Journal of Parapsychology, Vol. 50, December 1986

META-ANALYTIC PROCEDURES AND THE NATURE OF REPLICATION: THE GANZFELD DEBATE

BY ROBERT ROSENTHAL

ABSTRACT: This paper is a commentary on the valuable debate between Charles Honorton (1985) and Ray Hyman (1985) about the evidence for psi in the ganzfeld situation. Their debate was a creative, constructive, and task-oriented dialogue that served admirably to sharpen the issues involved. In my commentary I focus on the concept of replication, distinguishing the troublesome older view with a more useful alternative. Specific issues related to replication are discussed including problems of multiple testing, subdividing studies, weighting replications, and problems of small effects. The earlier meta-analytic work is summarized, evaluated, and compared with a meta-analysis of a different controversial area. Rival hypotheses of procedural and statistical types are discussed, and a tentative inference is offered. The conclusion calls for wider use of newer views of the success of replication.

Science in general and parapsychological inquiry in particular have been well served by the recent ganzfeld debate between Charles Honorton (1985) and Ray Hyman (1985) as organized by the *Journal*'s editor, K. Ramakrishna Rao. Two serious and highly knowledgeable scholars have invested a great amount of time, energy, and creative thought to produce a debate that is a model of task-oriented, constructive dialogue. It is clear that the participants have been devoted to clarifying and understanding the scientific issues rather than simply to "scoring points."

As a result of their efforts we have an excellent review of the issues to be considered in evaluating the data generated by the ganzfeld experiments. In addition, through their meta-analytic work, we have an enormously valuable quantitative summary of the ganzfeld studies. In the end, Hyman and Honorton have not resolved all their differences, nor is it likely that they will. Hyman has raised cogent and telling questions. Honorton has answered them in cogent and telling terms. I am sure that Hyman will have excellent

The preparation of this paper and the development of some of the procedures described within it were supported by the National Science Foundation. Much of the summary and interpretation of the meta-analyses will be included in a paper commissioned by the National Academy of Sciences that is in preparation by Monica J. Harris and Robert Rosenthal.

TABLE 1
COMMON MODEL OF REPLICABILITY: JUDGMENT IS DICHOTOMOUS AND BASED ON SIGNIFICANCE TESTING

		First study	
		$p > .05$[a]	$p < .05$
Second study	$p < .05$[b]	A. Falure to replicate	B. Successful replication
	$p > .05$	C. Failure to establish effect	D. Failure to replicate

[a] By convention .05 but could be any other given level, e.g., .01.
[b] In the same tail as the results of the first study.

replies to most of Honorton's rebuttals and that, in turn, Honorton will have excellent replies to most of Hyman's replies. Readers of their debate will tend to favor one position or the other. One hopes these tendencies will be based more on the issues raised by Honorton and Hyman and on the data they have arrayed than on the a priori grounds feared by open-minded "zetetic" scholars such as Marcello Truzzi (1981).

I am grateful to Editor Rao for the opportunity to comment on this lively debate. In his introduction to this dialogue, Dr. Rao (1985) saw very precisely the need to clarify the concept of replication, and it is with a consideration of this concept that I begin.

THE CONCEPT OF REPLICATION

The issue of successful replication is central not only to the ganzfeld debate but more broadly to the field of parapsychology, and still more broadly to the entire field of psychology. However, there is conceptual confusion over the actual meaning of replicability. Successful replication is ordinarily taken to mean that a null hypothesis that has been rejected at time 1 is rejected again, and with the same direction of outcome, on the basis of a new study at time 2. The basic model of this usage can be seen in Table 1. The results of the first study are described dichotomously as $p < .05$ or $p > .05$ (or some other critical level, e.g., .01). Each of these two possible outcomes is further dichotomized according to the results of the second study as $p < .05$ or $p > .05$. Thus, cells A and D of Table 1 are examples of failure to replicate because one study was significant

TABLE 2
ILLUSTRATIVE RESULTS OF AN EXPERIMENT IN PARAPSYCHOLOGY

	Investigator	
	Smith	Jones
Treatment mean	.38	.36
Control mean	.26	.24
Difference	.12	.12
t	2.21	1.06
df	78	18
Two-tailed p	.03	.30
Effect size d[a]	.50	.50
Effect size r[b]	.24	.24
Standard normal z	2.17[c]	1.03[c]

[a]Obtained from $2t/\sqrt{df}$ (Rosenthal, 1984).

[b]Obtained from $\sqrt{t^2/(t^2 + df)}$ (Rosenthal, 1984).

[c]These significance levels differ at $z = .81$, $p = .42$ [from $(z_1 - z_2)/\sqrt{2}$ (Rosenthal, 1984)].

and the other was not. Let us examine more closely a specific example of such a "failure to replicate."

Pseudo-Failures to Replicate

Smith has published the results of an experiment in parapsychology in which a certain treatment procedure was predicted to increase psi performance. She has reported results significant at $p < .05$ in the predicted direction. Jones publishes a rebuttal to Smith claiming a failure to replicate.

Table 2 shows the results of these two experiments in greater detail. Smith's results were more significant than Jones's, to be sure, but the studies were in perfect agreement about their estimated sizes of effect as defined either by Cohen's d [$(\text{Mean}_1 - \text{Mean}_2)/\sigma$] or by r, the correlation between group membership and psi performance score (Cohen, 1977; Rosenthal, 1984). Not only did the effect sizes of the two studies agree but also even the significance levels of .03 and .30 did not differ very significantly, $(z_{.03} - z_{.30})/\sqrt{2} = (2.17 - 1.03)/\sqrt{2} = z = .81$, $p = .42$; for details on the comparison of significance levels and effect sizes, see Rosenthal and Rubin (1979; 1982a) or a summary in Rosenthal (1984). Table 2 shows very clearly that Jones was very much in error when he claimed that his

study failed to replicate that of Smith. Such errors are made very frequently in most areas of psychology and the other behavioral sciences.

Pseudo-Successful Replications

Return now to Table 1 and focus attention on cell B, the cell of "successful replication." Suppose that two investigators both rejected the null hypothesis at $p < .05$ with both results in the same direction. Suppose further, however, that in one study the effect size r was .90 whereas in the other study the effect size r was only .10, significantly smaller than the r of .90 (Rosenthal & Rubin, 1982a). In this case our interpretation is more complex. We have indeed had a successful replication of the rejection of the null, but we have not come even close to a successful replication of the effect size.

"Successful Replication" of Type II Error

Cell C of Table 1 represents the situation in which both studies failed to reject the null hypothesis. Under those conditions investigators might conclude that there was no relationship between the variables investigated. Such a conclusion could be very much in error, the more so the lower the power of the two studies was low (Cohen, 1977). If power levels of the two studies (assuming medium effect sizes in the population) were very high, say .90 or .95, then two failures to obtain a significant relationship would provide evidence that the effect investigated was not likely to be a very large effect. If power calculations had been made assuming a very small effect size, two failures to reject the null although not providing strong evidence for the null would at least suggest that the size of the effect in the population was probably quite modest.

If sample sizes of the two studies failing to reject the null were modest so that power to detect all but the largest effects were low, very little could be concluded from two failures to reject except that the effect sizes were unlikely to be enormous. For example, two investigators with N's of 20 and 40, respectively, find results not significant at $p < .05$. The effect sizes *phi* (i.e., r for dichotomous variables) were .29 and .20, respectively, and both p's were approximately .20. The combined p of these two results, however, is $.035[(z_1 + z_2)/\sqrt{2} = z]$, and the mean effect size in the mid-.20's is not trivial (Rosenthal & Rubin, 1982b).

TABLE 3
COMPARISON OF TWO SETS OF REPLICATIONS

	Replication sets			
	A		B	
	Study 1	Study 2	Study 1	Study 2
N	96	15	98	27
p (two-tailed)	.05	.05	.01	.18
z (p)	1.96	1.96	2.58	1.34
r	.20	.50	.26	.26
z(r)	.20	.55	.27	.27
Cohen's q ($z_{r_1} - z_{r_2}$)	.35		.00	

Comparing Views of Replication

The traditional, not very useful, view of replication modeled in Table 1 has two primary characteristics:

1. It focuses on significance level as the relevant summary statistic of a study.

2. It makes its evaluation of whether replication has been successful in a dichotomous fashion. For example, replications are successful if both or neither $p < .05$ (or .01, etc.), and they are unsuccessful if one $p < .05$ (or .01, etc.) and the other $p > .05$ (or .01, etc.). Psychologists' reliance on a dichotomous decision procedure accompanied by an untenable discontinuity of credibility in results varying in p levels has been well documented (Nelson, Rosenthal, & Rosnow, 1986; Rosenthal & Gaito, 1963, 1964).

The newer, more useful view of replication success has two primary characteristics:

1. It focuses on effect size as the more important summary statistic of a study with only a relatively minor interest in the statistical significance level.

2. It makes its evaluation of whether replication has been successful in a continuous fashion. For example, two studies are not said to be successful or unsuccessful replicates of each other but, rather, the degree of failure to replicate is specified.

Table 3 shows two sets of replications. Replication set A shows two results both rejecting the null but with a difference in effect sizes of .30 in units of r or .35 in units of Fisher's z transformation of r (Cohen, 1977; Rosenthal & Rosnow, 1984; Snedecor & Cochran, 1980). That difference, in units of r or Fisher's z is the degree

of failure to replicate. That both studies were able to reject the null and at exactly the same p level is simply a function of sample size. Replication set B shows two studies with different p values, one significant at $< .05$, the other not significant. However, the two effect size estimates are in excellent agreement. We would say, accordingly, that replication set B shows more successful replication than does replication set A.

It should be noted that the values of Table 3 were chosen so that the combined probability of the two studies of set A would be identical to the combined probability of the two studies of set B; $(z_1 + z_2)/\sqrt{2} = z$ of 2.77, $p = .0028$, one-tailed.

The Metrics of the Success of Replication

Once we adopt a view of the success of replication as a function of similarity of effect sizes obtained, we can become more precise in our assessments of the success of replication. Figure 1 shows the "replication plane" generated by crossing the results of the first study conducted (expressed in units of the effect size r) by the results of the second study conducted. All perfect replications, those in which the effect sizes are identical in the two studies, fall on a diagonal rising from the lower left corner $(-1.00, -1.00)$ to the upper right corner $(+1.00, +1.00)$. The results of replication set B from Table 3 are shown to fall exactly on the diagonal of successful replication $(+.26, +.26)$. The results of replication set A are shown to fall somewhat above the line representing perfect replication. Figure 1 shows that although set B reflects a more successful replication than set A, the latter is also located fairly close to the line and is, therefore, a fairly successful replication set as well.

Cohen's q. An alternative to the indexing of the success of replication by the difference between obtained effect size r's is to transform the r's to Fisher's z's before taking the difference. Fisher's z metric is distributed nearly normally and can thus be used in setting confidence intervals and testing hypotheses about r's, whereas r's distribution is skewed, and the more so as the population value of r moves further from zero. Cohen's q is especially useful for testing the significance of difference between two obtained effect size r's. This is accomplished by means of the fact that

$$q \Big/ \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

is distributed as z, the standard normal deviate (Rosenthal, 1984;

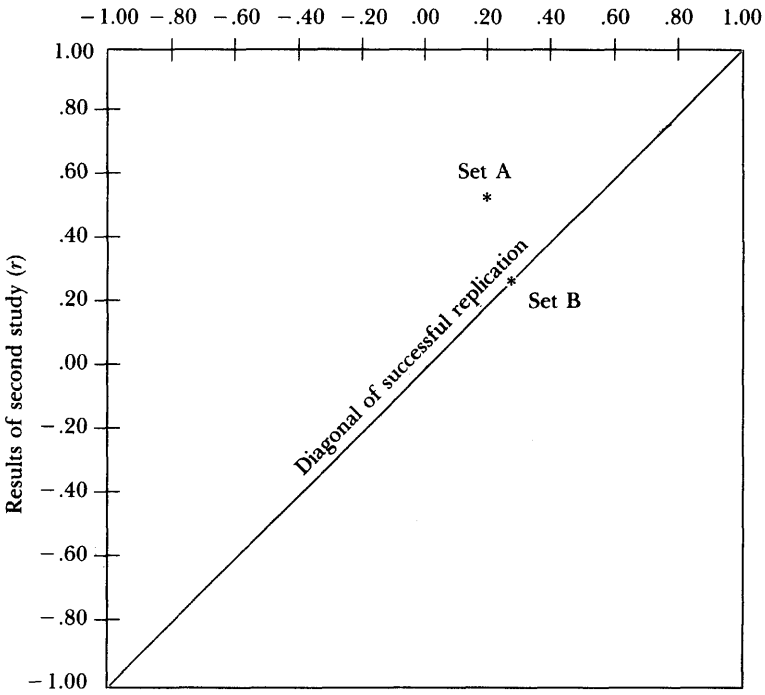

Figure 1. The replication plane.

Rosenthal & Rubin, 1982a; Snedecor & Cochran, 1980). When there are more than two effect size *r*'s to be evaluated for their variability (i.e., heterogeneity), the three references above all provide the appropriate formula for computing the test of the heterogeneity of *r*'s.

Issues Related to Replication

Multiple Testing

In ganzfeld studies, in parapsychological research more broadly, and, indeed, in most areas of behavioral science, it is common that more than one test of significance is computed to evaluate a research hypothesis. There may, for example, be a set of several dependent variables used to evaluate outcome. So long as there are

multiple questions, multiple dependent variables make good scientific sense. However, as both Honorton (1985) and Hyman (1985) point out, the use of multiple dependent variables may affect the accuracy of the p levels computed. For example, if five dependent variables are used and one of these is found to show an effect at $p < .05$, it would be misleading to say that an effect has been demonstrated at $p < .05$. That is because the actual p of finding *one* p significant at .05 (or any other chosen level) increases as the number of tests made increases. That is not a good reason to decrease the variety of dependent variables used, assuming there is a good theoretical basis for choosing to use each one.

Alternate procedures are available. Bonferroni procedures can be used to adjust for the number of tests made (Rosenthal & Rubin, 1983). To overcome the conservatism of this basic approach and decrease Type II errors, it is possible to weight the dependent variables according to their importance and apply a so-called ordered Bonferroni procedure (Rosenthal & Rubin, 1984, 1985). Perhaps it is most useful, however, to apply specially developed procedures that integrate all the information from all the dependent variables and obtain only a single overall test of significance and effect size estimate. This can be accomplished very easily so long as we have reasonable estimates of the intercorrelations among the dependent variables (Rosenthal & Rubin, 1986).

Subdividing Studies

An issue discussed in the ganzfeld debate has to do with the subdivision of studies into substudies as a function of different experimental procedures or individual difference variables such as sex, age, degree of belief in psi effects, and the like (Schmeidler, 1968). As long as all the data are preserved and entered into the meta-analysis, no harm is done by subdividing. Indeed, subdividing is very useful in the search for moderator variables (Rosenthal, 1984).

Subdividing could have a very biasing effect on the accuracy of a cited p value if the overall data are subdivided in various ways, significant results are reported for one or more substudies, and the rest of the substudies are "thrown away." In the ordinary more proper application of meta-analytic procedures, however, subdividing makes little difference. Consider a psi experiment with an overall nonsignificant effect ($p = .13$, two-tailed). After the study is over, it is noted that about half the subjects were favorable toward psi and half were not and that there had been both female and male sub-

TABLE 4
SUBDIVISION OF A LARGER EXPERIMENT

	Believing subjects		Disbelieving subjects	
	Two-tailed p	z	Two-tailed p	z
Females	.05	2.0	.62	0.5
Males	.32	1.0	.62	−0.5

Note: For the study as a whole, p was .13 and z was 1.5 before subdividing. Positive z's reflect results in the predicted direction; negative z's reflect results in the unpredicted direction.

jects. Suppose that a subgroup of subjects, say female believers, show a significant psi effect but the remaining groups do not. No harm is done by reporting that fact, though an adjustment is useful in reporting the obtained p that takes into account how many subgroups were tested. It is essential, however, that the results of significance tests for the nonsignificant subgroups also be entered into the meta-analysis.

Table 4 illustrates the situation; four substudies have been formed, only one of which was significant. When we combine the results of the four substudies, however, we find the overall z to be $[(2.0) + (1.0) + (0.5) + (-0.5)]/\sqrt{4} = 1.5$, $p = .13$, two-tailed. Essentially, subdividing makes little difference so long as no data are discarded. If a particular substudy showed great promise of evidencing psi, nothing would prevent the investigator from conducting new studies using only the preselected experimental conditions or types of subjects. It would also be appropriate to conduct a meta-analysis on all the substudies that could be found that met the promising condition. In that case, however, the initial "study of discovery" should be entered with an adjustment for the fact that several tests of significance were computed (Rosenthal & Rubin, 1983, 1984).

Flaw Effects and Weighting Replications

There are few flawless studies in the behavioral sciences. Flaws can increase Type I or Type II errors, and the wise meta-analyst would do well to note how well Hyman (1985) and Honorton (1985) have searched for and evaluated flaws. For each flaw, it would be desirable to make some estimate of how much difference it made to the outcome. In the present debate some flaws seemed to make a difference and others did not. When flaws matter we can adjust for

these flaws in our weighting of studies. For example, we can give weights of zero to truly terrible studies and lowered but nonzero weights to less than truly terrible studies. Such weighting may lead to less biased conclusions than simple discarding of studies for flaws (Fiske, 1978; Rosenthal, 1984; Rosenthal & Rubin, 1985).

Replication Difficulty and Small Effects

Although Hyman (1985) and Honorton (1985) disagree on the degree of confidence warranted by the ganzfeld literature, they agree that the results reported do not reflect an enormous magnitude of effect. In Cohen's (1977) terminology, the average size of the ganzfeld effect reported by Hyman (1985) and Honorton (1985) is on the small side. That, of course, is not surprising. Controversial research areas are characterized by small effect sizes. For example, in a recent review of five controversial areas of human performance research, Harris and Rosenthal (1986) estimated the actual effect sizes (r) to range only from .00 to .18 with a median of .10 and a 95% confidence interval ranging from .02 to .19.

Small effect sizes are just what we should expect from controversial areas. According to fundamental principles of statistical power (Cohen, 1977), if the true effect size were substantial, studies with only modest sample sizes would routinely be able to reject the null. For example, if the population value of r were .60, 90% of replication attempts would be significant at $p < .05$ with sample sizes of 24 (Cohen, 1977, p. 92). However, if the population value of r were .10, the median of our five controversial areas (Harris & Rosenthal, 1986), only 7% of replication attempts would be significant at $p < .05$ with sample sizes of 24. For the small population value of r (.10), it would require sample sizes of over 1,000 to achieve a 90% rate of rejecting the null at $p < .05$.

Even though controversial research areas are characterized by small effects (including zero as a possibility), that does not mean that the effects are of no practical importance. Indeed, the median small effect of five areas cited above ($r = .10$) is equivalent to improving our success rate from 45% to a success rate of 55% (Rosenthal & Rubin, 1982b).

Before leaving the topic of replication difficulty, it may help us to place this problem in useful perspective by noting that it is not only in the parapsychological or other behavioral sciences that replication difficulties emerge. Indeed, students of the physical sciences have pointed out failures to replicate the construction of TEA-lasers

despite the availability of detailed instructions for replication. Apparently TEA-lasers could be replicated dependably only when the replication instructions were accompanied by a scientist who had actually built a laser (Collins, 1985).

Summarizing the Meta-Analyses

Hyman (1985) and Honorton (1985) have done important meta-analytic work on the topic of the ganzfeld experiments; it is this work I summarize here.

Five indices of "psi" success have been used in ganzfeld research (Honorton, 1985). One criticism of research in this area is that some investigators used several such indices in their studies and failed to adjust their reported levels of significance (p) for the fact that they had made multiple tests (Hyman, 1985). Because most studies used a particular one of these five methods, the method of direct hits, Honorton focused his meta-analysis on just those 28 studies (of a total of 42) for which direct hit data were available.

The method of direct hits scores a success only when the single correct target is chosen out of a set of t total targets. Thus, the probability of success on a single trial is $1/t$ with t usually = 4 but sometimes 5 or 6. The other methods, using some form of partial credit, appear to be more precise in that they use more of the information available. Although they differ in their interpretation of the results, Honorton (1985) and Hyman (1985) agree quite well on the basic quantitative results of the meta-analysis of these 28 studies. This agreement holds both for the estimation of statistical significance (Honorton, 1985, p. 58) and of effect size (Hyman, 1985, p. 13).

Stem-and-Leaf Display

Table 5 shows a stem-and-leaf display of the 28 effect size estimates based on the direct hits studies summarized by Honorton (1985, p. 84). The effect size estimates shown in Table 5 are in units of Cohen's h, which is the difference between (a) the arcsine transformed proportion of direct hits obtained and (b) the arcsine transformed proportion of direct hits expected under the null hypothesis (i.e., $1/t$). The advantage of h over j, the difference between raw proportions, is that all h values that are identical are identically detectable whereas all j values that are identical (e.g., $.65 - .45$ and $.25 - .05$) are not equally detectable (Cohen, 1977, p. 181).

TABLE 5
STEM-AND-LEAF PLOT OF "DIRECT HIT" GANZFELD STUDIES: COHEN'S *h*

Stem	Leaf
1.4	4
1.3	3
1.2	
1.1	
1.0	
.9	
.8	
.7	3
.6	
.5	8
.4	0 2 2 2 4
.3	1 2 2 4 4 7 8
.2	2
.1	3 8 8
.0	7 7 9
−.0	5
−.1	0
−.2	
−.3	2
−.4	0
−.5	
−.6	
−.7	
−.8	
−.9	3

Tukey (1977) developed the stem-and-leaf plot as a special form of frequency distribution to facilitate the inspection of a batch of data. Each number in the data batch is made up of one stem and one leaf, but each stem may serve several leaves. Thus, the stem .1 is followed by leaves of 3, 8, 8 representing the numbers .13, .18, .18. The first digit is the stem; the next digit is the leaf. The stem-and-leaf display functions as any other frequency distribution but the original data are retained precisely.

Distribution of studies. From Table 5 we see that the distribution of effect sizes is unimodal, with the bulk of the results (80%) falling between −.10 and .58. The distribution is nicely symmetrical, with the skewness index (g_1 = .17) only 24% of that required for significance at $p < .05$ (Snedecor & Cochran, 1980, pp. 78–79, 492). The tails of the distribution, however, are too long for normality with

kurtosis index $g_2 = 2.04$, $p = .02$. Relative to what we would expect from a normal distribution, we have studies that show larger positive and larger negative effect sizes than would be reasonable. Indeed, the two largest positive effect sizes are significant outliers at $p < .05$, and the largest negative effect size approaches significance, with a Dixon index of .37 compared to one of .40 for the largest positive effect size (Snedecor & Cochran, 1980, pp. 279–280, 490). The total sample of studies is still small; however, if a much larger sample showed the same result, that would be a pattern consistent with the idea that both strong positive results ("psi") and strong negative results ("psi-missing") might be more likely to find their way into print or at least to be more available to a meta-analyst.

Distribution of subjects. It is useful to examine the distribution of effect sizes obtained in the summarized studies. It would also be useful to examine the distribution of effect sizes obtained by individual subjects *within* the studies summarized. For example, in a study with a mean h of .20, is the distribution of h fairly normal with centering at .20, or is the distribution skewed with the bulk of the subjects centered closer to zero but with a few subjects earning consistently high values of h?

Distribution of investigators. Just as it is useful to examine the distribution of the results of studies and of subjects within studies, it is also useful to examine the distribution of results obtained by different investigators (Honorton, 1985; Hyman, 1985; Rosenthal, 1969, 1984). The 28 direct hit studies were conducted by 10 different investigators (Honorton, 1985, p. 60). Four investigators conducted only one study each, two conducted two studies each, two conducted three studies each, one conducted five studies, and one conducted nine studies. Analysis of variance showed that these 10 investigators differed significantly and importantly in the average magnitude of the effects they obtained with $F(9,18) = 3.81$, $p < .01$, *eta* = .81. Interestingly, there was little relationship between the mean effect size obtained by each investigator and the number of studies conducted ($r = .11$; $t(8) = 0.31$, $p > .70$).

That different investigators may obtain significantly different results from their subjects is well known in various areas of psychology (Rosenthal, 1966). For example, in such a standard experimental area as eyelid conditioning, studies conducted at Iowa obtained results in the predicted direction 94% of the time, whereas those conducted elsewhere obtained such results only 62% of the time with $\chi^2(1) = 4.05$, $p < .05$, $N = 25$, $r = .40$ (Rosenthal, 1966, p. 24; Spence, 1964).

TABLE 6
STATISTICAL SUMMARY OF "DIRECT HIT" GANZFELD STUDIES

Central tendency (Cohen's h)		*Variability*	
Unweighted mean	.28	Maximum	1.44
Weighted mean	.23	Quartile 3 (Q3)	.42
Median	.32	Median (Q2)	.32
Proportion positive sign	.82	Quartile 1 (Q1)	.08
		Minimum	− .93
Significance tests		Q3 − Q1	.34
Combined Stouffer z	6.60	$\hat{\sigma}$: [.75 (Q3 − Q1]	.26
t test of mean z	3.23	S	.45
z of proportion positive	3.40		
		Correlation of h	
		With z	.86
Confidence intervals[a]		With raw j	.98
From	*To*		
80% .17	.39		
95% .11	.45		
99% .04	.52		
99.9% −.03	.59		

[a]Based on *N* of 28 studies.

Summary of Stem-and-Leaf Display

Table 6 provides a summary of the stem-and-leaf display of Table 5 and some additional useful information about central tendency, variability, significance tests, confidence intervals, and correlations between Cohen's *h* and (a) significance level (*z*) and (b) raw difference in proportions (*j*). Only a few comments are required.

Effect size. The bulk of the results (82%) show a positive effect size where 50% would be expected under the null ($p = .0004$). The mean effect size, *h*, of .28 is equivalent to having a direct hit rate of .38 when .25 was expected under the null. The 95% confidence interval suggests the likely range of effect sizes to be from .11 to .45, equivalent to accuracy rates of .30 to .46 when .25 was expected under the null hypothesis.

Significance testing. The overall probability that obtained accuracy was better than the accuracy expected under the null was a *p* of $3.37/10^{11}$ associated with a Stouffer *z* of 6.60 (Mosteller & Bush, 1954; Rosenthal, 1978a, 1984).

File-drawer analysis. A combined *p* as low as that obtained can be used as a guide to the tolerance level for null results that never found their way into the meta-analytic data base (Rosenthal, 1979,

1984). It has long been believed that studies failing to reach statistical significance may be less likely to be published (Rosenthal, 1966; Sterling, 1959). Thus it may be that there is a residual of nonsignificant studies languishing in the investigators' file drawers. With simple calculations, it can be shown that, for the current studies summarized, there would have to be 423 studies with mean $p = .50$, one-tailed, or $z = 0.00$ in those file drawers before the overall combined p would become just $> .05$, as Honorton (1985) has pointed out.

That many studies unretrieved seems unlikely for this specialized area of parapsychology (Honorton, 1985; Hyman, 1985). Based on experience with meta-analyses in other domains of research (e.g., interpersonal expectancy effects) the mean z or effect size for nonsignificant studies is not 0.00 but a value pulled strongly from 0.00 toward the mean z or mean effect size of the obtained studies (Rosenthal & Rubin, 1978).

Comparison with an Earlier Meta-Analysis

It is instructive to compare the results of the ganzfeld research meta-analysis by Honorton (1985) with the results of an older and larger meta-analysis of another controversial research domain—that of interpersonal expectancy effects (Rosenthal & Rubin, 1978). In that analysis, eight areas of expectancy effects were summarized; effect sizes (Cohen's d, roughly equivalent to Cohen's h) ranged from .14 to 1.73 with a grand mean d of .70. Honorton's mean effect size ($h = .28$) exceeds the mean d of two of the eight areas (reaction time experiments [$d = .17$], and studies using laboratory interviews [$d = .14$]).

The earlier meta-analysis displayed the distribution of the z's associated with the obtained p levels. Table 7 shows a comparison of the two meta-analyses' distributions of z's. It is interesting to note the high degree of similarity in the distributions of significance levels. The total proportion of significant results is somewhat higher for the ganzfeld studies but not significantly so ($\chi^2(1) = 1.07$, $N = 373$, $p = .30$, $\phi = .05$).

Interpreting the Meta-Analytic Results

Although the results of the meta-analysis are clear, the meaning of these results is open to various interpretations. The most obvious

TABLE 7
PROPORTION OF STUDIES REACHING CRITICAL LEVELS OF SIGNIFICANCE FOR TWO RESEARCH AREAS

Interval for z	Expected proportion	Expectancy research[a]	Ganzfeld research[b]	Difference
Predicted direction				
+ 3.72 and above	.0001	.07	.04	−.03
+ 3.09 and above	.001	.12	.18	.06
+ 2.33 and above	.01	.19	.25	.06
+ 1.65 and above	.05	.36	.43	.07
Not significant				
− 1.64 to + 1.64	.90	.60	.50	−.10
Unpredicted direction				
− 1.65 and below	.05	.03	.07	.04

[a]N = 345 studies; from Rosenthal & Rubin (1978).
[b]N = 28 studies; from Honorton (1985).

interpretation might be that at a very low p, and with a fairly impressive effect size, the ganzfeld psi phenomenon has been demonstrated. However, there are rival hypotheses that will need to be considered, many of them put forward in the detailed evaluation by Hyman (1985).

Procedural Rival Hypotheses

Sensory leakage. A standard rival hypothesis to the hypothesis of ESP is that sensory leakage occurred and that the receiver was knowingly or unknowingly cued by the sender or by an intermediary between the sender and receiver. As early as 1895, Hansen and Lehmann (1895) described "unconscious whispering" in the laboratory, and Kennedy (1938, 1939) was able to show that senders in telepathy experiments could give auditory cues to their receivers quite unwittingly. Ingenious use of parabolic sound reflectors made this demonstration possible. Moll (1898), Stratton (1921), and Warner and Raible (1937) all gave early warnings on the dangers of unintentional cueing (for summaries see Rosenthal, 1965, 1966). The subtle kinds of cues described by these early workers were just the kind we have come to look for in searching for cues given off by experimenters that might serve to mediate the experimenter expectancy effects found in laboratory settings (Rosenthal, 1966, 1985).

By their nature, ganzfeld studies tend to minimize problems of sensory cueing. An exception occurs when the subject is asked to choose which of four (or more) stimuli has been "sent" by another person or agent. When the same stimuli held originally by the sender are shown to the receiver, finger smudges or other marks may serve as cues. Honorton has shown, however, that studies controlling for this type of cue yield at least as many significant effects as do the studies not controlling for this type of cue.

Recording errors. A second rival hypothesis has nearly as long a history. Kennedy and Uphoff (1939) and Sheffield and Kaufman (1952) both found biased errors of recording the data of parapsychological experiments. In a meta-analysis of 139,000 recorded observations in 21 studies, it was found that about 1% of all observations were in error and that, of the errors committed, twice as many favored the hypothesis as opposed it (Rosenthal, 1978b). Although it is difficult to rule recording errors out of ganzfeld studies (or any other kind of research), their magnitude is such that they could probably have only a small biasing effect on the estimated average effect size (Rosenthal, 1978b, p. 1007).

Intentional error. The very recent history of science has reminded us that even though fraud in science is not quite of epidemic proportion, it must be given close attention (Broad & Wade, 1982; Zuckerman, 1977). Fraud in parapsychological research has been a constant concern, a concern found to be justified by periodic flagrant examples (Rhine, 1975). In the analyses of Hyman (1985) and Honorton (1985), in any case, there appeared to be no relationship between degree of monitoring of participants and the results of the study.

Statistical Rival Hypotheses

File-drawer issues. The problem of biased retrieval of studies for any meta-analysis was described earlier. Part of this problem is addressed by the 10-year-old norm of the Parapsychological Association of reporting negative results at its meetings and in its journals (Honorton, 1985). Part of this problem is addressed also by Blackmore (1980), who conducted a survey to retrieve unreported ganzfeld studies. She found that 7 of her total of 19 studies were judged significant overall by the investigators. This proportion of significant results (.37) was not significantly (or appreciably) lower than the proportion of published studies found significant (.43) in Honorton's (1985) meta-analysis of direct hit ganzfeld studies ($\chi^2(1)$ =

0.17, $\phi = .06$. Somewhat similar results were obtained by Sommer (in press) in her analysis of research on the menstrual cycle. She found 61% of the published results to be significant compared to 40% of the unpublished studies; $\chi^2(1) = 2.30$, $p < .065$, one-tailed, $\phi = .20$. The results of the Blackmore and Sommer studies did not differ significantly ($z = 0.69$). Taken together, these studies provide only modest evidence for a serious file-drawer problem.

A problem that seems to be a special case of the file-drawer problem was pointed out by Hyman (1985). That was a possible tendency to report the results of pilot studies along with subsequent significant results when the pilot data were significant. At the same time it is possible that pilot studies were conducted without promising results, pilot studies that then found their way into the file drawers. In any case, it is nearly impossible to have an accurate estimate of the number of unretrieved studies or pilot studies actually conducted. Chances seem good, however, that there would be fewer than the 423 results of mean $z = 0.00$ required to bring the overall combined p to $> .05$.

Multiple testing. Each ganzfeld study may have more than one dependent variable for scoring degree of success. If investigators use these dependent variables sequentially until they find one significant at $p < .05$, the true p will be higher than .05 (Hyman, 1985). This issue was discussed earlier; it is not an inherently intractable one (Rosenthal & Rubin, 1986).

Randomization. Hyman (1985) has noted that the target stimulus may not have been selected in a truly random way from the pool of potential targets. To the extent that this is the case, the p values calculated can be in error. Hyman (1985) and Honorton (1985) disagree over the frequency in this sample of studies of improper randomization. In addition, they disagree over the magnitude of the relationship between inadequate randomization and study outcome. Hyman felt this relationship to be significant and positive; Honorton felt this relationship to be nonsignificant and negative. Because the median p level of just those 16 studies using random number tables or generators ($z = .94$) was essentially identical to that found for all 28 studies, it seems unlikely that poor randomization procedures were associated with much of an increase in significance level (Honorton, 1985, p. 71).

Statistical errors. Hyman (1985) and Honorton agree that 6 of the 28 studies contained statistical errors. However, the median effect size of these studies ($h = .33$) was very similar to the overall median ($h = .32$), so that it seems unlikely that these errors had a major

effect on the overall effect size estimate. Omitting these six studies from the analysis decreases the mean h from .28 to .26. Such a drop is equivalent to a drop of the mean accuracy rate from .38 to .37 when .25 is the expected value under the null.

A Tentative Inference

On the basis of the preceding summary and the very valuable meta-analytic evaluations of Honorton (1985) and Hyman (1985), what are we to believe? It would be easiest to say, "Let's wait until more data have been accumulated from studies purged of the problems noted by Hyman, Honorton, and others." That is not a realistic approach. At any point in time some judgment can be made, and though our judgment might be more accurate later on when those more nearly perfect studies become available, the situation for the ganzfeld domain seems reasonably clear. We feel it would be implausible to entertain the null given the combined p from these 28 studies. Given the various problems or flaws pointed out by Hyman and Honorton, the true effect size is almost surely smaller than the mean h of .28 equivalent to a mean accuracy of 38% when 25% is expected under the null. We are persuaded that the net result of statistical errors was a biased increase in estimated effect size of at least a full percentage point (from 37% to 38%). Furthermore, we are persuaded that file-drawer and related problems are such that some of the smaller effect size results have probably been kept off the market. If pressed to estimate a more accurate effect size, we might think in terms of a shrinkage of h from the obtained value of .28 to perhaps an h of .18. Thus, when the accuracy rate expected under the null is 1/4, we might estimate the obtained accuracy rate to be about 1/3.

Conclusion

Parapsychologists in particular and scientists in general owe a great debt of gratitude to Ray Hyman (1985) and Charles Honorton (1985) for their careful and extensive analytic and meta-analytic work on the ganzfeld problem. Their debate has yielded an especially high light/heat ratio, and many of the important issues have now been brought out into bold relief.

In my commentary on the ganzfeld debate, I focused most closely on the concept of replication. That seemed appropriate, not

only because of the centrality of the problem of replicability in the parapsychological literature, but also because of the centrality of the problem in many sciences, especially when the effect sizes sought in the population are small. The effect size zero is only a special case of the class of small effect sizes.

In closing I want only to suggest that parapsychological and other behavioral sciences would be well served to modify their view of the success of replication in the direction of the following newer view:

1. A replication is successful to the degree that the second study obtains an effect size similar to the effect size of the first study.

2. Three or more investigations are successful replicates of one another to the extent that the effect sizes are homogeneous.

3. Significance testing has nothing to do with success of replication though it can be useful in many ways, including the assessment of the likelihood of the null given all prior research (weighted as desired and as reasonable) and the likelihood of real differences among the effect sizes of two or more studies.

References

BLACKMORE, S. (1980). The extent of selective reporting of ESP ganzfeld studies. *European Journal of Parapsychology*, **3**, 213–219.

BROAD, W., & WADE, N. (1982). *Betrayers of the truth.* New York: Simon and Schuster.

COHEN, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

COLLINS, H. M. (1985). *Changing order: Replication and induction in scientific practice.* Beverly Hills, CA: Sage.

FISKE, D. W. (1978). The several kinds of generalization. *The Behavioral and Brain Sciences*, **3**, 393–394.

HANSEN, F. C. C., & LEHMANN, A. (1895). Ueber Unwillkürliches Flüstern. *Philosophische Studien*, **11**, 471–530.

HARRIS, M. J., & ROSENTHAL, R. (1986). *Interpersonal expectancy effects and human performance research.* Report prepared for the National Academy of Sciences.

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

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

KENNEDY, J. L. (1938). Experiments on "unconscious whispering." *Psychological Bulletin*, **35**, 526. (Abstract)

KENNEDY, J. L. (1939). A methodological review of extra-sensory perception. *Psychological Bulletin*, **36**, 59–103.

KENNEDY, J. L., & UPHOFF, H. F. (1939). Experiments on the nature of extra-sensory perception: III. The recording error criticism of extra-chance scores. *Journal of Parapsychology*, **3**, 226–245.

MOLL, A. (1898). *Hypnotism* (4th ed.). New York: Scribner.

MOSTELLER, F. M., & BUSH, R. R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology: Vol. 1. Theory and method* (pp. 289–334). Cambridge, MA: Addison-Wesley.

NELSON, N., ROSENTHAL, R., & ROSNOW, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, **41**, 1299–1301.

RAO, K. R. (1985). The ganzfeld debate. *Journal of Parapsychology*, **49**, 1–2.

RHINE, J. B. (1975). Second report on a case of experimenter fraud. *Journal of Parapsychology*, **39**, 306–325.

ROSENTHAL, R. (1965). Clever Hans: A case study of scientific method. In O. Pfungst, *Clever Hans* (pp. ix–xlii). New York: Holt, Rinehart and Winston.

ROSENTHAL, R. (1966). *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.

ROSENTHAL, R. (1969). Interpersonal expectations. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 181–277). New York: Academic Press.

ROSENTHAL, R. (1978a). Combining results of independent studies. *Psychological Bulletin*, **85**, 185–193.

ROSENTHAL, R. (1978b). How often are our numbers wrong? *American Psychologist*, **33**, 1005–1008.

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

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

ROSENTHAL, R. (1985). Nonverbal cues in the mediation of interpersonal expectancy effects. In A. W. Siegman & S. Feldstein (Eds.), *Multichannel integrations of nonverbal behavior* (pp. 105–128). Hillsdale, NJ: Lawrence Erlbaum Associates.

ROSENTHAL, R., & GAITO, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, **55**, 33–38.

ROSENTHAL, R., & GAITO, J. (1964). Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports*, **15**, 570.

ROSENTHAL, R., & ROSNOW, R. L. (1984). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.

ROSENTHAL, R., & RUBIN, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *The Behavioral and Brain Sciences*, **3**, 377–386.

ROSENTHAL, R., & RUBIN, D. B. (1979). Comparing significance levels of independent studies. *Psychological Bulletin*, **86**, 1165–1168.

ROSENTHAL, R., & RUBIN, D. B. (1982a). Comparing effect sizes of independent studies. *Psychological Bulletin*, **92**, 500–504.

ROSENTHAL, R., & RUBIN, D. B. (1982b). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, **74**, 166–169.

ROSENTHAL, R., & RUBIN, D. B. (1983). Ensemble-adjusted *p* values. *Psychological Bulletin*, **94**, 540–541.

ROSENTHAL, R., & RUBIN, D. B. (1984). Multiple contrasts and ordered Bonferroni procedures. *Journal of Educational Psychology*, **76**, 1028–1034.

ROSENTHAL, R., & RUBIN, D. B. (1985). Statistical analysis: Summarizing evidence versus establishing facts. *Psychological Bulletin*, **97**, 527–529.

ROSENTHAL, R., & RUBIN, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, **99**, 400–406.

SCHMEIDLER, G. R. (1968). Parapsychology. In *International Encyclopedia of the Social Sciences* (pp. 386–399). New York: MacMillan & Free Press.

SHEFFIELD, F. D., KAUFMAN, R. S., & RHINE, J. B. (1952). A PK experiment at Yale starts a controversy. *Journal of the American Society for Psychical Research*, **46**, 111–117.

SNEDECOR, G. W., & COCHRAN, W. G. (1980). *Statistical methods* (7th ed.). Ames: Iowa State University Press.

SOMMER, B. (in press). The file drawer effect and publication rates in menstrual cycle research. *Psychology of Women Quarterly*.

SPENCE, K. W. (1964). Anxiety (drive) level and performance in eyelid conditioning. *Psychological Bulletin*, **61**, 129–139.

STERLING, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, **54**, 30–34.

STRATTON, G. M. (1921). The control of another person by obscure signs. *Psychological Review*, **28**, 301–314.

TRUZZI, M. (1981). Reflections on paranormal communication: A zetetic's perspective. In T. A. Sebeok & R. Rosenthal (Eds.), *The Clever Hans phenomenon* (pp. 297–309). New York: New York Academy of Sciences.

TUKEY, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

WARNER, L., & RAIBLE, M. (1937). Telepathy in the psychophysical laboratory. *Journal of Parapsychology*, **1**, 44–51.

ZUCKERMAN, H. (1977). Deviant behavior and social control in science. In E. Sagarin (Ed.), *Deviance and social change* (pp. 87–138). Beverly Hills, CA: Sage.

Department of Psychology
Harvard University
Cambridge, MA 02138