The Journal of Parapsychology, Vol. 66, March 2002 (pp. 73-82)

THE GANZFELD DEBATE CONTINUED: A RESPONSE TO MILTON AND WISEMAN (2001)

BY LANCE STORM AND SUITBERT ERTEL

ABSTRACT: Most researchers in parapsychological circles and beyond are familiar with the ganzfeld debate, which was revived in a series of articles that appeared in *Psychological Bulletin*. This article is a response to J. Milton and R. Wiseman's (2001) reply to L. Storm and S. Ertel (2001), who took issue with J. Milton and R. Wiseman's (1999a) claim that the evidence for psi in the ganzfeld was not replicable. The authors (Storm & Ertel) argue that in their reply, J. Milton and R. Wiseman (2001) misrepresented the issues raised in R. Hyman and C. Honorton's (1986) Joint Communiqué to their advantage. Milton and Wiseman wrongly took the standards of the Communiqué as implying low quality of all previous studies and downplayed the accumulated evidence that doubts about the credibility of pre-Communiqué ganzfeld researchers were unwarranted. They wrongfully belittled statistical significance, an important contributor to empirical evidence, and on mere circumstantial grounds, they ignored the necessity of the bidirectionality test, which is acknowledged as a unique psi indicator. The authors reassess the effect sizes for the various ganzfeld databases and conclude that Milton and Wiseman's critique is essentially out of place. For future ganzfeld and psi research in general, the authors recommend a process-oriented strategy.

The thrust of our critique of Milton and Wiseman's (1999a) article is condensed in our (Storm & Ertel, 2001) article's abstract:

> J. Milton and R. Wiseman (1999[a]) attempted to replicate D. Bem and C. Honorton's (1994) meta-analysis, which yielded evidence that the ganzfeld is a suitable method for demonstrating anomalous communication. Using a database of 30 ganzfeld and autoganzfeld studies, Milton and Wiseman's meta-analysis yielded an effect size *ES* of only 0.013 (Stouffer $Z = 0.70$, $p = .24$, one-tailed). Thus they failed to replicate Bem and Honorton's finding ($ES = 0.162$, Stouffer $Z = 2.52$, $p = 5.90 \times 10^{-3}$, one-tailed). The authors [Storm & Ertel] conducted stepwise performance comparisons between all available databases of ganzfeld research. Larger aggregates of such studies were formed, including a database comprising 79 ganzfeld/autoganzfeld studies (*ES* =

This article is part of an exchange, the previous components of which were published in *Psychological Bulletin*.

> 0.138, Stouffer $Z = 5.66$, $p = 7.78 \times 10^{-9}$). Thus Bem and Honorton's positive conclusion was confirmed . . . . The ganzfeld appears to be a replicable technique for producing psi effects in the laboratory. (p. 424)

By way of a reply to our (Storm & Ertel, 2001) article, Milton and Wiseman (2001) criticized some of our assumptions and procedures. Milton and Wiseman's (2001) response was as follows (taken from their abstract):

> [Storm and Ertel] ignored the well-documented and widely recognised methodological problems in the early studies, which make it impossible to interpret the results as evidence of extrasensory perception. In addition, Storm and Ertel's meta-analysis is not an accurate quantitative summary of ganzfeld research because of methodological problems such as their use of an inconsistent method for calculating study outcomes and inconsistent inclusion criterion. (p. 434)

We address Milton and Wiseman's (2001) criticisms below in the order that they appeared in their reply, which is not always in the order of importance.

Are Quotes Taken Out of Context Admissible As Evidence?

We criticize Milton and Wiseman (2001) for selectively picking quotes that feature "psi-questioning" content while ignoring "psi-supporting" accounts. That is, they started off by (a) strategically quoting, *at length*, skeptic Ray Hyman (see Milton & Wiseman, 2001, p. 434), (b) failing to represent Honorton's (1985) and Hyman and Honorton's (1986) positive views, and (c) using a misleading passage from Hyman and Honorton (see Milton & Wiseman, 2001, p. 435). Milton and Wiseman thus quoted Hyman and Honorton (1986) out of context (Milton & Wiseman, 2001, p. 435) and misrepresented Hyman and Honorton's (1986) joint overall conclusion, which was: "we agree that the overall significance observed in these studies cannot reasonably be explained by these selective factors [i.e., "multiple testing, retrospective experiments, . . . the file-drawer problem," etc.]" (p. 352). Two years later, after further testing, Harris and Rosenthal (1988b) reiterated this conclusion: "Our analysis of the effects of flaws on study outcome lends no support to the hypothesis that ganzfeld research results are a significant function of the set of flaw variables" (p. 3).

However, Milton and Wiseman (2001, p. 435) placed greater credence on the following statement from Hyman and Honorton (1986): "the final verdict awaits the outcome of future experiments—ones conducted by a broader range of investigators and according to more stringent standards" (p. 353). The fact is that Hyman and Honorton also "agreed that the significant outcomes have been produced by a number of different investigators" (p. 352) and that the argument over "stringent standards" was largely rhetorical (p. 353). (Note that we do not object to Milton and Wiseman's appeal to desirable future research for further evidence, but they make their point as if past research had been inconclusive.)

Thus, Milton and Wiseman (2001) clouded the waters and misled the unsuspecting reader into thinking that the statistically significant result of Honorton's (1985) database *taken at face value* contributed little, if anything, to the evidence for psi because the methodological issues (see Milton & Wiseman, 2001, pp. 434–435) were of greater concern. Yet, it is common knowledge that significance testing, aside from assessments of effect size, is an indispensable way of finding out whether experimental effects should be regarded as existent. The undoubtedly justified demand for replication, within and between investigators, cannot replace the equally important demand for statistical confidence of independent studies. Even if psi would seem to entirely disappear, like an ice-age climate in earth history, previous significant observations—ice formations as in our analogy—would not become invalid.

Even Hyman and Honorton (1986), it seems, disregarded this logic to some extent when they made the distinction between significant effects, on the one hand, and evidence for psi (i.e., a communications anomaly), on the other. Yet at a very early stage, Rosenthal (1986) insisted that the accumulated evidence should not be neglected: "At any point in time some judgment can be made . . . . We feel it would be implausible to entertain the null given the combined *p* from these 28 studies" (p. 333). Paraphrasing Rosenthal, our judgment is that psi effects have been evidenced by significant results so that we may rightfully defend our (Storm & Ertel, 2001, p. 424) quotes taken from Hyman and Honorton (1986).

Now, it seems, Hyman's personal communication to Milton and Wiseman (as of September 28, 2000, cited in Milton & Wiseman, 2001, p. 435) finds us guilty of a faulty interpretation of his original intent when he said that: "the ganzfeld data base had too many problems to be considered as evidence for the existence of psi." But this is Hyman's personal interpretation, and he cannot speak on Honorton's behalf. Hyman and Honorton (1986) actually disagreed over the "degree . . . [of] evidence for psi" (p. 352), and the two authors differed "about the extent and seriousness of [the] departures" from "ideal standards" for the ganzfeld (p. 352). Apparently, two stories are being told in the "Joint Communiqué." As stated above, Hyman and Honorton (1986) agreed that "significant outcomes have been produced by a number of different investigators" (p. 352). But then they seriously weakened this conclusion by saying: "If a variety of parapsychologists and other investigators continue to obtain significant results . . . a genuine communications anomaly will have been demonstrated" (p. 354). This reluctance (indicated by a conditional "if") to accept existing significant replications as grounds to remove uncertainty might be due to Honorton's, the psi-proponent's, probable dilemma: He was forced to accept some of skeptic Hyman's overskeptical formulations to get the "joint" Communiqué completed.

Did Honorton's (1985) Ganzfeld Database Have Flaws?

Milton and Wiseman (2001) contended that we "denied that there were problems in the early ganzfeld studies" and that our denial was made "in the face of so much documented evidence to the contrary" (p. 435). Did we really deny problems? Specifically, did we deny the fact that earlier methods had less controls or that actual controls were made less explicit? In fact, we did refer to those problems by stating that "claims of alleged flaws in Honorton's, 1985, ganzfeld studies, and his meta-analyses, have not been successfully defended" (Storm & Ertel, 2001, p. 426), and, prior to that statement, we backed up our argument by saying: "Numerous claims that flaws in Honorton's (1985) meta-analysis still exist have been debunked" (Storm & Ertel, 2001, p. 424), and we gave references to that effect—specifically, Atkinson, Atkinson, Smith, and Bem (1990), Harris and Rosenthal (1988a, 1988b), Saunders, (1985), and Utts (1991).

In other words, we fully acknowledged those "problems" (hypothesized artifacts), we did not deny them, we merely said they were solved. Milton and Wiseman's opposition is not based on facts, but on mere doubts over clean methods, which were brought forward in three earlier papers: Hyman (1985), Honorton (1985), and Hyman & Honorton (1986). Milton and Wiseman (2001) turn guesswork into "evidence" ("so much documented evidence to the contrary," p. 435) thus making a mountain out of a molehill, which no longer exists, while actually ignoring our list of five articles (just mentioned above) dating from 1985 to 1991. These up-to-date articles disprove Milton and Wiseman's conjectures (Storm & Ertel, 2001, p. 424).

Is Quality Rating of Earlier Studies Always Necessary?

Milton and Wiseman (2001) argued that the 11 pre-Communiqué studies used in our meta-analysis should not have been used at all because of ostensible "methodological weaknesses" (p. 435). To defend this argument, they referred to the Communiqué, not as "a mere documentation of traditional and uncontroversial research rules," as they should, but in order to "justify downgrading the quality of all research published before 1986," as already noted in Storm and Ertel (2001, p. 425). The Communiqué was not written as an indictment of prior ganzfeld research, and it should never be used as such.

In our own meta-analysis of 11 "newly found" studies, we endeavored to maintain Honorton's (reappraised and approved) standard in our search. Thus, we regard Milton and Wiseman's (2001) denouncing our practice by attributing "no value in performing such a meta-analysis" (p. 435) as not based on any factual arguments.

Milton and Wiseman (2001, p. 436) then criticized us for restricting our quality ratings to those 11 studies, without rating "the other 68 studies" (i.e., Honorton's 28 studies, plus Milton and Wiseman's 30 studies, plus Honorton's 10 studies). We undertook a quality assessment of those 11 studies because they had not yet been subjected to the rigors of Hymanian techniques of invalidation (as was done to Honorton's database). We also pointed out that the method of quality rating was not entirely our own, as Milton and Wiseman (2001, p. 436) falsely assumed, but was modeled on Radin and Ferrari's (1991, pp. 65–66) procedure.

We did not conduct a similar quality rating on Honorton's database because that database of direct-hit studies had already won acceptance from discriminating parapsychologists—it was actually never invalidated, neither by Hyman's nor any other researcher's analysis (as listed above). It should also be clear that quality rating of the post-Communiqué studies would be redundant—Milton and Wiseman themselves regarded them as flawless by their own standards. As for the inclusion of direct-hit studies only, that criterion was already explained (see Storm & Ertel, 2001, p. 427).

Effect size is another clarifying issue. We looked at effect sizes of the two databases ("pre-Communiqué" and "post-Communiqué") in a number of different ways (Storm & Ertel, 2001, pp. 427–429) and found that they did not differ significantly. We conducted performance comparisons of (a) pre-Communiqué studies with post-Communiqué studies and (b) pre-Communiqué authors with post-Communiqué authors, both of which yielded no statistical evidence that the guidelines in the Communiqué had any "influence on effect size outcomes" (p. 430) or any influence on principal authors. There was no indication that the mean effect size of the pre-Communiqué database was "inflated" (i.e., an artifact of flaws) because it compares favorably with the allegedly "flawless" post-Communiqué studies. And there was no evidence that the mean effect size of the post-Communiqué database was "deflated" because of the removal of purported flaws.

Apropos to our findings, we refer to Palmer (1986), who warned that false conclusions can be drawn on account of, and by appeal to, the Communiqué's guidelines—it should not be assumed that "past successes were due to the presence of the flaws" (p. 379). Thus we provided new evidence supporting our position that earlier studies do not show any effects of hypothesized methodological shortcomings. Milton and Wiseman ignored this evidence altogether.

Incidentally, an apparently perplexing contradiction escaped the notice of our two critics in their demand for more extensive quality ratings. They claimed there are "obstacles to using quality scales to detect and correct for methodological problems in studies" (p. 437) and referred extensively to some (irrelevant) medical study. How can Milton and Wiseman claim that pre-Communiqué studies had methodological problems (and must therefore be ignored) while claiming, a little while later, that the methods used to find that out are doubtful in the first place? For Milton and Wiseman, the null hypothesis of psi has never been and, apparently, will never be rejected—because there are "problems"—even if the claim of existing problems turns out to be itself unverifiable, in their view, because of further "problems."

Furthermore, Milton and Wiseman (2001) generally ignored important statements about quality assessment in other meta-analyses (except their own). For example, Lawrence (1993) said: "Neither the quality of studies nor their effect sizes, has significantly changed over the years" (p. 75). Milton ignored even her own coauthored finding of a meta-analytic quality assessment of forced-choice psi experiments: "There were no statistically significant correlations between the presence of procedural safeguards and effect size and hence no suggestion that methodological problems had played any strong and obvious role in the overall effects . . . although the small database would be expected to provide relatively low statistical power for detecting any such effects" (Steinkamp, Milton, & Morris, 1998, p. 193). (The sample consisted of 22 study pairs, i.e., of 44 studies!) Thus, if there really were such effects and if they were not revealed with 44 studies, the size of such effects must have been negligible.

Another indication of Milton and Wiseman's tendency to downplay psi comes in the form of their conclusion after conducting a meta-analysis of psi research via mass media channels (Milton & Wiseman, 1999b). The results did not support the psi hypothesis, so they deemed it possible that "ESP does not exist and that the mass-media studies accurately estimate its effect size as indistinguishable from zero. In this scenario, the positive results of the apparently successful meta-analyses would be due to methodological flaws" (Milton & Wiseman, 1999b, p. 237). Milton and Wiseman apparently did not consider the conflict of their "scenario" with the bulk of positive results accumulated over decades of parapsychological experimentation and, above all, the entire absence of empirical evidence for "methodological flaws." It is *lack* of such empirical evidence that has been accumulating. Milton and Wiseman took the liberty to ignore it.

Milton even downplayed "a highly significant cumulative effect (Stouffer $Z = 5.72$)" and an appreciable mean *ES* of 0.16 in her meta-analysis of free-response ESP studies (Milton, 1997, p. 279) by pointing out possibilities of artifacts (data analyses were possibly not preplanned, or authors failed to report whether their analyses were preplanned). However, she did not provide any empirical indication that studies without such reports gave rise to the suspicion that the authors of her sample were methodologically less sophisticated than her. Note also that Milton and Wiseman did not consider at all the possible lack of psi-conducive conditions ("flaws") in mass media studies. These were merely characterized as having very "different" conditions; they do not characterize them as "probably unfavorable" conditions.

Does a Revised Effect Size Calculation Make Any Difference?

Milton and Wiseman (2001, p. 436) took issue with the lack of *conservative* calculations of some *z* scores for studies in the 11-study database. In fact, only 3 of the 11 studies needed adjustment, thus reducing the quality-weighted mean *z* score from .32 ($ES = .14$; Stouffer $Z = 1.06$, $p = .144$) to .26 ($ES = .13$; Stouffer $Z = 0.87$, $p = .192$). The 11-study database is still not significantly different from Honorton's (1985) 28-study database, $t(37) = .61$, $p = .543$, two-tailed. Thus, the *old ganzfeld database* can still be formed. It has a mean *z* of .97 ($ES = .225$; Stouffer $Z = 6.05$, $p = 7.24 \times 10^{-10}$), results of which are comparable with Storm and Ertel's (2001, p. 429) original data for the old ganzfeld database: mean *z* of .99 ($ES = .227$; Stouffer $Z = 6.15$, $p = 3.93 \times 10^{-10}$).

The "old" and the "new" ganzfeld databases are significantly different, $t(77) = 3.04$, $p = .003$, $\omega^2 = .09$, but the omega-squared value (9%) is now exactly that of the critical value stipulated in Storm and Ertel's (2001) paper. But Cohen's (1988) test, as originally applied by Storm and Ertel, shows that the difference is again not significant. When the two databases are combined, the 79-study database has a mean *z* score only slightly reduced from .64 to .63 ($ES = .14$; Stouffer $Z = 5.59$, $p = 1.14 \times 10^{-8}$). This "revised" larger database, representing once again a unified ganzfeld domain, might indicate that over two decades of ganzfeld/autoganzfeld work, again dismissed by Milton and Wiseman (2001) almost out of hand, has in fact not been in vain.

Is Testing for Bidirectional Psi Not Legitimate Procedure?

We performed tests for bidirectional psi (Storm & Ertel, 2001, p. 429), and results were positive throughout. Bidirectional effects appeared in all four databases indicating that, had Milton and Wiseman proposed a bidirectional hypothesis, their results would have supported their replication trial even under their own unfavorable condition of limited data selection.

Milton and Wiseman's (2001, p. 436) reasons for disregarding bidirectionality are unacceptable. They referred to the fact that this form of analysis played no role in Hyman and Honorton's (1986) study or previous studies, and that "interest in testing for extreme dispersion" (p. 436) is a phenomenon that appeared only after their (Milton & Wiseman, 1999a) initial study. As it happens, bidirectionality has been regarded as a unique psi feature for five decades (Rao, 1965; Rhine, 1952). Thus we did not suggest a new approach to psi testing. Once attention is drawn to some relevant testing procedure, even if forgotten by most researchers, it must still be regarded as legitimate at any time.

In a final effort to downplay our finding of a significant bidirectional effect, Milton and Wiseman (2001) regarded the probability level of $p =$

.027 for that effect in their 30-study database as "marginal" (p. 436)—they considered that it did not "carry much weight" (p. 436) because it was a "post hoc analysis" (p. 436). Need we remind them that, by convention, chance explanations are rejected when p is less than or equal to .05 and that researchers are bound in that case to find explanations?

We regard Milton and Wiseman's wanton dismissal of the 5% rule, by labeling our decision "post hoc," as theoretically and statistically groundless. Not only were many of their own decisions made post hoc, but ignoring the 5% rule might be an act expected of skeptics in their burgeoning need for an ever-more "creative" and conditional interpretation of significant results whenever the need arises to undermine the evidence of an anomalous effect.

How Much Evidence Is Needed to Convince Neutral Scientists?

Milton and Wiseman (2001) reported the statistic that "only half" of a limited number of respondents (members of the ganzfeld research community) to an electronic mail forum (i.e., about 10 people!) "thought that the experimental evidence for psi was currently strong enough to convince a neutral scientist" (p. 437). Need we say that the bottle is half full? The fact that half of it is still empty appears reasonable after considering that respondents were not asked to indicate their own conclusions but were asked to guess conclusions by some "neutral," that is, skeptical, but unprejudiced researcher, implying that the neutral researcher's knowledge of the field was not broad-based. Those 10 respondents contributing to the "half-empty" kind of responses might have considered typical obstacles by "neutral" observers from a distance, while being convinced themselves—just as the other 10 respondents were—that the evidence was sufficient. Schmeidler and Edge (1999), whose article contained strong affirmative arguments, might nevertheless have replied to this questionnaire item that they doubt that "neutral" observers would regard the existing bulk of evidence as sufficient.

Milton and Wiseman (2001, p. 437) also informed us that only 17% of all respondents (fewer than 4 people!) "thought that the procedures necessary for producing a reasonably replicable ganzfeld psi effect had as yet been identified" (p. 437). But this common deficit of parapsychological research in general, which is due to our ignorance in this field of necessary causal agents cannot be reason to doubt psi phenomena.

Recommended Strategy for Looking Into Determinants of the Ganzfeld Effect

Milton and Wiseman's recommendation for future research includes the establishment of "a possible strategy for *attempting replication*" of ganzfeld effects (italics added; Milton & Wiseman, 2001, p. 437). In our view, the real focus should be on process-oriented work (cf. E. May's personal communication, March 8, 2001, in which he stated: "Further 'proof' oriented work is a waste of resources" given the weight of evidence for psi).

A movement toward identifying psi-conducive conditions must now be seen as more important than the ongoing debate over replication, which is often protracted by authors who, despite being well informed about the wealth of experimental findings in parapsychology, take pleasure in eroding away the psi findings by dubious means, apparently not unaware of immediate reinforcement from mainstream circles. We need to come closer to discovering elements of the true nature of ganzfeld phenomena, and psi in general. That shift in focus would be a movement away from unproductive styles of 20th-century research toward new horizons, it is hoped, of understanding the reality of the paranormal.

References

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

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

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

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

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

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

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

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

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

MILTON, J. (1997). Meta-analysis of free-response ESP studies without altered states of consciousness. *Journal of Parapsychology*, **61**, 279–320.

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

Milton, J., & Wiseman, R. (1999b). A meta-analysis of mass-media tests of extrasensory perception. *British Journal of Psychology,* **90**, 235–240.

Milton, J., & Wiseman, R. (2001). Does psi exist? Reply to Storm and Ertel (2001). *Psychological Bulletin,* **127**, 434–438.

Palmer, J. (1986). Comments on the "Joint Communiqué." *Journal of Parapsychology,* **50**, 377–381.

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

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

Rhine, J. B. (1952). The problem of psi-missing. *Journal of Parapsychology,* **16**, 90–129.

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

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

Schmeidler, G., & Edge, H. (1999). Should ganzfeld research continue to be crucial in the search for a replicable psi effect?: Part II. Edited ganzfeld debate. *Journal of Parapsychology,* **63**, 335–388.

Steinkamp, F., Milton, J., & Morris, R. L. (1998). A meta-analysis of forced-choice experiments comparing clairvoyance and precognition. *Journal of Parapsychology,* **62**, 193–218.

Storm, L., & Ertel, S. (2001). Does psi exist? Comments on Milton and Wiseman's (1999) meta-analysis of ganzfeld research. *Psychological Bulletin,* **127**, 424–433.

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

Anomalistic Psychology Research Unit
Department of Psychology
Adelaide University
South Australia 5005
Australia
lance.storm@psychology.adelaide.edu.au

Georg-Elias-Müller Institut für Psych.
Georg-August Universität
Gossler Strasse 14
D-37073 Göttingen
Germany
sertel@uni-goettingen.de