Psychological Bulletin
2001, Vol. 127, No. 3, 434–438

0033-2909/01/$5.00 DOI: 10.1037//0033-2909.127.3.434

Does Psi Exist? Reply to Storm and Ertel (2001)

Julie Milton
University College London

Richard Wiseman
University of Hertfordshire

The authors recently published a nonsignificant meta-analysis of 30 extrasensory perception ganzfeld studies, all conducted after the 1986 publication of important methodological guidelines aimed at reducing sources of artifact noted in earlier studies. In response, L. Storm and S. Ertel (2001) presented a meta-analysis of 79 studies published between 1974 and 1996. They argued that the positive and highly statistically significant overall outcome indicates a replicable paranormal effect. In doing so, they ignored the well-documented and widely recognized 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 criteria.

Storm and Ertel (2001) presented a meta-analysis of ganzfeld studies, which, they argued, provides strong evidence of a replicable and genuine communication anomaly. We disagree both with their conclusion and their view of the role of methodological quality in attempts to resolve whether extrasensory perception exists. In this article, we discuss the problems in the early ganzfeld studies that make it difficult to draw strong conclusions from meta-analyses such as Storm and Ertel's that include them, additional problems in Storm and Ertel's meta-analytic approach, the general difficulties in using meta-analysis to identify and correct for methodological problems in constituent studies, and a possible approach to the replication problem within parapsychology.

The Role of Study Quality in Ganzfeld Research

When experiments test for effects that are expected to be relatively small if they exist at all, stringent methodology is crucial. Otherwise, even small effects due to methodological weaknesses could plausibly account for any positive results. Particularly in controversial areas, however, high apparent study quality is not enough. An effect must be replicable by a broad range of investigators. This makes it much less likely that any positive results were due to error, design problems not apparent in the published report, or even fraud by participants or experimenters.

As we pointed out in our earlier article in *Psychological Bulletin* (Milton & Wiseman, 1999), the importance of study quality and interexperimenter replicability has long been recognized within parapsychology. These issues have had special prominence within ganzfeld research over the past 20 years. In 1981, Hyman, a psychologist skeptical of the existence of psi,[1] chose ganzfeld research as the domain representing parapsychology's strongest claim to produce results replicable by many experimenters under rigorous conditions (Hyman, 1985). He meta-analyzed and systematically assessed the quality of the 42 ganzfeld ESP studies known to have been conducted since the first was published in 1974. However, he concluded as follows:

> By now it is clear that I believe that the ganzfeld psi data base, despite initial impressions, is inadequate either to support the contention of a repeatable study or to demonstrate the reality of psi. . . . parapsychologists may be doing themselves and their cause a disservice by attempting to use these studies as examples of the current state of their field. . . . the present data base cannot by any stretch of the imagination be characterized as flawless. (p. 38)

In response, Honorton (1985), a parapsychologist who had carried out many ganzfeld studies, carried out his own examination of the experiments. He conceded that the studies contained methodological and reporting problems. Many studies had used several outcome measures, creating a possible problem of multiple analysis. To attempt to reduce the impact of this problem, Honorton restricted himself to examining the 28 studies that had reported the outcome measure of *direct hits* (the number of correct forced-choice guesses obtained in each study). This subset showed similar problems to those in the larger database examined by Hyman (1985). For example, Honorton (1985, p. 71) reported that only 36% of the studies used duplicate sets of pictures to avoid the target stimulus picture showing signs of having been handled that might have given away its identity, and 25% used an informal and possibly biased method such as shuffling to randomly select the target picture.

Julie Milton, Department of Paediatric Epidemiology and Biostatistics, Institute of Child Health, University College London, London, England; Richard Wiseman, Department of Psychology, University of Hertfordshire, Hatfield, England.

Correspondence concerning this article should be addressed to Julie Milton, Department of Paediatric Epidemiology and Biostatistics, Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, England. Electronic mail may be sent to j.milton@ich.ucl.ac.uk.

[1] *Psi* is a term used in parapsychology to denote anomalous processes of information transfer such as extrasensory perception that are not currently explicable in terms of known physical or biological mechanisms.

Hyman (1985) reported statistically significant correlations between the lack of certain procedural safeguards and effect sizes in the studies. Honorton (1985) contested these analyses and reported analyses based on his own quality assessment of the studies, which did not show such significant relationships. Rather than continue to focus their debate on these early studies, Hyman and Honorton (1986) noted that they did not agree on the extent to which such analyses within meta-analysis could identify and compensate for methodological weaknesses of individual studies. They decided that their most productive next step would be to describe guidelines for the conduct of future ganzfeld studies. Describing the 1974–1981 ganzfeld studies, they wrote as follows:

> We agree, as our earlier exchanges indicate . . . that the experiments as a group departed from ideal standards on aspects such as multiple testing, randomization of targets, controlling for sensory leakage, application of statistical tests, and documentation. . . . we agree that the final verdict [concerning the evidence for psi] awaits the outcome of future experiments conducted by a broader range of investigators and according to more stringent standards. (pp. 352–353)

Honorton himself conducted a series of 11 automated ganzfeld ("autoganzfeld") studies with a methodology designed to take Hyman's concerns into account (Bem & Honorton, 1994). The positive and statistically significant overall results were published in *Psychological Bulletin*. In that article, the authors cautiously reminded readers that "the autoganzfeld studies by themselves cannot satisfy the requirement that replications be conducted by a 'broader range of investigators' " (Bem & Honorton, 1994, p. 13). Close scrutiny of the methodology of Honorton's autoganzfeld studies since their publication by a number of researchers did not reveal any methodological weaknesses that could account in any obvious way for their positive results,[2] and so it appeared worthwhile to examine whether other apparently well-conducted studies replicated their findings. In 1997, we therefore meta-analyzed all of the ganzfeld studies conducted since the guidelines' publication to determine whether this requirement had been met. The cumulation of the 30 study outcomes was not statistically significant overall and had a near-zero mean effect size.[3] We concluded that the ganzfeld technique does not at present offer a replicable method for producing psi in the laboratory.

Storm and Ertel's (2001) Meta-Analysis

In their article, Storm and Ertel (2001) responded to our meta-analysis by denying that there were problems in the early ganzfeld studies examined by Hyman (1985) and Honorton (1985) and by using a set of rules to allow them to combine some of these early studies with later studies to form a large database of 79 studies. This large database has a highly statistically significant cumulation, but this is not surprising. Only 11 studies from a brief interim period in the mid-1980s not meta-analyzed by ourselves or Honorton are added to databases whose cumulated probabilities are published and that would clearly be highly statistically significant if combined. Storm and Ertel argued that this overall cumulation should be considered to be strong evidence for a genuine and replicable communication anomaly.

Because Storm and Ertel (2001) included in their meta-analysis so many studies known to have had methodological weaknesses, however, it is impossible to know what proportion of the outcome simply reflects methodological artifact. We can see no value in performing such a meta-analysis, nor can we account for Storm and Ertel's denial of the importance of quality problems in the early studies in the face of so much documented evidence to the contrary.

For example, Storm and Ertel (2001) stated that the studies in this early ganzfeld database were not "lacking in quality" (p. 424). Both Hyman (1985) and Honorton (1985) extensively documented the methodological problems in the database as is clear from the material that we have cited above. Storm and Ertel offered no argument or evidence to dispute their assessment. Storm and Ertel represented Hyman as agreeing with Honorton that the early studies were sufficiently well conducted to demonstrate that ESP existed but merely thought its effect size smaller. However, this was clearly not Hyman's view, as can be seen from the above excerpts from Hyman's published writings.[4] Storm and Ertel described Hyman and Honorton's (1986) guidelines for future research as "a mere documentation of traditional and uncontroversial research rules" (p. 425) that ganzfeld experimenters had been following all along. This is simply not the case as is clear from Hyman and Honorton's quality assessment of the early studies. The guidelines were launched in a special issue of parapsychology's leading journal devoted to the topic, with invited commentaries from prominent researchers in the field. Their publication was widely seen as a key moment in parapsychology—according

---

[2] Storm and Ertel (2001) inaccurately described us as referring to "Bem and Honorton's 'spurious' results" (p. 425) and as having "contrived possible deficiencies in the procedures applied in the studies" (p. 430). As is clear in our article, we described the possible sources of methodological artifact that a number of researchers have considered as possible sources of bias in Honorton's autoganzfeld studies. These researchers included Honorton himself, who described how the automated procedure introduced an unanticipated opportunity for auditory leakage (Honorton et al., 1990). However, as we stressed in our article, "none of the opportunities for sensory leakage appear sufficiently strong . . . to explain away the positive results of the autoganzfeld in any immediately compelling way" (Milton & Wiseman, 1999, p. 389). Nor is it the case that we "expressed doubts about the quality of Bem and Honorton's (1994) meta-analysis" (Storm & Ertel, p. 425) as is also clear from our article.

[3] Storm and Ertel (2001) stated that the effect size measure used in our article and theirs—$z/\sqrt{N}$, the standard normal deviate associated with the study's outcome divided by the square root of the number of trials in the study—is r, the correlation coefficient. Although this is true for the more common type of psychology study in which there is an experimental and control condition, it is not so in the special case when performance is being compared with a theoretical chance baseline as is the case in these ganzfeld psi studies. This can be demonstrated by considering examples in which $z/\sqrt{N}$ exceeds r's boundaries of -1 to $+1$. For example, if a participant correctly guesses 100 rolls of a 10-sided die, $z/\sqrt{N}$ is 2.98.

[4] In the joint communiqué, Hyman and Honorton (1986) wrote, "We continue to differ over the degree to which the effect constitutes evidence for psi" (p. 351). Storm and Ertel (2001) interpreted these words as "meaning that the unresolved issue between the two was over the actual size of the effect. In other words, there was agreement between the two that an effect existed" (p. 424). Hyman confirmed that Storm and Ertel's interpretation of this sentence is incorrect, stating, "My intention then, as it would be today, was to emphasize my belief that the ganzfeld data base had too many problems to be considered as evidence for the existence of psi" (Hyman, personal communication, September 28, 2000).

to one commentator, a "historical event" (Utts, 1986, p. 393). Experimenters cited the guidelines widely and made great efforts to ensure that their new studies complied with them. In denying that the early ganzfeld studies had methodological problems that were well documented and widely acknowledged, Storm and Ertel also deny the impressive willingness of parapsychologists in the mid-1980s to face up to a disappointing assessment of their own research and to embark on a field-wide effort to do something about it.

Problems in Storm and Ertel's (2001) Meta-Analytic Methodology

Quite apart from the issue of methodological problems in a large proportion of the studies in their database, there are problems in the methodology of Storm and Ertel's (2001) meta-analysis that prevent it from standing even as an accurate quantitative representation of ganzfeld research. We briefly deal with each in turn.

Inconsistent application of quality scale weights. Storm and Ertel (2001) reported a quality scale of their own devising that does not include important methodological safeguards, such as the use of duplicate judging sets to prevent handling cues, extensively discussed by Hyman and Honorton (Honorton, 1985; Hyman, 1985; Hyman & Honorton, 1986). They did not report clearly defined criteria for judging whether safeguards were present, and no interrater reliability data were reported. No summary of study quality was presented. Their only use of the quality scale was to weight the effect sizes of the 11 studies in the 1982–1986 database by quality but not the other 68 studies, including the heavily criticized 1974–1981 studies. When one is testing whether effect sizes replicate across studies or when one is attempting to obtain a quantitative summary of a group of studies, it makes no sense to distort those effect sizes by applying quality weights to them, nor can there be any justification for applying those weights to some studies and not others.

Inconsistent inclusion criteria. The standard method for unbiased retrieval of studies for a meta-analysis is to perform a systematic search of the literature and to apply consistent inclusion criteria to the results of such a search. Storm and Ertel (2001), however, performed a literature search for studies conducted or published during the mid-1980s and drew other studies from meta-analyses that have used different inclusion criteria. As a result, they excluded 14 studies from the 1974–1981 database because they did not report direct hits (Honorton, 1985) and an unknown number of such studies from the 1982–1986 database but included all 30 studies in our meta-analysis regardless of whether they reported direct hits.

Inconsistent method of outcome calculation. Storm and Ertel (2001) used the exact binomial method to calculate outcomes for some studies and the normal approximation to the binomial distribution for others.[5] They used the normal approximation even when the studies' authors reported sufficient data to calculate the exact binomial probability. In some cases the studies are small enough that approximation may be unwise.

Bidirectional Psi

An additional feature of Storm and Ertel's (2001) meta-analysis is the introduction of a test for bidirectional psi. They argued that if the study outcomes in our meta-analysis are analyzed for deviations both above and below mean chance expectation, the deviations are more extreme than would be expected by chance alone. Applying Timm's (1983) statistic, which tests for deviation from mean chance expectation regardless of direction, they obtained $p = .027$ for our database and argued that this "support[s] the psi hypothesis" (p. 429).

However, as Storm and Ertel (2001) themselves pointed out, none of the previous parapsychological meta-analyses have examined the data for evidence of bidirectional psi, although many parapsychologists have long believed psi to manifest as both above- and below-chance scoring (Wolman, 1977). Interest in testing for extreme dispersion from chance in meta-analyses has only appeared since the publication of our null cumulation of the recent ganzfeld studies. A number of commentators have noted that more of the 30 studies than would be expected by chance have statistically significantly below-chance outcomes (Edge & Schmeidler, 1999). Storm and Ertel's analysis is only one of a number of post hoc analyses that reflect this observation. However, we do not agree that a post hoc analysis with so marginal a probability value can carry much weight in this context. It should be noted that the statistically significant outcomes of Timm's (1983) test applied to their various other three ganzfeld subdatabases do not indicate replication of bidirectional outcomes in these databases. Timm's statistic is blind to direction rather than being a test for bidirectional effects as such. In these cases it merely appears to reflect these databases' statistically significant above-chance results.

The Role of Study Quality in Meta-Analysis

Although Storm and Ertel (2001) went to some lengths to contest our meta-analysis of ganzfeld studies, they stated that the failure of these new studies to replicate the results of earlier work is irrelevant, given what they believe to be overwhelming evidence for psi from other parapsychological research. They are not alone in this view. Since Hyman (1985) and Honorton (1985) used meta-analysis to attempt to resolve their debate over the ganzfeld studies, many other bodies of parapsychological research have been meta-analyzed (Honorton & Ferrari, 1989; Milton, 1997; Radin & Nelson, 1989). In general, they have shown highly statistically significant overall cumulations. As with the early ganzfeld studies, many of the meta-analyzed studies did not report a number of potentially important methodological safeguards. However, some of the authors of these meta-analyses have argued, as did Honorton (1985), that because study effect sizes mostly have not correlated with their methodological quality as measured by rating scales, study quality is not an issue and strong conclusions can be drawn even from databases of studies whose methods are not well documented. The popular book that Storm and Ertel cited in sole support of their claim that the evidence for psi is

[5] For example, a study by Braud, Ackles, and Kyles (1984; incorrectly reported by Storm and Ertel as being published in 1983 with Braude as first author) obtained 6 direct hits in 10 trials. Storm and Ertel's use of the normal approximation to the binomial to obtain the effect size yields a standard normal deviate of 2.19; use of the exact binomial test yields a more conservative 2.06.

already overwhelming (Radin, 1997)[6] shares this approach, and at least some parapsychologists appear to find it convincing (Edge & Schmeidler, 1999).

Outside parapsychology, however, it is well recognized that there are serious problems with this approach. The most important problem is that methodological quality scales are generally not evidence based. They typically consist of a list of desirable methodological safeguards such as prespecification of sample size, the use of certain randomization methods, and so on. A point is assigned for each safeguard reported in the study, and the points are summed to yield that study's quality score.

However, without any basis on which to weight the importance of the various safeguards, a study that is in fact more susceptible to bias than another may obtain a higher quality score. Juni, Witschi, Bloch, and Egger (1999) recently demonstrated the arbitrary nature of scales used to assess the quality of randomized controlled clinical trials. They meta-analyzed the results of 17 studies comparing the use of standard heparin with low-molecular-weight heparin (LMWH) for postoperative thrombosis, using 25 different quality scales proposed by different authors. For 12 of the scales, the two treatments showed no difference in effects, regardless of study quality as measured by the scales. For 6 of the scales, low-quality studies, but not high-quality studies, showed LMWH to be superior; for 7 of the scales, this result was reversed. Because of the obstacles to using quality scales to detect and correct for methodological problems in studies, we believe that the replicability of ganzfeld (or other) studies under stringent methodological conditions remains crucial. We cannot agree with Storm and Ertel (2001) that focusing attention on the replicability of psi studies under stringent conditions constitutes "unwarranted questioning of the existence of psi" (p. 425).

A Possible Strategy for Attempting Replication

In a recent electronic mail forum of 41 invited members of the ganzfeld research community (Edge & Schmeidler, 1999), approximately half of the participants responded to an exit questionnaire related to these issues. The results indicated that the evidence for psi in the ganzfeld is not yet regarded as overwhelming, even within the ganzfeld research community itself. Only half of the respondents thought that the experimental evidence for psi was currently strong enough to convince a neutral scientist. Only 17% thought that the procedures necessary for producing a reasonably replicable ganzfeld psi effect had as yet been identified.

Given that parapsychology, therefore, still appears to be faced with a replication problem, the question arises of what to do about it. Individual studies that are apparently highly successful do still appear. After our meta-analysis was submitted to *Psychological Bulletin*, Dalton (1997) published the results of a ganzfeld study with so highly statistically significant an outcome ($p = 7.2 \times 10^{-8}$) that it alone pulls the entire null database of post-1986 studies into overall statistical significance (Milton, 1999). However, unless researchers can identify the variables that produce such results, only a few experimenters will be able to conduct successful studies. The replication problem and the difficulty of convincing other scientists of the reality of psi will remain.

One of us (Milton, 1999) proposed a possible strategy for parapsychology to adopt in its attempts to produce a replicable effect under stringent methodological conditions. The first step would be to assess critically the experimental evidence for the effects of the many variables that have been suggested as affecting outcomes in ganzfeld studies. Storm and Ertel (2001) stated that there are "well-known psi-inhibiting effects that result from the introduction of stricter controls, automation, and so on" (p. 430) and that "a warm social ambience acts as a moderator variable in favor of a psi effect" (p. 426).[7] Although these are opinions often voiced within parapsychology, there is no direct empirical evidence to support them. The effects of many other variables that have been considered as possibly psi conducive have also not been formally tested and would have to be addressed in new studies.

If such variables can be identified, the second step in attempting to demonstrate a replicable effect would be possible. Researchers could incorporate in their new, stringently conducted studies those procedures that evidence shows are associated with success. The studies could be preregistered for inclusion in a meta-analysis with prespecified analyses and a prespecified end point, to preclude objections relating to post hoc data selection or optional stopping.

Such an approach would require considerable collaboration within the field, and it will be some time before it becomes apparent whether the will for such an endeavor exists. It may be that, in future, a collaborative and systematic approach to the replicability problem will be taken of the type suggested. In the meantime, we agree with Hyman and Honorton's (1986) position; the final verdict on psi depends on replication of an effect across experimenters under methodologically stringent conditions.

[6] For a detailed critique of the approach to the meta-analysis of parapsychological databases taken in this book, see Milton (1999).

[7] Storm and Ertel (2001) wrote, "Milton and Wiseman (1999, p. 390) noted that a warm social ambience should be created and maintained during the ganzfeld experiment" (p. 426), but this is not our opinion. We merely reported that "Ben and Honorton . . . stressed the importance of . . . a 'warm social ambience' " (p. 390). Similarly, Storm and Ertel wrote as follows: "Milton's opinion seems to waver (S. Ertel, personal communication, September 23, 1999). She noted that the similarity of mean effect sizes between two databases—Honorton's (1985) 28 studies and Bem and Honorton's (1994) 11 studies—'implies that the 'ganzfield effect' is fairly robust' " (p. 426). Again, this is not Milton's opinion but a description of an implication made by Honorton as can be seen from a fuller quote: "He [Honorton] points out . . . that the mean effect size of the PRL studies almost exactly replicates that of the 1974–1980 database . . . pointing out the similarity implies that the 'ganzfeld effect' is fairly robust" (J. Milton, personal communication, September 23, 1999).

References

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

Braud, L. W., Ackles, L., & Kyles, W. (1984). Free-response GESP performance during ganzfeld stimulation. In R. A. White & R. S. Broughton (Eds.), *Research in parapsychology 1983* (pp. 78–80). Metuchen, NJ: Scarecrow Press.

Dalton, K. (1997). Exploring the links: Creativity and psi in the ganzfeld. In *The Parapsychological Association 40th Annual Convention proceedings of presented papers* (pp. 119–134). Durham, NC: The Parapsychological Association.

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

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

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.

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.

Juni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association, 282,* 1054–1060.

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?: I. Discussion paper and introduction to an electronic-mail discussion. *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.

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

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

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

Timm, U. (1983). Statistische Selektionsfehler in der Parapsychologie und anderen empirischen Wissenschaften [Statistical selection errors in parapsychology and other empirical sciences]. *Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie, 25,* 195–230.

Utts, J. (1986). The ganzfeld debate: A statistician's perspective. *Journal of Parapsychology, 50,* 393–402.

Wolman, B. B. (1977). *Handbook of parapsychology.* New York: Van Nostrand Reinhold.

Received October 23, 2000
Accepted October 23, 2000 ■