Journal of Parapsychology, Vol. 50, December 1986

THE GANZFELD DEBATE: A STATISTICIAN'S PERSPECTIVE

By Jessica Utts

ABSTRACT: This paper is written in response to an invitation to comment on the meta-analysis of the ganzfeld research. Instead of responding to the details of the previous work, this paper focuses on the roles of testing and estimation in examining a data base of this kind. In particular, power is computed for the 24 direct-hit studies with $p = .25$, for two different alternative hit rates. Also, confidence intervals are computed for the probability of a direct hit in each of those studies. The issue of power is discussed as it relates to the concept of replicability. Specific comments on previous work are given in a few instances, but for the most part this paper is intended to shed even more light on the big picture rather than to focus on individual aspects of the analysis.

I am very pleased to be invited to participate in the historical event taking place with the joint publication by Hyman and Honorton on the psi ganzfeld controversy. As a statistician, I have been pleased and impressed with the sophistication of most of the analyses and rebuttals of Hyman (1985) and Honorton (1985). Furthermore, many of the issues raised are relevant not only to psi ganzfeld studies, but to other work in parapsychology and science in general.

In this paper, I have chosen for the most part not to respond to specific aspects of the previous work. This does not mean that I agree with everything that has been written. In particular, I object to many of the statistical methods used in Hyman's flaw analysis. But I do not think it would serve a useful purpose to pick on minor issues. Instead, I have tried to raise new issues that are relevant to meta-analysis in general and to the statistical treatment of a data base such as this one. These include comments on how meta-analysis is viewed by some statisticians, a discussion of the over-reliance on significance tests, an appeal to rely more on estimation, and an evaluation of distributional assumptions. I hope that my comments will be seen as being relevant to other areas of parapsychology and science, just as the previous work has been.

General Problems With Meta-Analysis

Meta-analysis is a relatively new field. It includes a rapidly expanding set of procedures for quantitatively combining results from several studies. It was for this reason that a group of prominent statisticians recently convened to discuss the value of these procedures and the likelihood that they would be properly developed and applied in the future. Although I did not attend the conference, I have received several reports from those who did.

Reactions of the participants were mixed, but everyone agreed that there are problems with the concept of meta-analysis in many cases. To paraphrase one participant: We thought we were comparing apples and oranges and studying fruit; we discovered that we were comparing apples and elephants and studying life.

The problem stems from the manner in which science generally proceeds. A study is done and gets a surprising result. Another study replicates the first, and is published as the definitive study. Future replications are either not done or are not published because they are "old news." New twists are added until someone comes up with one that seems to expand or contradict the original result. That study is then published, and the whole process begins anew. Thus, it is very rare to find in the literature several studies that are all measuring the same kind of effect under the same conditions.

The conference attendees agreed that more sophisticated meta-analytic techniques need to be developed. Models that incorporate factors such as laboratory differences should be used instead of treating all studies as if they were produced by a single source. After all, one can learn much about life by studying apples and elephants if one recognizes their differences as well as their similarities.

Other problems that have occurred in meta-analyses are discussed by Hedges and Olkin (1985). In the next section, I will outline the way in which some of these problems relate to the ganzfeld meta-analysis.

The Ganzfeld Studies

I do not intend to respond directly to the procedures used by Hyman (1985) and Honorton (1985) to summarize the ganzfeld studies because I think they have done a fine job of responding to each other. Instead, I have chosen to focus on some new ways of looking at the data, with an occasional reference to past work. I hope my suggestions will be useful for examining the results of future studies.

Replicability

Tversky and Kahneman (1982, Chapter 2) report an interesting phenomenon, which they call the belief in "the law of small numbers." They asked a group of 84 psychologists the question: "Suppose you have run an experiment on 20 subjects, and have obtained a significant result which confirms your theory (z = 2.23, $p < .05$, two-tailed). You now have cause to run an additional group of 10 subjects. What do you think the probability is that the result will be significant, by a one-tailed test, separately for this group?"

The median response was .85. In actuality, the true answer is somewhere around .48. They report the results of several similar experiments, all showing that our intuition is not very good at producing estimates of sampling variability.

In the ganzfeld data base, with all the emphasis on "successful outcomes," we would do well to see what kind of replication rate we should expect. Unless psi operates with 100% reliability, we should not expect every replication to reject the null hypothesis. Thus, we should not declare that a study "failed to replicate" the psi hypothesis just because the null hypothesis was not rejected.

The probability that a null hypothesis will be rejected is called the power of a study, and it depends on both the sample size and the true value of the parameter being tested. Very few researchers pay enough attention to the role of the sample size in determining power.

What does this mean for the ganzfeld studies? To see what we should expect for replication rates, I examined the power of all of the ganzfeld studies for which direct-hit information was available and for which the probability of a hit is .25. I obtained the data from Honorton's Table A1 of the Appendix (1985). I chose to restrict the results to those studies because I needed to choose specific alternatives at which to examine power. If the original expected hit rates were different, the power at a specific alternative would have meant something different.

The results are displayed in Table 1. I chose two specific alternatives to consider. The value p = .38 was chosen because it is mentioned by Hyman (1985, p. 13) as the weighted and unweighted hit rate for these studies. I chose p = .33 because that is the estimate given by Rosenthal (1986) after adjusting for the criticisms given by Hyman. The studies are ordered by increasing sample size so that the dependence of power on n can be seen. In each case, I obtained the exact critical region that would be used for a one-tailed test with

TABLE 1

POWER FOR THE DIRECT-HIT STUDIES WITH $p = .25$

Study	n	Critical region	Exact α	Power for $p = .33$	Power for $p = .38$
7	7	$\geqslant$ 5	0.013	0.043	0.078
19	10	$\geqslant$ 6	0.020	0.073	0.135
17	10	$\geqslant$ 6	0.020	0.073	0.135
16	10	$\geqslant$ 6	0.020	0.073	0.135
12	20	$\geqslant$ 9	0.041	0.181	0.334
24	20	$\geqslant$ 9	0.041	0.181	0.334
25	20	$\geqslant$ 9	0.041	0.181	0.334
21	20	$\geqslant$ 9	0.041	0.181	0.334
31	20	$\geqslant$ 9	0.041	0.181	0.334
30	26	$\geqslant$ 11	0.040	0.209	0.395
23	26	$\geqslant$ 11	0.040	0.209	0.395
38	27	$\geqslant$ 12	0.022	0.145	0.308
18	28	$\geqslant$ 12	0.029	0.181	0.364
26	30	$\geqslant$ 13	0.022	0.156	0.335
11	30	$\geqslant$ 13	0.022	0.156	0.335
8	30	$\geqslant$ 13	0.022	0.156	0.335
1	32	$\geqslant$ 13	0.038	0.230	0.445
28	32	$\geqslant$ 13	0.038	0.230	0.445
27	36	$\geqslant$ 14	0.046	0.279	0.519
29	40	$\geqslant$ 16	0.026	0.217	0.456
34	40	$\geqslant$ 16	0.026	0.217	0.456
41	48	$\geqslant$ 18	0.037	0.301	0.582
39	60	$\geqslant$ 21	0.030	0.316	0.632
33	100	$\geqslant$ 33	0.045	0.453	0.823

the largest $\alpha \leqslant .05$. I have listed cut-off points for the critical region as well as the exact value of α. Power is then calculated by computing the exact probability of falling in the critical region if $p = .33$ or .38, respectively. These values were obtained from *Tables of the Binomial Probability Distribution,* National Bureau of Standards, 1950, except for the last two rows, which were obtained using the Minitab Statistical Package.

Notice that for most of the studies, the power is relatively low. In other words, even if the probability of a direct hit is as high as 38%, we should not expect to reject the null hypothesis most of the time. For a sample of size 30, we would expect to obtain a significant result about one third of the time. Thus, we should be careful to define exactly what is meant by a "repeatable study."

Hedges and Olkin (1985, pp. 48–52) discuss the inadequacy of vote-counting as a method of determining whether an effect is zero. Their argument relates to the meta-analysis technique in which studies are categorized as being significant or not, and the category with the largest number of studies is declared to be the one in which a study is most likely to fall. The argument then follows that if there are more nonsignificant than significant studies, the effect must be zero.

The flaw in this method of vote-counting should be obvious after the above discussion of power. As Hedges and Olkin (1985, p. 51) point out, no matter how large the true effect size is, there are some sample sizes for which the power of this vote-counting technique will tend toward zero as the number of studies increases.

If the effect size and sample size remain constant across all studies, the number of significant studies expected out of m studies will follow a binomial distribution with p = power of each individual study. For example, if 100 ganzfeld studies are done, each with power of .33 (the case for $n = 30$ and true hit rate of .38), then we should expect to see about 33 significant studies with a standard deviation of 4.7 studies. In fact there is a 5% chance that there would be 25 or fewer significant studies, and a 30% chance that there would be 30 or fewer! Yet, many critics may not consider the ganzfeld paradigm to be successful if only 30 out of 100 studies produce significant results. In fact, Hyman and Honorton (1986) wrote, "If a variety of parapsychologists and other investigators continue to obtain significant results under these conditions, then the existence of a genuine communications anomaly will have been demonstrated" (p. 2). I would caution that a lack of significant results some of the time does not imply the lack of a genuine communications anomaly.

Hyman (1985, pp. 13–14) claims that there is evidence that the expected and actual number of significant ganzfeld studies differs more for small sample sizes than for large. But his results must be taken with a grain of salt. First of all, the expected number of studies in each category is only approximate. Second, it is not clear how Hyman computed his χ^2 value. There are specific distributional assumptions needed to use chi-square tests. If all studies in each category had the same probability of significance (which they do not), one could obtain a z score based on comparing observed and expected proportions, square each one, and add them up. The result would be χ^2 (4 *df*) = 24.56, the value reported by Honorton (1985,

p. 63) in his attempt to replicate Hyman's result. However, even this would be a questionable procedure for such small numbers, because it relies on the normal approximation to the binomial.

There is a further problem with this analysis. The way Hyman designated the four categories, and using the above method, even if all of the studies were significant the cell with $n =$ 45 to 184 could contribute only 3.00 to the χ^2 value. By contrast, the n = 5 to 19 cell could contribute as much as 46.85. This is because the expected proportion of significant studies is so much lower in the 5 to 19 cell.

Another way to see this is to recognize that even if all nine studies in the "large *n*" cell were significant, an exact binomial test of whether power = .75 (the value Hyman is apparently testing) would only have a p value of .0751. Thus, there is no possible way to have a "surprisingly high" number of significant studies in this category. By contrast, Honorton (1985, p. 63) reports that the exact test for the "small *n*" cell gives a p value of .0006, thus providing convincing evidence of a "surprisingly high" number of significant studies in that cell.

Effect Size and Other Estimation Procedures

In my opinion, there is too much emphasis on testing in parapsychology and in most other sciences and not enough emphasis on estimation. This obscures the relationship between sample size and accuracy and may lead to a false sense of confidence for studies with large sample sizes. For example, consider a binomial experiment for which the null hypothesis is $p = .5$. Suppose that for some reason (PK, mechanical deviations, etc.) the true $p = .52$. For a sample of size 100, one would expect z to be about .4, for a p value of .34. However, for a sample of size 10,000, z would be about 4, giving a p value of 3.17×10^{-5}, thus soundly rejecting the null hypothesis. In both cases the true difference is only .02, but the latter study appears to provide solid evidence of "an effect," whereas the former does not. Too often researchers confuse the magnitude of the p value with the magnitude of the effect. In this example, we would expect 95% confidence intervals for p to be about .42 to .62 for n = 100 and .51 to .53 for n = 10,000. Thus, even in the second case where the p value is extremely small, one can see that the magnitude of the effect is at most .03 (.53 − .50), and it can be left to the individual to determine whether that has any *practical* significance.

The recent emphasis on examining effect size may be necessary when one is doing a meta-analysis, but it obscures the interpretation

TABLE 2
95% CONFIDENCE INTERVALS FOR DIRECT-HIT STUDIES

Study	n	No. of hits	Point estimate	Lower limit	Upper limit
7	7	6	0.86	0.60	1.00
19	10	4	0.40	0.10	0.70
17	10	4	0.40	0.10	0.70
16	10	9	0.90	0.71	1.00
12	20	2	0.10	0.00	0.23
24	20	9	0.45	0.23	0.67
25	20	9	0.45	0.23	0.67
21	20	7	0.35	0.14	0.56
31	20	12	0.60	0.39	0.81
30	26	12	0.46	0.27	0.65
23	26	8	0.31	0.13	0.49
38	27	11	0.41	0.22	0.59
18	28	8	0.29	0.12	0.45
26	30	16	0.53	0.35	0.71
11	30	7	0.23	0.08	0.38
8	30	13	0.43	0.26	0.61
1	32	14	0.44	0.27	0.61
28	32	9	0.28	0.13	0.44
27	36	12	0.33	0.18	0.49
29	40	11	0.28	0.14	0.41
34	40	13	0.33	0.18	0.47
41	48	10	0.21	0.09	0.32
39	60	27	0.45	0.32	0.58
33	100	41	0.41	0.31	0.51

because most of us cannot intuitively interpret the results of, say, an arcsin transformation (c.f. Rosenthal, 1986, Table 5). Thus, in addition to combining effect sizes across studies, it is beneficial to examine estimates for each study.

Table 2 gives point estimates and 95% confidence intervals for the 24 direct-hit studies that used $p = .25$. These are based on normal approximations, and thus are not exact, but my aim here is to give a general picture of the magnitude of the effect and the accuracy with which we can estimate that magnitude. As in Table 1, the studies are organized by increasing sample size so that patterns can be seen easily.

It might be of interest to note that a 95% confidence interval for p based on the combined 722 trials is .345 to .415. Extreme caution must be used in quoting such a result, however. It assumes that all

722 trials follow the binomial assumptions of independence and constant success probability. In all likelihood, the success rate actually changes depending on the individual subjects, conditions, and experimenters. The same criticism can be leveled at the confidence interval for each individual study. But at least in that case each reader can determine the extent to which the assumptions were violated by reading the report of the study and by considering the issues raised by Hyman (1985) and Honorton (1985).

The focus on estimation also removes the criticism of multiple analyses. The criticisms showing that the probability of a direct hit should be higher than .25 can be evaluated by each reader, and the estimated magnitude of the effect can be viewed in light of those.

Distributional Assumptions

Exactly what can we assume about the distribution of the number of hits within each study and across studies? Hyman (1985, p. 8) claims that "studies in the data base are not independent (several coming from the same investigators) and are sampled from an unknown population."

The independence of random variables is a slippery concept. Formally, two random variables are independent if the distribution of one of them is unchanged, given knowledge about the value of the other one. I would argue that the ganzfeld trials are all independent because the outcome on one trial should not affect the outcome on any other trial. However, I would also argue that each trial comes from a different distribution or population. I would adopt the model that each trial comes from a binomial distribution, but with its own distinct probability of success. In statistical jargon, this means that they are independent but not identically distributed Bernoulli trials. An analogy would be an experiment in which each subject brought in a biased coin and flipped it. Even if the same subject flipped the same coin several times, those flips would be considered to be independent.

If we assume this model is correct, then the number of hits follows the "generalized binomial distribution of Poisson" (Patil et al., 1984, p. 16). The theoretical population mean for overall number of hits is Σp_j and the variance is $\Sigma p_j q_j$, where the p_j's are the individual success probabilities, and $q_j = 1 - p_j$. Furthermore, the central limit theorem still applies, so that we may use the normal approximation to do hypothesis testing and to get confidence intervals for the *average* hit rate.

A conservative approach would be to use the maximum possible value for the variance of the number of hits, which is $n/4$. Thus, a conservative 95% confidence interval for the average p_j, based on the 722 trials from the studies in Table 1 would be .344 to .416. Notice that this is very close to the interval derived using the regular binomial assumptions.

Of course this does not imply that psi is necessarily operating. There could be several other explanations for why the average hit rate has this magnitude, as Hyman (1985) has pointed out.

Future Work

Hyman and Honorton (1986) have done an admirable job of providing guidelines for future psi ganzfeld studies. By following their recommendations, it is possible that future meta-analysis of these studies may indeed allow fruitful comparisons.

However, I would caution against the strict use of hypothesis testing for these studies and would advocate the use of estimation. By removing the focus from the question of significance, which is highly dependent on sample size, and placing the emphasis on estimation, we will learn much more about the magnitude of the effect.

References

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press, Inc.

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.

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

Patil, G. P., Boswell, M. T., Joshi, S. W., & Ratnaparkhi, M. V. (1984). *Dictionary and classified bibliography of statistical distributions in scientific work: Volume 1. Discrete models.* Fairland, MD: International Cooperative Publishing House.

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

TVERSKY, A., & KAHNEMAN, D. (1982). Belief in the law of small numbers. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Division of Statistics
University of California
Davis, CA 95616