E. (1988–1989). Why the ganzfeld is conducive to ESP: A study of observational theory and the percipient-order effect. *European Journal of Parapsychology, 7,* 169–192.

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.

Kennedy, J. E., & Taddonio, J. L. (1976). Experimenter effects in parapsychological research. *Journal of Parapsychology, 20,* 1–33.

Lilienfeld, S. E. (1999). Research review: New analyses raise doubts about replicability of ESP findings. *Skeptical Inquirer, 23,* 9, 12.

* Milton, J. (1987). Judging strategies to improve scoring in the ganzfeld. In D. H. Weiner & R. D. Nelson (Eds.), *Research in parapsychology 1986* (pp. 100–103). Metuchen, NJ: Scarecrow.

* Milton, J. (1988–1989). A possible 'directive' role of the agent in the ganzfeld. *European Journal of Parapsychology, 7,* 193–214.

Milton, J. (1999). Should ganzfeld research continue to be crucial in the search for a replicable psi effect? Part I. Discussion paper and introduction to an electronic mail discussion. *Journal of Parapsychology, 63,* 309–333.

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

Morris, R. L., Cunningham, S., McAlpine, S., & Taylor, R. (1993). In N. L. Zingrone, M. J. Schlitz, C. S. Alvorado, & J. Milton (Eds.), Toward replication and extension of autoganzfeld results. In *The Parapsychological Association 36th Annual Convention: Proceedings of presented papers* (pp. 177–191). Durham, NC: Parapsychological Association.

* Murre, J. M. J., van Dalen, A. C., Dias, L. R. B., & Schouten, S. A. (1988). A ganzfeld psi experiment with a control condition. *Journal of Parapsychology, 52,* 103–125.

Nash, C. B. (1976). Group selection of target painting. *European Journal of Parapsychology, 1,* 37–39.

Palmer, J., & Broughton, R. S. (2000). An updated meta-analysis of post-PRL ESP-ganzfeld experiments: The effect of standardness. In *Proceedings of the 43rd Annual Convention of the Parapsychological Association* (pp. 224–240). Durham, NC: Parapsychological Association.

Radin, D. I. (1997). *The conscious universe: The scientific truth of psychic phenomena.* San Francisco: HarperCollins.

Radin, D. I., & Ferrari, D. C. (1991). Effects of consciousness on the fall of dice: A meta-analysis. *Journal of Scientific Exploration, 5,* 61–83.

Rao, K. R. (1965). The bidirectionality of psi. *Journal of Parapsychology, 29,* 230–250.

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

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118,* 183–192.

* Sargent, C. L. (1982). A ganzfeld GESP experiment with visiting subjects. *Journal of the Society for Psychical Research, 51,* 222–232.

* Sargent, C. L., & Harley, T. A. (1982). Precognition testing with free-response techniques in the ganzfeld and the dream state. *European Journal of Parapsychology, 4,* 243–256.

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

* Sondow, N. (1987). Exploring hypnotizability, creativity, and psi: Conscious and unconscious components to psi success in the ganzfeld. In D. H. Weiner & R. D. Nelson (Eds.), *Research in parapsychology 1986* (pp. 42–47). Metuchen, NJ: Scarecrow.

Storm, L. (2000). Replicable evidence of psi: A revision of Milton's (1999) meta-analysis of the ganzfeld databases. *Journal of Parapsychology, 64,* 411–416.

Timm, U. (1983). Statistische selektionsfehler in der parapsychologie und in anderen empirischen wissenschaften [Statistical selection errors in parapsychology and other empirical sciences]. *Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie, 25,* 195–230.

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

# Appendix

## Calculation of Statistics Used in the Analysis

### Calculation of Z Scores and Effect Sizes

If a $z$ score was not provided in a study, it was calculated from the hit rate by using the exact binomial test, which compares the number of hits obtained to the expected number of hits (i.e., $P_{MCE} = .25$). Effect size ($ES$ or $r$) calculations were made on Honorton's (1985) data using $z/\sqrt{n}$ (otherwise presented as $z/N^{1/2}$), where $n$ (or $N$) is the number of trials in each study (in the present study $k$ refers to the number of studies).

There are two reasons for using $z/\sqrt{n}$. First, Honorton and Ferrari (1989, p. 283) used it instead of Cohen's $h$ (as used by Honorton, 1985, Table A1, p. 84) because it is easier to work with. Second, Milton and Wiseman (1999, p. 388) also used it in their study, so for the sake of consistency the same formula is used here.

### Weighted Values

The quality-weighted $z$ was calculated using Rosenthal's (1984, p. 89) Formula 4.31: quality-weighted $z = \Sigma w_j z_j / [\Sigma w_j^2]^{1/2}$, where $w$ is the weight, and $j$ ranges from 1 to $k$.

The quality-weighted mean $ES$ was calculated using $ES = \Sigma w_j n_j r_j / \Sigma (w_j n_j)$, where $w$ is the weight, $n$ is the number of trials, $r$ is the unweighted effect size given by the $z$ score (see Table 1), and $j$ ranges from 1 to $k$. The standard error associated with this quality-weighted $ES$ is $[\Sigma (w_j^2 n_j)/(\Sigma (w_j n_j)^2]^{1/2}$ (these formulas are derived from Radin & Ferrari, 1991, p. 65).

### "File-Drawer" Statistics

The formula given by Rosenthal (1995, p. 189), $X = [(\Sigma Z)^2/2.706] - k$, was used to calculate estimates of the number of studies averaging null results needed to reduce significant probability values to chance values (i.e., $p = .05$). The $\Sigma Z$ value (i.e., sum of the standard normal deviates) can be found in the "Sums of $z$ scores ($\Sigma z$)" columns of the respective tables (see Tables 2, 3, and 4). The $k$ value refers to the number of studies actually retrieved for meta-analysis.

### Cohen's h Test

The first stage of the significance test uses Cohen's (1988, p. 200) Formula 6.3.1: $n' = (2n_1n_2)/(n_1 + n_2)$, where $n'$ is the harmonic mean of the two samples $n_1$ and $n_2$, which are the sample sizes for the two different databases (neither $n$ should be very small, i.e., <10). Thus, in the test, $n_1 = 39$, $n_2 = 40$, $n' = 39.49$.

The $ES$ difference ($h_\wedge$) between the two samples is found, and by locating the harmonic mean ($n'$) for the two samples, using Cohen's (1988, p. 194) Table 6.3.2, the viability of $h_\wedge$ is assessed by comparing it against a critical $ES$ difference ($h_c$). If $h_\wedge \geq h_c$, then the observed $ES$ difference between the two databases is significantly large and, therefore, not explainable by chance alone (in this case at the .05 level, directional), suggesting that there is a significant difference between the mean effect sizes of the two databases. The $h_c$ for the particular comparison is the minimal significant difference between mean effect size scores of the two databases that suggests the difference is not a chance occurrence but is due to some variable(s) not present in one or the other database.

Received November 5, 1999
Revision received June 23, 2000
Accepted October 26, 2000 ■