*The Journal of Parapsychology*, Vol. 63, December 1999 (pp.335-388)

# SHOULD GANZFELD RESEARCH CONTINUE TO BE CRUCIAL IN THE SEARCH FOR A REPLICABLE PSI EFFECT? PART II. EDITED GANZFELD DEBATE

By Gertrude R. Schmeidler and Hoyt Edge

Participants invited to join the three-week debate that was introduced in Part I (Milton, 1999) contributed 90 messages. These messages came to H. E. already numbered, so that he did not know the authors. H. E. first read the messages for courtesy. There was only one message that gave him pause, but in the spirit of open debate, this message was sent forward unchanged. Most of the problems encountered were practical ones, especially dealing with differences in formatting, with over half of the messages requiring reformatting before being sent out to the participants. Occasionally, there were additional technical problems which slowed down dispersing certain messages, and even assigning new numbers to them, but, in general, the process of H. E. checking the messages and redistributing them (by simply sending them back to the server) worked fairly well.

The edited messages thus came to G. S., numbered in the order in which they were received, except that numbers 2, 20-23, 32, and 45 did not correspond to new responses, and the messages listed here as 16a and 86a had not received a number.

Milton asked G. S. to edit the responses, putting them in intelligible order and keeping all substantive points while making them reasonably concise. G. S. decided on topics and subtopics, then listed under them in chronological order the relevant messages or parts of messages (each preceded by its number in boldface type). Parts of a single message often appear under different headings, either because the writer shifted to a new topic, or because a coherent argument cuts across the classifications.

The exact wording of the message was retained except for deletions and for a few trivial changes, like short insertions [in brackets] for clarity or using full terms instead of abbreviations. After the close of the debate,

---

The running of the debate was generously funded by the Fundação Bial and Society for Psychical Research.

G. S. forwarded this edited version to those who had participated and restored a few parts that the participants asked her to keep.

Several responses make substantially the same point, and a reader may ask whether they came from one person repeating an argument or from different discussants who agreed with each other. The reason that the transcript seldom gives this information is that Milton required G. S. to omit the discussant's name (except when the message identified the discussant). The transcript is followed by an Appendix that identifies the author of each message.

Almost half the messages clustered around criticisms or defenses of Milton's conclusion (Milton, 1999) that the Milton and Wiseman meta-analysis of 30 recent ganzfeld studies failed to replicate the evidence for psi in earlier work. This is the first topic below, followed by criticisms or defenses of other points in the Milton and Wiseman paper and in Milton's "Discussion paper." Then come the two next most frequent topics: attempts to define the ganzfeld and the desirable direction for future ganzfeld research. Other topics, such as the value of the debate, follow these.

## Critiques and Defenses of Milton's Conclusion

### *Misuse of Meta-analysis as a Proof*

**#5, part** 1. Meta-analysis is being misused by aiming specifically toward proof-of-existence of a phenomenon. The title of one of the papers forwarded for this debate begins with the inappropriate question,"Does psi exist?" Although there has been a widespread use of meta-analysis in various fields to argue such questions, its real power and value derive from the ability to usefully and quantitatively summarize large amounts of evidence. Meta-analysis should be used to learn something; it is not suitable for the simple-minded effort to prove something. Indeed, the latter effort is terribly vulnerable to contamination by the desire to prove something.

2. Meta-analysis must, in order to be useful, follow the intelligent dictum, "Concatenate widely and categorize wisely." This means to assemble, or fairly sample, all of the material that is pertinent to the topic, and then do an analysis that allows insight into the contributions of various factors. . . . Proof is not the point, but incisive assessment of the available data in order to understand better what the basic experiments are intended to explore. This point specifically abjures the mode employed in some recent meta-analyses, such as the Milton and Wiseman meta-analysis of ganzfeld, wherein arguably arbitrary criteria were set to include or exclude studies. The alternative, obviously, is to include all apparently or

arguably pertinent studies, and then to assess the data as a function of sensible categories. From this, one could learn a great deal that is lost when studies are excluded.

3. Proceeding to the "Discussion paper" per se, I disagree, as detailed in points one and two, with the concluding statement of the Abstract. Tests of replication must not be allowed to be the focus of meta-analysis. That is a bogus application of the tool.

10. The proposition and encouragement of a selective meta-analysis with exclusion criteria is troublesome, and I want to argue most strongly against it. I have given a number of reasons above. . . . The summary and conclusions section of this discussion paper are a splendid example of the misapplication of this tool. Its methods and selections force the indication of a "failure to replicate." But, worse, its methods and selectivity preclude any valuable examination of questions other than that simplistic question.

**#16a, part** The problem is that meta-analysis is expected to decide something, to determine once and for all whether "psi exists," or some such. Inclusion criteria can obviously determine the outcome of this kind of meta-analysis according to the tastes of the analyst.

A desirable approach to a solution is embedded in the root meaning of "meta-analysis": Take a larger perspective; take this larger picture apart to see what it is made of.

Following this approach, meta-analysis should be both broadly inclusive and intelligently categorical, in order to develop understanding of the effects of inclusion and exclusion of identifiable subsets within a well-defined area of study.

**#29** I think the debate member who provided message #5 has given us some very useful advice. . . . I would like to second what she or he presented about the importance of wide inclusions into meta-analysis—of not restricting, initially, studies to be included in terms of some narrow criteria. "Sample widely," and "categorize intensely and extensively" would be good watchwords for us in this endeavor. We can always apply whichever classificatory criteria we wish to studies within the database. Wide inclusions and extensive categories will help our understanding of the process and serve many purposes other than a single-minded dedication to "proof." In fact, we could even designate a certain category of a meta-analysis, rather than the entire analysis, to a proof function. For the latter, we could predecide the qualities we wish proof-oriented studies to possess (beforehand), and then include any and all such studies in the special "proof-focused" section of a more general meta-analysis. This section could be analyzed on its own, for its relevance to proof issues. Other sections of the entire meta-analysis, however, could continue to serve other purposes—very much in line with the spirit of the comments of

message #5. Our meta-analytical endeavors and purposes need not have an "either/or" quality.

**#31, part** Message #5 suggests that it is a misuse of meta-analytic techniques to aim such techniques specifically towards "proof" of a phenomenon. (Presumably we all accept that perfect proof is in the realm of mathematics and logic. What we are really talking about is establishing that the evidence supports the psi hypothesis beyond reasonable doubt.)

I think the point made by message #29—that proof- and process-oriented meta-analyses need not be mutually exclusive—is an extremely valuable suggestion for compromise.

However, message #5 begs the question (to which I don't know the answer), "If not meta-analysis for proof, then what?" Meta-analysis allows one to establish the degree to which an effect is replicating across labs and investigators—one of the criteria most often suggested as necessary to show parapsychologists have a genuine psi effect in their data.

**#53, part** (response to message #31, " . . . suggestion for compromise.")

That is half-right (as compromises necessarily are). The concept of learning what can be learned from the database seems the fully right approach. Meta-analysis, according to message #5, is being used as a sort of weapon, to attack and defend (debunk and prove, respectively, so to speak). This is the apparently easy road, with a particular destination shining in the distance. But the cost is not only the transparently obvious loss of information ignored in the search for "proof" but the hidden costs of servicing agendas that seem to be either half-conscious (half-witted?) or well understood but unscientific.

(response to last paragraph of message #31) Oh, there is no disagreement on the question of meta-analysis as a useful tool, merely the meta-question of what it should be used for. Some folks want to spend their time looking for proof (or lack of proof), and that is fine and dandy, with a few caveats. It is an inefficient use of the tool and the time spent applying it, given the options available, specifically, the option to do sensible and useful categorical analyses. The proof orientation is tantamount to counting on one's fingers, compared to the sophisticated work that could be done. Using your example of labs and investigators, consider what else might be discovered if you were not limited to "the degree to which an effect is replicating" but instead attempted to understand what variables affect that "replication" effect.

**#55, part** I'm not entirely sure that proof- and process-oriented research can go hand in hand. People are talking about "inclusive" rather than "exclusive" meta-analyses, but presumably even "inclusive" meta-analyses draw a line somewhere. And, even supposing that there is an intelligible way of having an "inclusive" meta-analysis that excludes but does not exclude too much of potentially relevant interest, I'm not sure

that such a meta-analysis could settle a "proof" question. Presumably the studies would vary too much to enable clear evidence of anything. . . . If, for example, who the experimenter is makes a difference in psi research (and I agree that the Wiseman and Schlitz research is making a good headway into this question), this itself will make the idea of using a subset problematic, I think. That is, if inclusive meta-analyses also use a subset for evidence-based research, that subset may have to be so selective (exclusive) as to be unpracticable (and perhaps so selective that it would in fact, ironically, not be useful evidence of anything).

Evidence-based meta-analyses would be useful only if we think we know to some, albeit limited, extent what does and does not work. . . . If we say we don't know enough about the process yet, it can be argued that previous ganzfeld meta-analyses have been misconceived: They should have been understood as process-oriented research and were conducted to see whether or not something worked (namely: the ganzfeld under such-and-such conditions).

(#56, listed under *Statistical Anomalies*, also speaks to this point.)

**#57** Another contributor has made the point that there may not be enough studies to do good process-oriented research, and that without such, proof-seeking meta-analysis is a questionable endeavor. That is a good point. The fact that the negative conclusions of the Milton and Wiseman study are vulnerable to the inclusion or exclusion of a very few studies (on the order of 10%) should give pause. The fact that those conclusions are likely to be destructive and misleading raises obvious concerns, some directed to the methodologies of meta-analysis, but others, understandably, to the difficulties frail and motive-driven humans have conforming to the ideals of science.

*Effect Size as the Only Measure of Replication*

**#5, part** 5. There is a curious implication that although the *z* score for the cumulated database is significant, the effect size is not—or rather, it is only 1/6 that of another meta-analysis. Here again, the proof-orientation fails us: We do not learn what is going on, but instead are asked to accept or reject a simplistic pronouncement on the validity or reliability of a potentially very interesting effect.

**#19, part** The title of the Milton and Wiseman paper, "Does psi exist? Lack of replication of an anomalous process of information transfer," suggests that the failure to replicate is total rather than only partial. In fact, however, only the effect *size* reported by Bem and Honorton has clearly failed to replicate! The jury is still out on the question of whether there is a significant psi effect at all in the new database.

*Statistical Anomalies*

**#42, part** A glance at the data presented in Table A1 of the Milton and Wiseman ganzfeld paper reveals that the distribution of *z* scores is anomalous. A chi-square test of the sum of *z* squares results in $\chi^2 = 46.67$, $p = .027$. Thus, a case can be made that this sample of studies is heterogeneous.

The source of the heterogeneity is clear. Three studies are significantly negative (those labeled Kanthamani & Broughton, 1994, Series 5b; Kanthamani & Palmer, 1993; Williams et al., 1994). When these three studies are removed, the remaining 27 studies are now homogeneous ($\chi^2 = 32.4$, $p = 0.35$), and the resulting Stouffer *z* of these 27 studies is $z = 1.99$, $p = .02$ (one-tail). Thus, upon removing three outlier studies from this meta-analysis, the overall result is a statistically significant replication.

If one prefers not to remove potential outlier studies, one may simply note that 6 of the 30 studies listed in Table A1 are independently significant in the predicted direction—one-tail. One would expect only 1 or 2 successful studies by chance. The binomial probability of obtaining 6 successful studies out of 30, where success is associated with $p = .05$, is $p = 0.0005$. This is sufficiently anomalous that it demands a serious explanation.

**#43, part** Dean Radin: Incidentally, an analysis of the Milton and Wiseman data by (unweighted) combined *z* score per principal investigator shows a significant heterogeneity of outcomes, $\chi^2 = 14.98$, $p = .04$.

**#44** Message #42 states, "The binomial probability . . . $p = 0.0005$ . . ."

Correction: That should be $p = 0.003$, not 0.0005.

**#47** Selectively trimming the bottom three studies in the Milton and Wiseman database is a little unsystematic. Surely it is more balanced and justifiable to trim the top and bottom three studies, otherwise it would appear one is simply cutting out unwanted data to improve the effect in a certain direction. What are the results of that kind of trim?

**#50** (response to message # 47) The point of trimming any studies from a heterogeneous meta-analysis distribution is to remove the most extreme sources of the heterogeneity. That is an algorithmic process, serving a completely specific purpose with proper justification. It is not "cutting out unwanted data" but iteratively qualifying the database as a proper representation of the nominal meta-analytic question.

**#51** Message 47 says, "Selectively trimming the bottom . . . is . . . unsystematic."

Actually, this trimming is indeed systematic. One perfectly reasonable method for determining heterogeneity in a meta-analysis database is to calculate the sum of the squared *z* scores, which is a chi-square value, then remove the largest contributor to this chi-square and recalculate. Keep doing this until the remaining studies become homogeneous. It turns out that in the Milton and Wiseman database the three largest

contributors to the chi-square value are the three significantly negative studies, thus trimming them is not arbitrary, but simply follows the algorithm.

**#52, part** Removing outliers may be accepted practice in meta-analysis, but that doesn't keep it from being a bad idea. I am particularly surprised to see parapsychologists embrace this practice, because its underlying rationale is exactly the same one skeptics use to reject psi data a priori: If it doesn't conform to the expected distribution, there must be something wrong with it, so throw it out. Remember that evidential psi results are outliers in the larger distribution of so-called random events. Is this a rationale we really want to buy into? If the data have to be normally distributed to use meta-analytic statistics, then apply transformations to the data or develop statistics that can handle the outliers; don't arbitrarily throw out data points that have as much right to be taken seriously as any others.

**#54** (response to message #52) Removing outliers is a fairly general statistical practice, and not unique to meta-analysis or social science analyses. It's often recommended as a way to keep outliers from making the bulk of the data uninterpretable. But it doesn't mean throwing out extreme cases simply because they're extreme—it's important to figure out *why* they are different from the rest of the data points being studied. If they're really intruders from a different distribution, then it makes sense to remove them (e.g., if . . . [the] context is different).

Radin's analysis is appropriate but not complete. The point is not that the strongly negative results should be dropped from consideration, but that the distribution of ganzfeld study results in Milton and Wiseman's analysis is heterogeneous and that the overall nonsignificant *z* depends largely on these 3 out of 30 results. The implications for interpretation are quite different than they would be if the nonsignificant *z* reflected a fairly even balance of positive and negative results. It's a common statistical injunction to look at the distribution, not just at the summary statistic, if you want to understand what's going on.

To complete the picture, however, we need to know *why* these three results are so different from the rest of the data points being studied. Perhaps something about their procedures militated against positive results. Wiseman and Milton rightly complain that there aren't enough studies using each procedural variation to see whether the variations have systematic effects (which is really what meta-analyses are for). Maybe there are no systematic effects, but 30 studies simply aren't enough to see that the true distribution is much more balanced than it currently appears. We can't tell from the current analyses. But Radin's point is that the picture here is more complicated than "the nonsignificant *z* indicates that . . . nothing's going on." There's a difference between "we don't know what's happening," and "there's nothing happening."

**#56** (response to message # 52) It is not necessary, nor appropriate, to "throw it out." What should be done in a good meta-analysis is to assess the fit of data to their proper expected distribution (based on the statistical model that the analyst must have if she or he wants to do a competent job). Then, the empirical distribution can be characterized, warts and all. Surely it is worth knowing, first that there are outliers, and second, how their presence affects the database.

**#59, part** (response to messages #42, #47, #50, #51) . . . There are plenty of justifiable algorithms and hence severe problems with removing outliers using any algorithm selected post hoc and claiming that the resulting database constitutes a successful replication. Even if one could justify removing outliers at all when asking the question of whether there is an overall significant effect, some algorithms would result in a statistically significant database and others would not. Let's have a look at the range of options:

1. Although some parapsychology meta-analyses have used *z* scores as the dependent variable, others have used effect size, for example by Milton (1993), Radin and Ferrari (1991), and Radin and Nelson (1989). Indeed, using effect size would seem to make more sense if the point is to produce a sample of studies that are similar in an obviously interpretable way.

2. In addition, some meta-analyses have not used the raw effect sizes as the dependent variable but have weighted them (e.g., Honorton et al., 1990; Radin & Ferrari, 1991; Radin & Nelson, 1989) by study size, quality, or both.

3. . . . a 10% bilateral trim (e.g. Honorton & Ferrari, 1989; Lawrence, 1993).

So, even if we just restrict ourselves to choosing from among the options used so far within parapsychological meta-analysis, there are 2 (*z* scores, effect sizes) times 3 (unweighted, N-weighted, quality-weighted) times 2 (chi square, 10% trim), that is to say, 12 choices of algorithm. It would be possible to make a justification for all 12. There are yet more variables that could be factored in, such as study outcome measure.

For example, Milton and Wiseman used each study's author's choice of outcome measure, but an argument could be made for imposing instead direct hits as the outcome measure, even on studies that did not preplan it as the sole outcome measure, as did Bem and Honorton, (1994) (see Honorton, Barker, Varvoglis, Berger, and Schechter, 1985, p. 40), Honorton (1985) and Kanthamani & Broughton (1994), or ranks (unprecedented but could be justified), leading to there being 36 algorithms for removing outliers, and so on and on.

It seems questionable to remove outliers when asking whether an effect exists, and even more questionable to try to draw a strong conclusion having removed outliers on a post hoc basis, given the enormous choice

of justifiable algorithms. Preplanned sensitivity analysis (i.e., applying a prespecified range of analyses to see if they make a difference to outcome) can be valuable, but trying to draw strong conclusions on the basis of a single analysis selected with sight of the study outcomes risks looking like post hoc hacking about in the data to try to rescue a disappointing result.

**#61** (response to message #59) Absolutely. I (author #47) am not suggesting that the author of message #42 is fiddling at all with the data to suit his or her ends, but merely that it can, with some strategies, look that way. A trim of the data typically "tops and tails" the dataset because this is "balanced." The iterative procedure described is another alternative to trimming whose rationale I perfectly appreciate, but confidence in the procedure is enhanced, even if only superficially, by a preemptive justification of its use before launching into a discussion of the results of its use. To use it, and only then justify its use, may to some appear to be a case of justification after the fact.

Additionally, whilst the iterative procedure used by author #42 is applied to the three studies with the largest absolute deviation from a null baseline, and one may therefore justify their specific "ordered" removal in terms of their rank order of absolute magnitude, one may be just as able to reach statistical homogeneity by removing the three largest positive *z*s. In fact, I've done this, by removing the top three most positive studies and leaving in the bottom three and $\chi^2 = 36.46$, $p = 0.11$. The consequence of this is that it is not absolutely "clear" that the heterogeneity issue is explained by the existence of three excessively negative studies, because the arbitrary removal of the three excessively positive studies removes the heterogeneity problem almost as well.

If you top and tail the top and bottom three studies from the database, $\chi^2(24df) = 22.21$, $p = 0.57$.

**#64** (author of message #52 responding to message #54) I agree that removal of outliers is not restricted to meta-analysis. . . .

If the inference that the three negative results are "intruders from a different distribution" is based on some other factor besides their being outliers (e.g., context . . .), then obviously my point does not apply. . . . On the other hand, if the inference is drawn merely from the fact that they are outliers, then the criticism does apply. . . . Exclusion on this basis makes the a priori assumption that nature must conform to some particular distribution, such as the normal curve, which simply is not defensible.

I agree that we should try to ascertain why the three outliers failed to conform to the rest of the distribution, but that should be done as the second step of the analysis, not the first. All studies that meet the substantive methodological criteria should be included for purposes of computing the Stouffer *z* (or equivalent), on the basis of which we decide whether the outcome is statistically significant and thus whether the earlier

ganzfeld studies have been successfully confirmed. . . . Then, as the second step, it is proper and highly advisable to try and figure out what factors led different studies to produce different results. The outcome of this inquiry could then be used to redefine the criteria for inclusion in the meta-analysis of the next generation of ganzfeld studies and *perhaps* to provide the reason for the lack of confirmation.

Incidentally, I think arbitrary exclusion of outliers is unjustified even if the exclusions from the two tails of the distribution balance each other off (e.g., the "10% trim" procedure), because the practice could lead to misleading conclusions about the homogeneity of the data. But when the exclusions are unbalanced, and particularly when they turn a nonsignificant analysis into a significant one, and on top of that the decision for exclusion is made after inspection of the data by someone whose preference would seem to be for the outcome to be significant, the reasons for the exclusion must be extremely compelling if the move is to be credible, and not damaging to the image of the field. So far, such a convincing rationale has not been supplied.

**#65** (response to message #43) I'm wondering whether, when categorizing the studies by principal investigator (senior author?), the following problems were taken into account: (a) In one study in the database (Kanthamani & Palmer, 1993), the principal author's identity depends on whether you rely on this *Journal of Parapsychology* version or the earlier *Parapsychological Association Convention Proceedings* version (Palmer & Kanthamani, 1990). (b) Studies are not always reported by their own experimenters. (The eight Kanthamani & Broughton, 1994, series were meta-analyzed by Kanthamani and Broughton, not carried out by them.)

**#66** (response to message #52) Deleting outliers is common practice not only in meta-analysis, but in virtually all scientific disciplines that rely on measurement and statistics. In any case, the method used to delete the outliers in this case is not arbitrary. It is based upon an algorithm that doesn't care which studies are removed. It just so happens that in this case, the three studies with the largest $z$ scores are all significantly negative.

**#67** In discussion of the wisdom of removing or not removing studies with outlier statistics, we seem to have overlooked the other anomaly: In the entire Milton and Wiseman list of reported studies the number of reported significant studies is in itself significant, binomial $p = .003$. This, of course, is a post hoc observation, but it is not a marginal effect.

**#68** (author of message #52 responding to message #66 ) . . . "arbitrary" . . . was a poor choice of word on my part, and I withdraw it. My withdrawal of this word has no effect on the main force of my argument, which is that removal of outliers lacks an adequate rational basis. The main defense I am hearing so far is that everybody does it, and I find that defense inadequate.

**#69** (author of message #54, responding to messages #52 and #64) . . . Agreed. I don't know whether [the three negative results are] "intruders" at all, and simply meant that there are situations in which trimming does make sense. The fact that data points are extreme isn't, by itself, evidence for any particular hypothesis about *why* they're extreme.

**#70** In message #54, I said [second paragraph is quoted, recommending interpretation of the whole distribution]. Message #61 indicates . . . that removing the three most positive results instead of the three most negative also leaves a homogeneous distribution. It's not an even balance, though: The distribution is still negatively skewed (clumped towards the positive side). For example, when the results are listed in order of size, it's easy to see that the highest ten *z*scores are all greater than +1.0, and nine of the ten associated effect sizes are greater than +.20, while only the four most negative *z* scores are more extreme than -1.0, and only the three most negative effect sizes are more extreme than -.20. It's not clear what this means. Are those strong negative results "intruders"? Are they even far enough out from the rest of the distribution to be considered real outliers? But, again, looking at the distribution does add something to what the summary statistic tells us. (The summary statistic in this case is chi-square used as a measure of heterogeneity.)

**#71, part** (response to message #59) From the point of view of an outsider, much of what #59 wrote would probably be seen as ironic, given the opinion expressed in the last paragraph: " . . . A single analysis selected with sight of the study outcomes risks looking like post hoc hacking about in the data to try to rescue a disappointing result." The irony is that message #59 itself looks quite like that. In an effort to reject an informative consideration of the distribution of studies (something that should have been done in the original meta-analysis since the analysis, ipso facto, relies on a model and hence on a distribution), various scare words are used to defend against the argument that the Milton and Wiseman conclusions are shaky, being vulnerable to the presence or absence of some potential outlier studies. A couple of examples:

"However, there are plenty of justifiable algorithms and hence severe problems with removing outliers using any algorithm selected post hoc and claiming . . . "; "Even if one could justify removing outliers at all . . . "

These don't look any better in context than out, but they do melt into the text, which consists of a longish list (message #59 counts 12) of "algorithms," which we are asked to consider as proper alternatives to the ordinary—and mathematically sound and generally accepted—heterogeneity testing using chi-square goodness of fit. This is not an effective defense of vulnerable conclusions. Instead, it has the appearance of a collection of unsupportable arguments in service of an unshakable conviction.

[quotes paragraph justifying at least 12 algorithms]

It goes on. Is this persuasive? Not to those who would like to understand the substance of the accumulated ganzfeld research. The problem, seemingly, is an opposition of two philosophies: one which seeks insight and one which seeks closure. . . .

From message #59 again, the final paragraph: [quoted in full].

The assumption here is that "asking whether an effect exists" is the only question worth our while. Defensive arguments about the multitude of possible algorithms for examining the distribution of results have the appearance of wishing the other questions would go away. But they will not, and the heterogeneity question in its general form should not obscure the more local questions on potential causes of heterogeneity.

**#72** Message #64 wrote: "I agree that we should try to ascertain why the three outliers failed to conform to the rest of the distribution, but that should be done as the second step of the analysis. . . . [First,] all studies should be included [to] decide whether the . . . earlier ganzfeld studies have been successfully confirmed."

This is a curious approach. We first decide whether earlier work is confirmed, then as a second step we address whether the test of that confirmation is sound? Looks backwards, sounds backwards, . . . Maybe it's a duck?

**#74** (author of message #64 responding to message #72) I see the author's point. My main point in response is that the second step does not involve confirmation, at least not directly. The initial question is whether the new data successfully replicate the results of the standard ganzfeld. This conclusion should be based on a global analysis of all the new studies that meet this definition. The second question is whether we can explain why some of the studies obtained negative results. If the negative studies shared something in common that distinguished them from the others, then this could be hypothesized as the constraining variable, but drawing a confident conclusion would require cross-validation. . . . If such commonality did not exist, it would be unwise to cite individual factors ad hoc within each negative study as constraining variables. No two studies are identical, which means this strategy could never fail. In other words, it is unfalsifiable. (I'm probably beating a straw man here, but just in case . . .)

Of course, if and when the constraining variable is confirmed, one could always redefine the standard ganzfeld and do a new meta-analysis based on the revised criterion. . . . This approach also fits in best with the philosophy of meta-analysis: As the author of message #5 put it, "Concatenate widely and categorize wisely."

**#76** Message #71 argues strenuously against the option-counting approach, previously taken in message #59. . . .

I agree that the chi-square heterogeneity test is ordinary, mathematically sound, and generally accepted, but so is the 10% bilateral trim: Each, in certain situations, has advantages over the other, so neither is a

clear-cut choice. The calculation of the number of justifiable options for outlier removal was based on the assumption that people may in general feel free to pick options from the menu in novel combinations, so that estimating the number of justifiable options by multiplying the number of options for each choice is reasonable. We have clear evidence that this is already happening: The very chi-square test we are discussing is being applied to *z*scores, not effect sizes. I can find no precedent for this combination in that parapsychological literature: All six meta-analyses that have used the chi-square heterogeneity test have used effect size as their dependent variable (Honorton et al., 1990; Honorton & Ferrari, 1989; Honorton, Ferrari & Bem, 1998; Milton, 1997; Radin & Ferrari, 1991; Radin & Nelson, 1989; Stanford & Stein, 1994).

Message #71 sees the calculation of the range of options to support the argument that there is a severe danger of post hoc selection as an unconvincing defense. . . . Our perspectives may differ because we differ in our estimation of how finely balanced the judgment calls are between the various options for each choice and how free we think people may be in picking and mixing from the menu. I don't understand why this appears so unlikely, given that the heterogeneity analysis we have been discussing involves an unusual judgment call and a novel option combination.

**#77** (response to message #76) This is a useful expression. It applies nicely to other issues, in particular a fundamentally important focus on the vulnerability of the Milton and Wiseman meta-analysis to exactly these kinds of "finely balanced" judgments, and "unusual judgment call(s)." The heterogeneity can hardly be called in question, and we should not be distracted by a plethora of choices of what homogenizing function to apply. The point is not to trim the database, but to establish what is in it and whether that content is a suitable representation of the ostensible question. Specifically, does the heterogeneous Milton and Wiseman collection include things it should not, and does it exclude things that should be a part of the database? If the researchers who know this research literature don't think the Milton and Wiseman collection properly represents the question it claims to ask, its conclusions should and will be rejected—not because those researchers don't like the conclusions, but because they are faulty.

**#78** (response to messages #19, #42, #43, #67)

I would agree that meta-analysis does, to a certain degree, involve making arbitrary decisions in the sense that there are not set rules. There are usually several justifiable options for each choice, and it is a matter of individual judgment to choose between them. The number of choices and of options for each choice creates a very large number of combinations. This is why prespecifying analyses is so crucial. When a meta-analysis with preplanned analyses comes up null but post hoc analyses come up significant, the post hoc ones can so easily appear to be

biased data selection that it is difficult to attach the weight to them that they might have deserved had they been preplanned.

Let's have a look at the number of choices . . . to get a cumulated outcome for the meta-analysis. . . . If we just restrict ourselves to options with a precedent in parapsychological meta-analyses, we have: 1. Four choices of outcome measure: (a) each study author's choice (Milton & Wiseman's approach); (b) imposed direct hits (Bem & Honorton, 1994; Honorton, 1985); (c) imposed ranks (never used for a main cumulation analysis but . . . produced a more statistically significant result and so could be justified for later work; see Honorton et al., 1990, p. 134); (d) imposed Stanford *z*-scores; justification as (c) above.

2. Three choices of cumulation: (a) Stouffer *z*, unweighted; (b) vote-count method (i.e., binomial applied to number of statistically significant upper-tail studies); (c) simple addition across studies followed by appropriate test statistic (Bem & Honorton, 1994; applicable to all except 1a above).

So there are 4 times 3 minus 1, that is 11 cumulation of outcome measure options and:

3. 13 choices of outlier removal method, as listed in message #59 (the 12 listed algorithms plus the option of not removing outliers).

Thus we have approximately 11 times 13, that is, 143 possible ways of obtaining a cumulation. . . . In this context I do not think that a significant cumulation on a post hoc basis calls the original result into serious question.

**#80** Message #76 made the point that it is difficult for post hoc analyses on Milton and Wiseman's database to carry much weight compared to the preplanned analyses because there is a large range of justifiable options . . . and . . . people are already picking and mixing. . . .

Message #77 seems to accept the multiple-option problem when it might appear to undermine the findings of the Milton and Wiseman meta-analysis and then to reject it when it appears to undermine the conclusion that the post hoc heterogeneity test is valid. I do not see how both opinions can hold. Moreover, the multiple-option problem applies to post hoc tests, not preplanned ones such as were used in the Milton and Wiseman meta-analysis.

Message #77 writes, "the heterogeneity (of the Milton and Wiseman database) can hardly be called in question" but offers no arguments to support that view, even while appearing to concede that there is a plethora of choices of homogenizing algorithms. I *am* calling it in question, however. How is it possible to support the use [of] a novel algorithm to examine heterogeneity without seriously addressing the issue of post hoc data selection? Why this certainty that that particular test is meaningful?

**#88** (response to message #80, last paragraph) This confuses the question of heterogeneity with the separate question of what to do about

it. There are not "a plethora of choices" for discovering whether the Milton and Wiseman database is heterogeneous. The meaning of that term is unambiguous: The data are *z* scores, whose expected distribution is defined; The Milton and Wiseman database, even though small, is a heterogeneous distribution by *the* canonical test.

The separate question of what next to do is not even at issue if one accepts #77's suggestion that the important question is what comprises the database, as opposed to the Milton and Wiseman question of proof. (Although, of course, the latter question itself depends on the former.)

**#92** In response to message #88, let me rephrase the question to see whether we can understand each other's perspectives better. The combination of a chi-square test with the dependent variable of *z* scores appears to be novel in parapsychological meta-analysis. This novelty of combination raises the problem that the other possible combinations can also be applied with justification, leading to an issue of post hoc data selection. I am not suggesting that this is what has gone on, but the situation clearly raises the question. Instead of referring to a "plethora of outlier removing algorithms," which I apologize for quoting out of context from elsewhere, I should have said "twelve." The *p* value attached to the test's outcome might have been meaningful (in the sense of not needing to be adjusted for multiple analysis) if the test had been preplanned, but it was not, calling the *p* value's interpretation into question. Effect sizes are also expected to have a normal distribution under the null hypothesis. So what makes the algorithm under discussion the canonical test?

*Selection or Omission of Particular Studies*

**#1, part** 1. My own work: This is misrepresented. . . . We have carried out five standard ganzfeld studies, each preset to 30 trials. The four standard studies which used auditory monitoring of the subjects' mentation reports gave an effect size (.33) and a hit rate (39%) close to, or higher than, the expected values from meta-analysis carried out by Bem and Honorton. Moreover, we estimate that one in six subjects amongst those who score hits, produce impressive qualitative hits, to the degree that the essential features of the film are represented in the mentation report.

2. The papers: Milton and Wiseman's interpretation of Table A1 (their report) and Milton's interpretation of Table 1 A1 (her report) seem at best arbitrary. Milton will exclude Symmons and Morris because they used drumming. On the same basis, the large study of Willin—which used music targets is a radical departure from standard ganzfeld—should be excluded.

**#3** Julie Milton: The author of message #1 states that his own work has been misrepresented but . . . does not say in what way he believes me to have misrepresented his work.

**#5, part** 2. . . . Arguably arbitrary criteria were set to include or exclude studies. . . .

4. In Table 1, there is an entry that excludes Dalton (1997). There is no apparent justification for this. . . .

**#13, part** Any study which did not have direct visual stimuli as its target type would not count as a standard ganzfeld—this counts the Willin study out straight away. I use the term direct because I submit that at a push the Willin study might just achieve success by the synaesthetic mapping of auditory stimuli (played music) in to musical notation (a visual stimulus), and that similar transmutations of primary stimulus types might also give rise to secondary visual stimulation in any study using nonvisual targets (at least where a sender is used). . . .

According to my criterion, the Willin study is a nonstandard *ganzfeld* study because it has the essential three ingredients which permit the title ganzfeld to be used, but it omits the conventional (and theoretically paradigmatic) use of directly visual target stimuli. However, the Symmons study is not a *ganzfeld* study at all because it does not use a monotonous auditory stimulus. (The stimulus has pattern, even if the pattern is repetitive.) The Symmons study is also driven by an alternate conception of psi enhancement: that psi may be enhanced by the "driving" of altered states of consciousness (created by the synchronization of brain waves with the beat frequency of the drum). So, the Symmons study really is a totally different testing paradigm and just isn't in quite the same game as the ganzfeld.

**#14, part** (response to message #13) Having outlined the criteria of the ganzfeld as: 1. . . . 2. . . . 3. . . . , it was stated that the Symmons study is therefore not a *ganzfeld.* However, the Symmons study not only used the ganzfeld to research a totally different paradigm, but it also outlined methods which streamline the ganzfeld procedure. To omit the Symmons study reflects the static nature of ganzfeld research.

**#16a, part** (response to message #14) This is a curious and bemusing example. It has the appearance of contradiction, considering the context, and suggests this interpretation: A proponent of "definitive" meta-analysis for ganzfeld wishes to include studies that clearly herniate the well-defined ganzfeld paradigm, justifying this by decrying the "static nature of ganzfeld research."

Certainly it is possible to distort the picture of a nominal field if inappropriate and nonrepresentative material is included—or, on the other hand, if some exclusion criterion keeps appropriate material out.

If an analyst wants to include nonstandard experiments in a more inclusive meta-analysis (which is not to be mistaken for a "ganzfeld" meta-analysis), then she or he must assume the responsibility for explication of differentiation attributable to the inclusion. If nonstandard experiments (by a reasonable, and largely agreed set of standards) are

included in a "ganzfeld" meta-analysis, it would be a dereliction of responsibility if the effect of such inclusion were not explicitly determined by appropriate categorical analysis.

**#18, part** (response to message #14) I don't have any problem with allowing the ganzfeld to develop in nonessential ways, such as testing multiple senders, with olfactory targets, or using virtual reality simulations as the target material. But when you break with the pattern of using monotonous auditory stimuli (forgive the pun), and choose drum beats instead, and have a different supporting theory to explain the way in which the drum beat changes one's state of consciousness, then you have changed one of the key features of the ganzfeld and it is no longer a ganzfeld. I'd call the Symmons study a ganzfeld/Auditory Driving study at best—a kind of hybrid. . . . Lastly, just in case anyone thinks it's possible to define a standard for a testing paradigm which changes its essential features over time, good luck.

**#28** I don't think Willin's studies should be excluded simply because he used auditory material in the ganzfeld. Most of the receivers' reports (and I have seen them) spoke in terms of visual material rather than musical imagery.

**#30** (response to message #28) I think it would be unnecessarily limiting to exclude studies that use other than visual target materials. I feel it was an historical accident that the ganzfeld (along with 99% of other free-response procedures) focused, and continue to focus, on visual target materials. This is especially ironic in that (and I think this is not well known) in the "context of discovery," Charles Honorton developed his own ganzfeld procedure of part of an (ultimately unsuccessful and aborted) attempt to operationalize the samyama notion (expressed in the Yoga Sutras attributed to Patanjali) by having research participants attempt to continue to detect a tone stimulus that was increasingly embedded in auditory noise, once they had attempted to "merge" with such a target tone (remotely presented) through an intense program of attentional training (as implied in the three major components of samyama).

**#35** (response to message #28) I am not sure that just because the receiver's mentation reports largely visual material that this is a good case for including the Willin study. There are at least two possibilities here: (a) The visual material is unrelated to the musical target and is in fact swamping the auditory ESP signal. (b) The visual material is related to the target in some synaesthetic way. . . .

If (a) is true, then the Willin study is out as it arguably uses a technique which boosts attention to visual imagination imagery to the detriment of the other senses.

If (b) is true, then clearly there is a certain theoretical and practical complication involved in accounting for the Willin study's modus

operandi, over and above the standard ganzfeld with visual targets. It is this complication that prevents the Willin study from achieving the standard ganzfeld. Remember that I am not excluding ganzfeld status to the Willin study outright. It does have three of the key features (1, 2, and 3 of my original posting [message #13] attempting to define a standard—which I took on for the fun of it). Willin strikes out on point 4 for me. I think point 4 (the preeminence of visual ESP in the ganzfeld) follows as a natural feature of the ganzfeld.

**#71, part** Why, really, if one is genuinely interested in the existence question, would one want to exclude clearly relevant, available data (which happen to show a positive yield), and why go to great lengths to include negative data whose relevance to the nominal existence question is questionable? Can objective observers be content with rigid cutoff dates as a definition of what should be included in a representative picture of the replication issue? Will such observers believe that auditory target studies should perform like the visually oriented primary ganzfeld paradigm, and hence should test the replicability of the ganzfeld experiment? Probably not, because the existence question is not one that can be settled conveniently by the calendar, or arbitrarily by inclusion standards.

**#73, part** I am the author of message #1. . . . You asked me to specify how my research was misrepresented. I thought I did: Serial ganzfeld is not a standard ganzfeld. Even nonmonitoring is not standard.

**#81, part** Julie Milton: In response to message #73 . . . Because I am not claiming that the studies in the meta-analysis are standard ganzfeld studies, I am not misrepresenting your serial ganzfeld series as standard by including it. My inclusion criterion was merely that a study be an ESP study and use the ganzfeld environment.

**#95, part** One issue that might need further comment regards the timing of meta-analyses. Analysts' timing may be driven by many factors, and it is hard to remain blind as to how things are likely to be going, if one is at all active in the field. If we consciously or unconsciously do analyses when we think things are "going our way," then we are more likely to be selecting occasions when the results are strong in one direction or the other. This may in turn prompt a kind of regression to the mean effect for the next one done.

## Criticisms and Defenses of the Presentation and of Its Omissions

### *General*

**#5, part** 5. . . . In a paragraph on "Implications of the current situation," the observation " . . . only show overall statistical significance to

date if one extremely successful study is included" is treated as a problem. In my opinion, it would be wiser and much more productive to regard it as a "solution," that is, a source of information.

8. Another nice example of a "problem" that may actually be better treated as information is the observation that only Honorton's PRL [Princeton Research Laboratories] ganzfeld meta-analysis was an exception to the rule of heterogeneity of effect sizes. It also is the only one of the databases listed generated by a single experimenter. Maybe there is some insight to be gained?

9. The last several interpretive paragraphs in the "Problems" section are littered with straw-man examples. I find the statement "I am not arguing that methodological problems clearly account . . . "to be disingenuous. I presume others will also, and I trust that more of the specific points that could be made will be. I'll note in passing that it is bemusing to find so clearly stated the assumption that "the only way to (provide strong evidence for psi) is by demonstrating a replicable, non-zero effect across a range of experimenters. . . . " We have a lot to learn before we can make such presumptions. Not least, we must come to respect the whole matrix of conditions—not only the experimenter, but the environment, and the experimental question. It is telling that the following bald statement is made: "It is not evident, at this point, what a replication of the PRL work in its essentials would have to consist of."

**#8, part** Julie Milton: Message #5 raises a lot of interesting points. . . . However, I would like to respond to the author's description . . . "disingenuous" and to his or her statement, "I presume others will [find it so]." I do not know why my statement appeared disingenuous to this participant, but it is, in fact, my honest opinion.

**#19, part** I am somewhat put off by Milton and Wiseman's (in press) bald and repetitive statements to the effect that the recent ganzfeld studies have failed to replicate Bem and Honorton (1994) ganzfeld results. . . . The answer to that question presently hinges on arbitrary decision rules about what outcome measures are used and on what studies are included or excluded from the analysis, as acknowledged by Milton in her predebate discussion paper. Yet nowhere in the (preview of the) Milton and Wiseman *Psychological Bulletin* article is there any acknowledgment of this fact. Even in the discussion paper, which reaches only a small audience of recently active ganzfeld investigators, this critical information is buried in the text and not even mentioned in the abstract or conclusion. Although Milton and Wiseman offer the usual sops about more research being needed, their paramount message to the wider scientific community is essentially "there ain't nothin' there folks." In fairness, shouldn't that wider community be informed about the ambiguities in the evidence on this most important question of whether there exists any anomalous

effect, at all, in the recent ganzfeld database, even if of much smaller size than previously believed?

**#42, part** Given the statistical peculiarities of the data reported by Milton and Wiseman, I find it most perplexing that the tone of their paper is so strongly negative.

*Handling of Procedural Flaws*

**#5, part** 6. Much of what is said about assessment of quality is, to be rather crude about it, crap. All of the quality scales I have seen imposed are arbitrary and redundant, and most of the analytical conclusions drawn are themselves more deeply flawed than the meta-analyses about which they allegedly inform us. Post hoc quality assessment is tough, as everyone knows, but we blithely go on as if the flaws in these quality assessments were minor. I quote, "The lack of evidence that these databases in general consist of high quality studies introduces the possibility . . . " This makes me think that the old saw, "lack of evidence is not evidence" should be written on the blackboard 100 times.

7. Of the eight enumerated points on flaws, the only one that impresses me as useful and important is number five, which points to a sort of "file drawer" at a different level: namely, the inclusion of information in reports (and the likelihood that the nature of the experimental results will influence the style and inclusiveness of the report).

**#6, part** As for the ganzfeld meta analysis, more important than discussing what is a standard ganzfeld procedure for inclusion must be to exclude all experiments that are flawed. Of course, criteria for inclusion must be formulated before the experiments are conducted. Otherwise the selection of experiments will be another kind of post hoc manipulation. However, flawed experiments must never be included, since they prove nothing but the flaws in the experiments of the meta-analysis.

**#8, part** Julie Milton: (response to message #5) I do not think that there is either evidence or argument to support a claim that methodological problems clearly account for the observed results. However, I do think that the lack of data on the methodological quality of the studies in many meta-analyses makes their interpretation extremely problematic, and that was my point.

**#17, part** An implicit assumption in the rationale behind Milton and Wiseman's meta-analysis appears to be that after the publication of the Hyman and Honorton guidelines and after Bem and Honorton's successful meta-analysis of Honorton's subsequent studies, most other experimenters will have increased their use of safeguards accordingly and have worked towards replicating Honorton's findings. I am a little skeptical that work in parapsychology has been so well organized.

Moreover, Milton and Wiseman do not provide a statistical comparison between the safeguards present in the pre-1986 studies and those that are post-1986, so there is no formal indication of whether the first assumption behind the rationale (that people significantly increased their use of safeguards) follows through.

**#41, part** (response to message #6, excluding all flawed experiments) This seems undeniable but does not undermine the search for a standard ganzfeld procedure.

**#49, part** (response to message #6) The author goes on to propose that all flawed experiments be excluded from meta-analyses. This is a contentious issue among meta-analysts in mainstream fields, but I understand that most meta-analysts maintain, contrary to the author, that at least some flawed experiments should be included and their impact on the results evaluated. This paragraph in the author's contribution suggests to me that he is defending Hyman's "dirty test tube" argument. It is not surprising that a dirty-test-tube proponent would embrace the exclusionary principle in meta-analysis, because whether a flaw actually did cause a result, or even whether it plausibly could have caused the result, is irrelevant for the dirty-test-tube argument as I understand it. But surely, if the purpose of evaluating the data at all is to make our best estimate of how the finding should be interpreted (as it surely must be), then the dirty-test-tube argument contradicts the very premise on which the analysis is justified in the first place. Perhaps the hidden corollary of the dirty-test-tube argument is that the existence of inconsequential flaws somehow implies that there must be consequential flaws that have not been identified. Although I don't see that this conclusion follows at all from its premise, adding this corollary at least makes the dirty-test-tube argument somewhat more sensible. . . .

Of course, the probability of any flaw being responsible for the result is not completely nil. We thus should strive to eliminate all artifactual interpretations as much as possible, even if they are implausible. . . .

The acknowledgment that flaws have some possibility, however minuscule, of explaining away the evidence for psi raises an important issue often overlooked in debates about the reality of psi. Conventionalist rhetoric has traditionally avoided addressing the possibility that the evidence for psi, while not "conclusive," might be such that the probability of psi's existence is high. The reader is asked in effect to make a choice between two extreme possibilities: the evidence is conclusive or there is no evidence at all; if it's not conclusive, it's worthless. This false dichotomy, which is very much to the conventionalist's rhetorical advantage, is never stated as starkly as I have here; if it were, its falsity would be immediately transparent. But by sliding it in the back door, so to speak, conventionalists (whether wittingly or not) have until recently tricked psi proponents into implicitly accepting the dichotomy as well, to the

detriment of their own position. To the degree that the alternative explanations of psi data have a low probability, the probability of a paranormal explanation becomes correspondingly high. Thus, I have a question to throw back to the conventionalists (and especially to the author of message #6). Do you accept that evidence is a matter of degree, and thus that a probabilistic judgment of the reality of psi, based on that evidence, is appropriate? If so, what probability, in your mind, does the evidence justify, and on what do you base that judgment?

**#63** Julie Milton: Point 6 in message #5 . . . [quoted in full]

In my paper I concluded that the parapsychological databases examined so far (not including the PRL or more recent ganzfeld studies) consist of studies of "uncertain or low methodological quality" (p. 5) because they do not generally perform as well as one would wish on the quality scales applied to them. Message #5's point 6 seems to be objecting to the use of "arbitrary and redundant" scales to reach such a conclusion.

I agree that the quality scales used in just about any meta-analysis, let alone parapsychological ones, are somewhat arbitrary in the important sense that methodological flaws that might lead to quite large effects are usually given equal weight to those that might be expected to have smaller effects. Several of us (Milton, 1997; Radin & Ferrari, 1991; Radin & Nelson, 1989) have discussed this problem and have attempted to deal with it in our meta-analyses by using expert weighting or Monte Carlo techniques, although arguments can be made against all of our attempts.

However, these quality scale data are the only data that we have available to assess the range within which mean study quality lies in the parapsychology meta-analysis. They . . . give us some basic data about the minimum frequency with which safeguards are implemented. I think it is well worth knowing also that two meta-analyses put forward as providing strong evidence for PK do not even report any measure of the mean quality of the studies that went into them.

I am not arguing that any database is of clearly low quality and have repeatedly stressed in my paper that the problems arise from having a database of low or "uncertain" quality . . . [and thus we have] a dangerous lack of the knowledge about their quality that would be needed to claim them as strong evidence for psi. Absence of evidence is obviously not evidence of absence, but it certainly isn't evidence of presence, either.

Apart from Hyman's (1985) scale which appears simultaneously partially redundant and lacking in some important quality measures, I have only observed a fairly trivial level of redundancy in most of these scales, and so I do not think that scale redundancy has led to any database's quality appearing very overconservative. It might be helpful if message #5's author could take maybe a couple of recent examples of typical quality scales and say which items appear redundant and what percentage of the

total number of scale items they constitute, so that we can get an idea of why our perceptions differ.

*Handling of Sensory Leakage*

**#10** In message #6 it was suggested that . . . "["psi" effects may be due only to procedural flaws]. Once parapsychologists succeed in producing pure chance results they will convincingly have shown their skills as experimenters."

Milton and Wiseman seem eager to demonstrate such skills. Intraparadigmatic science has once again succeeded in closing the gaps in our knowledge, all is quiet, there's nothing out there. In the preview of their meta-analysis publication, they mention that "Wiseman, Smith and Kornbrot (1996) suggested that the experimenters (in PRL ganzfeld sessions) who on each trial read the receivers' mentation back to them after the response period may have been unknowingly nonblind to the target's identity due to potentially inadequate auditory shielding between experimenter and sender."

They concede, however, that "none of the opportunities for sensory leakage appear sufficiently strong . . . to explain away the positive results of the autoganzfeld in any immediately compelling way and it is clear that Honorton and his research team went to considerable lengths to attempt to provide adequate sensory shielding."

Which does not keep them from concluding "this failure to replicate could indicate that the autoganzfeld's results were spurious, with the main effect having been due to very weak sensory leakage and the statistically significant internal effects resulting from correlations between psychological variables and performance in detecting weak sensory stimuli in some cases."

One way to strengthen the case of the sensory-leakage flaw, would be to demonstrate a pattern of lower hit rates in the PRL autoganzfeld database for those sessions where the subject was more confident about his ratings, since in those sessions one might expect the influence of the experimenter on the target selection to be relatively weak. In fact, a secondary analysis testing this prediction was presented by Dick Bierman in a 1995 PA poster-paper. (It can be found on his web pages: www.psy.uva.nl/bierman.)

The results showed an opposite trend, even when the analysis was repeated with the exclusion of those subjects that tended to give high ratings to all the targets in the set (in which case high ratings would not reflect high confidence, leaving subjects still vulnerable to subtle experimenter cues). This refutation was communicated to Wiseman et al. in 1995. It is not mentioned in the preview of Wiseman and Milton's meta-analysis publication. It seems that, in the neutral view propagated

above, Bierman's finding has been unidentified—has been turned into an Unidentified Experimental Effect, a flaw in a flaw-finding study.

**#11** Julie Milton: Message #10 puts the view that Dick Bierman's secondary analysis of the PRL autoganzfeld data refutes Wiseman et al.'s hypothesis that acoustic leakage from sender to experimenter may have allowed the experimenter unconsciously to cue the receiver during the judging process. . . . On the contrary, the results could be interpreted to support the sensory-leakage hypothesis.

Bierman's finding was that on trials where the receiver gave one of the judging set the maximum rating, the hit rate was higher on those trials than on other trials. No statistical test is reported. . . . Even if [the result] is not due to chance, the sensory-leakage hypothesis might be expected to produce such a pattern of results because one might expect both higher confidence about an individual cued target and higher performance on those trials where the experimenter had acquired unconscious information and was successfully passing it on unknowingly to the receiver. The cue would be expected to cause both the high confidence rating and the hit, leading to the correlation.

**#12** (Response to message #11) Remember that the cues we're talking about were supposedly acquired by an experimenter who was not aware of them, and were "successfully," though again "unknowingly," passed on to the receiver. Indeed, during a ranking process, such cues might turn the balance for an otherwise indecisive or indifferent subject. Extreme confidence is another matter, however. Though, of course, subtle cueing might stimulate maximum ratings, it seems unlikely that such optimal influencing behavior would have gone unnoticed by the PRL experimenters themselves. For those sessions that resulted in maximum rating (of the chosen alternative) plus large inter-rating variance, this leaves us with two alternatives:

1. Reflecting on their own behavior, PRL experimenters were generally aware of their hit-rate-boosting influence in these sessions (in which case they must have questioned their source of information).

2. Extreme scoring with large inter-rating variance reflects a (possibly psi-evoked) confidence that would make a subject relatively less vulnerable to experimenter influence.

Bierman's argument starts from this last alternative.

**#15** Dick Bierman's data that we're talking about consists (in his Table 1) of the number of hits and trials obtained in three types of trial. Group A (let's call it) consists of trials in which one of the set obtained the maximum rating of 40, Group B of trials on which one of the set obtained a rating of 39, and Group C of trials on which the highest rating was 38 or less. I can't see why the data of Group B is presented separately and not just lumped in with Group C because Dick's argument has to do with comparing trials on which targets are given the maximum rating with

other trials. He doesn't report any statistical tests, but a chi-square comparison of Group A with the hit rate of Groups B and C combined is nowhere near statistically significant, $\chi^2(1) = 0.528$, $z = 0.73$, $p = .23$.

Is there any purpose in discussing in depth a result that appears to be well within the range of chance variation? How can it be seen as a refutation of Wiseman et al.'s sensory-leakage hypothesis?

**#16** (Response to message #15) Well, here we go again.

The sensory-leakage hypothesis hinges on the assumption that the experimenter nonconsciously influences the subjects during judging, on the basis of sounds that are well below sensory thresholds. The calculation of sound levels under shouting conditions (shouting was never observed) shows signal-to-noise ratios which make the model on itself already extremely improbable (and some audio experts would say impossible).

A logical consequence of this "model" is that one would expect the experimenter to have the largest impact with this nonconscious "influencing" when the subject is rather insecure about what target to choose.

It is obvious that the sound leakage model has no impact if the subject is completely convinced about his choice (whatever the experimenter says). A straightforward prediction, therefore, would be that there will be no (or a much smaller) effect for trials where the subject is giving the highest possible rating because, whatever the nonconscious influencing, a subject who is giving a 40 (maximum) is hardly influencable. So, the sound-leakage model would predict a difference between the scoring rates where the subject just made it (e.g., gave a 10 to the target and 7, 8, and 9 to the decoys) and trials where the subject showed great confidence (e.g., 40 for the target and smaller numbers for the decoys). In other words, the sound leakage hypothesis results in a prediction that there will be a difference. So, the fact that the difference is "well within chance fluctuation" is exactly what undermines the sound-leakage hypothesis.

However, the major point is that the trials where the subject was convinced of the target (rating of 40) are already quite significant on their own (exact binomial $p = 0.007$). So, at least for these trials, where the highly improbable sound-leakage model is even more improbable, we see a significant anomaly which is larger (although not significantly so) than in trials where the sound leakage hypothesis would have been more probable. The effect size for the rating-40 trials is 0.27, which is generally considered to be a moderate effect. Not small or subtle by any means.

Note that the authors of the ganzfeld discussion paper knew about these data for a long time. It is possible that they too have similar or other ideas why the data do not undermine their hypothesis. Interestingly, they prefer to avoid discussing this. Not a very scientific attitude.

**#39** Messages #10, #11, #12, #15, and #16 engage in a debate. . . . As good empiricists, we could conduct some experiments to elucidate the position. On the other hand, is there—to quote message #15—"any purpose in discussing [this] in depth" if what we are trying to do is to plan for the future. Surely, our effort would be better spent in agreeing how to ensure that we have *no* sensory leakage in any new experiments, and then the argument will not arise.

**#60** Julie Milton: Messages #10 and #16 object to the omission of Dick Bierman's (1995) paper on the sound-leakage hypothesis from the section of the meta-analysis paper that mentions research on a number of potential methodological problems with the PRL work. That section was not intended to be comprehensive. . . .

[As stated in message #11] . . . it is by no means certain or even probable that a participant who gives one of the target set a rating of 40 believes that the match between its content and their mentation was extremely high, nor that they felt high confidence in their choice. Certain state or trait variables may instead lead some participants to tend to follow a pattern of assigning a maximum rating, regardless of the actual quality of the match. If the same variables also correlate with sensitivity to subtle cueing, it would not be surprising to find that trials in which a maximum rating was assigned were independently significant, and that they had a higher hit rate than other trials, as was observed (although the latter to a nonsignificant degree). Similarly, tight variance on ratings need not reflect doubt and an openness to an experimenter's unconscious influence on the participant's part, but a reluctance to use the whole scale. These problems in interpretation prevent Bierman's analyses from constituting a refutation of the sensory-leakage hypothesis.

## Definition of Ganzfeld

### *Standard Ganzfeld*

**#1, part** The conditions for a genuine ganzfeld replication: . . .

1. Introduction and supervision from previously successful experimenters on how to conduct sessions: We used a training tape recorded for us by Kathy Dalton.

2. Use of the correct population from which to recruit participants: avoid psychology students. Recruit through a carefully worded newspaper advertisement and even from New Age centers.

Some additional screening may give better results: Preferably, select those participants who have regular psi-experiences (not one-off crisis telepathy or such experiences as the telephone ringing when they happen

to think of the person), are not disturbed by them but have integrated them into a personal philosophy.

3. Auditory monitoring of the receiver's mentation.

4. Video targets with engaging contents and with a clear contrast between members of the series.

5. Use of standard Ping-Pong balls and white noise, alternatively seashore noise, during 30 minutes.

6. Elementary precautions: Separate rooms, preferably sound attenuated, and at least 20 meters separation; two-experimenter design, duplicate target material, isolation of sender and receiver teams, random number tables, or computerized randomization; double documentation; one target video sequence with 2-4 minutes in length per ganzfeld session. (We are currently experimenting with 2 targets from 2 separate series per session but obviously would not want this included in the database until at least a pilot study has shown it works.)

7. The experimenter leading the session should have experienced success with the technique, be convinced it works, and have sufficient skills to make the participants expect that it will work. Anticipating that some will object that this is a belief clause or even an incompetency clause, it should be stressed that the main experimenter should not abdicate control and should choose a coworker he considers fully trustworthy. If no one is considered trustworthy, conditions should be enforced that effectively rule out fraud.

8. The session has to be a social occasion. . . .

**#4** The essence of message #1 appears to be that there is presently such a thing as a "standard ganzfeld" ESP study and that studies that are not standard should not have been included in the Milton and Wiseman meta-analysis, nor in Milton's updated meta-analysis. The author lists the conditions for a genuine ganzfeld replication but does not state explicitly that these are also meant to be the conditions for a standard ganzfeld.

There seem to be serious problems with this approach. First, it is not clear on what basis one might define a standard ganzfeld. Is it on the basis of which procedures are common? If so, how can one decide which features matter? There are common features that might be unimportant.

Alternatively, is it on the basis of which features are likely to be psi conducive? If so, what are the rules for deciding? . . . A proof-oriented meta-analysis of existing studies that included only "standard" studies would appear so open to bias that its results would not seem credible. This is especially so, given that previous ganzfeld meta-analyses have not attempted to define a standard ganzfeld.

. . . Would anyone who holds this "standard" ganzfeld position like to respond to my questions? That is, first, what principle should be used to define a standard ganzfeld (or one expected to produce above-chance effects), and second, how could the credibility of that principle be

defended as the basis for a proof-oriented meta-analysis done on a post hoc basis?

**#13, part** To classify as a standard ganzfeld study the study must in practice involve three things;

1. A monotone visual stimulus—a colored light with eyes shut.
2. A monotone auditory stimulus—white or pink noise.
3. A free-response ESP testing environment.

In addition, because the ganzfeld is generally concerned with maximizing visual imagery (though I don't doubt it accentuates many other imagerial modalities as well) and because so much conscious psychic imagery is visual (say 80%, at a guess), I think one can argue for another defining feature of a standard ganzfeld.

4. A standard ganzfeld study must involve visual target stimuli as a minimum form of sensory stimulation.

Now, I would say that a ganzfeld study *must* have the above things in order to qualify as a standard ganzfeld study. I think 1 and 2 follow unequivocally from what "ganzfeld" means (whole or uniform field), and are in any case defined not just in principle but in practice also (the common practice of parapsychologists since Honorton and Harper in 1974). . . . Because of the particular research interests of any one parapsychologist we may expect ganzfeld studies to differ from each other in nontrivial ways (e.g., some have senders whilst others don't, some have dynamic targets whilst others have only static targets). These differences, providing the basic four conditions are met, are easily dealt with by meta-analysts by the use of blocking and moderating variables analysis.

**#14, part** In response to message #13, . . . sensory deprivation can be obtained in a number of different modalities, and allowing the ganzfeld procedure to be manipulated to test different methodology hypotheses will allow the ganzfeld research to develop. Variation in ganzfeld procedure will allow researchers to determine the best requirements for a standard ganzfeld.

**#18, part** Though I hate pedantry, I just have to be pedantic about a point of terminology which often passes unchecked in parapsychological research. Since when has a technique which bombards the participant with pink noise/pink light/drum beats been a "sensory deprivation" technique? The ganzfeld is strictly a "sensory monotonization" technique. The difference is subtle but worth noting, because it does have consequences for essentialist definitions of a standard.

**#25** (response to message #13) I'm confused about the "eyes shut" part. In the ganzfeld work with which I am familiar, much is made about keeping the eyes open,—except for necessary blinking. Indeed, that is the point of the uniform visual stimulation: There is sensory stimulation, but perceptual uniformity. (The incoming stimulation is "unpatterned.")

"Monotone auditory stimulus" also is puzzling. Early ganzfeld protocols included ocean sounds, instead of pink or white noise. Certainly, this is not "monotone." Indeed, ocean-wave sounds include patterns.

As typically used, the ganzfeld protocol includes free-response testing. Need this inevitably be the case, however? Could the ganzfeld set the stage for a wide variety of psi tests (not excluding, even, PK procedures)?

Finally, early ganzfeld protocols typically included an explicit relaxation component. Should the latter be part of its standard definition?

**#27** (response to message #25) O.K. folks, I've blended my recollection of the typical ganzfeld procedure with my own most recent work on an alternative psi-enhancement procedure in which the eyes are definitely kept shut. Point 1. of my message 13 should therefore read; 1. A monotone visual stimulus—a colored light with eyes open, open, open!

. . . The ocean sounds so wonderful and relaxing, but surely those early studies weren't quite taking the notion of stimulus habituation to a monotonous, unpatterned auditory stimuli as seriously as they should have. Now I believe most people do take it seriously—if not, why not? As far as I'm concerned, point 2. stands, and if it knocks some previously meta-analysed studies out the meta-analytic melting pot, then so be it. The ocean/white noise distinction is precisely the kind of subtle but nontrivial distinction which may account for heterogeneity in the ganzfeld outcomes. Who knows, unless people start getting wise to the background theory in the ganzfeld and block these variable stimulus types off in their future meta-analyses, *or* define a *standard* and stick to it. . . .

Of course [free-response testing] need not inevitably be the case. I have a student doing a ganzfeld PK project right now, but I wouldn't want to see it classed as a standard ganzfeld. Also, given the background theory to the ganzfeld, I can't see why one would bother testing forced-choice ESP in the ganzfeld. The ganzfeld was designed to make hypnagogic type imagery more direct and less symbolic (at least that was the thrust of the Bertini, Lewis, & Witkin paper from which I believe Honorton drew his support for the ganzfeld's psi-enhancing properties). We're talking about enhancing conscious imagery that may have a psi component with the ganzfeld. [In] forced-choice tests . . . consciousness of the target image is less paramount. . . .

[Relaxation as part of standard definition] Sorry folks, but I would say no. People have already mentioned the lack of controls or comparison procedures in the ganzfeld, and I think this is a *major* problem with the previous research. The ganzfeld is intrinsically a sensory monotonization technique which, we hope, leads to sensory habituation and the focusing of attentional resources on to internal sources of stimulation (which will include memories and imagination—the putative vehicles of ESP according to Roll, Irwin, etc.). However, internal

attention states also focus on nonrelevant internal stimulation, like the rising and falling of the breath, and stress in the muscles. Relaxation removes the amplitude of the breath, and relaxes the muscles so tension is reduced— but this is *not* a property intrinsic to sensory monotonization. It's more akin to sensory reduction or deprivation.

Now, if a ganzfeld has the standard ganzfeld stuff (as I have defined it in message #13) *and* relaxation, we have a problem with the attribution of successful outcomes. Did the ganzfeld (white noise, pink light) do it, or did the relaxation do it, or was it an interaction of both the ganzfeld and relaxation? We might believe that it must be an interaction of both variables, and I would agree, but the study design of most ganzfeld experiments does not meaningfully manipulate these variables so we can observe their separate or conjoint effects. So, we are left with a belief about what is going on, rather than some systematic understanding. . . . If we do some real science with it (and I'm guilty of following the typical pattern of previous studies as much as anyone), I think we'll soon find out just what, if any, its real psi-enhancing properties are.

**#33** (response to message #27) " . . . If [requiring a monotone auditory stimulus] knocks some previously meta-analyzed studies out . . . then so be it. . . ."

Studies that otherwise are arguably part of the ganzfeld corpus should not be "knocked out" of the meta-analysis. They should be included and categorically identified to determine what difference the difference makes.

" . . . block these variable stimulus types off . . ."

Yes, block them.

"*or* define a standard and stick to it . . ."

Defining a standard has little value compared with accepting the spectrum of more or less similar studies, and doing analyses that rely on blocking or categorization to impose standard definitions which have utility (or even aesthetic, preferential value).

"Sorry folks, but I would say no. . . ." It seems to me that nobody has the perspective to say no, or indeed yes, to such a question. But it is moot. Just do the meta-analysis so as to compare yes and no. What's the problem with that?

**#34** The suggestions of the writer of message #33 make a lot of sense to me: Include all that seem relevant, categorize or block appropriately, and note the patterns that emerge. (Find, empirically, whether a difference makes a difference.) These suggestions carry the flavor of "bottom-up" empiricism, rather than (possibly premature) "top-down" theorization. Once patterns (and their densities) have been identified, we can begin to understand, explain, and replicate them.

**#37** From message 35: " . . . (the pre-eminence of visual ESP in the ganzfeld) follows as a natural feature of the ganzfeld."

Is a visual imagery (versus other sense modalities) enhancement a "natural" feature of the ganzfeld? Or, could it be that we find a lot of visual imagery and visual mentation-target correspondence because we unconsciously arrange for that and expect it based on the types of targets we arbitrarily use, our nearly exclusive search for solely visual correspondences, and the subtle and not-so-subtle ways we might arrange the demand characteristics (subtle instructions, set, and setting factors that influence what the percipients expect and, hence, report to us) of our ganzfeld procedures? If we focus things so strongly in a visual direction, of course we will find much visual material. We could open our focus, however, and we might encounter much richer—and less expected—gifts.

**#48, part** (response following the sequence of message #25, "Is relaxation a component of the standard ganzfeld?"; and message #27, "Sorry folks, but I would say no. . . ."; and message #33 " . . . It seems to me that nobody has the perspective to say no. . . . Just do the meta-analysis to compare yes and no. What's the problem with that?")

The problem is that one can either do experimental science experimentally, or not. Meta-analysis is a post hoc quantitative review technique, not an experimental technique. Meta-analysts do not randomly allocate studies to fixed levels of an independent variable: They arbitrarily impose conceptual categories upon studies to observe (all other things being equal) the potentially causative effects of those blocked or moderated variables. But the problem is precisely that all other things are not equal in meta-analyses, . . . [for example] sample sizes, experimenters, testing environment. All of these things might act, alongside the blocked variable of interest, as covariates of either the blocking variable or the dependent variable (effect size)—and we are usually none the wiser at the end of the day. A good case of this is in the latest *Journal of Parapsychology*, where Palmer and Carpenter show that the supposedly confounded extroversion/forced-choice ESP effect . . . is reinstated. . . .

Now, blocking and moderating analyses often occur with incidental, undesigned features of studies and, in that sense, I think it is appropriate because it may be the only way of getting a handle on the diversity between studies and their consequence for study outcome: but whether relaxation is part of the ganzfeld or not is for me . . . to be determined . . . by the background theory of the ganzfeld. Relaxation is, for me, an additional feature of psi enhancement which augments the coherent features of ganzfeld procedures. But if you want to know whether the ganzfeld component works without the relaxation component and vice versa, then you are best off doing that within the experiment. . . .

Finally, if you think the ganzfeld *was* based around some kind of theory/model . . . [or] "perspective," then isn't it a simple matter of fact that

we do have the perspective to judge this issue Yes or no? If you think the theory is wrong, where's the alternative theory?

My core view on all this is [that one must] define a standard for the essential features of the ganzfeld, and block and moderate on the nonessential features. ...

**#73, part** I am the author of message #1. . . . You ([Julie] and Richard), in your replies, say that my specifications for a standard ganzfeld are not specific enough to prevent arbitrary decisions being made. But science is progressive, and they can be made specific—for instance, cut-off points on the sheep-goat scale. Even exposure to a training tape is an objective, yes or no?

Most, if not all, of the other requirements on my list are, I think, specific. If not, can you say which?

**#81, part** Julie Milton: In response to message #73, we appear to be at cross-purposes. You are arguing, as I understand it, that there is such a thing as a standard ganzfeld (one that replicates the PRL work in its essentials or that is psi conducive). . . .

I would argue that there is no current, generally accepted definition of a standard ganzfeld, and although one can come up with rules for one, such as you have done, others may disagree because there is no clear evidence basis at the moment on which to decide which features are crucial. As you will have seen, there has been plenty of discussion on this issue with a wide range of opinion on what might constitute a standard ganzfeld.

**#87** Adrian Parker: Question 1. Do we disagree about what is a standard ganzfeld? Obviously there is in terms of equipment and procedure: Ping-Pong balls, 30-minute ganzfeld, emotionally engaging targets, 1 in 4 binary choice, white noise or seashore noise. You would have a hard time arguing that this is not standard but might choose to do so, which prompts the question, "Do you?"

Question 2. Where I am more certain we disagree is concerning the importance of the selection of the right participants. I think this is unfortunate: It can be objectified. . . . Is there now enough literature in the history of parapsychology for us not to have to argue about the relevance of such factors, and [for us] . . . to agree that this should be the next priority in defining the successful ganzfeld-replication paradigm?

**#89** Julie Milton: Thanks, Adrian, for your questions. . . . In message #87, you are asking me whether I agree with your definition of a standard ganzfeld. My problem in answering is that you do not define what you mean by a standard ganzfeld before you ask whether I agree on its features. What is your rule for choosing what features are necessary for a standard ganzfeld? Do you mean features common to all the PRL series, for example, or features that you believe to be psi conducive?

*Psi-Conducive Conditions*

**#1, part** . . . No one has ever argued that the ganzfeld works for everyone under all conditions, so maybe if we are really interested in the answer we should turn to what Dalton, Parker (who has the experience of conducting one unsuccessful study—Parker, Millar & Beloff (1977), and Wezelman (who was involved in both successful and unsuccessful studies) have to say about it. This is to be conveniently found, all in one place, in the *European Journal of Parapsychology, 13* (1997). Perhaps we make so little progress in this field because we do not take seriously what others have written. Indeed many of the successful conditions were spelled out for us in the first *Journal of Parapsychology* and by Rhine in the 1948 *Journal of Parapsychology.*

**#17, part** Would it be fair play for only so-called psi-conducive experimenters (assuming we are able to know who they are) to register their studies? This approach using only so-called psi-conducive investigators may raise difficulties, of course. Novice experimenters may not yet know whether they are or are not psi conducive, and it may cause some people to feel resentful if their studies cannot be counted in the meta-analysis. Some people may be not clearly psi conducive nor clearly psi inhibitive. Should they be allowed to register? It would also spotlight those ostensibly psi-conducive experimenters who are included, and they may not want the inevitable close scrutiny that would result should the future meta-analysis be successful. . . .

Another problem that I think we just have to live with is that of interpersonal dynamics. There have been some studies where the dynamics amongst investigators have changed over time, and it has been argued that this change in dynamics resulted in a dampening of the study's results. . . .

**#24, part** The "ganzfeld" context includes the specific protocol, to be sure, but it also includes—and, I believe, quite importantly—the particular researcher, laboratory staff, and participants involved; the entire "atmosphere" of the laboratory; the expectations and qualities and histories of the researchers; and many other "out-of-protocol" features.

**#41, part** (response to message #17 on psi-conducive experimenters) This is a clear difficulty. Message #24 makes the point excellently.

**#49, part** (Response to message #6) At the end of his contribution, the author states that "a critical eye is perceived (by psi proponents) as a threat to the 'success' of the experiment." I think we must concede that psi results, like many results in behavioral science, are susceptible to experimenter effects: It is clear from the literature that some experimenters consistently get significant results in their experiments, whereas others just as consistently do not. I also think it is probably true, although I don't know of any systematic survey, that . . . successful experimenters are more

likely to come from among "believers" than "nonbelievers," and vice versa. So, let's grant for the sake of argument that this latter statement is true. First, belief cannot be an immediate cause of the diverging experimental outcomes; it must lead to something else as the more immediate cause. To conventionalists like the author, the immediate cause is presumably incompetent experimental technique. To most psi proponents, it is how well the experimenter is able to put the subject at ease and, in particular, inspire confidence in the subject that they can succeed in a task that even for psychics requires exercising a subtle and unreliable skill on cue. I think it is reasonable to suppose that (a) someone who believes that the skill is possible and can be produced by the subject in the experimental context will be better able to inspire confidence than a nonbeliever, even if both are making a sincere attempt; and (b) the amount of confidence the subject has in their ability to perform the task will, indirectly at least, affect that performance. In other words, assuming we don't prejudge the reality of psi, the confidence hypothesis is at least as plausible as the incompetence hypothesis, not to mention the fact that the incompetence hypothesis is not supported by correlations between study quality and effect size in meta-analyses. Thus, as the experimenter effect can be explained by both sides, it's existence does not argue either for or against the reality of psi.

**#73, part** But yes, surely what we are after are the psi-conducive conditions? You [Julie] talk about proof-orientated research—I thought we were talking about replications which surely are required to take into account process research in order to make them work?

**#79, part** The construction of "psi-conducive" experimental conditions is usually considered one of the basic defining features of the ganzfeld "sensory deprivation" or, if you prefer (see message #18), "sensory monotonization" technique. Presumably, this also counts for a number of other techniques that are used in parapsychological experiments in order to put subjects at ease. Consequently, both the Milton and Wiseman meta-analysis (1999 Psychological Bulletin, pp. 389-391) and Milton's updated meta-analysis (1999, pp. 321-322) pay some attention to the construction of "psi-conducive" conditions. "Conduciveness to psi" is also discussed . . . [by #17, #41, #48, #73].

However, what do we mean when we refer to "psi-conducive" conditions, states, or experimenters? In my opinion, the author of message #14 comes close to putting the right questions. She or he asks, . . . "What are the rules for deciding which features have evidence for being psi conducive?" I contend that we do not have any such rules. This is because we use the phrase "psi-conducive condition" in a way that prevents the identification of a condition that deserves that name before the experiment is conducted and evaluated. . . .

A surprising number of empirical studies that set out to put the concept of "psi-conducive states" to the test (e.g., Braud 1978; Braud & Braud 1973) have produced significant (and sometimes highly significant) results that seem to support the effectiveness, say, of relaxation or the ganzfeld technique. And, by implication, they also seem to have confirmed the usefulness and adequacy of the concept of "psi-conducive" experimental conditions. Other experiments, however, have produced mere chance results.

In the case of a study that has produced significant results in the predicted direction, parapsychologists claim (with some statistical justification) that "psi" (whatever "psi" might mean) was operating in the experiment. They are also tempted to say that the experimental manipulation (such as the ganzfeld technique) has been effective, that is, that it has managed to create "psi-conducive" conditions for the subject. And they seem confident that the concept of "psi-conducive states" also has received empirical support. But, has that concept really been supported?

What can we assert about the state of the subjects if the experiment only produced chance results? What can we say if the unsuccessful subjects assure us that they greatly enjoyed the experimental manipulations, that they held strong prior belief in the existence of psi, that they felt both entirely relaxed and calmly attentive, and that they were confident of their psi successes? Can we say anything, in this case, about the "conduciveness to psi" of the experimental manipulations such as muscular relaxation or the ganzfeld technique?

Were the experimental conditions "psi conducive"? Were the subjects in a "psi-conducive state," or weren't they? If we answer this question affirmatively, why don't we find "psi" in the data then? And if we decide negatively, what then becomes of the characteristic features of Braud's "psi-conducive syndrome" and of the experimental manipulations that were supposed to bring about a "psi-conducive state"?

My point is that the concept of "psi-conducive" states or conditions is begging the question, because our ascribing "conduciveness to psi" . . . is not independent of the empirical results of our experiments. "Conduciveness to psi" appears to be tacitly (and illegitimately) defined via the subjects' experimental success. . . .

If the only means that we have for establishing whether experimental conditions were "psi conducive" is that of looking at the empirical results of the experiment in question, the concept of "psi-conducive" states or conditions becomes thoroughly trivial and uninteresting . . . [and] can not be used for the definition of ganzfeld or other standard experimental procedures.

**#82, part** Regarding some of the comments of the writer of message #79: "Psi-conducive conditions"—by definition—can only be determined in terms of whether or not they promote psi. So, initially identifying them

depends on whether psi results obtain in association with such techniques. The apparent circularity issue can readily be handled by defining psi-conducive conditions in one context (set of studies) and then testing those same conditions in other, independent studies to learn whether the conditions are indeed (i.e., continue to be) psi favorable. A further way of assessing the relevance and role of psi-conducive factors is to manipulate them or assess the degrees to which they occur, and then use statistical contrasts or correlational methods to learn whether variations in psi-conducive condition or state indeed covary with psi measures. Rex Stanford, William Braud, and other investigators used precisely these strategies in their early ganzfeld and relaxation work, and did find the expected covariations.

Never has it been claimed that such psi-conducive conditions are the only factors that promote psi. Other things being equal, such psi-conducive factors help set the stage for psi or may be favorable to its occurrence, detection, or reporting. All other things (e.g., need conditions, investigator expectations, other psi-modulating factors) are rarely equal, however; so, it is not surprising that not all subsequent tests of psi-conducive conditions would not invariably yield the same or similar results.

**#84** The author of message #79 wrote: "What can we say if the unsuccessful subjects . . . [enjoyed the experiment, believed in psi, felt relaxed, attentive, confident]?"

. . . Interesting comments. I guess the answer is that "on the whole" (e.g., in a meta-analysis of studies meeting all those conditions) results should show an overall above-chance effect. If occasionally they did not . . . this would, presumably, just indicate that we do not know the whole story yet (i.e., there are psi-inhibitory factors that we do not know about and have, therefore, not categorized. . . .) or that psi is not wholly reliable. The problem comes if a meta-analysis shows that ideal studies (however defined) come up with an overall null outcome. The question, perhaps, is whether we know enough about what the conditions might be even *on the whole* to make an evidence-/proof-based meta-analytic approach a tenable one.

If we say "no," is this because—

(a) we don't think meta-analysis is the appropriate tool for "proof" (e.g., large well-controlled studies are better tools);

(b) we don't think we have yet shown a reliable psi effect;

(c) we think we have a psi effect, but we aren't sure what works and what does not (and thus there are too many unknown factors to make a meta-analysis possible);

(d) we think we do have a psi effect but we don't think we have enough studies that are similar enough with the relevant conditions for a valid proof-based meta-analysis to be conducted on them.

If we say "yes," is this because—

(a) we think the Milton and Wiseman meta-analysis is faulty (for reasons x, y, and z) and the other ganzfeld meta-analyses show there is a replicable effect;

(b) we accept the Milton and Wiseman meta-analysis but think it (and possibly others) was not conducted in a way that would answer a proof-oriented question because of the difference between the studies [in] the database. If conducted in a (yet-to-be-defined) way suitable for proof, a meta-analysis would provide good evidence for psi;

(c) regardless of the meta-analytic outcomes, there have been a number of experimenters who relatively consistently get above-chance results, and it is hard to explain those results away (and these people could be specified in advance and then their future work used in a meta-analysis).

There are doubtless other options I have missed. . . .

**#85** I strongly suspect that when all is said and done, the one "psi-conducive condition" that will prove reliable is the identity of the experimenter or principal investigator, with the other candidate conditions proving to be confounds of this. It could be that the reason ganzfeld results have declined in recent years (to the extent they have declined) is that psi-conducive investigators like Honorton, Braud, and (with due qualification) Sargent are no longer conducting them, and they have been replaced by only one psi-conducive newcomer. . . . I don't claim that the investigator hypothesis has been demonstrated by appropriate retrospective analyses. . . . I simply think that it is a possibility. . . . If the hypothesis is in due course confirmed, the next question will be, "What are these investigators doing to produce their superior results?"

**#86a, part** Message 64 wrote, "I agree that we should try to ascertain why the three outliers failed to conform to the rest of the distribution."

One key issue, brought up directly or indirectly in several messages, is the question of experimenter motivation. Regarding the Williams et al. study, the authors suggest that at least one factor may be the increasing interpersonal tension experienced by the group as the study progressed. A second potential factor involves the possibility of changing motivations during the course of the experiment if researcher or participant see the results developing in a consistent negative direction. This may be especially likely when the experimental design involves a comparison among conditions such that researchers are wanting at least one condition to be less successful. In the Williams et al. study, the authors note that: "We are not unduly disappointed by these results. We do not have a mere chance result, this much is clear. Furthermore, a missing effect of this magnitude stands in need of some discussion—were there any uncontrolled factors which could account for it?" Without continued failures as the study went on, the authors probably would have had less to discuss and would have

felt more disappointed. It could be difficult to resist a change in motivation under such circumstances.

Differential experimenter motivation is a difficult factor to control for and essentially impossible to measure directly. Some researchers have emphasized, at least informally, affirming for each session a commitment to obtaining the best possible results for that session, regardless of condition. If we view sessions as constituting systems, then this aspect of experimenter motivation and attitude may be important regardless of whether parties concerned are blind as to condition on a given session.

**#91** In two recent messages, the important roles of psi-favorable and psi-unfavorable investigators were mentioned. Several excellent suggestions were made. In addition to those, I add these two, as possible approaches:

1. Recall the work that Maher and Schmeidler did decades ago: They videotaped investigators and had judges rate their characteristics. Some interesting results emerged. This approach could be revisited and expanded.

2. Martin Johnson's Defense Mechanism Test could be modified, so that the stimulus items have relevance to the possible psi issues of investigators. By administering such an assessment to experimenters, investigators, and potential participants (as well as to counteradvocates of the field!), one might be able to begin to get at more "unconscious" motivations and other characteristics. This might be a particularly useful tool to use in attempting to identify possible unconscious resistances to psi (and "fear of psi") as well as other forms of investigator motivation.

I'm certain that investigators interested in exploring such issues would have no trouble designing many creative approaches to studying investigator qualities and motivations. . . .

## Future Research on the Ganzfeld

### *General Recommendations*

**#6, part** The reported nonrandom experimental effects are often interpreted as "psi" effects, whereas a more neutral view would be to consider them as Unidentified Experimental Effects (UEE), and then to work hard to identify any flaws (just as the UFO experts, most of the time, reveal interesting natural explanations). Once parapsychologists succeed in producing pure chance results, they will convincingly have shown their skills as experimenters. In case of extreme or consistent individual nonrandom results, there should be an active invitation of critics to follow the procedure—not the other way around, that a critical eye is

perceived as a threat to the "success" of the experiment. I mention this from personal experience.

**#7, part** . . . [quotes message #6, Milton and Wiseman study leading to the hypothesis that psi does not exist]

Your time could and should be spent to good research in order to get a better idea of the processes underlying psi phenomena. If we accept their reality as a working hypothesis that best fits the data, we have an obligation to continue our research, not to continue our debates with skeptics.

**#9** Very soon, my new ganzfeld experiment will *prove* its efficacy!!!!!!!!

**#17, part** Not everyone conducts a study with the view of it being later included in a meta-analysis, but rather with the view of finding out more about what works and what does not work. . . .This then raises the question (posed essentially by Milton) of whether we should be organizing ourselves more clearly for the next meta-analysis so that everyone is happy.

Perhaps it is what we should be doing in part. One possibility would be for those who are conducting a study with a view to it being included in a future meta-analysis to preregister this intention before the study begins. These people should, we hope, be those who will try to meet all the necessary procedural safeguards in their studies and should perhaps follow the defining ganzfeld criteria that are yet to be agreed upon. . . . [Include] only so-called psi-conducive experimenters? . . . If there are coexperimenters, they should register . . . only if they have worked successfully together before?

It depends, too, on what we would want a future meta-analysis to show:. . . that psychic functioning is possible, . . . that we are at least able to identify experimenters who can produce good results, . . . that it is the method (i.e., the ganzfeld) that works regardless of experimenter? . . .

**#24, part** I invite us to treat what might be some unexamined assumptions underlying this ganzfeld debate. I think we may [be] assuming that the various ganzfeld meta-analyses are really about the ganzfeld. Of course, that is a major component. However, because of the absence of control or contrast groups or conditions in nearly all ganzfeld investigations, we are quite uncertain whether the ganzfeld procedure has anything to do with the obtained results. My own opinion—based on decades of research in this and related areas, and my readings and discussions with others—is that all the ganzfeld studies really tell us is that sometimes psi effects occur in the laboratories of certain investigators. (Homogeneity or heterogeneity of effect sizes across different labs or experimenters does not fully address this issue.) To tie even the successful ganzfeld results to the ganzfeld itself is, I feel, premature and unwarranted by present evidence. . . . [see excerpt under Psi-Conducive Conditions] I suggest that we devote more attention to those features, other

than the ganzfeld protocol itself, that may really be important in the "success" or not of the complex "package" that we are calling "the ganzfeld." I would suggest that similar arguments apply, mutatis mutandis, to other ostensibly efficacious procedures such as "remote viewing" and so on.

**#26** Relying on the quantitative findings from individual ganzfeld studies (hit rates, significance levels, effect sizes) or from ganzfeld meta-analysis is one approach to seeking evidence for the existence of psi functioning. I urge that we not ignore the qualitative correspondences in such studies—and in a wide range of other research areas—as an additional, and perhaps equally or more important and potentially convincing, source of evidence for psi effects. I would extend this suggestion further. Laboratory studies are only one venue in which to look for psi-relevant evidence. Let's not forget other places to look: field studies of spontaneous cases, our own personal psi experiences, and so on.

**#41, part** Message #17 suggests that those wishing to be included in a future meta-analysis should pre-register. This seems eminently reasonable. As noted, that would still leave others free to "try variations."

. . . If our purpose is to "convince skeptics of the existence of anomalous interactions," then particular experimenters who get consistently unusual results may be evidence for just that. Perhaps even more important than this exercise, would be further *joint* investigations into the "experimenter effect" (with common subject pool, protocols, and so on)—like the Schlitz and Wiseman series.

**#42, part** It seems to me that the results of this meta-analysis—without reference to any previous studies, and without argument about studies that should have been added or subtracted from the analysis—indicate that continued serious study of the ganzfeld psi effect is definitely warranted.

**#48, part** My core view ... is define a standard for the essential features of the ganzfeld, and block and moderate on the nonessential features. Definition of a standard is not armchair philosophizing: It has in the context of this debate the purpose of setting certain basic standards for the inclusion of future ganzfelds into a meta-analysis, and is offered so that parapsychologists *and* skeptics cannot resort to questioning the absence/presence of certain studies because they don't meet their pet definition of ganzfeld. (Or being more cynical, because those studies do or don't get the desired effect sizes.)

**#55, part** I'm not sure we can talk about whether we want evidence-oriented meta-analysis until we have a better idea of whether we do have agreement about what the necessary procedures and characteristics of a "successful" ganzfeld experiment might be. Unless we want to say that the ganzfeld is generally successful for most people in most situations (which I rather doubt)?

**#75** If parapsychology is to be/remain on the cutting edge of science, it must not forget to maintain its passion about the phenomenon itself other than in terms of statistical anomalies. Bearing in mind psi's lack of regard for logic and replication, maybe we should "float" our feelings now and then and observe the results. Beethoven's greatest works would have been destroyed by statistics!!!!

**#82, part** In this "ganzfeld" debate, I feel a critical issue of whether the "ganzfeld meta-analyses" really have anything to do with the ganzfeld procedure, per se (as opposed to researcher or other characteristics), has not been addressed adequately. Until investigators begin including contrast or control conditions in their studies (e.g., ganzfeld versus non-ganzfeld controls, relaxation versus non-relaxation controls), the relevance of "ganzfeld" or other ostensibly psi-conducive conditions will remain needlessly shrouded in ambiguity.

**#83, part** The messagers have expressed a variety of viewpoints on what represents a standard ganzfeld procedure, which studies should be excluded, how or whether to trim the outliers, and a number of other contentious issues. We are far from agreement on the most fundamental question of whether there even exists an anomalous effect (psi) in the new ganzfeld studies. Milton and Wiseman conclude that there is no psi effect whatsoever in the new database. Other, equally valid, methods of analysis do find such an effect. . . . Personally, I think Milton and Wiseman overstate their case because they ignore the totality of the evidence. However, it is clear from the many messages that there is room for reasonable disagreement on this most important question. That is why I believe a consensus is emerging among parapsychologists and skeptics alike on Julie's call for a preplanned meta-analysis. A future meta-analysis, jointly designed by parapsychological researchers and skeptics, may be the only way to resolve this issue to the satisfaction of both groups.

**#86** Message #83 included this section: . . . [quotes last two sentences: emerging consensus for a pre-planned meta-analysis].

This suggestion sounds, to me, suspiciously like something that K. R. Rao planned many, many years ago, and also like something John Beloff recommended many, many years ago. The very formal, "predesigned" and "preregistered" effort that Ram planned seemed to dissolve away, and something very like John's "commission of inquiry" eventually occurred—in a way—in the form of the infamous National Research Council report—which, in turn led to counterreports, counter-counterreports, and so on, and so on. And here we are, once again. Seems like déjà vu all over again.

And, of course, we had the Bem and Honorton *Psychological Bulletin* report, and now we have the Milton and Wiseman report. Hmmmm.

**#86a, part** If we are looking for implications for future research and meta-analyses, it may be useful for researchers to consider in some detail

their own motivational states from session to session and consider some of the strategies articulated by psi-conducive researchers such as Schmidt and Braud who have expressed themselves publicly on these topics. In experimental write-ups, these factors could be discussed more, such as to present a more complete description of the experimental system, and to allow for possible future meta-analytic coding for such variables. Researchers may be able to include concrete strategies used to minimize decreases in motivation from session to session. Techniques for controlling the impact of session success and failure upon future events are increasingly available within the context of performance enhancement training. It may also be useful to code for the existence of conditions in which the research team has less admitted motivation for good results. Although, once again, it is not realistic to expect direct measures of motivation, there may be a variety of techniques for their indirect assessment such as to avoid after-the-fact registry biased by knowledge of results. If one is considering criteria for designating which studies are demonstrations of effect motivated rather than, or in addition to, process motivated, then perhaps researcher commitment to strong effects for all sessions should be included. Or within a study one could consider declaration in advance of those sessions for which researchers are committed to full strength results. This is in fact very similar to suggestions made previously but might be a slightly different way of looking at it.

**#90** Message #83 said . . .[quotes last two sentences: emerging consensus for a preplanned meta-analysis].

For sure if we don't try, we won't achieve anything: But as message #86 pointed out, we have been here before. Presumably, we would hope to ensure even tighter controls the next time around, but do we have reasonable grounds for optimism?

It seems to me that a possible analysis looks like this:

1. We are trying to demonstrate psi, reproducibly.

2. *Some* experimenters seem to be able to do this with reasonable consistency.

3. Skeptics take the fact that only some experimenters get good results as evidence against the psi phenomenon.

4. "Psi-conducive" experimenters seem to be doing one of two things (or, of course, both); namely, inducing "psi-conduciveness" in their subjects, by providing an encouraging environment, actually affecting the results by use of their own psi.

5. Both of the above are evidence for psi, but both militate against us being able to demonstrate a standard "psi-conducive" set-up.

6. To eliminate accusations of experimental artifact, we need to do more joint trials involving psi-inducers and skeptics.

In summary, establishing an "experimenter effect" looks increasingly like our best shot at demonstrating psi. (I realize this is something of a repeat of messages #24, #41 (my own) and #55.)

**#94** (response to message #90) Is the object (a) to demonstrate psi reproducibly, or (b) to understand what has produced the observed variance in results? I don't see how we can do (a) until we have accomplished (b). . . . A standard protocol, and limiting future reviews to studies that use it, only make sense if we agree that we have achieved (b). It seems clear that we do not. Let's not artificially limit the exploration.

**#95, part** Many of the suggestions involved looking at variables in order to detect pattern, answer questions, etc. In the most successful examples of this in other areas, the data bases have tended to be much larger than even our existing cumulating data base. Thus, one option for the future is to think less in terms of discrete future analyses and more in terms of the rules that might govern an ongoing analysis systematically updated at regular intervals. It would be inclusive more than restrictive and would allow researchers to see how large the data base was with respect to issues of importance, thus providing an additionally effective source of information about the state of the available information, including which areas are insufficiently researched to allow meaningful analysis and interpretation. New studies could be coded upon entry to the literature, perhaps by authors themselves and by an independent centralized entity. As has been argued before, having to code your own studies can be a great help in planning and conducting the studies themselves.

*Recommendations That Explicitly Consider Non-parapsychologists*

**#31, part** It may be simplistic to talk of "proving" a phenomenon's existence, and this may not be the most interesting question to many parapsychologists. But "Is the existence of ESP proven?" is probably the question most commonly put to me by people uninvolved in parapsychology. Others want to know the answer. I think parapsychologists should address this question, and I believe preplanning a meta-analysis, specifying inclusion criteria in advance, is a good starting point.

**#38** (response to messages #6 and #26) What are our aims? Presumably, we would all claim to be seeking scientific truth (within our conscious perception of a nondeterministic universe). Hopefully, too, we are prepared to follow where the science leads—even if that does not support our current hypotheses. Those of us who believe that existing theories do not explain all apparent information transfer between conscious bodies and their environment would hope to convince other scientists that there is an anomaly here that deserves study. In the words of message #6: "Scientists in other fields would certainly be prepared to accept experimental evidence for its existence if either of the above mentioned two effects

would have been successful (either a repeatable experiment or an individual experiment without crucial flaws)."

It is clear that any such evidence must conform rigorously to established empirical epistemology. Of course, it is possible to debate whether our epistemology is sufficient; but unless and until the whole of science agrees something different, the existing paradigm is all we have. Certainly, if we wish to be taken seriously by scientists in other fields, we must conform to their rules.

This is in contrast to the view in message #26, which said, "I urge that we not ignore . . . [qualitative correspondences, field studies, personal experiences, and so on]." . . .

Any or all of these things may suggest avenues of research; however, in the final analysis, these "experiences" can be held to have relevance in our "consensual universe" only if they are tractable by normal scientific principles.

As stated by Julie, the immediate aim of this debate is "[the] need to plan the next proof-oriented meta-analysis now, setting up agreed criteria to exclude unsuitable studies from that meta-analysis before they are conducted." Despite any objections which might be made to this approach, it is surely worth putting some effort into it in an attempt to produce some "experimental evidence for [psi's] existence."

**#46** Julie Milton: (response to message #38) The excerpted quote (from the invitation to the debate) is accidentally misleading. . . . I am not suggesting that we use this debate to actually plan such a meta-analysis. . . . The purpose of the debate is rather to allow people, as a first step, to question and discuss whether such a planned meta-analysis is necessary or desirable. Although I think it is, anyone . . . might well not think a preplanned meta-analysis necessary.

**#52, part** Knocking out three negative results from the meta-analysis is going to look like special pleading to outsiders, even with the rationale offered. If we want to challenge the conclusions of the Milton and Wiseman meta-analysis, let's find better ways of doing it than this.

**#53, part** . . . [quotes message #31, that to "prove" ESP, a preplanned meta-analysis, specifying inclusion criteria in advance, is a good starting point]

Depends on what you do next. We must not neglect the effect of the inclusion criteria themselves. But, as has been said so often it is surprising it needs to be repeated, you can have your cake and eat it too: Just make your inclusion criteria inclusive and pay attention to differences that might make a difference. You can even have an exclusive inclusion category (if you can defend it against criticism) which serves your interest in the existence conundrum. But we all should be aware that science is an inexorable force: It does not indefinitely abide false "answers," even to the difficult questions of parapsychology.

*Radin's Meta-Analysis*

**#36** As an important purpose of the debate seems to be to answer the question of whether the PRL autoganzfeld has been successfully replicated, it would help focus the discussion, as well as provide useful information, if we could do a detailed comparison between Milton and Wiseman's negative meta-analysis and the positive one conducted by Radin. . . .

**#43, part** (response to message #36) Dean Radin: Here's the approach I took for the analysis in my book, *The Conscious Universe.* Like Milton and Wiseman, I scanned the relevant journals and proceedings for ganzfeld studies, but unlike Milton and Wiseman, I (a) contacted all of the people I could think of who had conducted ganzfeld studies and had not published them yet, and (b) considered ganzfeld studies only using visual targets (thus excluding Willin's audio ganzfeld tests).

As I noted in my book (p. 87-88), the studies I considered were those from Edinburgh, Amsterdam, Utrecht, Cornell, Rhine Center, and Gothenburg. It appears from the Milton and Wiseman list that I missed the studies by Stanford and Frank, and McDonough et al. In all cases, I calculated hit rates from the published (or personally reported) number of hits and trials. I was able to retrieve a total of 1,432 trials (compared to Milton and Wiseman's 1,198): Edinburgh (331), Amsterdam (164), Cornell (25), Rhine Center (590), Gothenburg (90), Utrecht (232).

Bem's experiment was a differential ganzfeld study involving meditators (25 sessions) and nonmeditators (25 sessions). I did not include the nonmeditator data in my analysis because that group was predicted to not perform as well as the meditators (which is what happened), and I couldn't justify including in a proof-oriented meta-analysis a subset which was predicted to "not" perform.

**#58** [quotes message #43, on omitting those of Bem's subjects predicted to "not" perform]

Daryl Bem's study is unpublished, so I cannot check whether he predicted nonmeditators not to perform at all or to perform less well than meditators—two entirely different things. If the rule was to exclude conditions in which lower performance was predicted compared to another condition, was the rule applied consistently across the database, for example, to the static target condition of McAlpine's study (Morris et al., 1993)?

Also, in *The Conscious Universe,* the question posed is whether " . . . studies after the autoganzfeld studies continue(d) to successfully replicate the psi ganzfeld effect." (p. 87) In order to compare the two meta-analyses, it would be helpful to know whether the completion of the final autoganzfeld study was the relevant date after which studies were eligible and whether it was their publication or conduct date that counted.

**#62** [quotes message #43, on omitting the subset predicted to "not" perform]

This puzzles me. If a regular ganzfeld with nonmeditators is excluded, in this case, because such persons would not perform well, then what does one do with the many *other* ganzfeld studies in which "nonmeditators" were the participants, and in which they did perform well, and which were not removed from this and other global analyses? It strikes me as odd to exclude a particular subcondition in one study, simply because an author predicts its participants will not perform "as well as" those in another condition, when that same subcondition does result in positive results in other studies and is included in other analyses. An issue of consistency is involved here. Are we now saying that the "standard ganzfeld" cannot include nonmeditators as participants?

Additionally, as message #58 also suggests, it is not clear whether the prediction of Bem was that nonmeditators would not perform at all or whether they would simply perform less well than meditators. I agree with the writer of message #58: These are two entirely different things.

## Is Psi Real?

**#6, part** At any given time in the history of parapsychology there have been claims of having at least the hope for a repeatable experiment. These hopes have never been fulfilled. In the same vein, there have been claims that there is convincing experimental evidence in some of the best experiments. It has turned out, however, that no single important experiment has stood the test of a focused investigation of an independent analyst.

The Milton and Wiseman study shows that the situation now is just the same as it has been for the last 60 years. There is no repeatable experiment, and the experiments that gave rise to such expectations were flawed. All these failures by skilled experimenters to produce anything but elusive "evidence" of psi should alert parapsychologists to a neglected hypothesis in the field—the hypothesis that psi does not exist. In my opinion, the sum of parapsychological experiments strongly supports this "null hypothesis," based on these two results:

1. No individual experiment considered important by parapsychologists has passed the test of a close scrutiny by an independent expert.

2. No repeatable experiment has been produced in spite of constant efforts.

When confronting people in pseudoscientific areas they are sometimes asked to state under what conditions they would be prepared to reject their favored hypothesis. I think it is time for parapsychologists to

state under what conditions they are prepared to abandon their belief in the existence of psi.

**#49, part** The purpose of this contribution is to respond to message #6.

The author challenges parapsychologists "to state under what conditions they are prepared to abandon their belief in the existence of psi." My answer is simply the converse of what critics like the author would say in response to the corresponding question of what would make them believe in psi: I would abandon my belief in psi if I find there is no credible evidence for it. I would be surprised if this answer is not acceptable to all psi proponents in this debate. . . . The author's challenge begs the more important question of what constitutes credible evidence for psi. . . .

The author attempts to address this latter question in part by stating two criteria he or she thinks the data have failed to meet. Unfortunately, both statements need elaboration before they can be intelligently discussed:

1. "No individual experiment considered important by parapsychologists has passed the test of close scrutiny by an independent expert." We need to know what is meant by "passed the test" and who qualifies as an "independent expert." A conventionalist (a term for "skeptic")? Anyone other than the investigator? Something else?

2. "No repeatable experiment has been produced in spite of constant efforts." Here, we need to know what is meant by a "repeatable experiment," especially whether the author means statistical repeatability or repeatability on demand. . . .

In his final paragraph, the author makes a statement that I found astonishing: "Once parapsychologists succeed in producing chance results they will convincingly have shown their skills as experimenters." Although stated in a maximally polite way, the meaning of this sentence is clear: Obtaining nonchance results in a psi experiment, by virtue of this fact alone, demonstrates a lack of skill, that is, incompetence, on the part of the experimenter. Not only does this statement presume a point of contention in this debate (i.e., is psi real?), it leaves the psi proponent no way out: If you get positive results in a psi experiment, your experiment was incompetent, regardless of how many controls, and so forth, you may have employed. In conclusion, I will also be polite and merely say that the author's statement is not scientific.

## The Value of the Debate

**#1, part** I have to confess to a reluctance to spend time on this debate. . . . I find it hard to spend time on endless debates which are often ego based and as the author implies this paper does not take us beyond the

Honorton and Hyman debate of 15 years ago. My view is that the issue will resolve itself when conditions for strong effects are specified so as to enable us to unequivocally demonstrate psi and to discover something new about it . . . Much of the more suspect parts of the paper should have been examined beforehand and then our time could have been spent more productively.

**#7, part** I call on all excellent psi researchers . . . to reflect if they want to continue to devote time to this type of debate which has been going for the last 60 years with no substantial effect . . . I urge you to withdraw from this fruitless endeavor like I will do.

**#19, part** First, I'd like to applaud Julie for making this debate happen. I continue to believe that meaningful and constructive dialogue between parapsychologists and skeptics offers a way for moving ganzfeld research forward, and parapsychology generally. Even if the debate doesn't result in an overarching consensus, it is still valuable to highlight the areas of agreement or disagreement and to just frame the questions and issues that we feel are important at this stage.

**#40** (response to messages #7 and #19) I should like to applaud Julie's initiative also, and thank her for inviting me to the debate. The "Scientific Establishment" holds all the power here. We have to treat with them: Refusing to talk will profit us nothing at all.

**#41, part** This debate is clearly useful. . . .

**#71, part** This debate was organized to help set criteria and standards to define a research area so future meta-analyses could finally answer the question whether there is psi or not. That definition is not likely to emerge, nor should it.

**#79, part** Julie Milton's basic idea to have us join forces for this debate was a good one. Valuable insights might have resulted. However, following . . . [the debate] has been a peculiar (and not always rewarding) experience. If it had not been for some rhetorical highlights such as . . . message #49 . . . this debate would have turned into a rather tedious enterprise.

I concede that a variety of features of the ganzfeld procedure (cf. messages #4, #13, #14, #17, #18, #25, #27, #33, or #62) as well as the purposes and characteristics of meta-analyses (such as in messages #5, #22, #29, #31, or #78) have been ventilated at some length, and that some contributors (see messages #48, #55, or #74) even have tried to integrate insights from these two strings of discussion. Nevertheless, there seems to be little agreement among messagers even about the most basic issues: What constitutes at least a minimalist definition of the basic ganzfeld procedure? Which criteria may be used to decide about specific ganzfeld studies' exclusion from or inclusion into subsequent meta-analyses? What, if anything, does the Milton and Wiseman meta-analysis of ganzfeld studies prove?

On balance, therefore, I just don't think this ganzfeld debate has got us much closer to its stated aim [assessing the need for a pre-planned meta-analysis]. . . .

**#83, part** The ganzfeld debate has succeeded in stimulating a healthy dialogue between skeptics and parapsychologists . . . If the debate were to accomplish agreement on [a preplanned meta-analysis] and nothing else it will have been worthwhile.

**#93** I am convinced that after a few centuries, historians of science and philosophers will see much (not all!) of this discussion as we see discussions of scholasticism about how many angels can sit on the tip of a needle.

## References


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

Bertini, M., Lewis, H. B., & Witkin, H. A. (1972). Some preliminary observations with an experimental procedure for the study of hypnagogic and related phenomena. In C. T. Tart (Ed.), *Altered States of Consciousness* (pp. 95-114). New York: Anchor Books.

Bierman, D. J. (1995). The PRL autoganzfeld revisited: Refuting the sound-leakage hypothesis. *Proceedings of Presented Papers: The Parapsychological Association 38th Annual Convention*, 43-47.

Braud, W. G. (1978). Psi conducive conditions: Explorations and interpretations. In B. Shapin & L. Coly (Eds.), *Psi and states of awareness: Proceedings of an international conference held in Paris, France* (pp. 1-34). New York: Parapsychology Foundation.

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.

Dalton, K. (1997a). Exploring the links: Creativity and psi in the ganzfeld. *Proceedings of Presented Papers: The Parapsychological Association 40th Annual Convention*, 119-134.

Dalton, K. (1997b). Is there a formula to success in the ganzfeld? Observations on predictors of psi-ganzfeld performance. *European Journal of Parapsychology*, **13**, 71-82.

Druckman, D., & Swets, J. A. (Eds.). (1988). *Enhancing human performance: Issues, theories, and techniques.* Washington, DC: National Academy Press.

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

Honorton, C., Barker, P., Varvoglis, M. P., Berger, R. E., and Schechter, E. I (1985). "First-timers:" An exploration of factors affecting initial psi ganzfeld performance. *Proceedings of Presented Papers: The Parapsychological Association 28th Annual Convention*, 37-58.

HONORTON, C., BERGER, R. E., VARVOGLIS, M. P., QUANT, M., DERR, P., SCHECHTER, E. I., & FERRARI, D. C. (1990). Psi communication in the ganzfeld: Experiments with an automated testing system and a comparison with a meta-analysis of earlier studies. *Journal of Parapsychology,* **54**, 99-139.

HONORTON, C. & FERRARI, D. C. (1989). Meta-analysis of forced-choice precognition experiments. *Journal of Parapsychology,* **53**, 281-308.

HONORTON, C., FERRARI, D. C., & BEM, D. J. (1998). Extraversion and ESP performance: A meta-analysis and a new confirmation. *Journal of Parapsychology,* **62**, 255-276.

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**, 136-168.

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

HYMAN, R., & HONORTON, C. (1986). A joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology,* **50**, 350-364.

KANTHAMANI, H., & BROUGHTON, R. S. (1994). Institute for Parapsychology ganzfeld-ESP experiments: The manual series. *Proceedings of Presented Papers: The Parapsychological Association 37th Annual Convention,* 182-189.

KANTHAMANI, H., & PALMER, J. (1993). A ganzfeld experiment with "subliminal sending," *Journal of Parapsychology,* **57**, 241- 257.

LAWRENCE, T. (1993). Gathering in the sheep and goats . . . A meta-analysis of forced-choice sheep-goat ESP studies, 1947-1993. *Proceedings of Presented Papers: The Parapsychological Association 36th Annual Convention,* 75-86.

MCDONOUGH, B. E., DON, N. S., & WARREN, C. A. (1994). EEG in a ganzfeld psi task. *Proceedings of Presented Papers: The Parapsychological Association 37th Annual Convention,* 273-283.

MILTON, J. (1993). A meta-analysis of waking state of consciousness, free-response ESP studies. *Proceedings of Presented Papers: The Parapsychological Association 36th Annual Convention,* 87-104.

MILTON, J. (1997). Meta-analysis of free-response studies without altered states of consciousness. *Journal of Parapsychology,* **61**, 279-319.

MILTON, J. (1999). Should ganzfeld research continue to be crucial in the search for a replicable psi effect? Part I. Discussion paper. *Journal of Parapsychology,* **63**, 309-333.

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

MORRIS, R. L., CUNNINGHAM, S., MCALPINE, S., & TAYLOR, R. (1993). Toward replication and extension of autoganzfeld results. *Proceedings of Presented Papers: The Parapsychological Association 36th Annual Convention,* 177-191.

PALMER, J., & CARPENTER, J. C. (1998). Comments on the extraversion-ESP meta-analysis by Honorton, Ferrari, and Bem. *Journal of Parapsychology,* **62**, 277-282.

PALMER, J., & KANTHAMANI, H. (1990). A ganzfeld experiment with subliminal "sending." *Proceedings of Presented Papers: The Parapsychological Association 33rd Annual Convention;* 227-242.

PARKER, A., FREDERIKSEN, A. & JOHANSSON, H. (1997). Towards specifying the recipe for success with the ganzfeld: Replication of the ganzfeld findings us-

ing a manual ganzfeld with subjects reporting prior paranormal experiences. *European Journal of Parapsychology*, **13**, 15-27.

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.

RADIN, D. I. (1997). *The conscious universe: The scientific truth of psychic phenomena.* New York: HarperCollins.

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

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

RHINE, J. B. (1948). Conditions favoring success in psi tests. *Journal of Parapsychology*, **12**, 58-75.

SCHMEIDLER, G. R., & MAHER, M. (1981). Judges' responses to the nonverbal behavior of psi-conducive and psi-inhibitory experimenters. *Journal of the American Society for Psychical Research*, **75**, 241-257.

STANFORD, R. G., & FRANK, S. (1991). Prediction of ganzfeld ESP-task performance from session-based verbal indicators of psychological function: A second study. *Journal of Parapsychology*, **55**, 349-376.

STANFORD, R. G., & STEIN, A. G. (1994). A meta-analysis of ESP studies contrasting hypnosis and a comparison condition. *Journal of Parapsychology*, **58**, 235-269.

SYMMONS, C., & MORRIS, R. L. (1997). Drumming at seven Hz and automated ganzfeld performance. *Proceedings of Presented Papers: The Parapsychological Association 40th Annual Convention*, 441-453.

WEZELMAN, R., GERDING, J. L. F., & VERHOEVEN, I. (1997a). *Eigensender* ganzfeld psi: An experiment in practical philosophy. *European Journal of Parapsychology*, **13**, 28-39.

WEZELMAN, R., GERDING, J. L. F., & VERHOEVEN, I. (1997b). *Eigensender* ganzfeld psi: The practical philosophy of an experiment. *European Journal of Parapsychology*, **13**, 40-53.

WILLIAMS, C., ROE, C. A., UPCHURCH, I., & LAWRENCE, T. R. (1994). Senders and geomagnetism in the ganzfeld. *Proceedings of Presented Papers: The Parapsychological Association 37th Annual Convention*, 429-438.

WILLIN, M. J. (1996a). A ganzfeld experiment using musical targets. *Journal of the Society for Psychical Research*, **61**, 1-17.

WILLIN, M. J. (1996b). A ganzfeld experiment using musical targets with previous high scorers from the general population. *Journal of the Society for Psychical Research*, **61**, 103-106.

WISEMAN, R., & SCHLITZ, M. (1997). Experimenter effects and the remote detection of staring. *Journal of Parapsychology*, **61**, 197-208.

WISEMAN, R., & SCHLITZ, M. (1999). Experimenter effects and the remote detection of staring: An attempted replication. *Proceedings of Presented Papers: The Parapsychological Association 42nd Annual Convention*, 471-479.

WISEMAN, R., SMITH, D., & KORNBROT, D. (1996). Exploring possible sender-to-experimenter acoustic leakage in the PRL autoganzfeld experiments. *Journal of Parapsychology*, **60**, 97-128.

*14038 Sunset Drive*
*Whittier, CA 90602*
*USA*

*Department of Philosophy*
*Rollins College*
*Winter Park, FL 32789*
*USA*

## Appendix

### Table A1

### Key to Authors of Messages in the Debate

| Message[a] | Author |
|---|---|
| #1 | Parker, A. |
| #4 | Milton, J. |
| #5 | Nelson, R. D. |
| #6 | Wiklund, N. |
| #7 | Bierman, D. J. |
| #9 | Willin, M. |
| #10 | Wezelman, R. |
| #12 | Wezelman, R. |
| #13 | Lawrence, T. |
| #14 | Symmons, C. |
| #15 | Milton, J. |
| #16 | Bierman, D. J. |
| #16a | Nelson, R. D. |
| #17 | Steinkamp, F. |
| #18 | Lawrence, T. |
| #19 | McDonough, B. E. |
| #24 | Braud, W. G. |
| #25 | Braud, W. G. |
| #26 | Braud, W. G. |
| #27 | Lawrence, T. |
| #28 | Willin, M. |
| #29 | Braud, W. G. |
| #30 | Braud, W. G. |
| #31 | Watt, C. |

| | |
|---|---|
| #33 | Nelson, R. D. |
| #34 | Braud, W. G. |
| #35 | Lawrence, T. |
| #36 | Palmer, J. |
| #37 | Braud, W. G. |
| #38 | Menzies, S. |
| #39 | Menzies, S. |
| #40 | Menzies, S. |
| #41 | Menzies, S. |
| #42 | Radin, D. I. |
| #44 | Radin, D. I. |
| #47 | Lawrence, T. |
| #48 | Lawrence, T. |
| #49 | Palmer, J. |
| #50 | Nelson, R. D. |
| #51 | Radin, D. I. |
| #52 | Palmer, J. |
| #53 | Nelson, R. D. |
| #54 | Schechter, E. I. |
| #55 | Steinkamp, F. |
| #56 | Nelson, R. D. |
| #57 | Nelson, R. D. |
| #58 | Milton, J. |
| #59 | Milton, J. |
| #61 | Lawrence, T. |
| #62 | Braud, W. G. |
| #64 | Palmer, J. |
| #65 | Milton, J. |
| #66 | Radin, D. I. |
| #67 | Radin, D. I. |
| #68 | Palmer, J. |
| #69 | Schechter, E. I. |
| #70 | Schechter, E. I. |
| #71 | Nelson, R. D. |
| #72 | Nelson, R. D. |
| #73 | Parker, A. |
| #74 | Palmer, J. |
| #75 | Willin, M. |
| #76 | Milton, J. |
| #77 | Nelson, R. D. |
| #78 | Milton, J. |
| #79 | Hövelmann, G. |
| #80 | Milton, J. |
| #82 | Braud, W. G. |

| | |
|---|---|
| #83 | McDonough, B. E. |
| #84 | Steinkamp, F. |
| #85 | Palmer, J. |
| #86 | Braud, W. G. |
| #86a | Morris, R. L. |
| #88 | Nelson, R. D. |
| #90 | Menzies, S. |
| #91 | Braud, W. G. |
| #92 | Milton, J. |
| #93 | Gerding, H. |
| #94 | Schechter, E. I. |
| #95 | Morris, R. L. |

[a]Messages in which authors identified themselves during the debate are not listed.