Psychological Bulletin
2010, Vol. 136, No. 4, 486–490

0033-2909/10/$12.00 DOI: 10.1037/a0019676

Meta-Analysis That Conceals More Than It Reveals: Comment on Storm et al. (2010)

Ray Hyman
University of Oregon

Storm, Tressoldi, and Di Risio (2010) rely on meta-analyses to justify their claim that the evidence for psi is consistent and reliable. They manufacture apparent homogeneity and consistency by eliminating many outliers and combining databases whose combined effect sizes are not significantly different—even though these combined effect sizes consist of arbitrary and meaningless composites. At best, their study provides a recipe for conducting a replicable extrasensory perception experiment. This recipe includes following a design that employs the standard ganzfeld psi methodology and uses "selected" subjects. An experiment, having adequate power and that meets these criteria, has already been conducted and failed to produce evidence for psi. Parapsychology will achieve scientific acceptability only when it provides a positive theory with evidence based on independently replicable evidence. This is something it has yet to achieve after more than a century of trying.

Keywords: parapsychology, meta-analysis, ganzfeld, replicability

Regarding the meta-analysis of Storm, Tressoldi, and Di Risio (2010), I agree with the authors' warning that "some caution is warranted in the interpretation of these results" (p. 479). They follow with the statement that "the ganzfeld is one of the most consistent and reliable experimental paradigms in parapsychology" (p. 479). Perhaps the import of this latter statement should be placed in context. Some major parapsychologists argue that the evidence in their field is inconsistent, unreliable, contradictory, and elusive (Atmanspacher & Jahn, 2003; Bierman, 2001; Kennedy, 2001, 2003; Lucadou, 2001). If the ganzfeld is, indeed, "one of the most consistent and reliable experimental paradigms in parapsychology," it is unclear whether this should impress the scientific community.

The argument for the consistency and reliability of the ganzfeld experiment relies exclusively on meta-analysis. Meta-analysis is a reasonable way to search for patterns in previously published research. It has serious limitations, however, as a method for confirming hypotheses and for establishing the replicability of experiments. This is especially true for how Storm et al. (2010) have used it.

At best, their findings suggest a recipe that might enable parapsychologists to produce that hitherto elusive replicable experiment. This recipe consists of the following steps: (a) use the ganzfeld procedure, (b) make sure that the design and method closely follow the standard ganzfeld experiment, and (c) use "selected" subjects (ones who have been in previous parapsychological experiments and/or believe in psi and/or have had a psychic experience and/or practice one or more mental disciplines). Presumably, the experimenter should also include a sufficient number of trials to guarantee an adequate level of power.

Ray Hyman, Department of Psychology, University of Oregon.
Correspondence concerning this article should be addressed to Ray Hyman, 5205 Nectar Way, Eugene, OR 97405. E-mail: rayhyman@comcast.net

It is this kind of prospective evidence for replicability, not the retrospective kind that emerges from meta-analyses of previous experiments, that is required for scientific acceptability. Indeed, parapsychology's Achilles' heel is its persistent inability to come up with even one example of prospective replicability. Unfortunately for Storm et al. (2010), the outlook for their current solution to this problem is not promising. The reason I say this is that an experiment that meets the authors' requirements has already been conducted and failed.

Autoganzfeld II

The series of experiments known as the "autoganzfeld" database were conducted in a single laboratory beginning in 1983 and terminating in 1989. These included eight ganzfeld experiments (three pilot and five formal) that were praised by parapsychologists and others for having overcome the methodological weaknesses of the original ganzfeld database and for their use of state-of-the-art procedures (Bem & Honorton, 1994). The pooled findings based on 330 sessions yielded a hit rate of 32% (as compared with the chance level of 25%), which was statistically significant. Because the combined hit rate for Honorton's (1985) 28 studies from the original ganzfeld database was 35%, the similarities of these two hit rates encouraged parapsychologists to hail the autoganzfeld experiments as a successful replication of the original ganzfeld series.

I have more to say about these experiments later. Here I want to alert the reader to the fact that the autoganzfeld experiments used two types of targets: static and dynamic. The static targets were still pictures similar to those used in the original ganzfeld database. The dynamic targets were brief video clips with a sound track. The trials using static targets produced a hit rate consistent with chance (approximately 26%). Indeed, the hit rate for the static targets was significantly lower than the hit rate for the original ganzfeld database in which only static targets were used. The significant hit

rate obtained in the autoganzfeld experiments was due entirely to the hit rate of 37% for the dynamic targets.

In 1993 Broughton and Alexander (1997) initiated "an attempted replication" of the autoganzfeld research. During a period of 2.5 years, they completed 209 trials, 151 of which they considered to be the "formal" component of their replication attempt. Autoganzfeld II, as the authors designated their replication attempt, used the same design, software, and equipment that was used in Autoganzfeld I. Not only was it a direct attempt to replicate the allegedly successful original series, but it remarkably fulfills the very recipe that Storm et al. (2010) have proposed.

Autoganzfeld II meets the criteria that Storm et al. (2010) imply will achieve the desired outcome. It is a ganzfeld experiment that is clearly in line with the standard ganzfeld procedure. Although the participants did not have prior experience in parapsychology experiments, 91% reported having had psychic experiences, and 70% practiced a mental discipline. Just as important, this replication attempt had adequate power. If one focuses on the dynamic targets, which had a hit rate of 37%, and accounted for all the significant hitting in the Autoganzfeld I, the power was 94% for the formal sample of 151 trials and over 98% for the total sample of 209 trials.

For the formal trials, the hit rate was 26.5%, and for the total of 209 trials, the hit rate was 25.8% (chance = 25%). The authors correctly conclude that this attempted replication failed. Even the secondary analyses failed to support various other findings—personality correlates, role of selected participants, etc.—that had been reported for the Autoganzfeld I experiments. At the same time Broughton and Alexander were running the Autoganzfeld II trials, Bierman (2001) was conducting a series of ganzfeld experiments that were also aimed at replicating Autoganzfeld I. These experiments also failed to replicate the original autoganzfeld results.

Persistent Inconsistencies

Storm et al. (2010) emphasize the consistency of the ganzfeld databases. This position contrasts with that of some of their parapsychological colleagues, who are troubled by the pervasive inconsistencies in parapsychological data including the ganzfeld studies (Atmanspacher & Jahn, 2003; Bierman, 2001; Kennedy, 2001, 2003; Lucadou, 2001). These inconsistencies go beyond the decline effect over time. Kennedy (2003) supplied a comprehensive overview of the many ways in which the evidence for psi displays its frustratingly "capricious" nature. He covered such categories as psi missing and negative reliability, the shift from intended effects to unintended secondary effects, erosion of evidence and decline effects over time, the inverse correlation of effect size with improved methodology, and the lack of practical applications of psi.

The reliance of Storm et al. (2010) on meta-analysis masks rather than uncovers the actual situation. The consistency they find and report is a manufactured one. They create "homogeneous" databases by removing outliers. This practice makes the remaining effect sizes less variable but does not change the fact that the original populations of experiments are heterogeneous. They compound this problem by justifying the combination of databases whenever a statistical test fails to reject the null hypothesis of no difference among the effect sizes. Every introductory course in statistics emphasizes that failure to reject the null hypothesis is not the same as proving the null hypothesis.

Some of the joining of databases borders on incoherence. After making a point that the Milton–Wiseman database is an outlier with respect to other ganzfeld databases, Storm et al. (2010) do not hesitate to combine that database with their own. They fit a quadratic curve to the plot of the effect sizes for ganzfeld experiments over time to support the claim of a "rebound effect." A rebound effect implies that the effect sizes of the most recent studies are significantly higher than those of the immediately preceding studies. Again, this does not stop them from combining all these studies into one homogeneous database. They then compound this mistake by using the exact binomial to calculate a combined effect size and test it for significance.

One should remember that an effect size is simply a standardized discrepancy between an observed outcome and an outcome expected by chance. The combining of effect sizes from different studies makes sense only if one can show that the separate effect sizes are conceptually coherent—that they all can be attributed to the same underlying cause. For this purpose, the investigator must have a theory and a solid empirical rationale for assuming that the combined effect sizes truly belong to the same category. Even parapsychologists admit that they have no positive theory for deciding which departures from chance, if any, reflect the presence of a single entity called psi.

One way to judge whether a database is internally consistent is to see whether the effect sizes are homogeneous or heterogeneous. However, this makes sense only if it has already been determined that the separate effect sizes are qualitatively coherent, not just quantitatively so. It takes time and effortful scrutiny of the individual studies to make such a judgment. I have done this for both the original ganzfeld database and for the original autoganzfeld database (Hyman, 1985, 1994). For these databases, I can confidently state that they clearly do not form coherent collections. For the remaining and more recent studies, there is good reason to suspect that they differ qualitatively from the earlier databases. I provide some examples below.

Heterogeneity of the Original Ganzfeld Database

The original ganzfeld database was contained in a set of documents that Honorton supplied to me in 1982. By his count, this database consisted of 42 ganzfeld experiments described in 34 reports. As far as he could tell, these reports—some published and some unpublished—consisted of all the ganzfeld experiments that had been done at the time I had requested his help in locating them. I had accepted an assignment to do a critical evaluation of parapsychological research. However, the accumulation of experiments on psi by that time was far too large to be scrutinized in a reasonable amount of time. I decided to restrict my evaluation to the complete set of experiments that represented the most promising line of parapsychological research at that time. Honorton and the other parapsychologists agreed that this would be the ganzfeld.

In my critique of the ganzfeld experiments, I evaluated the quality of all 42 experiments. However, I conducted a meta-analysis using 36 of the 42 experiments (Hyman, 1985). Six studies were excluded because they used ratings rather than the number of hits as outcome measures. If one includes these six by measuring effect size based on a z score divided by the square root

of trials in the manner of Storm et al. (2010), the overall effect size for the database becomes notably smaller. In his rebuttal to my critique, Honorton (1985) used only those studies that used direct hits based on four targets. His reduced database of 28 experiments has subsequently been identified with the "original" ganzfeld database. The inconsistency in the database that I discuss here exists regardless of which of the original databases one uses—Honorton's set of 28, Hyman's meta-analytical sample of 36, or the full set of 42.

The surprising number of problems and flaws in the original database are well documented in my critique (Hyman, 1985). Here I discuss just the matter of experimenter effects. I reported an analysis of variance that obtained a significant outcome for effect size due to individual experimenters. The pooled effect size across all the investigators translated into a direct hit rate of 35% (chance level = 25%). Almost all the above-chance hitting came from four experimenters who contributed half the studies to this database. The composite effect size for these four investigators translated into a direct hit rate of 44%. The contributions from the remaining experimenters, which composed the remaining 50% of the trials in the database, yielded a composite hit rate of 26%.

Clearly, this database consists of two populations: one from four experimenters who consistently contributed experiments with above-chance results and the other from several other experimenters who consistently obtained results consistent with chance. The composite hit rate for the entire database is an arbitrary mixture of at least two sources. Its average size is arbitrary. If a larger proportion of the studies had been contributed by the "successful" experimenters, the average of the composite could have been much higher. It could have been much lower if the "unsuccessful" experimenters had provided a larger proportion.

Autoganzfeld as a Failed Replication of the Original Ganzfeld Database

I have already mentioned that the combined hit rate for the original autoganzfeld experiments consisted of the nonsignificant hit rate for static targets and the significant hit rate for dynamic targets. Surprisingly, parapsychologists still treat the autoganzfeld experiments as a successful replication of the original ganzfeld experiments. In fact, the trials in the autoganzfeld series with static targets, the type of still images used in the original database, were not only consistent with chance but also significantly different from the hit rate in the original series. This situation represents a failure to replicate the findings in the original database. The parapsychologists, with their trust in meta-analysis, apparently were persuaded that the hit rate of 35% in the original series was sufficiently similar to the hit rate of 32% in the autoganzfeld to justify declaring the outcomes equivalent. These procedures demonstrate that the blind use of meta-analysis can convince some parapsychologists that two arbitrary and conceptually meaningless composites reflect the same underlying reality just because they have approximately the same quantitative size.

Autoganzfeld II as a Failed Replication of Both the Original and the Autoganzfeld Series

I used the Autoganzfeld II experiment as an example of one that, according to Storm et al. (2010), meets the criteria for a reliable, independently replicable study. Despite having adequate power, the results not only failed to replicate the key findings of the original autoganzfeld series but also were inconsistent with the original ganzfeld database.

Recent Ganzfeld Databases Inconsistent With the Preceding Databases

One peculiarity that was consistent across the early ganzfeld databases was the finding of a negative correlation between the z score and the square root of the number of trials across experiments (Hyman, 1985; Kennedy, 2001, 2003). This is a peculiar finding because the statistical theory underlying hypothesis testing depends upon the assumption that the z scores are positively correlated with the square root of the number of trials (which has interesting implications for issues of power that I do not discuss here).

As Storm et al. (2010) reported in an earlier version of their article, the correlation between the z scores and the number of trials in their most recent database is positive. This suggests another way in which the most recent ganzfeld databases may qualitatively differ from the earlier ones. I do not have sufficient space to describe other reasons for questioning the authors' combining of databases and conducting tests of significance on composite effect sizes that are of dubious provenance. However, the examples I have provided should suffice to make the point.

Use of Meta-Analysis to Support Replicability of Psi Evidence Is Fallacious

Storm et al. (2010) conclude their article as follows: "We emphasize how important it is to free up this line of investigation from unwarranted skepticism and hasty judgments, so that these communication anomalies might be treated and investigated in like manner with other psychological functions" (p. 480). This statement begs the question. It assumes that "communication anomalies" have been demonstrated. The statement also implies that parapsychological claims are dismissed for reasons other than the adequacy of their evidence.

Since the beginnings of modern science, scientists have been confronted with many claims of anomalies. Typically, these claimed anomalies were presented as clearly defined discrepancies from theoretical baselines within a given discipline. For each such claim, the scientific community reacted with caution. The first inclination was to look for flaws in the evidence and the arguments used to support the discrepancy. If the claim passed this test, the next step was to demand independently replicable evidence to justify the claim (as well as some coherent explanatory support).

For some claims, such as those for meteorites, general relativity, quantum mechanics, and the planet Neptune, the claims passed these tests, and the scientific theories were adjusted or revised to accommodate the claimed anomaly. For many more claims, such as those for N-rays, polywater, mitogenetic radiation, and Martian canals, the evidence could not be independently replicated, and the claimed anomalies were rejected. This conservative approach to claims of an anomaly has served the scientific community well. Philosophers and historians of science, among others, often credit this conservative approach for the huge success of modern science (Hyman, 1964).

If the scientific community has a bias against parapsychology's claim of a communications anomaly, this is as it should be. Parapsychology, during its more than a century of existence, shares many similarities with the many other failed claims for a scientific anomaly. A key similarity is its failure to provide independent, replicable evidence. Another similarity is its defense that psi is an inherently elusive phenomenon and that only certain conditions and individuals have the ability to produce or observe the evidence.

A puzzling difference between parapsychological claims and other failed claims of anomaly is that despite over a century of failing to come up with even one replicable experiment, parapsychological claims are still with us, whereas the others occupy science's discard pile. One reason for this state of affairs is that the other failed claims all originated within a given scientific program and the implications of the claimed anomaly were clearly apparent within that program. Parapsychological claims, on the other hand, originate outside existing scientific programs. They do not arise as a clearly delineated discrepancy from a specific hypothesis within a given scientific domain. Rather, the claim for a communications anomaly is based on an amorphous discrepancy from a generic statistical baseline. As such, its specific implications, if true, for a given scientific discipline are unclear.

In relying on their meta-analyses to justify their claim for a communications anomaly, Storm et al. (2010) argue that the results indicate that the ganzfeld experiments, especially those that follow their recipe, are coherent and consistent. They do not actually go so far as their colleagues Utts and Radin, who use meta-analysis to justify their claim that the ganzfeld experiment, along with other parapsychological experiments, is replicable. Utts (1995) has written, "Using the standards applied to any other area of science, it is concluded that psychic functioning has been well established" (p. 289). Radin (1997) has put the claim more forcefully: "We are forced to conclude that when psi research is judged by the same standards as any other scientific discipline, then the results are *as consistent* as those observed in the hardest of the hard sciences!" (p. 58, emphasis in the original).

This reliance on meta-analysis as the sole basis for justifying the claim that an anomaly exists and that the evidence for it is consistent and replicable is fallacious. It distorts what scientists mean by confirmatory evidence. It confuses retrospective sanctification with prospective replicability. An example from one of the failed claims of anomaly may be instructive. The French physicist René Blondlot announced his discovery of N-rays early in the 20th century (Stradling, 1907). During the period from 1903 through 1907, 325 reports on N-rays appeared in scientific journals. Of these, approximately 180 were accounts of experimental investigations into N-rays. During the first year and a half after Blondlot's announcement, 128 of the N-ray experiments were successful, and around 37 failed to find evidence for N-rays. It was not until 1905 that the number of unsuccessful experiments began to outnumber the successful ones. By 1907, it had become clear that almost no one could successfully replicate the original N-ray findings.

The claim for N-rays was rejected because the results were inconsistent and could not be reliably replicated. However, by the time that N-rays were clearly rejected, the accumulated database consisted of 180 successful and 100 unsuccessful experiments supporting the existence of N-rays. If someone decided to use a meta-analysis on this database, the combined effect size would have undoubtedly been significant. Of course, the scientific community would not consider such an outcome as evidence for the replicability of N-rays, nor should one accept that a significant effect size in a parapsychological database is evidence that the evidence is replicable. Required, of course, are demonstrations that the claimed evidence can be prospectively obtained by independent investigators, given appropriately designed experiments with adequate power.

Conclusions

My first attempt to critique parapsychological research (Hyman, 1957) motivated me to scrutinize actual experimental reports rather than general summaries. Such scrutiny takes time and effort. I was surprised to discover that parapsychologists were more statistically and methodologically sophisticated than critics had portrayed them. However, I also found disturbing examples of methodological oversights along with otherwise admirable controls. Perhaps the most puzzling experience came from my hours devoted to a detailed inspection of the experiments in the original ganzfeld database (Hyman, 1985). Even today I find it difficult to understand how parapsychologists could have tolerated so many obvious flaws in what was claimed to be their most successful database. I found the autoganzfeld experiments greatly improved in methodology over the original ganzfeld experiments. On the other hand, my careful analysis of the actual data from these experiments uncovered peculiar patterns that could possibly point to some subtle biases (Hyman, 1994).

To me the most bothersome aspect of parapsychological research during the century or so of its existence is its persistent inconsistency. During the past 50 years, I have become acquainted with many parapsychologists who agreed with my assessment. They were understandably distressed by this state of affairs. I have already referred to some contemporary parapsychologists who acknowledge the elusiveness and inconsistency of parapsychological evidence (Atmanspacher & Jahn, 2003; Bierman, 2001; Kennedy, 2001, 2003; Lucadou, 2001).

Bierman (2001) provides an account of one of the kinds of inconsistencies that frustrate parapsychologists:

> A rather spectacular example of the decline in effect size happened when in 1995 two independent groups, one from Durham, NC, the other from Amsterdam, published the data of the first part of their respective ganzfeld series (Broughton & Alexander, 1995; Bierman, 1995). The over-all hit rates were: 33% ($N = 100$) and 38.2% ($N = 76$). One year later the series were finished with the following astonishing figures for the second part: Durham 13.7% ($N = 51$) and Amsterdam 15.6% ($N = 32$) (Broughton & Alexander, [1997]; Wezelman & Bierman, 1997). Thus the results of the first and second part of both series differed significantly between the years while within the same year the groups replicated each other as if some outside factor in 1996 had influenced both groups to go from hitting into missing. . . . (p. 274)

For parapsychologists who believe in psi, such inconsistencies must be discouraging, indeed. Their typical remedy is to propose that such inconsistencies are an inherent property of psi. This not only begs the question but makes it impossible to prove the existence of psi within the framework of science. Science cannot

investigate a phenomenon that is inherently unpredictable and evasive.

In direct opposition to those parapsychologists who emphasize the elusive and inconsistent nature of their evidence, those such as Storm et al. (2010) believe that the evidence is consistent and can meet accepted scientific criteria, if only the scientific community can set aside its prejudices and fairly examine the data. I think it is relevant that the parapsychologists of this second group rely on meta-analysis for their arguments.

Meta-analysis can mask the underlying contradictions of the original experiments by providing a buffer between the original data and the investigators. Once the effect sizes are abstracted from the original data, the only information that is preserved is a dimensionless index of size. Such indices can display heterogeneity only as variations in effect sizes. When the appropriate statistics indicate that a collection of effect sizes is heterogeneous, the authors can make databases homogeneous by simply removing a sufficient number of the more extreme cases. They can further create consistency by combining databases from different sources whenever the combined effect sizes do not differ significantly. All these manipulations of the data, including finding significant effect sizes, can be done without any reference to the bothersome inconsistencies that abound within the actual studies.

In the final analysis, parapsychology will succeed in its quest to demonstrate its communications anomaly only when it can generate specific hypotheses that predict patterns of outcomes that are consistent, lawful, and independently replicable by parapsychologists and others. So far, careful assessment of the parapsychological literature does not justify optimism on this matter.

References

Atmanspacher, H., & Jahn, R. G. (2003). Problems of reproducibility in complex mind–matter systems. *Journal of Scientific Exploration, 17,* 243–270.

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

Bierman, D. J. (1995). The Amsterdam ganzfeld series III & IV: Target clip emotionality, effect sizes and openness. In *Proceedings of the 38th Annual Convention of the Parapsychological Association* (pp. 27–37). Durham, NC: Parapsychological Association.

Bierman, D. J. (2001). On the nature of anomalous phenomena: Another reality between the world of subjective consciousness and the objective world of physics? In P. Van Loocke (Ed.), *The physical nature of consciousness* (pp. 269–292). New York, NY: Benjamins.

Broughton, R. S., & Alexander, C. H. (1995). Autoganzfeld II: The first 100 sessions. In *Proceedings of the 38th Annual Convention of the Parapsychological Association* (pp. 53–61). Durham, NC: Parapsychological Association.

Broughton, R. S., & Alexander, C. H. (1997). Autoganzfeld II: An attempted replication of the PRL ganzfeld research. *Journal of Parapsychology, 61,* 209–226.

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

Hyman, R. (1957). Review of *Modern Experiments in Telepathy,* Second Edition. *American Statistical Association Journal, 52,* 607–610.

Hyman, R. (1964). *The nature of psychological inquiry.* Englewood Cliffs, NJ: Prentice Hall.

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

Hyman, R. (1994). Anomaly or artifact? Comments on Bem and Honorton. *Psychological Bulletin, 115,* 19–24.

Kennedy, J. E. (2001). Why is psi so elusive? A review and proposed model. *Journal of Parapsychology, 65,* 219–246.

Kennedy, J. E. (2003). The capricious, actively evasive, unsustainable nature of psi: A summary and hypotheses. *Journal of Parapsychology, 67,* 53–74.

Lucadou, W. V. (2001). Hans in luck: The currency of evidence in parapsychology. *Journal of Parapsychology, 65,* 3–16.

Radin, D. (1997). *The conscious universe: The scientific truth of psychic phenomena.* San Francisco, CA: HarperEdge.

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010). Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin, 136,* 471–485.

Stradling, G. F. (1907). A résumé of the literature of the N rays, the N_1 rays, the physiological rays and the heavy emission, with a bibliography. *Journal of the Franklin Institute, 164,* 57–74, 113–130, 177–189.

Utts, J. (1995). An assessment of the evidence for psychic functioning. *Journal of Parapsychology, 59,* 289–330.

Wezelman, R., & Bierman, D. J. (1997). Process oriented ganzfeld research in Amsterdam: Series IV B (1995): Emotionality of target material, Series V (1996) and Series VI (1997): Judging procedure and altered states of consciousness. In *Proceedings of the 40th Annual Convention of the Parapsychological Association* (pp. 477–491). Durham, NC: Parapsychological Association.

Received March 12, 2010
Revision received March 15, 2010
Accepted March 15, 2010 ■