The Journal of

Parapsychology

Volume 49 Number 1 March 1985

THE GANZFELD DEBATE

This issue of the *Journal* is devoted to discussions of ganzfeld ESP studies. The papers by Ray Hyman and Charles Honorton are invited contributions. Therefore, they are published here without any refereeing and with very little editing. The basis for invitation is the interesting exchange between Hyman and Honorton on the question of replication of ganzfeld ESP studies at the joint conference of the SPR and the PA held at Cambridge University during August 1982. Summaries of their presentations are published in *Research in Parapsychology, 1982*.

Repetition of an experiment or the possibility of it is not a matter of primary importance in *normal* science. Only on rare occasions do we find reputed journals reporting straightforward replications. When experiments are repeated, it is for secondary reasons such as (a) improving experimental techniques, (b) accumulating more data for greater generalizability, or (c) checking on the competence of the experimenter.

The role of replication in controversial areas is somewhat different and paradoxical. If the results are readily replicable, then there should be little room for controversy about them. If they are not easily replicable, we cannot simply reject them as spurious because there are phenomena such as planetary positions that do not repeat themselves. Thus, in a sense, replicability is an inappropriate criterion for distinguishing between the genuine and the spurious in science.

For a subject like parapsychology, however, the replication question has a special significance. First, much of the serious research in parapsychology is laboratory-oriented. As a laboratory science, parapsychology presupposes that psi phenomena are replicable in principle. Second, the rate of replication is a fair index of the frequency of occurrence of a given phenomenon, which is necessary for a systematic study. A knowledge of the frequency with which one may obtain psi in a laboratory is helpful for making an intelligent choice of a career in the field. Third, there is much interest in the possibility of applying psi for practical use. Applying psi for pragmatic use depends largely

on our success in obtaining reliable results. Fourth, several parapsychologists have indeed made claims that their results are replicated to a degree. And these claims deserve consideration.

It is therefore important that the question of replication be fully discussed. The ganzfeld ESP studies seem to be especially appropriate for this purpose because (a) recent reviews have suggested a fair amount of replication (Blackmore, 1980; Honorton, 1977; Sargent, 1980); (b) these studies are of recent origin, which makes it easier to have access to the original data; and (c) the rationale behind them fits very well with the widely held belief that psychic abilities are manifested better under conditions of reduced sensory input.

It is our belief that a comprehensive discussion of the replication question in relation to ganzfeld ESP studies is important for the following reasons: to clarify the concept of replication itself, to examine the nature of evidence in controversial areas, to resolve certain methodological problems, and to deal effectively with issues relating to meta-analysis. It is indeed our hope that the two papers published in this issue will start a dialogue that will be continued further. We invite further exchanges and responses to these papers, which we hope will be published in a subsequent issue.

K. R. Rao
Editor

Journal of Parapsychology, Vol. 49, March 1985

THE GANZFELD PSI EXPERIMENT: A CRITICAL APPRAISAL

By Ray Hyman

ABSTRACT: The paper describes a critical evaluation of 42 ganzfeld psi studies reported from 1974 through 1981. Allegedly, 55% of these studies achieved significance on the primary index of psi. The first part of the critique challenges this claimed rate of successful replication. Taking into account ambiguities and inconsistencies in what is counted as an independent ganzfeld study, and citing evidence suggestive of a bias in reporting the studies, it is argued that the actual rate of success was at most 30%. The second part points out that, because of multiple testing, the true significance level was much higher than the assumed .05 level, perhaps .25 or higher. The third part tallies a number of procedural flaws involving inadequate randomization, potentials for sensory leakage, statistical errors, and the like, and strongly suggests that most of the studies in this data base were originally intended to be exploratory investigations rather than well-planned, confirmatory experiments. The final part is a meta-analysis based on indices of significance and effect size as they relate to the various categories of flaws. The flaws of inadequate security, possible sensory leakage, and multiple testing did not correlate with significance and effect size. But the flaws involving inadequate randomization and insufficient documentation did correlate with these indices. Both effect size and *Z* scores become approximately zero when regression equations are used to predict their values for the case in which these latter types of flaws are zero. It is concluded that this data base is too weak to support any assertions about the existence of psi.

In the latter half of 1981, I found myself with two assignments to provide a critical assessment of the field of parapsychology. In both cases, I had initially refused because the task seemed beyond my available resources. But I was urged to reconsider on the grounds that no other qualified critics were available. The option of trying to review the entire research literature was impractical, even if I restricted myself only to papers that had been published in refereed journals, such as the *Journal of Parapsychology*, *The Journal of the American Society for Psychical Research*, and *The European Journal of Parapsychology*. I also

An earlier version of this paper was presented at the joint meetings of the Society for Psychical Research and the Parapsychological Association in Cambridge, England, August 1982. Part of the preparation of the paper was done during the academic year 1982-1983 while I occupied the Thomas Welton Stanford Chair for Psychical Research at Stanford University. The present manuscript benefited from the comments made on earlier versions by Susan Blackmore, Irvin Child, Robyn Dawes, Persi Diaconis, Piet Hein Hoebens, Charles Honorton, J. E. Kennedy, Adrian Parker, and Christopher Scott.

rejected the more feasible alternative of evaluating a random sample of this literature. Such a sample could supply a picture of the adequacy of the average research report in parapsychology, but I felt it would be fairer to try to assess the case for parapsychology at its best. I suspect that the typical contribution to any research enterprise is mediocre and that the viability of a research program is best judged by its strongest representatives. Nevertheless, I did not want to follow Hansel's (1980) approach of focusing on only a handful of the "best" individual experiments.

My compromise was to look for a research program in parapsychology that consisted of a series of studies by a variety of investigators and was considered by parapsychologists as especially promising. As a result both of reading some of the parapsychological literature and of talking with some parapsychologists, I chose the ganzfeld psi paradigm as the most appropriate. This paradigm consists of a systematic body of research that covers a span of 10 years, involves a number of highly respected investigators, and has allegedly produced significant psi scores in over half of the experiments. In addition, I was intrigued by some of the claims made about the high level of research sophistication and rigor that had been achieved in these experiments. In presenting his case for parapsychology, Rogo (1977), for example, chose the Honorton and Harper ganzfeld psi experiment (1974) as an example of a good ESP experiment. "To me," Rogo wrote, "a good experiment is one that is designed to safeguard against fraud and experimental error and uses a clear-cut method of analysis to see whether or not ESP actually occurred during the tests" (p. 41).

On August 18, 1981, I wrote to Charles Honorton to request his help in obtaining access to the ganzfeld psi data base. Honorton phoned to tell me that it would take some time but that he would gladly undertake the mission of getting me copies of every relevant study. He felt it was important to get an outside critic such as myself to assess this body of literature. He hoped that it might lead to cooperative ventures in which critic and parapsychologist could attempt careful replications. In January 1982, I received from Honorton a copy of every ganzfeld study known to him, along with his detailed analysis of the various characteristics of this sample. Honorton also included a number of papers that criticized aspects of the ganzfeld research or commented on it (e.g., Blackmore, 1980; Honorton, 1978, 1979, 1981; Kennedy, 1979a, 1979b). All told, the entire package consisted of 600 pages of reports.

The Data Base

The data base was extracted from 34 separate reports written or published from 1974 through 1981. By Honorton's count, these 34 reports described 42 separate studies. Of these, he classified 23 as having achieved overall significance on the primary measure of psi at the .05 level. This successful replication rate of 55% is consistent with earlier estimates of 54% (Honorton, 1978), 58% (Sargent, cited in Blackmore, 1980), and 50% (Blackmore, 1980). If we treat each study as a unit, then we find that 15 (36%) appeared in refereed publications (11 in the *Journal of the American Society for Psychical Research*, 3 in the *European Journal of Parapsychology*, and one in the *Journal of the Society for Psychical Research*); 5 (12%) appeared in a published monograph; 20 (48%) appeared only in the form of abstracts or papers delivered at meetings of the Parapsychological Association; and 2 (5%) were part of an undergraduate honors thesis in biology. The studies were authored by 47 different investigators. Carl Sargent's 9 studies and Charles Honorton's 5 account for one third of the total. Other major contributors were John Palmer with 4, Scott Rogo with 4, W. G. Braud with 3, and Rex Stanford with 3. These six parapsychologists account for two thirds of the data base. (See Appendix.)[1]

Procedure

Prior to the evaluation that I made for the present paper, I had made two prior analyses of the same data base. The first analysis was done for a paper that I presented at the combined meetings of the Society for Psychical Research and the Parapsychological Association in August 1982. As a result of comments on that paper made by Honorton and others, I reanalyzed the data base again in November 1982. For the purposes of the present critique, I began a new and more systematic analysis of this data base in July 1983 and finally finished the task in January 1984.

In the present evaluation, I tried to take into account the comments and disagreements generated by my previous two evaluations. Honorton, for example, disagreed with many of my assignments of

[1]To meet the requirements of this report, the editors have amended the usual style for references: When a reference is cited for general or historical purposes, it is followed as usual by the year of publication in parentheses and is listed in the Reference section. For the 42 reports in the data base, however, study *numbers* are used (e.g., Study 1) and are referenced in the Appendix.

flaws to the studies. I hope to reduce some of this disagreement by using more specific and refined categories for encoding the flaws. For example, in the previous critiques, I assigned a flaw in the category "multiple testing" to any study that was guilty of this flaw, regardless of the specific way this was manifested. The current analysis now uses six different and narrower categories to cover the same problem. I also refined and made more operational the categories for procedural flaws. In all, I used 12 systematic categories for assigning flaws to studies. These categories and their assignment to each study in the data base are listed in the Appendix.

I also noted a variety of other defects, which I did not formally assign because of a variety of reasons. Some depended on my subjective impression; others were flaws only under some circumstances; some were unique to a given context; and so on. Many of these will be discussed in appropriate places in the manuscript.

General Guidelines

A reviewer can try to use the ganzfeld psi data base to answer a variety of questions. Many of the individual studies, for example, examined the relationship between psi scores and personality variables such as extraversion. Although most investigators talk about the ganzfeld procedure as "psi conducive," only a few have actually tried to test this hypothesis by including nonganzfeld control conditions. Following Honorton's lead, I ignored questions about ganzfeld psi and personality correlates (although I could not help being startled by investigators' routinely doing factor analyses on sample sizes of 30 or less). My review focuses on two questions: (1) Does this data base, taken as a whole, supply evidence for the existence of psi? (This is the question of major concern to the outsider.) (2) Does the ganzfeld psi study yield evidence for psi that is replicable? (This is the question of major concern to Honorton and other parapsychologists.)

The basic index for both these questions is some measure of hitting or target matching compared with a chance baseline. This creates special problems when compared with more conventional measures of effect that depend on empirical comparisons between two or more groups. In particular, assumptions about probability distributions take on a greater burden in the decision about whether a given discrepancy between the observed and the theoretical value is significant.

The prototype or basic pattern against which to judge a replication is the Honorton and Harper study (Study 8). Not only was this the

first published ganzfeld psi study, but also it served as the model for subsequent studies. In addition, it is relatively uncomplicated in that it consists of a single uniform condition for all subjects. No comparison or control conditions exist, and the only meaningful way to evaluate the observed number of hits is against some theoretical expectation.

In evaluating the adequacy of each study, I have tried to use criteria that I believe every competent parapsychologist, including most of the authors of papers in the present data base, would endorse. I have been guided by parapsychologists such as Palmer (1983), who has written:

> How do parapsychologists define psi? . . . One [definition] which I think most of us would accept is the following: psi is a statistically significant departure of results from those expected by chance under circumstances that mimic exchanges of information between living organisms and their environment, *provided that* (a) proper statistical models and methods are used to evaluate the significance, and (b) reasonable precautions have been taken to eliminate sensory cues and other experimental artifacts. (p. 54)

Palmer's definition also provides reasonable standards against which to judge the adequacy of a psi study. To his two general criteria of appropriate statistics and adequate controls, I would add two more. One is related to the statistical criterion. It would require that treatments and targets be assigned in such a way as to guarantee the major assumptions on which the statistical tests are based (such as independence of trials, appropriate distributions, and so forth). The other involves the inclusion of sufficient documentation to enable the reader to judge if certain departures from the model experiment do or do not make a difference (such as having some trials using friends of the percipient as agent and others using strangers).

Using these general guidelines, my critique involved the following phases:

1. *Rechecking the vote count.* The replication rate of 55% in this data base depends critically on what qualifies as an independent study. Should the sampling unit be the study, the cells of the study, the report, the laboratory, or the investigator? What counts as an acceptable ganzfeld study? To what extent could the apparent success rate be biased by unreported studies?

2. *Assessing the actual as opposed to the assumed level of significance.* Honorton accepts as "successful" those studies that achieve overall psi-hitting at the .05 level, one-tailed, as well as studies that achieve overall psi-missing at the .05 level (two-tailed). In addition, a replication is considered successful if it achieves significance on at

least one of a number of different possible indices. Many of the studies exhibit other statistical practices that increase the probability of achieving significance well above the reported .05 criterion.

3. Assigning procedural flaws to studies. The preceding phase assesses the ways in which experimenters unwittingly inflate the true significance level by multiple testing. Another threat to the validity of the statistical inferences is the failure to insure that randomization of targets is adequately carried out or that the statistical tests are correctly used (proper degrees of freedom, correct calculation of the *p*s, and so on). Other procedural flaws involve failure to secure against witting and unwitting sensory leakage and failing to adequately document information that could potentially affect the interpretation of the results. To the extent that such procedural flaws characterize the data base, the suspicion is justified that the studies were not carefully planned and executed.

4. Making a meta-analysis of the relationships among the flaws, effect size, and significance. For each study in the data base, the Appendix lists, in addition to the flaws, Honorton's original and revised classification of each study as significant or not. Also, for the 36 studies for which the appropriate data were available, I have supplied a common effect size as well as the *Z* score.

Perhaps this is the place to point out that the several statistical evaluations involving effect size, significance, and flaws were all carried out in the spirit of a meta-analysis (Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1982). Because the studies in the data base are not independent (several coming from the same investigators) and are sampled from an unknown population, drawing valid statistical inferences is an uncertain procedure. In addition, the many tests, although converging and consistent among themselves, can provide only suggestive hypotheses about possible relationships. For these tests, I have used the conventional significance levels only as a convenient yardstick for suggesting possible relationships that may be worth further exploration.

The Vote-Counting Problem

As already mentioned, Honorton identifies 42 separate studies in this data base, 55% of which he classifies as successful in terms of achieving overall significance on the primary measure of psi. Light and Smith (1971), as well as subsequent writers, have indicated that this method of "vote-counting" raises many problems. Nevertheless, if

the count is correct, and if the rate of success expected by chance is truly .05, then such a replication rate is impressive. In the next section of the paper, I shall discuss whether the actual level of significance in these studies is higher than the advertised .05 level. In this section, I deal with the question of whether the replication rate should be considered much lower than the claimed 55%.

For the sake of the present discussion, I shall ignore the vexing and potentially important question of what the appropriate sampling unit should be for aggregating the findings across this data base. Within the separate papers, the investigators seem to treat the single trial as the independent sampling unit regardless of whether the total set of trials comes from the same or different subjects. In addition to indiscriminately pooling within- and between-subject contributions, both Honorton and the individual experimenters sometimes also pool across separate experimental conditions without trying to deal with the problem of interdependencies among the sampling units. This issue of interdependencies may also matter in deciding what the sampling unit should be for counting successes in the data base—the individual study, cells of the study, the investigator, the laboratory, or the report. A discussion of some of these issues and the problems they raise can be found in Glass et al. (1981) and in Hedges and Olkin (1982).

Honorton has opted for the study as the unit of replication. For much of the data base, this creates little ambiguity. Even when a report contains more than one study, these usually are easy to identify. But Honorton is not consistent. In the experiment by Braud and Wood (Study 3), which contains several different ganzfeld conditions, Honorton pools the data over trials within and between subjects as well as across conditions to come up with one successful replication. This seems to be his typical response to studies with multiple conditions. However, Honorton treats Raburn's study differently by partitioning it into its separate cells, discarding two of the cells as being too atypical, and counting one cell (Study 16) as a significant experiment and the other as an insignificant experiment (Study 17). No reason is proffered for treating the cell as the unit in this case and the total study as the unit in the other cases.

By itself, this particular inconsistency in the vote-counting would not alter the results very much. But it does begin to suggest some of the problems involved in post hoc attempts to define and assess a body of research literature. Honorton and I are both concerned with the probability of successfully replicating the ganzfeld psi study. Using the Honorton and Harper study (Study 8) as the prototype, one

could argue that each experimental condition of a psi study in which all subjects are treated uniformly and produce their mentations under ganzfeld conditions can be taken as a replication. This amounts to treating each separate cell of a study in which the subjects are run under the ganzfeld conditions as the replication unit. In this framework, Honorton is correct in treating the individual cells of Raburn's study as separate replications. He is incorrect, in my opinion, in discarding two of the cells on the grounds that no other ganzfeld psi studies have run subjects under conditions in which they did not realize their guesses were being scored for psi. One could just as easily argue, remembering that all of these decisions are post hoc, that many of the other replications should be discarded because they, too, contain conditions that do not appear in any of the other ganzfeld studies. Which decision one makes in the case of Raburn's report makes a difference between scoring it as contributing one successful and one unsuccessful replication or as contributing one successful and three unsuccessful replications.

In addition to Braud and Wood (Study 3), 10 other studies in the data base contain multiple conditions that could be considered as separate replications (Studies 6, 9, 10, 13, 29, 32, 33, 35, 37, 41). The issue raised by all of them can be illustrated by discussing the Braud and Wood experiment (Study 3). These investigators divided their sample of 30 subjects into two groups of 15. Each subject served in six experimental sessions. The first session for each group was essentially a replication of the Honorton and Harper experiment (Study 8). In both groups, the results on the primary measure of psi were insignificant. It would seem reasonable to argue that here we have two clear-cut failures to replicate. Each subject returned for four practice sessions. Each of these sessions differed from the original session in that two practice targets were "sent" in addition to a regular target. In the feedback group, the practice responses were accompanied by immediate feedback via the intercom system. In the control groups, no such feedback was given during the practice responses. Following the two practice periods, each session ended with the subjects responding to a target just as in the original ganzfeld session. Again, it can be argued that each of these eight separate practice sessions constitutes a separate replication. The same can be said of the two postpractice sessions, one of which gave significant results. All told, this one study, which is counted as a single successful replication by Honorton, could be viewed—with equal justification, and consistent with his treatment of Raburn—as contributing one successful and 11 unsuccessful replications to the total. Following this logic with the

other 10 studies with multiple cells, I achieve a count of 25 "successes" out of 80 replications, for a success rate of 31%.

One could argue, of course, that treating the cells as the unit has drawbacks. It raises, again, the question of independence of units, especially when the same subjects appear in different cells. But this problem of independence of units plagues Honorton's procedures, and other vote-counting procedures as well. As I have already indicated, this problem runs rampant through this data base in a variety of ways. At this point, it is unclear just how it affects the various analyses. One could further argue that the use of the cell rather than the total study as the unit lowers the sample size per unit, thereby lowering power. Ordinarily this would be a reasonable objection; but, as we will see shortly, a peculiarity of this data base is that significance is uncorrelated with sample size.

The File-Drawer Problem

Even if we accept the present data base as complete, an argument can be made for asserting that the successful replication rate is more like 31% than 55%. But it is likely that the data base is incomplete. Parker and Wiklund (1982) included in their survey 11 ganzfeld psi studies that are not in the present data base. Honorton has subsequently added some of these to his data base (personal communication, October 2, 1982). And Blackmore's survey (1980) uncovered 19 other unpublished studies. If we add these to Honorton's initial count, we come up with a total of 53 studies, with an apparent success rate of 43%. If we add these, instead, to my adjusted count, we come up with a total of 110 studies, with a success rate of 30%. This latter estimate is an upper bound because I have not seen these 31 additional studies and do not know how many of them contain multiple conditions that have been pooled together.

The fact that the success rate decreases as we find and add previously unknown studies to the data base is consistent with both the general belief and the empirical finding (Glass et al., 1981) that unreported studies tend to be those with lower effect sizes. To the extent that this is so, we would expect the success rate to be even lower if we could find and include all the currently unknown ganzfeld psi studies. Rosenthal (1978, 1979) has suggested ways to estimate the seriousness, for any data base, of this "file-drawer problem." Honorton (1979), at a time when the ganzfeld data base included 28 studies, used Rosenthal's procedure to estimate that it would require 275 unreported and nonsignificant studies "to reduce the overall signifi-

cance of the reported psi ganzfeld work to $p < .05$, two-tailed." He adds:

> Considering that the average ganzfeld experiment involves 40 trials with a per-trial time investment on the order of one hour, it would take 12 laboratories nearly six months each to accumulate this number of unreported failures (assuming eight-hour days, no coffee breaks or vacations). Considering the small number of active researchers in parapsychology, and the meager resources that are available to them, we can confidently reject the hypothesis that the ganzfeld success rate is due to selective reporting of significant studies. (p. 388)

Honorton's argument seems to gain even more force when we apply Rosenthal's technique to the current data base. Using either Honorton's count or my adjusted count, I come up with estimates of between 440 and 580 unknown and nonsignificant ganzfeld psi studies needed to reduce the current data base to one selected from a distribution centering around chance expectancy. But the critic can find a number of reasons to question this argument. One of the reasons will be the subject of the next section. It deals with the assumption that studies in this data base are operating at the .05 level. If this is not so, the application of the procedure for estimating the maximum number of fugitive studies has to be either appropriately adjusted or abandoned.

Evidence for Biased Reporting

Much of the force of the argument against the seriousness of the file-drawer problem depends on the assumption that the ganzfeld study is time consuming and requires special resources to conduct. There is some merit to this argument, but it is not quite so strong as Honorton asserts. For one thing, his calculations of effort involved assumes that the typical study contains 40 trials. In fact, for this data base, over half of the studies have 30 or fewer trials. But more important is the possibility that many of the unreported studies were aborted before many trials were completed. So far as I can tell, every ganzfeld psi study in this data base was conducted in such a way that the experimenters were aware of the cumulative number of successes and failures as each new subject was added to the data base. Indeed, such awareness is reflected in the reports. Palmer et al. (Study 11), as a result of noting that the accumulating hits were at a chance rate, made drastic changes midway in the study to try to increase the hit rate. Perhaps it is unlikely that a large number of completed ganzfeld

studies remain unreported. But it is easier to imagine that a large number of experimenters, after reading about the ganzfeld, might have begun conducting some trials and then abandoned the study when the first few trials turned out to be unpromising. On the other hand, a few of these exploratory ventures might have started with initially successful trials, encouraging the experimenter either to continue or to stop and write up the result as a successful replication.

Is there any basis for such a suggestion? In fact, this suggestion occurred to me as a result of examining the relation between sample size and significance in this data base. One of the problems of the vote-counting method, of course, is that it ignores sample size. In the current data base, the studies vary in size all the way from 6 to 180 trials. (Because both Honorton and the experimenters seem to treat trials as the sampling unit regardless of whether they are within or between subjects, I shall also ignore this distinction for purposes of the present discussion.) We would normally expect to find the probability of obtaining a significant result, all other things being equal, to increase with the square root of the sample size. We can calculate theoretically expected proportions of significant studies (power) for each sample size if we know the true effect size. For this purpose, I used the data reported from all the studies in the data base that reported direct hits based on chance outcome of $P = 1/4$. This involved 22 of the studies and eight different investigators. All told there were 746 trials (48% of all the trials in the data base). For this situation, the estimated number of hits is 38%. Fortunately, this estimate is the same regardless of whether we compute the weighted or unweighted average of hit rates for each individual investigator. Using 38% as the theoretical true hit rate and 25% as the chance rate, we can calculate the power for various sample sizes.

To evaluate how the actual proportion of significant studies departed from the theoretically expected proportion, I grouped the 42 studies into four classes of sample size. For each class, I obtained the observed and theoretical proportion of significant studies (using Honorton's original classification given as SIG-1 or NSIG-1 in the Appendix). For the class with 5 to 19 trials (median = 10), 5 of the 7 studies were significant as compared with a theoretical expectation of 0.91 significant studies. For the class with 20 to 29 trials (median = 20), 6 of the 12 studies were significant as compared with a theoretical expectation of 3.96 significant studies. For the class with 30 to 34 trials (median = 35), 7 of the 14 studies were significant as compared with a theoretical expectation of 6.58. And for the class with 45 to 184 trials (median = 72), 5 of the studies were significant as compared with a

theoretical expectation of 6.75. This tendency for the studies with the fewer trials to have a higher proportion of significant outcomes than predicted is highly significant ($\chi^2[4\ df] = 31.42$) with most of the contribution coming from the class with a median of 10 trials.

The most obvious conclusion is that such a strange relationship is due to a selective bias. It suggests a tendency to report studies with a small sample only if they have significant results. This is understandable in that a significant outcome is likely to be accepted for publication even if the sample size is small, but a nonsignificant study with only 5 to 19 trials is easy to dismiss as having inadequate power. Another consequence of such a selective bias would be that effect size should be greater in this data base for the experiments with the smallest number of trials. As we will see later, this, indeed, is the case.

The "Retrospective" Study

This proposed bias toward reporting small studies only if they succeed is related to what I refer to as the "retrospective study." This is the tendency to decide to treat a pilot or exploratory series of trials as a study if it turns out that the outcome happens to be significant or noteworthy. Such a tendency, if it exists, operates to inflate the apparent success rate in a way different from the file-drawer problem. In the latter case, the observed data base has an inflated rate of success because many studies that did not achieve significant outcomes are not reported. The retrospective study inflates the success rate by adding to the data base studies that were not originally intended to be studies.

I have not formally scored the retrospective study as a flaw because of the difficulty of clearly drawing a line between it and a planned study. Two studies in the data base are clearly retrospective. Honorton (Study 7) constructed a retrospective study out of seven psi ganzfeld trials, each of which had originally been conducted as demonstrations for television film crews over a period of 1⅓ years. The justification for doing so was that, "these sessions involved the same procedures as [the] formal experimentation and included the same precautions against sensory leakage" (p. 185). But it does not matter how rigorously these demonstrations were carried out. The critic can justly point out that if the demonstrations had not resulted in significant psi-hitting, we probably would never have heard of them. After all, they were simply demonstrations. Nor is it clear at what point one stops collecting demonstrations and decides a sufficient number have accumulated to call the collection a "study."

The Child and Levi study (Study 4) also clearly qualifies as a retrospective study. These authors make it clear that this was not intended to be a formal replication:

> The instance of apparent psi-missing we report here is one that occurred with the ganzfeld procedure. The data did not emerge from systematic research, but from use of the ganzfeld procedure to demonstrate methods and perhaps outcomes of psi research in a college course. We have no way of immediately testing the replicability of the findings; indeed, anybody's attempt to do so will probably differ at least in involving a research setting rather than a completely educational one. But the outcome seems sufficiently striking to justify reporting it for its possible value in stimulating more definitive research. (pp. 279–280)

It is clear from what the authors write that they are reporting this classroom demonstration just because the results seem "sufficiently striking." And it is just as clear that if these results had not been significant they would not have been reported and would not thereby have become a member of this data base.

Strong circumstantial evidence exists to suggest that four others of the "significant" studies were also retrospective: Studies 2, 33, 34, 37. Study 2 was published almost 3 years after it was conducted. In marked contrast with the prototypical ganzfeld study, a single individual served as the experimenter and agent. In the other three, the authors referred to their studies as "preliminary," "exploratory," or "pilot." This again suggests that the only reason we are reading about them is because they gave significant results. In a few studies (e.g., Study 22), the author referred vaguely to a pilot study that presumably gave negative results and thus was not worth reporting. And the only explanation I can find to account for why so many of these studies exhibit the glaring flaws that will be discussed later on is that originally they were never planned as formal experiments.

Summary

Many different reasons strongly suggest that the actual rate of successful replication is much less than the 55% reported by Honorton. By counting as the unit the experimental conditions that replicate most closely the original ganzfeld study, I find an apparent replication rate of 31% for this data base. By taking into account the evidence for a selective bias to report only significant outcomes, the reasonable argument can be made that this success rate would be much lower if we could include the currently unknown ganzfeld psi studies. In addition, it is clear that at least some, and perhaps even

many, of the studies included in the current data base were not planned as formal experiments and have been given this status retrospectively just because they yielded significant results.

These considerations make it highly likely that the apparent rate of successful replications must be well below 30%. But even a success rate approaching 30% might be encouraging *if* the rate of success on the chance hypothesis is the advertised 5% level. The next section examines the possibility that the actual chance level might be much higher.

The Effective Error Rate

The task of the reviewer or investigator who is searching for the pattern or aggregate story in a body of literature would be fairly straightforward (a) if there were not the problem of which studies to include and exclude from the data base; (b) if each study used the same or overlapping independent variables; (c) if each used the same dependent variable; and (d) if each used the same planned test of significance. Under these circumstances, what Glass et al. (1981) refer to as the "primary analysis," "the secondary analysis," and the "meta-analysis" would be consistent with one another.

But when the studies in the same data base vary widely in independent and dependent variables and in the questions being asked by the original experimenters as opposed to those of the secondary analyst, then many confusing questions arise about what probability levels to assign to the various tests of significance. Such confusion is rampant in the attempts to find a coherent picture in the ganzfeld psi data base.

From the statistical inference viewpoint, the original Honorton and Harper study (Study 8) is appealing in its directness and simplicity. A total of 30 trials were carried out, each one on a separate subject and under identical conditions. If we can assume that the randomization of target and foils at the time of judging was properly carried out (unfortunately, there is some question about this), then the subjects' correct choices of the target, on the null hypothesis, would yield a binomial distribution with an expected value of 7.5 hits. Honorton's subjects achieved 13 direct hits. The probability of obtaining this many or more hits just by chance is .0216, which is considered to be significant. (If the subjects had obtained one less hit, then the probability would have been .0507, or just barely not significant.)

Braud, Wood, and Braud (Study 2) published the first replication

of the Honorton and Harper study. Both Braud et al. and the parapsychologists who have published vote counts consider the Braud study to be a successful replication of the Honorton and Harper ganzfeld psi study. Yet, in their ganzfeld condition, Braud et al. obtained only 3 direct hits out of 10 trials. Because the probability of obtaining 3 or more direct hits just by chance in 10 trials is .22, this outcome can by no stretch of the imagination be designated a successful replication.

So, how does it happen that Honorton and other parapsychologists treat this as a "success"? The answer is that Braud et al. used an alternative criterion for scoring "hits." They used binary (or partial) hits. At the end of the ganzfeld session, the subject was presented with the target picture and five foils to rank in terms of how well each picture matched the mentations during the ganzfeld period. A direct hit was scored if the target was ranked first. A binary hit was scored if the target was ranked in the top half (here in the top 3) of the set. Obviously, direct hits and binary hits have to be correlated because every direct hit is also a binary hit. But such a correlation is not perfect, and the two measures can give different results, as was the case in this study. In terms of binary hits, all 10 subjects succeeded, and this is highly significant ($p = .002$).

Honorton (personal communication, October 15, 1982) has defended counting the Braud et al. study as a successful replication on the grounds that "the Brauds have always used 'binary hits' as their ESP index." But this is not the issue. The point is that if the Braud et al. study is to be counted as a separate replication, it logically implies that the parapsychologists who accept this are guided by the following rule: A successful ganzfeld psi study is one that obtains results significant at the .05 level on *either* the number of direct hits *or* the number of binary hits.

Notice that such a rule implies that the actual significance level for achieving success must be greater than .05. If the two indices were independent, the effective significance level would be approximately .10. Because they are correlated, the actual level is somewhere between .05 and .10.

But this definition must be generalized because Honorton and the other vote-counters also accept as successful replications studies that achieve significance at the .05 level on at least one of the following alternative indices: direct hits, binary hits, sum of ranks, rating score, or binary coding.

This criterion obviously implies an effective error rate or chance level in excess of the advertised .05 level. Before considering just how

much in excess, we need to consider the use of one-tailed versus two-tailed tests of significance in this data base. The hypothesis in most of these studies, sometimes explicit but mostly implicit, is that psi-hitting will be obtained under ganzfeld conditions. When the deviations from chance are positive, the investigators seem to uniformly use one-tailed tests at the .05 level. When the deviations are negative, the experimenters do not hesitate to test for psi-missing. However, as far as I can tell, whenever the experimenter tests for psi-missing, a two-tailed test at the .05 level of significance is used (psi-missing is considered as a possibility in 24% of the studies in the data base). This implies that within this data base the following rule applies: If the number of observed hits exceeds the chance baseline, then test for "psi-hitting" with a one-tailed test at the .05 level of significance; but if the number is less than the chance level, then test for "psi-missing" with a two-tailed test at the .05 level.

Note that this rule also implies that the actual significance level is greater than .05. In fact, in this case it is .075. When we combine these two procedural rules, we find that in this data base, the vote-counters are in fact accepting as a successful replication any study that achieves a significant departure from the chance level at the .075 level on at least one of the five indices: direct hits, binary hits, sum of ranks, ratings, and binary coding.

A Simulation to Estimate the Error Rate

It would be easy to estimate the actual error rate for this rule if the indices were independent. But clearly the direct hits, binary hits, and sum of ranks must be intercorrelated. And it is reasonable to assume that these three indices will be highly correlated with the rating index. Because Palmer et al. (Study 11) report that they found no correlation between the binary coding and their rating score, I will assume that binary coding is relatively independent of the other four indices. By making a few reasonable assumptions, I was able to generate a set of simulated experiments to estimate the effective error rate.[2]

In the simulation, each experiment had the following characteristics:

1. The number of trials was fixed at 30 because the median number of trials per study in the current data base is 30.

2. For each trial, an integer from 1 to 4 was randomly generated. This represented the subject's ranking of the actual target for that trial

[2]Ron Friedland aided me with both the programming and the running of the computer simulations. The program was written in Pascal.

(assuming the target was randomly selected from a set of four candidates).

3. In addition, a standardized rating of the target was generated for each trial. This was done by generating four random integers from 0 to 100, standardizing these four to a mean of zero and a standard deviation of 1, and then selecting as the target rating that rating whose rank order matched the ranking assigned to the target for that trial.

For each study, I obtained four statistics: (a) the number of direct hits; (b) the number of binary hits; (c) the sum of ranks; and (d) the mean of the normalized ratings. Because three of the measures are discretely distributed, it was not possible to set the significance levels exactly to .075 (.05 for psi-hitting and .025 for psi-missing). The actual levels used turned out to be .0613 for direct hitting, .0708 for binary hitting, .064 for sum of ranks, and .075 for ratings.

Two separate batches of 1,000 simulated studies were conducted. Each batch yielded the following data: (a) the intercorrelations among the statistical indices; (b) the number of "significant" outcomes on each of the four indices; (c) the number of studies with at least one significant statistic.

If the four indices were independent, we would expect the probability of obtaining at least one significant outcome per experiment to be .24. In fact, the intercorrelations among these four indices ranged from approximately .62 between direct hits and binary hits to .95 between sum of ranks and ratings. And the actual probability of achieving at least one significant outcome per study on these four measures is approximately .152. Assuming that the binary coding index is uncorrelated with these other four indices, I estimate that the probability of achieving a significant outcome on at least one of these measures is approximately .22.

In other words, by using the procedural rule that appears to be followed by the vote-counters on this data base, we obtain an effective level of significance of .22, or over four times the assumed level of .05.

At this point, the objection might be raised that it is unrealistic to charge each study with the usage of five different indices. The binary coding system, for example, can be used only with the special set of slides created by Honorton and his coworkers; and, in fact, only a handful of studies actually used it. And in some studies, they used it in such a way that it was clearly the only possible index they could compute. In still other studies, the data were collected in such a way as to preclude either the calculation of a rating score, or a sum of ranks, or both.

Such an objection has merit, but it also raises again the problem of assigning error rates on the basis of the primary analysis conducted by

the original investigator, the secondary analysis conducted by a reviewer, and the meta-analysis produced by combining the results of several studies on some assumed common metric. The procedural rules used by the vote-counters of the ganzfeld psi studies, accepting as they do the actual index used by the original investigator (rather than attempting to reanalyze the results, where possible, to put them on a common footing), logically imply this .22 error rate.

Multiple Indices

The problem raised by multiple indices is just one of a number of ways in which multiple testing can occur. I shall mention some of the other ways shortly. However, the different forms of multiple testing intertwine, and it is difficult to consider them in isolation. Also, as I have just indicated in the preceding paragraph, it is difficult to keep separate the error rate for the testing by the original investigator from the error rate for the subsequent uses of the original study by others. In the case of multiple indices, 55% of the original experimenters actually used two or more such indices, and no one actually used more than three. It is possible to assign a separate error rate, based on the simulations, to each study and then to compute an average error rate per study in terms of the criteria used by the original investigator. This error rate works out to approximately .10 per study.

It is important to keep in mind that this error rate applies only to the usage of two or more of the five common indices. As we will see, other factors contribute to inflating the actual error rate well beyond this level. But even if we consider an error rate based only on the use of multiple indices, this estimate is probably too low. Rarely did an investigator make it clear that he had decided on his primary index or test prior to conducting the study. And it is only in those studies in which the authors have given us sufficient documentation that can we be sure we know all the indices that were considered and perhaps later discarded. For some of the studies that were published only as an abstract in *Research in Parapsychology*, Honorton supplied me with a longer, unpublished version. In one or two of these, the longer report contained information that was missing from the published, shorter version. For example, if I had to rely only on the report in York (Study 42), I would not have assigned it a flaw in the category of "multiple indices" because the published report is written as if only direct hits were scored and tested. The longer, unpublished account, however, makes it clear that the primary measure was intended to be a rating score and that the direct hits were intended to be only a secondary

index. It is possible that many of the other reports have not fully reported all the indices they actually tried or would have tried had their original index not worked out.

Some other indications in the current data base seem to support this last speculation. For example, Honorton and his colleagues have uniformly used either direct hits or binary coding in their studies (Studies 7, 8, 32). However, Terry and Honorton (Study 38) inexplicably departed from past practice and used binary hits as their primary index. This usage seems especially peculiar because their customary measure, direct hits, did not achieve significance in this study; and in their very next study (Study 39), described in the same report, the authors reverted to direct hits as the primary measure, with no further mention of binary hits. Even though Honorton (1979) later tried to justify this behavior, it cannot help but reinforce suspicions that even more potential multiple testing is going on than appears on the surface.

Honorton's sudden switch of dependent variable is also inconsistent with his own professed standards. For example, Braud and Wood (Study 3) used both binary coding and their customary binary hits. In a previous study involving the same two investigators (Study 2), Honorton justified using the significant outcome on binary hits rather than the nonsignificant outcome on direct hits to classify the study because the "Brauds have always used 'binary hits' as their primary measure." In the second study, Braud and Wood failed to find significance with their customary binary hits, but did find it (in one of 12 ganzfeld conditions) on the binary coding measure. Consistency would dictate classifying this study as nonsignificant. But Honorton classifies it, without explanation, as significant. Not only is such inconsistency perplexing to the outsider, but it also greatly strengthens the suspicion that the multiple options that are due to the availability of several indices of psi are underestimated by my assignments in the Appendix.

Considerations such as these suggest that the error rate per study in this data base, on the basis of the use of multiple indices, is at least .10 and almost certainly higher. But the use of multiple indices is just one of several ways that multiple testing occurs in this data base.

Multiple indices was but one of six formal categories of multiple testing on which I judged the studies in the data base. As indicated, over half the studies clearly used multiple indices without taking this into account in computing their statistical significance. A flaw was also assigned on one of the following categories of multiple testing, if it occurred without corresponding adjustments in the significance level.

Alternative tests. In five studies (12% of the total), the experimenters used more than one statistical test on the same index. For example, Raburn (Studies 16, 17) applied both Fisher's exact test and the chi-square test (which is an approximation to the exact test) to the direct hits in her studies. The results of such tests will be highly intercorrelated, but they do increase the actual level of significance above the assumed .05 level. In a few cases, the assignment to this category could just as well have been made to the category "groupings."

Multiple baselines. For the most part, the investigators in this data base tested their index of psi against a theoretical baseline. But many also included control groups of various kinds and also tested the same index against the control comparison. Such use of multiple baselines occurred in 17 (40%) of the 42 studies. As one example, Terry et al. (Study 40) used both a ganzfeld psi condition and a control (guessing) condition. The number of binary coding hits for each of these conditions was tested separately against the chance baseline, with significance being obtained for the ganzfeld but not for the guessing condition. The numbers of hits for each condition were also tested for significance against one another, with a nonsignificant outcome. Although this is counted as a significant outcome, one can just as logically argue that if the hitting in the ganzfeld condition is to be attributed to psi we should demand that it be significantly greater than a control condition in which the hitting is apparently due entirely to guessing. For our purposes, however, the point is that having the option of testing for psi by testing an index against both a theoretical and an empirical baseline effectively increases the true significance level beyond the .05 level used for each test separately.

Multiple dependent variables. Only three (7%) of the studies were assigned this flaw. In retrospect, because of the small frequency of occurrence, it might have been better to absorb this category into one of the others. As one example, Stanford (Study 35) had targets matched not only against the entire transcript but also against only the second half of the transcript. Again this increases the number of options for obtaining a significant result.

Multiple groupings. This cause of multiple testing was the most frequent one in this data base (27, or 64% of the studies). The possibility for this form of multiple testing exists whenever the study has more than one condition. Smith, Tremmel, and Honorton (Study 32), as one example, tested their two conditions separately as well as pooled.

Independent judges. The typical ganzfeld psi study (as contrasted with the typical remote-viewing experiment, for example) uses the

percipients as their own judges. Some of the studies used independent judges instead. The use of independent judges was scored as a flaw here only if they were used in addition to subjects acting as their own judges in such a way as to produce unadjusted multiple testing. In this data base, 5 (12%) of the studies exhibited this flaw.

No significant differences on any of these six categories of multiple testing exist between the studies classified as "significant" and those classified as "nonsignificant." Nor can I think of any good reasons to expect such differences. One could imagine a scenario in which some investigators habitually indulge in multiple testing whereas others rigorously avoid it. We would then expect to find a greater proportion of multiple-testing flaws among the significant studies. However, we could just as easily imagine a scenario in which the typical investigator uses one option at a time until he either obtains a significant outcome or exhausts all the available options. In the latter scenario, we would predict a greater proportion of multiple-testing flaws in the nonsignificant studies. And, in still a third scenario, we might suppose that the data base contains a mix of the first two scenarios: In this case, one would not predict a difference among the significant and nonsignificant studies on these flaws.

The True Error Rate

The various flaws attributable to multiple testing in this data base and the arguments made in this section suggest that the true error rate is much higher than the assumed .05 level. The question is, how much higher? Unfortunately, we can only make some crude guesses at this time. One problem involves how to reconcile the probable error rate under which the individual experimenter was operating with the error rate that the secondary or meta-analyst is using. The typical experimenter in this data base seems to be operating at the .10 level of significance if we consider just the problem of multiple indices. But the vote-counters, who are trying to find a consistent pattern in the total set of experimenters, are operating at the .22 level or higher if we consider their criteria for accepting individual studies as successful replications of one another.

These two estimates, made on the basis of computer simulations, use just one of the six ways that multiple testing occurs in this data base. How much more should these estimates be enlarged to take into account the additional five flaws? Perhaps some additional computer simulations based on various plausible assumptions could set some upper and lower bounds on these figures. And perhaps I or some

other interested investigators will conduct such investigations in the future. But I think it is important to emphasize that the outcome of such investigations is likely to be surprising and somewhat counterintuitive to many investigators.

I shall use an example to indicate why I believe that the actual effective error rate is probably much higher than anyone has previously suggested. Consider the Braud and Wood study (Study 3); and to keep matters simple, I shall omit the four practice sessions through which they conducted their two groups. They ran two independent groups of subjects through pretreatment and posttreatment ganzfeld psi sessions. They used two indices of psi: the number of binary hits, and the number of hits on the binary coding scale. The test for psi was made within each of the four conditions, the two pretreatment and the two posttreatment sessions. The test on each index can be considered to be at the .075 level for reasons already indicated (there seems little doubt that the typical experimenter will test for psi-missing if the number of hits is substantially less than chance). In addition, these two indices, binary coding and binary hitting, are apparently uncorrelated. This means that there is a probability of .144 of obtaining at least one significant result in a given condition just by chance. Furthermore, under the reasonable assumption that the tests in each of the four cells are independent, the probability that significance will be found in at least one of the four conditions becomes .464.

But this estimate of the effective significance level in this case is still too low. The experimenter also has the option of pooling the data over the pretreatment and posttreatment trials within each group (indeed, this was actually done). Although these new tests are not independent of the previous ones, they further increase the effective error rate, almost surely beyond .50. We still have to consider the option, used by Honorton, of pooling the data over both conditions and over all trials to make a further statistical test. In addition, we have the options, actually used by Braud and Wood, of testing the conditions not only against the chance baseline, but also against one another. And for simplicity, I omitted eight treatment conditions that could also be considered replications of the ganzfeld study.

In other words, this one study can be considered to be operating under a true significance level of well over .50. Indeed, if we consider the eight intervening practice conditions, the chances of coming up with a significant outcome are well over .80! And this is just one of many studies in this data base that exhibit such complex options, either explicitly or implicitly.

One possible protest against my treatment of the Braud and Wood study might be that the logic of the study suggests that they were predicting a significant outcome mainly in the posttreatment condition, which had been preceded by practice sessions with feedback. But this would be a feeble protest for a number of reasons. Significance was obtained only with the binary-coding index. The binary-hitting index, which Braud had consistently relied on in the past, was insignificant. No hint was given in the write-up that Braud and Wood had suddenly found grounds for predicting success on binary coding and not on their favored measure. At any rate, the outcome again raises questions about what constitutes a successful replication.

What if significance had been obtained in one of the pretrial conditions rather than one of the posttreatment conditions? Would Braud and Wood or Honorton have considered this a failure to replicate? This is highly unlikely when it is realized that the pretreatment conditions are closer replicas of the original Honorton-Harper study than either of the posttreatment conditions. And what if the posttreatment condition without the preceding feedback practice had been the only significant condition? Again, this surely would have been considered a successful replication since it is a closer replica of the original ganzfeld psi study than is the one condition for which significance was claimed.

Summary

In the paper I presented to the combined meetings of the Society for Psychical Research and the Parapsychological Association in August 1982, I estimated that the effective error rate was closer to .25 than to .05. Honorton attacked this estimate as based on subjective speculation. He also suggested it was based on a worst-case scenario. Any estimate of the effective error rate is, of course, speculative. But I believe the arguments I have made in this section make a strong case that the overall effective error rate per study could easily be this high or higher. When taken together with the arguments in the preceding section on the actual rate of success, the claims made for the ganzfeld psi data base have been premature, to say the least. Many considerations indicate that the actual rate of successful replication is less than 30%. And the arguments in this section strongly suggest that this rate of "successful" replication is probably very close to what should be expected by chance given the various options for multiple testing exhibited in this data base.

Procedural Flaws

As parapsychologists as well as their critics have frequently remarked, the evidence for psi consists of statistically significant deviations from a chance baseline. Presumably, parapsychologists, like critics, would agree that before such deviations are interpreted as being due to psi, certain elementary safeguards must be met. These safeguards are as follows: (a) randomization of targets and conditions in such a way as to guarantee, on the chance hypothesis, that the resulting distribution of hits and misses will be consistent with the assumptions underlying the statistical tests; (b) given the underlying assumptions, the use of appropriate statistical tests in such a way as to guarantee that the assumed error rate is, in fact, the actual one; and (c) use of experimental controls to eliminate obvious possibilities for sensory leakage.

The last section listed several reasons for concluding that insufficient attention had been given to the second safeguard. At most, 3 of the 42 studies in the data base were entirely free of multiple-testing flaws. I did not try to assign a flaw for an additional aspect of this safeguard—the need to insure that sampling units are independent. The potential violation of independence was extremely widespread in the data base. The most common violation was the indiscriminate pooling over the within- and between-subject trials. But pooling also took place over separate experimental conditions without any attempt to segregate within- and between-condition variance. And a very common practice was to include within a single experimental condition trials in which agents were friends of the percipient along with trials in which agents were members of the laboratory staff. Just how serious such violations of independence are is difficult to decide. One can imagine possible models in which they make no difference. But all such models assume that randomization has been optimal and that, on the null hypothesis, no psi exists.

Procedural Categories

The first and third safeguards entail what I will designate as procedural safeguards. These involve conducting the study in such a way as to insure proper randomization and to eliminate obvious possibilities for sensory communication between target and percipient. To these safeguards, I have added two other components: reporting the results in such a way that the reader can tell if the safeguards have been used, and conducting the statistical analyses correctly. I assigned

flaws to the studies (see Appendix) if they were deficient on any of the following six procedural categories:

Randomization (R). Typically, the ganzfeld psi study uses a number of target pools. Each pool, in turn, contains, say, four pictures or slides that are candidates for the target on a given trial. On a given trial, a pool is selected. Then, a target within that pool is chosen. Later, all the members of the pool are given to the subject or the judge for evaluation. Randomization refers to the procedure for making the selection at the first two stages. (I include problems of replacing the target among the other members of its pool for judging under the category Feedback.) The most critical aspect of the randomization procedure is the selection of the target from its pool. When the experimenter reported using an inadequate measure of randomization, such as hand-shuffling of cards or reels, or tossing coins, I assigned the study an "R–." I also assigned this flaw to studies in which no randomization at all was used for selecting the target. For example, in Studies 7, 8, 38, and 39, the experimenter simply took the uppermost reel in a packet of four slide reels as the target. Altogether, 15 (36%) studies were assigned this flaw. When the experimenter reported using a table of random numbers or a random number generator to select the specific target from a pool, I assigned the study an "R+." Only 11 (26%) of the studies met this minimal standard. The remaining 16 studies were assigned an "R?" to indicate that they supplied insufficient information about how they were randomized. Because adequate randomization is so basic, one might assume that it can be taken for granted that an investigator has properly carried it out. So the failure to fully describe how the target was selected perhaps ought not to be tallied as a "flaw." But the fact that in those studies in which the target selection is fully described, 58% use a clearly inadequate method of randomization suggests that it would be unwise to assume that randomization was adequate in the questionable cases. For purposes of subsequent analysis, then, I have assumed that randomization was suboptimal in all those studies (74% of the total) in which it is not clear that it was conducted adequately. Fortunately, all of the correlations with this index come out in the same direction regardless of whether I use this stringent criterion or simply use only the studies that clearly describe their procedures. Only the studies involving Stanford (Studies 31, 35, 36) seem to treat the problem of randomization as something to be taken seriously.

Single target (ST). Obviously, the use of the same target that has been separated from its pool and later replaced for judging purposes allows various possibilities for sensory leakage. At least, it was obvious

to a critic such as myself when I first encountered the Honorton and Harper study (Study 8) (Hyman, 1977). Yet, it was not until 1980 that this flaw disappeared from ganzfeld psi studies. I assigned the flaw "ST" to each of the 23 studies (55% of the total) that used a single target.

Feedback (FB). The 10 studies to which I assigned this flaw, in addition to possible handling cues because of the use of a single target, typically did not use an adequate procedure to insure that the target was properly randomized among the other candidates in the pool before being presented for judging. Although only 24% of the total data base exhibited this flaw, the flaw could occur only in those cases in which a single target was used. Of these cases, 43% exhibited this flaw.

Documentation (DOC). This could refer to inadequate reporting of many critical details needed for assessing the adequacy of the procedures. But most of the assignments of this flaw had to do with failure to report the number of times the agent was a friend of the percipient or to provide data on whether this made a difference in those studies in which subjects were encouraged to bring their own agents. As might be expected, this flaw was much more prevalent in the unpublished studies. Inadequate documentation was a serious problem in 81% of the unpublished studies as opposed to 38% of the published studies.

Security (SEC). The prototypical ganzfeld psi study corresponds to what Rhine and Pratt (1957) refer to as "the two-experimenter plan." With one experimenter monitoring the agent and one monitoring the subject, there is increased security against a variety of potential threats to the validity of the study. But some of the studies depart from this safeguard. In the study by Braud et al. (Study 2), a single person, an undergraduate student, plays the role of both experimenters as well as the agent—roles that are enacted by three different individuals in the typical study. A single person also plays this part in Studies 19 and 41. Such departures from the prototype lessen security. I assigned SEC to studies for other reasons also, such as failing to monitor the agent or rolling a clay ball over the target, etc. A total of 10 (24%) studies were assigned this flaw.

Statistics (STAT). I assigned 12 (29%) studies the flaw STAT because of what appeared to be an erroneous use of the statistical procedure. Apparently some investigators used Fisher's exact test without adding in the probabilities of getting results even more extreme than the actual outcome (Studies 16, 17, 31, 33). Such a mistake can greatly exaggerate the actual significance. In one study,

this error was compounded because a two-tailed test had clearly been intended. The other studies involved inappropriate pooling over trials and using wrong degrees of freedom. As mentioned previously, inappropriate pooling and violations of independence were quite common in this data base. But I assigned a statistical error only in the most blatant violations—those in which the correct alternative should have been applied.

Additional Problems

The assignment of flaws according to the preceding six categories was conservative on a number of grounds. The occurrence of some flaws, for example, is contingent on other circumstances. For example, the FB flaw could occur only in studies that also have the ST flaw. And in some of the early studies, the absence of this flaw did not mean that the experimenter had taken the appropriate precautions, but rather that the procedure—the use of the binary coding method—made it irrelevant. Most of the statistical flaws occurred because of inappropriate handling of between- and within-subject trials. Such an error cannot occur in a simpler study that has a single condition and no repeated measures. Thus, the absence of some statistical flaws may not indicate that the experimenter was statistically sophisticated but rather that the simplicity of the design precluded his committing certain types of mistake.

I did not systematically score a variety of flaws because they either depended on suspicions or hard-to-objectify criteria, or were not too common. I have already mentioned the "retrospective study," which can be charged to at least three studies but probably includes several more. Another problem was the fact that several studies used inconstant conditions. One changed the procedure halfway through the study, but did not include this change as a variable. Several studies allowed percipients to bring their own agents, but supplied an agent for those who did not. These potentially different conditions were typically not analyzed separately.

In any case, the existence of so many elementary defects in this data base is both disturbing and surprising. Only two studies were entirely free of the six procedural flaws. And if we include multiple-testing errors, not a single study in this data base was flawless. It is important to realize that the defects being discussed are not obscure or subtle. Rather, I suspect that a typical parapsychologist would spontaneously list them as being unacceptable in a psi experiment. Given the central role of statistical assumptions, it is distressing to

discover that only 26% of the studies in this data base clearly used appropriate randomization. And, further, it should be of little comfort to find that 29% made statistical blunders.

I would like to emphasize that I do not view the existence of these flaws as causal in the sense that their presence accounts for the significant results. Rather, I see them as symptoms. I know, and have a great deal of respect for, the experimental competence of many of the investigators in this data base. I have little doubt that most of them know full well how to conduct a planned and well-controlled study. I believe that just about all of them would agree that the use of random numbers or random generators is superior to hand-shuffling of cards. Yet, despite the universal availability of such optimal procedures, several experiments use suboptimal procedures.

Meta-Analysis of Flaws and Successful Outcomes

The surprising number of defects in the ganzfeld psi studies have been pointed out by critics within the field of parapsychology (Akers, 1984; Ballard, unpublished; Kennedy, 1979a, 1979b; Parker & Wiklund, unpublished; Sargent, 1980a, 1980b). However, I have not heard any reasons offered for the occurrence of these defects. Instead, attempts have been made to dismiss the criticisms on the grounds that the flaws, in fact, make no difference. Characteristic of this approach to minimizing the defects is that of Honorton (1979, 1981).

Honorton argues that if a critic wishes to fault a study because of the possibility of handling cues, for example, then he is obliged to demonstrate empirically that such cues do, in fact, make a difference. Honorton deals specifically with the flaw I have called Single Target (ST), which allows the possibility of handling cues. Honorton first reviews some studies that suggest that even when deliberately introduced, subliminal or sensory cues are rarely exploited. Then he demonstrates that, if anything, there is a slight tendency for ganzfeld studies that do not have this flaw to produce higher Z scores on overall hits.

My analysis agrees with Honorton in showing no correlation between the use of single targets and significance. In the current data base, 52% of the significant studies were assigned ST as compared with 58% of the nonsignificant studies. But this empirical relationship hardly justifies retrospectively sanctifying studies that committed this blunder. Honorton (1981) concludes:

> Since there was no difference in the results of the two groups of studies, the handling cue hypothesis was rejected. As John Stuart Mill put it, "A difference, in order to be a difference, must make a difference." . . . The moral here is simply this: Disputes over empirical claims can only be resolved through empirical methods. This is the hallmark of science and what differentiates it from other approaches to knowledge such as religion. (p. 159)

But such a defense will not do. Even if a first-order correlation is zero between a flaw and significance, this does not mean that no relationship exists or that a causal connection is absent. Much can depend on how the flaws intercorrelate with each other and other variables. For example, if, when all other factors are controlled, ST correlates positively with significance and we also have a second flaw that not only correlates highly with significance but also correlates negatively with ST, then we could easily find that the first-order correlation between ST and significance is zero or even negative. This is just one of the many problems of trying to use statistics to substitute for empirical controls.

In addition, correlation deals with symmetrical relationships and overlooks more complicated possibilities. For example, ST and FB are intercorrelated in this sample, but the relationship is asymmetric. Because of the way I scored it, FB can occur only if ST has also occurred. The probability of FB when ST is absent is zero, but is approximately .43 in this sample, given that ST is present. Furthermore, FB *does* seem to be correlated with significance. Of the significant studies that have ST, eight (or 67%) were assigned FB. Of the nonsignificant studies that were assigned ST, two (or 18%) were also assigned FB. In other words, even though, by Honorton's criterion, ST does not correlate with significance, its presence enables the occurrence of another flaw that does seem to correlate with significance.

In addition to the previous two considerations, as already indicated, such flaws are signs that the study was probably not carefully planned or properly carried out.

Dependent Variables for the Meta-Analysis

Although the retrospective correlations between flaws and experimental outcomes cannot be used to salvage improperly executed studies, it still may be helpful to follow Honorton's lead and, in the spirit of a meta-analysis (Glass et al., 1981), examine the pattern of

relationships among indices of success and various flaws. Such an examination should be taken in the spirit of exploratory data analysis, and its goal is to suggest hypotheses about what may be going on. For the purposes of this analysis, I used the following dependent variables:

1. Honorton's original classification as significant or nonsignificant. Using Honorton's original criterion (personal communication, November 30, 1981), I assigned the notation SIG-1 to studies he deemed to be significant overall and NSIG-1 to the remaining studies (see Appendix for these assignments).

2. Honorton's second classification after adjusting for multiple testing. Here I used Honorton's classification of June 2, 1982 (personal communication) to assign the notation SIG-2 to those studies that still achieved significance at the .05 level after adjustment of significance levels for multiple testing. The remaining studies were assigned NSIG-2.

3. Effect size in degrees (E). One problem with Honorton's classifications of significance is that he accepts the original investigator's primary index even though the index varies from experimenter to experimenter. This not only logically commits Honorton and the other reviewers who keep box scores in this manner to an absurdly high error rate, but it also violates the spirit of meta-analysis, which seeks to find a common index or scale by which to compare different studies. In addition, vote counting based on significance ignores the fact that different studies vary in number of trials (and, hence, power). Consequently, meta-analysis tries to focus on a common index of effect size rather than level of significance.

I needed an index that could be used to characterize most of the studies in this data base. The index I chose was the number of direct hits. When it was impossible to obtain the number of direct hits for a given study, I used binary hits instead. In a few cases, I used the number of hits on the binary coding scale when nothing else was available. I omitted six studies from this meta-analysis because they did not supply data that would enable me to compute an index comparable to the number of direct hits.

To convert the hits into a reasonably comparable measure of effect size, I used the Freeman-Tukey arc sine transformation for binomial proportions (Freeman & Tukey, 1950) on both the number of hits and the expected number of hits. The effect size was then the difference between these two transforms in degrees. In the Appendix, when BH or BC appears in parentheses after the effect size, it indicates that it was computed on the basis of binary hits or binary coding, respectively.

If no such designation is given, then the effect size was calculated on the basis of direct hits.

Z score. The Freeman-Tukey transform supplies, as a side benefit, a theoretical standard deviation for each sample size. This was used to obtain a *Z* score for each effect size. The *Z* score can be referred to the tables of the normal curve for computing significance.

Absolute effect size and Z score. Finally, in many of the analyses, the absolute effect size and *Z* score were also used because graphic examination seemed to indicate that on some comparisons—say, effect size with sample size—the relationship was in terms of variance rather than algebraic mean.

Independent Variables

The 12 formal categories of flaws served as the major source of independent variables for finding correlates of effect size and significance. In addition, other variables were studied, such as number of trials, published or unpublished results, investigator, and year of report.

Because the 12 flaws are scored only as dichotomies, the first step was to find some rational or empirical means of combining the flaws into relatively continuous scales. For this purpose, I discarded three of the categories—Alternative Tests, Multiple Dependent Variables, and Independent Judges—because their frequency of occurrence was too low. I then obtained the intercorrelations (point-biserial *r*) between each pair of the remaining nine flaws and carried out both a factor analysis and a cluster analysis on the resulting matrix. Fortunately, both the cluster analysis and the factor analysis agreed in partitioning the nine flaw measures into three overlapping factors or clusters. For each cluster, I obtained the first principal component and used this as a basis for obtaining regression weights to use for combining the indices for each cluster into a single composite cluster score. On the basis of the weights, I gave each of the cluster scores a title:

Cluster I: "General security." This cluster was composed of five indices, but the major contributors were ST, FB, and SEC. These all seemed to be related to problems of security or possibilities of sensory leakage. The actual indices and their weights were:

$$\text{C-I} = -.33\text{MI} + .52\text{ST} + .54\text{FB} + .47\text{SEC} + .31\text{STAT}$$

Cluster II: "Statistics." This cluster also consisted of five indices, with the three strongest contributors being Multiple Baselines, Multi-

ple Groupings, and Statistics. All these seem to be flaws that contribute to an inflated level of significance. The indices and their weights were:

$$\text{C-II} = .42\text{MB} + .48\text{MG} - .41\text{MI} - .41\text{R} + .51\text{STAT}$$

Cluster III: "Controls." This cluster consisted of four indices, with the strongest contributions coming from Randomization, Feedback, and Documentation. The simplest word I could find to characterize this dimension was *Controls*. The common aspect seems to be a lack of care in carrying out various procedural controls. The indices and their weights were:

$$\text{C-III} = .16\text{MI} + .64\text{R} + .48\text{FB} + .57\text{DOC}$$

Correlations Between Flaws and Significance

Of the three cluster scores, only Controls correlates significantly with both the original and adjusted classifications of significance. The group of studies that were significant in both classifications had the highest mean score on this index of flaws. The group that was significant on the original but not on the revised classification had the second highest mean score. And the group that was significant on neither classification had the lowest score on this cluster. This finding is confirmed by the meta-analysis. The correlation between effect size and the Controls cluster was .37, and that between Z score and the Controls cluster was .44 (both significant).

The individual flaws that seem to correlate (using conventional significance levels as a convenient criterion) with the three different measures of significance are Randomization, Feedback, Documentation, and Statistics. The more likely a study was to be assigned any of these flaws, the more likely it was to be classified as significant. The same pattern, but somewhat weaker, is to be found in the case of effect size. In the case of this latter index, the two best single predictors are Feedback and Documentation. Interestingly enough, the fewer the trials in an experiment, the *larger* the effect size!

Experimenter Effects

Parker (1978) has written, "The present crisis in parapsychology is that there appear to be few if any findings which are independent of the experimenter. Indeed, it can be claimed that the experimenter effect is parapsychology's one and only finding." Because several experimenters appeared in more than one study in the current data base, some of my analyses examined the relationship of variables such as flaws, effect size, and significance to experimenter. In this data

base, experimenters tend to be quite consistent in their experimental designs from study to study. Indeed, most of the variation and covariation between studies is due to differences in experimenters. An analysis of variance on effect size with investigators as the independent variable yielded a significant outcome. Seven investigators who had two or more studies in the meta-analysis were included as levels. An additional level was constructed by combining the eight remaining studies in which the experimenters appeared only once. For the group of 36 studies as a whole, the average effect size was 5.98 degrees, which corresponds to a direct hit rate of 35% (as compared with the chance rate of 25%). However, four experimenters who contributed 18 studies, or 50% of the total data base, accounted for almost all of this effect size. These four experimenters (Honorton, Sargent, Sondow, and Raburn) reported results whose average effect size was 11.31 degrees, which corresponds to a direct hit rate of 44%. The remaining studies yielded an average effect size of 0.76 degrees, which corresponds to a hit rate of 26%.

Other analyses confirm that an experimenter effect does exist in these studies. Indeed, two experimenters account for much of the apparent success of this paradigm. Honorton and his coworkers dominate the successes in the first 4 years of the studies. In the past few years, Sargent and his coworkers have carried the burden. Strangely, no contributions have come from Honorton and his laboratory during the latter 4 years of the span covered by the present data base.

The experimenter effect is confounded with the fact that the experimenters also differ significantly in their patterns of flaws. Honorton's studies, which have the highest average effect size, also have the highest score on the Controls flaw cluster. Palmer's studies, which have the lowest average effect size (in fact, slightly negative), also have the lowest score on the Controls flaw factor. Therefore, it would be premature, to say the least, to look for the source of the experimenter effect in either the personality of the investigator or the social context in which he operates. The first place to look, it would seem, is in the differences in the way the investigators carry out their studies, such as use of a single condition or multiple conditions, care in procedural details, emphasis on security, and the like.

Factor Analysis of Variables

Several other relationships were examined. Because most of the findings seem to be captured by a factor analysis I conducted on

several variables, I shall use the results of that analysis to summarize my findings from the meta-analysis. The factor analysis used 17 variables—the three cluster flaw scores, the logarithm of the number of trials, the year of the report, whether repeated measures were used, the five investigators with the largest number of studies in the data base (Honorton, Sargent, Rogo, Palmer, and W. G. Braud), effect size, *Z* score, the absolute values of the latter two indices, and a dummy variate to take into account that effect sizes on binary hits and binary coding tend to be lower than those for direct hits.

Four factors were extracted (using principal components and the varimax method of rotation). One factor was characterized by a high positive-loading (0.85) on year and a high negative-loading on the General Security cluster (–0.88). This reflects the fact that security, as measured by the increasing use of duplicate target sets, has been increasing in the past few years. Indeed, the possibility of sensory leakage owing to the use of a single target-set has now disappeared. A second factor has high loadings on all the measures of effect size and significance level along with a positive loading (0.49) for the statistics cluster. The two indices with highest loading on this factor are absolute effect size (0.74) and absolute value of *Z* score (0.82). A third factor involves the number of trials. The major loadings on this third factor are logarithm of trials (0.91), cluster score on Statistics (0.53), and repeated measures (0.57). Part of this is consistent with other analyses that strongly suggest that statistical flaws that unwittingly inflate the significance level tend to increase with the complexity of the study. (Complexity is measured here by number of trials and use of repeated measures.)

But it is the fourth factor that holds most interest in the present context. The key loadings are effect size (0.55), *Z* score (0.59), and cluster score on Controls (0.81). In addition, this factor contrasts the experimenters Honorton (0.54) and Palmer (–0.71), This factor is simply consistent with the other analyses that indicate that both effect size and significance are correlated with the existence of flaws which indicate insufficient care with respect to experimental procedures. And it is also consistent with the prior findings that experimenters who pay the most attention to such controls also report the smallest effects.

Predicted Significance and Effects for Flawless Studies

The average *Z* score for this data base is 1.04. The composite *Z* score (using the method suggested by Rosenthal, 1978) is 6.27, which would suggest that for all practical purposes the possibility of getting

such a combined result by chance is zero. In addition, the average effect size is 5.98, which corresponds to a direct hit rate of 34%. Another way to emphasize the magnitude of this composite picture, one favored by Honorton, is to use Rosenthal's (1979) procedure for dealing with the "file-drawer problem." The idea is to use the composite *Z* score to estimate the number of unknown or unreported studies with nonsignificant results that would have to exist to make the present 36 studies simply a selection from a population with zero effects. In this instance, the procedure informs us that it would take 486 such studies. Such considerations seem to indicate that the data base provides a robust and compelling argument for the existence of psi.

But, as we have seen, both *Z* score and effect size, in this data base, correlate with the cluster score on Controls. The three flaws that contribute most to this score are Randomization (R), Feedback (FB), and Documentation (DOC). These are scored as dichotomies, a "1" indicating the presence of the given flaw and a "0" indicating its absence. A regression equation relating these dummy variates to both *Z* score and effect size separately was computed to predict what the corresponding *Z* scores and effect sizes might be for studies that were free of these flaws. A dummy variate for *Z* scores and effect sizes that were not based on direct hits was also included in the equations to take into account the fact that binary coding and binary hits tended to give smaller effect sizes in this data base.

The regression equation for the *Z* score gave the following weights:

$$Z' = 0.03 + 0.74\text{R} + 0.28\text{FB} + 0.91\text{DOC} - 1.27\text{D(BH/BC)}$$

with a multiple correlation of .53. The corresponding equation for effect size was:

$$E' = 1.52 + 2.55\text{R} + 3.68\text{FB} + 4.50\text{DOC} - 7.40\text{D(BH/BC)}$$

with a multiple correlation of 0.48.

These equations should cause us to reconsider the seemingly impressive implications of the composite *Z* score and the estimates based on the file-drawer problem. The first equation informs us that for studies in which the flaws R, FB, and DOC are eliminated, we can expect the *Z* score based on direct hits to be zero. The second equation, when properly translated, tells the same story. It predicts an effect size of 1.52 degrees for experiments that have none of these three flaws. Such an effect size corresponds, in terms of direct hits, to a hit rate of 27%, which is well within the statistical neighborhood of the 25% chance rate.

Published versus Unpublished Studies

About half of the studies in the data base actually were published. The rest appeared as abstracts or papers presented at meetings of the Parapsychological Association. Unlike the findings of meta-analyses in other bodies of research literature (Glass et al., 1981), no difference in effect size or other variables, with one exception, between published and unpublished studies showed up in this sample. The one exception was documentation. As would be expected, unpublished studies were much more likely to lack adequate documentation than published ones.

Conclusions

By now it is clear that I believe that the ganzfeld psi data base, despite initial impressions, is inadequate either to support the contention of a repeatable study or to demonstrate the reality of psi. Whatever other value these studies may have for the parapsychological community, they have too many weaknesses to serve as the basis for confronting the rest of the scientific community. Indeed, parapsychologists may be doing themselves and their cause a disservice by attempting to use these studies as examples of the current state of their field.

I have no doubt that most of the investigators who have contributed to this data base are fully capable of conducting well-planned and relatively flawless studies. Whatever the reasons, the 42 studies in the present data base cannot by any stretch of the imagination be characterized as flawless, and I suspect that most of them were not well planned. If these investigators wish to impress outsiders and critics with their efforts, they will have to present them with studies that have reasonable controls against sensory leakage, adequate procedures to insure that the underlying statistical assumptions have been met, and evidence that the advertised error rate is, in fact, the actual one. If a body of such studies can be carried out, and the results come out as successful as many parapsychologists believed the ganzfeld psi studies were, then the time will have come for the scientific community to sit up and take notice.

Exploratory versus Confirmatory Studies

One problem with the current data base seems to be a confusion between exploratory and confirmatory studies. In fact, many of the

authors explicitly announce their studies as exploratory. In responding to Sargent's (1980b) critique of her study, Sondow (1980) points out, "Finally, I must point out that asking more than one question in an exploratory study is hardly a 'shortcoming'" (p. 272). Indeed, asking many questions in carrying out exploratory data analysis is greatly to be encouraged. But such exploratory investigations should be used as the basis for generating testable hypotheses, not for both generating and testing the hypotheses. Any findings from an exploratory data analysis need to be confirmed with new data that are collected for the explicit purpose of testing one or more given hypotheses under a controlled and specified error rate.

One suggestion might be for the Parapsychological Association to attempt to draw up a set of standards for designating studies as confirmatory. Ideally, exploratory studies, as such, should not be published, except possibly as the context for a confirmatory study that was a consequence of the exploratory study. By requiring confirmatory studies both to be labeled as such and to conform to a given set of standards, the parapsychological community will not only help to dispel much confusion, but it will also help both parapsychological and critical reviewers to select the appropriate studies to include in a meta-analysis.

Comments on the Individual Flaws

Would data bases in other fields of research withstand the same sort of scrutiny to which I have put the ganzfeld psi data base? I suspect, for example, that if I devoted the same effort to a critical evaluation of a set of studies in an area of psychology, I would find many of the same sorts of flaws. I am almost certain, based on my years as a referee for several scientific journals, that I would find the same sorts of multiple-testing flaws that I report in the ganzfeld psi studies. This is because parapsychologists have been trained as psychologists, biologists, and physicists and have been taught statistics from the same textbooks. And it is sad to report that most of this training is inadequate and even inconsistent in dealing with the problem of simultaneous statistical inference. Both textbooks and journal editors seem to have no hesitation, for example, in allowing investigators to test the several separate lines of an analysis of variance table each at the .05 level. Yet, when the same investigators attempt to compare the individual means within a given line of that same table, they are often required to use post hoc tests that penalize them for the implicit number of comparisons that might be made.

In other words, the problem of multiple testing and inflated error rates is by no means unique to the ganzfeld psi data base. But this should be of little comfort to the parapsychologist. Nor should this lessen the force of my arguments based on the existence of these inflated significance levels. In addition, an argument might be made that parapsychologists should be more vigilant in this regard than scientists in other areas. After all, the argument for psi is currently based on the assumption that the statistical probabilities as reported are reasonably close to what is actually the case.

Again, a body such as the Parapsychological Association might lessen the seriousness of multiple testing by establishing guidelines. And the various journals might make it clear to potential authors that such guidelines must be followed for all papers that are submitted as confirmatory studies. As well as setting a model for other fields to emulate, the development of such guidelines might also break new ground in other ways. One would be to suggest ways that authors can most efficiently and legitimately make multiple comparisons and tests and appropriately take this into account in drawing their conclusions.

I am not so sure about how much the procedural flaws in this data base would also characterize studies in other fields. I suspect that randomization in many areas of psychology is a casual affair. As such, it would be considered "flawed" by my criterion. But again, appropriate randomization is more critical in a parapsychological study than it probably is in a learning or perceptual study. The very definition of psi involves a statistically significant departure from a value in a specified distribution. Both the meaning of the departure and the interpretation of the significance level depend crucially on the underlying assumptions' being correct. And careful attention to randomization is one of the few ways to guarantee the adequacy of such assumptions.

Again, perhaps the Parapsychological Association in conjunction with various journal editors can lead the way in emphasizing the critical importance of randomization. The one investigator in the current sample who seems to realize the importance of randomization is Stanford (Studies 31, 35, 36). The three studies in which he was investigator or coinvestigator were the only ones about which I was confident that randomization had truly been carried out correctly. The indication that inadequate randomization in this sample is correlated with significance further stresses the importance of making it clear to future investigators that this is not a casual matter.

One promising trend in this data base is the significant reduction over the 8 years in the flaws relating to security. Most of this reduction

is due to the replacement of single target-sets with duplicate ones. But the parapsychologists ought not to be content simply to point out that this flaw has been banished. They ought to ask themselves, in my opinion, why this flaw was allowed to persist so long. Studies continued exhibiting this flaw right through 1979. Only in the last 2 years of the data base did it finally disappear. Why did it take the parapsychological community 6 years to finally recognize and abolish this flaw that so readily strikes the eye of an outsider? If such an obvious flaw can persist in a major body of parapsychological literature for 6 years, what other, perhaps more subtle, flaws still abound?

The flaws classified as Statistics (STAT) raise other challenges. Twelve, or 28%, of the studies were faulted for such mistakes. This could be an underestimate for several reasons. I could detect some of the errors only when enough data were provided for me to recalculate the statistics. In some cases I suspected that the reported statistics could not be correct, but I was unable to verify this suspicion because of insufficient data. And because the occurrence of this flaw was correlated with the size and complexity of the study, it could be the case that other experimenters might also have exhibited such errors if they had conducted more complicated studies.

All this suggests that parapsychologists no longer should rest content with Burton Camp's 1937 assertion that for Rhine's work "the statistical analysis is essentially valid" (cited in Mauskopf & McVaugh, 1979). Because there is no common curriculum for those who end up as parapsychologists, it is probably not safe to assume that a parapsychologist automatically knows how to conduct and compute the appropriate statistical analysis properly. As paradigms change and complexity increases, statistical competence that was adequate for previous research may no longer apply. Again, both the Parapsychological Association and journal editors might take the lead in trying to insure that future work uses correct statistical analyses.

Here, too, is an opportunity to break new ground. In many of the ganzfeld psi studies, there is a gray area in which it is unclear what the appropriate sampling unit should be. Many questions arise about when it is safe to pool over conditions, trials, subjects, and other potentially correlated dimensions. Both theoretical and empirical work could be useful here in establishing some guidelines.

It is no surprise that unpublished abstracts often are inadequately documented. But 38% of the published papers were also faulted for insufficient documentation. Much of this involves the introduction of changes or variations in conditions without the inclusion of information about the effects of these changes. This is closely related to

another problem, for which I did not create a formal category, but which strongly suggests that these studies were conducted on a rather informal basis. Percipients, for example, were often encouraged to bring friends as agents. For those who did not, the experimenter supplied an agent. Here we have a variable treated as a constant because, with one or two exceptions, all these situations were treated as a single condition.

In conclusion, the current data base has too many problems to be seriously put before outsiders as evidence for psi. The types of problems exhibited by this data base, however, suggest interesting challenges for the parapsychological community. I would hope that both parapsychologists and critics would wish to have parapsychological studies conducted according to the highest standards possible. If one goal is to convince the rest of the scientific community that the parapsychologists can produce data of the highest quality, then it would be a terrible mistake to use the current ganzfeld psi data base for this purpose. Perhaps the Parapsychological Association can lead the way by setting down guidelines for what should constitute an adequate confirmatory study. And then, when a sufficient number of studies that meet these guidelines have accumulated, they can be presented to the rest of the scientific community as an example of what parapsychology, at its best, can achieve. If studies carried out according to these guidelines also continue to yield results suggestive of psi, then the outside scientific community should be obliged to take notice.

References

Akers, C. (1984). Methodological criticisms of parapsychology. In S. Krippner (Ed.), *Advances in parapsychological research* (Vol. 4). Jefferson, NC: McFarland.

Ballard, J. A. (1981). *Relaxation and the ganzfeld as psi-conducive states.* Unpublished manuscript.

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

Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, **21**, 607–611.

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

Hansel, C. E. M. (1980). *ESP and parapsychology: A critical reevaluation.* Buffalo, NY: Prometheus Books.

Hedges, L. V., & Olkin, I. (1982). Analyses, reanalyses, and meta-analysis [Review of *Meta-analysis in social research*]. *Contemporary Education Review*, **1**, 157–165.

HONORTON, C. (1978). Psi and internal attention states: Information retrieval in the ganzfeld. In B. Shapin & L. Coly (Eds.), *Psi and states of awareness* (pp. 79–90). New York, NY: Parapsychology Foundation.

HONORTON, C. (1979). Methodological issues in free-response experiments. *Journal of the American Society for Psychical Research.* **73**, 381–394.

HONORTON, C. (1981). Beyond the reach of sense: Some comments on C. E. M. Hansel's *ESP and parapsychology: A critical reevaluation. Journal of the American Society for Psychical Research*, **75**, 155–166.

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

HYMAN, R. (1977, November/December). The case against parapsychology. *The Humanist*, **37**, 47–49.

KENNEDY, J. E. (1979a). Methodological problems in free-response ESP experiments. *Journal of the American Society for Psychical Research*, **73**, 1–15.

KENNEDY, J. E. (1979b). More on methodological issues in free-response psi experiments. *Journal of the American Society for Psychical Research*, **73**, 395–401.

LIGHT, R. J., & SMITH, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, **41**, 429–471.

MAUSKOPF, S. H., & MCVAUGH, M. R. (1979). The controversy over statistics in parapsychology, 1934–1938. In S. H. Mauskopf (Ed.), *The reception of unconventional science* (pp. 105–123). Boulder, CO: Westview Press.

PALMER, J. (1983, August). In defense of parapsychology: A reply to James E. Alcock. *Zetetic Scholar*, No. 11, 39–70.

PARKER, A. (1978). A holistic methodology in psi research. *Parapsychology Review*, **9** (2), 1–6.

PARKER, A., & WIKLUND, N. (1982). *The ganzfeld: A methodological evaluation of the claims for a repeatable ESP experiment.* Manuscript submitted for publication.

ROGO, D. S. (1977, November/December). The case for parapsychology. *The Humanist*, **37**, 40–44.

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

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

SARGENT, C. L. (1980a). Exploring psi in the ganzfeld. *Parapsychological Monographs*, No. 17. New York: Parapsychology Foundation.

SARGENT, C. L. (1980b). [Letter to the editor]. *Journal of the American Society for Psychical Research*, **74**, 265–267.

SONDOW, N. (1980). [Letter to the editor]. *Journal of the American Society for Psychical Research*, **74**, 267–272.

Department of Psychology
University of Oregon
Eugene, Oregon 97403

Appendix
The Ganzfeld Psi Data Base

The data base consists of 42 separate ganzfeld psi studies as defined by Charles Honorton (personal communication, November 30, 1981). These include all the studies conducted between 1974 and 1981 known to Honorton and meeting his criteria of acceptable ganzfeld psi studies. The studies are numbered for referencing in the text of the article. They are listed alphabetically. Because the same report sometimes describes more than one study, the reference is repeated separately for each study.

A list of coded descriptors follows each entry. These are defined briefly here but are described more fully in the text:

AT. Alternate tests used without adjusting significance.

DOC. Inadequate documentation, usually with respect to trials that differ on how the agent is related to the percipient.

E. Effect size defined as the difference in degrees between the transformed proportion of hits and transformed proportion expected by chance. Proportions were converted to degrees by the Freeman-Tukey arc sine transformation for binomial proportions (Freeman & Tukey, 1950). When the effect size is followed by a BH or a BC in parentheses, it indicates that it was calculated either on the basis of binary hits or binary coding, respectively. When neither of these descriptors is given, the effect size is based on direct hits. Where no effect size is given, it indicates that the report gave insufficient data for calculating an effect size.

FB. Inadequate randomization of target and foils at judging, or inadequate precautions against communication from percipient to agent at feedback.

IJ. Independent judges used in addition to subjects as own judges without adjusting significance.

MB. Multiple baselines (testing the same index against both a chance and a control baseline without adjusting significance).

MDV. Multiple dependent variables used without adjusting significance.

MG. Multiple groupings in testing for psi against control comparisons without adjusting significance.

MI. Multiple indices used without adjusting significance.

NSIG-1. Not classified as significant on the first (November 1981) data base.

NSIG-2. Not classified as significant after the June 1982 adjustments.

R(+). Appropriate randomization.

R(–). Inadequate randomization.

R(?). Randomization procedures inadequately described.

SEC. Inadequate security, usually in monitoring crucial phases of the study or in having only one experimenter.

SIG-1. Classified as significant overall on the primary index by Honorton at the .05 level as of November 30, 1981.

SIG-2. Classified as significant by Honorton after adjusting for multiple testing (personal communication, June 2, 1982).

ST. Single target used, allowing sensory cueing.

STAT. Inappropriate statistics, such as wrong degrees of freedom or failing to calculate p for Fisher's exact test appropriately.

Z. Critical ratio or normal deviate based on the effect size E and its theoretical standard deviation.

The Studies

1. Ashton, H. T., Dear, P. R., Harley, T. A., & Sargent, C. L. (1981). A four-subject study of psi in the ganzfeld. *Journal of the Society for Psychical Research,* **51**, 12–21. [Experiment 4 of Sargent, 1980]
   SIG-1, NSIG-2, AT, MG, MI, R(?), DOC, E = 11.03, $Z = 2.19$

2. Braud, W. G., Wood, R., & Braud, L. W. (1975). Free-response GESP performance during an experimental hypnagogic state induced by visual and acoustic ganzfeld techniques: A replication and extension. *Journal of the American Society for Psychical Research,* **69**, 105–113.
   SIG-1, SIG-2, MB, MI, R(?), ST, FB, SEC, E = 8.08, $Z = 0.91$

3. Braud, W. G., & Wood, R. (1977). The influence of immediate feedback on free-response GESP performance during ganzfeld stimulation. *Journal of the American Society for Psychical Research,* **71**, 409–427.
   SIG-1, SIG-2, MB, MDV, MG, MI, R(?), ST, DOC, STAT, E = –2.82 (BH), Z = –0.77

4. Child, I. L., & Levi, A. (1979). Psi-missing in free-response settings. *Journal of the American Society for Psychical Research,* **73**, 273–289.
   SIG-1, NSIG-2, R(?), ST, FB, SEC, STAT, E = –20.43, $Z = -2.71$

5. Dunne, B. J., Warnock, E., & Bisaha, J. P. (1977). Ganzfeld techniques with independent rating for measuring GESP and precognition. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1976* (pp. 41–43). Metuchen, NJ: Scarecrow Press.
   SIG-1, NSIG-2, MI, R(–), DOC

6. Habel, M. M. (1976). Varying auditory stimuli in the ganzfeld: The influence of sex and overcrowding on psi performance. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975* (pp. 181–184). Metuchen, NJ: Scarecrow Press.
   NSIG-1, NSIG-2, MG, R(?), ST, DOC, E = –0.63(BH), $Z = -0.21$

7. Honorton, C. (1976). Length of isolation and degree of arousal as probable factors influencing information retrieval in the ganzfeld. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975* (pp. 184–186). Metuchen, NJ: Scarecrow Press.
   SIG-1, SIG-2, R(–), ST, FB, DOC, E = 32.76, $Z = 3.13$

8. Honorton, C., & Harper, S. (1974). Psi-mediated imagery and ideation in an experimental procedure for regulating perceptual input. *Journal of the American Society for Psychical Research*, **68**, 156–168.
SIG-1, SIG-2, R(–), ST, FB, DOC, SEC, E = 10.77, $Z = 2.08$

9. Keane, P., & Wells, R. (1979). An examination of the menstrual cycle as a hormone related physiological concomitant of psi performance. In W. G. Roll (Ed.), *Research in parapsychology, 1978* (pp. 72–74). Metuchen, NJ: Scarecrow Press.
SIG-1, NSIG-2, MB, MDV, MG, R(+), DOC, STAT

10. Palmer, J., & Aued, I. (1975). An ESP test with psychometric objects and the ganzfeld: Negative findings. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1974* (pp. 50–53). Metuchen, NJ: Scarecrow Press.
NSIG-1, NSIG-2, MB, MG, MI, R(+), ST, SEC, E = –2.68, $Z = -0.60$

11. Palmer, J., Bogart, D. N., Jones, S. M., & Tart, C. T. (1977). Scoring patterns in an ESP ganzfeld experiment. *Journal of the American Society for Psychical Research*, **71**, 121–145.
NSIG-1, NSIG-2, MI, IJ, R(+), ST, E = –1.07, $Z = -0.21$

12. Palmer, J., Khamashta, K., & Israelson, K. (1979). An ESP ganzfeld experiment with Transcendental Meditators. *Journal of the American Society for Psychical Research*, **73**, 333–348.
NSIG-1, NSIG-2, MB, MI, IJ, R(+), ST, E = –10.67, $Z = -1.69$

13. Palmer, J., Whitson, T., & Bogart, D. N. (1980). Ganzfeld and remote viewing: A systematic comparison. In W. G. Roll (Ed.), *Research in parapsychology, 1979* (pp. 169–171). Metuchen, NJ: Scarecrow Press.
NSIG-1, NSIG-2, MG, R(+)

14. Parker, A. (1975). Some findings relevant to the change in state hypothesis. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1974* (pp. 40–42). Metuchen, NJ: Scarecrow Press.
NSIG-1, NSIG-2, MG, R(?), ST, SEC, E = –7.48(BH), $Z = -1.44$

15. Parker, A., Millar, B., & Beloff, J. (1977). A three-experimenter ganzfeld: An attempt to use the ganzfeld technique to study the experimenter effect. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1976* (pp. 52–54). Metuchen, NJ: Scarecrow Press.
NSIG-1, NSIG-2, MG, R(+), DOC

16. Raburn, L. (1975). *Expectation and transmission factors in psychic functioning.* Unpublished honors thesis, Tulane University, New Orleans, LA. [Experiment 1 = Cell with informed Ss and an agent].
SIG-1, SIG-2, AT, MG, R(–), ST, FB, DOC, SEC, STAT, E = 37.20, $Z = 4.21$

17. Ibid. [Experiment 2 = Cell with informed subjects but no agent]
NSIG-1, NSIG-2, AT, MG, R(–), ST, FB, DOC, SEC, STAT, E = 8.33, $Z = 0.94$

18. Rogo, D. S. (1976). ESP in the ganzfeld: An exploration of parameters. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975* (pp. 174–176). Metuchen, NJ: Scarecrow Press. [Experiment 1]
NSIG-1, NSIG-2, MI, R(–), ST, DOC, E = 2.21, $Z = 0.41$

19. Ibid. [Experiment 2]
NSIG-1, NSIG-2, MI, R(–), ST, DOC, SEC, E = 8.33, $Z = 0.94$

20. Rogo, D. S. (1977). A preliminary study of precognition in the ganzfeld. *European Journal of Parapsychology*, **2**(1), 60–67.
NSIG-1, NSIG-2, AT, MG, R(+), E = –3.72(BC), $Z = -1.84$

21. Rogo, D. S., Smith, M., & Terry, J. (1976). The use of short-duration ganzfeld stimulation to facilitate psi-mediated imagery. *European Journal of Parapsychology*, **1**, 72–77.
NSIG-1, NSIG-2, MI, R(–), ST, FB, DOC, E = 5.93, $Z = 0.94$

*22. Roney-Dougal, S. M. (1982). A comparison of psi and subliminal perception: A confirmatory study. In *Research in parapsychology, 1981* (pp. 96–99). Metuchen, NJ: Scarecrow Press.
SIG-1, NSIG-2, MB, MG, MI, IJ, R(?), DOC, SEC, STAT, E = 6.09, $Z = 1.35$

23. Sargent, C. L. (1980). Exploring psi in the ganzfeld. *Parapsychological Monographs*, No. 17. [Experiment 1]
NSIG-1, NSIG-2, MI, R(–), E = –3.53, $Z = 0.63$

24. Ibid. [Experiment 2]
SIG-1, SIG-2, MI, R(?), E = 11.51, $Z = 1.82$

25. Ibid. [Experiment 3]
SIG-1, NSIG-2, MI, R(?), E = 11.51, $Z = 1.82$

26. Ibid. [Experiment 5]
SIG-1, SIG-2, MB, MI, R(?), E = 16.33, $Z = 3.15$

27. Ibid. [Experiment 6]
NSIG-1, NSIG-2, MB, MG, MI, R(?), E = 5.10, $Z = 1.07$

*28. Sargent, C. L., Bartlet, H. J., & Moss, S. P. (1982). Response structure and temporal incline in ganzfeld free-response GESP testing. In W. G. Roll, R. L. Morris, & R. A. White (Eds.), *Research in parapsychology, 1981* (pp. 79–81). Metuchen, NJ: Scarecrow Press.
SIG-1, NSIG-2, MG, MI, IJ, R(?), DOC, E = 1.95, $Z = 0.39$

*At the time this report was written, I had access to the paper as submitted rather than the version published in *Research in Parapsychology*. I assume that the two versions differ in no essential respect.

29. Sargent, C. L., Harley, T. A., Lane, J., & Radcliffe, K. (1981). Ganzfeld psi-optimization in relation to session duration. In W. G. Roll & J. Beloff (Eds.), *Research in parapsychology, 1980* (pp. 82–84). Metuchen, NJ: Scarecrow Press.
NSIG-1, NSIG-2, MB, MG, MI, R(?), DOC, E = 1.58, $Z = 0.35$

*30. Sargent, C., & Matthews, G. (1982). Ganzfeld GESP performance in variable duration testing. In W. G. Roll, R. L. Morris, & R. A. White (Eds.), *Research in parapsychology, 1981* (pp. 159–160). Metuchen, NJ: Scarecrow Press.
SIG-1, SIG-2, MB, MG, MI, R(?), DOC, E = 12.28, $Z = 2.21$

31. Schmitt, M., & Stanford, R. G. (1978). Free-response ESP during ganzfeld stimulation: The possible influence of menstrual cycle phase. *Journal of the American Society for Psychical Research,* **72**, 177–182.
SIG-1, SIG-2, MB, MG, R(+), ST, STAT, E = 19.74, $Z = 3.12$

32. Smith, M., Tremmel, L., & Honorton, C. (1976). A comparison of psi and weak sensory influences on ganzfeld mentation. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975* (pp. 191–194). Metuchen, NJ: Scarecrow Press.
SIG-1, SIG-2, MG, R(–), DOC, STAT, E = 3.01(BC), $Z = 2.10$

33. Sondow, N. (1979). Effects of associations and feedback on psi in the ganzfeld: Is there more than meets the judge's eye? *Journal of the American Society for Psychical Research,* **73**, 123–150.
SIG-1, SIG-2, AT, MB, MG, IJ, R(–), ST, FB, DOC, STAT, E = 9.70, $Z = 3.40$

*34. Sondow, N., Braud, L., & Barker, P. (1982). Target qualities and affect measures in an exploratory psi ganzfeld. In W. G. Roll, R. L. Morris, & R. A. White (Eds.), *Research in parapsychology, 1981* (pp. 82–85). Metuchen, NJ: Scarecrow Press.
SIG-1, NSIG-2, MB, MG, MI, R(+), DOC, STAT, E = 4.62, $Z = 1.03$

35. Stanford, R. G. (1979). The influence of auditory ganzfeld characteristics upon free-response ESP performance. *Journal of the American Society for Psychical Research,* **73**, 253–272.
NSIG-1, NSIG-2, MB, MDV, MG, R(+), STAT

36. Stanford, R. G., & Neylon, A. (1975). Experiential factors related to free-response clairvoyance performance in a sensory uniformity setting (ganzfeld). In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1974* (pp. 89–93). Metuchen, NJ: Scarecrow Press.
NSIG-1, NSIG-2, MB, MG, R(+), ST

37. Terry, J. C. (1976). Comparison of stimulus duration in sensory and psi conditions. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975* (pp. 179–181). Metuchen, NJ: Scarecrow Press.
NSIG-1, NSIG-2, MG, R(–), E = –0.34(BC), $Z = -0.22$

38. Terry, J. C., & Honorton, C. (1976). Psi information retrieval in the ganzfeld: Two confirmatory studies. *Journal of the American Society for Psychical Research*. **70**, 207–217. [Experiment 1]
SIG-1, SIG-2, MG, MI, R(–), ST, FB, DOC, E = 9.28, $Z = 1.70$

39. Ibid. [Experiment 2]
SIG-1, SIG-2, MG, MI, R(–), ST, FB, DOC, E = 11.91, $Z = 3.23$

40. Terry, J., Tremmel, L., Kelly, M., Harper, S., & Barker, P. L. (1976). Psi information rate in guessing and receiver optimization. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975* (pp. 194–198). Metuchen, NJ: Scarecrow Press.
SIG-1, SIG-2, MB, R(–), DOC, STAT, E = 4.19(BC), $Z = 1.80$

41. Wood, R., Kirk, J., & Braud, W. (1977). Free response GESP performance following ganzfeld stimulation vs. induced relaxation, with verbalized vs. nonverbalized mentation: A failure to replicate. *European Journal of Parapsychology*, **1**, 80–93.
NSIG-1, NSIG-2, MB, MG, MI, R(?), ST, DOC, SEC, E = –2.76, $Z = -0.67$

42. York, M. (1977). The defense mechanism test (DMT) as an indicator of psychic performance as measured by a free-response clairvoyance test using a ganzfeld technique. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1976* (pp. 48–49). Metuchen, NJ: Scarecrow Press.
SIG-1, SIG-2, MI, R(?), ST, DOC, E = 11.07, $Z = 2.72$

Journal of Parapsychology, Vol. 49, March 1985

META-ANALYSIS OF PSI GANZFELD RESEARCH: A RESPONSE TO HYMAN

By Charles Honorton

ABSTRACT: In response to Hyman's (1985) critique of psi ganzfeld studies, an evaluation is reported that eliminates multiple-analysis problems. The evaluation is restricted to the 28 studies (of the 42 considered by Hyman) that reported the number of direct hits. A uniform test (*Z* score associated with the exact binomial probability) is applied to a uniform index (proportion of direct hits). The mean *Z* score is 1.25 (SD = 1.57) with .76 as the lower bound of a 95% confidence interval estimate of the true population mean. The composite (Stouffer) *Z* score for the 28 studies is 6.6 ($p < 10^{-9}$), and 43% of the studies were independently significant at the 5% level. Six of the ten investigator groups reported significant outcomes, and cumulation by investigator yields a composite *Z* of 6.16; the significance of the psi ganzfeld effect does not depend on any one or two investigators.

A number of considerations mitigate against selective reporting bias as a viable explanation of these findings: (a) publication policies and practices in parapsychology show that null findings are frequently reported, (b) a large number of the studies under consideration *do* report null findings, and (c) Rosenthal's "file-drawer" estimate of the number of fugitive null studies needed to jeopardize the known results requires 15 fugitive studies for each one known.

Contrary to Hyman's claim, no significant relationship is found between study outcomes and measures of study quality (cue control and method of randomization). Hyman's "procedural flaws" analysis is discussed; ambiguities in the flaw criteria are noted, and examples of inconsistent or inappropriate assignment of flaw ratings are given.

In the early 1970s, a number of investigators were independently led to explore the effects of perceptual isolation techniques on performance in an ESP task (Braud, Wood, & Braud, 1975; Honorton & Harper, 1974; Parker, 1975). The psi ganzfeld research developed out of earlier research suggesting that successful performance in psi tasks is frequently associated with internal attention states brought about through dreaming, hypnosis, induced physical relaxation, and related procedures involving perceptual restriction. (For reviews, see Braud, 1978; Honorton, 1977; Honorton & Krippner, 1969.) The

I wish to thank four colleagues who have contributed in various ways in the preparation of this paper: Donald McCarthy for help on a variety of statistical matters, David Saunders for contributing the appendix on Hyman's factor analysis, and George Hansen for helpful comments on an earlier draft. I want to especially acknowledge the help of Ephraim Schechter for his many and varied contributions of time and expertise.

initial success of several different investigators with the ganzfeld technique stimulated wider interest, which led to a shift in emphasis from the process-oriented origins of the research to one focusing on replication rates. Earlier reviews of ganzfeld replication rates suggested a success rate over 50% in studies using the technique (e.g., Blackmore, 1980; Honorton, 1978).

Hyman's critique of the psi ganzfeld research (Hyman, this number of the *Journal*) is concerned with two issues: (a) whether the psi ganzfeld experiment supplies evidence for the existence of psi and (b) whether the effects obtained in psi ganzfeld experiments are replicable. I believe the existence of psi will remain in dispute until putative psi effects can be produced and studied with some specifiable degree of replicability. I am therefore primarily concerned with the extent to which the psi ganzfeld paradigm represents a step in that direction. The central claim under discussion is a replicability claim and, as such, will eventually be resolved through future replications.

The data base I am using comprises the 42 psi ganzfeld studies reported between 1974 and 1981 that reflect the scope of Hyman's review. Subsequent to the time I received Hyman's request for assistance with his review, I learned of additional studies that either had been unknown to me or were reported later; but because Hyman elected to freeze his analysis to the initial set of 42 studies, I shall do so too.

To facilitate comparison of the two papers, I am also adopting Hyman's format of referring to specific studies by the use of study numbers and an appendix for referencing the studies. Since there are numerous and often major points of disagreement between us over the interpretation of individual studies, I have documented my coding and major analyses in a way that will allow readers to reconstruct the process by consulting the original reports.

The following discussion is divided into four major sections. In the first section, I focus on whether there is, in fact, a statistically significant effect in the psi ganzfeld data base. The second section assesses the likely impact of reporting bias in psi ganzfeld studies. The relationship between study outcomes and various potential threats to their internal validity is explored in the third section. Finally, in the fourth section, I shall discuss Hyman's classification and analysis of flaws.

Is There an Effect?

Hyman devotes over half of his paper to the question of whether, after taking into account the effects of multiple analysis, choice of

sampling units, and possibilities of selective reporting bias, there is an aggregate psi ganzfeld effect. He suggests on the basis of his assessment of these factors that the actual rate of success (i.e., proportion of "significant" experiments) is at most 30% and that owing to multiple testing options, the true chance rate may be 25% or higher, not 5% as assumed in earlier replication rate estimates. In this section, I present an evaluation that is not influenced by multiple analysis.

Sampling Units

Hyman (pp. 9–11) raises a valid concern regarding ambiguities in the definition of study units where multiple conditions are involved. This problem, unfortunately, is endemic to meta-analysis, and there is no easy solution. As Cooper (1984) comments, "Obviously there will be some subjectivity in the reviewer's judgment of what constitutes a study. For instance, one reviewer might consider all results in a single report as one study. But another reviewer might consider a report that divides results into separate studies as containing more that one study" (p. 75).

Raburn's study (Studies 16 and 17) illustrates the problem. Ten subjects were assigned to each of four cells, in a 2 × 2 factorial design. One factor varied the presence of a sender (sender/no sender), and the other varied subjects' awareness of the ESP task (informed/uninformed). As in all of the other psi ganzfeld studies in this data base, subjects in the informed condition knew they were participating in an ESP task. Subjects in the uninformed condition, however, were led to believe that the experiment was "merely an attempt to elucidate physiological functions associated with sensory deprivation" (Raburn, 1975, pp. 11–12). This represented a radical departure from the standard ganzfeld procedure, one that is unique within the ganzfeld data base. I therefore excluded the uninformed condition, considering only the two cells in which subjects knew they were participating in an ESP task as valid ganzfeld studies.

Inevitably, there will be differences of opinion on decisions of this sort, and Hyman (p. 9) disagrees with my exclusion of Raburn's uninformed condition. Adrian Parker (personal communication, January 10, 1983), on the other hand, agrees that the uninformed condition should be excluded, and he would also exclude the informed–no-sender condition on the grounds that, unlike other ganzfeld studies with clairvoyance procedures, subjects in Raburn's no-sender group were misinformed that there would be a sender.

Though acknowledging that my classification of Raburn's study has, in itself, little effect on the overall evaluation of the data base as a

whole, Hyman (pp. 9–11) suggests that similar ambiguities exist in the classification of all studies with multiple conditions. Readers are invited to examine the original study descriptions and draw their own conclusions about the appropriateness of study classifications over which Hyman and I disagree.

Indices of Success

Five indices of success have been used in psi ganzfeld research. Four are based on blind-judging procedures in which the subject or judge is presented with a set of pictures ("judging pool") consisting of the target and a number of control pictures ("decoys"). The judge ranks (or rates) the degree of similarity between each picture and the verbal impressions ("mentation report") elicited during the ganzfeld session.

Direct hits. The most widely used variation is the direct-hits index. Here credit is given only when the target is correctly identified from a judging pool of N elements; so the probability of a hit on each trial is $1/N$. The overall success rate for a series is estimated through the binomial critical ratio (CR) or the exact binomial probability is calculated directly. The direct-hits measure is the simplest index of success, but also the most conservative since it discards most of the rank data.

Binary (partial) hits. The binary-hits index represents a crude weighting scheme that gives credit if the target is in the lower half of the ranks. If, for example, there are six elements in the judging pool, a partial hit is counted when a rank of 1, 2, or 3 is assigned to the target, and the probability of a hit is 1/2. Though giving some weight to partial hits, this measure still discards half of the rank data, and direct hits receive no more credit than partial ones.

Sum of ranks. The sum-of-ranks statistic is the most powerful rank index because it uses all the rank data and differentially rewards lower rank values. It became widely used following publication of the paper by Solfvin, Kelly, and Burdick (1978), which provided formulas and convenient tables.

Standardized ratings. A related index, which has been used in a few studies, uses standardized ratings (Stanford & Sargent, 1983). Here the target and decoys are rated, for example, on a scale from 0 to 100; and the rating assigned to the actual target is standardized. This measure has primarily been used to provide more continuous psi scores in studies correlating psi performance with psychological variables. Standardized ratings can also be reduced to ranks for analysis by one of the rank methods described above.

Binary coding. A very different index, used in eight studies in this data base, involved the Maimonides binary coding system (Honorton, 1975). This required a specially constructed target set defined in terms of the presence or absence of features in each of 10 categories such that the content of each target could be encoded as a 10-digit binary number. In contrast to the other indices described above, determination of success with this method does not involve comparison with decoys. The subject's mentation is coded with respect to the 10 target descriptors, thus producing a 10-bit response, which is simply matched against the target code. Because the target is selected from a pool containing all possible combinations of the presence or absence of the 10 features, each experimental trial constitutes 10 independent binary trials with a chance expectation of 5.

A number of factors could influence the relative sensitivity of the various indices. Although a detailed consideration of these issues is beyond the purview of our present discussion, one example, the composition of judging pools in studies with blind-judging procedures, illustrates some of the problems. Except for studies using the binary coding system, each investigator contributing to this data base has used a different target set and uniquely composed judging pools. To maximize the subject's ability to distinguish targets from decoys, the investigator must make the content of pictures in the judging pool as dissimilar as possible. If the subject's mentation report includes impressions of people and there are people in several pictures in the judging pool, there is a judging problem. Indices based on binary hits, sum of ranks, or standardized ratings are likely to be more sensitive than a direct-hits measure in such cases since lower ranks (or ratings) can be spread over the pictures containing mentation-related content. Avoiding overlap of content among the elements in a judging pool becomes more of a problem as the number of pictures in the judging pool increases, and it is probably for this reason that most investigators have limited their judging pools to four elements.

Multiple Analysis

There is no doubt that many investigators have applied multiple statistical tests or indices without adjusting their significance criteria for the number of tests used. This practice *could*, as Hyman argues, dramatically alter our estimates of overall significance. Although Hyman and I agree that there is a multiple-analysis problem in this data base, we have taken different approaches to evaluating its impact.

Multiple-analysis flaw ratings. Hyman's approach has been to assign

flaw ratings to studies that used unadjusted multiple analysis. He has coded studies for the presence or absence of six categories of multiple-analysis errors. Many of these "flaws" are irrelevant since they do not affect assessment of an overall effect. For example, multiple groupings, which Hyman finds to be one of the most frequent causes of multiple-testing errors in this data base, was charged against a study "whenever the study [had] more than one condition" (p. 22). This would be an appropriate charge if we were concerned with assessment of statistical significance of any kind in these studies, but is irrelevant to the assessment of an overall psi ganzfeld effect.

Though comparisons between ganzfeld and other conditions often supplied the motivation for the original investigators in conducting their studies, the previous reviews of psi ganzfeld research, which serve as the source of claims for a psi ganzfeld effect, have counted only overall effects in their estimates of psi ganzfeld success rate (Blackmore, 1980; Honorton, 1977, 1978). I do not believe it is appropriate to charge studies with multiple-analysis flaws simply because the interest of the reviewer differs from that of the original investigator, and this seems to be what Hyman has done in his multiple-grouping and multiple-baseline flaw assignments.

It is not necessary to pursue Hyman's sixfold approach to multiple analysis much further because he found that "no significant differences on any of these six categories of multiple testing exist between the studies classified as 'significant' and those classified as 'nonsignificant'" (p. 23). (Curiously, he then describes three "scenarios" to explain why his six categories should not be expected to predict psi ganzfeld success in the first place, leaving this writer somewhat baffled about the purpose of the exercise.)

There remains, however, a genuine multiple-testing problem in this data base. A number of authors did use multiple tests or indices without applying suitable corrections. I have taken two approaches in evaluating the impact of multiple analysis on the assessment of overall outcome: (a) adjusting alpha levels to correct for the number of analyses of overall success rate in studies originally classified as significant, and (b) focusing on studies that use a uniform index. The first approach was included in my response to Hyman's (1983) earlier critique (Honorton, 1983) and constitutes what he refers to as my revised classification; the second approach is presented in this paper.

Alpha adjustment for multiple testing. This approach used the Bonferroni inequality (Rosenthal & Rubin, in press) to guarantee an alpha level no larger than 5%. The adjusted alpha level is calculated

by dividing the criterion of significance (i.e., .05) by the number of tests from which an overall effect might be claimed. The Bonferroni correction requires that at least one of the tests remain significant at the adjusted level. If, for example, three separate (albeit not necessarily independent) tests were conducted to assess an overall effect, at least one must be associated with a p level not larger than .05/3 (i.e., .0167) for the study to survive the adjustment for multiple testing.

This analysis led to a revision of the number of significant studies in this data base from 23/42 to 19/42; that is, approximately 45% of the studies remained significant at the 5% level. Hyman's response (personal communication, November 29, 1982) was that it is possible to extract a much larger number of analyses (implying a substantially larger correction factor) by counting *all possible* analyses in a given study, including those that are irrelevant to assessment of an overall effect. In one case (Study 22), he counted 513 "implicit" analyses in the study!

Uniform test and index. Fortunately, there is another solution, one that removes all reasonable doubt about the nonchance status of the psi ganzfeld effect. This is to apply a uniform test on a single index across all studies that used that index. The direct-hits index was chosen because it was by far the most popular index and could be applied to the largest number of studies. Direct hits was also the index used in the first published psi ganzfeld study (Study 8), which Hyman and others have taken as the model or prototypical psi ganzfeld experiment.

The Direct-Hits Studies

Of the 42 studies included in Hyman's review, 14 did not provide direct hits. These include 10 blind-judging studies in which the primary analysis was based on sum of ranks, binary hits, or standardized ratings (Studies 3, 5, 6, 9, 13–15, 22, 35, 36). In four other studies, the Maimonides binary coding system provided the only index possible (Studies 20, 32, 37, 40). The number of direct hits is available for the remaining 28, or two thirds of the 42 studies examined by Hyman (Studies 1, 2, 4, 7, 8, 10–12, 16–19, 21, 23–31, 33, 34, 38, 39, 41, 42).

These 28 studies were reported by investigators in 10 different laboratories and comprise a total of 835 psi ganzfeld sessions. The reports appeared in the following forms: 12 (43%) were published in journals, 11 (39%) were reported at Parapsychological Association

TABLE 1
COMPARISON OF THREE *Z* SCORE ESTIMATES

	Z score estimates		
	Hyman	*CR*	Exact *p*
Mean	1.31	1.37	1.25
Median	1.05	1.03	.95
SD	1.61	1.54	1.57
Stouffer *Z*	6.95	7.22	6.60
File-drawer	472	512	423

Note. *Z* (Hyman) was estimated from Freeman-Tukey test. *Z* (*CR*) is the binomial *CR* with correction for continuity when $np < 10$. *Z* (exact *p*) was obtained by converting the exact binomial *p* into a *Z* score.

conventions with abstracts appearing in *Research in Parapsychology,* and 5 (18%) appeared in a monograph.

Test of direct-hits index. Several options are available for obtaining *Z* scores based on the number of direct hits in each study. The authors of the studies generally used the binomial *CR* approximation, though some calculated the exact binomial probabilities. Hyman uses an approximation based on the Freeman-Tukey effect-size estimate (Freeman & Tukey, 1950). The *Z* scores produced by these three methods are almost perfectly correlated, although the magnitude of individual *Z* scores does vary somewhat. In one case (Study 4), Hyman's estimate yields a *Z* of –2.71, and the exact binomial gives a much smaller *Z* of –1.71. In another case (Study 38), the differences alter the study's classification as significant or nonsignificant. For this study, Hyman obtains a *Z* of 1.7, which is significant on a one-tailed test. The *CR* approximation is also significant ($Z = 1.67$). The *Z* score based on the exact binomial probability, however, falls short of significance ($Z = 1.62$). Both the *CR* and Hyman's method are approximations, and considering the relatively small sample sizes of the studies in this data base, the use of the exact binomial would seem to be the most appropriate method. Therefore, I calculated the exact binomial probability for each study and obtained its associated *Z* score. As shown in Table 1, this produces slightly more conservative estimates of overall significance and tolerance for the file-drawer problem (to be discussed later) than either the *CR* method or Hyman's method.

The expected *Z* score, on the null hypothesis, is zero. Of the 28 studies, 23 (or 82%) have positive *Z* scores ($p = .00046$, exact binomial test with $p = q = .5$). The mean *Z* score is 1.25 (SD = 1.57). The one-tailed 95% confidence interval (Kirk, 1982) yields .76 as the

TABLE 2
PUBLICATION SOURCE AND SUCCESS RATE

Source of publication	No. of studies	Mean *Z* score	Stouffer *Z*	% Sig. .05	1-Way ANOVA, source × *Z*
Journals	12	1.02	3.53	42%	
Monograph	5	1.61	3.60	60%	$F(2,25) = 0.28$
Research in Parapsychology	11	1.33	4.41	36%	
All sources	28	1.25	6.60	43%	

estimate of the lower limit of the true population mean. (When Hyman's *Z* estimate is used, the 95% confidence interval gives a lower limit of .81.)

A composite *Z* score was computed by the Stouffer method recommended by Rosenthal (1978). This involves dividing the sum of the *Z* scores for the individual studies by the square root of the number of studies. The resulting *Z* score is 6.60 ($p < 10^{-9}$). (If we assume average *Z* scores of *zero* for the 10 additional blind-judging studies, which did not provide direct-hits information, the combined result across all 38 studies to which direct hits *could possibly have been applied* gives a Stouffer *Z* of 5.67 [$p = 7.3 \times 10^{-9}$]. Thus, whether we stick to the studies for which the relevant information is available or include a null estimate for the additional studies where the information is not available, the aggregate result cannot reasonably be attributed to chance fluctuation.)

Of the 28 studies, 12 (or 43%) have *Z* scores independently significant at the 5% level (Studies 1, 7, 8, 16, 24–26, 30, 31, 33, 39, 42; $p = 3.5 \times 10^{-9}$, exact binomial test with $p = .05$ and $q = .95$). Twenty-five percent of the studies (7/28) are significant at the 1% level (Studies 7, 16, 26, 31, 33, 39, 42; $p = 9.8 \times 10^{-9}$, exact binomial test with $p = .01$ and $q = .99$).

It is clear that the cumulative outcome of this set of studies cannot be attributed to the inflation of alpha levels through multiple analysis. When a single test is used on a uniform index of success, the result indicates a strong and highly significant overall psi ganzfeld effect.

Success rate and source of publication. As indicated above, there were three different sources of publication for these 28 studies: journals, monograph reports, and *Research in Parapsychology* abstracts. Table 2 shows the number of studies for each source along with the mean and combined *Z* scores and the percentage of significant studies. A

TABLE 3
OUTCOME (DIRECT-HITS STUDIES) BY INVESTIGATOR

Investigator	No. of studies	Study[a]	Stouffer *Z* by investigator
Child & Levi	1	4	–1.71
Schmitt & Stanford	1	31	3.11
Sondow	1	33	3.41
York & Morris	1	42	2.89
Braud & Wood	2	2,41	– .04
Raburn	2	16,17	3.38
Palmer et al.	3	10–12	–1.69
Rogo	3	18,19,21	1.04
Honorton et al.	5	7,8,34,38,39	4.82
Sargent et al.	9	1,23–30	4.28
		Stouffer *Z* (investigators) =	6.16

[a] Study refers to the study number used in Appendix A.

one-way analysis of variance (source of publication × study *Z* score) shows no significant effect of publication source ($F[2,25] = 0.28$).

Interlaboratory replicability. A valid objection made to estimates such as this is that they are based on the success rate of studies rather than on the number of successful investigators or laboratories (Parker, 1978). Since experimenter effects appear to play a prominent role in psi research, high success rates by a small proportion of the contributing investigators are less germane to an assessment of replicability than are the estimates based on success rates across investigators.

Following Rosenthal (1984, p. 128), a combined (Stouffer) *Z* score was obtained individually for each investigator. These results are shown in Table 3. Significant outcomes are reported by 6 of the 10 investigators and the combined result across investigators yields a *Z* of 6.16 ($p < 10^{-9}$). Even though half of the studies ($n = 14$) were contributed by two investigators, Carl Sargent and me, and account for 8/12 (67%) of the significant studies, the interlaboratory replicability of the psi ganzfeld effect does not depend on Sargent's work or my own: if we remove the Sargent and Honorton studies, the Stouffer *Z* across the eight other investigator teams remains safely significant ($Z = 3.67$; $p = .0001$). Four (29%) of these studies (Studies 16, 31, 33, 42) are significant at the 1% level ($p = 9.2 \times 10^{-6}$; binomial test with $n = 14$, $p = .01$, and $q = .99$), and *each* was contributed by a different investigator.

Thus, though the total number of investigators in this data base is small ($n = 10$), a majority of them have reported significant studies, and the significance of the overall effect is not dependent on one or two investigators.

Summary on Existence of an Effect

I have limited my analysis to a subset of psi ganzfeld studies to resolve issues raised by Hyman concerning multiple analysis and the effective error rate (alpha level). The analysis was restricted to the blind-judging psi ganzfeld studies that supplied the number of direct hits. A uniform test (Z score associated with the exact binomial probability) was applied to a uniform index (proportion of direct hits). The analysis shows that (a) the cumulative Z score across all studies that met or could in principle have met these criteria is associated with a probability not larger than one part in 100 million, (b) the cumulative Z score by investigators is likewise highly significant and does not depend on any one or two laboratories or investigators, and (c) 43% of the studies (by 60% of the investigators) are significant with the expected chance level safely set at 5%.

Reporting Bias

The File-drawer Problem

Can a good case be made for attributing these findings to selective reporting of "successful" studies? The question is a reasonable one. Selective reporting is a well-known and pervasive problem in the behavioral sciences (Bozarth & Roberts, 1972; Sterling, 1959), and there have been numerous calls for remedial action from investigators in fields ranging from the neurochemistry of learning (Dunn, 1980) to abnormal and developmental psychology (Sommer & Sommer, 1983).

In this regard, parapsychology has set a precedent that other areas of behavioral research would do well to emulate. Recognizing the importance of negative results in assessing research findings, the Parapsychological Association Council in 1975 adopted a policy opposing the selective reporting of positive outcomes. As a consequence, negative findings are routinely reported at Parapsychological Association meetings and in its affiliated journals and other publications.

It is therefore not surprising that approximately half of the known psi ganzfeld studies have, in fact, reported nonsignificant outcomes. Nor is it surprising that Blackmore's (1980) survey of unreported psi ganzfeld studies failed to support the hypothesis that we are dealing here with a biased sample of studies. Blackmore found that, of the 19 completed but then unreported studies elicited through her survey, 7 (or 37%) claimed significant overall outcomes, and she reported a chi-square analysis indicating that the outcome status (significant or nonsignificant) was not significantly related to publication status. She concluded that "the bias introduced by selective reporting of ESP ganzfeld studies is not a major contributor to the overall proportion of significant results, and the apparent success of the technique" (pp. 217–218).[1,2]

Taken to an extreme, the appeal to unknown or unreported studies is a fundamentally nonfalsifiable claim. We can never know, with anything approaching finality, the extent of reporting bias in any research domain. The viability of such claims can be evaluated by attention to the research and reporting practices in the research domain being examined and through estimation of the extent of selective reporting that would be necessary to jeopardize the existing data base. Rosenthal's "file-drawer" statistic (Rosenthal, 1979) estimates the number of unreported studies with Z scores averaging zero that would be required to cancel out the significance of an existing data base. For the direct-hits ganzfeld studies, the file-drawer statistic leads to an estimate of 423 such studies needed to raise the cumulative probability of the 28 known studies to a p of .05 that is, a ratio of unreported-to-reported studies of approximately 15 to 1. Given the mean ganzfeld duration for the existing studies of approximately 28 min, an additional time expenditure of 30 min per session for set-up, instructions, randomization, judging, feedback, and so forth, and an average of 30 trials per study, this translates into more than 12,000 fugitive sessions—one psi ganzfeld session per hour for over 6 years, assuming 40-hour weeks and no vacations! (As for Hyman's comment that "strangely, no contributions have come from Honorton and his

[1]It is inappropriate to add these 19 studies to the current data base as Hyman has done in obtaining his "adjusted count" (p. 11) because some of these studies were subsequently published and they duplicate those already in the data base of 42 studies (personal communication, S. J. Blackmore, November 3, 1982). As for the 11 other studies Hyman includes (p. 11) in his "adjusted count," I simply point out that it was he, and not I, who decided to freeze the analysis to the 42 studies I initially sent to him.

[2]Readers who know of other unreported psi ganzfeld studies that were not registered in Blackmore's survey should contact the Editor of the *Journal*, who will send them a questionnaire for documenting details of their study.

laboratory during the latter 4 years of the span covered by the present data base" [p. 35], I draw your attention to Study 34.)

Hyman's "Retrospective Study" Hypothesis

Hyman reports a significant negative relationship between sample size and study outcome.[3] Dividing the studies into four classes of sample size and performing a power analysis, he finds a significant tendency for studies in the class with the smallest sample sizes (< 20 trials, $n = 7$) to have significant outcomes. [4,5] Since, as he says (p. 13), "We would normally expect to find the probability of obtaining a significant result, all other things being equal, to increase with the

[3]Using the figures Hyman cites for the observed and expected number of significant studies (pp. 13–14) and performing a standard chi-square calculation yields a chi-square of 24.56, not 31.42 as reported by Hyman. The chi-square test is not really appropriate here, however, because of the extremely small expected frequency (< 1) in the cell representing studies with less than 20 trials, which is the cell responsible for the significant chi-square value. Cochran (1954) is widely cited in this context and cautions that for chi-square tests with $df > 1$, no cell should have an expected frequency < 1. Therefore, I used an exact binomial test instead of the chi-square test used by Hyman. This confirmed the apparent clustering of significant studies in the cell with < 20 trials ($p = .0006$).

[4]For the benefit of readers who may want to conduct their own analyses of the data, it should be noted that there appear to be a number of errors in the specific figures Hyman cites. I find slightly different numbers of studies and significant studies in the four classes of sample size that Hyman used. For the class with 6 to 19 trials, Hyman reports 5 of 7 studies significant; I find 6 out of 8 studies significant. For the class with 20 to 29 trials, Hyman reports 6/12 rather than 5/11. For classes 30 to 44 and 45 to 180, Hyman reports 7/14 and 5/9, respectively, whereas I find 6/12 and 6/11. Since Hyman does not document his assignment of studies to classes, it is not possible to resolve these discrepancies with the information provided in his paper. Another discrepancy occurs in the number of direct-hits studies and trials: Hyman uses the direct-hits studies with p(hit) = .25 to estimate the "true" hit rate for use in his power analysis. He says (p. 13) that there are 22 studies, comprising 746 trials. As the table in Appendix A shows, of the 28 direct-hits studies, there are actually 24 with p(hit) = .25, comprising 722 trials. The proportion of hits in these 722 trials, however, is .38, the figure Hyman reports. I also disagree with Hyman's treatment of studies that used the binary coding index. He appears to treat each binary coding trial as a single trial. This seems inappropriate for power analysis because each binary coding trial actually comprises 10 independent binary trials and the significance tests whose power is being evaluated were based on 10 binary trials per session.

[5]Hyman does not document the precise method by which he calculated the expected number of successful studies obtained in each class, but it appears that the assumed "true" hit rate of .38, the obtained hit rate for each class, and the median sample size for the class are used to obtain the power for each class, and the expected number of successful studies is obtained by multiplying the power by the median sample size. He appears to have used this estimate of the "true" hit rate even when analyzing the 18 studies that have hit probabilities *other than .25* (i.e., studies in which the chance probability of a hit was .5, .2, and .167). No rationale or justification is given for doing this, and none is apparent to me.

square root of the sample size," his reaction to the negative relationship is that "the most obvious conclusion is that such a strange relationship is due to a selective bias" (p. 14). He suggests that it is due to selective reporting of significant small studies. "This is understandable," Hyman suggests (p. 14), "in that a significant outcome is likely to be accepted for publication even if the sample size is small. But a nonsignificant study with only 5 to 19 trials is easy to dismiss as having inadequate power."

This argument, however, ignores the fact that free-response psi experiments (both significant and nonsignificant) typically have small sample sizes. The Maimonides ESP-dream studies, from which the ganzfeld work developed, generally involved seven or eight trials per series (Ullman, Krippner, & Vaughan, 1973). Sample sizes under 20 trials also characterize psi relaxation studies (Braud & Braud, 1973, 1974a, 1974b) and remote-viewing studies (Puthoff, Targ, & May, 1981). Studies failing to replicate free-response psi effects are also characterized by small sample sizes. For example, three reported failures to replicate the Maimonides dream ESP studies (Belvedere & Foulkes, 1971; Foulkes et al., 1972, and Globus, Knapp, Skinner, & Healy, 1968) involved, respectively, 8, 8, and 17 trials. Nine of 13 reported failures to replicate remote-viewing effects involved 12 or fewer trials (Allen, Green, Rucker, Goolsby, & Morris, 1976; Marks & Kammann, 1980; Rauscher, Weissman, Sarfatti, & Sirag, 1976; Solfvin, Roll, & Kreiger, 1978).

Although a selective bias of the type suggested by Hyman is a possibility, it is not the only interpretation possible nor is it strongly supported by the examples he cites. Hyman describes his retroactive-study hypothesis as follows:

> This proposed bias toward reporting small studies only if they succeed is related to what I refer to as the "retrospective study." This is the tendency to decide to treat a pilot or exploratory series of trials as a study if it turns out that the outcome happens to be significant or noteworthy. (p. 14)

He says that "two studies in the data base are clearly retrospective" (p. 14). One (Study 4) is described by its authors as a "class demonstration," and the other (Study 7) is my seven-session demonstration series with TV film crews. "If the demonstrations had not resulted in significant psi-hitting," he says, "we probably would never have heard of them" (p. 14). I disagree, but it is a moot point, and my only comment is that the TV sessions could be conducted only when a TV film crew was present and, as Hyman himself notes (p. 14), it took 16 months to collect just seven trials.

"Strong circumstantial evidence exists," Hyman continues (p. 15), "to suggest that four others of the 'significant' studies were also retrospective: Studies 2, 33, 34, 37." One of these (Study 2), as we will see later, was a small but systematic and thoughtfully conceived study which seems to have aroused Hyman's suspicions because it was published "almost 3 years after it was conducted" and "a single individual served as the experimenter and agent" (p. 15). "In the other three," Hyman says, "the authors referred to their studies as 'preliminary,' 'exploratory,' or 'pilot.' This again suggests that the only reason we are reading about them is because they gave significant results" (p. 15). But one of these (Study 37) was in fact a *nonsignificant* study and the remaining two (Studies 33 and 34) were complex and elaborate experiments with large sample sizes. Study 33, with 100 trials has the *largest* sample size of all the direct-hits studies and Study 34 has 40 trials. Both were exploratory in the sense that they introduced and attempted to assess the effects of novel conditions or experimental manipulations on psi ganzfeld performance, but neither fits Hyman's description of a "retrospective" study.

Nor does the retrospective bias hypothesis receive encouragement from Blackmore's (1980) survey of reporting practices. In addition to the 19 studies that had not (at the time of her inquiry) been published, she also found 12 studies that had not been completed. "In no case," she said, "was 'results not significant' given as the sole reason for failing to complete [the study] and therefore no selection at this stage was apparent" (p. 216). Thus, neither the examples cited nor what is known about reporting practices in this area strongly support a reporting bias interpretation of Hyman's finding.

As Hyman says, *all other things being equal,* statistical power should increase with the square root of the sample size. That all things cannot be assumed to be equal across the psi ganzfeld studies is evident by the fact that these studies varied greatly in specific instructions given, use of naïve or experienced subjects, ganzfeld duration, sender conditions (lab sender, friend of subject, no sender), use of preparatory relaxation exercises, type and orthogonality of target sets used, and so on. Indeed, we cannot assume that all things are equal within a single study.

Consider two studies in the psi ganzfeld data base that provided information on the number of sessions run per day by an individual experimenter. Habel (Study 6), with 90 sessions, found a drop in subjects' performance as the number of sessions run per day was doubled to expedite completion of the study. Habel used a partial-hits index with $p(\text{hit}) = 1/2$. She reported a decline in scoring rate from 55% in the period with 1 to 3 sessions per day, to 41% in the later

period with 5 to 8 sessions per day. Sondow [33], with 100 sessions, also found a decline in performance as the number of sessions per day increased. With a chance expectation for each session of 25%, a 51% hit rate occurred on days when she ran only one session ($n = 41$), compared to 39% on days when she ran two sessions ($n = 38$) and 24% when she ran 3 sessions per day ($n = 21$).

The Habel and Sondow studies are among the largest studies in the data base, and in both cases, a single individual served as primary experimenter. It is not implausible that an experimenter's enthusiasm and interest might change over the course of a long workday and that such changes might be communicated to subjects and reflected in their subsequent performance. Like Hyman's selective-bias interpretation, this possibility must remain conjecture until it has been explicitly studied, though it is hoped that it will be taken into consideration by future experimenters in the design of new ganzfeld studies. Even as conjecture, however, it does illustrate the danger of assuming "all things are equal" in situations involving repeated interactions among human participants and experimenters.

Regardless of the interpretation of the excess of significant outcomes in studies with small sample sizes, this finding does not materially affect the overall significance of the ganzfeld data base. Even if we use only studies with 20 or more trials, the success rate of the larger sample size studies is not substantially diminished. Using the original classification (used in Hyman's analysis) we find that 17 of the studies with 20 or more trials are significant and 17 are nonsignificant, a success rate of 50%. Using the revised classification, which corrects for multiple analysis, the result is 14 significant and 20 nonsignificant experiments (41%). Of the 28 direct-hits studies, 22 have 20 or more trials, and the cumulative results for these is nearly the same as for all 28 studies. The mean Z score is 1.24 (SD = 1.48) and the Stouffer $Z = 5.83$ ($p = 2.8 \times 10^{-9}$). Ten of the 22 studies, or 45%, are independently significant ($p = 3.62 \times 10^{-8}$). (See table in Appendix A.)

Summary. Although selective reporting bias can never be conclusively refuted, a number of considerations strongly mitigate against the likelihood of a serious reporting problem in this area: (a) the publication policies and practices in parapsychology show that reporting of null results is commonplace; (b) a large number of the existing ganzfeld studies report null results; (c) the file-drawer estimate of the number of fugitive null studies required to wash out the known results requires 15 such fugitive studies for each one known; and (d) Hyman's hypothesis that there may be a tendency to report

small studies only if they are significant is not strongly supported by the examples he cites and is inconsistent with the literature on free-response psi research, which shows that both significant and nonsignificant free-response studies typically have small sample sizes.

Study Quality

We have seen that the cumulative psi ganzfeld effect remains highly significant when evaluated by a single uniform test and index, that the effect is not dependent on the studies of one or two investigators, and that the cumulative effect cannot be attributed to selective reporting bias without assuming the existence of a large number of unreported studies averaging null outcomes. The second stage of our meta-analysis attempts to account for some of the variability in study outcomes by examining their relationship to procedural variations across studies. Specifically, we will be concerned with the relationship between study outcome and procedural variations related to study quality. One of the principal advantages of meta-analysis over traditional narrative reviews is that it seeks empirical assessment of methodological issues rather than relying on a priori judgments of research quality. Glass, McGaw, & Smith (1981) express the attitude of meta-analysis as follows:

> An important part of every meta-analysis with which we have been associated has been the recording of methodological weaknesses in the original studies and the examination of their covariance of study findings. Thus, the influence of "study quality" on findings has been regarded consistently as an empirical a posteriori question, not an a priori matter of opinion or judgment used in excluding large numbers of studies from consideration. (pp. 221–222)

The general procedure is to define and encode relevant study features, then use statistical analysis to evaluate the impact of variations across studies on their outcomes. A number of problems arise in the course of this process. Some are due to ambiguous specification of features to be encoded and inconsistencies in encoding them. In meta-analysis of controversial research domains such as psi research, it is especially important that the study variables to be encoded be defined as unambiguously as possible to allow independent reexamination by other reviewers. The criteria used by a meta-analytic reviewer should be specified (and documented) in such a way that others can, by going to the original research reports, reconstruct the

analysis and satisfy themselves as to the appropriateness of the original coding and analysis. As Cooper (1984) advises, reviewers should "open their rules of inference to public inspection" (p. 111).

For the purpose of my own meta-analysis of study quality, I have defined variables to be encoded in terms of procedural descriptions (or their absence) in the research reports, and I have avoided as much as possible, making inferences that go beyond what is given in the reports. As in the preceding section, my analysis is limited to studies that used a uniform test and index, which will eliminate concern over multiple-analysis options. After presenting my own meta-analysis, I shall describe what I believe are serious problems in Hyman's approach and document specific instances to illustrate the problems.

Sensory Cues

Since the ganzfeld is a perceptual isolation procedure, it eliminates potential sensory contact between percipient and target during the session. The percipient's auditory and visual input typically consists of white noise and an unpatterned visual field. Sender and target are isolated in a different room. Except for the possibility of deliberate electronic sabotage, these procedures prohibit conventional information exchange between sender and receiver during the psi ganzfeld session.

A channel for potential sensory cues does exist, however, in the judging phase of blind-judging studies in which the same target set used by the sender (or by an experimenter in clairvoyance studies) during the session is then used by the percipient for blind-judging at the end of the session. In these cases, a sender or experimenter may have physically handled the target, enabling transmission of potential cues concerning target identity in the form of fingerprints, smudges, or other markings that might differentiate target from decoys. Though it might be argued that the likelihood of handling cues would be diminished in clairvoyance studies where there was no sender, there is no way to eliminate the possibility of such cues in studies that used single target-sets, and I have made no attempt to differentiate studies within that class.

CUE ratings. For the purpose of the present analysis, a cue rating (CUE) was assigned to each study on the basis of procedural descriptions in the reports. A CUE rating of 2 was assigned to studies reporting the use of duplicate target-sets, which eliminate possible handling cues. Studies using single target-sets were given a CUE rating of 1. A CUE rating of 0 was assigned to one study (Study 10)

that, in addition to the use of a single target pack, provided two other opportunities for sensory cues: (a) During the sending period, the experimenter-sender rolled a clay ball over the target, increasing the possibility of handling cues; and (b) after the sending period, the experimenter-sender had sensory contact with a second experimenter who later supervised the percipient's judging. (The outcome of the study was nonsignificant, $Z = -.57$.)

Effect of CUE control procedures on study outcome. CUE ratings for all 28 direct-hits studies are given in column 2 of Table 4. Although studies with better controls against sensory cues (CUE = 2) were slightly more successful than those permitting handling cues (CUE < 2) (t [26] = .318; p = .687), this finding is limited because all but one of the studies with a CUE of 2 come from one laboratory. The correlation between cue control and study outcome is nonsignificant (r[26] = .134).

Studies eliminating potential handling cues. Ten studies used duplicate target-sets for sender and percipient judging (CUE = 2). The mean Z score was 1.38, and the combined (Stouffer) Z was 4.35 ($p = 6.8 \times 10^{-6}$). Half the studies in this group (5/10) were independently significant (p = .000064, exact binomial test with p = .05 and q = .95).

Binary coding studies. From the standpoint of cue control, studies using the binary coding index deserve special attention because the procedure does not involve exposing the subject to a judging pool, so no opportunity for transmission of handling cues exists. This index was used in five studies that do not overlap with the direct-hits studies described above (Studies 3, 20, 32, 37, 40) and was the only index possible for all but one of these studies (Study 3). In Study 3, a blind-judging partial-hits index was also used, with the judging taking place *after* the binary coding (p. 415). The combined (Stouffer) Z score for the binary coding studies[6] was 2.84 (p = .0023). Three of the five studies (by two of the three contributing investigators) were independently significant (p = .0012, exact binomial test with p = .05 and q = .95).

Randomization

Psi ganzfeld studies have used a number of methods for target selection. For the direct-hits studies, tables of random numbers or

[6] I have not cumulated Z scores for binary coding and direct-hits blind-judging studies because the number of trials going into the two cases differ so much. The direct-hits studies have sample sizes ranging from 7 to 100 trials, and the binary coding studies have sample sizes ranging from 150 to 1800 trials.

random number generators were reported in 16, or 57%, of the studies; hand-shuffling and related methods (die-casting and coin-flipping) were used in 7, or 25%, of the studies. Numbered poker chips were shaken together and selected by hand in 2 of the remaining 5 studies (Studies 16, 17). Randomization was not described in the other three reports.

What would constitute a randomization problem in these studies is not entirely clear. In 20 studies, 71% of the total, subjects contributed only one trial each (for study documentation, see the table in Appendix A). For each subject, then, this amounts to one random selection with, usually, $P = 1/4$. It is not clear that random number tables provide better randomization than shuffling techniques when a separate randomization is used for each trial. The single-trial-per-subject studies are independently quite significant (Stouffer $Z = 4.61$; $p = .000002$), with 8 of the 20 studies, or 40%, individually significant at the 5% level (exact $p = .000003$, with $p = .05$ and $q = .95$), and the mean Z score for this group does not differ significantly from that for studies with multiple trials per subject ($t[26] = 1.17$; $p = .126$, one-tailed). Further, I suspect that if all the studies under consideration had used random number tables as the method of target selection, questions might once again be raised concerning peculiarities of random number tables (e.g., see Spencer Brown, 1953, 1957). Similarly, if all the studies used random number generators, critics pursuing alternative explanations of putative psi effects might reasonably request specifications of generator characteristics and performance. As it is, it might be best that a variety of methods have been used, *if* it can be shown that study outcomes are independent of the method of randomization. For the purpose of analysis, however, I have adopted what is surely the most popular opinion, that use of random number tables or generators is superior to hand-shuffling and related methods.

RAND ratings. Studies reporting the use of random number tables or random number generators for target selection were assigned a RAND rating of 2. Studies in which target selection was based on card-shuffling, coin-flipping or die-casting were given RAND ratings of 1. RAND ratings of 0 were assigned for any other method of target selection or when the method of randomization was not specified.

Effect of randomization procedures on study outcome. RAND ratings are given for each study in column 3 of Table 4. The correlation between RAND ratings and study outcome (Z scores) is nonsignificant ($r[26] = -.095$). The outcome of studies using random number tables or generators (RAND = 2) does not differ significantly from that of

TABLE 4
METHODOLOGY RATINGS AND STUDY OUTCOME

Study	CUE[a]	RAND[b]	*Z* score
	Studies with Z > median		
16	1	0	4.02
33	1	1	3.41
39	1	1	3.24
26	2	2	3.15
31	1	2	3.11
7	1	1	3.00
42	1	2	2.89
30	2	2	2.16
1	2	2	2.15
8	1	1	2.02
24	2	2	1.74
25	2	2	1.74
38	1	1	1.62
27	2	2	.97
Means	1.43	1.50	
	Studies with Z < median		
34	2	2	.92
21	1	1	.79
2	1	2	.76
17	1	0	.76
19	1	0	.76
23	2	2	.48
18	1	0	.25
28	2	2	.24
29	2	0	.21
11	1	2	− .39
10	0	2	− .57
41	1	2	− .82
4	1	1	−1.71
12	1	2	−1.97
Means	1.21	1.28	

[a]CUE ratings are as follows: 2 = use of duplicate target-sets documented in report; 1 = single target-set described; 0 = other potential sources of cues evident in report.

[b]RAND ratings are as follows: 2 = report describes target selection using random number tables or generators; 1 = target selection involving shuffling techniques described; 0 = any other randomization technique, or method not described.

studies using other randomization methods or studies not specifying the method of randomization (RAND < 2) (t [26] = –.824, p = .422).

Studies using random number tables or generators. Of the 28 direct-hits studies, 16 reported target-selection procedures based on random number tables or random number generators[7] (RAND = 2). The mean Z score for these studies was 1.04, and the combined (Stouffer) Z was 4.14 (p = .000017). Seven of the 16 studies, or approximately 44%, were independently significant at the %5 level ($p = 5.98 \times 10^{-6}$). The studies in this group were contributed by six different investigators.

Covariation of Study Quality and Outcome

The joint effects of cue control and randomization method on study outcome were evaluated through a multiple regression analysis, with CUE and RAND ratings as the independent variables and study Z score as the dependent variable. The resulting multiple correlation of .195 is nonsignificant (F[2, 25] = .493, p = .613). Thus, there appears to be no systematic relationship between these indices of study quality and study outcomes.

Hyman's Flaw Classification

Hyman's Initial Tally of Flaws

Hyman first reported an analysis of flaws in his earlier review of psi ganzfeld research (Hyman, 1983, p. 23). "In the data base of 42 studies," he said, "the three most common flaws were multiple tests of significance (64%), possibilities of sensory leakage (60%), and inadequate randomization (45%)." He obtained an overall flaw count by tallying the number of flaws he found in the reports of each study, and then he compared the average number of flaws in the significant versus the nonsignificant studies. He reported a significant difference in the average number of flaws for the two groups (t[40] = 2.85; p <

[7]The reports of 7 of the 16 studies in this group provided specific descriptions of how the random number table or generator was used in target selection (Studies 2, 11, 12, 24, 30, 31, 34), whereas the reports of the remaining 9 studies simply stated that such methods were used (Studies 1, 10, 23, 25–28, 41, 42). This difference was not related to study outcome (Z score) (t[14] =–.3, p = .77).

.01) and concluded, "There is a strong tendency for the rate of success to increase with the number of obvious defects."

This analysis was straightforward and easy to interpret. If the classification of flaws was correct, Hyman would have clearly demonstrated a link between successful outcomes and procedural flaws in the studies. Because my own analysis of the covariation of study quality and outcome differed so markedly from Hyman's, I requested study-by-study documentation of his flaw classification prior to the cojoint SPR and PA meeting in Cambridge. The document I received (personal communication from Ray Hyman, July 29, 1982) contained a large number of errors in the assignment of flaws to studies, which, I was later informed (personal communication from Ray Hyman, November 29, 1982), were typing errors rather than errors in classification.

Hyman's Second Classification of Flaws

The November 1982 communication was accompanied by a revised classification and analysis based on five categories of flaws: the three just described, plus inadequate documentation (defined the same way as in his current analysis) and feedback (assigned when "no precautions had been taken to insure that the target reported by the agent at feedback had, indeed, been the target actually used"). Hyman's flaw assignments again led to a significant difference, with more flaws assigned to significant than to nonsignificant studies ($t[40] = 2.89$; $p = .006$), and I again seriously objected to many of his classifications, such as the assignment of a feedback flaw to a clairvoyance study in which there were no senders (agents).

Hyman's Current Classification of Flaws

Hyman's present classification of flaws is thus the third iteration, and it is appropriate that our differences in the assessment of the quality of psi ganzfeld research be made available for evaluation by the research community.

In his present paper, Hyman has six categories of procedural flaws. As described in the previous section, I have dealt with multiple analysis through the use of a uniform index and test and am therefore omitting consideration of Hyman's six multiple analysis flaws (three of which he also discards from his analysis, and none of which, as we have seen, correlates with study outcome). Hyman's procedural flaw

categories now include sensory cues, randomization, feedback (with a revised definition), security, documentation, and statistical errors. I have serious objections to Hyman's definition and coding of some of these flaws. In what follows, I shall briefly discuss his categories and provide examples illustrating problems in flaw definition and attribution. I shall then examine in some detail Hyman's flaw ratings of a single study to show how his count greatly inflates the number of flaws in the study. Finally, I shall consider his current analysis.

I begin by noting two areas in which Hyman and I are in agreement.

Single target (ST). The definition of ST is very similar to my CUE criteria, and the two indices are highly correlated ($r[26] = .94$). Hyman says, "My analysis agrees with Honorton in showing no correlation between the use of single targets and significance" (p. 30). For the direct-hits studies, the correlation between ST and study outcome (Z score) is close to zero ($r[26] = -.062$).

Statistical errors (STAT). I agree with Hyman that six of the direct-hits studies contain statistical errors (Studies 4, 16, 17, 31, 33, 34).

Feedback (FB). Hyman claims that although ST does not correlate with study outcome, the FB flaw that can only occur with ST does correlate with success. He assigned an FB flaw to ST studies that "typically did not use an adequate procedure to insure that the target was properly randomized among the other candidates in the pool before being presented for judging" (p. 28). His appendix definition, however, adds a second condition for the assignment of FB: "Inadequate randomization of target and foils at judging, *or* inadequate precautions against communication from percipient *to* agent at feedback" (p. 44, emphasis added). Hyman provides no elaboration or commentary on this second condition, how it was evaluated from the research reports, or why he included it under FB rather than under security (SEC), for which he has another flaw category. It should be noted that this second definition is the same definition of feedback given in Hyman's second classification (November 1982) and that the assignment of FB in the second classification was not limited to studies using single target-sets. The formulation given implies deliberate fraud: The sender, having received feedback of the percipient's choice, substitutes the target chosen by the percipient for the actual target, thereby creating a spurious hit. (The first condition for FB, which is the judging order, may also imply cheating, with the sender [-experimenter] in single-target experiments placing the actual target in a constant location in the judging pack.) I shall use the designations

TABLE 5
ORDERING OF TARGETS IN DIRECT-HITS STUDIES WITH SINGLE TARGET-SETS

Study	JORD rating[a]	*Z* score
16	0	4.02
33	2	3.41
39	1	3.24
31	2	3.11
7	1	3.00
42	2	2.89
8	1	2.02
38	1	1.62
21	0	.79
2	2	.76
17	0	.76
19	0	.76
18	0	.25
11	2	− .39
10	2	− .57
41	2	− .82
4	1	−1.71
12	2	−1.97

[a] JORD refers to the method of ordering the targes for judging. The JORD ratings are as follows: 2 = report describes target and decoys presented in numerical or alphabetical order; 1 = report describes judging order via hand-shuffling; 0 = order of targets at judging not described.

FB-1 and FB-2 to identify the two parts of Hyman's definition of this flaw.

Hyman is not consistent in his flaw assignments. He assigned FB to 10 of the 18 direct-hits studies using single target-sets (Studies 2, 4, 7, 8, 16, 17, 21, 33, 38, 39). The reports of all but three of these (Studies 16, 17, 21) describe procedures for ordering targets at judging. And he did not charge an FB flaw against two studies that failed to describe the method of ordering targets at judging (Studies 18, 19).

Assessment of FB-1. To evaluate the effects of these procedures, I coded each of the 18 direct-hits studies as follows: A rating of 2 was given for each study that described the arrangement of targets in numerical (or alphabetical) order; a rating of 1 was given to studies describing the reordering of targets by hand-shuffling, and studies that did not report the method of ordering targets for judging were

given a rating of zero. These ratings are listed in Table 5. There is no significant relationship between target-ordering procedures and study outcome (*Z* score): The correlation (with 16 *df*) is –.138.

Except for two cases (Studies 7, 21), the studies to which Hyman has assigned FB flaws explicitly report experimenter monitoring of feedback. Examination of some of the studies assigned FB flaws makes it clear that if the study is to be faulted for FB, it is FB-2 that is operative and that the experimenter would have to be implicated in any cheating scenario. In Study 33, for instance, significance was contributed primarily by subjects in one of the two experimental groups, who received feedback only after completing four of their five trials. The report also indicates that the targets were displayed in numerical order at judging:

> Subjects in the Association group were asked to read a set of instructions before looking at the four pictures, *which were laid out on the desk in numerical order.* Their associations to each of the pictures were then recorded by the experimenter on a separate sheet of paper. Subjects were allowed to read through their complete records while making these associations, and were encouraged to use them in making their ranking decisions. After the experimenter recorded these decisions, the subject left the building. *The experimenter then opened the envelope containing the number of the target picture and recorded it.* (Sondow, 1979, p. 132, emphasis added)

Randomization (R). Hyman defines a randomization flaw (p. 44) as "inadequate randomization" or "inadequately described" randomization. He says that he considered studies to have adequate randomization when the authors describe "using a table of random numbers or a random number generator to select the specific target from a pool" (p. 27). He now finds that 74% of the studies failed these criteria and were guilty of "suboptimal" randomization—up from 45% in his first and 52% in his second evaluations of the same studies. For the direct-hits studies, Hyman considers that only 5 studies or 18% used "adequate" randomization (Studies 10–12, 31, 34). A typical example is Study 10: "J.P. selected both the set and the picture within the set that was to be the target by referring to a random number table" (p. 50).

Again, Hyman has not consistently applied his own criteria in coding the studies; many of those he cites as having "inadequate" or "suboptimal" randomization contain descriptions that satisfy his stated criteria (e.g., Studies 1, 2, 23, 41, 42). The authors of Study 41, for instance, report that "the [experimenter-]agent used card cuts and random number tables to choose one slide out of 1,024" (p. 84).

Similarly, the randomization procedure in Study 42 describes "individual targets . . . selected using a random number generator from each mini-target pool by an otherwise uninvolved assistant" (pp. 48–49).

Security (SEC). Hyman's definition of a security flaw is "inadequate security, usually in monitoring crucial phases of the study or in having only one experimenter" (p. 44). SEC was also assigned for "failing to monitor the agent" (p. 28). I will discuss the single-experimenter issue below in regard to a study (Study 2) that illustrates a valid one-experimenter design. It is unclear both from Hyman's definition and his assignments just what other "crucial phases" of an experiment he has in mind that are not already covered by his flaw categories for sensory cues, "feedback," and randomization problems.

Documentation (DOC). Concern over possibilities of sender-receiver cheating appears to supply the impetus for another Hyman flaw category, documentation (DOC), which we are told was usually assigned for "failure to report the number of times the agent was a friend of the percipient or to provide data on whether this made a difference in those studies in which subjects were encouraged to bring their own agents" (p. 28). Since Hyman's interest is whether there is a psi effect, not with its relationship to psychological variables, his focus on sender-receiver relations appears to be motivated by security concerns. As we have seen, security is a separate category of flaw that already covers monitoring the sender. One of the studies faulted was a clairvoyance study in which there were no senders (Study 17), and Hyman does not specify what other conditions prompted his assignment of DOC flaws.

Assessment of sender-receiver documentation. To evaluate study outcomes in relation to the one positive condition that Hyman specifies for the DOC flaw, I performed a sender-receiver DOC (SR-DOC) rating on each of the direct-hits studies. Ratings of 1 were given to studies that satisfied one of three conditions: (a) The report specified that the sender was always the experimenter; (b) sender-receiver breakdowns were provided; or (c) the study used clairvoyance procedures and there were no senders. The correlation between SR-DOC and study outcome is nonsignificant (point-biserial $r[26] = -.165$). My SR-DOC ratings are given in Table 6.

Why There Are So Many Flaws: A Detailed Example

I have described several examples of what I consider inappropriate flaw coding. To illustrate how these problems can grossly overesti-

TABLE 6
SENDER/RECEIVER DOCUMENTATION IN DIRECT-HITS STUDIES

Study	SR-DOC rating[a]	*Z* score
16	1	4.02
33	1	3.41
39	0	3.24
26	1	3.15
31	1	3.11
7	0	3.00
42	1	2.89
30	1	2.16
1	1	2.15
8	0	2.02
24	1	1.74
25	1	1.74
38	0	1.62
27	1	.97
34	0	.92
21	0	.79
2	1	.76
17	1	.76
19	1	.76
23	1	.48
18	1	.25
28	1	.24
29	0	.21
11	1	− .39
10	1	− .57
41	1	− .82
4	1	−1.71
12	1	−1.97

[a]The rating for sender-receiver documentation (SR-DOC) is as follows: 1 = experimenter always sender, or no sender (clairvoyance), or sender/receiver pairing specified; 0 = sender/receiver pairing not specified.

mate the flaw count for an individual study, I will use one of Hyman's frequently used examples, the study by Braud, Wood, & Braud (1974). This is Study 2 in the present data base and is one that Hyman suggested was a "retrospective" study. Hyman charges this study with four different procedural flaws: ST, R, FB, and SEC. As with all of the early ganzfeld studies, a single target-set was used by the experimenter-sender and by the subject for judging, so there is no question that this

study should be faulted for ST. The remaining three flaws Hyman assigns to this study are, I believe, inappropriate. The study is charged with "suboptimal" randomization. Yet, ironically, it has one of the most complete descriptions of randomization using random number tables in the data base, one that is as complete as any that Hyman has acquitted of randomization flaws:

> Target preparation and selection techniques were identical to those described in Braud and Braud (1974b). The [experimenter-]agent randomly selected (using card cuts, coin tosses, and a table of random numbers) one out of a pool of 20 packs, then one of the six pictures within that chosen pack as the actual target. (Braud, Wood, & Braud, 1975, p. 108)

The auxiliary citation (Braud & Braud, 1974b) provides details of the method used to obtain an entry point in the random number table and how the entry point was used to select a specific target and decoys for each session:

> After leaving the subject, [the experimenter-agent] randomly selected first a pack and then an envelope within that pack through the use of a 40-row × 40-column table of random numbers in which the entry point was determined by two cuts of a well-shuffled deck of cards bearing the numbers 1 through 40 and a coin toss to determine row vs. column. The chosen envelope contained the target for that session and the others were the controls (p. 232).

Compare this description with that for Study 10 which Hyman approved of (see the paragraph entitled *Randomization* in the preceding section).

Hyman's assignment of FB and SEC flaws to Study 2 appears to be due to the use of a single experimenter or agent but is clearly inappropriate given the protocol described in the report. Following Hyman's definition, FB should not be charged against this study because the report states that the target and decoys were replaced in their original numerical order and because there was no opportunity for communication between subject and experimenter-agent until after the subject completed judging:

> At the termination of the psi impression period, the subject self-terminated his hypnagogic state [signaled by five thumping sounds recorded at the end of a continuously playing tape], recorded his impressions on paper, and read the instructions inside an envelope which had been placed in his room before the session began. At the same time, the [experimenter-] agent recorded the code number of the correct target on his data sheet, replaced all six pictures in their individual

> envelopes (in their original numerical order) back into their larger envelope. He then placed the packet of six pictures on a stool outside the still-closed door of the subject's room and returned to his room without sensorially encountering the subject. (Braud, Wood, & Braud, 1975, p. 108)

It is clear from this description that considerable care was taken by the investigators to eliminate security problems. Elsewhere, the authors describe the one-way signaling system used to mark the beginning and end of the 5-min sending–psi-impression period. It seems excessive to charge this experiment with a security lapse only because there was a single experimenter. If anything, the design of this experiment illustrates how an isolated investigator working alone might conduct an adequately controlled psi ganzfeld experiment. If a duplicate target-set had been included in the subject's postsession judging packet, I believe the Braud-Wood-Braud protocol would be completely adequate.

In summary, whereas Hyman has charged this study with four procedural flaws, consistent application of his flaw criteria suggests that it should be charged only with one.

Summary on Hyman's Flaw Classification

It is clear that Hyman's assignment of flaws is itself seriously flawed. There are problems in the definition of several of his flaw categories, largely owing to vagueness in specifying codable characteristics of the flaw (e.g., "inadequate security, usually in monitoring crucial phases of the study . . ."), and to the use of disjunctive definitions (FB-1 *or* FB-2). In addition, Hyman has been inconsistent in his assignment of flaws, with the effect of spuriously increasing the flaw count in some studies that appear to satisfy his stated criteria and decreasing the flaw count in other studies that fail his criteria. Interested readers who are willing to consult the research reports can verify this for themselves.

Hyman's current analysis of flaws. Given these problems, any statistical analysis involving Hyman's flaw ratings would be uninterpretable. My response, however, would be incomplete without a brief comment on Hyman's current analysis. Unlike his two earlier evaluations, which involved a straightforward *t* test of the relationship between flaws and study outcome, Hyman now performs a factor analysis to demonstrate such a relationship. In view of his earlier statement about "being startled by investigators routinely doing factor analyses on sample sizes of 30 or less" (p. 6), it comes as somewhat of a surprise to

find him now basing his own evaluation on a factor analysis involving 36 cases! His analysis is sufficiently complex that it seemed advisable to me to have it evaluated by someone well versed in factor analysis, and I have asked a psychological statistician, David R. Saunders, to examine it. Saunders's evaluation is presented in Appendix B; his conclusion is that both the factor analysis and Hyman's interpretation of it are faulty.

Discussion

Is there a significant psi ganzfeld effect? I believe my evaluation of direct-hits studies justifies an affirmative answer to this question. When multiple analysis problems are eliminated through use of a uniform test and index, the effect remains highly significant. And though selective reporting bias cannot be conclusively ruled out, consideration of reporting practices in this area and the file-drawer estimate of the extent of selective reporting necessary to jeopardize the known data base indicate that selective reporting bias does not pose a serious problem.

Does the ganzfeld paradigm represent a step toward replicability of psi effects? Significant direct-hits studies have been reported by 6 of the 10 contributing investigators. Even though the number of investigators may be too small to allow a firm conclusion to be reached, this result is certainly encouraging. New replication efforts are, however, clearly needed and, ideally, by as many new replicators as possible.

On the evaluation of study quality, many readers will find it disconcerting that two reviewers should come to such divergent conclusions in evaluating the same set of studies. Neither Hyman nor I conducted our evaluations of study quality on a blind basis, and it would not be unreasonable for readers to suppose that the disagreement mainly reflects our respective a priori views. This is a matter that the reader will have to decide. I am hopeful that at least a few readers will want to consult the original studies and make their own determination.

If we are to be more successful in achieving consensus over future studies, we must be able to agree in advance on the criteria that will be used to assess them. This is crucial, and the absence of such agreement in the past has, in my opinion, contributed heavily to the perennial stand-off between psi researchers and critics. Hyman and I are in substantial agreement on the need to improve study documentation. Clearly there is work to be done both in improving the level of procedural description and in specifying the potentially

important moderators. The psi ganzfeld research has been underway now for a decade, and it is reasonable to expect some degree of standardization in reporting work in this area. For this reason, the Council of the Parapsychological Association has commissioned a study group to develop specific guidelines for reporting psi ganzfeld studies and research in other areas that have been ongoing over a substantial period of time. The study group will consist of researchers with varied outcome histories, critics, and editors of PA-affiliated journals. I am pleased that Hyman has agreed to participate in the development of guidelines for future ganzfeld studies.

Appendix A
The Data Base

Study numbers in this data base are the same as those used by Hyman. References preceded by an asterisk represent the ones for which longer unpublished reports were provided to Hyman.

I. Studies Using Direct-Hits Index

1. Ashton, H. T., Dear, P. R., Harley, T. A., & Sargent, C. L. (1981). A four-subject study of psi in the ganzfeld. *Journal of the Society for Psychical Research*, **51**, 12–21.
2. Braud, W. G., Wood, R., & Braud, L. W. (1975). Free-response GESP performance during an experimental hypnagogic state induced by visual and acoustic ganzfeld techniques: A replication and extension. *Journal of the American Society for Psychical Research*, **69**, 105–113.
4. Child, I. L., & Levi, A. (1979). Psi-missing in free-response settings. *Journal of the American Society for Psychical Research*, **73**, 273–289.
7. Honorton, C. (1976). Length of isolation and degree of arousal as probable factors influencing information retrieval in the ganzfeld. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975*(pp. 184–186). Metuchen, NJ: Scarecrow Press.
8. Honorton, C., & Harper, S. (1974). Psi-mediated imagery and ideation in an experimental procedure for regulating perceptual input. *Journal of the American Society for Psychical Research*, **68**, 156–168.
10. Palmer, J., & Aued, I. (1975). An ESP test with psychometric objects and the ganzfeld: negative findings. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1974* (pp. 50–53). Metuchen, NJ: Scarecrow Press.
11. Palmer, J., Bogart, D. N., Jones, S. M., & Tart, C. T. (1977). Scoring patterns in an ESP ganzfeld experiment. *Journal of the American Society for Psychical Research*, **71**, 121–145.

12. Palmer, J., Khamashta, K., & Israelson, K. (1979). An ESP ganzfeld experiment with Transcendental Meditators. *Journal of the American Society for Psychical Research*, **73**, 333–348.
16. Raburn, L. (1975). *Expectation and transmission factors in psychic functioning*. Unpublished honors thesis, Tulane University, New Orleans, LA. [GESP cell]
17. Ibid. [Clairvoyance cell]
18. Rogo, D. S. (1976). ESP in the ganzfeld: An exploration of parameters. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975* (pp. 174–176). Metuchen, NJ: Scarecrow Press. [Experiment 1]
19. Ibid. [Experiment 2]
21. Rogo, D. S., Smith, M., & Terry, J. (1976). The use of short-duration ganzfeld stimulation to facilitate psi-mediated imagery. *European Journal of Parapsychology*, **1**, 72–77.
23. Sargent, C. L. (1980). Exploring psi in the ganzfeld. *Parapsychological Monographs* (No. 17). [Experiment 1]
24. Ibid. [Experiment 2]
25. Ibid. [Experiment 3]
26. Ibid. [Experiment 5]
27. Ibid. [Experiment 6]
*28. Sargent, C. L., Bartlett, H. J., & Moss, S. P. (1982). Response structure and temporal incline in ganzfeld free-response GESP testing. In W. G. Roll, R. L. Morris & R. White (Eds.), *Research in parapsychology, 1981* (pp. 79–81). Metuchen, NJ: Scarecrow Press.
29. Sargent, C. L., Harley, T. A., Lane, J., & Radcliffe, K. (1981). Ganzfeld psi-optimization in relation to session duration. In W. G. Roll & J. Beloff (Eds.), *Research in parapsychology, 1980* (pp. 82–84). Metuchen, NJ: Scarecrow Press.
*30. Sargent, C. L., & Matthews, G. (1982). Ganzfeld GESP performance with variable duration testing. In W. G. Roll, R. L. Morris, & R. White (Eds.), *Research in parapsychology, 1981* (pp. 159–160). Metuchen, NJ: Scarecrow Press.
31. Schmitt, M., & Stanford, R. G. (1978). Free-response ESP during ganzfeld stimulation: The possible influence of menstrual cycle phase. *Journal of the American Society for Psychical Research*, **72**, 177–182.
33. Sondow, N. (1979). Effects of associations and feedback on psi in the ganzfeld: Is there more than meets the judge's eye? *Journal of the American Society for Psychical Research*, **73**, 123–150.
*34. Sondow, N., Braud, L., & Barker, P. (1981). Target qualities and affect measures in an exploratory psi ganzfeld. In W. G. Roll, R. L. Morris, & R. White (Eds.), *Research in parapsychology 1981* (pp. 82–85). Metuchen, NJ: Scarecrow Press.
38. Terry, J. C., & Honorton, C. (1976). Psi information retrieval in the ganzfeld: Two confirmatory studies. *Journal of the American Society for Psychical Research*, **70**, 207–217. [Experiment 1].

TABLE A1
THE DIRECT-HITS STUDIES

Study	No. of subjects	No. of trials	Direct hits	p(hit)	Proportion of hits	Z score	Effect size
1	4	32	14	.25	.44	2.15	.37
2	10	10	3	.167	.30	.76	.31
4	14	14	0	.20	.00	−1.71	− .93
7	4	7	6	.25	.86	3.00	1.33
8	30	30	13	.25	.43	2.02	.38
10	40	40	6.5[a]	.20	.16	− .57	− .32
11	30	30	7	.25	.23	− .39	− .05
12	20	20	2	.25	.10	−1.97	− .40
16	10	10	9	.25	.90	4.02	1.44
17	10	10	4	.25	.40	.76	.32
18	28	28	8	.25	.29	.25	.09
19	1	10	4	.25	.40	.76	.32
21	20	20	7	.25	.35	.79	.22
23	26	26	8	.25	.31	.48	.13
24	20	20	9	.25	.45	1.74	.42
25	20	20	9	.25	.45	1.74	.42
26	30	30	16	.25	.53	3.15	.58
27	3	36	12	.25	.33	.97	.18
28	32	32	9	.25	.28	.24	.07
29	40	40	11	.25	.28	.21	.07
30	26	26	12	.25	.46	2.16	.44
31	20	20	12	.25	.60	3.11	.73
33	20	100	41	.25	.41	3.41	.34
34	40	40	13	.25	.33	.92	.18
38	12	27	11	.25	.41	1.62	.34
39	6	60	27	.25	.45	3.24	.42
41	24	48	10	.25	.21	− .82	− .10
42	49	49	18.5[a]	.20	.38	2.89	.40

Note. Effect size index is Cohen's h. See Cohen (1977), pp. 179–213.
[a] Ranks derived from tied ratings.

39. Ibid. [Experiment 2]
41. Wood, R., Kirk, J., & Braud, W. (1977). Free response GESP performance following ganzfeld stimulation vs. induced relaxation, with verbalized vs. nonverbalized mentation: A failure to replicate. *European Journal of Parapsychology*, **1**, 80–93.
*42. York, M. (1977). The defense mechanism test (DMT) as indicator of psychic performance as measured by a free-response clairvoyance test using a ganzfeld technique. In W. G. Roll & R. L. Morris (Eds.), *Research in parapsychology, 1975* (pp. 48–49). Metuchen, NJ: Scarecrow Press.

II. Studies Using Binary-Coding Index

3. Braud, W. G., & Wood, R. (1977). The influence of immediate feedback on free-response GESP performance during ganzfeld stimulation. *Journal of the American Society for Psychical Research*, **71**, 409–427.
20. Rogo, D. S. (1977). A preliminary study of precognition in the ganzfeld. *European Journal of Parapsychology*, **2** (1), 60–67.
*32. Smith, M., Tremmel, L., & Honorton, C. (1976). A comparison of psi and weak sensory influences on ganzfeld mentation. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975* (pp. 191–194). Metuchen, NJ: Scarecrow Press.
37. Terry, J. C. (1976). Comparison of stimulus duration in sensory and psi conditions. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975*. Metuchen, NJ: Scarecrow Press.
40. Terry, J. C., Tremmel, L., Kelly, M., Harper, S., & Barker, P. (1976). Psi information rate in guessing and receiver optimization. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975* (pp. 194–198). Metuchen, NJ: Scarecrow Press.

III. Studies Using Other Indices

5. Dunne, B. J., Warnock, E., & Bisaha, J. P. (1977). Ganzfeld techniques with independent rating for measuring GESP and precognition. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1976* (pp. 41–43). Metuchen, NJ: Scarecrow Press.
6. Habel, M. M. (1976). Varying auditory stimuli in the ganzfeld: The influence of sex and overcrowding on psi performance. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975* (pp. 181–184). Metuchen, NJ: Scarecrow Press.
*9. Keane, P., & Wells, R. (1979). An examination of the menstrual cycle as a hormone related physiological concomitant of psi performance. In W. G. Roll (Ed.), *Research in parapsychology, 1978* (pp. 72–74). Metuchen, NJ: Scarecrow Press.
13. Palmer, J., Whitson, T., & Bogart, D. N. (1980). Ganzfeld and remote viewing: A systematic comparison. In W. G. Roll (Ed.), *Research in parapsychology, 1979* (pp. 169–171). Metuchen, NJ: Scarecrow Press.
14. Parker, A. (1975). Some findings relevent to the change in state hypothesis. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1974* (pp. 40–42). Metuchen, NJ: Scarecrow Press.
15. Parker, A., Millar, B., & Beloff, J. (1977). A three-experimenter ganzfeld: An attempt to use the ganzfeld technique to study the experimenter effect. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1976* (pp. 52–54). Metuchen, NJ: Scarecrow Press.
22. Roney-Dougal, S. M. (1982). A comparison of psi and subliminal

perception: A confirmatory study. In W. G. Roll, R. L. Morris, & R. White (Eds.), *Research in parapsychology, 1981* (pp. 96–99). Metuchen, NJ: Scarecrow Press.

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

*36. Stanford, R. G., & Neylon, A. (1975). Experiential factors related to free-response clairvoyance performance in a sensory uniformity setting (ganzfeld). In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1974* (pp. 89–93). Metuchen, NJ: Scarecrow Press.

Appendix B
On Hyman's Factor Analyses

By David R. Saunders†

Even under the most favorable conditions, factor analysis is rarely used as a tool for statistical inference. The only aspect of factor analysis for which there are recognized significance tests is for the number of factors, and these tests depend on assumptions about normality of distributions that are most unlikely to be met in small samples of interesting data. Even when conditions are ideal, the chances are good that the experimental hypotheses can be addressed by using multiple correlation or canonical correlation, and these methods are operationally better defined and conceptually more straightforward. Thus, it is our opinion that factor analysis is never a *necessary* tool for statistical inference, although the method remains a powerful approach for the exploration of multivariate data (see, e.g., Bartlett, 1950; Harris, 1976).

Factor analysis makes its best contribution in really large problems, where it can serve to simplify the conceptualization of substantial bodies of data. In this context, "really large" means at least 20 to 25 experimentally independent variables and several times as many cases. Under such circumstances, the conventional algorithms for communality estimation and definition of simple structure that are built into the computerized statistical packages such as BMD, SPSS, and SAS are sufficiently robust to be trusted; that is, though they may occasionally "fail," their success rate is high enough to warrant their use as a first and even as a blind approximation. The risks of failure increase as the sample of variables or of cases is reduced. Conventional factor analytic wisdom holds that, regardless of the number of variables, the number of *cases* should be at least 10 times the number of *factors retained* in order to establish a reasonable probability that the factors are replicable. This is the rule for analyses based on Pearson product-moment correlations with *continuous* variables. An equivalent guideline for dichotomous variables would require at

†MARS Measurement Associates, P. O. Box 6396, Lawrenceville, NJ 08648-0396.

least 40 cases per factor, assuming the dichotomies divide the sample near 50-50. If the typical dichotomy is, say, 70-30, the guideline now requires at least 80 cases per factor, and the need increases inexorably as the dichotomies get worse. In the limit, when the dichotomies reach 100-0, the sample requirement becomes infinite.

It should be clear from the foregoing that the specific analyses we have been asked to examine must be regarded with extreme caution. At one point (1985, pp. 33–34) Hyman alludes to a 9-variable factor analysis yielding three factors, and at another point (1985, pp. 35–36) he provides somewhat greater detail about a 17-variable analysis yielding four factors (Hyman actually identifies only 16 of these 17 variables). Many of the variables in these analyses are dichotomous, and several of them are experimentally dependent; the sample consists of just 36 cases. However, since three cluster scores based on the smaller analysis appear as variables in the larger one, with the same sample, the analyses do not provide independent information. The results Hyman finds most interesting are drawn from the factor labeled as "Cluster III" in the smaller analysis, supported by the fact that the composite score on that cluster appears in Factor IV of the larger analysis. However, there are major logical problems with all of this.

First, we simply observe that the size of the available data base marginally suffices to support one factor. Even if the variables do represent more than one factor, the sample size is too small to allow more than one factor to appear above the noise level of either analysis.

We must also consider the effects of experimental dependence within the set of variables used in the larger analysis. Among the 17 variables in this analysis, Hyman has included 5 that code the identity of the experimenter for each of the 36 cases (studies) being analyzed. Because these codes are mutually exclusive, the correlation between any two of them ought to be −1.00. If Hyman has used tetrachoric correlations (which are preferred when dichotomous variables are to be factor analyzed [Carroll, 1961]), these coefficients are indeed −1.00. If he has used phi-coefficients, which would result from routine application of the Pearson product-moment formula to pairs of dichotomous variables, the numbers will be smaller but will still be relatively large negative values; and the fact that the variables are mutually exclusive will make each totally predictable from the other four. Thus, no matter what correlation coefficient has been used, the presence of these five coded variables is, in and of itself, enough to force the usual canned algorithms for factor extraction by the principal components method to find four "significant" factors, that is, factors associated with "latent roots" greater than 1. Since there is no way for the analysis to yield fewer than four factors, while the data remain insufficient to support more than one factor, the entire analysis is meaningless. Hyman's report provides no reason for believing that the number of nonartifactually significant factors is greater than zero. Therefore, the larger analysis cannot really be used to emphasize any of the results of the smaller analysis.

Actually, the only use that Hyman makes of either factor analysis in his larger argument is to rationalize selecting a particular subset of his predictor variables for use in two regression equations purporting to estimate ganzfeld effect size as a function of procedural flaws (Hyman, 1985, p. 37). This subset consists of the three strongest contributors to "Cluster III" (from the smaller analysis) plus a fourth variable that was not involved in either of the factor analyses. Limiting the regression analyses to these three flaw categories makes it appear that the multiple correlations associated with the regressions are significant. If the reported correlations of 0.53 and 0.48 had been derived from the *only* three predictors available in a sample of 36 cases, they could be regarded as conventionally significant. However, these are merely three of the nine flaw categories. Since there are 84 ways to select three predictors from the nine, and the factor analysis cannot be relied on to guide the selection, we have a clear example of implicit multiple analysis. That is to say, it is reasonable to suppose that Hyman would have been willing to use any of the 84 possible analyses that permitted the same interpretation, not to mention other possible analyses using different numbers of predictors. Under the circumstances, the multiple correlations cited above must be regarded as nonsignificant, and any interpretation drawn from them must be regarded as meaningless.

References

Allen, S., Green, P., Rucker, K., Goolsby, C., & Morris, R. L. (1976). A remote viewing study using a modified version of the SRI procedure. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975*. Metuchen, NJ: Scarecrow Press.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, Statistical Section*, **3**, 77–85.

Belvedere, E., & Foulkes, D. (1971). Telepathy and dreams: A failure to replicate. *Perceptual and Motor Skills*, **33**, 783–789.

Bozarth, J. D., & Roberts, R. R. (1972). Signifying significant significance. *American Psychologist*, **27**, 774–775.

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

Braud, W. G. (1978). Psi conducive conditions: Explorations and interpretations. In B. Shapin & L. Coly (Eds.), *Psi and states of awareness* (pp. 1–34). New York: Parapsychology Foundation, Inc.

Braud, L. W., & Braud, W. G. (1974a). The influence of relaxation and tension on the psi process. In W. G. Roll, R. L. Morris & J. D. Morris (Eds.), *Research in parapsychology, 1973* (pp. 11–13). Metuchen, NJ: Scarecrow Press.

Braud, L. W., & Braud, W. G. (1974b). Further studies of relaxation as a

psi-conducive state. *Journal of the American Society for Psychical Research,* **68**, 229–245.

BRAUD, W. G., & BRAUD, L. W. (1973). Preliminary explorations of psi-conducive states: Progressive muscular relaxation. *Journal of the American Society for Psychical Research,* **67**, 26–46.

BRAUD, W. G., WOOD, R., & BRAUD, L. W. (1975). Free-response GESP performance during an experimental hypnagogic state induced by visual and acoustic ganzfeld techniques. A replication and extension. *Journal of the American Society for Psychical Research,* **69**, 105–113.

CARROLL, J. B. (1961). The nature of the data, or how to choose a correlation coefficient [Presidential address]. *Psychometrika,* **26**, 347–372.

COCHRAN, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, **10**, 417–451.

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

COOPER, H. (1984). *The integrative research review: A social science approach.* Beverly Hills, CA: Sage Publications.

DUNN, A. J. (1980). Neurochemistry of learning and memory: An evaluation of recent data. *Annual Review of Psychology*, **31**, 343–390.

FOULKES, D., BELVEDERE, E., MASTERS, R., HOUSTON, J., KRIPPNER, S., HONORTON, C., & ULLMAN, M. (1972). Long-distance, "sensory bombardment" ESP in dreams: A failure to replicate. *Perceptual and Motor Skills,* **35**, 731–734.

FREEMAN, M. F., & TUKEY, J. W. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, **21**, 607–611.

GLASS, G. V., MCGAW, B., & SMITH, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage Publications.

GLOBUS, G. G., KNAPP, P. H., SKINNER, J. C., & HEALY, G. (1968). An appraisal of telepathic communication in dreams. *Psychophysiology*, **4**, 365. (Abstract)

HARRIS, R. J. (1976). The invalidity of partitioned-U tests in canonical correlation and multivariate analysis of variance. *Multivariate Behavioral Research*, **11**, 353–365.

HONORTON, C. (1975). Objective determination of information rate in psi tasks with pictorial stimuli. *Journal of the American Society for Psychical Research*, **69**, 353–359.

HONORTON, C. (1977). Psi and internal attention states. In B. B. Wolman (Ed.), *Handbook of parapsychology* (pp. 435–472). New York: Van Nostrand Reinhold.

HONORTON, C. (1978). Psi and internal attention states: Information retrieval in the ganzfeld. In B. Shapin & L. Coly (Eds.). *Psi and states of awareness* (pp. 79–90). New York: Parapsychology Foundation, Inc.

HONORTON, C. (1983). Response to Hyman's critique of psi ganzfeld studies. In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in parapsychology, 1982* (pp. 23–26). Metuchen, NJ: Scarecrow Press.

HONORTON, C., & Harper, S. (1974). Psi-mediated imagery and ideation in an

experimental procedure for regulating perceptual input. *Journal of the American Society for Psychical Research*, **68**, 156–168.

HONORTON, C., & KRIPPNER, S. (1969). Hypnosis and ESP performance: A review of the experimental literature. *Journal of the American Society for Psychical Research*, **63**, 214–252.

HYMAN, R. (1983). Does the ganzfeld experiment answer the critics' objections? In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in parapsychology, 1982* (pp. 21–23). Metuchen, NJ: Scarecrow Press.

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

KIRK, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole Publishing Company.

MARKS, D., & KAMMANN, R. (1980). *The psychology of the psychic*. Buffalo, NY: Prometheus.

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

PARKER, A. (1978). A holistic methodology in psi research. *Parapsychology Review*, **9**, 1–6.

PUTHOFF, H. E., & TARG, R. (1976). A perceptual channel for information transfer over kilometer distances: Historical perspective and recent research. *Proceedings of the IEEE*, **64**, 329–354.

PUTHOFF, H. E., TARG, R., & MAY, E. C. (1981). Experimental psi research: Implications for physics. In R. G. Jahn (Ed.), *The role of consciousness in the physical world*. Boulder, CO: Westview Press.

RAUSCHER, E., WEISSMAN, G., SARFATTI, J., & SIRAG, S.-P. (1976). Remote perception of natural scenes, shielded against ordinary perception. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in parapsychology, 1975*. Metuchen, NJ: Scarecrow Press.

RABURN, L. (1975). Expectation and transmission factors in psychic functioning. Unpublished honors thesis, Tulane University, New Orleans, LA.

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

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 Publications.

ROSENTHAL, R., & RUBIN, D. (in press). Multiple contrasts and ordered Bonferroni procedures. *Journal of Educational Psychology*.

SOLFVIN, G. F., KELLY, E. F., & BURDICK, D. S. (1978). Some new methods of analysis for preferential-ranking data. *Journal of the American Society for Psychical Research*, **72**, 93–110.

SOLFVIN, G. F., ROLL, W. G., & KRIEGER, J. (1978). Meditation and ESP: Remote viewing. In W. G. Roll, (Ed.), *Research in parapsychology, 1977*. Metuchen, NJ: Scarecrow Press.

SOMMER, R., & SOMMER, B. (1983). Mystery in Milwaukee. *American Psychologist*, **38**, 982–985.

SPENCER BROWN, G. (1953). Statistical significance in psychical research. *Nature*, **172**, 154–156.

SPENCER BROWN, G. (1957). *Probability and scientific inference.* New York: Longmans, Green.

STANFORD, R. G., & SARGENT, C. L. (1983). Z scores in free-response methodology: Comments on their utility and correction of an error. *Journal of the American Society for Psychical Research*, **77**, 319–326.

STERLING, T. C. (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.

ULLMAN, M., KRIPPNER, S., & VAUGHAN, A. (1973). *Dream telepathy.* New York: Macmillan.

Psychophysical Research Laboratories
301 College Road East
Princeton, NJ 08540